Ranking scientific publications using a model of network traffic

Dylan Walker[1,2]**, Huafeng Xie**[2,3]**, Koon-Kiu Yan**[1,2] **and Sergei Maslov**[2]

[1] Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA
[2] Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, NY 11973, USA
[3] New Media Lab, The Graduate Center, CUNY, New York, NY 10016, USA
E-mail: dwalker@grad.physics.sunysb.edu, hxie@bnl.gov, kyan@bnl.gov and maslov@bnl.gov

Abstract. To account for strong ageing characteristics of citation networks, we modify the PageRank algorithm by initially distributing random surfers exponentially with age, in favour of more recent publications. The output of this algorithm, which we call CiteRank, is interpreted as approximate traffic to individual publications in a simple model of how researchers find new information. We optimize parameters of our algorithm to achieve the best performance. The results are compared for two rather different citation networks: all American Physical Society publications between 1893 and 2003 and the set of high-energy physics theory (hep-th) preprints. Despite major differences between these two networks, we find that their optimal parameters for the CiteRank algorithm are remarkably similar. The advantages and performance of CiteRank over more conventional methods of ranking publications are discussed.

Contents

1. Introduction

Due to their rapid growth and large size, many information networks have become untenable to navigate without some sort of ranking scheme. This is particularly evident in the example of the World Wide Web, a network of pages connected by hyperlinks. A successful solution to the problem of ranking the Web is Google's PageRank algorithm [1]. Another class of information networks that could benefit from such a ranking method are citation networks. These networks are comprised of scientific publications connected by citation links.

Current methods of ranking publications based on the total number of citations received are rather crude. They are too 'democratic' in treating all citations as equal and ignoring differences in importance of citing papers. One of the advantages of Google's PageRank algorithm is that it implicitly accounts for the importance of the citing article in a self-consistent fashion. Authors of [2] proposed using the PageRank algorithm to improve the formula used to calculate the impact factor of scientific journals. In [3] some of us directly applied this algorithm to individual papers published in all American Physical Society journals. This allowed us to discover a set of highly influential papers ('scientific gems') that would be undervalued based on just their number of citations. However, there exist significant differences between the World Wide Web and citation networks that suggest a modification of the original PageRank algorithm. The most important difference is that, unlike hyperlinks, citations cannot be updated after publication. This makes ageing effects [4, 5] in citation networks much more pronounced than in the WWW. The other consequence is the inherent time-arrow present in the topology of citation networks, due to the constraint that, at the time of publication, a paper may only cite earlier works. This significantly alters the spectral properties of the adjacency matrix which lie at the heart of the PageRank algorithm. In particular, the absence of directed loops means that the adjacency matrix can have only zero eigenvalues.

The success of the PageRank algorithm can be attributed, in part, to its ability to capture the behaviour of people randomly browsing the network of web pages. Indeed, the PageRank of a given web page can be interpreted as the predicted traffic (quantified, for example, by the rate of downloads) for that page if every WWW user follows a random path of (on average) $1/\alpha$ hyperlinks starting from a randomly selected webpage. The assumption that a typical web surfer starts at a randomly selected webpage might be not completely unreasonable for the WWW, but it needs to be modified for citation networks.

As all of us know, researchers typically start 'surfing' scientific publications from a rather *recent* publication that caught their attention on a daily update of a preprint archive, a recent volume of a journal or, perhaps, was featured in a news article in the popular media. Thus a more realistic model for the traffic along the citation network should take into account that researchers 'surfing' the citation network preferentially start their quests from recent papers and progressively get to older and older papers with every step.

In this work we introduce the CiteRank algorithm, an adaptation of the PageRank algorithm to citation networks. Our algorithm simulates the dynamics of a large number of researchers looking for new information. Every researcher, independent of one another, is assumed to start his/her search from a *recent* paper or review and to subsequently follow a chain of citations until satisfied or saturated with information. Explicitly, we define the following two-parameter CiteRank model of such a process, allowing one to estimate the traffic $T_i(\tau_{\mathrm{dir}}, \alpha)$ to a given paper i. A recent paper is selected randomly from the whole population with a probability that is exponentially discounted according to the age of the paper, with a characteristic decay time of τ_{dir}. At every step of the path, with probability α the researcher is satisfied/saturated and halts his/her line of inquiry. With probability $(1-\alpha)$ a random citation to an adjacent paper is followed. The predicted traffic, $T_i(\tau_{\mathrm{dir}}, \alpha)$, to a paper is proportional to the rate at which it is visited (downloaded) if a large number of researchers independently follow such a simple-minded process.

