INFORMATION AND SOCIAL SYSTEM INTERACTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Asif-ul Haque

August 2011

© 2011 Asif-ul Haque

ALL RIGHTS RESERVED

INFORMATION AND SOCIAL SYSTEM INTERACTION

Asif-ul Haque, Ph.D.

Cornell University 2011

Ever increasing participation has made the interaction between information and social systems not only interesting to observe but essential to quantify and analyze. This dissertation presents methods for understanding such interaction through combined analysis of metadata, networks, text and log data. ArXiv, an open and highly influential scholarly communication system, served as the testbed for these methods.

In the first part of this dissertation we examine in depth interesting phenomena such as self-promotion, procrastination, visibility and geographic differences. We have confirmed the predictive power of early readership through regression and discussed undesirable effects of recommendation and possibilities of new impact metrics.

In the second part we demonstrate extraction of subtopical concepts, characterized by phrases, through a statistical method for vocabulary selection and a network based ranking. Validation via search query and click logs is advocated as relevant and scalable. A clustering scheme to summarize temporal patterns of topic clicks is also presented.

In the last part of this dissertation we present a name disambiguation algorithm and a novel evaluation method using node role based sampling in the context of network analysis. Finally we provide guidelines on performing large scale graph computation using the Map-Reduce framework.

BIOGRAPHICAL SKETCH

Asif-ul Haque grew up in the dense bustling metropolis of Dhaka, Bangladesh. He attended St Joseph High School and Notre Dame College in Dhaka till 12th grade. He completed his undergraduate studies at the Bangladesh University of Engineering and Technology (BUET) where he received Bachelor of Science in Computer Science and Engineering. Asif joined Cornell University as a graduate student in 2005. He received a PhD in Computer Science with a minor in Operations Research and Information Engineering in August 2011.

Dedicated to my father *Abdul Haque* and my mother *Kohinoor Haque*

ACKNOWLEDGEMENTS

I thank my committee members Eric Friedman, Paul Ginsparg and David Williamson for guiding me through my graduate studies. It has truly been a pleasure to be able to work with Paul Ginsparg whose vision and multi-disciplinary interests I admire. I am much indebted to Eric Friedman for helping me realize my interests and accommodating them. I am grateful to David Williamson for supervising my minor in Operations Research and providing helpful suggestions about my research.

I thank my collaborators Carl Lagoze, William Arms and my dear friend Theresa Velden for constant motivation. I thank all the professors, graduate students, researchers and staff in the departments of Computer and Information Science for providing a stimulating environment for research. I acknowledge with gratitude that my research was made possible by National Science Foundation Awards 0404553 and OCI-0926550.

I thank my parents, Abdul Haque and Kohinoor Haque, for all the love, sacrifice and lifelong support of instilling confidence in my creativity and freedom. I also thank my younger brother Enamul Haque for sharing his interest and passion in technology, music and sport as we grew up, and my younger sister Tasmira Haque for always listening to my concerns. I am grateful to Dr M Kaykobad for being a fatherly figure during my undergraduate studies and inspiring me throughout my graduate studies. I thank all my wonderful friends in the USA, Bangladesh and abroad for taking care of my well-being.

Last but not least I thank my wife Ooyshee whom I met at Cornell. Her constant attention during my final year of PhD ensured that I was in a positive frame of mind. I am grateful to her for being a part of my life through happy and difficult times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part I

Introduction

CHAPTER 1

INTRODUCTION

1.1 Information Systems

Information systems of our daily lives are at the confluence of computation and communication. Rapid progress in producing faster, cheaper computers with larger storage and faster connectivity has made computing devices as common as household items. And with the explosive growth of hand-held devices such as smart phones and tablets, our connectivity with information systems has made them ubiquitous tools for work and entertainment beyond. This dissertation explores and discusses methods to mine data from these systems.

Systems that we rely on everyday are email systems, instant messenger systems, search engines, social networking sites, personal blogs and discussion forums, photo and video sharing sites. In areas of arts and sciences, information systems have revolutionized modes of public communication. With the advent of many music sharing websites new and young artists are able to produce music at home and reach millions globally. Similarly indispensable tools for many areas of science are the scholarly communication systems that accumulate and disseminate knowledge. In this dissertation, information systems where humans generate and consume information are considered relevant. Computerized systems such as weather analysis, air traffic control or distributed cluster management systems are not discussed here as these systems have neither significant human participation nor mass accessibility.

For large systems, it is preferable to employ simpler methods for data mining, sacrificing a fraction of the quality if necessary, but lowering the turnaround time that would otherwise be prohibitively large for complicated algorithms. As the amount of data increases, these simple methods often approach sophisticated ones in terms of accuracy. With this important observation in perspective, the methods discussed in this dissertation are deliberately aimed to be simple yet effective. Scalability of computation and storage is also an important concern for large systems. For services that involve sifting through terabytes of data in millisecond time, cloud computing is becoming the preferred framework for computation. Open source implementations of the Map-Reduce paradigm have made efficient data aggregation easy for application programmers. We have discussed the possibility of using cloud computing whenever appropriate.

1.2 Social Interaction through Information Systems

On social networking and media sharing systems, social interaction is obvious. Search engines utilize the "wisdom of the crowds" by building services, from spelling correctors to translators, using probabilistic language models engineered from human generated text and patterns of submitted queries. For some tasks, social interaction, quantified through features, is combined with domain knowledge. Sentiment analysis is one such area where understanding of natural languages is supplemented by statistical models derived from large corpora. For other tasks, human interaction at the social level is exclusively used. Automatic friend recommendation on social networks, as an example, does not (yet) attempt to understand the nature of human friendship, its cultural significance or evolution – the simple basis for predicting friendship is individual interaction trends in the recent past.

Measurement of social interaction on information systems is interesting for two reasons. Firstly, quantification of medium to large scale human interaction permits understanding ourselves better. Sociological hypotheses and queries that could only have been proved or satisfied through careful reasoning in the past are now much easier to establish by mining large amounts of data – simple conjectures can be validated empirically. Implications in policy making and advertising are far reaching.

Secondly, the use of common patterns of social interaction in prediction reinforces certain human behavior while discounting others. Through these information systems we are learning how to search, recommend and filter. At the extreme, there is a constant narrative on our minds of what message to broadcast to our friends everyday, which, as a social phenomenon, may have been present since prehistory, but realized only recently due to the ease with which we can reach our friends and acquaintances whenever we wish. We are learning to perform our duties of taking turn in sharing expressions of such narratives and being a listener for others. Whether this is something positive or negative goes well beyond the scope of this dissertation. But the importance of measuring social interaction on information systems is easily understood.

The focus of this dissertation is to explore measurement of social interaction through information systems. We discuss a set of methods to understand how and which human tendencies are getting reinforced through these systems and how to obtain interesting features to perform tasks such as prediction, aggregation and generalization. We do not attempt to present a comprehensive set of tools for all data mining needs on information systems. Instead we present simple methods to expose social behavior not investigated adequately in the past, and to combine orthogonal sources of information to vary the granularity of tasks already established as necessary.

1.3 Scholarly Communication

The methods proposed in this dissertation are aimed to be as generalizable as possible. Scholarly communication sytems have the appropriate mixture of components from various information systems that nurture social interaction. The articles on arXiv are interlinked with each other via citations. Thus the corpus of full-text is very similar to corpora that search engines crawl, index and mine, except that within areas of research the articles are homogeneous and higher quality, either due to direct moderation, even minimal, or professional ethics of not promoting inferior material. Article abstracts are analogous to document summaries and keywords to tags. While structured search is easier on scholarly systems, full-text search is similar to search engine algorithms. Most scholarly systems have a log of user access, identified through IP addresses or http cookies, that can be used as a source of implicit information. Search query logs are similarly useful for certain tasks, as we will discuss in chapter 4.

Specificity of domains for scholarly articles allows exploration of entity extraction that is much more difficult for heterogeneous collections of web documents. One noteworthy difference between the interlinking pattern of web documents and scholarly articles is that the citations have a temporal dimension associated with them. So the network formed is a directed acyclic graph whereas a static snapshot of the world wide web contains a large giant component that is strongly connected. In this respect, a corpus of articles is similar to blogs. However, blogs evolve much faster than research articles and temporal normalization for scholarly systems is thus more challenging than blogs.

Scholarly communication systems also provide a social network of researchers. Authorship represented as a bipartite graph between authors and articles can be used to induce a co-authorship network where two individuals authoring an article are connected by an edge. Furthermore, the number of articles co-authored by two individuals can be used to infer strength of their connection. The structure of this social network is of much interest to both social scientists and researchers working on prediction and recommendation algorithms. In this dissertation, we propose methodologies to mine metadata, full-text interlinked via citations, various log data and the social network of individuals. We shed light on interesting social phenomena, extract and track concepts from full-text combined with network analysis, and resolve entities using network features.

There are many scholarly communication tools, some of which are very domain specific. PubMed[1] for example deals with biomedical literature exclusively, while others, such as the ISI Web of Knowledge[2], cover many areas of scholarly interest. These systems mostly deal with bibliometric data and make it easier for researchers to manage bibliographies and access full-text from publishers. Some of the systems, such as CiteSeer[3] for Computer Science, crawl author websites and link to the articles available there. Systems such as the SAO/NASA Astrophysics Data System[4] maintain bibliometric data, link to both publisher sites and preprint versions of articles, and allow personalized access. ArXiv[5] is one of the most prominent and comprehensive communication sys-

[1] http://www.ncbi.nlm.nih.gov/pubmed/

[2] http://apps.isiknowledge.com/

[3] http://citeseerx.ist.psu.edu/

[4] http://adswww.harvard.edu/

[5] http://arXiv.org/

tems, and stores full-text of preprint versions of articles. For many areas of science, it is the most important tool for knowledge dissemination. In this dissertation we have relied on data from arXiv and additional information from the SAO/NASA Astrophysics Data System, the High-Energy Physics Literature Database[6] and the ISI Web of Knowledge.

1.4 arXiv.org

Since its creation by Paul Ginsparg 20 years ago, arXiv rapidly became the most important scholarly communication system. Authors usually submit preprint versions of their articles that are under review in journals and conferences. These articles are then made globally available via daily listings. This way arXiv has been serving two very important purposes for two decades : rapid dissemination of knowledge and open access to scholarly articles. For background information about arXiv and its vision, see [36, 37, 38, 35]. Currently Cornell University library maintains arXiv with the help of a number of supporting institutions.

ArXiv currently contains over 650,000 articles with about 6,000 submissions per month. The steady increase in monthly submission is shown in fig. 1.1. In fig. 1.2, the monthly averages for high-energy physics, condensed matter physics, astrophysics and mathematics (including mathematical physics) is shown over 20 years. We observe that high-energy physicists embraced arXiv early and submissions have been very steady, with an increase in recent years. Condensed matter physics and astrophysics subject areas have shown constant

[6] http://www.slac.stanford.edu/spires/

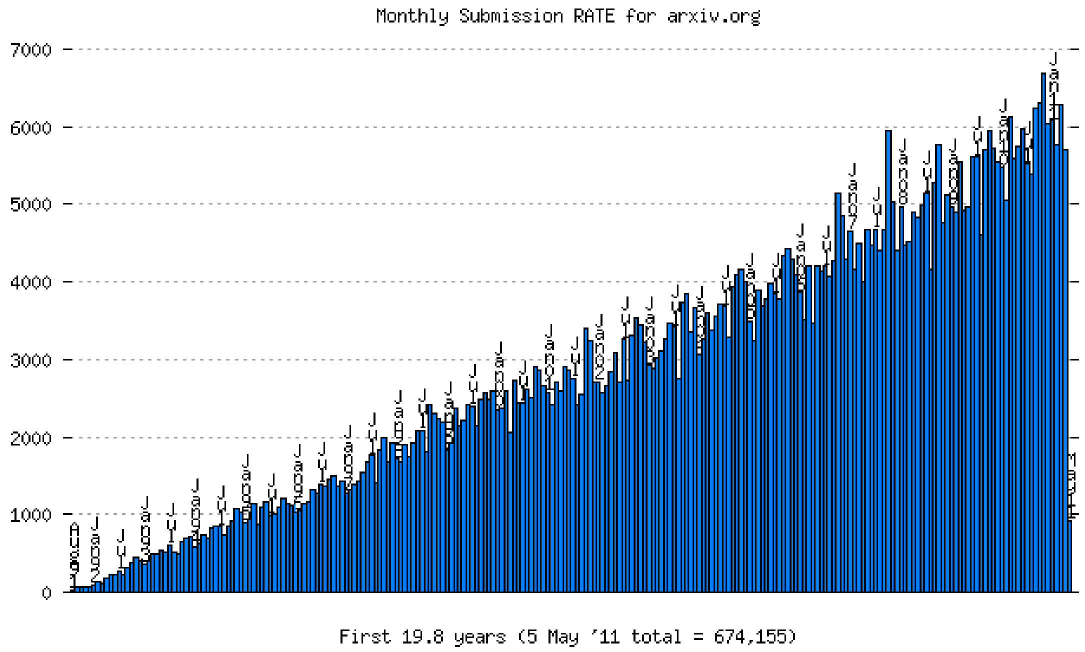


Figure 1.1: Monthly submission rates on arXiv for the period 1991–2011.

increase throughout. Fig. 1.2 also shows that although mathematics and mathematical physics trailed for the first 15 years, average monthly submission in that area is currently about 1.5 times the larger areas of physics.

On an average weekday in early May 2011, the number of hits on the main arXiv site passed one million around 6 pm EST. A wealth of information on user activity is recorded every day. ArXiv has a version of the Osmot[7] search engine which logs user submitted queries. As a system that has been functional, growing and used heavily, arXiv provides an ideal testbed to invesigate scholarly behavior and social interaction on information systems in general.

[7] http://radlinski.org/osmot/

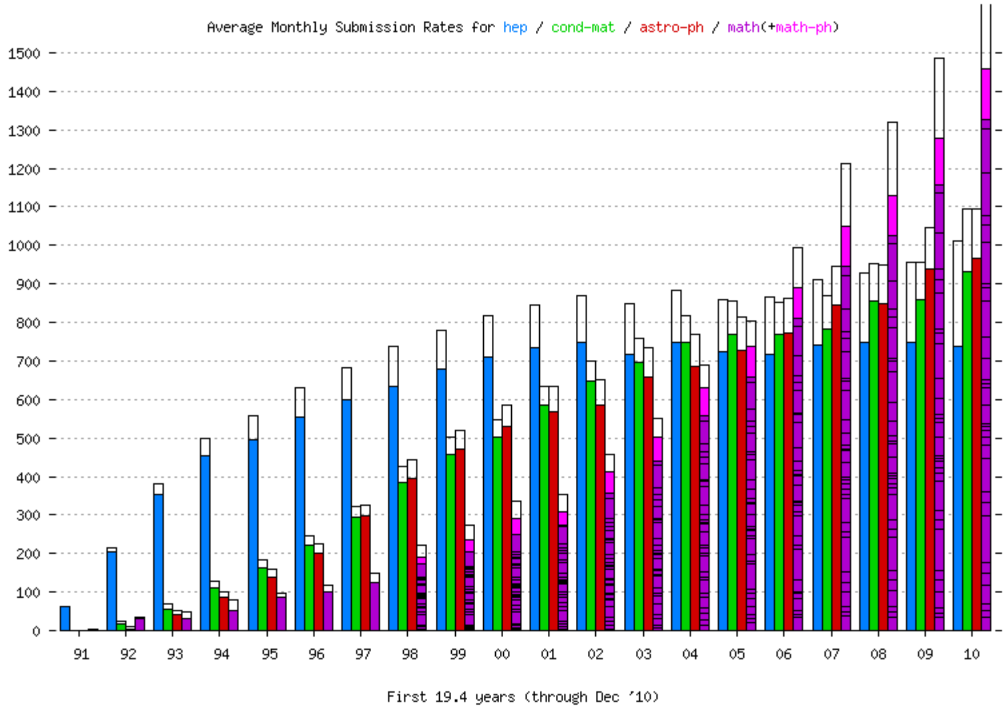


Figure 1.2: Average monthly submissions on arXiv for high-energy physics, condensed matter physics, astrophysics and mathematics combined with mathematical physics. Standard deviations are shown in white.

1.5 Contributions

This dissertation presents a set of methods to perform interesting tasks using various sources of information on scholarly communication systems. Such systems are specializations of modern information systems with many similar components. The important contributions of this dissertation are as follows.

- Chapters 2 and 3 discuss quantification of social phenomena such as self-promotion, visibility, geographic effects and procrastination through metadata analysis and demonstrates the predictive power of user clicks by supervised machine learning.

- Chapter 4 presents an algorithm to extract subtopical concepts from scientific text using a combination of statistical and network analysis. A novel paradigm of validating through implicit user feedback is also presented.

- Chapter 5 presents a clustering scheme to summarize and aggregate topic interest patterns in scientific corpora.

- Chapter 6 presents an author name disambiguation algorithm that utilizes simple features such as co-authorship and self-citation. A novel approach of network node role based sampling and validation is shown to be appropriate for network analysis.

- Chapter 7 briefly discusses how graph algorithms can be adapted for the Map-Reduce framework of distributed computing. An efficient Map-Reduce formulation of Pagerank is presented as an example.

The content of this dissertation overlaps in part with the articles [43, 44, 45, 94, 95, 17] we have published in various journals and conferences in the recent past.

Part II

Positional Effects

CHAPTER 2

POSITIONAL EFFECTS ON CITATION AND READERSHIP

arXiv.org mediates contact with the literature for entire scholarly communities, both through provision of archival access and through daily email and web announcements of new materials, potentially many screenlengths long. In this chapter we confirm and extend a surprising correlation between article position in these initial announcements, ordered by submission time, and later citation impact, due primarily to intentional "self-promotion" on the part of authors. A pure "visibility" effect was also present: the subset of articles accidentally in early positions fared measurably better in the long-term citation record than those lower down. Astrophysics articles announced in position 1, for example, overall received a median number of citations 83% higher, while those there accidentally had a 44% visibility boost. For two large subcommunities of theoretical high energy physics, hep-th and hep-ph articles announced in position 1 had median numbers of citations 50% and 100% larger than for positions 5–15, and the subsets there accidentally had visibility boosts of 38% and 71%.

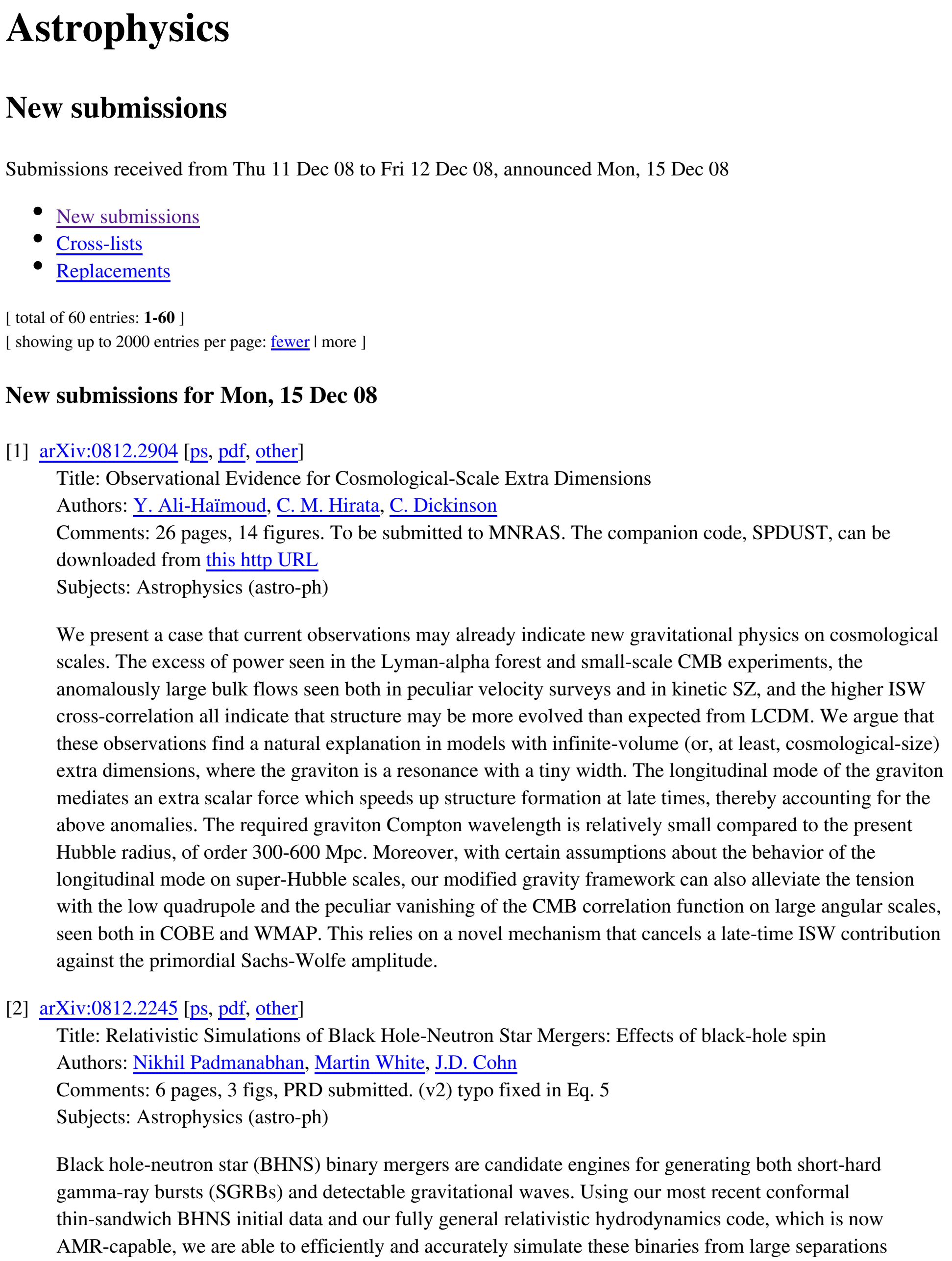
We also consider the positional effects on early readership. The median numbers of early full text downloads for astro-ph, hep-th, and hep-ph articles announced in position 1 were 82%, 61%, and 58% higher than for lower positions, respectively, and those there accidentally had medians visibility-boosted by 53%, 44%, and 46%. Finally, we correlate a variety of readership features with long-term citations, using machine learning methods, thereby extending previous results on the predictive power of early readership in a broader context. We conclude with some observations on impact metrics and dangers of recommender mechanisms.

2.1 Introduction

The arXiv[1] repository currently contains over 600,000 documents and is growing at a rate of over 60,000 new submissions per year. For two decades it has been the primary means of access to the research literature in many fields of physics and in some related fields. Its log data provides the basis for many studies of user behavior during this unique transition period from print to electronic medium. The arXiv corpus is divided into different subject areas, with corresponding constituent subcommunities. Each of these subcommunities receives notifications each weekday of new articles received in the relevant subject area, either by subscription to email announcements or by checking the web page of newly received submissions in the relevant subject area, updated daily (or, equivalently, through the associated RSS feed). These daily listings, viewed either through a web browser or email client, consist of standard metadata, including title and author information, and as well the full abstracts. As depicted in fig. 2.1, this means that it is necessary to scroll down to see beyond the entry in the second position, and to scroll down many times to see the entries in positions near the end of the daily announcements. While the overall order of articles is retained when browsing through the archival monthly listings, no trace of the boundaries between days is retained, hence the daily positional information is lost, and of course articles retain no vestige of their position in original daily announcement when retrieved via the search interface.

In this chapter we investigate the effect of article position in these daily an-

[1] http://arXiv.org/. For a recent overview, see [35].



Astrophysics

New submissions

Submissions received from Thu 11 Dec 08 to Fri 12 Dec 08, announced Mon, 15 Dec 08

- New submissions
- Cross-lists
- Replacements

[total of 60 entries: **1-60**]
[showing up to 2000 entries per page: fewer | more]

New submissions for Mon, 15 Dec 08

[1] arXiv:0812.2904 [ps, pdf, other]
Title: Observational Evidence for Cosmological-Scale Extra Dimensions
Authors: Y. Ali-Haïmoud, C. M. Hirata, C. Dickinson
Comments: 26 pages, 14 figures. To be submitted to MNRAS. The companion code, SPDUST, can be downloaded from this http URL
Subjects: Astrophysics (astro-ph)

We present a case that current observations may already indicate new gravitational physics on cosmological scales. The excess of power seen in the Lyman-alpha forest and small-scale CMB experiments, the anomalously large bulk flows seen both in peculiar velocity surveys and in kinetic SZ, and the higher ISW cross-correlation all indicate that structure may be more evolved than expected from LCDM. We argue that these observations find a natural explanation in models with infinite-volume (or, at least, cosmological-size) extra dimensions, where the graviton is a resonance with a tiny width. The longitudinal mode of the graviton mediates an extra scalar force which speeds up structure formation at late times, thereby accounting for the above anomalies. The required graviton Compton wavelength is relatively small compared to the present Hubble radius, of order 300-600 Mpc. Moreover, with certain assumptions about the behavior of the longitudinal mode on super-Hubble scales, our modified gravity framework can also alleviate the tension with the low quadrupole and the peculiar vanishing of the CMB correlation function on large angular scales, seen both in COBE and WMAP. This relies on a novel mechanism that cancels a late-time ISW contribution against the primordial Sachs-Wolfe amplitude.

[2] arXiv:0812.2245 [ps, pdf, other]
Title: Relativistic Simulations of Black Hole-Neutron Star Mergers: Effects of black-hole spin
Authors: Nikhil Padmanabhan, Martin White, J.D. Cohn
Comments: 6 pages, 3 figs, PRD submitted. (v2) typo fixed in Eq. 5
Subjects: Astrophysics (astro-ph)

Black hole-neutron star (BHNS) binary mergers are candidate engines for generating both short-hard gamma-ray bursts (SGRBs) and detectable gravitational waves. Using our most recent conformal thin-sandwich BHNS initial data and our fully general relativistic hydrodynamics code, which is now AMR-capable, we are able to efficiently and accurately simulate these binaries from large separations

Figure 2.1: New astro-ph listings, from http://arXiv.org/list/astro-ph/new. Note that a standard sized Web or e-mail browser window may not accommodate even the full entries in the first two positions without requiring scrolling down. The astro-ph listings averaged roughly thirty such entries every weekday during the period studied here.

nouncements for certain physics subfields of arXiv, a purely short-term phenomenon, on citations received over the long-term. This effect for the astro-ph subject area, primarily used by astrophysicists, was first considered in [26, 25]. Here we will consider as well two other communities of users, those of the hep-th and hep-ph subject areas ("High Energy Physics – Theory" and "High Energy Physics – Phenomenology"). The hep-th subject area is the original arXiv subject area initiated in mid 1991, covering highly theoretical areas of particle physics such as string theory. The hep-ph subject area was started in early 1992, covering areas of theoretical particle physics more directly related to experiment. During the 2002–2004 periods to be studied here, hep-th and hep-ph received an average of roughly 3320 and 4110 new submissions per year, respectively. The astro-ph area, started later in 1992, is an amalgam of many types of relevant theory and experiment, from stellar to galactic to cosmological, and by 2005 had grown to exceed the combined size of the High Energy Physics subject areas.[2] The astro-ph subject area averaged roughly 7720 new submissions per year from 2002–2004, and grew to over 9000 new submissions per year in 2005–2006.

A strong correlation between the position of articles in their initial announcement and the number of citations later received was found in [26, 25]. Since position in the daily announcement of newly received submissions is a one-day artifact, visible only that day and with no trace afterwards, it is extraordinarily surprising that it could nonetheless be correlated with long-term citation counts, accumulated years later. Due to the weight given to citations as a measure of research impact, it is important to verify such an unexpected effect by different methods, and assess whether some analog exists as well in other communities. Our results here confirm the effect discovered in [26, 25], and suggest that arXiv

[2] http://arxiv.org/Stats/hcamonthly.html

subject area organization and interface design should be reconsidered either to utilize or counter such unintentional biases.

It is evident to readers that a fraction of authors, working entirely within the established operating procedures for the site, has been jockeying for top position in the daily announcements. Since late 2001, the policy has been that submissions received until 16:00 US eastern time (EST/EDT) on a given weekday are announced at 20:00 eastern time, and submissions received after that deadline are announced the following day, in rough[3] order of receipt. Articles submitted shortly after 16:00 will thus be listed at or near the top of the next day's announcement, and will potentially receive greater visibility. Submitters are evidently conforming their schedules to take advantage of some presumed benefit to the greater visibility afforded by submitting within this time window.

Fig. 2.2 shows the submission counts, broken down by the time of submission, of arXiv:astro-ph from the beginning of 2002 through the end of Mar 2007. The spike in submissions corresponds to the period 16:00–16:10. That ten minute bin contains 5 times as many submissions as any other bin outside of the 16:00–16:30 period. Other variations during the day visible in the figure correlate with periodicity of overall activity levels, resulting from the effects of users in different timezones. The period between 10 a.m. and noon eastern time, for example, corresponds to late afternoon in western Europe and early morning in western U.S. The server itself is not affected by any excessive operating load during 16:00 period, since the submissions are automatically serialized by time of receipt. Typical submissions take under a second to process, and no noticeable processing queue develops from the [at most] few tens of submissions in that initial minute on a busy day, while the server simultaneously processes

[3] See sec. 2.2.1 for important exceptions.

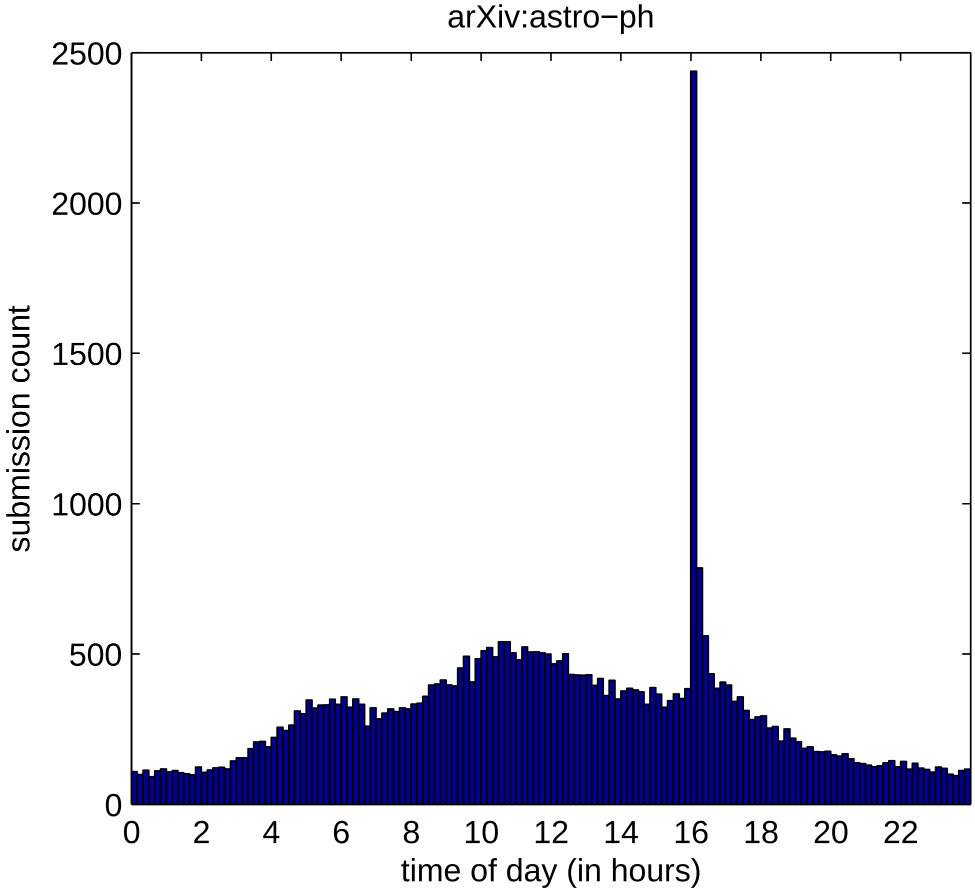


Figure 2.2: Number of astro-ph submissions by time of day, in 10 minute bins, during the period Jan 2002 – Mar 2007.

multiple retrievals and searches per second. The average submission rate during the rest of the day is roughly one new submission every six minutes.

