Scientometrics (2015) 105:1577–1604
DOI 10.1007/s11192-015-1737-9

Modelling citation networks

S. R. Goldberg[1] · **H. Anthony**[2] · **T. S. Evans**[2]

Received: 11 January 2015 / Published online: 5 September 2015

Abstract The distribution of the number of academic publications against citation count for papers published in the same year is remarkably similar from year to year. We characterise the shape of such distributions by a 'width', σ^2, associated with fitting a log-normal to each distribution, and find the width to be approximately constant for publications published in different years. This similarity is not surprising, after all, why would papers in a given year be cited more than another year? Nevertheless, we show that simple citation models fail to capture this behaviour. We then provide a simple three parameter citation network model which can reproduce the correct width over time. We use the citation network of papers from the hep-th section of arXiv to test our model. Our final model reproduces the data's observed 'width' when around 20 % of the citations in the model are made to recently published papers in the entire network ('global information'). The remaining 80 % of citations are made using the references from these papers' bibliographies ('local searches'). We note that this is consistent with other studies, though our motivation to achieve the above distribution with time is very different. Finally, we find that, in the citation network model, varying the number of papers referenced by a new publication is important as it alters the parameters in the model which are fitted to the data. This is not addressed in current models and needs further work.

Keywords Complex networks · Directed acyclic graphs · Bibliometrics · Citation networks

Mathematics Subject Classification 91D30

✉ S. R. Goldberg
s.r.goldberg@qmul.ac.uk

[1] School of Physics and Astronomy, Queen Mary University of London, London E1 4NS, UK

[2] Centre for Complexity Science and Physics Department, Imperial College London, London SW7 2AZ, UK

Introduction

A citation network is defined using a set of documents as vertices with directed edges representing the citations from one document to another document. Examples include networks from patents (see Sternitzke et al. 2008; Clough et al. 2014; Clough and Evans 2014) and court decisions (see Clough et al. 2014; Clough and Evans 2014; Fowler and Jeon 2008) but in this paper we will work with data from academic papers (Solla Price 1965, 1976) and our language will reflect this context. Since citation networks capture information about the flow of innovations, understanding the large scale patterns which emerge in the data is of great importance.

The citation distribution, namely the number of papers with a given number of citations against citation count, is one of the simplest features and it has long been known to be a fat-tailed distribution (Solla Price 1965, 1976), a few papers garner most of the citations. See Perc (2014) for an overview. However, the focus of our work is the more recent observation that the shapes of these distributions are surprisingly stable over time (Radicchi et al. 2008; Evans et al. 2012; Goldberg and Evans 2012). This is a feature that many simple models (of the process of citation) fail to capture, for an example refer to model A in "Model A: The price model" section. Our aim is to produce a simple model which will enable us to understand the origin of this feature. In order to create a model close to the real world we will work with a real citation network derived from papers posted between 1992 and 2002 on the hep-th section of arXiv. We will use this data to provide both key input parameters for our models and as a test of the output from our models.

In "Analysis of hep-th arXiv data" section, we analyse the key features of our real citation network. Then, we define our models A–C in "Model A: The price model", "Model B: Time decay of core papers" and "Model C: Copying" sections. In each case we identify strengths and weaknesses, using the latter to provide motivation for improvements to the models. Finally, we conclude and discuss further work in "Conclusions" section.

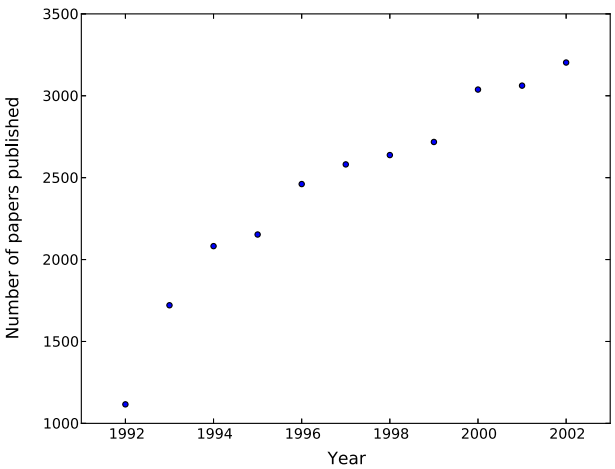

Fig. 1 This is a graph of the number of papers published in a given year against year for the hep-th arXiv dataset. An increase is clear, the number of papers published per year increases in an approximately linear fashion, it is not constant. The number of publications released from 1992 to 2002 approximately triples from 1116 to 3203 papers. When year to year comparisons are made from the hep-th data to our citation network models, it is important to incorporate literature growth with year

Analysis of hep-th arXiv data

In this section we will describe the data and analyse its key features. In doing so, we will define our notation and analysis methods. The data comes from the hep-th section (high energy physics theory) of the arXiv online research paper repository [11] citation network from 1992 to 2002. As this section only started in 1991, we suggest that the early data may have some initialisation effects. For example, the number of publications released per year increases rapidly at first, but after 1994 the rise is a more gentle one, see Fig. 1. Anecdotally, hep-th rapidly became the defacto standard for the field (it has always been free to both authors and readers with easy electronic access) and we suggest that essentially every paper in this field produced after 1994 is in our data set. This completeness gives us over 27,000 publications in our data, so we have enough information to draw useful statistical inferences. Different fields have a different number of papers published per year and different distributions for the number of references associated with these papers (Radicchi et al. 2008). By creating a citation network model associated with one field alone (hep-th) we eliminate any bias in our model due to field dependence. Overall, while all citation data sets miss citations to documents outside the data set, we are confident that the post 1994 parts form a reliable single field citation network. Another by-product is that our data is open source allowing independent verification; we took our copy from KDD Cup (2003).

An important feature of arXiv is that papers are given a unique identifier when they are first submitted. This identifier records the order in which papers are submitted within a particular section, a larger number indicates a later submission, and the year and month of this first submission are also simply encoded in the identifier. For instance arXiv:hep-th/9803184 was submitted in March 1998 just after arXiv:hep-th/9803183 but just before arXiv:hep-th/9803185. It is this first submission date which we take to define the publication date of each article.

Number of papers published per year

The increase in the number of publications released per year is shown in Fig. 1. This is expected intuitively and in the literature. Intuitively, because we live in a more highly connected era where an increasing number of publications are written. Many authors, e.g. Simkin and Roychowdhury (2007), Golosovsky and Sorin (2013), find that the number of papers published increases with time. We also find that the number of publications in our data set increases markedly over time, even if we ignore the initial couple of years corresponding to the start of the arXiv repository, as Fig. 1 shows. Therefore, when making year-to-year comparisons from the models to the hep-th data we compare the same number of publications in both cases to avoid any bias. For example, if we compared some quantity from a model in year 10 to the corresponding 10th year in the hep-th dataset, 2002, there may exist a bias due the hep-th citation network and model having a different number of papers published in that year. We treat this growth in paper numbers as a separate external aspect of the citation process and do not attempt to model this growth.

A measure of the hep-th citation network data may be biased near to 1992 because this is when then network started,[1] i.e. there may exist some unknown biases which are due to the initialisation of the network. These initialisation effects may require a settling down period before they stop biasing the measure of the network. The initialisation effect that

[1] We omit year 2003 from our analysis because it is incomplete.

could bias our measures is the number of papers published per year. From 1992 to 1994 the number of publications released per year increases rapidly, then approximately linearly after 1994, Fig. 1. This initial increase is explained by two effects: the effect of the arXiv becoming a more popular network to reference and the quantity of literature published per year increasing generally, irrespective of the citation database, due to increase in the spread of knowledge. After the arXiv had gained popularity, after 1994, only the second effect controls the increase. To eliminate the initialisation effect and the effect of the general growth in the number of papers published per year on the measures of the model and hep-th network we simply compare our models (of hep-th) to hep-th by comparing the first, second (and so on) 1000 papers, not the number of papers in each year of the hep-th network, Fig. 1.

In our models we create 28,000 nodes, the same number of nodes as the hep-th arXiv citation network (27,000 nodes) plus 1000 nodes. Then, we deleted the first 1000 nodes and directed edges to those nodes. This corresponds to deleting the first year (as 1000 papers corresponds to approximately 1 year, Fig. 1). This replicates the hep-th data.

A parameterisation of the fat-tail

The citation or in-degree distribution is one of the most fundamental features of any citation network and is the focus of our attention here. It is well known that these are invariably fat-tailed distributions with the tails often described in terms of power-laws. Such distributions are formed from all the citations between papers published in many different years and many different research areas.

Our principal concern is the general shape of citation distribution of papers published in the same field and in the same year; such distributions are fat-tailed which we characterise by a log-normal fit. In contrast, there are many others (below) concerned about the best way to characterise such distributions. All we require is a functional form of few parameters which we can use to fit these fat-tails, a form that has proven effective on real data elsewhere and which we can apply to our hep-th arXiv data. In this way we can capture the behaviour of these fat-tails, for papers published in the same field and year, in just a few parameters. Our aim is to study the time evolution of these tails. From this we can address our key question: what are the basic features required in a citation model in order to reproduce the correct time evolution of the citation distribution?

Our approach is to use a lognormal distribution to describe the tail of the citation distribution. The lognormal form has been used effectively in many studies of the fat tails of citation data (Radicchi et al. 2008; Evans et al. 2012; Goldberg and Evans 2012; Brzezinski 2015; Stringer et al. 2008, 2010). In general the lognormal has been shown to be one of the best fitting forms for many types of data with fat-tailed distributions, see (Stringer et al. 2010; Clauset et al. 2009; Newman 2010; Mitzenmacher 2004). We note there are criticisms of certain claims based on lognormal fits to citation distributions (Waltman et al. 2012). Indeed, many other forms have been used successfully to describe the fat-tailed citation distributions, for example power-laws (Brzezinski 2015; Redner 1998; Seglen 1992; Eom and Fortunato 2011), stretched exponentials (Laherraére and Sornette 1998; Wallace et al. 2009) and modified Bessel's functions (Raan 2001).

As this range of results suggests, it is difficult to find statistically significant differences in the quality of fits made by different functional forms to citation data. We expect different functional forms may give a reasonable description of the tail of the citation distribution (Brzezinski 2015; Clauset et al. 2009). However, all we require in our work is that the lognormal distribution is one form which captures the behaviour of the tail of the

citation distribution. Since we are working with real data, we have an explicit check that our procedure is reasonable. The debate on wider issues, such as which characterisation of the data's citation distribution fits best, does not invalidate our method.