While we interpret the output of the CiteRank algorithm as the traffic, its utility ultimately lies in the ability to successfully rank publications. High CiteRank traffic to a publication denotes its high relevance in the context of currently popular research directions, while the PageRank number is more of a 'lifetime achievement award' [3]. It is fruitful to compare the CiteRank traffic to a paper, T_i, with the more traditional method of ranking publications, the number of citations received. Indeed, the two are highly correlated; a result easily understood on the basis that the larger the number of citations a paper has, the more likely it will be visited by a researcher via one of the incoming links.

However, the more refined CiteRank algorithm surpasses the conventional ranking, by number of citations, in its characterization of relevancy on two accounts:

- Like the original PageRank algorithm [1, 2], in CiteRank, the popularity of papers is calculated in a self-consistent fashion: the effect of a citation from a more popular paper is greater than that of a less popular one. A citation from a paper that is 'highly visible' will contribute more to the visibility of the cited paper.
- The age of a citing paper is intrinsically accounted for: the effect of a recent citation to a paper is greater than that of an older citation to the same paper. New citations indicate the relevancy of a paper in the context of current lines of research.

2. Results

An algorithmic description of the aforementioned model can be understood as follows. The transfer matrix associated with the citation network is

$$W_{ij} = \begin{cases} \dfrac{1}{k_j^{\mathrm{out}}} & \text{if } j \text{ cites } i; \\ 0 & \text{otherwise}, \end{cases}$$

3

where k_j^{out} is the out-degree of the jth paper. Let ρ_i represent the probability of initially selecting the ith paper in a citation network:

$$\rho_i = e^{-\text{age}_i/\tau_{\text{dir}}}. \tag{1}$$

The probability that the researcher will encounter a paper by initial selection alone is given by $\vec{\rho}$. Similarly, the probability of encountering the paper after following one link is $(1-\alpha)W \cdot \vec{\rho}$. The CiteRank traffic of the paper is then defined as the probability of encountering it via paths of any length. That is, given an initial distribution of new papers, $\vec{\rho}$, and transfer matrix, W, the CiteRank traffic is given by:

$$\vec{T} = I \cdot \vec{\rho} + (1-\alpha)W \cdot \vec{\rho} + (1-\alpha)^2 W^2 \cdot \vec{\rho} + \cdots. \tag{2}$$

Practically, we calculate the CiteRank traffic on all papers in our dataset by taking successive terms in the above expansion to sufficient convergence ($<10^{-10}$ of the average value).

In order to assess the viability of this ranking scheme and to select optimal parameters $(\tau_{\text{dir}}, \alpha)$, we need a quantitative measure of its performance on real citation networks. Two real citation networks are evaluated:

- *Hep-th*: An archive snapshot of the 'high energy physics theory' archive (http://arxiv.org/archive/hep-th) from April 2003 (preprints ranging from 1992 to 2003). This dataset, containing around 28 000 papers and 350 000 citation links, was downloaded from [6]. We know the actual date of appearance of each of the entries in the preprint archive and thus the age of each node is known with the resolution of 1 day.

- *Physrev*: Citation data between journals published by the American Physical Society[4]. This dataset contains around 380 000 papers and 3 100 000 citation links. We know only the year in which each paper was published and it ranges from 1893 to 2003.

Of course, evaluating the performance of any ranking scheme is a delicate, but often necessary, matter. One way to select the best performing α and τ_{dir} is to optimize the correlation between the predicted traffic, $T_i(\tau_{\text{dir}}, \alpha)$ and the actual traffic (e.g. downloads). Unfortunately, the actual traffic data for scientific publications are not readily available for these networks. However, it is reasonable to assume that traffic to a paper is positively correlated with the number of new citations it accrues over a recent time interval, Δk_{in}.