It is important to note that the positional effects are potentially much more dramatic than, say, the corresponding effects in presentation of search results. In the latter case, typically ten results are presented on a single web page, with each result entry reduced to a small number of lines of key text. Eye-tracking studies [39] have shown the extent to which users nonetheless tend to focus only on the top few entries. In the case of arXiv announcements, on the other hand, the entries consist of entire abstracts (see fig. 2.1). Only the first two entries are visible in a standard sized Web or e-mail browser window, and it is necessary to scroll down to see the remainder. The situation is thus more comparable to viewing successive pages of search results, where for example analysis of log data in [32] suggested a click probability that decreased with result rank as $r^{-1.63}$.

In the sections below, we consider the positional effects on both citation and readership, in an attempt to understand author and reader behavior, and ascertain whether the policies of the arXiv system itself need modification to counter any unexpected long-term consequences of a seeming short-term artifact.

2.2 Effects on Citation

2.2.1 Previous Work

[26] used the SPIRES High-Energy Physics Literature Database[4] to reconstruct the daily arXiv astro-ph mailings from Jul 2002 through Dec 2005, giving the articles at least a year to gather citations. The citations were collected from the SAO/NASA Astrophysics Data System (ADS) bibliographic services[5] in December 2006. It was inferred that on average, articles in positions 1 get 89.8 ± 9.0 citations, while those in positions 10–40 get 44.6 ± 0.9 citations. Three possible explanations were suggested for this: self-promotion bias (SP), visibility bias (V), and geographic bias (G). The self-promotion argument assumes that authors can intuit in advance the quality of their articles and specifically aim to promote the better ones through early submission. This is related in spirit to the

[4] http://www.slac.stanford.edu/spires/hep/

[5] http://adsabs.harvard.edu/

‘self-selection’ postulate [57] , which suggests that more prestigious articles, i.e., those more likely to be cited, are more likely to be made freely accessible. In the current context, the suggestion is that those articles are as well promoted by authors to the top of a daily list of new freely accessible articles. Enough of these higher quality articles are submitted in the critical time window to result in the measured citation advantage for submissions in the first few positions. The visibility argument is that the initial higher visibility translates to higher readership, and some fraction of that higher readership translates to higher citations later on. The geographic argument is that articles submitted during the critical period are more likely to come from North America due to timezone differences, and those might be more likely to be cited for other reasons. Comparing overall citation trajectories of submissions from Europe and North America, however, permitted exclusion of the geographic bias in [26]. Our investigation of this bias is presented in chapter 3 and is not considered further here.

Using submission times later provided from arXiv log data, a subsquent comparison of three sets of articles was undertaken in [25] to disentangle the SP and V biases. The first set contained articles that appeared in the first three positions and were submitted within the first five minutes after the deadline, hence inferred to have been submitted with an intention to be listed at or near the top. The second set contained articles that were submitted after the first ninety minutes, and yet appeared in the first three positions.[6] These are assumed not to be self-promoted. The last set contained articles in positions 26–30. It was observed that the self-promoted articles received more citations than those in the other two sets. The articles that fortuitously appeared near the top, however,

[6] This can happen either because there were few or no early submissions, or because an administrative removal of an early article caused a later submitted article to be shifted to that earlier position to fill the gap.

also appear to receive more citations than had they appeared in a lower position, indicating as well some visibility bias. The increase in citations due to the visibility bias was found to be smaller than that due to the self-promotion bias.

The methodology used in [26, 25] to quantify the citation effects involves fitting the citation distributions to a power law, excluding the regimes of data that do not follow the power law (the head and the tail of the distributions), and averaging the rest. Power law fitting can be tricky [71], and as described in Appendix A.1, the above methodology results in inadvertent biases, including using only a portion of the data. Due to the sociological importance of the result, it is useful to reconsider the results of [26, 25] using slightly different methods.

2.2.2 Methodology

For heavy-tailed distributions such as power laws, the mean can be strongly affected by the large values at the tail. A more robust statistic is the median, which is not affected by the large values, and is also representative of the large number of small values in the sample set. For this reason, nonparametric statistical methods often use the median. More generally, we can consider the k^{th} percentile as the aggregate measure of a set of values. If the quartiles (25th and 75th percentiles, usually denoted $Q1$ and $Q3$, respectively) and the median of a distribution are larger than the same quantities of another distribution (at a statistically significant level), then stochastic dominance (see Appendix A.1) is likely. The interquartile range (the difference $Q3-Q1$) measures the spread of the distribution, analogous to the variance of a normal distribution. We analyze the citation data by presenting plots of the median and the quartiles and check for statistical significance, using the nonparametric Mann-Whitney U (also known as Wilcoxon rank-sum test) and Kolmogorov-Smirnov tests [33].

We consider the 23,165 arXiv astro-ph articles from the beginning of 2002 through the end of 2004, announced in 777 daily announcements (via one-time email announcements and web pages daily updated), with an average mailing containing 29.8 papers. The citations were collected from NASA's Astrophysics Data System (ADS) Bibliographic Services in August 2008, giving the articles over three and half years to gather citations. There are thus 777 articles in each the top positions and roughly that number in the rest of the positions, at least up to the typical number per announcement.

Fig. 2.3 shows the median citations and quartiles for each position. The later positions are binned to reduce noise. From position 1, the median decreases until position 5, and beyond position 7 the medians effectively cease changing. The upper quartile (upper boundary of boxes) shows a more pronounced decreasing trend. Even the lower quartiles (lower boundary of boxes) show a decreasing trend. Statistical significance of these differences is assessed in Appendix A.2.

2.2.3 Self-Promotion vs. Visibility

We now consider the SP and V contributions to increased citations, taking a different approach from that of [25], as described in sec. 2.2.1.

In the astro-ph dataset, we mark those articles submitted in the first 10 minutes after the deadline as "early" (E), a time period chosen from fig. 2.2. Of the 23,165 articles, 1049 were marked as E, and the vast majority of those are

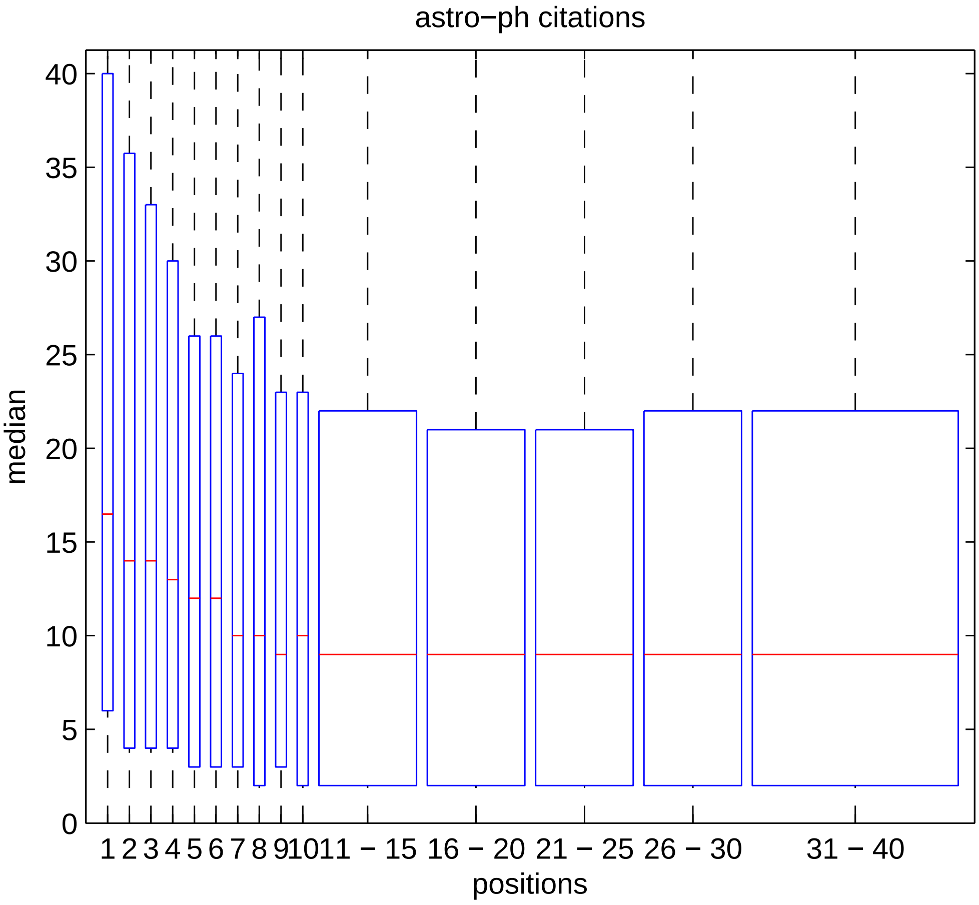


Figure 2.3: Box plot of citations for different positions in astro-ph. Boxes represent the interquartile range, bounded above and below by the third and first quartiles, and the red horizontal lines mark the medians.

likely to be self-promoted. The articles submitted after the first 30 minutes after deadline were marked “not early” (NE). The submitters of these are inferred to be indifferent about the position in the announcements. 643 articles submitted after the first 10 minutes but before the first 30 minutes after deadline were considered as ambiguous in author intent, so omitted from the analysis (which biases the results in neither direction).

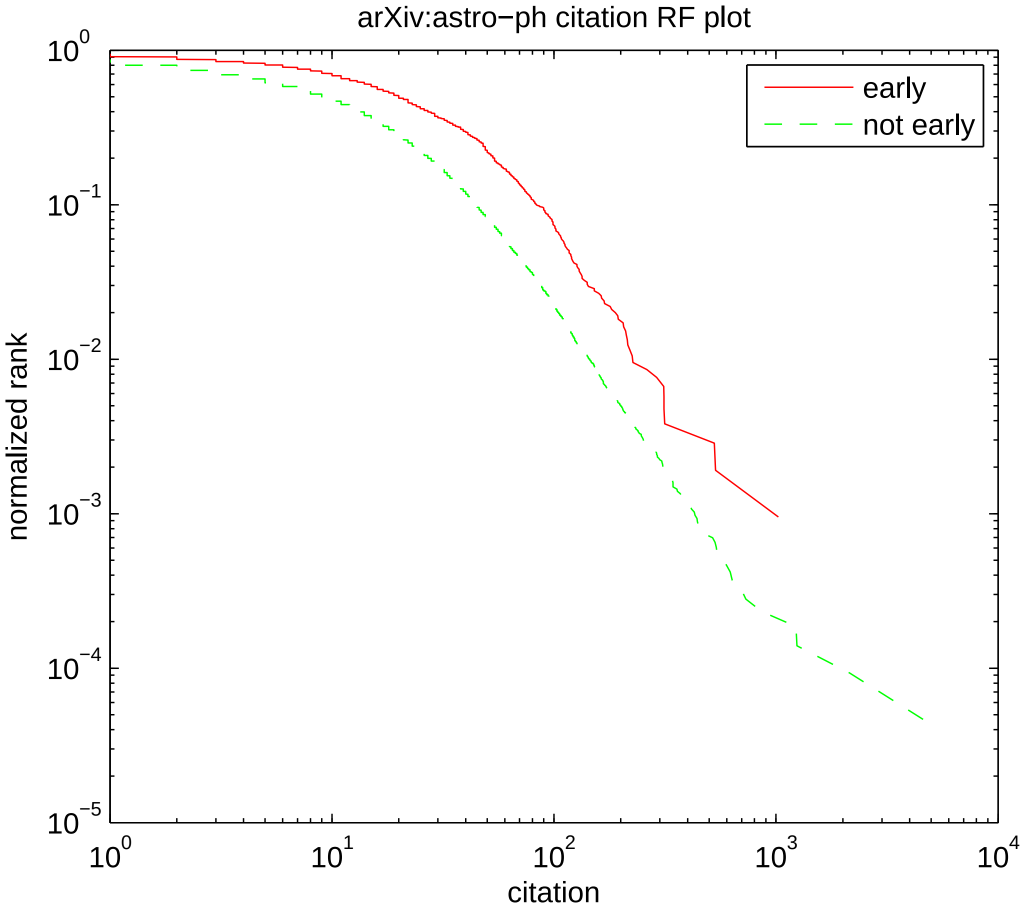


Figure 2.4: Rank-Frequency (RF) plot for astro-ph citations. The solid line is for articles submitted within the first 10 minutes after the weekday deadline of 16:00 eastern time. The dashed line is for the articles submitted after the first 30 minutes.

The median citation of the E articles is 20 while that of the NE articles is 9, and the difference in medians is significant using the MWU test at 1% significance level. The KS test at 1% significance level shows that the E citation distribution is as well higher than the NE distribution, in the global sense described in Appendix A.2. The rank–frequency (RF) plots of the two citation distributions, depicted in fig. 2.4, indicate that self-promoting submitters by-and-large do have a good intuition for the likely future impact of their articles. Not all self-promoted articles, however, receive high citations: roughly 10% of the E articles in position 1 have no more than 1 citation.

	Position	1	2	3	4	5
astro-ph:	Early	510	289	146	64	24
	Not Early	147	299	484	613	694

Table 2.1: Number of E and NE articles in arXiv:astro-ph listed at positions 1–5 during the 2002–2004 timeframe.

To further probe the two biases, we separate out the E articles at each position. For the top 5 positions, the numbers of articles are shown in table 2.1. Fig. 2.5 shows the median number of citations for each position. The red bars (E articles) characterize the self-promotion effect, while the green bars (NE articles) characterize the visibility effect. At every position, we see that the effect of self-promotion is much stronger than that of visibility, a difference significant at 1% level (MWU test) for the first 4 positions. The citation advantage of the top few positions is thus largely due to self-promotion, but as we shall see there is as well a visibility effect.

The differences between the blue bars in positions 1 and 5 in fig. 2.5 is statistically significant (MWU test at 5% level), and while it is likely that this difference is entirely due to the SP effect, there is not enough data in the green bars to make a statistically significant statement (at the same 5% level). But we can compare these articles to ones that appeared in lower positions. The median citation of articles in positions 10–40 is 9, while the median citation of non-SP articles in positions 1-3, i.e., submitted after the first 30 minutes, is 12. This difference of 3 citations (significant at the 1% level) is the extent to which visibility bias contributes to citations. The non-SP articles are randomly selected, independent of authorship, length, subject area with Astrophysics, or other confounding qual-

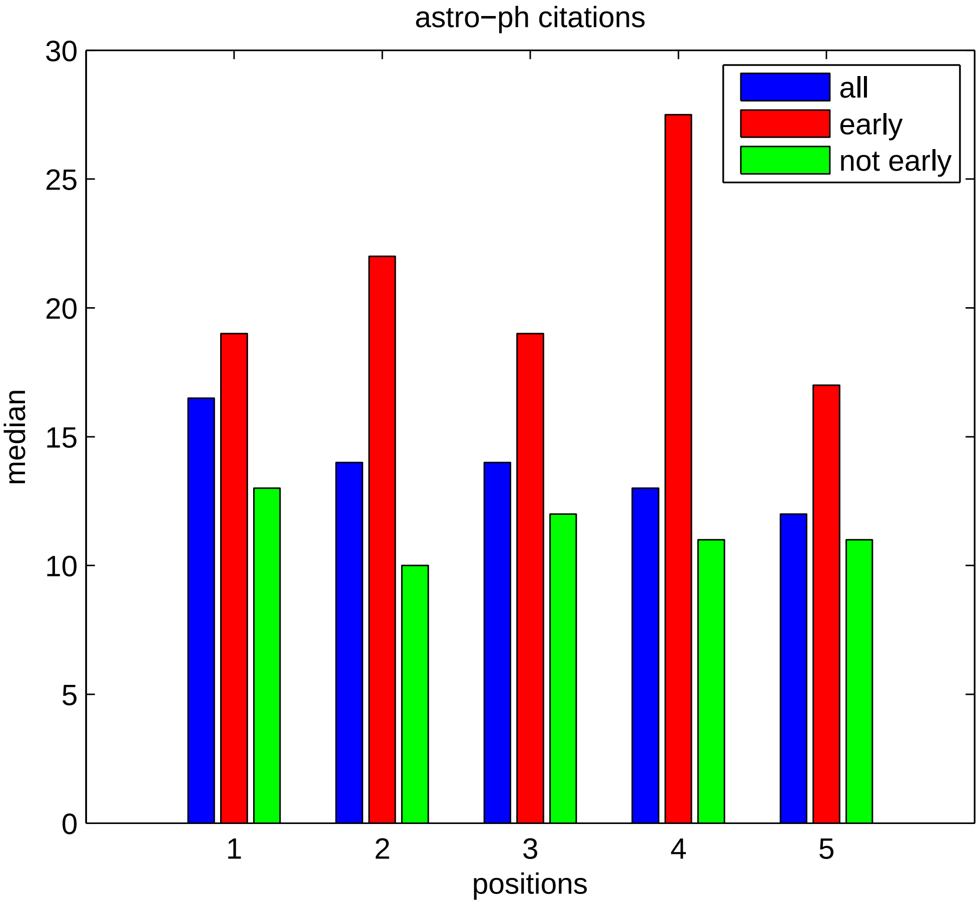


Figure 2.5: Median citations for each position for astro-ph announcements from the beginning of 2002 through the end of 2004. The red bars represented the 'self-promoted' articles. The non-self-promoted articles in the top few positions, represented by the green bars, nonetheless receive more median citations than those lower down in announcements.

ity factors, yet solely by virtue of having appeared near the top of a web page or email announcement on one single day, are measured to receive significantly more median citations many years later.[7]

[7] For comparison with the bins used by [26, 25], articles announced in astro-ph positions 1–6, received a median of 14 citations, 55% higher than the median of 9 for those in positions 10–40. The NE articles in positions 1–6 received a median of 11 citations, pure visibility still giving 22% more median citations than those lower down.

We have also analyzed the data for the full period in fig. 2.2, i.e., the 43,686 astro-ph articles from the beginning from 2002 till the end of March 2007 announced in 1350 mailings. With citations again collected from ADS in August 2008, this gave at least roughly a year and half for the most recent articles to gather citations. The resulting graph has the same general form as fig. 2.5, with greater significance and medians only 10% to 20% smaller. But the "median number of citations" for the enlarged dataset doesn't correspond to any particular set of articles, because it involves an average over articles of vastly different ages, with as much as six and a half years to as little as a year and a half to collect citations. For this reason we used the smaller data set for which the medians do correspond to median numbers of citations for 4.5–6.5 years old, and don't change appreciably when we further restrict the time window of articles considered. The early timeframe was chosen for stable citation data, although the SP effect became increasingly pronounced in the later data.

2.2.4 hep-th and hep-ph

Having confirmed the self-promotion phenomenon in the astro-ph subject area, we now consider the hep-th and hep-ph subject areas: the largest and most active of arXiv's high energy physics areas. The 776 daily announcements for those areas during the Jan 2002–Dec 2004 period had averages of 12.8 and 15.9 articles, respectively.

Figs. 2.6, 2.7 show the number of hep-th and hep-ph submissions from the beginning of 2002 through Mar 2007, in 20 minute submission bins. The first 20 minutes after 16:00 eastern time have exceptionally high submission rates,

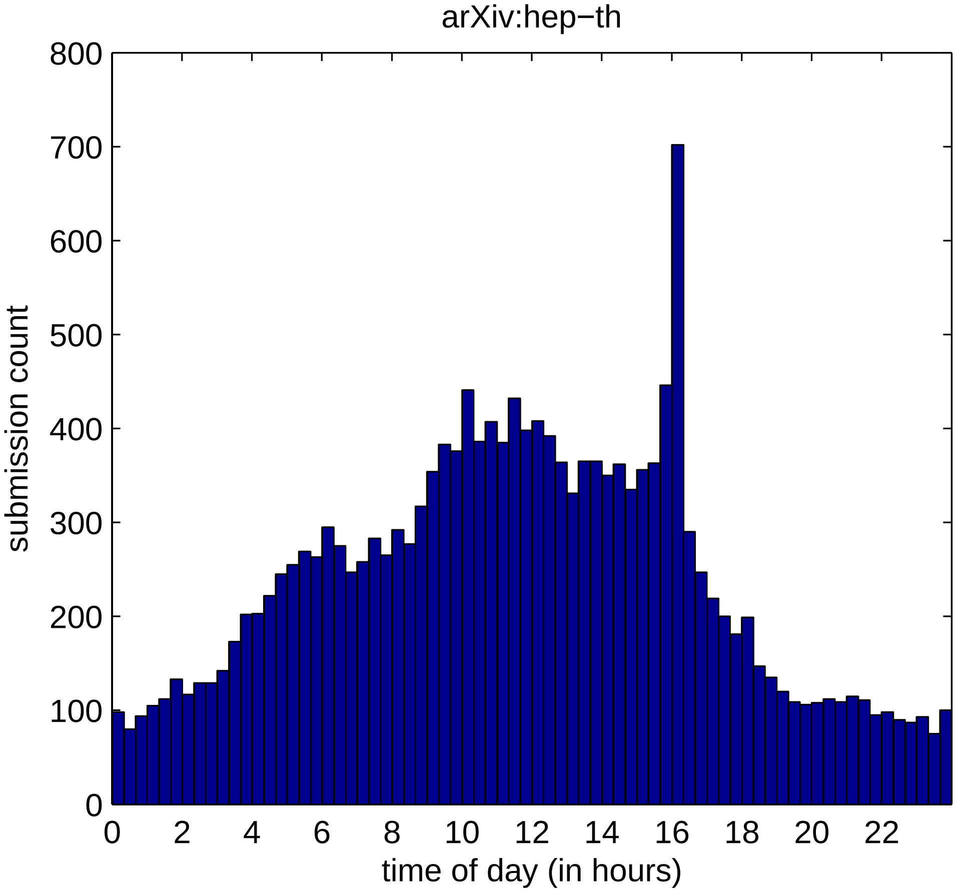


Figure 2.6: Number of hep-th submissions by time of day, in 20 minute bins, during the period Jan 2002 – Mar 2007.

although not as high as astro-ph (fig. 2.2). Articles submitted in this 20 minute period are considered early (E) and the rest are considered not early (NE). We use the articles submitted from Jan 2002 till Dec 2004 for our analysis, for reasons discussed at the end of the previous subsection. Of the 9,932 total hep-th submissions during this period, 309 were submitted during the first 20 minutes and marked as E; and of the corresponding 12,281 hep-ph articles, 363 are marked as E, a similar percentage as for hep-th. Citations were collected from the SPIRES High-Energy Physics Literature Database in September 2008, giving the articles over three and half years to accumulate citations. The high energy physics lit-

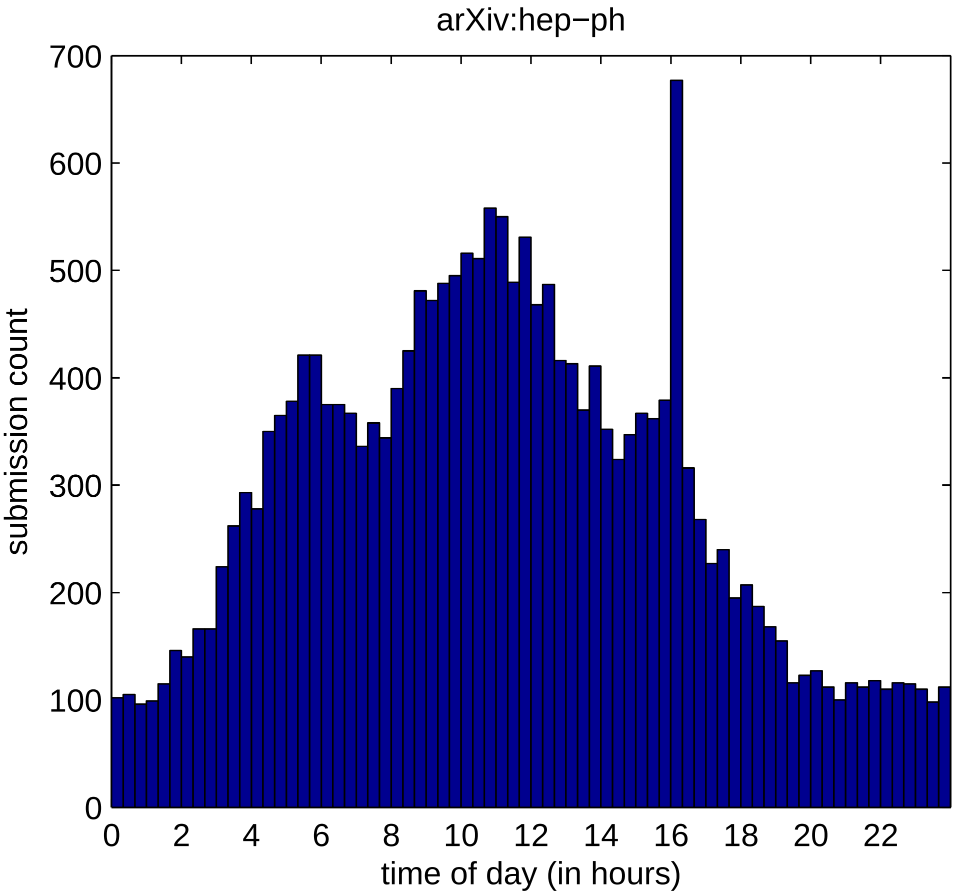


Figure 2.7: Number of hep-ph submissions by time of day, in 20 minute bins, during the period Jan 2002 – Mar 2007.

erature, like the astrophysics literature, is served by a relatively small number of conventional published journals, and dominated by a very small number of very large ones.

The early hep-th and hep-ph articles are interpreted as self-promoted and, as seen in figs. 2.8, 2.9 their citation distribution stochastically dominates the rest (KS test at 1% significance level). The median citation for hep-th position 1 is 12, while the median citation for positions 4–15 is a significantly lower 8. Similarly for hep-ph, articles at position 1 have a median citation of 14, while articles at positions 4–15 have a median citation of 7.

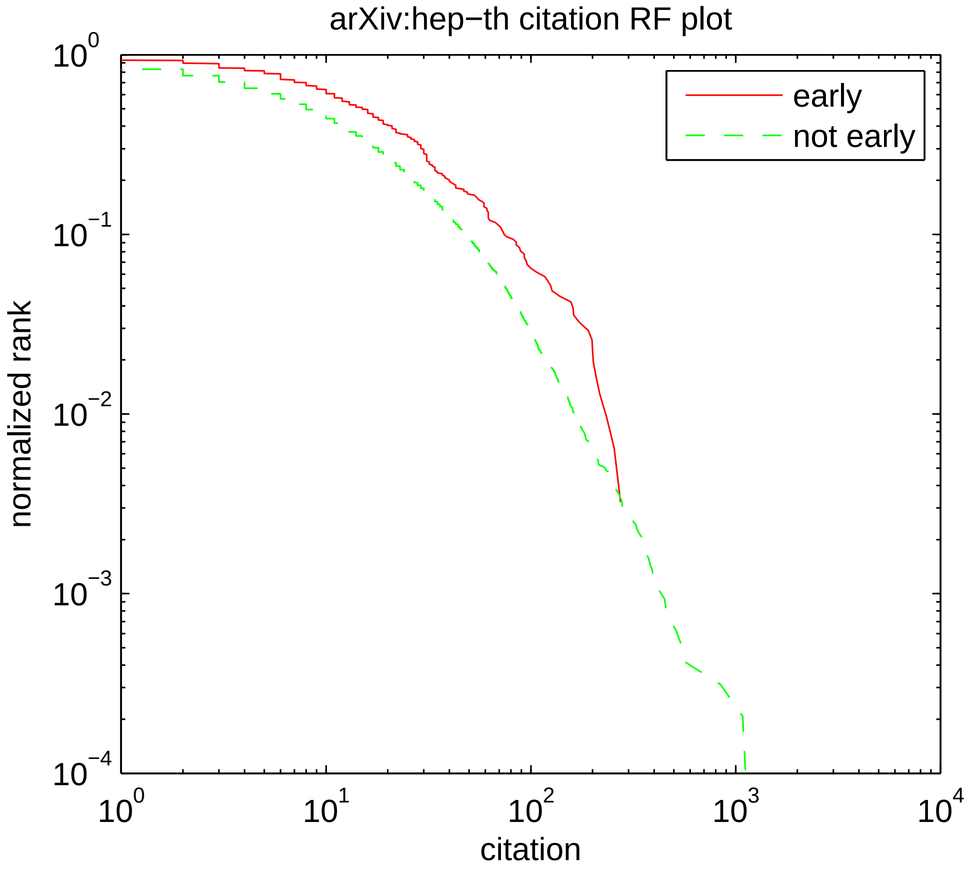


Figure 2.8: Rank-Frequency (RF) plot for hep-th citations. The solid line represents E articles, submitted within the first 20 minutes after the 16:00 eastern time weekday deadline. The dashed line is for the remaining articles.

hep-th:

Position	1	2	3
Early	237	58	11
Not early	537	715	759

hep-ph:

Position	1	2	3
Early	282	67	12
Not early	492	703	756

Table 2.2: Number of articles in arXiv:hep-th and arXiv:hep-ph listed at positions 1–3 during the 2002–2004 timeframe.