We use the approach discussed in Evans et al. (2012); Goldberg and Evans 2012). First we select all the papers published in one period of time, typically one calendar year. We then calculate the citation count for this year, the in-degree for the corresponding vertex in the citation network. From this we get the average citation count for this subset of papers and we denote this as $\langle c \rangle$. Finally we fit the tail of the citation distribution, that is for $c > 0.1\langle c \rangle$ the statistical model for the number of papers with c citations is

$$n(c) = (1+A)N \int_{c-0.5}^{c+0.5} \mathrm{d}c \; \frac{1}{\sqrt{2\pi}\sigma c} \exp\left\{ -\frac{(\ln(c/\langle c \rangle) + (\sigma^2/2) - B)^2}{2\sigma^2} \right\}. \quad (1)$$

Here N, the total number of papers published by the end of our time period, and $\langle c \rangle$ are fixed by the data. This leaves us with three parameters to fit: A, B and σ. A perfect lognormal fit occurs when $A = B = 0$. As noted in Evans et al. (2012), if the number of zero and low cited papers, those with $c \leq 0.1\langle c \rangle$, follows a different distribution, then neither A nor B will be zero. This is typically found to be the case but this is not the focus of our work. The only output parameter we use is σ^2 which is simply used as a measure of

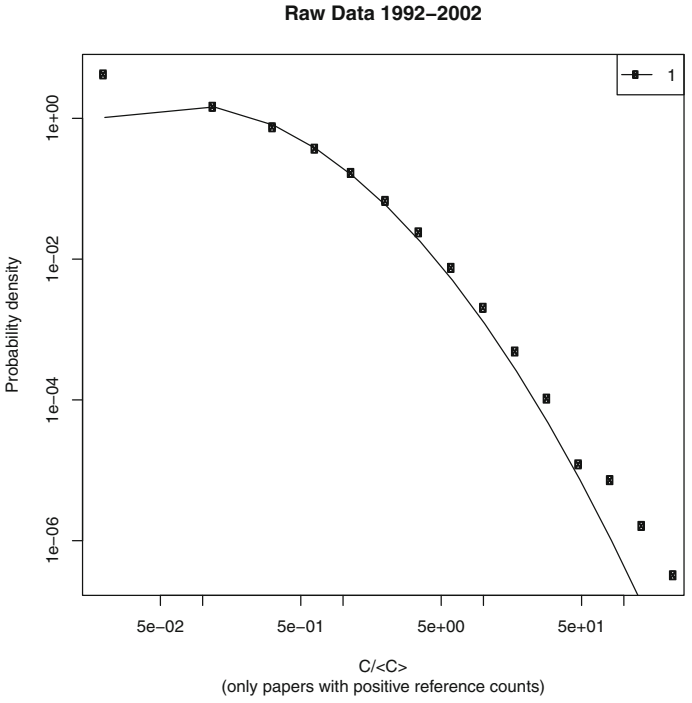

Fig. 2 The citation distribution for all papers in the hep-th arXiv repository between 1992 and 2002. The points are the counts from logarithmic sized bins and plotted in terms of $c/\langle c \rangle$ where $\langle c \rangle$ is the average citation count taken over all papers. The *line* shows the form of the best fitting lognormal curve, Eq. (1) (Evans et al. 2012; Goldberg and Evans 2012). The width of the distribution is $\sigma^2 = 1.78 \pm 0.14$, large compared to the $\sigma^2 \approx 1.3$ found by Radicchi et al. (2008) and Waltman et al. (2012)

the 'width' (tail) of the $\ln(c)$ distribution. Note that since we have a fat-tailed distribution, it makes sense to work in terms of the width of a $\ln(c)$ distribution.

It is important to note that σ is not the square root of the variance of either the citations c nor of the logarithms of citation count, $\ln(c)$. This is because we only fit part of the lognormal distribution to part of the actual citation distribution, namely we use reasonably well cited papers with $c > 0.1\langle c\rangle$ to estimate σ^2. By finding a value for σ^2 from the fitting procedure we merely get a measure of the range of citation values found in any data set.

By working with the normalised citation count $c/\langle c\rangle$ we account for one of the major differences between papers published at different times, namely that older papers have more citations. What we are focusing on is on the temporal evolution of the width of the distribution.

We first look at the fat-tail of the distribution for all the papers in our data set, shown in Fig. 2. We used logarithmic binning excluding zero cited papers.[2] We used the width of the in-degree distributions of papers published in the *same year* for years 1992–2002 as a measure to compare our model to the hep-th network. We find the width $\sigma^2 = 1.78 \pm 0.14$ of the entire hep-th arXiv data set, large compared with the literature where $\sigma^2 \approx 1.3$ (Waltman et al. 2012). For different fields some widths are simply larger than others because there is a larger spread in citation count. σ^2 has not been measured for this data set before therefore this high σ^2 is not unexpected. This difference does not effect our model building: we simply want to build a model that reproduces the data's fat tail over time. One can clearly see that the lognormal is a good way of characterising this tail because it is a very good fit of the data, Fig. 2.

If we now we fit our lognormals to the in-degree distribution for papers deposited on arXiv in same calendar year, we find a reasonable fit. Moreover, we confirm the results of Evans et al. (2012) that these fits are approximately the same with our measure of the width of the distribution, σ^2, remaining roughly constant. In Fig. 3 most of the error bars lie within one standard deviation of the $\sigma^2 = 1.78 \pm 0.14$ for the *whole* data set, we use this as an 'average' of the distribution in Fig. 3. At the very least there is no evidence for any systematic change in the width over time. The challenge now is to find a simple model which can reproduce this effect.

Zero cited papers

The primary focus of our work is to ensure that we get the correct behaviour for the width of the tail of the citation distribution, as captured by our σ^2 parameter. However, there are large numbers of low cited papers not included in that aspect of our analysis (based only on papers with $c > 0.1\langle c\rangle$). To make sure our models give a realistic result for the low cited papers we simply look at the proportion of zero cited papers in the citation network, z, produced in each model. In previous work this has been calculated and can be up to 40 % of the citation network data analysed, refer to Goldberg and Evans (2012) and Waltman et al. (2012). We note that our approach is similar to many other analyses of fat-tailed distributions where the low valued part (low citation count in our case) is typically excluded, for example see Clauset et al. (2009).

[2] The bin scale was chosen to ensure there were no empty bins below the bin containing the highest citation values.

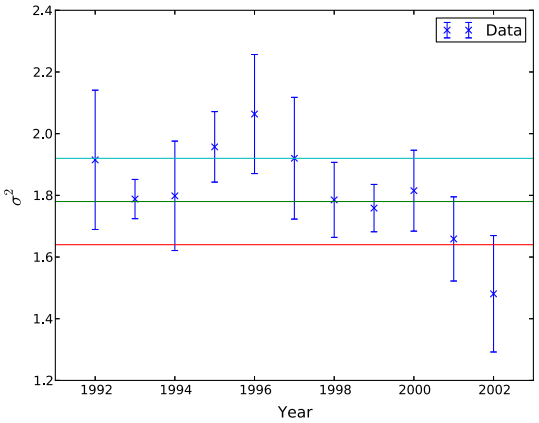

Fig. 3 The σ^2 plot is a plot of σ^2 against year for the hep-th data. To calculate this we plot 11 in-degree distributions (corresponding to citations gained by papers published in years 1992–2002), each against the normalised citation count. We fit a lognormal to each of these distributions. Each lognormal gives a signature width of the distribution, σ^2. We find the σ^2 and their error bars for each year are all consistent with a constant σ^2. In particular, the σ^2 for each year are consistent with the constant value of 1.78 ± 0.14 obtained for the σ^2 for the entire network, shown here by the *horizontal lines*. The last 2 years are lower than the rest because they have three times as many publications as years 1992 or 1993 and they pull the average down. 2002 has a σ^2 much lower than other years and the trend from 2001 is downward. This is because the dataset used half of 2003 (we analysed up to 2002 and omitted the last year because it was incomplete) therefore the last year or so had very little time to gain citations, less than a year. Hence, most papers had few citations and the spread, σ^2, was smaller

Length of bibliographies

While the citation count or in-degree of a paper is an important measure, for citation network models we also need to understand the relationship between this distribution and the length of the bibliographies (the out-degree of publications). The out-degree distribution is again a fat-tailed distribution though not nearly as broad as the in-degree one, Fig. 4, and is noted in the literature (Vázquez 2001). We find that a lognormal distribution gives a reasonable description of the out-degree data (Fig. 21 in the "Appendix").

Many models choose the simplest approach and use a constant out-degree for all publications created in the model (Simkin and Roychowdhury 2007; Zhu et al. 2003). Another approach is to duplicate the out-degree distribution of the data set, see Ren et al. (2012). As we show in "Lognormal out-degree distribution" section, changing the distribution of the length of bibliographies (number of references made by a paper), see Fig. 4, alters the results of the final model—so this is an important point. Refer to Goldberg (2013) for more discussion. In particular we found that our model only gave good fits for the citation distribution when we used a normal distribution to the out-degree distribution, with a mean 12.0 and standard deviation 3.0 equal to that found from the hep-th data.

We now turn to look at how three different models perform.

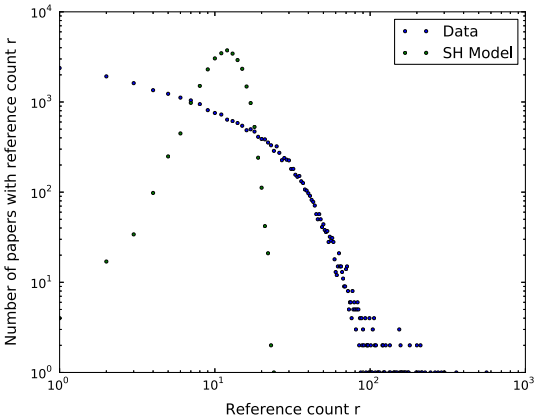

Fig. 4 This is a log–log plot of the out-degree distribution: the number of papers which make r references (reference count) against reference count r. The data is plotted in *blue*. Superimposed in *green* is a plot of the out-degree distribution used in our models, a normal distribution with the same mean 12.0 and standard deviation 3.0 as the whole hep-th data for the same number of publications, 27,000. This normal distribution used in our models is clearly not a good fit for the out-degree distribution found in the hep-th data. (Color figure online)

Model A: The price model

An obvious model to start from is the Price model of cumulative advantage (Solla Price 1965, 1976). While the limitations of such a model are well known (see Smolinsky et al. 2015 and our discussion in the next section) we wish to show that this simple model also fails in terms of having a constant width of the citation distribution for papers published in the same year for different years. This section will also define aspects of our method used in our later more sophisticated models.