For lack of better intuition we first assume a linear relationship between actual traffic and number of recent citations accrued. This corresponds to a simple-minded scenario in which every researcher downloading a paper will, with a certain small probability, add it to the citation list of the manuscript he/she is currently writing. It should be noted that we make no attempt to model network growth in this paper.

In order to compare CiteRank with actual citation accrual, we constructed an historical snapshot of both networks used in this study. In both cases, the most recent 10% of papers are pruned from the network. This corresponds to the last 4 years (2000–2003)

---

[4] The APS journals include Phys. Rev. Series I (1893–1912), Phys. Rev. Series II (1913–1969), and Phys. Rev. Series III (1970–present). This latter series includes the five topical sections: Phys. Rev. A, B, C, D and E (the latter from 1990 to the present). Also included are Phys. Rev. Lett., Rev. Mod. Phys. and Phys. Rev. Special Topics, Accelerators and Beams (1998–present).
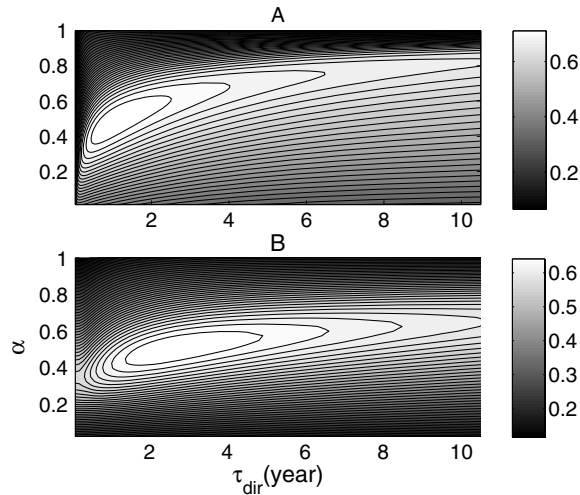
Figure 1. The Pearson (linear) correlation coefficient between the number of recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) is calculated over the parameter space of the CiteRank model for the Hep-th (A) and Physrev (B) network. Both networks exhibit peaks in correlation coefficient in the α-τ_{dir} plane. The highest correlation is achieved for $\alpha = 0.48$, $\tau_{\mathrm{dir}} = 1$ year in the Hep-th network and $\alpha = 0.50$, $\tau_{\mathrm{dir}} = 2.6$ years in the Physrev network.

in the Physrev network and last 1 year in the Hep-th network. The CiteRank traffic, T_i, of the remaining 90% of the papers is then evaluated and correlated with their actual accrual of new citations, Δk_{in}, originating at the most recent 10% of papers.

It is important to note the qualitative and quantitative differences between the two citation networks considered. The Physical Review citation network (Physrev) is comprised of a large number ($\sim 400\,000$) of peer-reviewed publications acquired over a period close to 100 years. The high-energy physics archive citation network (Hep-th) is completely comprised of a much smaller number ($\sim 28\,000$) of electronically submitted publication preprints, with no associated form of peer review. Despite these significant differences in the nature of the networks considered, the general features of their correlation contours are outstandingly similar. In both cases, a single sharp peak in correlation is evident for particular values of the parameters. The value of the optimal parameters for both networks are:

- hep-th: $\alpha = 0.48$, $\tau_{\mathrm{dir}} = 1$ year
- physrev: $\alpha = 0.50$, $\tau_{\mathrm{dir}} = 2.6$ years.