Figs. 2.10, 2.11 show the medians and the quartiles of different hep-th and hep-ph positions. The first two positions have median number of citations significantly higher (at the 1% level) than the lower positions, and the difference between positions 1 and 2 is particularly striking. Figs. 2.12, 2.13 disentangle

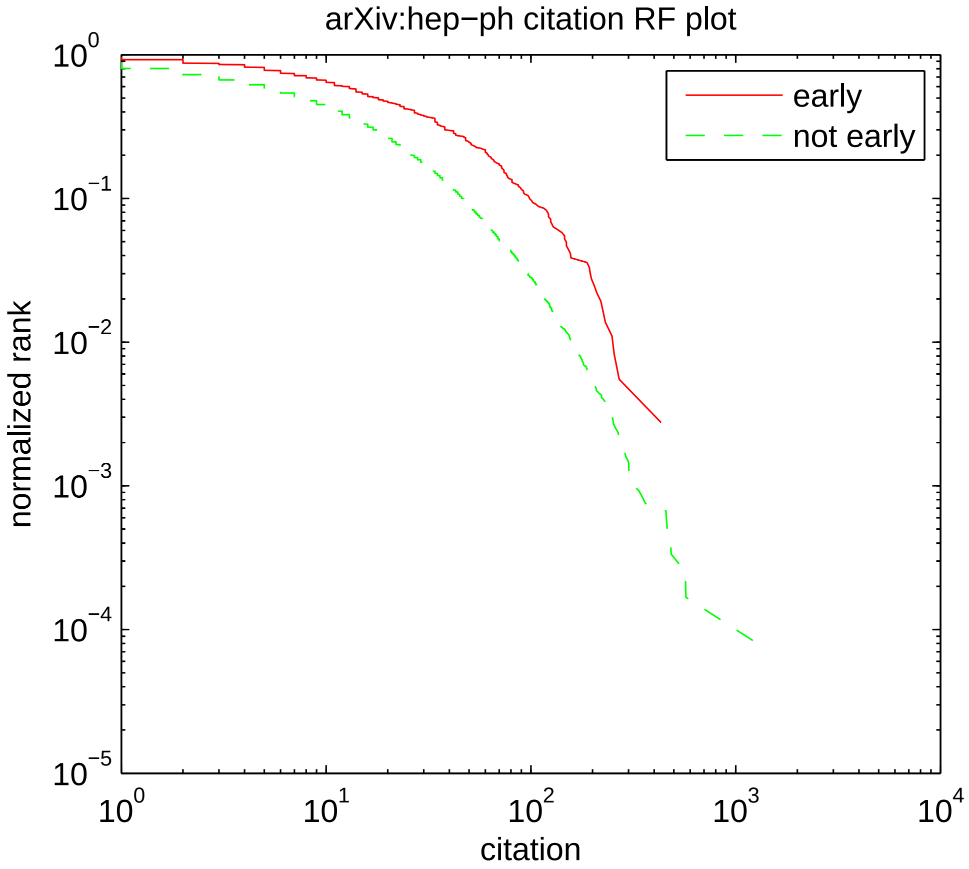


Figure 2.9: Rank-Frequency (RF) plot for hep-ph citations. The solid line represents E articles, submitted within the first 20 minutes after the 16:00 eastern time weekday deadline. The dashed line is for the remaining articles.

self-promotion and visibility effects and, as in astro-ph, the self-promotion effect (red bars) dominates over the visibility effect (green bars), significant (1% level) for the first 2 positions. The effect is quite striking for the first position. The number of articles at each position is shown in table 2.2. Note that since there were only 11 and 12 early articles at position 3, respectively, for hep-th and hep-ph, the red bars for this case in Figs. 2.12, 2.13 are not statistically significant (and similarly for positions 4 and beyond).

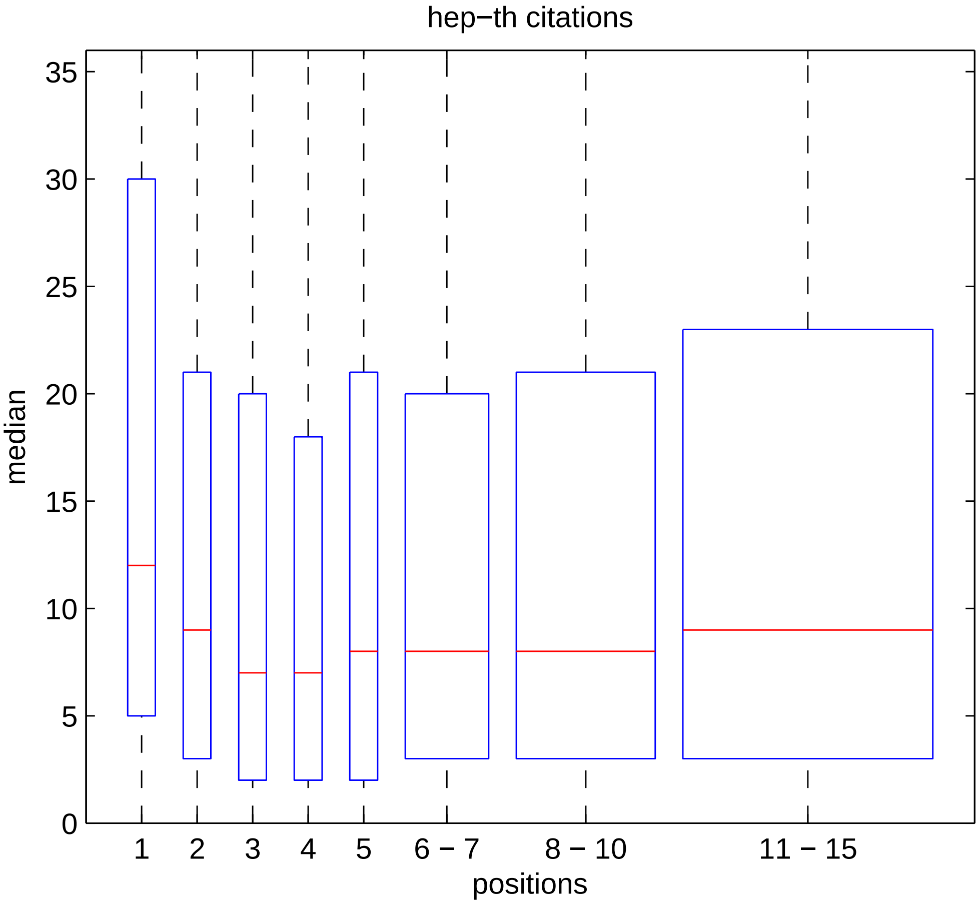


Figure 2.10: Box plot of citations for different positions in hep-th. Boxes depict the interquartile range, and the red lines mark the medians.

Visibility

Although self-promotion is the dominant effect in the positional citation advantage in each of astro-ph, hep-th and hep-ph (figs. 2.5, 2.12, 2.13), there was a pure visibility effect in the astro-ph data and here we find it as well in the hep-th and hep-ph data. For hep-th, the articles in position 1, but not early (green bar in fig. 2.12), have a median of 11 citations. Articles in positions 5–10 have a median of 8 citations. This difference is significant at the 1% level. Similarly,

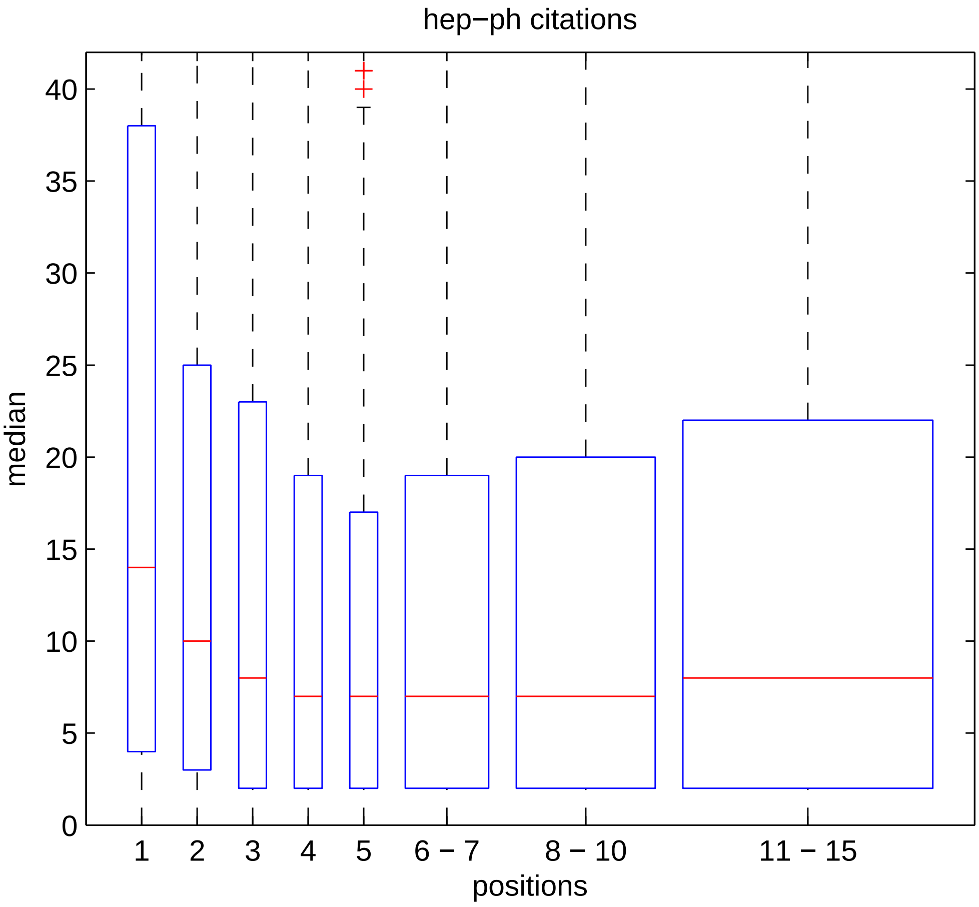


Figure 2.11: Box plot of citations for different positions in hep-ph. Boxes depict the interquartile range, and the red lines mark the medians.

for hep-ph, the articles in position 1, but not early (green bar in fig. 2.13), have a significant median citation advantage of 12 – 7 = 5 citations over the articles in position 5–10. The falling trends in the green bars in figs. 2.12, 2.13 capture the beginning of this visibility effect.

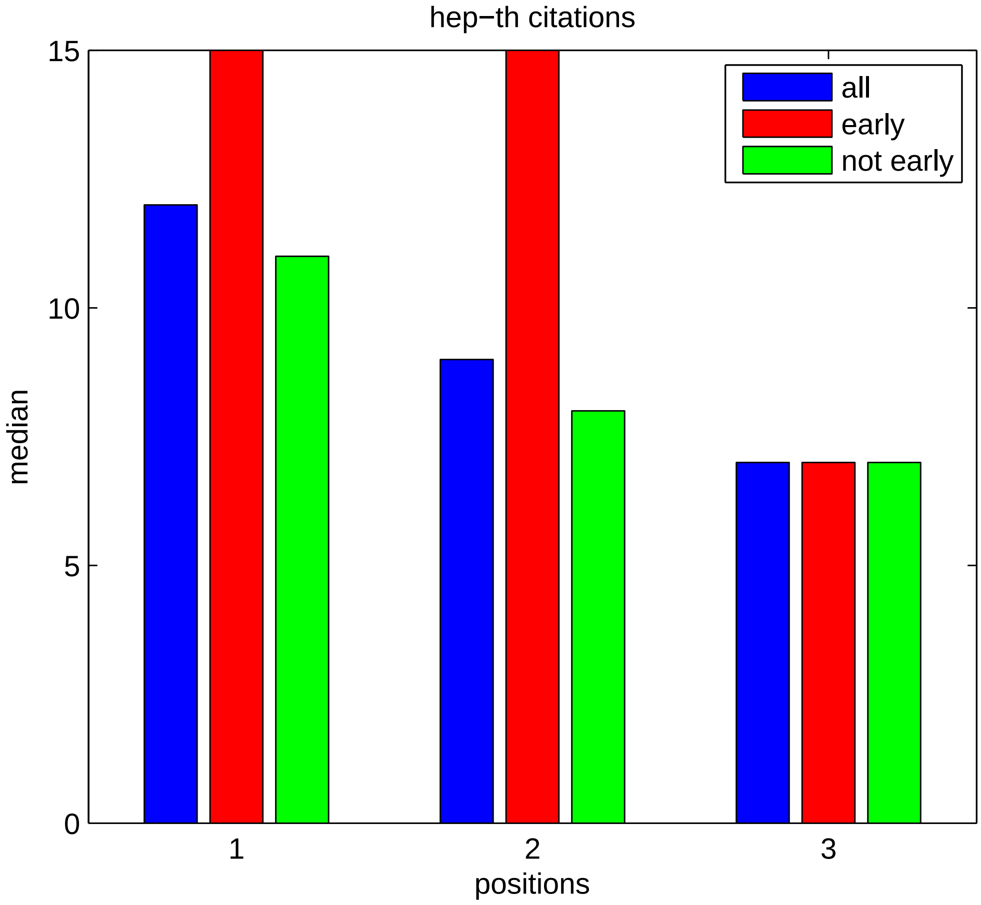


Figure 2.12: Median citations for each position in hep-th, for announcements from the beginning of 2002 through the end of 2004. The red bars represented the 'self-promoted' articles.

2.2.5 Discussion

It is not within the purview of this dissertation to attempt a detailed explanation of why a one-time visibility would leave its trace in the citation record years later. As we shall see in the readership data in the next section, articles in the top few positions receive more initial downloads, whether or not submitted early (i.e., self-promoted). The extra initial readership may probabilistically translate into a few early citations, which in turn could cascade into more cita-

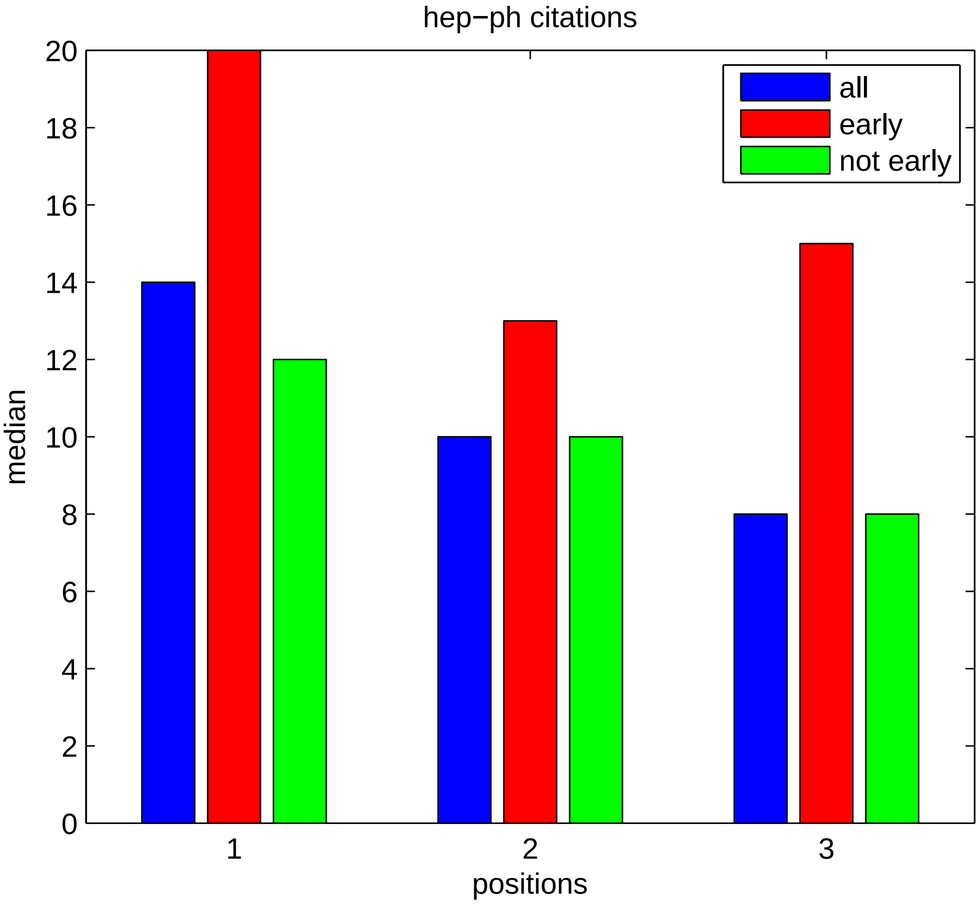


Figure 2.13: Median citations for each position in hep-ph, for announcements from the beginning of 2002 through the end of 2004. The red bars represented the 'self-promoted' articles.

tions later on. We could hope to model this in terms of some set of "fungible" articles, more or less similar in quality and subject area, with the ones cited determined by something of a social convention, based on artifactual collective effects within the citing community. This would parallel the behavior seen in studies of how social influence affects individual decisions and collective outcome in social networks[80].

Citation practices differ from discipline to discipline, and there are many known pitfalls of citation as measure of quality. Studies of subsets of geoscience [87], astrophysics [2], and demography [93] do at least suggest that citations primarily indicate some form of direct intellectual acknowledgement and information flow, rather than primarily reflecting reputational or other secondary social factors.

But other features are known to be correlated to increased citation, including number of authors[8], number of pages, and also specifically visibility factors such as mainstream media coverage or being featured on a journal front cover. For example, it was shown in [74] that major media coverage alone could lead to increased citations. Control for other factors in that study was provided by a serendipitous period for which there is a newspaper archive of stories that would have appeared, but were not disseminated due to a distribution strike: the journal research articles that would have been featured in those stories do not exhibit the same citation boost as did articles covered during periods of normal newspaper distribution. A similar effect can now be expected from visibility in blogspace, or via publicity in either blogspace or the media and amplified through feedback loops between them. The analyses of [25] and this chapter were similarly able to isolate the role of visibility by exploiting the serendipity of randomly selected articles accidentally accorded high visibility without the conscious intent of the authors.

Since a significant component of the citation effect is nonetheless due to intentional self-promotion, it is natural to wonder whether other forms of additional care taken during the submission process as well correlate with early submission, and hence with more citations in the long term. For example, it is op-

[8] Larger groups could be correlated with more funding and hence better equipment and past track record; see also sec. 2.3.4.

tional for authors to provide their institutional affiliations parenthetically along with their names in the Author field. We find that 63% of the early astro-ph submitters provided affiliations, compared to only 43% of the not early ones. The total length of the metadata fields in arXiv has always been limited to prevent any one submission from monopolizing too much screen space. (Submissions exceeding the limit are automatically rejected until they are within the limit.) But early submitters nonetheless took maximal advantage within the guidelines: the median length of the title for early submissions was 70 characters, compared to 66 for not early ones (the difference significant at 1% level KS), and the median length of abstract was also greater for the earlier submissions, 1177 compared to 1014 characters (i.e., 16 lines compared to 14 in the email announcements, with lines wrapped at the nearest whitespace to under 80 characters per line).

By contrast, early and not early submissions had the same median number of authors (three), the same likelihood of providing initials rather than full first names of authors, and (reassuringly) there was no tendency for authors of early submissions to have longer last names, so the increased length of the overall author field (median of 70 characters compared to 62) was due entirely to the increased tendency of early submitters to provide author affiliations. The greater completeness of metadata and inferred submitter effort also correlates with greater citation impact even among only the non self-promoted articles: for not early submission with author affiliations provided, the median number of citations in the 2002–2004 astro-ph dataset was 10 compared to 9 for those without, a statistically significant difference (1% MWU). (For the early submissions, the median number of citations was also greater for the submissions that provided affiliation, 20 compared to 19, but the difference was not statistically significant in that case.)

The considerations in this chapter are also in principle independent of the 'citation advantage' sometimes postulated for open access articles, since all of the articles in arXiv are equally open access. But if the existence of this one-time visibility effect suggests the possibility of an open access advantage, then any analog of the self-promotion effect (i.e., that articles more likely to be cited are *a priori* more likely to be deposited in an open access site) would have to be eliminated as the underlying cause.

The latter self-selection effect [57] was considered further as a 'quality bias' in [69] and [20], which studied respectively the citation impact of those articles in the Condensed Matter (cond-mat) and Mathematics (math) sections of arXiv later published in journals, as compared to articles in the same journals but not deposited in arXiv. Both 'early view' (advance availability on arXiv prior to publication in journal) and 'quality bias' (higher quality articles more likely to be posted on arXiv) are potential confounding effects that could lead to an artifactual citation advantage, and it was found that correcting for those left no general 'open access advantage' for articles deposited in arXiv. Similarly, in a study of open access articles published in eleven scientific journals, [21] used a randomized controlled trial to eliminate biases from other quality indicators: whether self-archived, featured front cover of journal, received press-release, and other confounding attributes (nature of article, number of authors and geographic location, number of references, article length, journal impact factor), and later estimated their effect. This study as well found that any citation differences were due to factors other than open access *per se*: while those articles randomly assigned open access status received more full text downloads, they were no more likely to be cited a year later.

In Jan 2009, the astro-ph section of arXiv was subdivided into six smaller subsections. It remains possible to receive the combined daily listings for all subsections, but many users expressed a preference to be able to browse only the restricted subsets. This division into smaller announcements will in principle ameliorate some of the positional effects, but not all, since the larger of these subsections still average more than ten new submissions per day. Some users have suggested randomizing the daily order entirely, either uniformly for everyone, or individually for each user. Others have pointed out that such a methodology would potentially do a disservice to readers, who may indeed be benefitting from having self-promoted articles brought preferentially to their attention (presuming those really are the more likely to be of importance in the long-run). Perhaps a better methodology is afforded by personalization, by which users can register to receive daily announcements based on their preferences, and ordered accordingly.[9] These preferences can be indicated via a controlled vocabulary of keywords, or via arbitrary search terms, and can be implemented in combination with data from a user's own past on-line reading behavior at the site, on an opt-in basis.

2.3 Readership Data

Since citations can signify some long-term reflection of quality (positive or negative), it is reassuring that the positional advantage of citation is primarily due to self-promotion, rather than to a one-time visibility effect. In this section, we

[9] Such a personalization system has been available to the subset of readers using the myADS features of the NASA ADS system, at http://myads.harvard.edu/.

consider the visibility effect on readership, and more generally consider how readership features can be used to predict the number of citations of an article. We will use full-text downloads as a proxy for readership. The download data is from the main arXiv site only, though constitutes a representative sample. It is cleaned of robotic accesses and multiple repeat accesses from the same domain within a small timeframe. Many articles are made available at the arXiv site in advance of publication by a peer-reviewed journals, though some authors await the results of peer review and make them available at arXiv.org more or less simultaneously with their appearance in a conventional journal.

2.3.1 Previous Work

Past studies have explored the relationship between downloads and citations. Using ADS data from 7.66 months of 2001, including more than 1.8 million "reads", [58] studied, among other things, the mean relation between reads and cites, and estimated roughly twenty ADS reads per citation for that period. [73] investigated the relationship between citations and first week's downloads for 153 articles in the British Medical Journal (vol. 318 from 1999), and found that the first week's download activity appeared to capture subsequent article citability. [70] computed the correlation between downloads and citations using a larger sample from the journal Tetrahedron Letters: 1,190 short articles published during the first half of 2001, with about 410,000 total downloads and 4,300 total citations. [13] discussed the correlation between early downloads (minus the first seven days) and citations of arXiv articles deposited 2000–2002. The data in this case came only from a single arXiv mirror, since the more voluminous data from the main site was not publicly available. [29] considered the relation between early downloads (first 90 days) and future citations for a few hundred articles that appeared in Nature Neuroscience during the period Feb–Dec 2005, and found a usefully predictive correlation, despite a comparatively small level of download activity.

In this section we use a data set considerably larger than the data used in these studies, and moreover a different methodology. Downloads and citations are typical heavy-tailed rather than normal distributions, so measures such as mean and standard deviation are less useful. Instead of computing a simple correlation between two variables, we consider the problem as a prediction task and use modern machine learning tools. Finally, we focus on the positional effect on readership, an effect not considered at all in the above, although any general relation between readership and citation, combined with a positional effect on citations investigated in the previous section, would naturally imply a positional effect on readership.

2.3.2 General Pattern

We use the readership for articles in the astro-ph, hep-th and hep-ph subject areas of arXiv received from Jan 2002 through Mar 2007. The dataset contains the date and time of every full-text download for each article through the end of 2007. There is great variation in the temporal readership pattern of articles, but the general feature is a burst of initial readership during an "active" period, and only sparse readership thereafter.[10] The existence of such an "active" period is an indication of the extent to which readers track the research via the daily announcements of new submissions. In fig. 2.14, we see that almost all articles

[10] This permits use of the full 5+ years of data, unlike the citation study of the previous section.

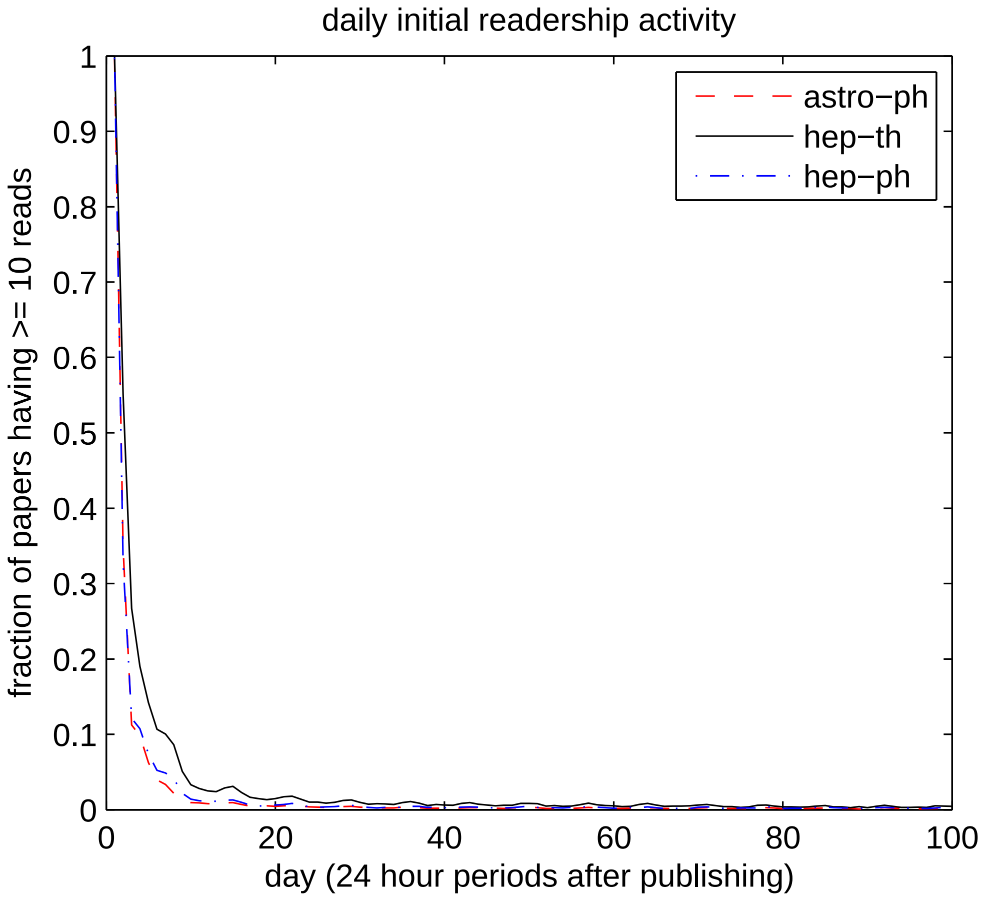


Figure 2.14: Fraction (in the subject area) of articles having ≥ 10 reads on a day.

are downloaded at least 10 times on the day they are first made public, and that fraction then falls rapidly.[11] For astro-ph, less than 1% of the articles have 10 or more downloads per day after the first 10 days. We take 1% to be the threshold of activity, so the active period for astro-ph is taken to be roughly 10 days. For hep-th, this period is 25 days, while for hep-ph it is 15 days. The total number of downloads in the active period can be taken as a measure of the initial popularity of an article.

[11] The seven day periodicity in fig. 2.14 results from the confluence of lower weekend readership with announcements of articles being made only on the five weekdays. We also checked for a possible "day of the week" bias, but found that the particular day of the week that an article is announced has no effect on the median number of citations.

Beyond the active period, typical articles receive no downloads on most days.[12] In astro-ph, for example, an average article is downloaded at least once during 12% of the days of its lifetime. For hep-th and hep-ph, this number is 13% and 17% respectively, with a standard deviation of about 10%. Readership can therefore be characterized by the number of days an article gets at least some downloads. Since the articles are of varying age in our dataset, we compare their readership activity beyond the active period by using the fraction of days an article gets downloaded at least once. It is natural to ask if there is a correlation between total initial reads and later (long-term) fraction of days getting some reads. Table 2.3 shows that indeed the fraction of later days getting some downloads is quite strongly correlated with initial popularity, by two common statistical measures.[13]

	astro-ph	hep-th	hep-ph
Pearson	0.5861	0.7436	0.6625
Spearman	0.6716	0.7525	0.6750

Table 2.3: Correlation ($P = 0$) between the number of downloads in the active period with the fraction of days, beyond the active period, an article is downloaded at least once.

[12] The articles that tend to have the most usage in the long-term are review articles and other pedagogical resources such as lecture notes. Ironically, this long-term usage is frequently not reflected in the citation record. These articles constitute a small enough fraction of the total that they do not skew the data.

[13] The Pearson correlation coefficient is a parametric statistic computed directly using the values. The Spearman correlation coefficient is the nonparametric version of Pearson, replacing the values with their ranks in sorted order. Correlation coefficients range from -1 and $+1$, where $+1$ indicates a linear correlation, 0 no correlation, and -1 linear anti-correlation. A value of 0.5 or more is ordinarily considered high.

2.3.3 Positional Effects

We now examine the relation between article position on the day of announcement and the total number of downloads in the initial active period.

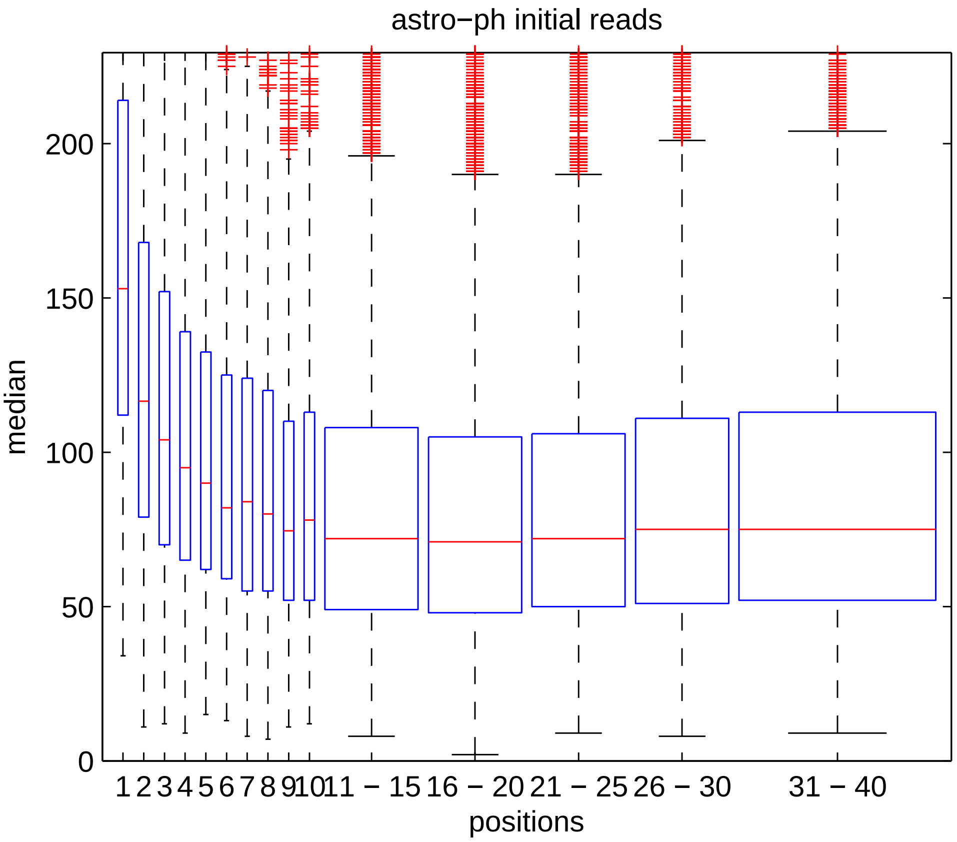


Figure 2.15: Box plot of total astro-ph downloads in the active period for different positions. Each box extends from the first through the third quartile, and the red line marks the median. The vertical dashed lines extend above and below to the largest and smallest values within 1.5 times the interquartile range from the respective quartile. The red '+' signs represent "outlier" points above this range.