In the Price model new papers reference existing publications in proportion to their current number of citations (cumulative advantage) and this is well known to generate a fat-tailed distribution for citations. Our version, model A, is based on a generalisation of the original model of Price, see ch. 14 of Newman (2010) for further discussion and references. At each step one new publication is added to the N pre-existing publications in the network. The number of papers in its bibliography, r, is chosen from a normal distribution, described above. The new publication then makes references to different pre-existing publications in one of two ways: with probability p it uses cumulative advantage (referencing a publication chosen in proportion to its current citation count) otherwise with probability $(1-p)$ it chooses uniformly at random from all existing publications. References are chosen repeatedly until r distinct pre-existing papers have been chosen and these then define the r new links in the citation network. The probability of making a link to an existing node i is $\Pi_A(i)$ where roughly[3]

[3] There is a small correction to this form due to the possibility of that the same vertex i could be chosen more than once which is excluded in the actual model.

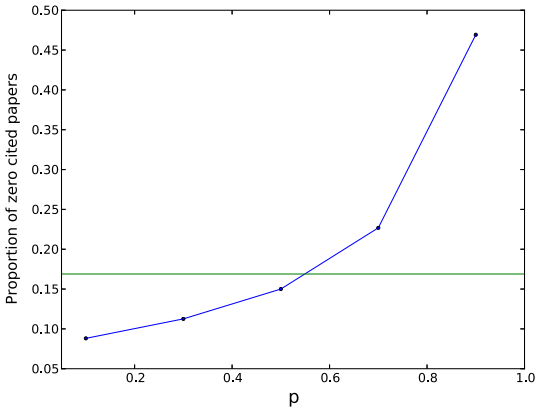

Fig. 5 The *blue curve* is a plot of the proportion of zero cited papers, z, in model A for a given probability of cumulative advantage p, against varying p. As expected, as p decreases uniform random attaching increases, this gives more opportunity for zero cited papers to attach. The *green horizontal line* is a constant line where the z is 0.169, the proportion of zero cited papers in the hep-th arXiv data set. Where the *lines cross* is where the proportion of zero cited papers in model A is equal to that of the data. This occurs at the desired p, 0.55. (Color figure online)

$$\Pi_A(i) = \frac{pk_i^{(\text{in})}}{\langle k^{(\text{in})} \rangle N} + \frac{(1-p)}{N}. \tag{2}$$

Here $k_i^{(\text{in})}$ is the in-degree of node i, $\langle k^{(\text{in})} \rangle$ the average of the in-degree and N is the total number of papers.

Our starting configuration is no network. The first node decides to make r_1 references (according to the normal distribution), but as the first node has nothing to reference a new node is created. This second node decides to make r_2 references (via the normal distribution, above) and associates a probability to the first node according to Eq. (2), however, it may not reference the first node. This second node cannot make any more references and therefore a new node is created and so on. The first few nodes in the network will have very little choice in what to reference but these initialisation effects will only effect the first papers. In any case we do not consider the first 1000 papers created in any of our models (see "Number of papers published per year" section) so this initial condition has no effect on any of our results. Note that this initialisation effect is present in the arXiv data too, therefore we ignore the initial publications in the hep-th data. Also, note that the Price model implies the papers with the very largest citation count are the oldest (Newman 2010; Dorogovtsev et al. 2000).

To determine the parameter p we use the observation that the proportion of zero cited papers over all years in hep-th is $z = 0.169$. We find how the total fraction of zero cited papers found in model A varies with p, as shown in Fig. 5. We then chose the parameter $p = 0.55$ so that model A gives the same number of zero-cited papers overall as found in our data.[4]

[4] The full Price model may be solved exactly within the mean-field approximation in the infinite time limit. Those solutions are very close to numerical results found for finite sized simulations such as ours. For $\langle k^{(\text{in})} \rangle = 12.0$, these formulae give $z = 0.156$ if $p = 0.55$ and we find we need $p = 0.59$ in order to get the same value of $z = 0.169$ found in our data. However, we remove the first thousand papers created in our simulation so we do not expect an exact match with the theoretical expressions.

We then observe the fat-tail of the citation distribution from model A: the number of papers with a given normalised citation count against normalised citation count (the normalised citation count is simply the citation count divided by the average citation count in the network).

Firstly, we analyse the citation (in-degree) distribution *for all years* for which model A and cumulative advantage models are designed to give the expected long-tail. We do indeed find a fat tail, the overall width of this tail, Fig. 6, is $\sigma^2 = 1.05 \pm 0.18$. This is significantly different to the data's $\sigma^2 = 1.78 \pm 0.14$, Fig. 6.

We are more interested in the shape of the in-degree distribution *for papers published in the same year* for different years. Using $p = 0.55$ we plot the in-degree distribution graphs with a lognormal fit for each year which gives the parameter σ^2 describing the tail. The σ^2 value associated with each year's lognormal fit is plotted against year, the 'σ^2 plot', is shown in Fig. 7. For a good model we need constant $\sigma^2_{\text{Model}}(t) \approx \sigma^2_{\text{data}}(t) = 1.8$ (from the data's σ^2, in green, Fig. 7). Model A is not consistent with the hep-th data because $\sigma^2_{\text{ModelA}}(t) \approx 0.33 \pm 0.15 \neq 1.78 \pm 0.14$. Also, $\sigma^2_{\text{ModelA}}(t)$ seems to have a downward trend for older papers. In fact, we find that for all p values ($0 < p \leq 1$) in model A the σ^2 plots do not differ significantly from one another and therefore were very similar to Fig. 7.

Why is $\sigma^2 \approx 0.33 \pm 0.15$ *so small* for model A? This is because in our citation network the oldest papers (e.g. papers in years 0 and 1) will *all* gain many citations through cumulative advantage (if $p \neq 0$) as they have been around the longest and will have had more chances to accumulate citations. In fact, our model shows that generally the oldest papers will be those in the fat-tail with many citations each. In terms of the width σ there is relatively little variation around the mean. Likewise the youngest papers (e.g. papers in years 9 and 10) will lose out and all will have a similarly low number of citations. The principle is the same for all years with relatively little variation around the means in the citation counts of papers published in the same year. Thus the Price model and the variation used for our model A is likely to have little variation and low σ^2 values for all years.

Our simple cumulative advantage model A with best parameter $p = 0.55$ does produce an overall fat-tailed in-degree distribution for all years Fig. 6 (although not quite fat-tailed enough because model A's overall $\sigma^2 = 1.05 \pm 0.18$, Fig. 6, is not within the data's overall $\sigma^2 = 1.78 \pm 0.14$, Fig. 2). It also produces a constant σ^2 plot, see previous section. However, model A does not produce the long-tail for *individual* years. The average σ^2 with time for model A 0.33 ± 0.15 does not lie within the data's average σ^2 with time 1.78 ± 0.14. Note that no matter how we changed p, we could not get our model A to give a σ^2 similar to that of hep-th. Therefore, to increase the σ^2 values (over different years) we need a new parameter.

Model B: Time decay of core papers

One well known problem with the Price model and our model A is that cumulative advantage gives the oldest papers too much of an advantage. That is, all the papers in the oldest years tend to have a large citation count so that there is too little variation in their citation counts. So for our second model, model B, we suppress the probability of adding a citation to an older paper, a feature often known as 'ageing' in the literature.

Looking at our data, Fig. 8 shows how citations are gained over time for papers published in the first (1992) and fourth (1995) years of our hep-th data. This shows a general

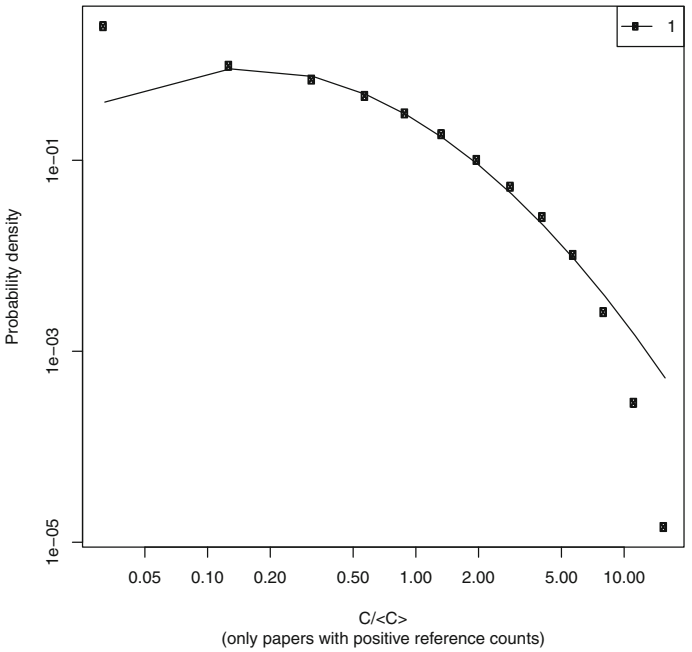

Fig. 6 This is the usual in-degree distribution (for normalised citation counts) for all years with a lognormal fit for model A. The axes are log–log, probability density against normalised citation count. We observe the long-tailed in-degree distribution goes up to approximately 20 (120 citations). The *corresponding plots* for the hep-th data goes up to 50 (500 citations), Fig. 2. This implies the width is not large enough. Another factor needs to increase the σ^2 values, next section

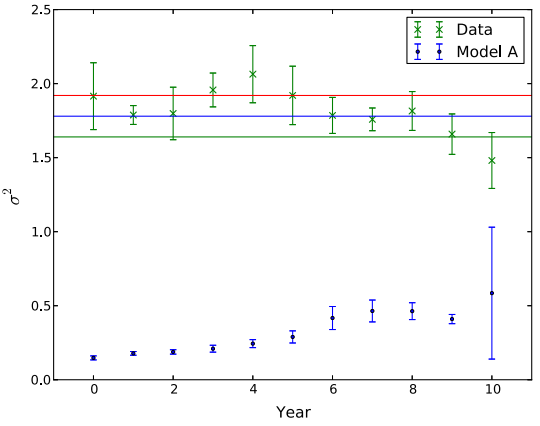

Fig. 7 This is the σ^2 plot, the plot of σ^2 against year for both the data, in *green*, and model A, in *blue*. σ^2 is the width associated with the lognormal fitted to the in-degree distribution for papers published in the same year. On this plot the data's years from 1992 to 2002 are relabelled years 0–10 respectively. The data's average and one standard deviation are plotted as *horizontal lines* 1.78 ± 0.14. Model A and hep-th's σ^2 values are very different. (Color figure online)

decay in the rate at which citations are accumulated, also found in Pollmann (2000). In order to keep our model simple we will use an exponential decay form in model B, giving it an additional parameter over model A. Such exponential decays are widely used in citation modelling (Eom and Fortunato 2011; Zhu et al. 2003; Ren et al. 2012; Dorogovtsev et al. 2000; Dorogovtsev and Mendes 2000, 2001; Hajra and Sen 2005, 2006;