Remarkably, the value of α is nearly the same for the rather different networks considered and is in agreement with that proposed in [3] on purely empirical grounds. The difference in optimal parameter τ_{dir} for these networks is in agreement with the commonsense expectation of faster response time (and hence faster ageing of citations) in preprint archives compared to peer-reviewed publications. Another feature of figure 1 is that, in both networks, large values of the correlation coefficient are concentrated along a diagonally positioned ridge. In other words, the best choice of α for a given τ_{dir} seems to rise linearly with τ_{dir}, a behaviour that will be revisited later in this text.
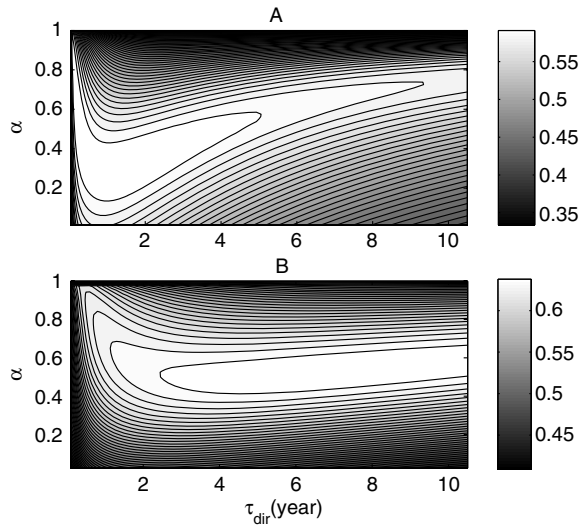
Figure 2. The Spearman rank correlation coefficient between recent citations accrued (Δk_{in}) and CiteRank traffic (T_i) for the Hep-th (A) and Physrev (B) network. Both networks exhibit similar behaviour. There are more extended regions of good correlation relative to the linear correlation contours of figure 1. This broadening is expected as a consequence of the more relaxed correlation measure. The highest rank correlation occurs for $\alpha = 0.31$, $\tau_{\mathrm{dir}} = 1.6$ years, in the Hep-th network and $\alpha = 0.55$, $\tau_{\mathrm{dir}} = 8$ years, in the Physrev network.

While the correlation contour plots shown in figure 1 are a promising indication that the CiteRank model of traffic with optimized parameters provides a good zero-order approximation to the actual traffic along a citation network, they are, to some extent, predicated on the assumption of a linear relationship between actual traffic and Δk_{in}. One might readily ask how this model fares in the absence of such an assumption. While the assumption of a *linear* relationship may be unreasonable, a positive, monotonic relationship between these quantities is certainly expected. There is a statistical correlation method precisely adapted for such a situation, namely, the Spearman rank correlation. Under this relaxed correlation measure, only the rank of T_i is correlated with the rank of Δk_{in}. Numerical changes in T_i that do not lead to reordering have no effect on the value of the rank correlation coefficient. Another rationale for using rank correlations is that our ultimate goal is ranking publications, not modelling the traffic. Thus, we are currently not interested in individual T_i's, but only in their relative values. Spearman correlation contour plots are constructed for both networks and shown in figure 2. The optimal values for both networks are:

- hep-th: $\alpha = 0.31$, $\tau_{\mathrm{dir}} = 1.6$ year
- physrev: $\alpha = 0.55$, $\tau_{\mathrm{dir}} = 8$ years.

These results roughly confirm the prediction of $\alpha \sim 0.5$ from figure 1: however, there is a more appreciable discrepancy in τ_{dir} between linear and rank correlation for both networks.
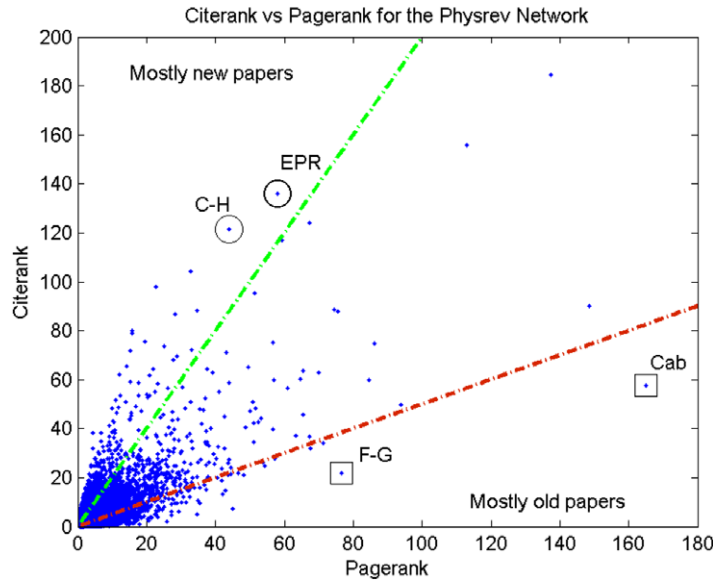
Figure 3. The scatter plot of CiteRank versus PageRank for all papers in the Physrev network. Two sectors of the data are distinguished according to the CiteRank to PageRank ratio: CR/PR > 2 (above the dashed green line) and CR/PR < 1/2 (below the dashed red line). The average publication year of papers in these sectors is 2000 and 1973, respectively. The sophistication of CiteRank goes beyond simple age classification, however. Particular examples that illustrate this sophistication are marked above and discussed in the main body of the text.