For astro-ph, we see from fig. 2.15 that the number of downloads is higher for the top positions, and the median number declines with position for the early positions. The difference between the first and the second positions is quite

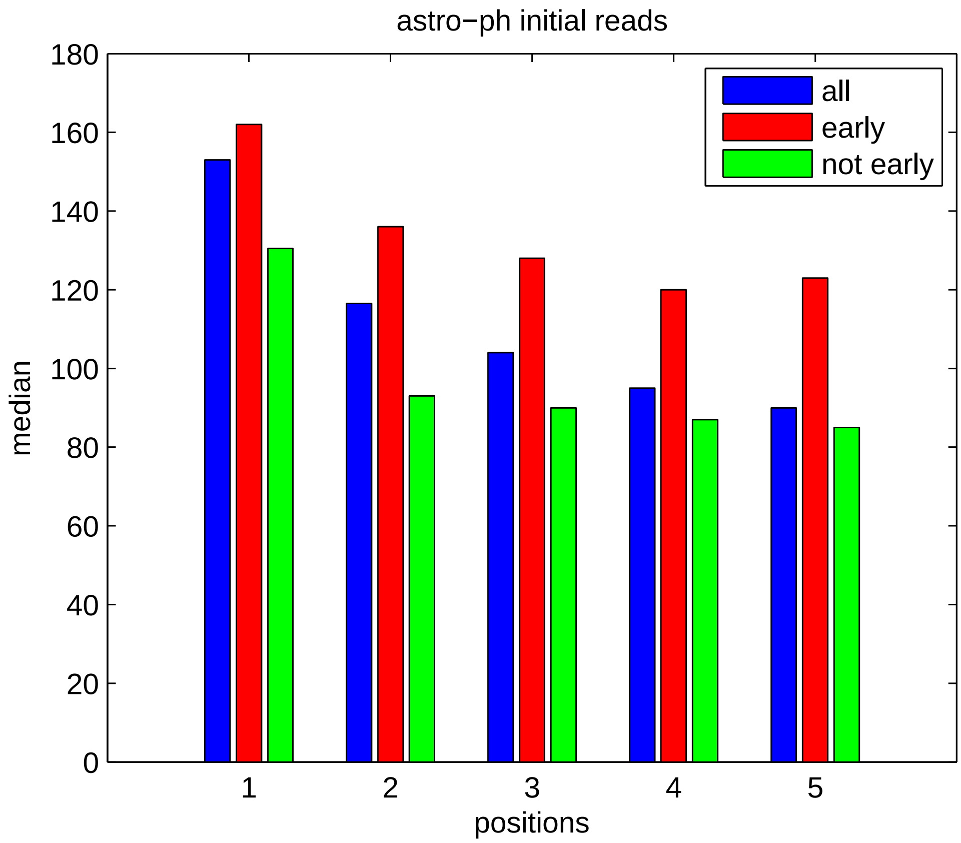


Figure 2.16: Median total reads for each position, with the red bars isolating the SP effect and the green bars the V effect.

striking. The differences in medians for the first six positions are statistically significant at the 1% level. The stochastic dominance of the distributions for different positions is also significant. Position 1 receives roughly twice the median number of initial reads as positions 10–40, indicating a very strong positional effect.

Fig. 2.16 shows that the positional effects in readership are dominated by self-promotion: the difference between the red and green bars is significant at the 1% level for each of the first 5 positions. Comparing with fig. 2.5, we also see that the visibility bias is much stronger in the initial popularity of an article

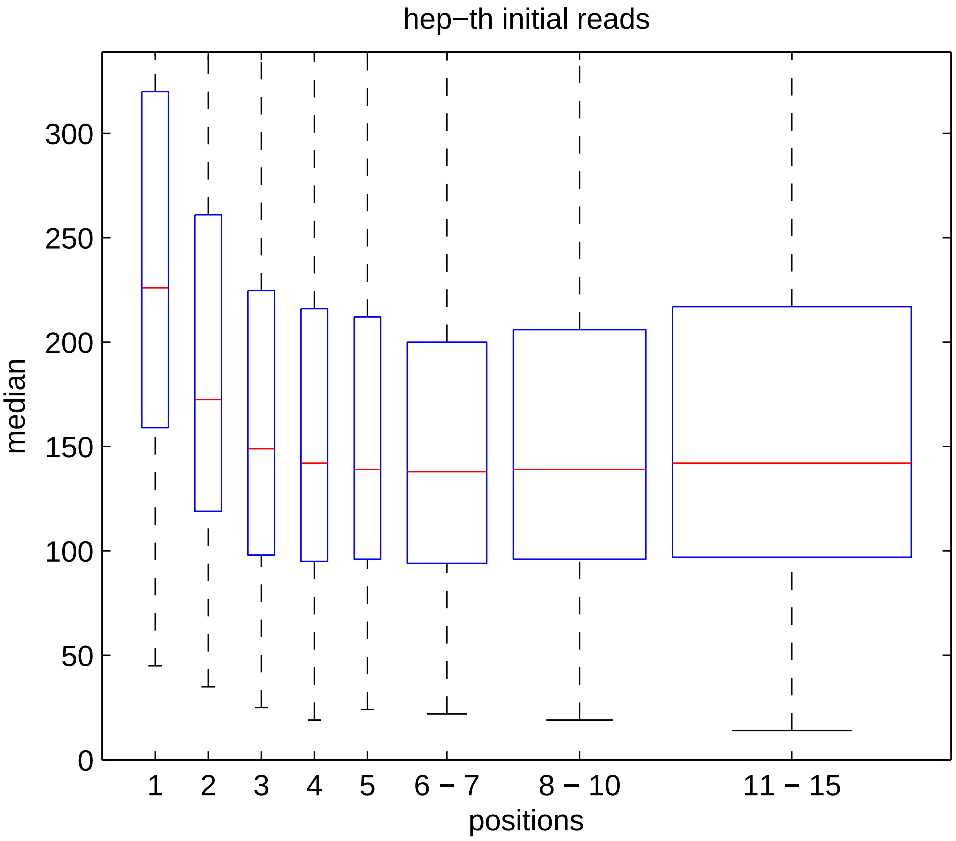


Figure 2.17: Box plot of total hep-th downloads in the active period for different positions, as in fig. 2.15 for astro-ph.

than in its long term citations, especially for the first position: the green bar, representing "not early" submissions, shows a significant drop from the first position to the next four in fig. 2.16.[14]

For the relation between article position and initial downloads for hep-th and hep-ph, figs. 2.17, 2.19 show a strong initial download advantage for the first two positions, and the green bars in figs. 2.18, 2.20 indicate a strong visibility effect for them. We confirm that the visibility effect can play a strong role

[14] For comparison with the larger bins mentioned in subsection 2.2.3, astro-ph articles in positions 1–6 received a median of 105 downloads, 44% higher than the median of 73 for those in positions 10–40. NE astro-ph articles in positions 1–6 received a median of 88 downloads, still 20% higher than for those in positions 10–40.

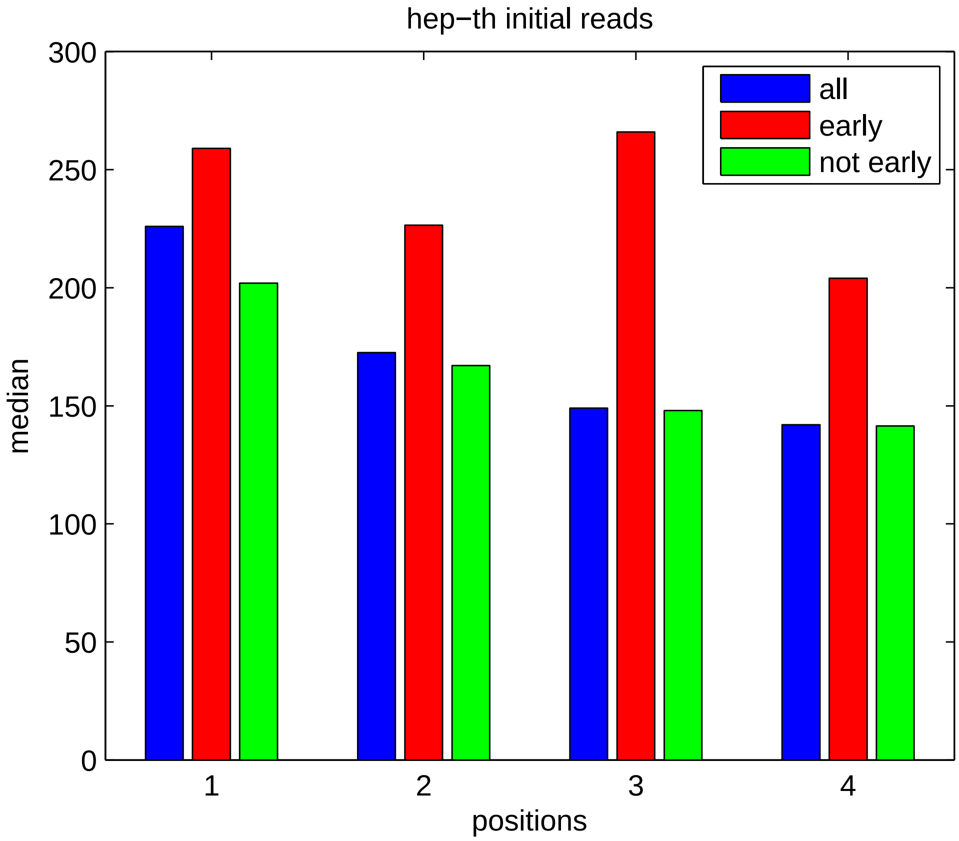


Figure 2.18: Median hep-th reads for each position, as in fig. 2.16 for astro-ph.

in the number of early reads even in the smaller hep-th and hep-ph announcements.

As pointed out earlier, article position is a one-time artifact of the initial announcement, persisting only for a single day. It is very difficult, if not impossible, to imagine any positional effect on citations in the absence of even stronger positional effects on initial reads. The above initial readership data for astro-ph, hep-th, and hep-ph provide a consistent underpinning for the citation results of the previous section, and are certainly consistent with some form of causal relationship.

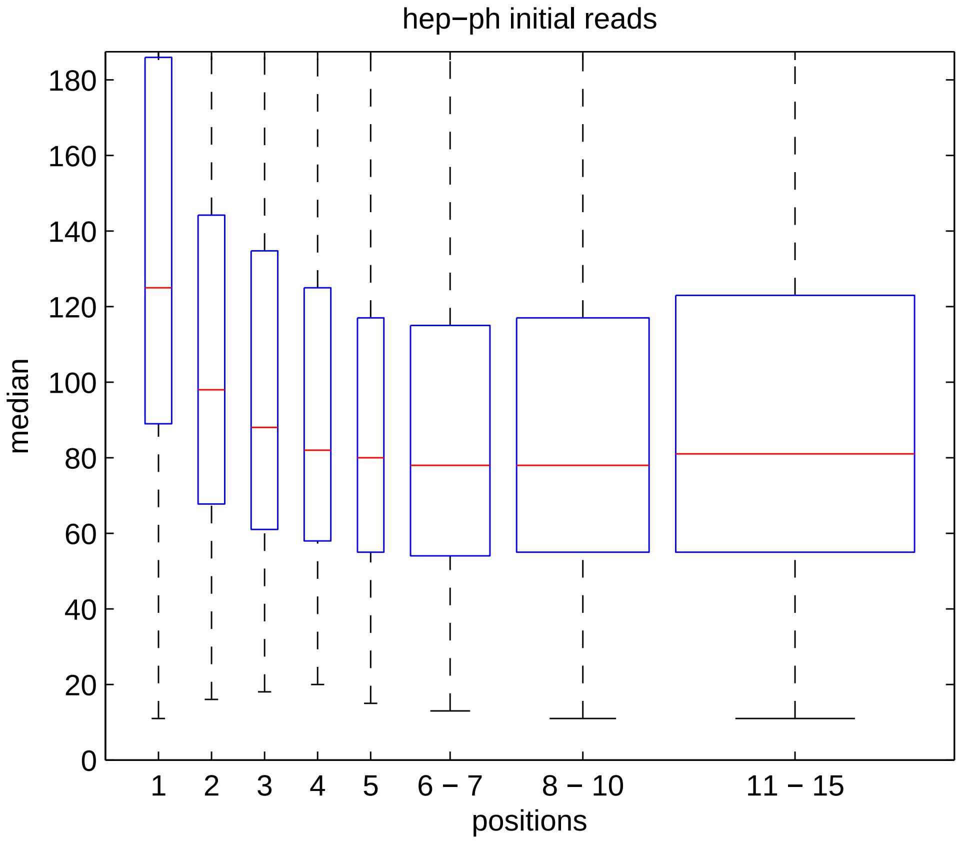


Figure 2.19: Box plot of total hep-ph downloads in the active period for different positions, as in fig. 2.15 for astro-ph and fig. 2.17 for hep-th.

2.3.4 Correlating Citation with Readership Features

The download data was also analyzed to discover the extent to which article readership predicts citations, and in principle gives some initial measure of article quality. Obvious features that could potentially be correlated with citations are the total downloads, total downloads in the active period, and total number of days getting some downloads. Articles whose initial active period is much shorter than average (e.g., 3 days rather than 10) do tend to get somewhat fewer citations in the long run, as would be expected for lower quality articles, rapidly

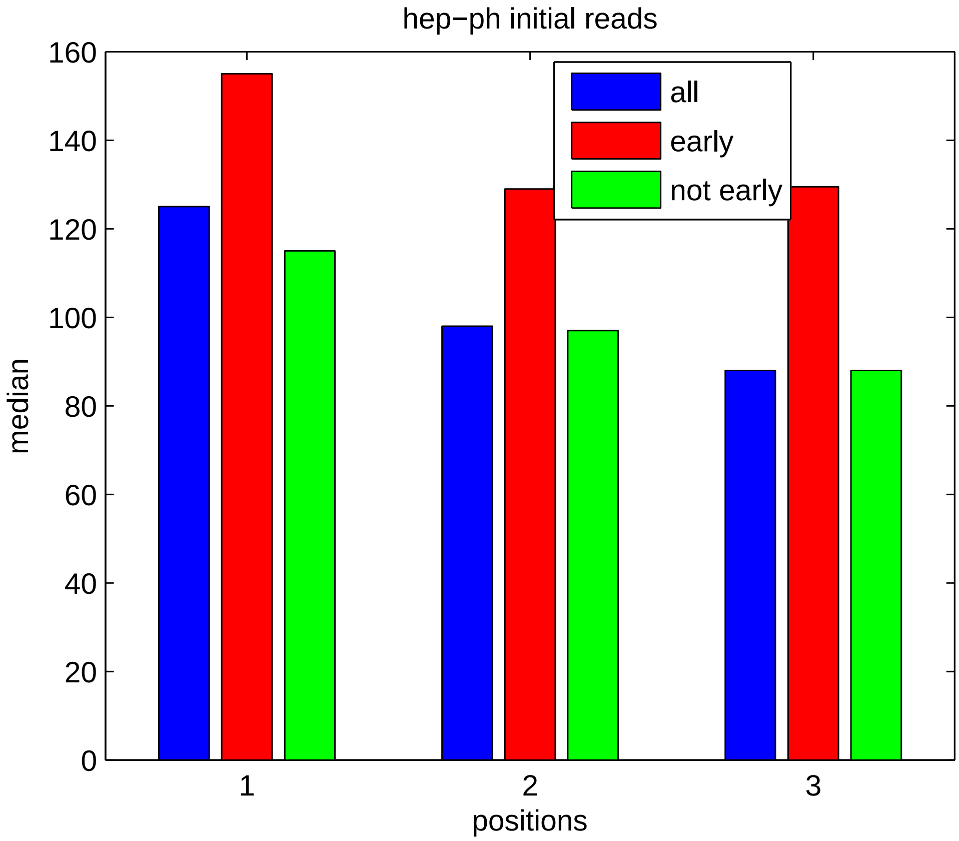


Figure 2.20: Median hep-ph reads for each position, as in fig. 2.16 for astro-ph and fig. 2.18 for hep-th.

identified as such by discerning readers. In astro-ph, for example, roughly 2.5% of the articles have 95% or more of their initial active period downloads during the first 3 days. These receive a median of 4 citations, whereas the remaining articles have a median of 7 citations, a difference statistically significant at the 1% level. The fraction of active period downloads occurring in the first 3 days could thus be another predictive feature.

It has been observed [87, 2, 93] that the number of citations is positively correlated with the number of authors of an article. Since articles accumulate citations with time, their age will have some correlation with the number of citations. As discussed earlier, self-promoted early articles receive more citations, perhaps due to higher quality, and position in the mailing may result in a visibility effect: thus whether or not an article is early and its position are important features.

	E	P	A	AR	F	D	TR	AG
astro-ph	0.113	-0.087	0.25	0.2753	0.069	0.326	0.328	0.086
hep-th	0.07	0.013	0.256	0.4825	0.25	0.61	0.593	0.07
hep-ph	0.092	-0.02	0.27	0.41	0.212	0.642	0.61	0.08

Table 2.4: Spearman rank correlation between the number of citations with different features: early or not (E), position in mailing (P), number of authors (A), reads in the active period (AR), fraction of active period reads outside the first 3–5 days (F), number of days beyond the active period getting some reads (D), total reads during lifetime (TR), age in days (AG).

These features are all correlated in some way with the number of citations. We use the citation and readership data for papers submitted between Jan 2002 and Dec 2004 for our analysis. Table 2.4 shows the rank correlation between number of citations and the different features individually. The feature most correlated with the ultimate number of citations is the number of days beyond the active period an article gets some downloads. Steady reads beyond the initial period are thus most predictive of citations, although initial reads are useful as well. Reads in this case can even be a consequence of the citations, since citations can lead readers directly to the arXiv site, hence the correlation.[15] The total number of reads is also well correlated with citations.

[15] Whether or not citation or use of bibliographic database leads readers to a journal site after publication or still to the arXiv site depends on how an article is cited, and also on the readership habits of the community, which could differ between the high energy physicists and the astrophysicists. Even the initial period in astro-ph is more likely to share readership with a journal version, since astrophysicists occasionally make arXiv submissions simultaneous with journal acceptance, while high energy physicists tend to make arXiv submissions hot out of the word-processor.

	E	P	A	AR	F	D	TR	AG
astro-ph	0	16	3	69	0.15556	96	185	1598
hep-th	0	7	2	137	0.17966	162	346	1636
hep-ph	0	9	2	79	0.16949	118	228	1630

Table 2.5: Medians of the quantities in table 2.4.

For completeness, in table 2.5 we give the medians of the quantities in table 2.4, but emphasize that the details of the distribution reflected in the rank correlation are not well captured by an aggregate quantity like the median. In addition, many of the medians are intrinsically unilluminating. For example, whether or not an article is early (E) is a binary feature, and the median is 0 since more than half the articles are not early. The median position (P) will be very close to half of the average mailing length since each of the positions has the same number of articles up to that length. The median age (AG) is constrained to be roughly 5 years for articles that range from 3.5–6.5 years old, and the median reads in the active period (AR) have already been given in figs. 2.15–2.20. Apart from the small fraction of articles that lose readership very quickly, the distribution of the fraction of active period reads after the first 3–5 days (F) will not differ substantially from the overall pattern of exponential falloff in readership.

Is there a meaningful way to harness the combined predictive capacity of the above features? The next logical step beyond correlation is to use regression. In addition to the above features, we have used the daily number of downloads for each of the first 100 days since the initial period is of much interest, and used the Support Vector Machine implementation SVM$^{\text{light}}$[16] [51], a modern supervised machine learning tool (see Appendix A.3), to predict citations. The methodol-

[16] http://svmlight.joachims.org/

ogy involved normalizing every feature by the 95th percentile of its set of values, to avoid convergence problems due to features having values that differ by several orders of magnitude. Since features like the initial and total reads have heavy-tailed distributions, norms like 1-norm, 2-norm or the ∞-norm would be dominated by the few large values, and hence normalization by any of these norms would result in setting the small values effectively to zero.

After normalization, the data set was randomly split into five equal parts. Then we ran SVM$^{\text{light}}$ in its regression mode [84] (with the default linear kernel) five times, using in turn each of the five parts as test set, and the remaining 80% as training set in each case. This is the standard 5-fold cross-validation procedure to ensure no overfitting of the data. For every run, the predicted citations were compared against the true citations to compute the predictive accuracy. Once again, since citations follow power law distributions, it is preferable to compare the *ranking* of the articles by the predicted citations and the *ranking* produced by the true citations, rather than comparing the actual magnitudes. This was done by computing the Spearman rank correlation coefficient between the predicted citations and the true citations, with the numbers then averaged for the 5 runs.

	astro-ph	hep-th	hep-ph
Average	0.3930	0.5998	0.6326
Standard Deviation	0.0211	0.0074	0.0168

Table 2.6: Spearman rank correlation coefficient between the actual citations and the citations predicted by the SVM regression.

Table 2.6 shows the extent to which regression was successful in ranking the articles. For hep-th and hep-ph the correlation is indeed quite high. For astro-ph the correlation is smaller, but still substantial. One possible explanation for this smaller correlation is that astro-ph citations more frequently lead to readership of the journal version, and are not captured by arXiv readership data as well are citations in the hep-th and hep-ph literatures, whose readers are by habit more likely to consult the version resident on the arXiv server. To assess this possibility, we folded in data, kindly provided by ADS, giving the number of full text downloads directed to the publishers via ADS (rather than to arXiv). This number is strongly correlated with the number of citations (roughly Spearman 0.5 for articles eventually published in a journal). Used as an additional feature in our SVM setting, the rank correlation in table 2.6 shifts to 0.7 for astro-ph, now comparable to and even slightly higher than the hep-th and hep-ph correlations.

	astro-ph	hep-th	hep-ph
Average	0.3869	0.577	0.5812
Standard Deviation	0.0075	0.0214	0.0200

Table 2.7: Spearman rank correlation coefficient between the actual citations and the citations predicted by the SVM regression, but without using the total reads and the long-term fraction of days receiving downloads.

As noted earlier, reads beyond the initial period, characterized both by the number of days beyond the active period when a paper gets some reads, and by the total number of reads, are most strongly correlated with citations. These two features are not necessarily *predictive*, however, since later reads at the arXiv site can result in future citations but can also result from citations, either directly or indirectly due to increased interest in an article. Table 2.7 shows the results of removing these two features and again running the SVM regression again with a 5-fold cross-validation. The correlation weakens slightly, as would be expected, but the early number of reads remains highly predictive of the long term citation behavior.[17]

[17] Another highly predictive feature that we did not analyze in detail here is the time to the

2.3.5 Discussion

There is no direct analog in other on-line resources for the positional effects on readership considered in sec. 2.3.3, both due to the nature of the arXiv daily announcements and the central notification role arXiv plays for entire research communities. We've seen that visibility plays a strong unintentional role as a recommender. The readership effects of the top few positions can be understood in terms of a stochastic decay-of-attention model, in which there is some probability of distraction at each entry, either by pausing to read the associated article full text, or by some external event. The reader either never returns to the original window to read the rest of the list, or having already spent time looking at full text becomes less likely to retrieve later full texts for perusal. The difficulty in eliminating such effects provides an additional rationale for offering personalization services to readers: when different readers view customized announcements ordered according to their individual preferences, the artifactual visibility biases of a single global list no longer play a dominant resonant role for the full research community.

The overall correlation we have found between citation and various readership features in table 2.4 confirms in a modern electronic context the primary intellectual role played by citation. Rather than playing some symbolic or primarily social role, or thoughtlessly propagated without consultation of sources, citations both appear clearly as a consequence of readership, and lead to further readership. The relation found here between readership and later citations amplifies the results of previous studies [73, 70, 13, 29] on the highly predictive role played by early readership. It is thus tempting to try to incorporate early

first citation.

readership and other newly available measures of popularity, such as blog commentary or other ‘Web 2.0’ commentary mechanisms, into some form of early guide to readers; and later in an article’s lifetime into some more generalized impact metric, incorporating citations as well. The visibility effects seen here, on the other hand, should give some pause in this regard, since they show that early readership driven by accidental forms of visibility can easily initiate feedback loops, which can leave significant and measurable traces in the citation record. Thus while citations are not primarily used for social purposes, they may nonetheless be subject to indirect influences familiar from studies of social networking effects [80], and thereby not provide an impact metric with the desired objectivity.

Other early activity measures correlated with long-term popularity have recently been considered for on-line sites such as YouTube and Digg [89], where the effect of early feedback mechanisms is found to be even more pronounced. There are many areas of superficial similarity between on-line scholarship sites on the one hand, and news and commerce sites on the other, but in the context of the results presented here it is important to recall their very different motivations for recommender mechanisms. An on-line newsite that draws the attention of readers to popular articles increases the number of article reads and hence chances to bring advertisements in front of readers. An on-line commerce site that successfully recommends other popular items increases its number of products ordered and gross revenues. By contrast, a scholarly site that focuses attention on a smaller number of articles, either intentionally or otherwise, could do an inadvertent disservice to both its authors and readers.

CHAPTER 3

LAST BUT NOT LEAST: ADDITIONAL POSITIONAL EFFECTS

We continue investigation of the effect of position in announcements of newly received articles, a single day artifact, with citations received over the course of ensuing years. Earlier work [26, 25] and chapter 2 of this dissertation focused on the “visibility” effect for positions near the beginnings of announcements, and on the “self-promotion” effect associated to authors intentionally aiming for these positions, with both found correlated to a later enhanced citation rate. In this chapter we consider a “reverse-visibility” effect for positions near the ends of announcements, and on a “procrastination” effect associated to submissions made within the 20 minute period just before the daily deadline. For two large subcommunities of theoretical high energy physics, we find a clear “reverse-visibility” effect, in which articles near the ends of the lists receive a boost in both short-term readership and long-term citations, almost comparable in size to the “visibility” effect documented earlier. For one of those subcommunities, we find an additional “procrastination” effect, in which last position articles submitted shortly before the deadline have an even higher citation rate than those that land more accidentally in that position. We consider and eliminate geographic effects as responsible for the above, and speculate on other possible causes, including “oblivious” and “nightowl” effects.

3.1 Introduction

In chapter 2, we considered a surprising correlation between article position in the initial announcements of new articles and later citation impact. As first described in [26, 25], articles that appeared at or near the beginnings of the simultaneous web and e-mail announcements, appearing every weekday, received substantially higher median citations due to a combination of “self-promotion” and “visibility” effects. “Self-promotion” reflected the tendency of some submitters to aim their submissions for just after the deadline, when they were most likely to appear near the beginning of the next day’s announcements. “Visibility” reflected the extent to which articles that serendipitously appeared near the beginning, by no conscious intent on the part of the submitter, nonetheless collected more median citations than had they appeared lower down in the listings. As emphasized in chapter 2, it is not immediately intuitive that position in the daily announcement of newly received submissions, a one-day artifact leaving no trace afterwards, could nonetheless leave its mark in long-term citation counts, accumulated years later.

While [43] was in preparation, a local physics grad student suggested that some submitters might instead sometimes aim for the *bottom* of the list. One reason intimated for doing so was that some readers use a different URL from the website to track the newly announced submissions. In addition to the http://arXiv.org/list/.../new URL[1], there is an additional URL of the form http://arxiv.org/list/.../recent, also prominently linked from the homepage, which collects the previous five days of announced articles. These are separated by day, with the most recent day at the top, so it was natural to present articles *within* each day as well in reverse order, in order that the numbering be continuous through the day boundaries. The other feature of the “recent” URL is that it displays Title and Author information, with only a link to the full abstract, so that many more entries appear within a browser window before paging down. Following the appropriate links to read only those abstracts

[1] described in chapter 2, where “...” denotes a specific subject class, e.g., astro-ph, hep-ph, hep-th

with relevant sounding titles may avoid some of the “fatigue” that contributes to the “visibility effect.” In any event, the logic was that if enough readers access the announcements via the latter URL, and submitters are aware of this, then there would be an incentive to aim submissions for the bottom of the list, as an alternative distinguished position.

But what fraction of readers view the daily listings on the pages with reversed order? During the 2002–2004 period studied in chapter 2, the webserver logs indicate that just over 80% of the hep-th (“High Energy Physics — Theory”) and hep-ph (“High Energy Physics — Phenomenology”) readers, and roughly 75% of the astro-ph (Astrophysics) readers access the announcements via the /list/. . . /new URL, and hence read in the canonical order. An even greater number of readers continued to receive the announcements via the legacy e-mail subscription[2], and hence the vast majority of practitioners of these three subject areas read the announcements in the standard (forward) order. “Reverse-visibility” effects were consequently not pursued in [43]. (Interestingly, the pattern was very different in some other subject areas, with roughly 75% of mathematics, 85% of computer science, and 50% of condensed matter web interface users reading the arXiv daily listings in reverse order, via the /list/. . . /recent URLs.)

Here we return to the “reverse-visibility” issue more systematically, and find that there is nonetheless a statistically significant enhancement of both readership and long-term citation behavior for submissions appearing near the end of the daily announcements. That both the top and bottom of lists can be distinguished positions is familiar from cognitive studies, in which people asked to

[2] roughly 1500 for each of hep-th and hep-ph, and about 3500 for astro-ph — this was the timeframe during which the primary usage began to migrate to the web interface. Use of RSS feeds had not yet become widespread.

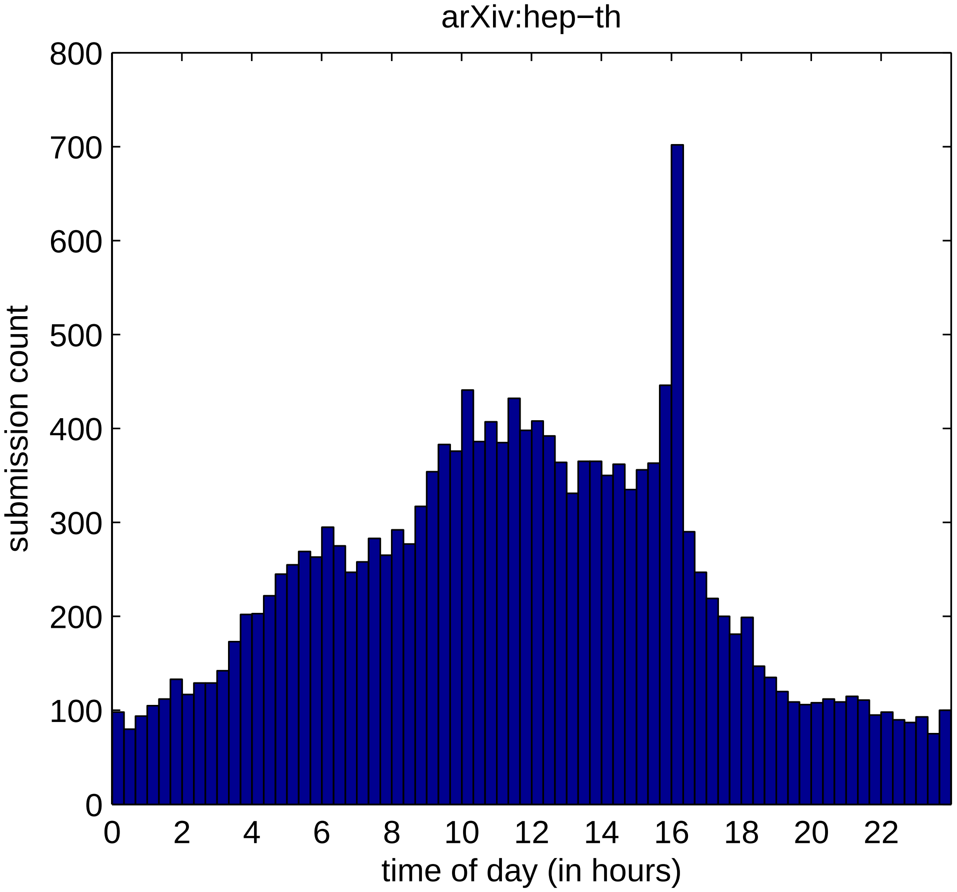


Figure 3.1: Figure 2.6 from chapter 2. Number of hep-th submissions by time of day, in 20 minute bins, during the period Jan 2002 – Mar 2007.

recall a list of items tend to recall most easily items near the beginning (primacy effect) and toward the end (recency effect) of the list [28]. This "serial position effect" remains a very active area of research in cognition and memory, and is likely due to a form of interference, in which items cognitively processed in the middle of a list receive confounding interference from those on both sides, whereas those at the beginning or end receive less interference, only from succeeding or preceding items.