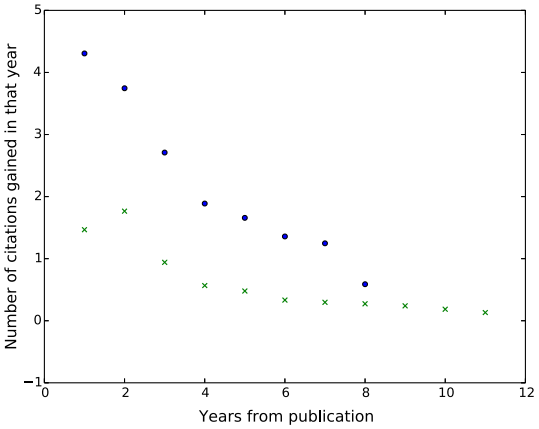

Fig. 8 The average number of citations gained each year for hep-th arXiv papers published in 1992 and 1995, shown as crosses and dots respectively. The *horizontal axis* gives the number of years since publication, the *vertical axis* is the number of citations gained in that 1 year. The first 2 years of the data, 1992 and 1993, have anomalous distributions with a peak in year 2, see "Number of papers published per year" section. For later years, e.g. 1995 onwards, we find a sharp decrease which may be characterised using an exponential decay. The average number of citations accumulated in 1 year falls by around a half over 3.5 years or around 5000 papers, Fig. 1

Maslov and Redner 2008; Geng and Wang 2009; Wu et al. 2014). Alternatives, such as only referencing papers if they are less than a year old, also creates an effective time decay (Simkin and Roychowdhury 2007). A similar decay over time in attachment probability has been used for other types of citation networks such as patents (Bentley et al. 2004).

In our model B we add new publications one at a time as before. With probability p the new publication chooses to reference an existing paper i chosen with a probability proportional to the current in-degree of paper i multiplied by an exponential decay factor dependent on i's age. Alternatively, with probability $(1-p)$ a papers are chosen with uniform probability multiplied by the same exponential decay factor. Thus the probability of attaching to node i, $\Pi_B(i)$, is roughly[5]

$$\Pi_B(i) = p\left(\frac{k_i^{(\mathrm{in})} 2^{(N-i)/\tau}}{Z_{\mathrm{B,ca}}}\right) + (1-p)\left(\frac{2^{(N-i)/\tau}}{Z_{\mathrm{B,ua}}}\right), \tag{3}$$

where $k_i^{(\mathrm{in})}$ is the in-degree of node i, τ is defined as the 'attention span' parameter in the *model*,[6] this is a time decay parameter in the model that acts like the time it takes for a paper to gain half the total number of citations it ever will, when $p = 0$). N is the total number of pre-existing nodes a publication can reference. Note that we are using the 'rank' time to determine the age of the paper, that is paper i is added at a time equal to i.

The normalisation factors, $Z_{\mathrm{B,ca}}$ and $Z_{\mathrm{B,ua}}$, are

[5] Again there is a small correction to this form to allow for the fact that we do not allow the same vertex i to be chosen more than in model B.

[6] This is different from the 'half-life' values referred to later, which are *measured* from the data.

$$Z_{\mathrm{B,ca}} = \sum_{j=1}^{j=N} k_j^{\mathrm{in}} 2^{-(N-j)/\tau}. \tag{4}$$

$$Z_{\mathrm{B,ua}} = \sum_{j=1}^{j=N} 2^{-(N-j)/\tau} = \frac{1-2^{-N/\tau}}{1-2^{-1/\tau}}. \tag{5}$$

Thus our model B has just two parameters: p and τ. To determine p we use the fraction of zero cited papers found in the whole of our hep-th data, $z = 0.169$. With a fixed attention span we vary p and calculate z for each of these models. We worked with a value of $\tau = 2000$ paper while determining p, see Fig. 9. From this we find $p = 0.80$. We found that this result for p does not change significantly for different attention span values τ. For example, for attention span values of $\tau = 200$ and $\tau = 5000$ papers we need to choose $p = 0.81$ and $p = 0.79$ respectively to get the correct zero-cited paper fraction, a mere 1.25 % change from our chosen value of $p = 0.80$.

Given $p = 0.80$ we now wish to determine the best half-life value for model B. To do this we choose the value of the attention span parameter τ such that *measured* half-lives of the hep-th data are as close as possible to the *measured* attention span outputted by the model. This process of trying to match the half-life derived from the output of a citation network model to that observed in data is original. However, note that there is no guarantee that the measured attention span parameter from our model will be identical to the half-life $T_{1/2}$ measured from the decay in citations seen in the data. This is because τ and p cause recent papers to be more and less likely to be referenced, respectively, therefore the *measured* half-life from the model $T_{1/2}$ is not equal to the inputted attention span τ. Our approach to find the inputted attention span of our model is as follows.

We first take each paper in hep-th and find the median citation time (or median half-life) T_{med} for that paper in a given year. This is the time it took for a paper in the nth year (where $1 \leq n \leq 11$) to accumulate half of its final citation count such that is final citation count is the sum of all of it's citations at the end of the 11th year. Histograms of these median citation times were considered for papers published in the same year, with examples of these distributions shown in Fig. 10. Note that the median citation times are smaller for 2001 than for 1992 because the 2001 papers had only about 2 years to gain citations. The time a 2001 paper takes to reach half of its total citation count (counted at the end of 2002 in our data) is small.[7]

We will just use the average median citation time for each year y in our data set, $\langle T_{\mathrm{med}} \rangle_y$, to characterise when a given year's papers have gained half of the citations they're ever going to gain, by the 11th year. For an exponential model, the total number of citations gained over a period T is proportional to $\int_0^T dt\, 2^{-t/T_{1/2}} = (1-2^{-T/T_{1/2}})(T_{1/2}/\ln(2))$. Since the number of citations at the median time T_{med} is half that of the total time T_{tot} available for a paper to collect citations in a given data set, we have that

[7] Note this is why we call it the *median half-life*: if you plot the number of citations gained by a paper against year and take the median, this is the value we call the median half-life T_{med}. Thus the value of T_{med} for a given paper increases over time. We call it the *median half-life* not a *half-life* because the *half-life*, as defined in a process with an exponential decay, is a fixed value, only equal to our median half-life in the limit of an infinitely old paper. We expect any estimate of the half-life of a paper's citations to be roughly constant whereas our *median half-life* T_{med} measurement varies from year to year, increasing until it reaches the formal half-life value.

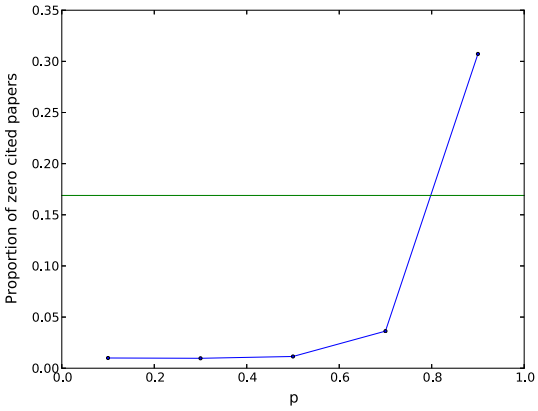

Fig. 9 The *blue* data points show the z found for model B for a fixed τ of 2000 papers and given probability of cumulative advantage p, against varying p. The *green horizontal line* gives the value for z found in the hep-th data, namely $z = 0.169$. Interpolating form the model output suggests a value of around $p = 0.80$ is needed. (Color figure online)

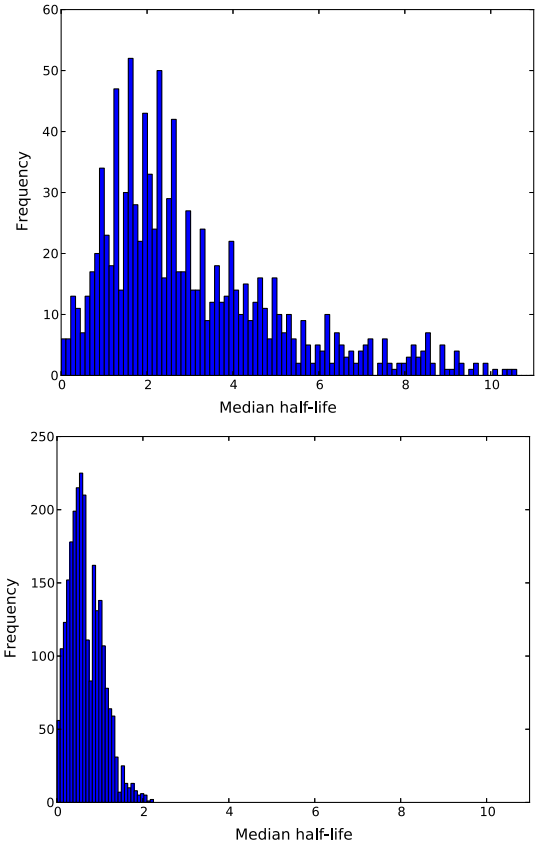

Fig. 10 These are two histograms to contrast. For an early year, above, $y = 1992$, and a late year, below, $y = 2001$ we plot the number of publications created in year y with a given median half-life T_{med} against the median half-life value. Both histograms have the same binning. They are similar because they both follow an approximate skew normal distribution with a mean of 2 and 0.5 years for 1992 and 2001, respectively

$$2^{1-T_{\text{med}}/T_{1/2}} - 2^{-T_{\text{tot}}/T_{1/2}} = 1. \tag{6}$$

Using the average median time $\langle T_{\text{med}} \rangle_y$ for papers published in 1 year of our hep-th data we solve (6) numerically to find a data half-life value $T_{1/2}$.

Figure 11 shows that the half-lives characterising the data, $T_{1/2,\text{data}}$, decreases for later years. This is because the number of papers published per year is increasing with time. This implies the number of citations gained per paper decreases with time because there is more literature to read. To factor this growth rate out, we can work in 'rank time'. That is, since arXiv automatically provides the order in which papers were first submitted, we use this order as a time parameter, so the earliest paper in the data has rank time 1, the n-th paper submitted has rank time n. In this case we define our 'years' to be collections of 2000 papers, so year 0 contains the first 2000 papers, year 1 has papers with rank times 2001 to 4000, and so forth. This gives us roughly the same number of years as the calendar years in our data. Working out the data half-life $T_{1/2,\text{data}}$ using rank time we find this is now roughly constant, see Fig. 11, confirming our suggestion that the number of papers in each year is an important factor here.