3. Discussion

A qualitative examination of CiteRank performance over the unmodified PageRank algorithm can be accomplished by direct comparison on the networks in question. As an example of this, the scatter plot of CiteRank versus PageRank for all papers in the Physrev network is shown in figure 3. The positive correlation between the two algorithms is clearly evident in the plot. Two sectors of the data are distinguished according to the ratio of CiteRank to PageRank, CR/PR. Papers with a relatively large (small) ratio have been marked comparatively higher (lower) by the ageing effects inherent to the CiteRank algorithm. These sectors are distinguished in figure 3 above (below) the dashed green (red) line, respectively. In accordance with our claim that the CiteRank algorithm ranks papers of current relevance in research higher, the average publication year of papers in the high-(low-) ratio sector is found to be 2000 (1973). Of course, CiteRank is more sophisticated than a simple re-ranking according to publication age. For one thing, recent citations contribute greatly to a paper of *any* age. A particularly good example of this is the famous 1935 Einstein, Podolsky and Rosen (EPR) paper [8] which receives both a large CiteRank and CR/PR ratio despite its age. A quick glance at citing papers reveals approximately fifty citations to this paper throughout this year (2007) alone, indicating its clear connection to current lines of research. Another notable example of a publication in the high-ratio sector is a review paper of out-of-equilibrium pattern formation [9] by

Cross and Hohenberg (C–H). It has a high ratio, despite being significantly older than papers in this sector. This paper is a good example of a class of *review* papers that serve to summarize the state of affairs regarding a particular topic that is of continuing interest to research. They are clearly of great use to the modern researcher and thus obtain their high CR/PR by virtue of recent citations. Of further interest are papers that received high PageRank (lifetime achievement) but have a relatively low CR/PR ratio. The wealth of these papers cover undeniably fundamental advancements in physics. Two explicit examples of this are the Feynman and Gell-Mann paper on Fermi interactions [11] (F–G) and the well known Cabibbo paper on leptonic decay (Cab) [10]. The low CR/PR of these papers can be explained by a dearth of recent citations, which in turn is likely due to the incorporation of fundamental discoveries and advancements into textbooks and other published works that include more recent developments, in addition to historical context.

A better physical understanding of the sophistication of the CiteRank algorithm may be gleaned from a simple quantitative analysis of the traffic dynamics in terms of its parameters. In both panels of figure 1, over a broad range of parameters, the optimal value of $\alpha(\tau_{\rm dir})$ for a given value of $\tau_{\rm dir}$ is positively correlated with $\tau_{\rm dir}$. This is an indication that these two parameters are entangled. In fact, this is to be expected as it is some admixture of the two parameters which leads to the exposure of a given paper to the researcher. An intuitive picture of this entanglement can be understood in terms of the penetration depth, which is a measure of how far back in time a random surfer following the rules of the CiteRank algorithm is likely to get. The penetration depth is affected by both $\tau_{\rm dir}$—the average age of the initial paper at which he/she started following the chain of citations—and $1/\alpha$—the mean number of steps on this chain of citations. For small $\tau_{\rm dir}$ and large α, the penetration depth is small, implying that only very recent papers receive traffic. On the other hand, for large $\tau_{\rm dir}$ and small α, the penetration depth is very large, indicating that most of the traffic is directed towards older papers.