Choosing whether or not to follow a link from an abstract to retrieve a full

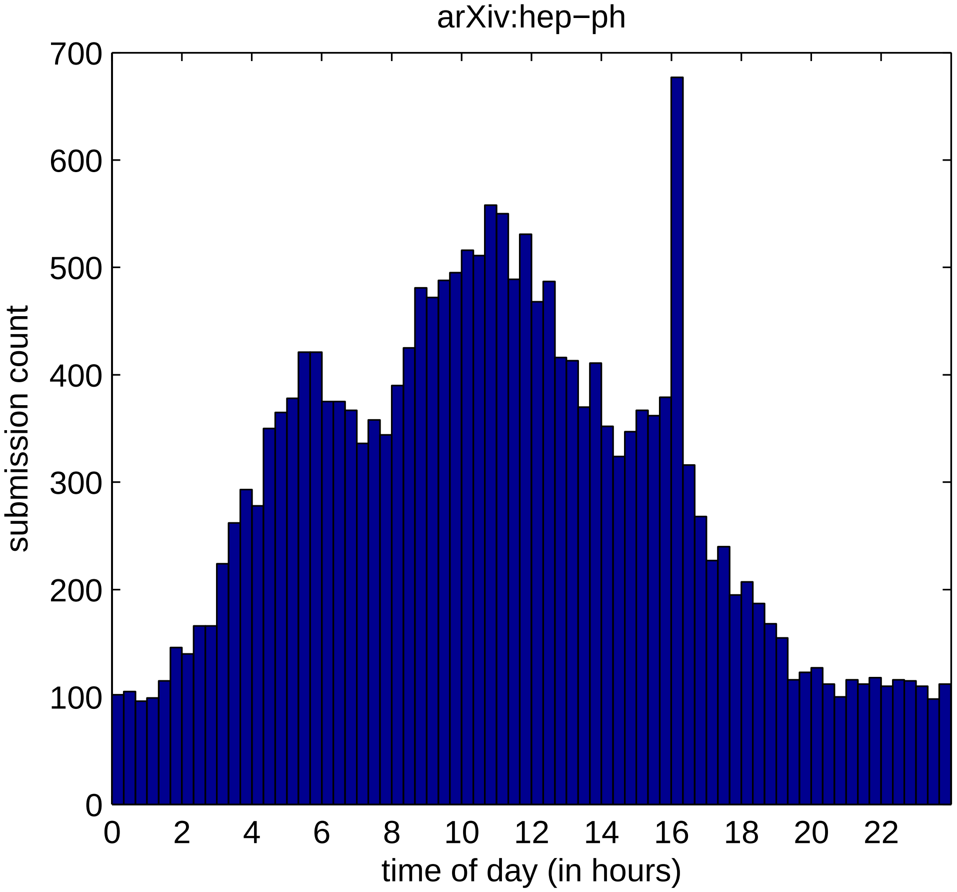


Figure 3.2: Figure 2.7 from chapter 2. Number of hep-ph submissions by time of day, in 20 minute bins, during the period Jan 2002 – Mar 2007.

text is a different exercise than recall of a randomly presented list, but nonetheless similar mechanisms may be at play. For example, some readers may occasionally find sifting through the daily announcements to be something of a chore, paying more attention in the beginning, scanning more quickly through the middle, only to concentrate again towards the end for some sense of closure. The task here, however, permitting real-time actions to be taken by readers while scanning the list (e.g., retrieving further information, as measured through full-text readership data), and as well permitting reconsultation of the list throughout the day, is sufficiently different that we maintain the visibility-related terminology used in [26, 25] and chapter 2. An additional twist is that submitter behavior can affect placement of items within the next day's list, leading to the notions of "willful primacy" or self-promotion, and of "willful recency" or procrastination.

Another suggested motivation for posting near the 16:00 U.S. eastern time deadline is simply that it is a deadline: submitters with neither interest nor motivation to aim for the beginning of the following day's listings may nonetheless have a strong desire for their article to appear that night, perhaps to stake a precedence claim in a fast-moving field, or perhaps to offload a psychological weight. For various reasons, including other commitments throughout the day carried by the busier and potentially higher profile researchers, or interactions with co-authors, etc., the submitter may delay posting until very close to the deadline. The deadline itself can even serve as the motivation for a final burst of focused activity to finish the text, rather than let it linger to the next day ad infinitum. Indeed in figs. 3.1, 3.2, we have reproduced for convenience figs. 2.6, 2.7 from chapter 2, showing the number of hep-th and hep-ph submissions received in 20 minute bins throughout the day. While the largest spike appears in the 20 minute bin just after the 16:00 deadline, there is also evidence for a smaller burst in submissions leading up to that time.[3] We refer to this effect as the "procrastination effect", a flip-side of the "self-promotion" effect which, as we shall see, shares some of its essential features.

[3] Note that those are not typically submissions piling up in the final seconds before 16:00, which might instead indicate a premature submission effect: submitters with bad aim due to poor clock synchronization. While most such submitters are probably not aiming for the final position, such delayed submission behavior has the same effect.

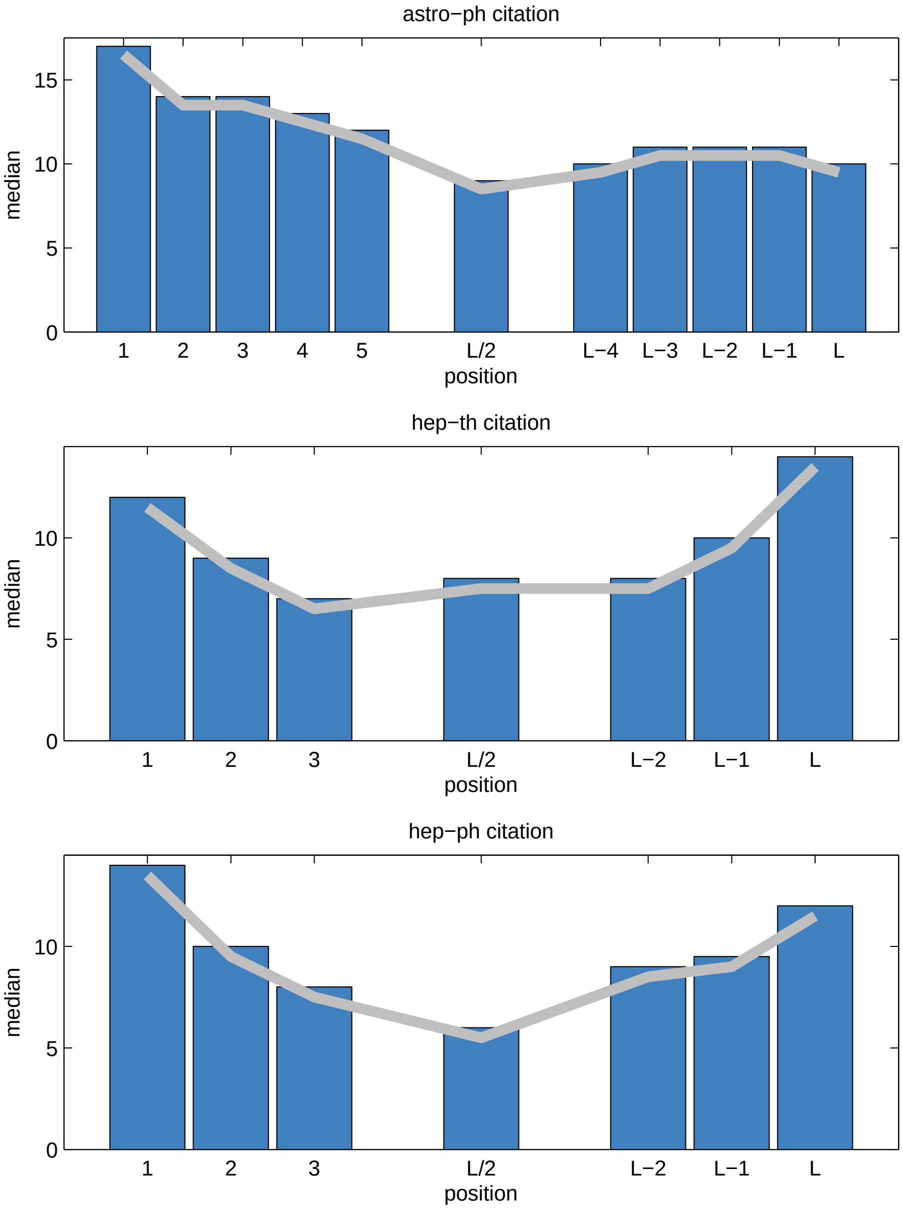


Figure 3.3: Citation statistics for the period 2002–2004. L denotes the mailing length: L is the last article, L-1 is the second-to-last, L/2 is the middle, and so on.

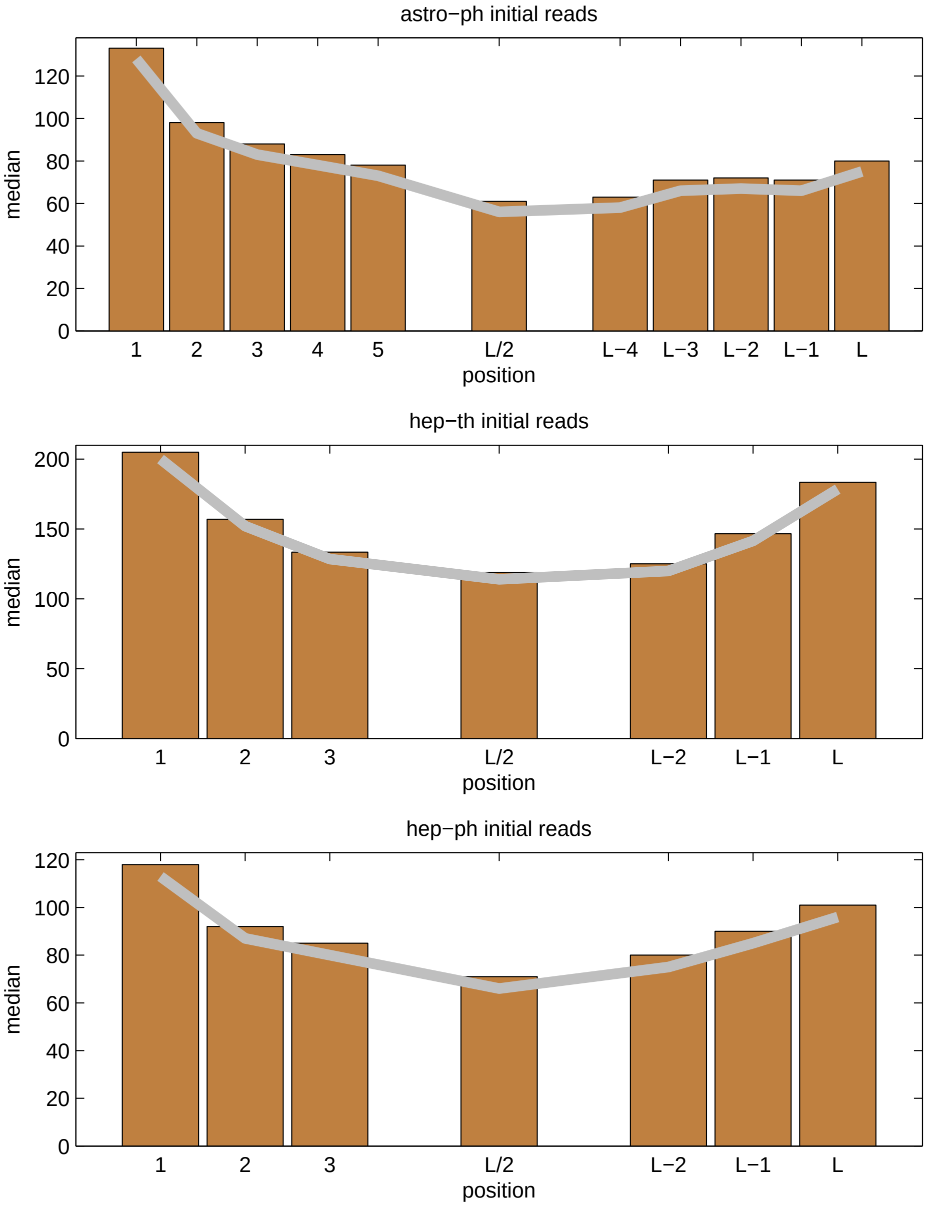


Figure 3.4: Readership statistics for the period 2002–2004. As in 3.3 L denotes the mailing length: L is the last article, L-1 is the second-to-last, L/2 is the middle, and so on. The readership data is for the initial active period of 10, 25 and 15 days respectively for astro-ph, hep-th and hep-ph, as explained in chapter 2.

3.2 Citation and Readership data vs. position

In order to assess any systematic effects for articles appearing near the ends of the mailings, we present in figs. 3.3, 3.4 three sets of articles for any mailing of size L. The articles near the beginning of the announcement are labelled according to position 1,2,3,. . ., those near the end of the announcement are labelled . . .,L-2,L-1,L, and the one (or two, for L even) in the middle is labelled L/2. Isolating the submissions in reverse order near the end of the announcements permits identifying any coherent end effects, independent of the varying length L. The underlying dataset is the same as used in chapter 2, except that mailings shorter than 11 were ignored for astro-ph (requiring at least five articles before and after the middle), and similarly mailings shorter than 7 were ignored for hep-th and hep-ph (requiring at least three before and after middle). Of the initial total of 776 mailings, this left roughly 770 mailings for astro-ph, 720 for hep-th and 750 for hep-ph, and the average mailing size L remained 30, 13, and 16, respectively, for that period. The number of articles in each position is the same as the number of mailings, except the middle position has roughly 1.5 times as many articles as mailings since both middle positions are kept for L even.

We see visual evidence in figs. 3.3, 3.4 for a "reverse-visibility" effect near the end of the hep-th and hep-ph mailings, both in readership during the initial active periods of the first few weeks after announcement, and as well in median citations received years afterwards. The effect is most prominent in the median citations received by articles near the end of the hep-th mailings (middle figure on left), coinciding with the noted pile-up in submissions just *before* the 16:00 deadline, reproduced here in fig. 3.1. Overall the full-text readership data confirms that the majority of readers peruse the lists in the standard forward order, with the greatest number of accesses to articles in the first few announcement positions.

The effect is less dramatic for the (longer) astro-ph announcements, so we restrict attention in what follows to the hep-th and hep-ph subject areas. As an aside, these two disciplines are very closely related, having emerged from the same theoretical particle theory research community following the re-emergence of string theory in the mid 1980's, with the more mathematically inclined represented in hep-th, and the more phenomenologically oriented in hep-ph, and a small subset of researchers contributing to both.

For further clarification of the effects in figs. 3.3, 3.4 we separate out in figs. 3.5, 3.6 the contributions to the first position according to whether the articles were received "early" (E), within the first 20 minutes after the 16:00 eastern time deadline, and the contributions to the last position according to whether the articles were received "late" (L), within the last 20 minutes before the next 16:00 eastern time deadline. By this criterion, 32% and 37% of the first position submissions to hep-th and hep-ph, respectively, were early, and 22% of the last position submissions to each were late.

In principle, one might expect no difference between either 1-NE or L-NL submissions and L/2 submissions, since all were submitted far enough from the deadline to be insulated from "self-promotion" or "procrastination" effects. Yet the median citation differences between submissions appearing at L/2 and the other two positions are significant at the 1% level[4]. The difference between the 1-NE and L/2 citations is the visibility effect discussed in [26, 25] and chapter

[4] We have used the non-parametric Mann-Whitney U (MWU) and Kolmogorov-Smirnov (KS) tests.

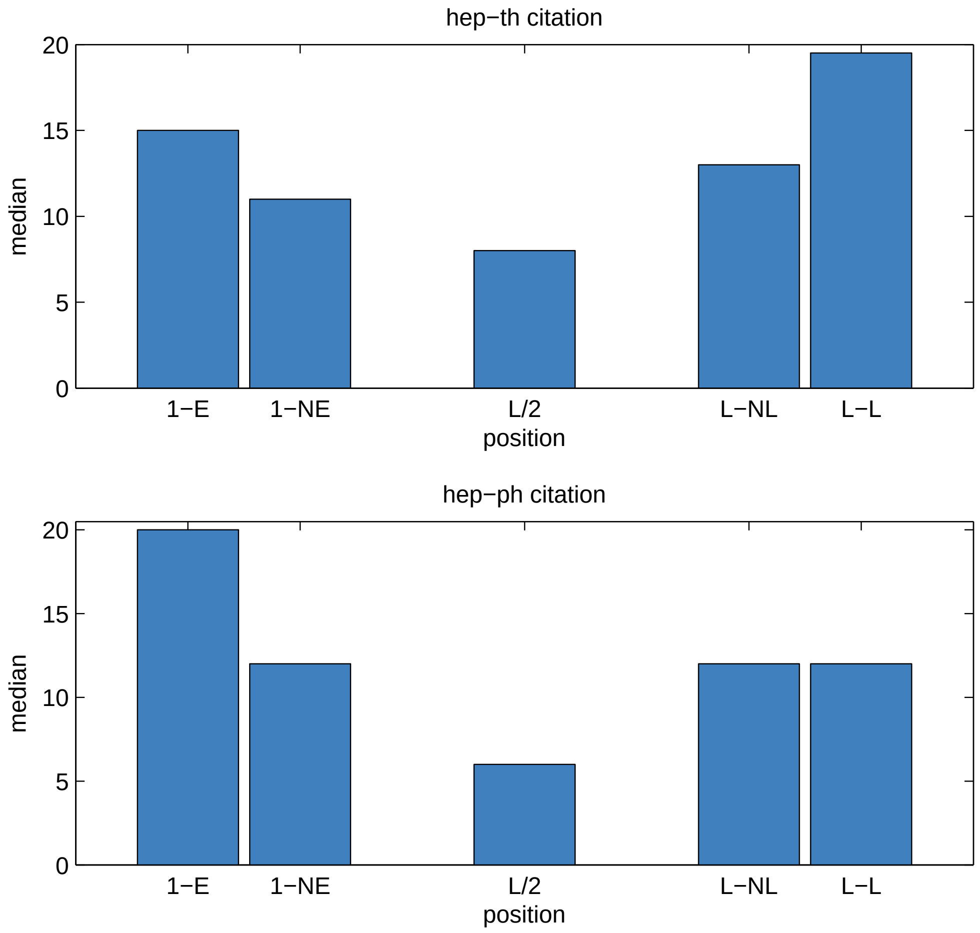


Figure 3.5: Citation plots for hep-th and hep-ph isolating early and late contributions. Here 1-E and 1-NE denote first position "early" and "not early", L/2 denotes the middle position, L-NL and L-L denote last position "not late" and "late" submissions, respectively.

2, and the difference between the L-NL and L/2 citations is part of the new "reverse-visibility" effect. This latter is likely some combination resulting from the smaller percentage of readers who access the lists in reverse order via the /list/.../recent URLs, from readers whose attention lapses in the middle of lists to refocus when the end is in sight, and from readers who may consciously

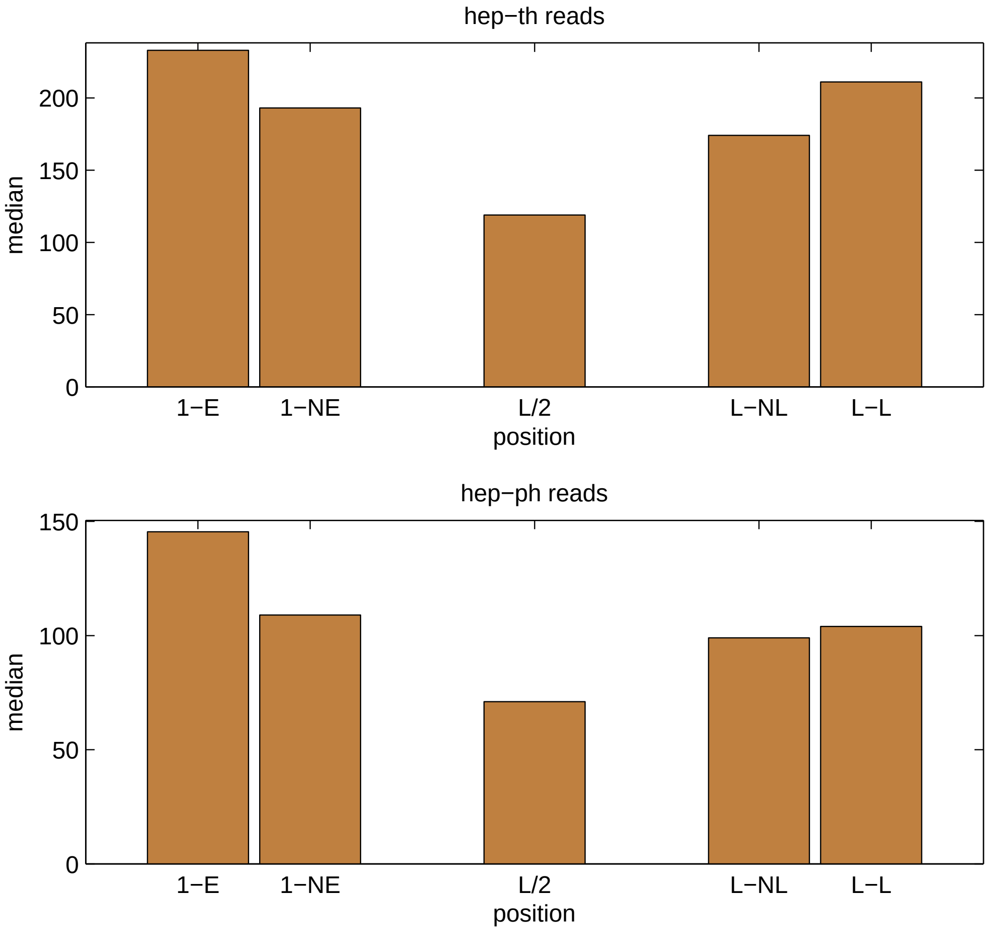


Figure 3.6: Read plots for hep-th and hep-ph isolating early and late contributions. Here 1-E and 1-NE denote first position "early" and "not early", L/2 denotes the middle position, L-NL and L-L denote last position "not late" and "late" submissions, respectively.

or subconsciously be expecting to finding higher quality articles near the end, due to the next effect we discuss.

The distinction between 1-E and 1-NE submissions is attributed to "self-promotion", i.e., the 1-E articles are distinguished by having been intentionally targeted by submitters to appear early in the announcements. The distinction between the L-NL and L-L articles is what we've termed "procrastination", with the latter articles having been more likely to appear in the last position due to submission within the final 20 minutes before deadline. (We note that "procrastination" might involve a slightly different mentality in Europe, where the deadline occurs in the late evening rather than during the working day.) For hep-th, the median citation differences between submissions in the 1-E and 1-NE positions, and between those in the L-NL and L-L positions in figs. 3.5, 3.6 are significant at the 1% level. While the L-L submissions appear to have a strikingly higher median citation rate of 19.5 compared to the self-promotion enhanced 1-E rate of 15, the 4.5 citation difference is not significant even at the 10% level by MWU test (P=0.1138 or only at the 11.38% level), so the 1-E and L-L positions should be considered as statistically similar.

For hep-ph, all of the 1-NE, L-NL, and L-L positions in figs. 3.5, 3.6 have 12 median citations, while 1-E submissions have a median of 20 (the enhancement corresponding to the previously noted self-promotion effect). Only hep-th appears to have the "procrastination" effect, in which submissions made just before the deadline tend to receive more citations. It is not entirely clear why two such similar disciplines would exhibit this distinction — perhaps practitioners of the less experimentally oriented discipline, operating on a shorter timescale and hence feeling more competition, perceive more of a need to stake their last minute precedence claims to avoid being scooped.

3.3 Geographic/Timezone effects

The other curious feature of figs. 3.3, 3.4, 3.5, 3.6 is the dip in the middle, i.e., the lower median citation rate for articles appearing in the vicinity of position L/2. This raises the question of whether submissions that appear in the middle of the announcement lists are subject to some other systematic bias, such as geographic. Suppose that researchers located in timezones whose workday is far displaced from 16:00 U.S. eastern time also happen to receive systematically lower citations, for some related or unrelated reason. Then a geographic bias would explain not only that salient feature of figs. 3.3, 3.4, 3.5, 3.6, it might also be partly responsible for what was identified as a visibility effect in [26, 25] and chapter 2.

The end-of-workday 16:00 U.S. eastern time deadline for submissions corresponds (during daylight savings time) to afternoon (13:00–16:00) in the continental U.S. (and to 10:00 and noon in Hawaii and Alaska, resp.), to late afternoon in South America (15:00–17:00), late evening in Western Europe, Middle Africa, and Scandinavia (21:00-23:00), around midnight in the Middle East and Western Asia (23:00–01:00), to middle of the night in Eastern Asia (01:00–05:00 for India, China, Korea, Japan), and to early morning in Australia and New Zealand (04:00–08:00). The deadline corresponds to the middle of the night for much of Asia, and the natural end of the workday in those regions corresponds instead to the middle of the daily submission period, causing submissons to land closer to the middle of the daily announcements.

To begin assessing whether the larger admixture of such submissions is responsible for the dip, fig. 3.7 further subdivides the bins of figs. 3.5, 3.6 accord-

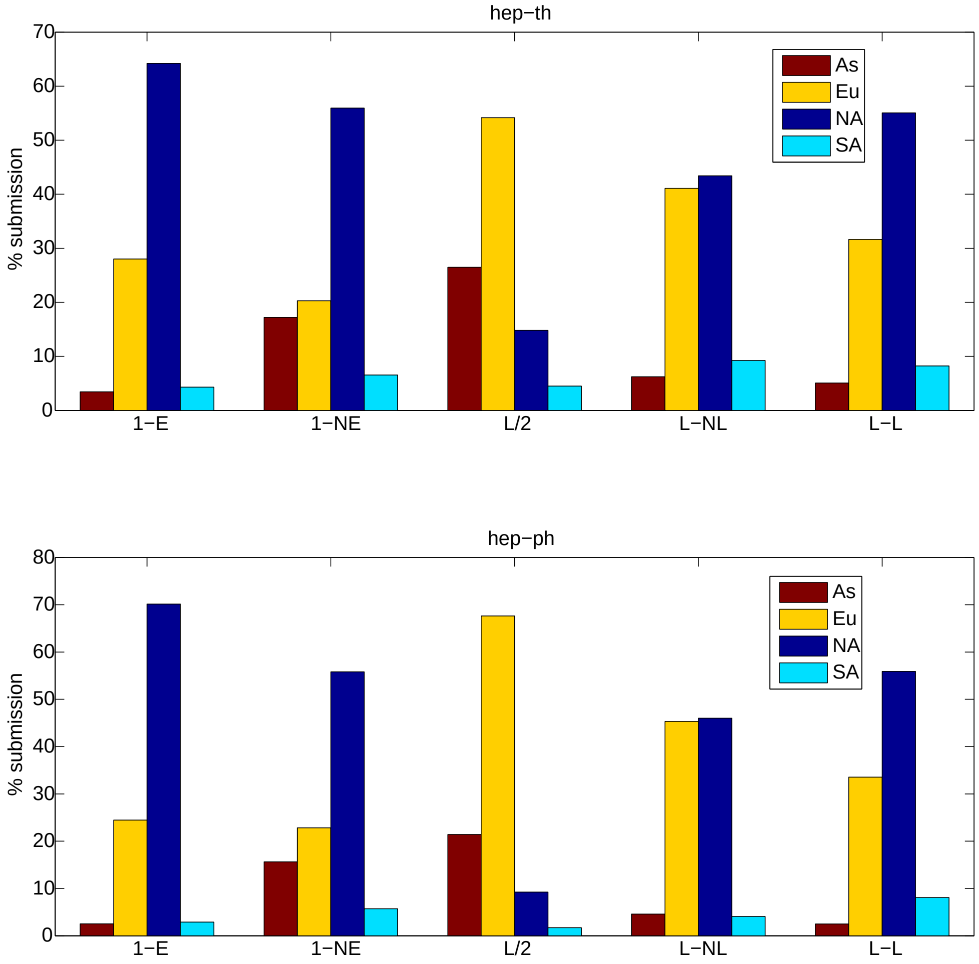


Figure 3.7: Geographical distribution of 1-E, 1-NE, L/2, L-NL, L-L articles for hep-th and hep-ph, using the country specified by the domain of the submitter's e-mail address. North America (NA) is out of phase with Asia (As), although a few dedicated submitters from Asia submit early/late. A substantial fraction of the middle articles are from Europe (Eu). Contributions from South America (SA) are under 10% for each bin.

ing to the registered e-mail address of the submitter.[5] The geographic distribu-

[5] This is in rough correspondence with the geographic location of the submitter, with a few qualifications: (i) .com and .org addresses were taken as US and hence North America. This should not be problematic since during the 2002–2004 submission period in question, only a very small percentage of submitters were using .com addresses and there should be no effect on the medians or statistical significances of differences. (More care would be required for analysis of astro-ph, since the world's largest astronomy research facility, eso.org, is a European center.)

tions for the 1-E vs. 1-NE and L-NL vs. L-L submissions are clearly different. The middle L/2 position decomposes most distinctly from the distributions for first and last positions, with the majority at L/2 coming from Europe, and with a much larger percentage coming from Asia.

If we consider the analog of figs. 3.1, 3.2, but only for As submissions, then neither hep-th nor hep-ph show spikes at the 16:00 deadline. The overall submission pattern is as expected from fig. 3.7, out of phase with NA: with a smooth peak in the 4:00 range (early in the morning on the East coast), about 7 hours before the 11:00 aggregate peaks in figs. 3.1, 3.2. We have verified the geographic locations of the internet IP addresses used to upload each of the 1-E submissions from submitters registered with As e-mail addresses, and the majority did indeed arrive from middle-of-the-night timezones. The small remainder came from As registered users temporarily operating from NA or Eu locations, as temporary visitors or attending summer workshops or schools. (Curiously, the submission pattern for As submissions to astro-ph does show a significant spike at the 16:00 deadline, so self-promotion in astro-ph is sufficiently attractive to entice submitters to operate at inconvenient hours. The roughly 100 such submissions over the greater than 5 year time period nonetheless corresponds to a low rate, of only one per every 2.5 weeks. About 75% of those did arrive from internet IP addresses in As timezones, mainly from Israel, Japan, and India — at local times ranging from 23:00 to 5:00.)

(ii) The e-mail address only corresponds to the location of the submitting author, so this form of classification ignores the effects of cross-continental collaborations. Our experience is that the choice of submitting author nonetheless typically reflects the majority point of origin. (iii) The submitting author may no longer be physically at the same location as the e-mail address first used to create the registration. For the most part, however, submitters have kept their contact information up-to-date. (iv) The boundary between Europe and Asia can be defined either politically or geographically. Here we use the political definition, including Russia with Europe. (v) Asia is far from homogeneous, even familiar countries such as China, Japan, South Korea, Israel and India are at differing levels of economic and hence academic development. The coarse aggregate measures used here accurately reflect these weighted heterogeneous distributions.