In order to set the model parameter τ we now try different values of τ for model B and use the output from the model to determine a model output half-life $T_{1/2,\text{model}}$ in exactly the same was as we did for the arXiv data, working now in rank time. To find the best τ parameter value we minimised χ^2 where

$$\chi^2(\tau, p = 0.80) = \sum_y \left(T^{(y)}_{1/2,\text{data}} - T^{(y)}_{1/2,\text{model}}(\tau, p = 0.80) \right)^2. \tag{7}$$

and $T^y_{1/2,\text{data}}$ and $T^y_{1/2,\text{model}}$ (calculated above) are the average half-lives outputted of the data and model B (for $p = 0.80$ and given τ) for rank year y, respectively. Figure 12 shows the results with a minimum χ^2 found for $\tau = 2000$ papers.

As a final check, Fig. 11 shows the half-lives $T_{1/2}$ found from the arXiv data and from the output of model B with its optimal input parameter values $p = 0.80, \tau = 2000$, both in normal time and rank time. Observe that model B does not fit the data well (right). We also note our earlier assertion that the *internal* attention-span parameter τ, which is set here to our optimal value of 2000 papers, does not need to match the *measured* half-life value derived from the output because they are different quantities. The processes associated to

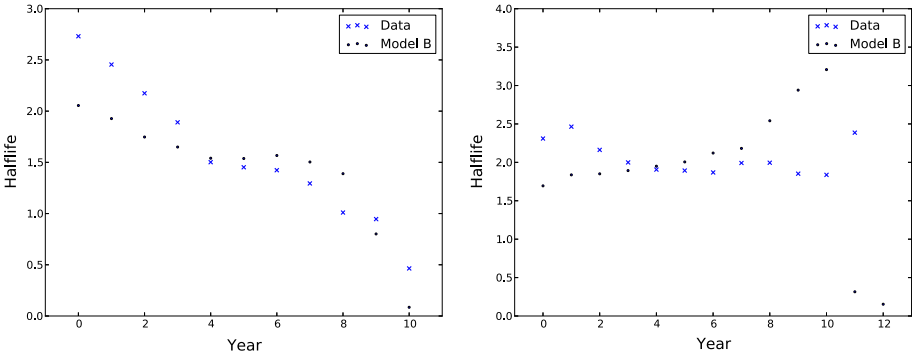

Fig. 11 In these figures we compare/contrast the measured half-life $T_{1/2}$ from the hep-th data (*crosses*) and model B using the best parameter values of $p = 0.80$ and $\tau = 2000$ (*dots*), for normal time (*left*) and rank time in years (*right*). On the left time is taken to be the calendar year, on the right time is in rank time (where 1 year equals 2000 papers). As expected, rank time implies the data's half-life is roughly constant (*right*). We also observe the similarity between model B and the data for the first 0–9 years and the discrepancy in later years. The latter implies model B is not a good model of the data

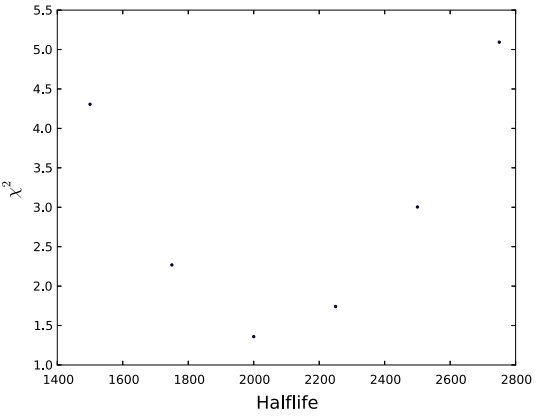

Fig. 12 This is a plot of χ^2, see Eq. 7, against varying input parameter τ, measured in rank time (number of papers), constant $p = 0.80$ in model B. There is a clear minima, minimum difference, between model B and the hep-th data when $\tau = 2000$ papers. Therefore a half-life of 2000 papers is chosen as our final τ. The minima approaches zero but is non-zero, meaning the half-life versus time plot for the best model B is close to but not exactly equal to the data, Fig. 11

our τ and p parameters cause recent papers to be more and less likely to be referenced, respectively, ensuring that half-life *measured* from the model output is not equal to the attention span parameter which inputted into the model. For instance, using rank time, the measured half-lives are around 1.5–2.0 years, so 3000–4000 papers. As explained above,

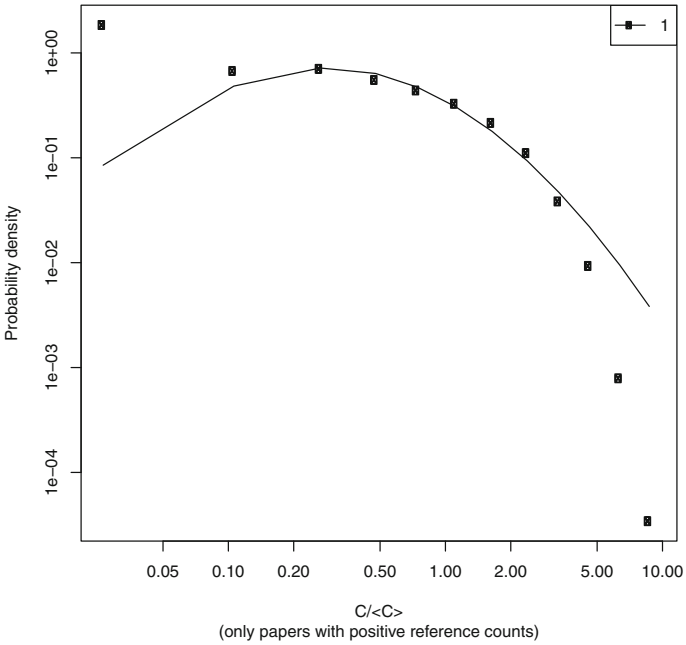

Fig. 13 This is the overall in-degree distribution of probability density against normalised citation count for model B with best parameters $p = 0.80$ and $\tau = 2000$ papers. This σ^2 of this distribution is 1.05 ± 0.18. Compared to model A, Fig. 6, model B has more papers with high citation counts (the mid part of the distribution) making the tail longer and therefore increasing σ^2 (although the distribution maximum only goes up to approximately 10 citations like model A, Fig. 6). However, compared to the data, Fig. 2, the tail of model B needs to be longer, it needs to have a non-zero probability of having a normalised citation count greater than 50

cumulative advantage favours older papers which extends the time scale measured from the output compared to the time-scale used as an input parameter.

Results for Model B

We can now look at the shape of the citation distribution for model B with the optimal parameter values, $p = 0.80$ and $\tau = 2000$ papers. We plot the in-degree distribution for the whole data set and observe the characteristic large width, Fig. 6 as expected by the hep-th data, Fig. 2. The width of the in-degree distribution is a slight improvement from that found with model A, changing from $\sigma^2 = 1.05 \pm 0.18$ in model A to $\sigma^2 = 1.14 \pm 0.17$ in model B, closer to the arXiv data's $\sigma^2 = 1.78 \pm 0.14$. However, as our error estimates show, this is not *significantly* better statistically. In fact we tried varying the parameters p or τ but found no way to increase the σ^2 significantly in model B. See Fig. 13 for the in-degree of all years for model B.

For model B, the width of the citation distribution for papers in each year with its optimal parameter values, $p = 0.80$ and $\tau = 2000$ papers, are shown in Fig. 14. We find that model B is now producing a distribution with a roughly constant width σ^2. This is an improvement on model A with its decreasing width with age. Unfortunately, model B is producing a width that is about half that seen in the data. The σ^2 values associated with model B's σ^2 plot (Fig. 14 shows most values are between 0.7 and 0.9) are better than those found in model A (typically averaging at 0.33 ± 0.15 as seen in Fig. 7) because model B's values are closer to those found in the hep-th data (most values between 1.7 and 2.0 in Fig. 14). However, no amount of variation in the values p or τ in model B gave a significant increase in the σ^2 plot. The model B σ^2 plot values always lie well outside the

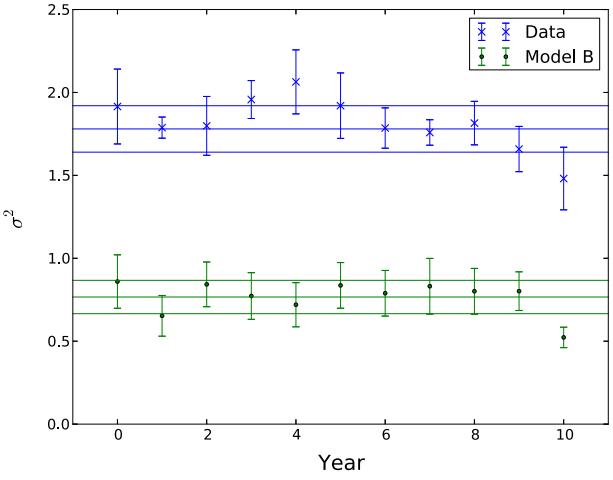

Fig. 14 This is a σ^2 plot of the data, *blue crosses*, and model B for $p = 0.80$ $\tau = 2000$ papers, *green dots*. The average, *upper and lower bounds* (one standard deviation away form the mean) of the data and model B at 1.78 ± 0.14 and 0.77 ± 0.10 in *blue* and *green*, respectively. We observe model B is not a good fit of the data as the data's average σ^2 is approximately 2.3 times the corresponding average for model B. No variation in the parameters p or τ altered the σ^2 plot of model B significantly. We can conclude from this graph model B is not a good citation network model of the hep-th data. The signature width is not large enough, however, it is constant. (Color figure online)

range of uncertainty in our values for the σ^2 obtained from the actual data in any one year. Therefore we need to add a further parameter to model B in order to reach a wider range of citations in each year, giving larger σ^2 values.