To better understand how α and $\tau_{\rm dir}$ influence the age distribution of CiteRank traffic, we performed the following quantitative analysis. Let $T_{\rm tot}(t)$ denote the CiteRank model traffic to papers written exactly t years ago, where the meaning of the additional subscript shall be made clear in the lines that follow. As described by equation (2), two distinct processes contribute to $T_{\rm tot}(t)$. The first is the 'direct' traffic $T_{\rm dir}(t)$ due to the initial selection of papers in this age group, which is proportional to $\exp(-t/\tau_{\rm dir})$.[5] The second is the 'indirect' traffic $T_{\rm ind}(t)$ arriving via one of the incoming citation links, which is given by $T_{\rm ind}(t) = (1-\alpha) \int_0^t T_{\rm tot}(t') P_c(t', t) \, {\rm d}t'$, where $P_c(t', t)$ is the fraction of citations originating from papers of age t' that cite papers of age t. It should be noted that $P_c(t', t)$ is an *empirical* distribution and, as such, is a *measured* property of the citation network under consideration. The integral takes into account the fact that incoming links to papers of age t can originate from all possible intermediate times. According to [5] and our own findings, $P_c(t', t)$ is reasonably well approximated by the exponential form $(1/\tau_c) \exp(-(t - t')/\tau_c)$. Taking the Fourier transform of the equation $T_{\rm tot}(t) = T_{\rm dir}(t) + T_{\rm ind}(t)$, we have

$$T_{\rm tot}(\omega) = T_{\rm dir}(\omega) + (1 - \alpha) T_{\rm tot}(\omega) P_c(\omega). \tag{3}$$

---

[5] Precisely speaking $T_d(t)$ in the CiteRank model is given by $N_p(t)$—the number of papers of age t—multiplied by the exponential probability of selection $\exp(-t/\tau_{\rm dir})$. Since $N_p(t)$ itself often has an approximately exponential form with time constant τ_p, $\tau_{\rm dir}$ used in the following equations should be 'renormalized' to $\tilde{\tau_{\rm dir}} = \tau_{\rm dir} \cdot \tau_p/(\tau_p + \tau_{\rm dir})$. However, τ_p is usually rather large ($\sim$28 years in the PhysRev network). Thus, except for very large $\tau_{\rm dir}$'s this renormalization can be safely ignored.
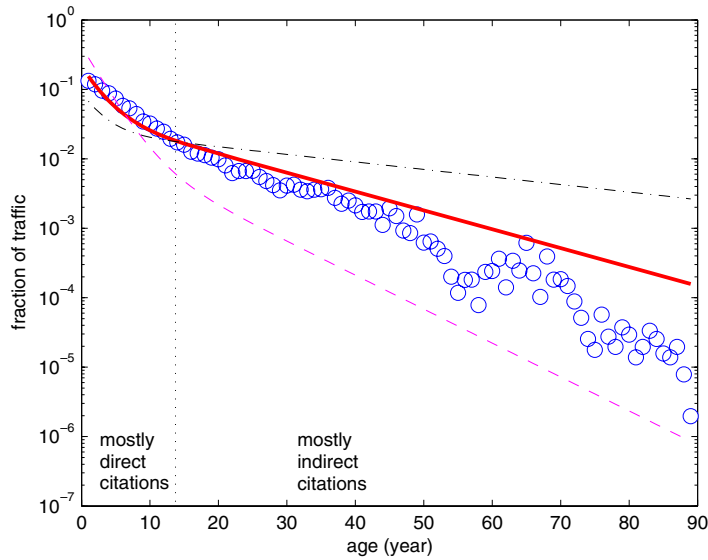
Figure 4. The age distribution of newly accrued citations Δk_{in} (blue) for the Physrev network. Theoretical predictions (4) for the CiteRank traffic are calculated for the optimal $\tau_{dir} = 2.6$ and three values of $\alpha = 0.2$ (dotted–dashed line), 0.5 (thick solid line) and 0.9 (dashed line). In agreement with figure 1, the optimal value, $\alpha = 0.5$, provides the best agreement with Δk_{in}. All curves are normalized so that the sum of all data points is equal to 1.