	Eu	NA	As	SA	EuNA
astro-ph	9	12	7	6	10
hep-th	8	13	7	4	9
hep-ph	7	12	6	5	10

	Eu	NA	As	SA
astro-ph	10523	9272	2296	452
hep-th	4495	2688	2030	606
hep-ph	6537	3100	2220	322

Table 3.1: (a) table of median citations of 02-04 articles. All pairwise differences are significant at the 5% level. (b) number of submissions from each continent.

Table 3.1 shows the geographic citation differences for submissions associated to the 4 most active continents, independent of announcement position (i.e., aggregated over all submissions from each of the four regions during the 2002–2004 timeframe). The median citations of 7, 7 and 6 of Asia (As) articles for astro-ph, hep-th and hep-ph, respectively, are about 30% less than the median citations of 10, 9 and 10 for the same subject areas from the combination of Europe and North America (EuNA).

We now assess to what extent geographic/timezone effects, causing those Asian submissions with fewer citations to be disproportionately represented among the middle submissions, might account for the dip in the middle of figs. 3.3, 3.5. In the following, statistical significance limits are set at the 5% level, and we ignore South America.

First we compare citations of European (Eu) and North American (NA) submissions. For the five subsets 1-E, 1-NE, L/2, L-NL and L-L in fig. 3.5, the differences between NA and Eu median citations are not statistically different, except for the hep-ph L/2 submissions. In that bin, NA submissions have a median of 8 citations while Eu submissions have a median of 6 (and even there only significant by the MWU test, not by the KS test — by contrast, the corresponding medians of 11 and 8 for hep-th at L/2 are not significant by either test at the 5%

hep-th	1-NE	L/2
EuNAAsSA	11	8
EuNA	12	8
EuNAAs	11	8
NA	12	11
Eu	12	8
As	8	8

hep-ph	1-NE	L/2
EuNAAsSA	12	6
EuNA	13	6
EuNAAs	12	6
NA	14	8
Eu	12	6
As	9	7

Table 3.2: Median citations to hep-th and hep-ph submissions at positions 1-NE and L/2, for various subsets of geographic regions. The data from these bins succinctly capture any geographic bias and visibility effects.

level). Overall there is little evidence of geographic bias between Eu and NA submissions, in accord with the observation in [26, 25].[6]

Next we include Asian (As) submissions and compare citations of the combined EuNAAs to those of the combined EuNA submissions. The first rows in Table 3.2 show the median citations for the combined submissions from the four continents (EuNAAsSA) for articles in positions 1-NE and L/2 (just two of the bins from fig. 3.5), for hep-th and hep-ph. Restricting to just the combined EuNA submissions in the second rows, only the hep-ph 1-NE median has a slight statistically significant increase (from 12 to 13 in the first column of the table on the right). Reincluding the As submissions in the third row, we see that the median citation for articles in the L/2 position is unaffected, and consequently those are *not* responsible for the smaller median for articles in that position. The median of 8 for As submissions to hep-th at the L/2 position is actually the same as that for EuNA (none of As, Eu and NA have a statistically significant difference at that position). The As L/2 median of 7 for hep-ph is even

[6] As described in Appendix A.1, the cuts in [26, 25] resulted in focusing on higher cited articles with potentially attenuated geographic bias, so it is worthwhile to confirm for the larger sample.

slightly larger than the EuNA L/2 median of 6 (the small median advantage for NA submissions over those from Eu and As at L/2 is not statistically significant, with EuAs constituting over 80% of the submissions at that position). Since the Asian submissions are concentrated near the middle of the mailings and receive similar median citations as Eu or NA submissions at L/2, they do not add a geographic bias to the positional effects analyzed in chapter 2.

By contrast, there is a drop by one median citation at the 1-NE position between rows 2 and 3 of table 3.2, confirming that adding the lower median citations of As submissions to those from EuNA do lower the median for positions other than L/2 (though with only about 80 As submissions at that position, the signal is weak, i.e., significant by only one of the two tests, at the 5% level). A similar comparison between EuNA and EuNAAs submissions at the 1-E and L-L positions is not possible since there are even fewer As submissions (under 10) at those positions.

We close here with some additional comments about data in the tables. It might seem odd at first sight that the overall median NA hep-th citation in table 3.1 is 13, while the corresponding NA median hep-th citations at the 1-NE and L/2 positions in table 3.2 are 12 and 11, but the former is pulled up by other positions, including a 1-E median citation of 16. In table 3.2, the majority of the 1-NE vs. L/2 effect in hep-th median citations arises from Eu submissions. Since 1-NE submissions from Eu are typically submitted long after ordinary working hours, this difference suggests an additional “nightowl” effect (not further investigated), complementary to the “procrastination” effect, that researchers who habitually work late into the night receive more median citations. It would be instructive to find further quantitative evidence that researchers with obsessive work habits (working to meet deadlines, or working through the night) ultimately have more impact.

In summary, the results of this section clarify the question posed by table 3.1, suggesting that any period with fewer NA submissions would have a geographically induced diminution of median citations. Even though NA articles in aggregate tend to receive more median citations, the specific subset that appears in the middle of mailings perform about the same as those from elsewhere. This suggests yet another curiosity in aggregate citation behavior, an "oblivious" effect: researchers who operate oblivious to deadlines, submitting neither shortly before nor shortly afterwards, tend to get fewer median citations!

3.4 Discussion

In chapter 2, the effects of visibility and self-promotion were disentangled by comparing the subset of articles that serendipitously appeared in early positions (due to administrative moves or slow submission days) with those targeted to appear there. An analogous procedure here would be to consider articles originally slated for final positions but shifted to middle positions by administrative moves. A statistically significant lower median citation rate for this subset compared to those that remained in final positions would provide alternate confirmation of the "reverse-visibility" effect. Similarly, if articles submitted shortly before the deadline and administratively moved to middle positions continued to show a higher median citation rate than either middle or late position articles, this would provide alternate confirmation of the "procrastination" effect. While the medians in these cases did tend strongly in the expected directions, the sets in question were unfortunately not large enough to make statistically significant statements (even relaxing the definition of “middle” to be neither first three nor last three positions).

As discussed in chapter 2, the effects uncovered here result from unintentional properties of the announcement system during the timeframe studied. There have since been some changes in the system, including subdivision of astro-ph into subcategories, but many of the positional effects remain possible in the current system. Some users have suggested randomizing the daily order entirely, either uniformly for everyone, or individually for each user. Others have pointed out that might render an unintentional disservice to readers, who perhaps benefit from seeing self-promoted or procrastinated articles brought preferentially to their attention. More modern presentation systems have also been suggested, such as more subgrouping by topics or enhanced graphical representations (2d concept maps, etc.). The most likely remediation of these issues remains some form of personalization system, in which preferences actively registered by users via controlled keywords or search terms, combined with passively collected past usage data (from the same user on an opt-in basis), provide user-specific highlighting or reordering of entries. This would ultimately mitigate the global resonance phenomenon, unique to this resource, of entire research communities viewing the same material in the same order on a daily basis.

The citation effects analyzed in this chapter and in chapter 2 have been formulated in terms of the median, because the mean of these heavy-tailed distributions would be strongly affected by the highly cited articles in the tail. Those heavily cited “elite” articles are moreover less likely to be subject to the various visibility and related effects in the long-term. These effects, however, do extend beyond the median, or typical, article, as seen in the upper quartile distributions of figs. 2.3, 2.10, 2.11 in chapter 2. It is also important to note that the difference in read and citation rates as a function of list position is large, in some cases a factor of two, and, for the most part, independent of the quality of the paper. This has substantial implications for use of these metrics in assessing individuals and organizations.

Finally, since such intriguing differences in behavior between practitioners of such closely related disciplines (hep-th and hep-ph) are seen here, it will be informative to assess the behavioral characteristics of other disciplines within the arXiv dataset. Further insight can be obtained by tracking the behavior of individual (anonymized) readers in the usage logs. It is also possible to consider the time dependence of these effects, using datasets after the 2002–2004 sample used here, now that their long-term citations have stabilized. Other short-term visibility-type effects on long-term citation rates can as well be investigated, including work on whether an article's lucky appearance in a smaller daily announcement list, or unlucky co-occurrence with an article destined to be highly cited, have measurable long-term citation effects.

Part III

Concept Extraction and Tracking

CHAPTER 4

PHRASES AS SUBTOPICAL CONCEPTS IN SCHOLARLY TEXT

Retrieval of subtopical concepts from scholarly communication systems is now possible through a combination of text and metadata analysis, augmented by user search queries and click logs. In this chapter we investigate how a “phrase”, defined as a variable length sequence of vocabulary words, can be used to represent a concept. We present a method to extract such phrases from a text corpus, and rank them using a citation network measure. We validate the ranking with actively and passively determined metrics: comparison with human-assigned keywords, and comparison with passively harvested terms from search query logs.

A vocabulary of significative words is first identified by contrasting the unigram word probability distribution of the scientific corpus of interest with that of a non-scientific corpus. Phrases, as sequences of these vocabulary words are then systematically extracted, and ranked using a network measure, the “compensated normalized link count” (CNLC), which measures the extent to which they are propagated, or “conserved”, along the network structure of the citation graph of articles.

We demonstrate this method on full texts and abstracts from 7 years of high energy physics articles from the arXiv preprint database. Evaluation was performed by comparing to topic keywords assigned by Deutches Elektronen-Synchroton (DESY) library staff, and to the query terms submitted to arXiv’s search engine, classified as coming from readers in the high energy physics community. These correspond, respectively, to explicit and implicit forms of human annotation, with the latter providing a more direct window into the cognitive representation of concepts employed by researchers. Our method, applied to either abstracts or full-texts, is found to perform better than comparable network and text-based metrics using either of the above forms of human annotation as baseline.

4.1 Introduction

As digitized text becomes increasingly available, from libraries and scholarly journals to news, blogs and microblogs, many changing on very short timescales, filtering and understanding the incoming text stream is an increasingly daunting task. New tools to trace the history of a certain piece of news, or the development of an idea through large text repositories, are consequently of great interest. It was shown in [61] how "memes" — quotations and their variations in the news and blog text — can be efficiently tracked through time, leading to interesting models of news and blogspace interactions.

While short-lived and dynamically changing content is useful to trace, comparable tools for tracking over longer timescales could as usefully be applied to scholarly text repositories for use in scholarly communications infrastructure. Experts in a subject area are ever pressed to maintain a global understanding of the trends and to track the parallel time evolution of multiple concepts. Recommender systems can tend to narrow focus by suggesting excessively similar articles within a narrow range, rather than trying to provide a broadened perspective.

The essential step in tracing concepts is to characterize their textual representation. In this chapter we present a method of extracting subtopical concepts from scientific text. In chapter 5 we demonstrate how reader interest can be tracked through topics and summarized via clustering.

Scientific concepts, such as ideas and techniques, are often represented by sequences of terms. In this work, meaningful sequences of words are called “phrases”. Phrases are coarser than the longer news quote “memes” in [61], but are finer than topic categories that contain collections of related subtopical concepts. As an example, “learning query intent classification” is a phrase within the topic of “query classification” in computer science. Words co-occurring anywhere in documents is often used for clustering, but for finer concepts, the word co-occurrences need to be restricted to smaller windows. *n*-grams are the most restricted examples of such word co-occurrence, and variable length *n*-grams can be used to extract concepts of varying specificity.

Our method of identifying phrases proceeds in two phases. First we restrict attention to a smaller set of words, or vocabulary, computed from the corpus. Contrasting with a non-technical reference corpus permits identifying the vocabulary without domain specific knowledge or language understanding. For example, a standard corpus of english literature serves as a baseline for identifying field specific vocabularies for physics or mathematics. Even a corpus of a subfield of physics could be the contrasting corpus of choice for another area of physics, if the differences between the two subfields is the primary interest. Words in the main corpus whose usage is statistically different between the two are selected as the vocabulary words, with the intuition that the different statistical usage implies a novel connotation of a word. For example, the word “gravity” has very different meanings in Shakespearean literature and astrophysics, and the differences are reflected in the distributional usage of the word.

Sequences of vocabulary words, extracted as phrases, are then ranked using citation information. Each phrase is associated with a set of articles in which it appears, and these sets of articles for different phrases overlap. If we consider articles as nodes of a graph, then phrases function as hyperedges, connecting the articles in such sets. In this chapter, we propose ranking of phrases through the density of links, after subtracting the components expected by chance in citation subgraphs for each phrase. The main idea is that if one article containing a phrase cites another article containing the same phrase, then it is likely that the citation indicates subtopical continuity characterized by the phrase. This proposed network measure, named “compensated normalized link count” (CNLC), is computed for citation subnetworks associated with each phrase. Within each article, CNLC can be combined with phrase frequencies to provide good local rankings.

To evaluate our ranking, we have first compared to a set of manually assigned topic keywords. Since such explicit human annotation is not always available, we have explored alternative sources of human assessment, including search query and click logs of arXiv. Heuristic classification of search queries using click data allows us to consider queries submitted by users as phrases of interest. For each query, we have used Jaccard similarity to find the best matching phrase, using the rank of that phrase as the rank of the query. We have then used the average rank of the subset of queries relevant to our corpus to quantify the effectiveness of phrase ranking algorithms. This method of evaluation, using such accumulated implicit information, can provide a baseline for other tasks for which the creation of benchmark test sets can be time-consuming or expensive.

For scholarly articles, abstracts are frequently assumed to be well-written summaries of the full texts, and as such are used as proxies for full text in many text mining tasks. The efficiency of our phrase extraction and ranking method permits comparing the performance on full text versus just abstracts. We have observed that although abstracts are better for finding user search queries, full texts are better for discovering subtopical concepts. This likely results from restrictions on abstract length making it difficult to accommodate all the concepts in a long article.

4.2 Related Work

4.2.1 Topic Detection

There has been ever-increasing activity in finding topics in text corpora over the past two decades. Early non-probabilistic methods such as Latent Semantic Indexing [23] employed the still widely used bag-of-words model, and an important early variant was probabilistic Latent Semantic Indexing [47]. With the explosion of electronic text available for analysis in the past decade, a variety of topic-detection algorithms have been developed. Recently, Latent Dirichlet Allocation (LDA) -based methods (see [9] for a comprehensive overview) have been applied to find topics in large longitudinal scientific text corpora such as *Science* [8]. LDA based text analysis of scientific and wikipedia texts was augmented with link analysis in [15, 16], using available citations, hyperlinks or inferred linkages, within a language modeling framework. The problem of retrieving subtopical documents, with respect to a query, was addressed in [99], where the topics and subtopics, however, were assumed as inputs to the algorithm.

The concepts considered here are intended to be somewhat finer-grained than the topics typically considered in topic detection. Our object is to extract phrases that characterize scientific concepts or techniques, to give a subtopical projection of the documents. Most of the research on topic detection limits the number of detected topics to a few hundred, whereas our method detects hundreds of thousands of phrases, well beyond the number of documents in the corpus. While phrases potentially representing the concepts contained in a topic are often implicit in the topic computation, explicit extraction of the phrases ordinarily entails a sophisticated language parsing methodology, moving far beyond the more generic bulk text-mining of bag-of-words document representations. Here we present a method of retaining just enough of the essence of the semantic relationship between words for the task of detecting subtopical phrases.

Algorithms incorporating language modeling usually assume a generative model for the documents, and compute the parameters of the model through different inference procedures [85, 9]. Here we use neither ordinary language modeling nor bag-of-words representation, nor even the basic techniques of stemming and stop word removal. Instead we emphasize word co-occurrence in near proximity to infer a relationship between words to construct phrases, and use the underlying citation network to gauge the significance of phrases.

The use of citation network of documents here is closest in spirit to the approach of [50], on which this work builds. There bigrams are taken as topics, and ranked using a log odds ratio computed from the citation network. Our ranking of phrases uses a different measure of interconnectedness, to avoid biases of the log odds ratio which may skew the rankings. Such biases include ignoring the actual number of links of a document in a network, and weighting disconnected nodes much higher than connected ones. Phrase extraction using the bigram co-occurrence in [50] also does not scale well, so primarily used abstracts of scientific texts, whereas our attempt employs the full text of the documents.

4.2.2 Keyphrase Extraction

The basic idea behind extracting keyphrases from documents [92, 4, 49, 98, 56, 66] is use of supervised methods trained to classify which keyphrases are important. Candidate phrases were often extracted using natural language processing techniques, such as identification of parts of speech. Common features used to train classifiers were term-frequency multiplied by the inverse document frequency (tf-idf), length of phrase, and position of first occurrence in the document. Different learning methods were used: e.g., decision tree learning and genetic algorithm in [92], naive bayes in [98], and neural networks in [66]. In [49], a ranking approach using Ranking SVM was shown to be quite effective. In [56] restricted length *n*-grams were used but no learning algorithm was employed.

Our method of finding important phrases is fundamentally different from these supervised learning approaches. We use network information to find a global ranking of the phrases whereas previous research on keyphrase extraction has mostly concentrated exclusively on the text. Not only is our dataset two orders of magnitude larger than most of the work in that area, we also present an alternative evaluation method that does not rely on hand-curated data, and is thus scalable to larger systems.

The most important distinction between keyphrase extraction research and this work is once again the granularity of the phrases identified here. While keyphrases are finer than broad topics, the phrases employed here are even further fine-grained, and much more numerous than the handful of keyphrases typically assigned by authors or curators.

4.2.3 Search Queries

There is a wealth of literature on using search engine logs to collect implicit feedback from users to improve search result quality [75, 76, 53, 52, 96]. Classification of queries is a very important step in modern search engines to disambiguate queries with multiple meanings [64, 63, 82, 12, 62, 24]. Query similarity measures were often computed from the click log to help cluster the queries [6, 97]. Use of query classification to learn user preferences was demonstrated in [14].

In this work, we have used search queries as an alternative to manually labeled data to test the quality of our ranking. For that purpose, the queries from the search log were classified using a simple voting heuristic derived from the click log so that the correct subset of queries, relevant to our corpus, could be extracted. Unlike the above-cited work on queries and click data, these data sources were not used for training the algorithm here.

4.3 Methodology

The characterization of a subtopical concept that we have used in this work is through a phrase defined as **a sequence of vocabulary words**.

Every *n*-gram of an article is a potential phrase. We can discard *n*-grams longer than the usual sentence length, which would traverse sentence boundaries. But a simpler approach is to use a set of selected words, a vocabulary, whose sequences we will consider as phrases. Non-vocabulary words will form natural boundaries for phrases.

One drawback of this characterization is that depending on the set of words chosen as vocabulary, meaningful phrases may be split into two or more phrases. For example "anomalous dimensions of baryon operators" would be split into two phrases "anomalous dimensions" and "baryon operators" if the word "of" is not in the vocabulary. Exclusion of words like "of", however, helps to prevent other less meaningful conjunctions of individually contentful phrases. In our method, the parts get similar ranking due to co-occurrence and partial match with human annotation is used in the evaluation procedure. (Clustering based on co-occurence of phrases within sentence length windows could be used to merge the split parts, but that is not discussed any further in this dissertation.)

4.3.1 Vocabulary Selection

Vocabularies are ordinarily used for text mining tasks. In [8], a vocabulary was computed by removal of "common function words" as well as frequency-based pruning. In [4, 49, 66], natural language processing (NLP) techniques were employed to perform parts of speech tagging so that the extracted noun phrases could be used to identify keyphrases. In [98, 56], predefined patterns of parts-of-speech were used to identify phrases. In general, vocabulary selection for NLP-based methods relies on stemming and parts -of-speech tagging.

Vocabulary selection is undertaken differently here. Stemming is more difficult for scientific text, and the different variations of the same stem may have significantly different meaning in technical literature. So a statistical approach to the task was chosen, partially incorporating semantic meaning by using a non-technical reference corpus to remove uninteresting linguistic components.

Suppose $\mathbf{p}(w)$ is the probability of a word w in our corpus and $\mathbf{q}(w)$ is the probability of the same word in the reference corpus. The KL-divergence is an asymmetric measure of distance between two probability distributions:

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_{w:\mathbf{p}(w)>0} \mathbf{p}(w) \log \frac{\mathbf{p}(w)}{\mathbf{q}(w)} \, . \tag{4.1}$$

The contribution of a word w to the KL-divergence is thus $\mathbf{p}(w) \log \frac{\mathbf{p}(w)}{\mathbf{q}(w)}$. Words with large contribution to this distance are most discriminating between the two corpora, and we have selected those words to form the vocabulary for phrase extraction.

Note that if a word is rare in the reference corpus, then $\mathbf{q}(w)$ is small and thus $\frac{\mathbf{p}(w)}{\mathbf{q}(w)}$ is large. But the factor of $\mathbf{p}(w)$ still ensures that words rare in the main corpus will not be given large weight. This is the case for typographic errors or very rare non-english words. The most important words according to this measure should be the words that are relatively rare in the reference corpus while relatively frequently used in the corpus of interest.

4.3.2 Phrase Ranking

Sequences (or *n*-grams) of vocabulary words as phrases are easily identified, but require a meaningful ranking methodology. Term (or phrase) frequencies can be used for a simple document-specific ranking, and for corpus-wide global ranking inverse document frequency (idf) is often employed. The tf–idf combination is the most important feature for keyphrase extraction methods in [4, 49, 98, 56, 66]. Here we use instead the citation link structure of the articles to rank phrases, in principle a more straightforward use of citation data than previously employed to compute the log-of-odds-ratio (LoOR) for topic ranking in [50].

The central idea for our ranking is that if a phrase represents a concept, then it is highly likely that the articles in which the phrase appears cite each other more than they would by chance. So for each phrase, we count the number of citation links in the subgraph of articles where it appears, and subtract the number of links we would expect by chance. We call this the *compensated link count* for the phrase. Finally we normalize this quantity by the size of the phrase subgraph (number of articles) so that phrases that represent broad and specific concepts may be compared against each other. We call this final quantity the *compensated normalized link count* (**CNLC**) and use it for ranking. Important phrases should have higher **CNLC** values.

Suppose Q is the set of articles in our corpus, and $E_Q = \{(i, j)\}$ is the set of citation links such that for $i, j \in Q$, i cites j. Let $Q_p \subseteq Q$ be the set of articles in which a phrase p appears. Let $n = |Q|$ and $n_p = |Q_p|$. For an article $i \in Q_p$, let

$$k_i = |\{(i, j) \in E_Q : j \in Q\}|$$

$$a_i^p = |\{(i,j) \in E_Q : j \in Q_p\}| \ .$$

k_i is the number of citation links from each article i to other articles in the corpus, and a_i^p is the number of citation links from i to articles in the subgraph Q_p. The probability of a citation to Q_p is $\frac{n_p}{n}$, and if we assume citations are independent of each other, then we should expect article i to have $k_i \frac{n_p}{n}$ links to Q_p. The compensated link count for i is thus

$$a_i^p - k_i \frac{n_p}{n} \ .$$

For phrase p, we obtain the compensated link count by summing over Q_p

$$\sum_{i \in Q_p} a_i^p - k_i \frac{n_p}{n} \ .$$

And finally we compute the *compensated normalized link count* (**CNLC**) for p by dividing the above by n_p, which simplifies to

$$\mathbf{CNLC}_p = \frac{1}{n_p} \sum_{i \in Q_p} a_i^p - \frac{1}{n} \sum_{i \in Q_p} k_i \ . \tag{4.2}$$

The first term is the average number of citations an article in the phrase subgraph receives from articles in that subgraph, while the second part subtracts out the average number of citations an article, in the whole corpus, receives from the phrase subgraph. If a phrase is very common and has a large subgraph, then the subtracted part will be large. In the hypothetical extreme case that the phrase subgraph is the whole corpus, the two terms are equal and the CNLC is zero. For phrase subgraphs of small size, the subtracted part is small since n is large compared to the phrase subgraph size.

Note that articles have a particular sequence in time, and ordinarily citations would necessarily be directed towards earlier articles. But occasionally authors later update articles with citations to articles that appeared after the original submission time of the earlier article. Such patterns may introduce cycles in the citation graph. Models that use the acyclic property of the citation network will need to discard the seemingly inconsistent links. But these links do convey useful information and we have used all the links, whether temporally consistent or not.

Note also that there can also be links to articles outside of the ingested corpus. Since arXiv is relatively comprehensive for the subject areas under consideration here, we have decided to ignore citations to articles outside our corpus in order to avoid the task of identifying and ingesting texts from external sources that would have no systematic effects on the results.

4.4 Evaluation

4.4.1 Data

The corpus used in this work is from arXiv High Energy Physics – Theory (hep-th) from January 2000 through December 2006. At the beginning of this timeframe, arXiv had a stable user base. The corpus consisted of about 1Gb of text from 22,712 articles, which was pared down to about 740Mb of text after heuristically removing the lists of references at the ends of the articles. Abstracts were obtained from the curated metadata.

The sample non-physics corpus was comprised of the Bible (King James version), the complete works of William Shakespeare, War and Peace by Leo Tolstoy, and Ulysses by James Joyce, all obtained from Project Gutenberg.[1] The citation information for these articles was obtained by crawling the SLAC Spires High-Energy Physics Literature Database[2] in late September of 2010. Topics for each article, assigned by the German research center for particle physics Deutches Elektronen-Synchroton (DESY)[3] library staff, were also obtained from this database.

In addition to the full text, data from arXiv's search query log was used. The selected segment is from March 2008 through August 2009, and contains about 180,000 queries submitted from over 48,000 different IP addresses. Download data for articles from the same time period, cleaned of robotic access such as search engine crawlers, was used to categorize the search queries.

4.4.2 Vocabulary

We have applied our method of phrase extraction and ranking on the corpus of abstracts as well as on the full-text corpus of the hep-th articles. By contrasting with the English literature corpus consisting of the Bible, Shakespeare's work, War and Peace, and Ulysses, we have ranked the words according to their KL-divergence contributions. In the process of computing KL-divergence contributions, the English literature distribution was smoothed, through additive smoothing with $\alpha = .001$. Words of length 1 and 2 were ignored since these are very often stop words or mathematical symbols like ij. Words present in more than 95% of the articles were also ignored as being too pervasive to be informative, although their usage distribution may be different between the two

[1] http://www.gutenberg.org/

[2] http://www.slac.stanford.edu/spires/

[3] http://www.desy.de/

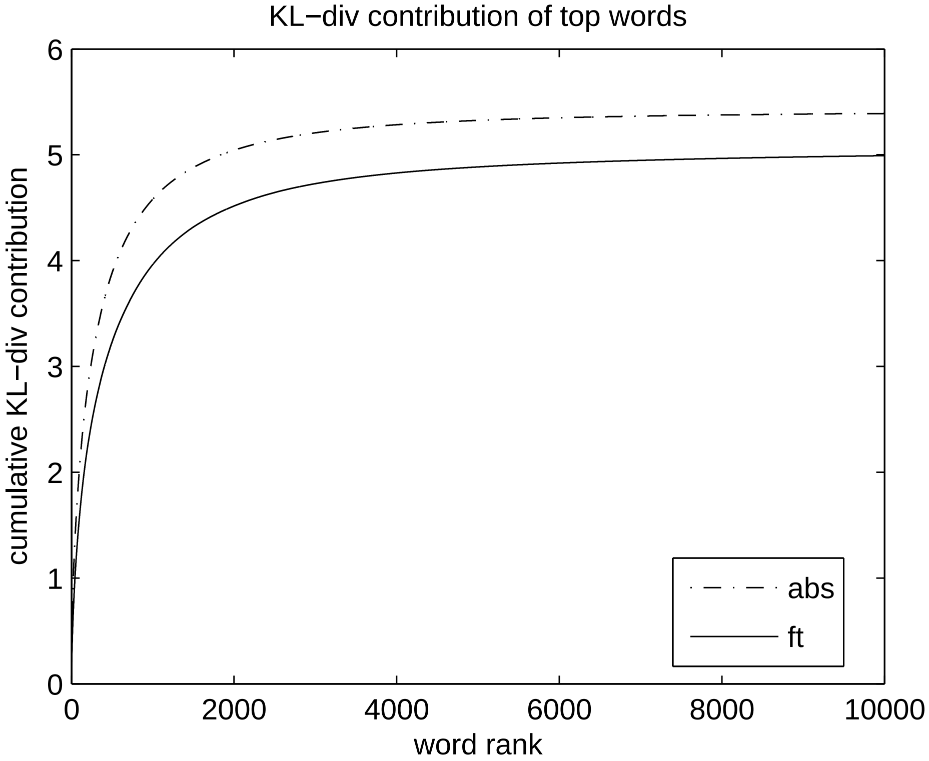


Figure 4.1: Cumulative KL-div contribution of the top 10,000 words for the abstracts corpus (abs) and the full-text corpus (ft).

Top 1–5	Top 6–10	Last 10–6	Last 5–1
brane	*string*	*when*	*have*
dimensional	*supersymmetric*	*was*	*all*
theory	*scalar*	*their*	*that*
quantum	*model*	*they*	*not*
gauge	*noncommutative*	*but*	*and*

Table 4.1: Words in the corpus of abstracts ranked by KL-divergence contribution.

corpora.

The corpus of abstracts had about 19,000 different words, while the corpus of full-texts had about 145,000 words. The top-ranked and last 10 words of the corpus of the abstracts are shown in table 4.1. The effectiveness of the method in identification of words important in high energy physics is apparent.

The cumulative KL-divergence contributions of the top words for the two corpora are shown in fig. 4.1. Beyond a certain point, marginal KL-divergence contributions become small enough that the cumulative contribution is effectively flat. We have chosen to cut-off at the point where the cumulative KL-divergence becomes 90% of its maximum value. For the two corpora, we obtained vocabularies consisting of the top 1470 and 2165 words, respectively. Over 92% of the words in the vocabulary from abstracts were also present in the vocabulary from the full-text.

4.4.3 DESY Topics

In the keyphrase extraction literature [92, 4, 49, 98, 56, 66], the extraction method is evaluated using keyphrases assigned to a document by humans, either authors or independent judges. In the same spirit, we have used topic keywords assigned by the DESY staff, publicly available on the SLAC Spires database. About 4.5% out of the 22,712 articles did not have topics assigned to them and were thus ignored. The remainder of the articles had 7.7 topics on average (median of 7). The distribution of the number of topics per article is shown in fig. 4.2. There were 9500 different topics with an average length of 2.65 words (median of 3). Over 70% of the topics had 2 words, and about 10% had 3 words.