Model C: Copying

A general problem with the Price model and related models, such as our models A and B, is that the cumulative advantage and random attachment processes require global knowledge of the whole network. This can be seen in the normalisation of the two contributions in Π_A of (2): the number of citations for the cumulative advantage process and the number of papers for the uniform random attachment process. Neither of these is needed or known by authors looking to cite new papers. The addition of a time decay in model B means that the emphasis is on a smaller set of recent papers, something authors are more likely to be aware of, but the normalisations in Π_B (3) indicate that authors still require global knowledge if the processes are taken literally. There is, however, a very natural process based on local knowledge which reproduces the long tails and that is to use random walks (Simkin and Roychowdhury 2007; Vazquez 2000; Krapivsky and Redner 2001; Vázquez 2003; Chung et al. 2003; Saramäki and Kaski 2004; Evans and Saramaki 2005). That is an author will find papers of interest by following the references of papers already known to the author. In terms of a model, we will assume that authors find new papers of interest by choosing uniformly at random from the references listed in one paper; authors are making a

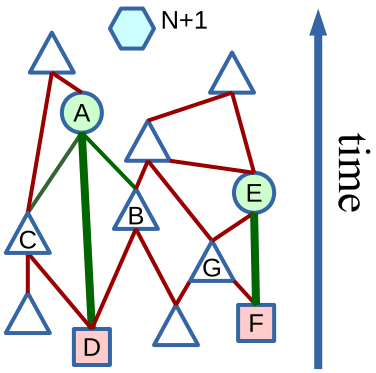

Fig. 15 An illustration of the processes in model C. A new paper $(N+1)$, the *blue hexagon*, is added to the network. Then the length of the bibliography r is chosen from a normal distribution of the same mean and standard deviation as the arXiv hep-th data. Suppose $r = 4$ here. A first 'core' paper, paper A (*green circle*), is then chosen exactly as in model B, using global information from all previously published papers (*triangles, circles and squares*). Next, all the papers cited by paper A, are considered and each is added to the bibliography of the new paper $(N+1)$ with probability q. These are our 'subsidiary' papers in the reference list of the new paper $(N+1)$. In this case paper D (*red square*) is chosen as indicated by the *thick green link* while the *thin green links* indicate that neither paper B nor C (*white triangles*) are chosen this way. Now a second core paper is chosen via global information (model B), say paper E (*green circle*), and added to the bibliography of the new paper $(N+1)$. We again consider the papers cited by paper E, selecting them with probability q. Here paper F is considered first and is selected becoming our second subsidiary paper citation. At this point we have the four papers needed for the new paper so we do not consider paper G. The new paper cites two core papers, A and E (*circles*), and two subsidiary papers, D and F (*squares*). The new citation network shown in Fig. 17. (Color figure online)

random walk on the citation network. To do this, an author need only use the the local information available in currently known papers. The properties of the rest of the network are irrelevant to this process as indeed they will be for real authors. What is particularly interesting, is that random walks of any length or type, even of length one, will generate the fat-tailed power-law like distributions with a wide range of characteristics (Evans and Saramaki 2005).

In looking to improve upon model B, we will therefore add a local search process based on a single step random walk to find some of the citations to be added to a new paper. We will start these walks from papers chosen (and cited) using the same mechanism as model B, given the partial success obtained there. That is we are assuming that some citations are derived using global knowledge and some from local searches. This is meant to mimic the fact that authors do come to a new paper with some limited knowledge of the whole network, obtained by looking at recent posts on arXiv or from conversations with colleagues and so forth, while local searches will also reveal new relevant material to an author.

Model C is defined as follows, see Fig. 15. A new paper, paper number $(N+1)$, is added to the network and the length of its bibliography, r is chosen randomly via a normal distribution as before. Next, we use the model B processes, Π_B of (3), to find the first 'core' paper to be cited (through global knowledge of the entire network). We then look at *each* of the papers cited by this first core paper, adding each of these 'subsidiary papers' to the bibliography of the new paper $(N+1)$ with probability q (through local knowledge of the bibliography of this core paper). We will then repeat this whole process, choosing another 'core' paper c' using the global knowledge processes of model B, followed by more 'subsidiary' papers found using the local knowledge process of a one-step random walk from the new core paper. Papers will only be cited once and the whole process stops as

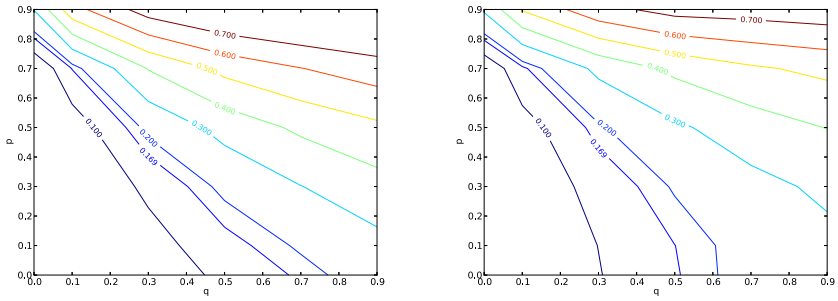

Fig. 16 The fraction of zero cited papers found in the output from model C. The *lines* represent different fractions of zero cited papers z_c found for varying (p, q) values for a fixed attention span (defined below) of $\tau = 200$ papers on the *left* and $\tau = 2000$ papers on the *right*. The *line* of constant z_c equivalent to the hep-th data's z value, $z_{\text{data}} = 0.169$. There are a family of (p, q) solutions ranging from $(p, q) = (0.8, 0)$ to $(p, q) = (0, 0.45)$, one of the *purple contour lines*. Note that $(p, q) = (0.8, 0)$, is simply model B, i.e. there are no q steps only a mixture of cumulative advantage and uniform attachment with an overall time-decay factor. For the 'unreasonable' value of $\tau = 2000$ on the *right*, the solution of $(p, q) = (0.0, 0.80)$ for $z_{\text{modelC}} = z_{\text{data}} = 0.169$ is the same as in contour for $\tau = 200$ papers. Along the q axis the solution for $z_{\text{data}} = 0.169$ when $\tau = 2000$ is $(p, q) = (0, 0.51)$, as opposed to (0, 0.66) for $\tau = 200$. Although the attention span τ has changed by a factor of 10, q has only changed by 0.15 absolutely. For $q < 0.3$ this contour can be treated as approximately equivalent to that for $\tau = 2000$ papers. Therefore it is valid to just use one contour plot for $\tau = 200$ papers

soon as the new paper $(N+1)$ has a total of r references. For more on declustering refer to Goldberg (2013).

Note that the time scale used when choosing the core papers using global information, the attention span τ, is applied only to the selection of core papers in model C. This is to be contrasted with model B where all papers cited are core papers selected using global information with one time scale τ. There are many possible variations of our model C but we expect that similar results can be obtained as suggested in Evans and Saramaki (2005). We prefer to keep the model as simple as possible. This leaves us with just three parameters for our model C: p and τ from model B and the additional probability q.

To fix the unknown parameters we start by using the constraint that the total fraction of zero cited papers found in the output of model C, z_c, should be equal to that found in the data, $z = 0.169$. To find these optimal values in Fig. 16 we show contours indicating the fraction of zero cited papers z_c in the output of model C for a given p and q with the third parameter τ fixed. By looking at the widths of the citation degree distributions σ^2 against year in the model output, we found that 'reasonable' attention spans lay between 150 and 300 papers. Further in this range of τ the contour plots of z_c for different (p, q) values did not change significantly. Even when we tried much larger values for τ, we found that the contour plots for z_c against (p, q) were approximately the same for $\tau = 2000$ and 200 papers if q was less than about 0.3. In fact we find that $q = 0.2$ is our chosen value suggesting that z provides only a weak constraint on τ in the region of interest. However, we will take the attention span to be $\tau = 200$ papers because it works for all values of (p, q) and doesn't bias the determination of the best (p, q) values.

To fix p and q we need a further piece of information. To do this we conjectured that the value of q required will depend on the fraction of references to core papers in each paper. In the model C there will be a number of references made from new publications of which around $q\langle k^{(\mathrm{out})}\rangle = q\langle k^{(\mathrm{in})}\rangle$ go to subsidiary papers for every citation to a core paper. So a

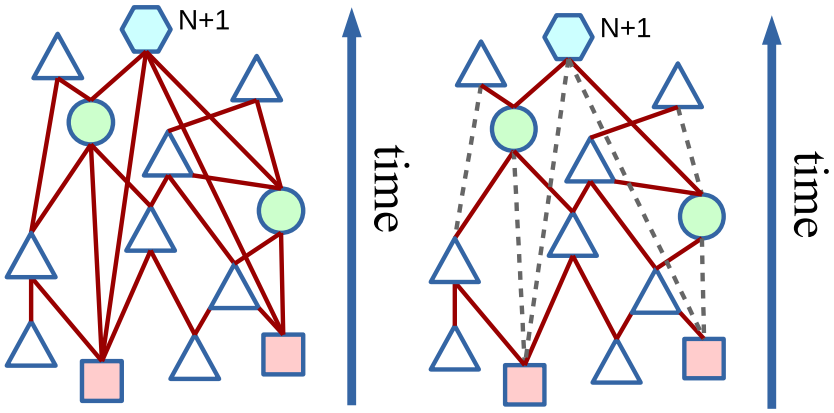

Fig. 17 An illustration of the transitive reduction to identify citations to 'core' papers. On the left is the citation network of Fig. 15 with new paper $(N+1)$ (*blue hexagon*) and its references (*green circles* for its core references, *red squares* for its subsidiary references). On the right is the transitive reduction of this network where the *dashed grey edges* are the ones removed under transitive reduction. In this case, after transitive reduction the new paper $(N+1)$ is only linked to the core papers in its references. The declustering method described in the text produces the same result in this case, but not in general. (Color figure online)

low q will require a large faction of core papers to make up the bibliography, and vice versa.

To find these references from papers to their core papers in the hep-th network we used a declustering method, a quicker and simpler version of transitive reduction (Clough et al. 2014; Clough and Evans 2014). For each paper X we start from the oldest paper referenced. We delete any references from nodes X to Z if there exists a reference route from X to Y and Y to Z, i.e. we remove the long edge of any triangles found between paper X and its references, Fig. 17. This will definitely remove all of the links from a paper to its subsidiary papers, but it is possible to remove other types of links. While not perfect, we use this to get an estimate for the average number of citations to core papers in the hep-th data, and we found this to be 3.9.[8] This is approximately what we expect as we expect that an author references a few core papers, on average, and a proportion of the core papers' bibliographies q. By running the same process on the output from model C, using parameters where $\tau = 200$ papers and line of (p, q) values in Fig. 16 from the line giving the correct z value, we determine that $p = 0.55$ and $q = 0.2$ are the best values.

At this point we recall that our analysis of the fraction of zero cited papers z which is used to choose the attention span τ is actually relatively insensitive to changes in τ when $q = 0.2$. So to find the optimal τ value for our best choice of $(p, q) = (0.55, 0.2)$ we now look at the σ^2 plot produced by model C.[9] We find τ by varying it in our model C [for $(p, q) = (0.55, 0.2)$] and find the τ which gives us the σ^2 plot closest to that of the data. We define 'closest' by, in a standard way as before using Eq. (7), minimising χ^2 against τ. In Fig. 18 we clearly observe a χ^2 minima at $\tau = 200$ papers, this corresponds to our most optimal and final τ because it gives the closest σ^2 plot to that of the data.

Checks on the optimal values

We have arrived at an optimal set of values for our model C: $p = 0.55, q = 0.20$ and $\tau = 200$ papers. It is worth looking to see how these values compare with what we know from elsewhere.