Solving equation (3) and taking the inverse Fourier transform, yields

$$T_{tot}(t) \sim (\tau_c - \tau_{dir}) \exp(-t/\tau_{dir}) + (1 - \alpha)\tau_{dir} \exp(-\alpha t/\tau_c). \qquad (4)$$

Thus, the traffic arriving at the subset of papers of age t is given by the superposition of two exponential functions.

Having an approximate analytical expression for $T_{tot}(t)$, we are now in a position to better understand what determines the optimal values of α and τ_{dir}. Figure 4 shows the age distribution of the number of recently acquired citations, Δk_{in}, for papers in the Physrev dataset. The approximate CiteRank traffic, given by equation (4), is also displayed. It is calculated using the empirically determined value $\tau_c = 8$ years, optimal $\tau_{dir} = 2.6$ years and three values of $\alpha = 0.2, 0.5$ and 0.9. As one would expect, the profile of $\langle \Delta k_{in} \rangle$ versus t best agrees with the CiteRank plot for the optimal value $\alpha = 0.5$.[6] Figure 4 also provides some clues to the positive correlation between near-optimal choices of α and τ_{dir}, visible as diagonal 'ridges' in figures 1(A) and (B). Indeed, if the value of α is chosen to be large, the contribution from the second term is diminished; the use of a larger value of τ_{dir} could partially compensate for the loss of CiteRank traffic to older papers, and would thus be in reasonably good agreement with the Δk_{in} data.

Another encouraging observation is that, like equation (4), the age distribution of recently acquired citations shown in figure 4 has two regimes characterized by two different decay constants of about 5 and 16 years, with a crossover point around $t = 15$ years. Our

---

[6] The apparent disagreement in the tail involves profound dips due to World War II and I [5], which of course cannot be explained by any theoretical model.

interpretation of this fact is that papers are found and cited via two distinct mechanisms: researchers can either find a paper directly or by following citation links from earlier papers. For each of these mechanisms, the probability that a given paper is found decays with its age but the characteristic decay time for the direct discovery is shorter. While very recent papers, especially the ones altogether lacking citations, are for the most part discovered directly, older papers are mostly discovered by following citation links.

The optimal values of α in the two very different citation networks considered are remarkably close to each other. In both cases it appears that, on average, the length of chains of citations followed by a typical researcher is close to $1/\alpha \simeq 2$. Since this chain includes the original starting point (the paper from which the researcher started his quest), the length of around 2 means that the average cited paper is just one link away from the starting point. This raises a disconcerting possibility that many of the papers cited in any given manuscript are not read by the authors but instead their citations are copied from other papers they know. Sadly, this was recently proven to be a very common scenario [7].

Acknowledgments

References

[1] Brin S and Page L, 1998 *Comput. Netw. ISDN Syst.* **30** 107
[2] Bollen J, Rodriguez M A and Van de Sompel H, 2006 *Preprint* cs.DL/0601030
[3] Chen P, Xie H, Maslov S and Redner S, 2006 *Preprint* physics/0604130
[4] De Solla Price D J, 1965 *Science* **149** 510 (The PDF-file of this pioneering study can be downloaded e.g. from http://garfield.library.upenn.edu/papers/pricenetworks1965.pdf)
[5] Redner S, 2005 *Phys. Today* **58** 49 [physics/0407137]
[6] This hep-th dataset was used in the KDD Cup 2003, a knowledge discovery and data mining competition, and can be downloaded from their website http://www.cs.cornell.edu/projects/kddcup/
[7] Simkin M V and Roychowdhury V P, 2003 *Complex Syst.* **14** 269 [cond-mat/0212043]
[8] Feynman R P and Gell-Mann M, *Can quantum-mechanical description of physical reality be complete?*, 1935 *Phys. Rev.* **47** 777
[9] Cross M C and Hohenberg P C, *Pattern formation outside of equilibrium*, 1993 *Rev. Mod. Phys.* **65** 851
[10] Cabibbo N, *Unitary symmetry and leptonic decays*, 1963 *Phys. Rev. Lett.* **10** 531
[11] Feynman R P and Gell-Mann M, *Theory of the Fermi interaction*, 1958 *Phys. Rev.* **109** 193