4.4.4 Ranking

After computational identification of the vocabulary words, maximal sequences of these words, delimited by non-vocabulary words or mathematical symbols,

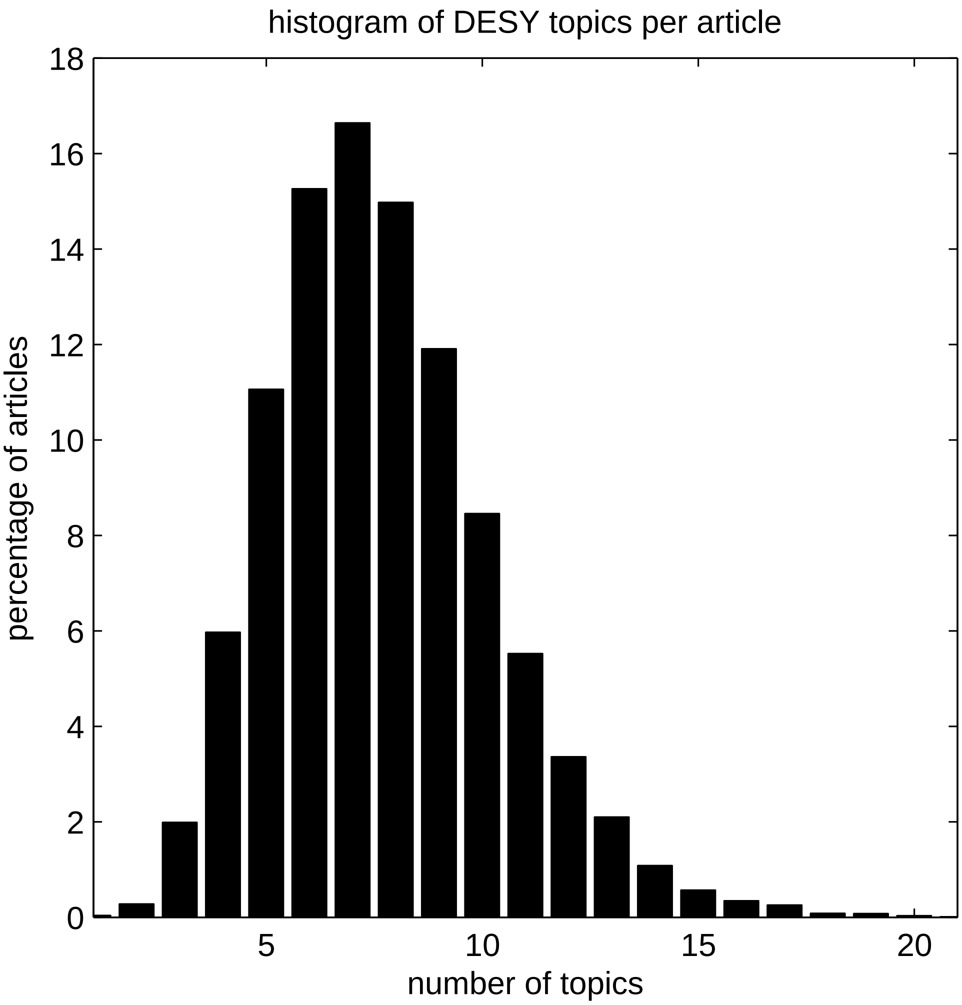


Figure 4.2: Distribution of number of DESY topics per article, with a mean of 7.7 and a median of 7.

were extracted as phrases of interest. The number of phrases from the corpus of abstracts was over 150,000, while the full-text corpus had over 2 million phrases. Although some of the phrases were subsequences of longer phrases, all phrases that appeared at least once, delimited by non-vocabulary words, were kept. For each of these phrases, the citation subgraph of articles containing it was used to compute the network metric *compensated normalized link count* (**CNLC**) via

abstracts	full-text
tachyon	*loop dilation*
string field theory	*mhv*
wave background	*wrapping numbers*
vacuum string field theory	*wave background*
higher spin	*tadpole conditions*
anomalous dimensions	*vacuum string field theory*
quantum einstein gravity	*level truncation*
penrose limit	*integrable spin chain*
string gas cosmology	*bmn operators*
twistor	*negative helicity gluons*

Table 4.2: A few of the top phrases from abstracts and full-text, ranked by our metric CNLC.

eqn. (4.2). Note that if a phrase is a subsequence of another phrase, then the citation subgraph for the shorter phrase is a superset of the citation subgraph for the longer phrase.

We obtained a global ranking of the phrases by ordering them according to decreasing CNLC values. Table 4.2 shows a few of the top phrases. For comparison, we have also ranked the phrases by the only other network metric used in a similar context of topic ranking, the log of odds ratio (**LoOR**) as described in [50]. For each article, the top k phrases contained in the article were extracted using these global rankings. A **tf–idf** ranking of phrases per article formed a convenient baseline for comparison with the network-based rankings. Rankings by **tf–CNLC** (term frequency times CNLC) and **tf–LoOR** (term frequency times LoOR) were also computed.

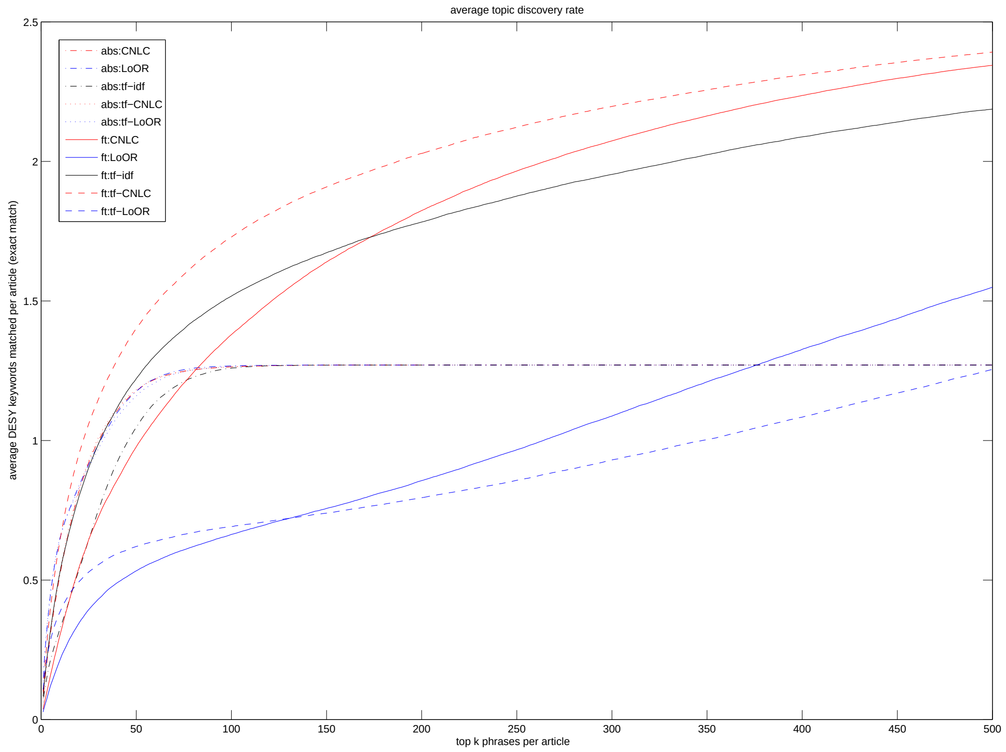


Figure 4.3: Comparison of rankings by CNLC, LoOR, tf-idf, tf-CNLC and tf-LoOR on the two corpora of abstracts (abs) and full-text (ft) using DESY topic keywords. The curves display the average number (per article) of DESY keywords matched *exactly* in the top *k* phrases extracted by the rankings.

4.4.5 Exact Match Evaluation

The first comparison among the different rankings is shown in fig. 4.3. For the top *k* phrases extracted from each article, according to each ranking, we have computed the number of DESY topics for that article *exactly* matched i.e. contained the same words, in the same sequence and neither of them longer or shorter than the other. The curves in fig. 4.3 show the number of topics matched, averaged over all the articles.

For abstracts, we see in fig. 4.3 that the network-based metrics, CNLC, LoOR, tf-CNLC and tf-LoOR, have similar performance and are better than the simple tf–idf. The flattening out of the curves for abstracts results from the word limit on the abstract lengths. All the rankings on abstracts perform better than LoOR and tf–LoOR rankings on full-text.

The tf–idf ranking on full-text initially performs very similarly to abstracts, and then continues to improve. But beyond the top 200 phrases, CNLC on full-text performs better than tf–idf. The tf–CNLC ranking on full-text, however, performs better than all other rankings, on abstracts and full-text. CNLC and tf–CNLC seem to converge asymptotically as more phrases per article are considered.

Note that the LoOR in [50] was used for both extracting and ranking topics. In our method, there is a separate vocabulary selection step that excludes many phrases, so the curves in fig. 4.3 compare only the ranking part of the two methods. Direct comparison with the keyphrase extraction literature [92, 4, 49, 98, 56, 66] is difficult for two reasons. First, there are different preprocessing steps for these algorithms (noun phrase extraction for example). Second, most of these algorithms are supervised learning methods, whereas our method is unsupervised. Nevertheless, the low precision of all of these algorithms underlines the difficulty in exactly matching human-labeled topics, and we see a reflection of that in our results as well.

4.4.6 Jaccard Similarity Evaluation

Not only is the exact match of phrases with the DESY topics difficult, it may also be less useful for phrases longer or shorter than the topics, denoting different levels of granularity of the same general concept. For this reason, we have relaxed the matching condition by first considering the phrases as sets instead of sequences, and then using a similarity measure on the sets. For similarity measure, we use the well-known **Jaccard similarity**, defined as the ratio of sizes of set intersection and union: if two phrases p_1 and p_2 have sets $\mathbf{P}_1$ and $\mathbf{P}_2$ of words, then the Jaccard similarity is

$$\text{sim}(\mathbf{P}_1, \mathbf{P}_2) = \frac{|\mathbf{P}_1 \cap \mathbf{P}_2|}{|\mathbf{P}_1 \cup \mathbf{P}_2|} , \tag{4.3}$$

taking values between 0 and 1. Even the maximum value of 1, however, is relaxed compared to the exact match, since the former indicates identical sets whereas the order of the words may still be different. For each of the DESY topics, we have identified the phrase with the highest Jaccard similarity, in the top k phrases per article. The number of topics matched by an article is then the sum of the similarities of these phrases. The average number of topics matched by an article is shown in fig. 4.4.

The relaxation from exact match to Jaccard similarity in fig. 4.4 results in more DESY topics matched on average per article. The patterns nonetheless remain very similar to the exact match curves in fig. 4.3. For abstracts, CNLC and tf–CNLC show better performance, followed by LoOR, tf–LoOR and tf–idf. For full-text, tf–CNLC once again shows the best performance. We also notice that the performance of LoOR and tf–LoOR is inferior to that of CNLC, tf–CNLC and tf–idf. As in fig. 4.3, CNLC begins to outperform tf–idf after about the top 200 phrases, and CNLC and tf–CNLC ultimately seem to converge.

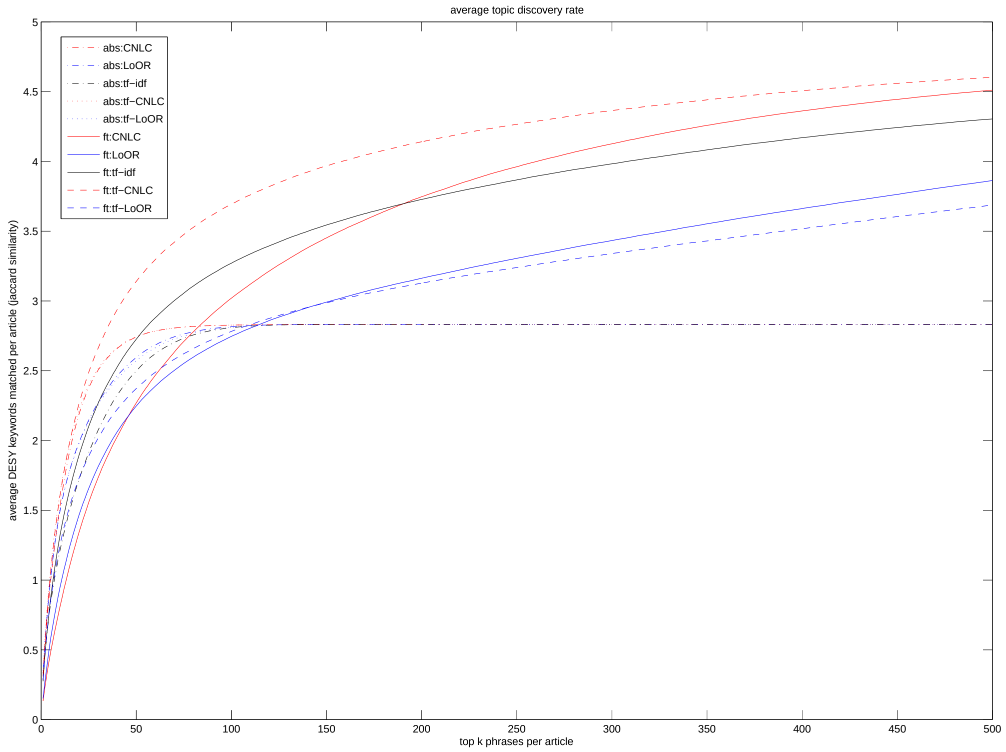


Figure 4.4: Comparison of rankings by CNLC, LoOR, tf-idf, tf-CNLC and tf-LoOR on the two corpora of abstracts (abs) and full-text (ft) using DESY topic keywords. The curves display the average number (per article) of DESY keywords matched using Jaccard similarity in the top *k* phrases extracted by the rankings.

Interestingly, tf–CNLC on full-text in figs. 4.3, 4.4 discovers more topics than all other methods on abstracts. This indicates that although abstracts should in principle be well-written summaries of the full-text, they do not in practice contain all the relevant phrases. So many text-mining tasks clearly benefit if run on full-text: not surprisingly the length restriction on abstracts prevent inclusion of many relevant phrases, particularly for longer articles.

In figs. 4.3, 4.4, we have counted the actual number of DESY topics matched by the top *k* phrases. Different articles have different number of topics associated with them (fig. 4.2 shows the distribution). So the quantity of interest is the average number of topics matched, but as a fraction of the total number of topics for each article. The curves for such fractions turn out to be sufficiently similar to the curves in figs. 4.3, 4.4 that little new signal emerges, and thus for brevity are not presented here.

4.4.7 Queries

Although human annotation can be high quality, it has its drawbacks. Ref. [4] mentions problems with author-provided keyphrases, such as 25% of such keyphrases not taken from the text itself, only 2–3 keyphrases provided per article, and keyphrases aimed towards classification rather than content summarization. Author-provided keyphrases may also not be available for many article sets. Use of human judges has the difficulty of being time consuming and expensive [4]. In addition there is the problem of inter-judge agreement. Even ignoring the issues with human annotation, our goal of finding subtopical phrases demands evaluation beyond coarse topics. Assessment of phrases not matching any DESY topics in figs. 4.3, 4.4 demonstrates the crucial difference in the level of granularity of our method vs. previous works on keyphrase extraction and topic detection.

Queries submitted by users provide implicit human annotation, as well as giving an implicit window into the way users cognitively represent concepts. Our goal is to use these queries as phrases of interest. Since arXiv's search engine defaults to encompass all subject areas, for present purposes it is necessary

queries
non-extremal black hole
noether theorem
wilson fermionic action
zeta-function regularization
rieman-cartan
tunneling
wignar rotation
gribov horizon
scalar quark
zero point energy

Table 4.3: Sample search queries classified as hep-th and perfectly matching some phrase from full-text.

to extract only those queries that were submitted with the intention of finding high energy physics theory (hep-th) articles. For a small subset of the queries, the subject areas were explicitly passed as an option passed to the search engine, but the majority of queries require classification.

One simple way of classifying a query is to resubmit it to the search engine and examine results retrieved. If most of the articles are from a particular subject area, we can classify the query to that area. This can be slow, however, and as well the results may be time-dependent, and moreover might not unambiguously capture the intent of the user. Instead we have mined the logs and tried to match the search query log with the clickthrough data.

For each query, we have examined the clicks from the same ip address within 30 minutes following the query submission. This 30 minute session length heuristic was shown to be effective in [75]. Examining the click log during the session, we determined which subject area was clicked most often (abstract or pdf download) and assigned that as the desired class of the query. The idea is to use the click trail following a query to infer the user's likely initial subject area of interest, in order to restrict attention to hep-th practitioners to evaluate their query string usage. About 6700 search queries were classified as high energy physics theory (hep-th), of which about 4000 had some common word with the phrases extracted, and were thus used for evaluation. Table 4.3 shows a few of these queries that perfectly match phrases in our corpus.

This method as well has issues of course: multiple users behind a single proxied ip address will have their queries and click trails inadvertently intermingled. Even a single user with interests in multiple fields, and short search sessions in quick succession, may result in incorrectly classified queries. The multitude of internet-connected devices in modern life, and dynamic addressing of these devices, make identification and aggregation of a single user behavior through ip addresses even more challenging.

Although sophisticated learning algorithms could be employed for the task of query classification, the simple method used here is commensurate in simplicity with the method of finding and ranking phrases that is to be evaluated. Furthermore, the metric we use to compare search queries with extracted phrases is robust to the presence of some noise in the query classification.

4.4.8 Search Query Evaluation

The rank of a phrase is its position in a global ranking, and is normalized to have values between 0 and 1, where low values are ranks near the top and high values are near the bottom. We define the rank of a search query to be the rank of the most similar phrase, as measured by Jaccard similarity. In other words, a query is considered equivalent to a phrases to which it's highly similar.

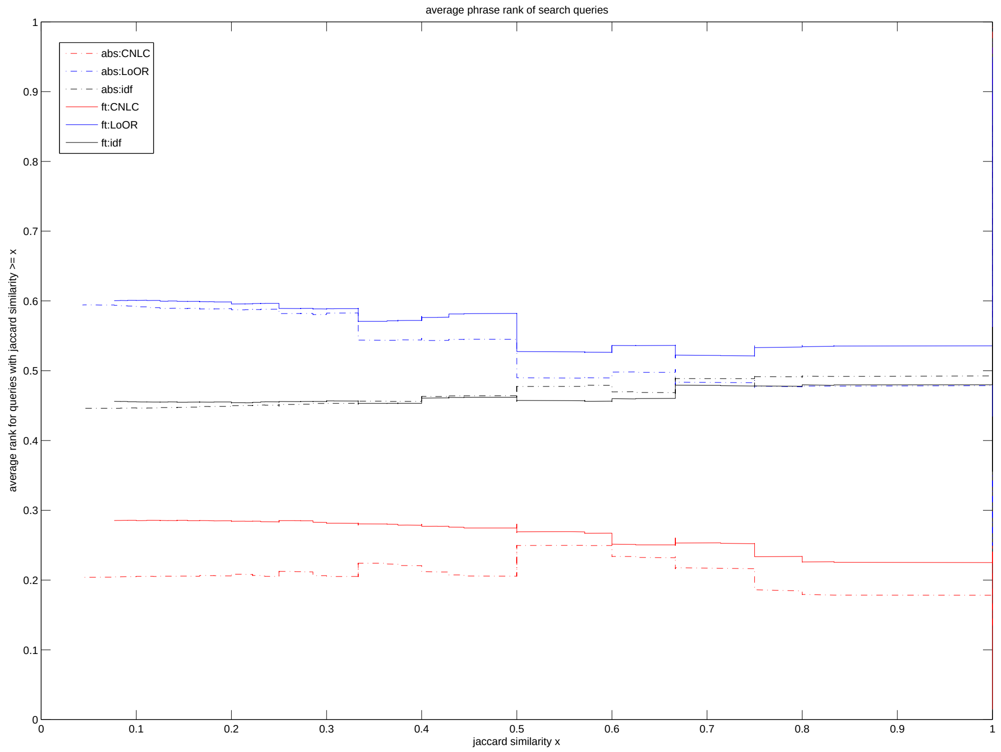


Figure 4.5: Comparison of search query rankings by CNLC, LoOR and idf on the two corpora of abstracts (abs) and full-text (ft). For each value of Jaccard similarity x on the x-axis, the y-axis shows the average ranks of search queries having a Jaccard similarity equal or greater than x. Ranks are normalized to have values between 0 and 1, where low values are ranks near the top of the list and high values near the bottom.

Mathematically, if $r(\mathbf{p})$ is the normalized rank of a phrase $\mathbf{p}$, then the rank of a search query $\mathbf{q}$ is

$$r(\mathbf{q}) = r(\mathbf{p}^*_{\mathbf{q}}) \text{ where } \mathbf{p}^*_{\mathbf{q}} = \text{argmax}_{\mathbf{p}} \text{sim}(\mathbf{p}, \mathbf{q}) \ . \tag{4.4}$$

If there are m search queries, then the average rank of the queries is $\bar{r} = \frac{1}{m} \sum_{\mathbf{q}} r(\mathbf{q})$, giving their expected rank. A lower average query rank is preferable, indicating that the set of search queries inherits a high ranking from the corre-

	average		median	
	abstract	full-text	abstract	full-text
LoOR	0.59	0.60	0.84	0.81
idf	0.45	0.46	0.42	0.43
CNLC	**0.21**	**0.29**	**0.04**	**0.13**

Table 4.4: Aggregate query rank on abstracts and full-text. The best (smallest) value of each column is highlighted.

sponding phrase ranking. The first two columns of numbers in table 4.4 show the average rank of those queries that were classified as hep-th, as described in the previous subsection. Since the standard deviations were high, the medians are also shown in the last two columns of table 4.4. Results overwhelmingly favor the CNLC ranking metric.

We need a similarity cut-off threshold, so that a search query with only very low similarity to a high-ranking phrase is not considered equivalent to it. To find a natural value of the similarity cut-off, we investigated the effect of similarity cut-off threshold on the average query rank (see fig. 4.5).

For a threshold x, let S_x be the set of queries with similarity at least as high as x.

$$S_x = \{\mathbf{q} : \mathrm{sim}(\mathbf{p}^*_{\mathbf{q}}, \mathbf{q}) \geq x\}$$

If $x_1 \leq x_2$, then $S_{x_1} \supseteq S_{x_2}$. The average query rank at x is

$$\bar{r}_x = \frac{1}{|S_x|} \sum_{\mathbf{q} \in S_x} r(\mathbf{q})$$

The lines in fig. 4.5 show the average rank of search queries for various cut-off thresholds.

In fig. 4.5, we observe that CNLC rankings produce lower average query ranks for all similarity thresholds, while LoOR performs worse than idf. It is reassuring to observe the downward trend with increasing threshold in the full-text CNLC curve. This indicates that queries with low phrase similarities have overlaps with less interesting phrases. The step patterns of the curves are due to the integer step values in the size of the set intersections.

We note that the average rank of queries in the corpus of abstracts is lower than that of full-text. It is tempting to interpret this through two mutually dependent expectations: of authors putting phrases likely to be searched in the abstract, and users searching through the abstract more than the full-text for faster retrieval. Although fig. 4.4 shows better topic discovery on full-text than abstracts, even for the top few phrases per article, the lower average rank of queries for abstracts in fig. 4.5 indicates the difference between topic labels and user search queries, rather than any difference in the effectiveness of our method of phrase extraction on full-text vs. abstract corpora.

4.5 Conclusion

In this chapter, we have demonstrated how subtopical concepts, characterized by phrases, can be extracted and ranked. Using both explicit and implicit forms of human assessment, the method has been shown to be effective, and superior to word frequency and network-based rankings. The implicit assessment, using search terms harvested from query logs, suggests that the phrases extracted by our method play an important role in researchers' internal cognitive representations of the associated concepts. Although abstracts are frequently considered to be well-written summaries of the full text, and thereby usable as effective proxies in many text-mining contexts, we have found that certain tasks require analysis of the full text, and thus simple and fast algorithms are necessary for very large corpora.

In chapter 5 we show how concepts can be traced by tracking the readership patterns of the topics. If we cluster such patterns, we obtain a global understanding of trends for different concepts. We conclude with some possibilities for future extensions of this work. Phrases can be used to compute subtopical projections of documents that may be useful for a variety of tasks, such as the recommender systems in [59, 46], in which keywords assigned to documents are used to classify reader interests. Document representations can also be refined by adding phrases to the bag-of-words model. Query completion without building a history of queries is as well possible through use of these phrases. Finally, it will be interesting to apply the methodology used here to temporal tracking of phrases in other linked corpora, including focused subnetworks of the WorldWideWeb, to provide useful temporal overviews of them.

CHAPTER 5

CLUSTERING TOPIC CLICK TRENDS

Temporal tracking of scholarly topics is essential for understanding the dynamics of research fields. In chapters 2, 3 we have noted that online readership, approximated by clicks, has potential use in impact metrics and recommender systems. In this chapter, we present a method of aggregating accumulated readership information by topic. Our method involves coarse binning to mitigate the effects of sparseness and sharp spikes, and hierarchical clustering of normalized trends to discover general patterns of topic interest. We demonstrate our method on 7 years of click data for 23,000 arXiv articles containing 10,000 distinct human assigned topics.

5.1 Introduction

Generation and dissemination of electronic information has continued to accelerate over the past decade, furthering the establishment of a "creative commons". Increasing information overload argues both for better filters and for content summarization, exemplified by the "one minute world news" on popular news sites. Such condensation of information through direct human labor, however, is not generally scalable. Recently, it was shown [61] how systematic tracking of news quotes can provide an overview of evolving trends, and such tools may play a role in future information filtering methodology.

Although scholarly communication typically proceeds on longer timescales than newspace/blogspace interactions, the large amount of scholarly content on the web has long called for similar automated filtering and recommendation systems. With acceleration of human scientific endeavors, it has become difficult even for expert researchers, let alone lay readers, to acquire a broad understanding of an active area. Suitable tools for providing broad overviews would be of use for the full spectrum of readers, and as well to help policymakers determine where to allocate funds.

Traditional bibliometrics has emphasized active indicators of scholarly consumption, primarily citations. Ref. [78], for example, analyzes citations for articles spanning a timescale of over a century. Use of readership data as an additional metric has become feasible in the past two decades, as a greater percentage of literature usage has moved on-line. Bibliometric properties of readership have been investigated in [60], in which various hybrid assessments metrics were considered. Readership metrics are typically strongly correlated with citation metrics [60, 70, 13], and are hence highly predictive of data only later available, after some months to years. The predictive power of user clicks for individual articles, as early as within the first two weeks of on-line availability, was confirmed in chapter 2 of this dissertation. Recommender systems using readership information have been described in [59, 46].

In this chapter, we present a method of aggregating readership patterns of scientific articles for use in understanding more global properties of research areas. We have temporally tracked interest in topics and subtopics, as characterized by human-assigned keywords, and clustered them to understand readership patterns and to provide smoothed representations of individual topic trends. In [81], topic interest was measured via the volume of publication. A more immediate and dynamic temporal overview of topic activity is possible through user clicks. We have used the arXiv preprint system as the source of readership information. 7 years of click data for 23,000 arXiv high energy physics articles, containing 10,000 distinct human-assigned topics, are analyzed in this article. Detailed click trends for each of the 10,000 topics are available at the website http://www.cs.cornell.edu/~asif/clicktrends/.

Public presentation of topic click trends has its drawbacks. One undesirable effect of topic clicks, if used for recommendation, could be an excessive reenforcement of existing trends. We have already documented the long-term effects of presentation bias earlier in chapters 2, 3. If readership information is publicly available, it also becomes vulnerable to manipulation. In this chapter, we have normalized the temporal patterns such that trends are presented without magnitudes. This makes manipulation harder, although not impossible, but reduces utility of these trends for recommender systems.

5.2 Data

Click data, excluding robotic access, between January 2000 and December 2006 for almost 23,000 arXiv High Energy Physics – Theory (hep-th) articles published during the same time period was used for our experiments. About 10,000 distinct topic keywords for these articles were obtained from the SLAC Spires High-Energy Physics Literature Database[1] in late September of 2010. These topics were assigned by the German research center for particle physics Deutches Elektronen-Synchroton (DESY)[2] library staff. Articles under consideration had roughly 7 topics on average. About 4.5% of the 23,000 articles did not have DESY topics associated with them, and were thus ignored.

[1] http://www.slac.stanford.edu/spires/

[2] http://www.desy.de/

5.3 Method and Results

Daily click patterns for hep-th articles were examined in chapter 2. The most common pattern was high levels of activity during the first several weeks after announcement, but then rapid decline. Beyond this initial period, click data can be sparse, with only occasional clicks due to uncorrelated activity from individual researchers, or can have large spikes within short time periods, due to prominent citation by a highly active article, or mention in some popular blog or news. (Some articles, typically review articles, tend to have very long persistent tails of clicks, even over timescales of longer than a decade.) Instead of undertaking the difficult task of modelling bursty traffic [55], we have used bin sizes much larger than a day to smooth out both of these effects. Occasional clicks are accumulated in larger bins while large spikes are spread out. Use of larger bins also helps to reduce the dimensionality of the data, and hence to aggregate the patterns into fewer clusters. We have used bins of duration 4 months, so there are $7 \times 3 = 21$ bins in total for the 7 years of data considered here.

While click trends for individual articles are informative to examine, it is also useful to understand the overlying topical dimensions of these trends. For this purpose, we have aggregated the clicks for the DESY topics by summing up the clicks of all articles associated with each topic. The same bin size of 4 months was used for these topic click trends. In order to compare trends of different magnitude, the patterns are L2 (square root of the sum of squares) normalized. Fig. 5.2 shows the click trends for two topics in high energy physics. Graphical representation of the click curves for each of the 10,000 DESY topics is available at http://www.cs.cornell.edu/~asif/clicktrends/.

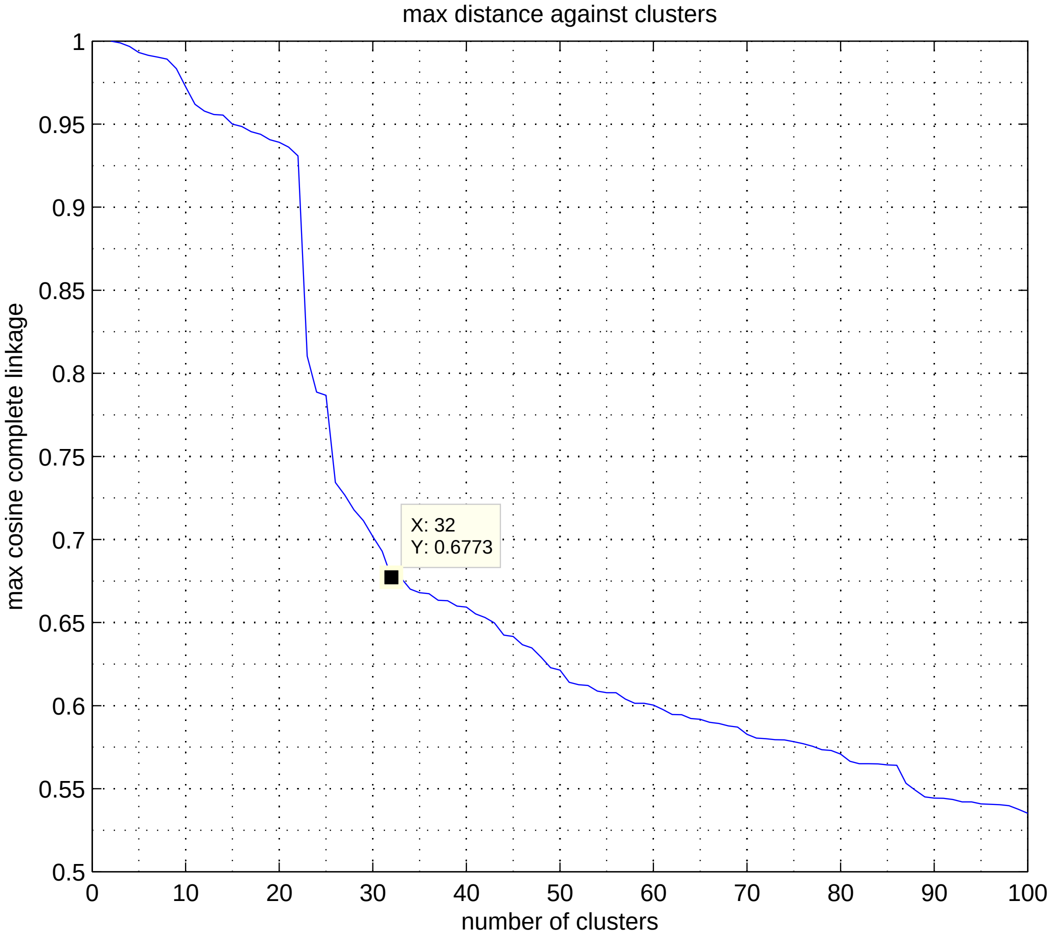


Figure 5.1: Maximum complete linkage among clusters for different number of clusters. The chosen number of clusters is shown as the dark bullet.