For the processes described by model C we might expect (assuming no correlation between in- and out-degree)[10] that

$$\langle k^{(\text{out})} \rangle = C(1 + q\langle k^{(\text{out})} \rangle) \tag{8}$$

where $C = 3.9$ is the average number of core papers in a bibliography and $\langle k^{(\text{out})} \rangle = 12.0$. This gives us $q = 0.17$ which is very close to the value we extracted numerically.

Our value for q also compares well with values used by Simkin and Roychowdhury (2005a) in their model and their different data set. They create a simple model in which a new publication references a few core papers and a quarter of the core papers' bibliographies. Therefore, they expect 25 % of a core paper's bibliography to be copied which is not too different from the 20 % we have arrived at for model C. Simkin and

[8] Note that the average in-degree of the full un-reduced arXiv network is 12.82. The average in-degree after two-step declustering is much less, 3.9. The fully transitively reduced network has an even smaller average in-degree of 2.27 (Clough et al. 2014; Goldberg 2013), as expected.

[9] For model C our final comparison tool is the σ^2 plot of model C and the hep-th data. Note that so far we have used z and the number of core papers C as our comparison tools.

[10] This estimate assumes no correlation between in- and out-degree as a better estimate for the average in-degree of core papers chosen using cumulative advantage is $\langle (k^{(\text{in})})^2 \rangle / \langle k^{(\text{in})} \rangle$.

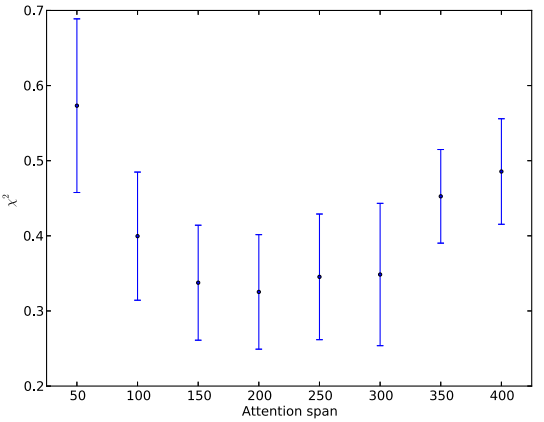

Fig. 18 Plot of χ^2 [the squared difference between the σ^2 of the data and the model C for fixed $(p, q) = (0.55, 0.20)$] against varying attention span τ measured in papers. Results show the mean and standard deviation for 10 runs of model C at each parameter value. The minimum is at at 200 ± 50 papers

Roychowdhury (2005b) also studied errors in bibliographies. Using a statistical model for this process they again arrive at the conclusion 70–90 % of the references to papers are literally copied from existing papers. We find this value to be, on average, $Cq\langle k^{(\text{out})}\rangle / \langle k^{(\text{out})}\rangle = Cq \approx 0.67 \approx 0.70 = 70\,\%$ or $0.78 \approx 80\,\%$ from the second term on the right side of Eq. 8, for $R = 3.9$ and $q = 0.173$ or $q = 0.20$ (which are the mathematically expected value of q and the q derived from fitting the model C to the hep-th data, respectively). Both values are compatible with our choice of $q = 0.20$.[11] Analysis of similar arXiv data sets using the different technique of transitive reduction by Clough et al. (2014) is also consistent with these values.

The value of the attention span model parameter $\tau = 200$ is a fraction of a year (a year is 2000 papers in rank time) and therefore at first appears to be surprisingly low. This result may be necessary if a low τ leads to the larger σ^2 values. The idea is that if a paper has not gained citations in this short time span, then it is very unlikely to gain many citations later. On the other hand, if a paper gains a few citations in this short time, then they are likely to be referenced again and again as a citation to a subsidiary paper in the future publications. Not only does this gives us the 'rich-get-richer' effect and so the long-tail of the in-degree distribution, more importantly the short attention span is exacerbating the difference between high and low citations, exactly the type of effect we need to widen the distribution of citations for papers published in the same year, i.e. it makes τ higher. Note that within the literature an attention span of less than a year is also found (Simkin and Roychowdhury 2007).

Another reason why the attention span parameter is so low could be that the academic field covered by the hep-th arXiv, which is largely the string theory approach to particle physics, genuinely has an unusually short time scale. The focus in hep-th is on particle physics theory but the bulk of papers are on topics which are highly theoretical and often weakly constrained by experimental data. This allows for a rapid response to the largest,

[11] We have also done another check. The proportion of core papers referenced per new publication is therefore, on average $= \frac{C}{\langle k^{(\text{out})}\rangle} = \frac{C}{C(1+q\langle k^{(\text{out})}\rangle)} = \frac{3.9}{12.0} \approx 0.3$ for $q = 0.17$ or $q = 0.20$ (which are the mathematically expected value of q and the q derived from fitting the model C to the hep-th data, respectively). In *A Mathematical Theory of Citing* and Solla Price (1965) this was found to be 0.1 and 0.15 for their models, respectively. Again, our values are consistent with these as they are low.

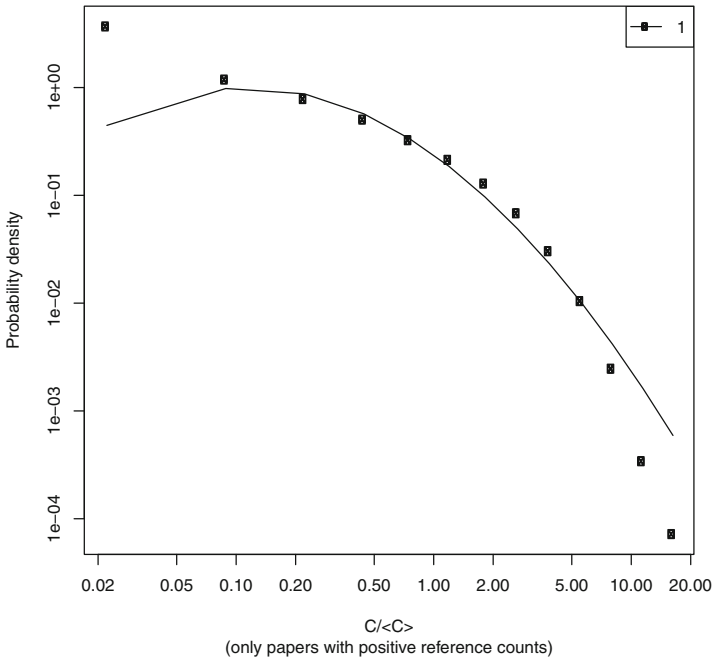

Fig. 19 This is the citation distribution for all papers for our final model C, $(p, q) = (0.55, 0.20)$ and $\tau = 200$ papers, plotted against normalised citation count, on a log–log plot. The tail has a large with of 1.68 ± 0.27 consistent with the data 1.78 ± 0.14. The fitted lognormal reaches a normalised citation count of approximately 50, the same as the hep-th data, Fig. 2, and 3/2 times that of model A and B, see Figs. 6 and 13, respectively

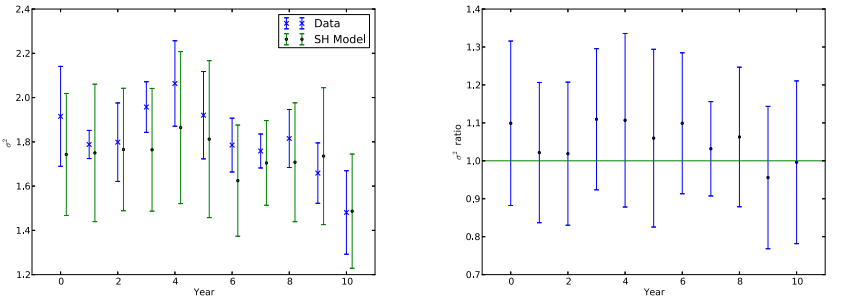

Fig. 20 Plots of width of the long-tail citation distribution, σ_y^2, for each year for data and model C. On the *left* are the absolute values for the σ_y^2 values for model C (*green*) and for the hep-th data (*blue*). On the *right* is a plot of the ratio $\sigma_{y,\text{data}}^2/\sigma_{y,\text{model}}^2$, the σ^2 value from the hep-th data divided by that found in model C with optimal parameters. The data and model C results for σ are consistent with each other, being within ± 0.2 of one another and within each others error bars. The data's σ^2 values are slightly higher for years 0–8 (1992–2000) but slightly lower for years 9 and 10 (2001 and 2002) though differences are not statistically significant. (Color figure online)

mostly theoretical, developments. Another factor here is that almost all work in this area is read on the hep-th and not in journals—arXiv was set up specifically for this area of research by a string theorist Paul Ginsparg. Again, this open access allows for, indeed was designed, to speed up access to work before slower paper-and-post based journals.

To see if our low attention span value reflects particular properties of the hep-th research area, we would need to extend our approach to different fields of research and different journals. Furthermore, different models of the fall-off in citation count over time would be a good test for whether this low attention span is just a limitation of Model C.

We can now look at the output from our model C with the optimal parameter values, $(p, q) = (0.55, 0.20)$ and $\tau = 200$ papers. For the citation distribution for the whole time period, shown in Fig. 19, we find a long-tailed distribution with $\sigma^2 = 1.68 \pm 0.27$, consistent with the hep-th data 1.78 ± 0.14. Also, the shape of the citation distribution for publications published in a given year, the widths σ_y^2, against time are all consistent with the data, see Fig. 20. That is we find the large and constant widths $\sigma_y^2 \approx 1.8$ seen in the data, but not seen in model A or model B.

Conclusions

We have created three citation network models and compared them to the hep-th arXiv data. Our aim has been to find what processes are sufficient to produce the long-tailed citation distribution for papers published in a given year against year which, when rescaled by the average citation count in that year, is large and constant $\approx$1.8 for different years.

The difficulty in achieving our aim is illustrated by failure of the first two simpler models (A and B). Our first was a simple variation of the Price model, while our second model added a preference to cite recent papers.

Our successful model is still relatively simple with just three parameters. It shows that around 70–80 % of papers cited are 'subsidiary papers', found by doing local searches from known recently published ('core') papers. The latter are recent papers identified using knowledge of the entire network (global information). Similar results have been seen using different methods by Simkin and Roychowdhury (2005b) and by Clough et al. (2014) but our model highlights the role of local and global information in the microscopic process leading to these citation patterns.

Our work suggests that in the future it is worth considering the role of the length of bibliographies, the out-degree distribution. While our normal distribution gives us variations in $k^{(\text{out})}$, which is often not present in other models (Simkin and Roychowdhury 2007; Zhu et al. 2003), it is not the fat-tailed distribution seen in the data. Our preliminary investigations found that the shape of the probability distribution of the bibliography length did alter our results, though our results were broadly similar. It would be interesting to find an effective but simple model with realistic distributions for both in- and out-degree distributions.

Acknowledgments We would like to thank James Gollings and James Clough for allowing us to use their transitive reduction code from which we created our own declustering code. We would like to thank Tamar Loach for sharing her results on related projects and M. V. Simkin for discussions about his work.

Appendix

Fitting procedure

We follow the procedure used in Evans et al. (2012) and Goldberg and Evans (2012). We use logarithmic binning so that the citations in bin b $c_b \in \mathbb{Z}$ with c_{b+1} equal to Rc_b rounded to the nearest integer or to $(c_b + 1)$, whichever is the highest, where R is some fixed bin scale chosen to ensure there are no empty bins below the bin containing the highest citation values. The edge of the first bin is chosen to be the lowest integer above the value $0.1\langle c\rangle$. In order to make the fit we compare the total value in the bth bin, $n_b = \sum_{c=c_b}^{c_{b+1}} n(c)$, against the expected value

$$n_b^{(\text{expect})} = (1+A)N \int_{c_b-0.5}^{c_{b+1}+0.5} \mathrm{d}c \; \frac{1}{\sqrt{2\pi}\sigma c} \exp\left\{ -\frac{(\ln(c/\langle c\rangle) + (\sigma^2/2) - B)^2}{2\sigma^2} \right\}. \tag{9}$$

This gives us a sequence of data and model values which are compared using a non-linear least squares algorithm to give us values for σ^2, A and B.

Lognormal out-degree distribution

In all above models the citation networks were created by determining the number of references a new node would create (via a normal distribution mean 12.0, standard deviation 3.0 references) and then having a method of deciding which nodes to reference. However, we found that a lognormal fits the out-degree distribution of the hep-th data better than a normal distribution, Figs. 4 and 21, respectively.

As further work we inputted this fitted lognormal to determine the number of references created by a new node into the model C and ran it for our final parameters $(p, q) = (0.55, 0.20)$ and $\tau = 200$papers. The ratio of the σ^2 values associated with the in-degree distribution of papers published in the same year for the data is divided by the corresponding year's σ^2 for this modified model C and plotted against year in Fig. 22. We find that the ratio is close to 1, however, it is not as close as the original model C, Fig. 20. Therefore the σ^2 plot *does* depend on in-degree, contradicting (Ren et al. 2012), who say it is 'innocuous' to the in-degree distribution of the citation network. Although this out-

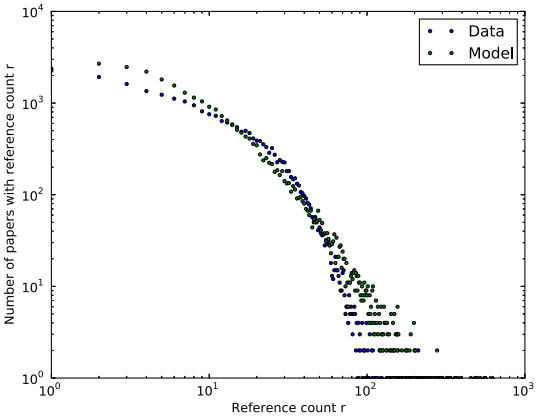

Fig. 21 This is a plot of the out-degree distribution of 27,000 publications from the hep-th arXiv data, in *blue*, on a log–log plot. Superimposed, in *green*, is a plot of 27,000 numbers generated by the lognormal distribution fitted to the out-degree of the hep-th arXiv data. We observe that the lognormal is a better fit to the data than the normal in Fig. 4. (Color figure online)

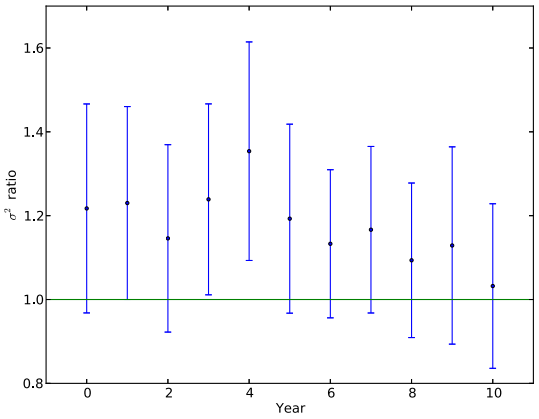

Fig. 22 This is the ratio of σ^2 associated with the in-degree distribution of papers published in the same year for the data (years from 1992, 1993 etc. relabelled to 0, 1 etc.) divided by that of the modified model C (where the out-degree is determined by a lognormal distribution, above). The *plot* and *error bars* lie within 1.0 therefore the modified model C is consistent with the data. Therefore the modified model C is promising, a significant improvement on model A and B, Figs. 7 and 14, respectively. However, the data is always lower than the modified model C; the points are always above the 1.0 line and not as close to 1.0 as the model C which implies the need for modification of the parameters in model C. So modifying the out-degree does change the in-degree, which contradicts (Vázquez 2001). We conjecture that by changing the attention span parameter this model's σ^2 plot could increase to match the data

degree distribution has been observed by the literature (Vázquez 2001) its use in a citation network model is novel and original.

Although the σ^2 plot is lower than that of the original model C we conjecture that by varying the τ of the model C the σ^2 plot could match that of the data's, this may also increase the attention span to something closer to a year as expected by Simkin and Roychowdhury (2007).

References

Bentley, R., Hahn, M., & Shennan, S. (2004). Random drift and culture change. *Proceedings of the Royal Society B*, *271*, 1443–1450.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*(1), 213–228.

Chung, F., Lu, L., Dewey, T. G., & Galas, D. J. (2003). Duplication models for biological networks. *Journal of Computational Biology*, *10*, 677–687.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *Siam Review*, *51*, 661–703.

Clough, J. R., & Evans, T. S. (2014). What is the dimension of citation space? arXiv:1408.1274.

Clough, J. R., Gollings, J., Loach, T. V., & Evans, T. S. (2014). Transitive reduction of citation networks. *Journal of Complex Networks*. arXiv:1310.8224.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, *149*, 510–515.

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*, 292–306.

Dorogovtsev, S., & Mendes, J. (2000). Evolution of networks with aging of sites. *Physical Review E*, *62*, 1842–1845.

Dorogovtsev, S., & Mendes, J. A. A. A. (2001). Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, *63*, 056125.

Dorogovtsev, S., Mendes, J., & Samukhin, A. (2000). Structure of growing networks with preferential linking. *Physical Review Letters*, *85*, 4633–4636.
Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *Plos One*, *6*, e24926.
Evans, T. S., Hopkins, N., & Kaube, B. S. (2012). Universality of performance indicators based on citation and reference counts. *Scientometrics*, *93*, 473–495. doi:10.1007/s11192-012-0694-9. arXiv:1110.3271.
Evans, T. S., & Saramaki, J. (2005). Scale-free networks from self-organization. *Physical Review E*, *72*, 026138. doi:10.1103/PhysRevE.72.026138. arXiv:cond-mat/0411390.
Fowler, J. H., & Jeon, S. (2008). The authority of supreme court precedent. *Social Networks*, *30*, 16–30.
Geng, X., & Wang, Y. (2009). Degree correlations in citation networks model with aging. *EPL*, *88*, 38002.
Goldberg, S. R. (2013). Modelling citation networks. figshare. doi:10.6084/m9.figshare.1134542.
Goldberg, S. R., & Evans, T. S. (2012). Universality of performance indicators based on citation and reference counts. *figshare*. doi:10.6084/m9.figshare.1134544. Retrieved 12 Aug 2014.
Golosovsky, M., & Sorin, S. (2013). The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. *Journal of Statistical Physics*, *151*, 340–354.
Hajra, K., & Sen, P. (2005). Aging in citation networks. *Physica A*, *346*, 44–48.
Hajra, K. B., & Sen, P. (2006). Modelling aging characteristics in citation networks. *Physica A*, *368*, 575–582.
KDD Cup. (2003). Network mining and usage log analysis. http://www.cs.cornell.edu/projects/kddcup/datasets.html. Accessed 1 Oct 2012.
Krapivsky, P., & Redner, S. (2001). Organization of growing random networks. *Physical Review E*, *63*, 066123.
Laherraére, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, *2*, 525–539.
Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of Neuroscience*, *28*, 11103–11105.
Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, *1*, 226–251.
Newman, M. (2010). *Networks: An introduction*. New York: Oxford University Press.
Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface*, *11*, 20140378–20140378.
Pollmann, T. (2000). Forgetting and the ageing of scientific publications. *Scientometrics*, *47*, 43–54.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 17268–17272.
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, *4*, 131–134.
Ren, F.-X., Shen, H.-W., & Cheng, X.-Q. (2012). Modeling the clustering in citation networks. *Physica A*, *391*, 3533–3539.
Saramäki, J., & Kaski, K. (2004). Scale-free networks generated by random walkers. *Physica A*, *341*, 80.
Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*, 628–638.
Simkin, M. V., & Roychowdhury, V. P. (2005a). Copied citations create renowned papers? *Annals of Improbable Research*, *11*, 24–27.
Simkin, M. V., & Roychowdhury, V. P. (2005b). Stochastic modeling of citation slips. *Scientometrics*, *62*, 367–384.
Simkin, M. V., & Roychowdhury, V. P. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology*, *58*, 1661–1673.
Smolinsky, L., Lercher, A., & McDaniel, A. (2015). Testing theories of preferential attachment in random networks of citations. *Journal of the Association for Information Science and Technology*. doi:10.1002/asi.23312.
Sternitzke, C., Bartkowski, A., & Schramm, R. (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information*, *30*, 115–131.
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, *3*(2), e1683.
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology*, *61*, 1377–1385.
van Raan, A. F. J. (2001). Two-step competition process leads to quasi power-law income distributions: Application to scientific publication and citation distributions. *Physica A*, *298*, 530–536.

Vazquez, A. (2000). Knowing a network by walking on it: Emergence of scaling. arXiv:cond-mat/0006132.
Vázquez, A. (2001). Statistics of citation networks. arXiv:cond-mat/0105031.
Vázquez, A. (2003). Growing networks with local rules: preferential attachment, clustering hierarchy and degree correlations. *Physical Review E*, *67*, 056104.
Wallace, M. L., Lariviere, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*, 296–303.
Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.
Wu, Y., Fu, T. Z. J., & Chiu, D. M. (2014). Generalized preferential attachment considering aging. *Journal of Informetrics*, *8*, 650–658.
Zhu, H., Wang, X., & Zhu, J. (2003). Effect of aging on network structure. *Physical Review E*, *68*, 056121.