Further aggregation of topic clicks is possible by clustering the individual patterns. In [1], a similar task of clustering web revisitation patterns was undertaken via exponential binning and normalization by bin averages. We have experimented with symmetric distance measures such as euclidean distance and linear correlation, and asymmetric measures such as KL-divergence, but discovered cosine distance to be most meaningful in our setting. (This is in accord

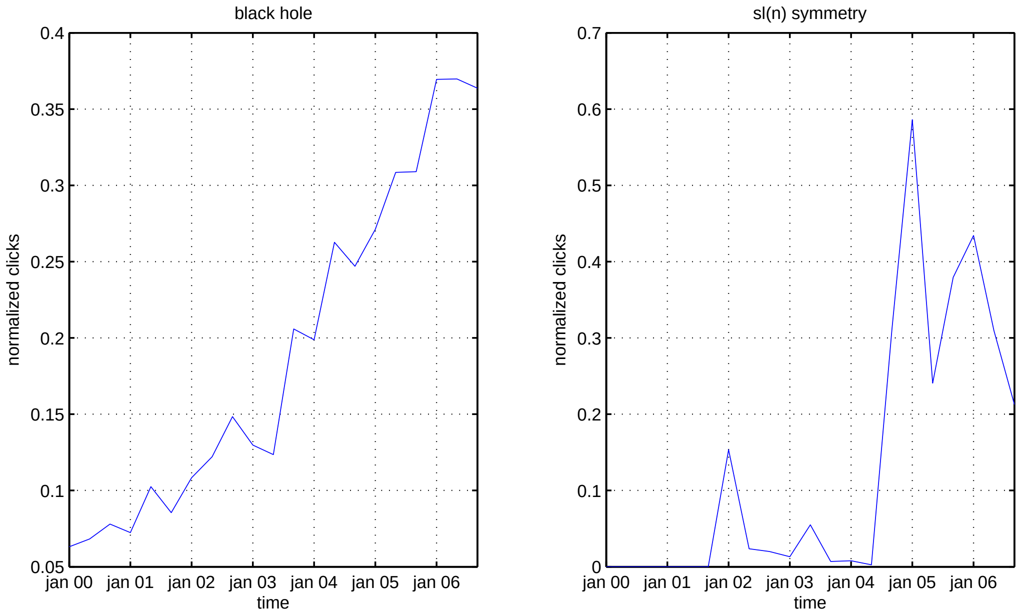


Figure 5.2: Click trend for (a) *black hole* showing steady growth of interest, and (b) *sl(n) symmetry* showing sharp increase in mid 2004 and decline afterwards.

with [1], although web revisitation patterns are different from our click trends.) We have employed hierarchical clustering since it has two desirable features: efficiency and the ability to zoom in and out. Finally, we have chosen to use complete linkage on cosine distances for merging. If X, Y are two clusters then the complete linkage between the two clusters is

$$\mathbf{D}(X, Y) = \max_{x \in X, y \in Y} \mathrm{d}(x, y) ,$$

where $\mathrm{d}(x, y)$ is the cosine distance between the click patterns for topics x and y.

Our means of finding the appropriate number of clusters is shown in fig. 5.1. Starting with all topics as one cluster, we successively split the cluster with the largest complete linkage between its two parts during hierarchical merging. Decrease in the maximum complete linkage was rapid until reaching 32 clusters, and the rate of decline continued to slow down beyond 32 clusters. Since our data is of dimensionality 21 (7 years × 3 four-month bins), the 32 clusters roughly correspond to patterns with a single bump at various places, and a few other important combinations of bumps, including a slightly increasing trend (examples shown in fig. 5.2). A number of clusters roughly on the order of the data dimensionality is consistent with the observations in [1].

Fig. 5.3 shows the 32 clusters as L2 normalized curves averaged over the topics in each cluster. Roughly two-thirds of the patterns have one single bump, either due to a burst of publication for a topic, or mention of an article containing that topic in popular media, resulting in the "slashdot effect". There are a few clusters whose curves are increasing throughout. The total readership of arXiv increased slightly with time during this period, and these curves represent topics with sustained research activity. While it is possible to normalize these curves even further by compensating for this increase in overall click volume, we have chosen not to perform such normalization so that click trends for the same topics on different scholarly systems may be directly compared.

In fig. 5.3, we see various interesting patterns. Full commentary on these trends is outside the scope of this chapter, but a few clusters deserve brief comment. Clusters in [row 2, column 1], [row 6, column 4], and [row 8, column 2] show the aforementioned increase with time. The readership pattern for *sl(n) symmetry*, shown in fig. 5.2(b), belongs to the cluster shown in [row 6, column 1]. The 17 topics in [row 8, column 4] received steady clicks, except in early 2001 and late 2005, when more readers followed these topics. The cluster of topics in [row 8, column 1] similarly contains two smaller bumps, except that they are separated by a year and half, rather than the nearly 5 years in [row 8, column

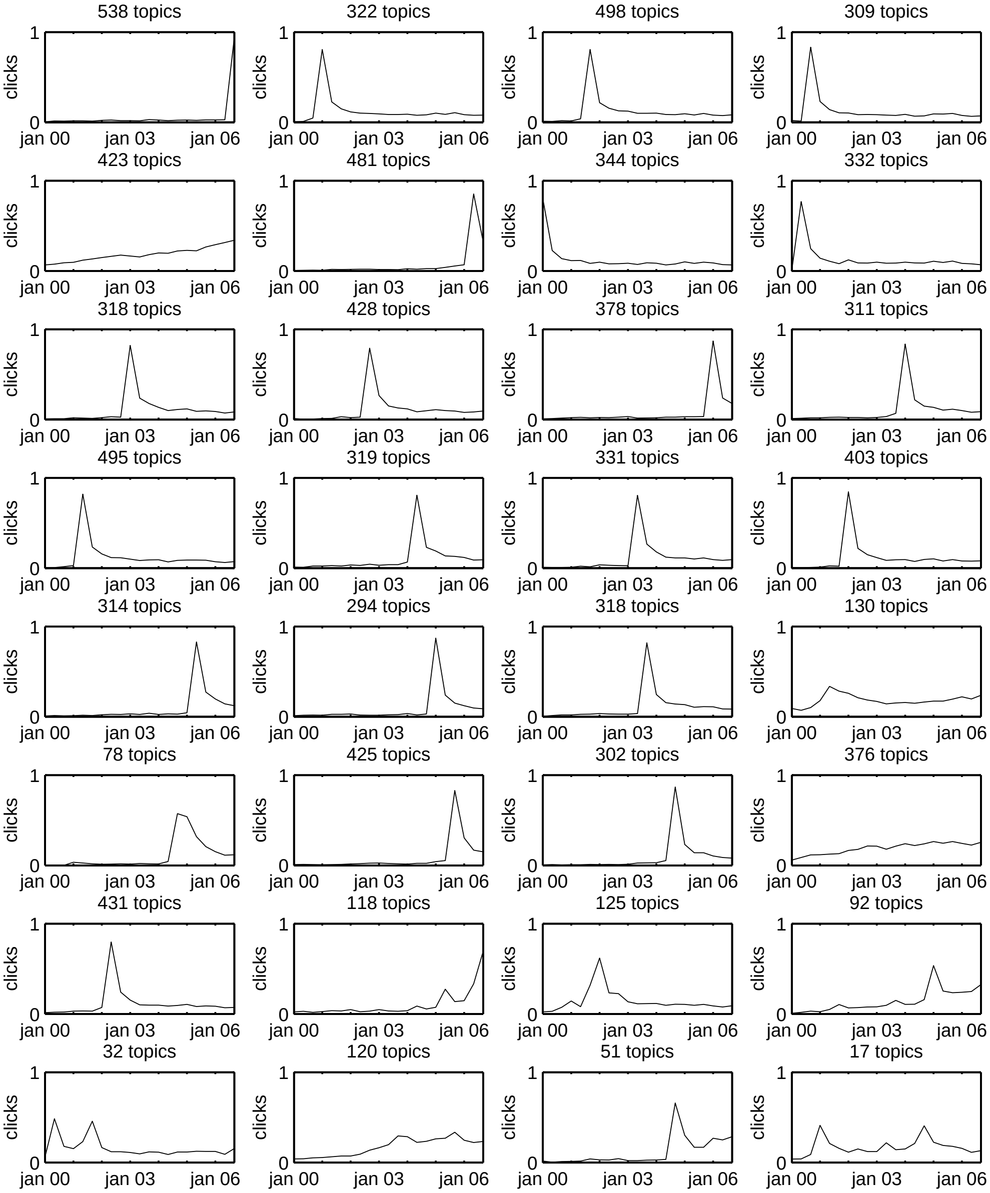


Figure 5.3: L2 normalized click trends averaged for each of the 32 topic clusters. The size of each cluster is mentioned on top of the corresponding figure. The x-axis shows time, marked by the first month of each year, while the y-axis shows normalized values of clicks between 0 and 1.

4]. It is possible to zero in on the individual articles within these clusters, and, using some combination of co-readership and text similarity metrics, easily disentangle any independent topic contributions for use in recommender systems or more refined temporal tracking.

5.4 Conclusion

In this chapter, we have advocated the utility of analyzing readership information in the form of user clicks, and shown how the patterns can be aggregated to obtain overviews of research areas and smoothed representation of topic trends. We have used the click log of the arXiv preprint system to obtain readership patterns associated with human-assigned topics, and clustered them using hierarchical clustering. The data for our experiments involved 7 years of click data for 23,000 articles containing 10,000 distinct topic keywords.

Obtaining human-assigned keywords for scientific texts does not scale to large corpora. The aggregation and clustering presented here can also be applied to topics and subtopics discovered algorithmically. If implemented in online recommender methodology, it will be important to investigate how the presentation of click information reenforces existing scholarly trends (or fads), and whether manipulation of click data can be screened.

Part IV

Network Analysis

CHAPTER 6

RESOLVING NAME HOMONYMY FOR MESOSCOPIC ANALYSIS

In this chapter we investigate the issue of author name homonymy in the context of co-author network analysis, and present a simple, effective, scalable and generalizable disambiguation algorithm. We evaluate the performance of the algorithm to improve the resolution of mesoscopic network structures. To this end, we establish the ground truth for a sample of author names that is statistically representative of different types of nodes in the co-author network, distinguished by their role for the connectivity of the network.

6.1 Introduction

A nascent stream of research in scientometrics, policy research, and social studies of science and technology analyzes co-author or citation networks to obtain a better understanding of scientific collaboration and the social organization of science. Author name ambiguity compromises this analysis and it is essential to remove this noise as the study of network structures becomes more sophisticated and moves beyond global measures of network topology to mesoscopic network features. Whereas in the past, e.g. for the evaluation of scientists based on their publication output, manual disambiguation of author names was feasible, large scale network studies require automated methods.

We present here a simple, effective, scalable and generalizable algorithmic approach for name disambiguation, and evaluate its performance in the particular use context of co-author network analysis. Based on our observations we suggest a new approach to assessing the quality of name disambiguation in co-author networks that does not require the expensive investment of establishing the ground truth for a representative sample, but builds exclusively on measures that can be derived from a structural analysis of the network itself.

Name ambiguity can be classified into two kinds of problems: synonymy and homonymy. Here we focus on name homonymy, which in the remainder of the chapter we refer to as “name disambiguation”. In name homonymy, different individuals have the same name, either due to coincidence or abbreviations of names such as using initials for given names instead of using the full name. Homonymy is a problem especially for names coming from naming practices, such as those in Korea or China, that may have uniquely identifiable full names but very common last names.

Effective and generalizable author name disambiguation remains a generally unsolved problem for the following reasons. First, different databases provide different kind of information about articles and authors (the feature-set used for disambiguation), making it hard to devise a general algorithm. Second, the tolerance for errors and for different types of errors will differ between use contexts. Third, the methods for evaluating the effectiveness of a disambiguation algorithm are not well-established. No comprehensive, standardized set of benchmark data exists due to the variety in use contexts, the range of possibly relevant features of a dataset, and the costs of manually establishing ground truth. Finally, some algorithms do not scale for large data sets. All these concerns have resulted in a variety of algorithms for name disambiguation.

We consider here the problem of name ambiguity in the context of earlier work presented in [94]. In this earlier work, we analyzed co-authorship networks to better understand patterns of scientific collaboration in different scientific fields. We combined ethnographic methods with network analysis to identify co-author clusters in a co-author network as the smallest collective units of research in a field, and to extract linking patterns that represent different kinds of cooperative relationships between such collectives. A subnetwork of particular interest are the co-author clusters in a specialty field that show intensive inter-group collaboration. This chapter addresses the fact that this network, which was based on non-disambiguated author names, showed peculiarly dense clustering for research groups with Asian affiliations, suggesting distortions due to name homonymy.

Our evaluation method is relatively novel compared to previous approaches because it takes network structural properties explicitly into account. We have extracted and quantified mesoscopic network features by classifying the nodes in a clustered co-author network into seven different classes of node roles based on their cluster internal and cluster external linking, following a classification scheme introduced in [40]. Given our suspicion of network distortions due to homonymy, we are interested in learning how those classes of nodes are affected by name homonymy. To establish the ground truth, for each class of nodes we have sampled a representative set of author names and manually disambiguated them. Based on this node role stratified sample, we evaluated the node role specific performance of our disambiguation algorithm, and obtained estimates of the network distortions due to name homonymy.

Our algorithm for name disambiguation is fairly simple, yet effective, and can easily scale up for large networks. We consider two articles with the same name to be by the same individual if either there is a co-author that is common in both the articles, following an approach by [54], or if there is a citation from one article to the other, which we interpret as a self-citation. Co-author overlap is easy to compute and very effective, while self-citation leverages an author's research continuity. One novel feature we have used in our algorithm is the commonality of last names, which we operationalized as author name redundancy by counting the number of variations of initials of a last name within our data set. Hence the distribution of name redundancy can be easily obtained from the data set itself, and we present a principled way of using this information for excluding less common names from unproductive disambiguation attempts.

6.2 Related Work

There is a large body of work on name disambiguation which falls under the general area of entity resolution (see [83] for a broad overview). These methods employ either supervised or unsupervised learning.

In supervised learning a smaller set of names is manually disambiguated so that a classification model can be trained. In [41] techniques such as naive bayes and support vector machines were employed effectively. The drawback of such methods is that the training set needs to be large enough for the classifier to extrapolate unseen data accurately. This re-introduces the problem of manual disambiguation of large sets of names.

Unsupervised learning uses clustering based on similarity metrics between names [42]. Generative models such as latent dirichlet allocation and topic-based probabilistic latent semantic indexing have also been used [7, 48]. The tricky part of using unsupervised learning is to judiciously choose the similarity metric and the clustering algorithm. In [48] the similarity metric was learned from a set of similarity metrics via online active learning.

There are methods that tried to combine the benefits of supervised and unsupervised learning. In [91, 31] training sets were generated automatically from the data. Such training sets have noise in them and algorithms must not overfit by learning the noise.

Whether learning is supervised or unsupervised, feature availability in the data and feature selection is of paramount importance. Features regularly employed are co-author names, affiliation, article title, journal names and topic keywords [91, 83, 31, 41, 42, 48, 54]. Unfortunately, affiliation on an author basis is not regularly available, nor are standardized keywords. Co-author names have been shown to be extremely effective [91, 54], even by itself [54], and they provide a feature that is generally available in any data set of interest to author name disambiguation. Topics from article text were used in [86] while random walks on co-author networks were used in [67]. An entirely different set of features arises from reference or citation networks. For example, self-citation was used in [68] and co-reference was used in [90].

Because of overwhelming evidence in favor of co-author names we have chosen it as the main feature. We have also used self-citation to gain more accuracy on top of co-author patterns. By using both co-author and citation based features we have broadened the grasp of our algorithm. One novel feature introduced in this work is the quantification of the variety of first name initials associated with last names as an indicator of last name commonality.

Our algorithm falls under the category of unsupervised learning where we have blocked the authors by their names and clustered them using co-authorship and self-citation. It relies on clustering as simple as finding connected components on co-author overlap graphs, making it useable for large scale network analysis (see chapter 7 for insights into fast distributed network analysis). The necessity for simplicity in large scale disambiguation was correctly noted in [83] and a recent attempt of disambiguation in the context of network analysis was presented in [88].

We do, however, have one parameter in our algorithm that was learned from a small set of manually disambiguated names. So our method is semi-supervised in some sense. But this parameter is based on a straightforward intuitive consideration, and the empirical determination mainly served to verify this intuition. We suggest that the learned value for this parameter can be safely applied to other data sets, so that our algorithm could be run in an unsupervised manner.

Because of the context of network analysis, our evaluation method is significantly different from previous works. Although name ambiguity is apparent in most standard bibliographic datasets, the importance or effect of disambiguating these authors is not apparent. In our evaluation we have taken into account the role of an author in a network and sampled authors from the seven roles (as presented in [40]) for manual disambiguation so that network structural effects of disambiguation can be assessed.

	Node	Characterization	Total	Samples
Non Hubs	R1	“ultra-peripheral nodes”	5167 (30.3%)	102 (1.97%)
	R2	“peripheral nodes”	8245 (48.4%)	102 (1.24%)
	R3	“connector nodes”	2527 (14.8%)	102 (4.04%)
	R4	“satellite connector nodes”	611 (3.6%)	89 (14.57%)
Hubs	R5	“provincial hubs”	195 (1.1%)	72 (36.92%)
	R6	“connector hubs”	257 (1.5%)	77 (29.96%)
	R7	“global hubs”	34 (0.2%)	28 (82.35%)

Table 6.1: Node role type distribution in the whole network and ground truth sample. The proportions with respect to the network size (in nodes) is also shown.

6.3 Data

The publication data used in this study has been obtained from the Web of Science database by Thomson Reuters using a lexical query to capture the publications of a specialty field in physical chemistry over a period of 22 years (1987-2008). The co-author network constructed from this data set of 29,905 publications, identifying individuals based solely on first name initials and last name, was introduced in [94]. When building the co-author network we filtered out and excluded from the network author names that had only one paper associated with them, and ended up with 18,419 nodes, representing authors linked by co-authorship, with a giant component of 17,250 nodes (93.7%).

Clustering of the co-author network using the information theoretic clustering in [79], exposes the modular structure of co-author relationships, and results in a network of clusters of closely collaborating authors. Each author node in such a clustered network can be classified into one of seven node role types introduced in [40]. A node is classified as a hub node or a non-hub node based on a first parameter, the number of its cluster internal links relative to the average inside-the-cluster degree of the nodes in the respective cluster. This means a hub node in a cluster has more cluster internal links than the average node of that cluster. A second parameter quantifies how a node distributes its outside links among the clusters and subdivides hub nodes into three groups, and non-hub nodes into four groups, both of which are ordered by increasing outside linking. See table 6.1 for characterizations of those type of nodes and their frequency in the giant component of our network. As reported in [94], based on this distinction between node roles, we can find a typical principal investigator (PI) led, hierarchically organized research group as a starlike structure, represented by a hub node in the center of a cluster with smaller nonhub nodes around, or a field-specific research institution or funded research network as a more egalitarian organized cluster with several hub nodes involved.

In this chapter we focus on the giant component of the coauthor network, and population statistics are based on all nodes in the giant component that can be classified according to *Guimera et al.'s* role type classification[1]. This population comprised 92.5% of the nodes in the entire (undisambiguated) network. For this population at least 75% of papers were published by coauthor teams of 5 or less authors (median 3, mean 3.8). The maximum number of coauthors found was 34.

As described below, the classification of author nodes is significantly distorted by author name homonymy, affecting in particular externally linking node role types (R3, R4, R6, R7). This is of concern; for the study of collaboration between groups, the resolution of nodes with role types characterized by high between-cluster linking is crucial, since they determine the connectivity of the inter-group collaboration network.

[1] For a few clusters zero standard deviation of the inside-the-cluster degree prevents calculation of the first parameter needed in the classification, resulting in the exclusion of 1.2% of nodes in the giant component from the population.

6.3.1 Name Redundancy

To capture the ambiguity of an author name due to homonymy we have introduced a measure of a name's commonality that we have derived from the data set itself. We call it "raw name redundancy", and it is obtained by examining how numerous variations of initials with the same last name are. For example, for the Chinese last name "WANG" we have found in our data set 740 instances of names containing the last name "WANG" that can be distinguished by their initials, like "WANG, CH" can be distinguished from "WANG, XL". Another example for a high scoring last name is the Korean name "LEE" with raw name redundancy of 511. A large portion of the last names appearing in our data set, 91.7%, have raw name redundancy of 3 or less. It is worth noting though that of the 86,389 co-authorship instances (an author being listed as a co-author for a paper), 52,913 (61.2%) are attributed to authors with raw name redundancies greater than 3, suggesting the larger number of actual authors represented by that smaller proportion of names.

If we observe a last name L to have $r_n(L)$ different initials associated with it in the dataset then we define its "name redundancy" $s_n(L)$ to be the cumulative normalized $r_n(L)$ value:

$$s_n(L) = \mathbf{Pr}[X \leq r_n(L)]$$

Here X is the random variable on $r_n(.)$ distribution, and $r_n(L)$ the "raw redundancy" of L. Last names with small raw redundancy will have name redundancy close to 0 while last names with many different initials will score close to 1.

Building on this definition we introduce as "article redundancy" the combined name redundancies of the co-author team writing an article, defined as

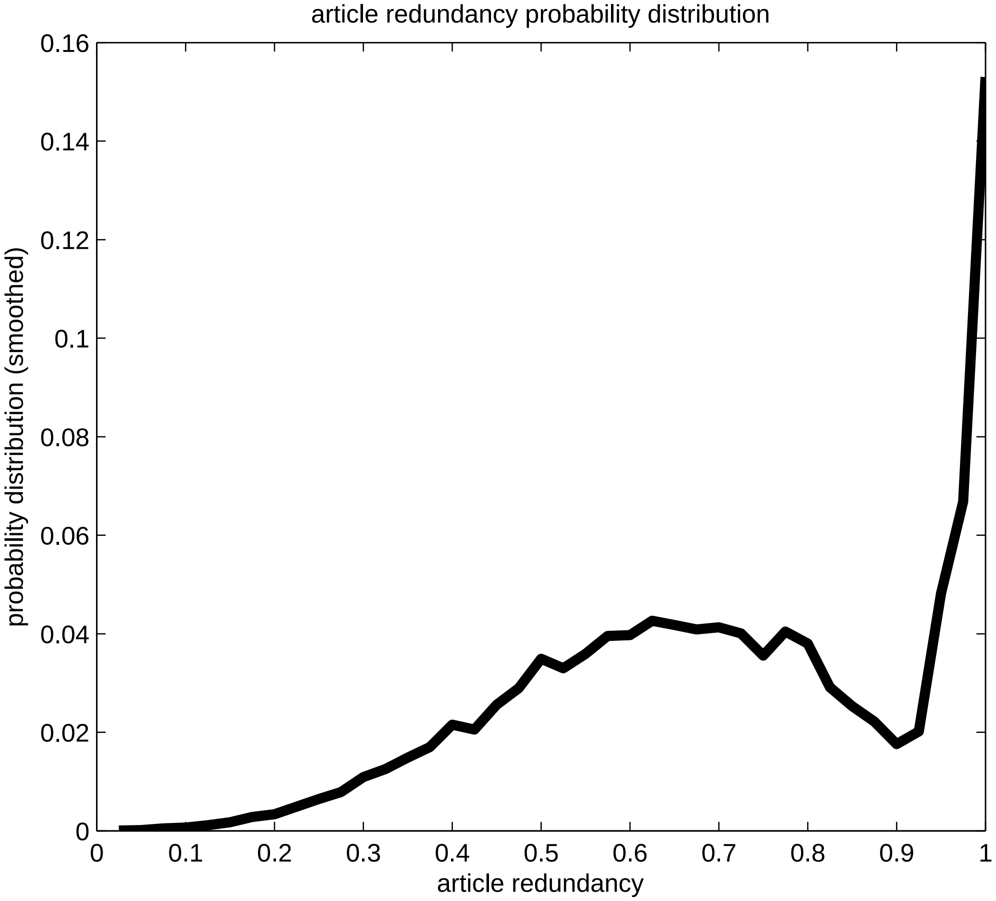


Figure 6.1: Smoothed probability distribution of article redundancy for the population data set resulting from the combined name redundancies of the authors of every article.

the product of name redundancies of the last names of the authors. The distribution of article redundancies for the articles of the authors included in our population data set shows two distinct regions, one symmetric broad distribution, and one narrow peak, in fig. 6.1. Those can be conceptualized as the overlap of two distributions. The broad distribution comprises articles with author teams that include one or several author last names with low name redundancy. Assuming an average number of co-authors per paper of roughly four authors, this distribution would result from the 4-fold convolution of distributions representing the independent, random choice of last names from the name redundancy distribution. The narrow peak on the other hand can be interpreted as the result of the convolution of distributions representing the independent choice of last names exclusively from the heavy tail of the name redundancy distribution. Upon inspection of manually selected samples we concluded that these are mainly East Asian, specifically Chinese and Korean, last names. Hence we suggest that the shape of the distribution in this diagram highlights the division of our data set into two components that are culturally (naming traditions) and geographically (co-location of closely collaborating authors) distinct.

6.3.2 Ground Truth

To estimate the error made by not correcting for homonymy in author names, and to quantify the improvement made by our disambiguation approach, we randomly sampled a subset of 571 author names from the population for manual disambiguation of author identities. To account for systematic differences between the different node role types, we stratified the sample by node role type and sized the sample strata to be able to make statements on sample proportions with at least a confidence interval of 10%, and a level of confidence of 95%. Sample sizes are reported in table 6.1. We sampled an additional 33% of author names for each groundtruth stratum to obtain a training set for verifying our intuition about a low-name-redundancy cut-off parameter that excludes extremely uncommon names from any disambiguation attempt.

To find information on the actual identities of authors with the same combination of last name and initials, we looked up full names and institutional affiliations, if given, in the full text version of articles. We further used biographic information and affiliation information gleaned from personal homepages and institutional web pages, as well as topic information from article titles and abstracts to establish topical closeness.[2]

Note that the groundtruth sample when aggregated across the node role strata does not reflect the actual proportions of node role types in the population (shown in table 6.1), simply because their relative proportions in the ground truth sample are not representative for their relative proportion in the population. Consequently, when interpreting results for the aggregate groundtruth set one has to keep in mind that one can make straightforward statistical estimates only within each stratum, i.e. for a specific node role type.

6.4 Algorithm

The basic idea of our algorithm is simple: two papers authored by an author with the same name are highly likely to be works of the same author if the two papers share common co-authors. Following [54] we use overlap of two coauthor sets by at least one last name as sufficient to merge two author identities. The result is the growth of connected components in co-author overlap graphs.

Furthermore, if a paper cites another, and both papers are authored by an author with the same name, then very often this a is self-citation reflecting the

[2] Obviously, even the "ground truth" is not necessarily the truth, because due to lack of evidence legitimate merges of identities may have been left out, and occasionally subjective judgements on topic closeness or similarity of institutional affiliation may have led to invalid merge decisions.

research continuity of that author. Although weaker than co-authorship, we have found signal from self-citation to be very accurate.

Finally, we have found that authors with last names that are unique in our data set are best disambiguated by considering every occurrence of such a name as referring to the same individual. The most uncommon last names will show up in our data set with raw name redundancy of 1. Intuitively, because often the same name is written with last name plus 2 or 3 different variations of initials, such as first initial, first and middle initials, or solely middle initial, we might want to include names with raw name redundancies of 2 or even 3 into that set of "unique" names.

We do not use affiliation and city information when available in our dataset since it is difficult to associate those with authors in a principled manner. We also do not use any text or topic content such as title, journal or keywords because of our dataset being from a narrow subfield of chemistry. These features may be discriminative for a large heterogeneous dataset like PubMed, but are less useful for a narrow research area where a lot of articles share the same keywords and are published in a few journals. We have investigated the applicability of tf-idf similarity of the abstracts and it indeed turned out to be less informative.

Thus we use in our method of disambiguation co-authors and self-citation on those names whose redundancy is beyond a certain value that we call the low redundancy cut-off, which we determine from the training data set to verify our intuition.

6.4.1 K Metric

The ground truth specifies for a set of articles with the same author name subgroupings or clusters of articles, each cluster for a different individual with that author name. In order to compare this “true” clustering with either the trivial clustering for the undisambiguated data (all papers with the same author name form one group) or with the clustering resulting from an automated disambiguation attempt, we need a measure of the agreement between those clusterings. The accuracy of a clustering with respect to the true clustering, can be quantified in a number of different ways. The metric we found most relevant is the “K metric” used in [31]. Given the true clusters for a name there are two quantities of interest for an empirical clustering: the average cluster purity (ACP) and the average author purity (AAP).

Cluster purity is high when an empirical cluster contains articles mostly by the same individual. But cluster purity does not quantify how fragmented a cluster is. In the extreme case a true cluster may be split into many singleton clusters, each with high cluster purity. Author purity quantifies the correctness of the splits. For a true cluster if all the articles are in the same empirical cluster the author purity is perfect. The K metric combines the cluster and author purities. It is defined as the geometric mean of the average cluster purity and the average author purity.

For a name let there be N articles (N nodes in the article graph constructed by our algorithm) which in reality represent t individuals. Suppose the j^{th} individual, or cluster, contains n_j articles. So $\sum_{j=1}^{t} n_j = N$. Suppose the grouping of the same articles produced by our algorithm has e clusters where the i^{th} cluster has n_i articles. Thus $\sum_{i=1}^{e} n_i = N$. The *average cluster purity*(**ACP**) and the *average* *author purity* (**AAP**) are defined as follows.

$$\mathbf{ACP} = \frac{1}{N}\sum_{i=1}^{e}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_i}$$

$$\mathbf{AAP} = \frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{e}\frac{n_{ij}^2}{n_j}$$

Here n_{ij} is the number of articles that are in true cluster j as well as in empirical cluster i. So $\sum_{i=1}^{e}\sum_{j=1}^{t} n_{ij} = N$.

$$\mathbf{K} = \sqrt{\mathbf{ACP} \times \mathbf{AAP}}$$

The K values for our data are widely distributed. For this reason we have used quantiles in parameter learning and algorithm evaluation, rather than averages to aggregate the K distributions. Further, we have weighted the distribution of K values with the size of the article set for each names since this size is indicative of the importance of disambiguating that name.

6.4.2 Parameter Learning

Our disambiguation algorithm has one parameter, the low redundancy cut-off. Last names with redundancy scores below this threshold are assumed to refer to the same individual. This parameter was learned from the training set of author names without using self-citation information. The result of a series of runs with different low name redundancy cut-off on the training data is shown in fig 6.2. For each cut-off value, last names with raw redundancy less than or equal to it were trivially disambiguated by considering each of them to be one single identity. For last names above the cut-off, co-author overlap was used for disambiguation. A cut-off value of zero meant all names were disambiguated via co-author overlap.

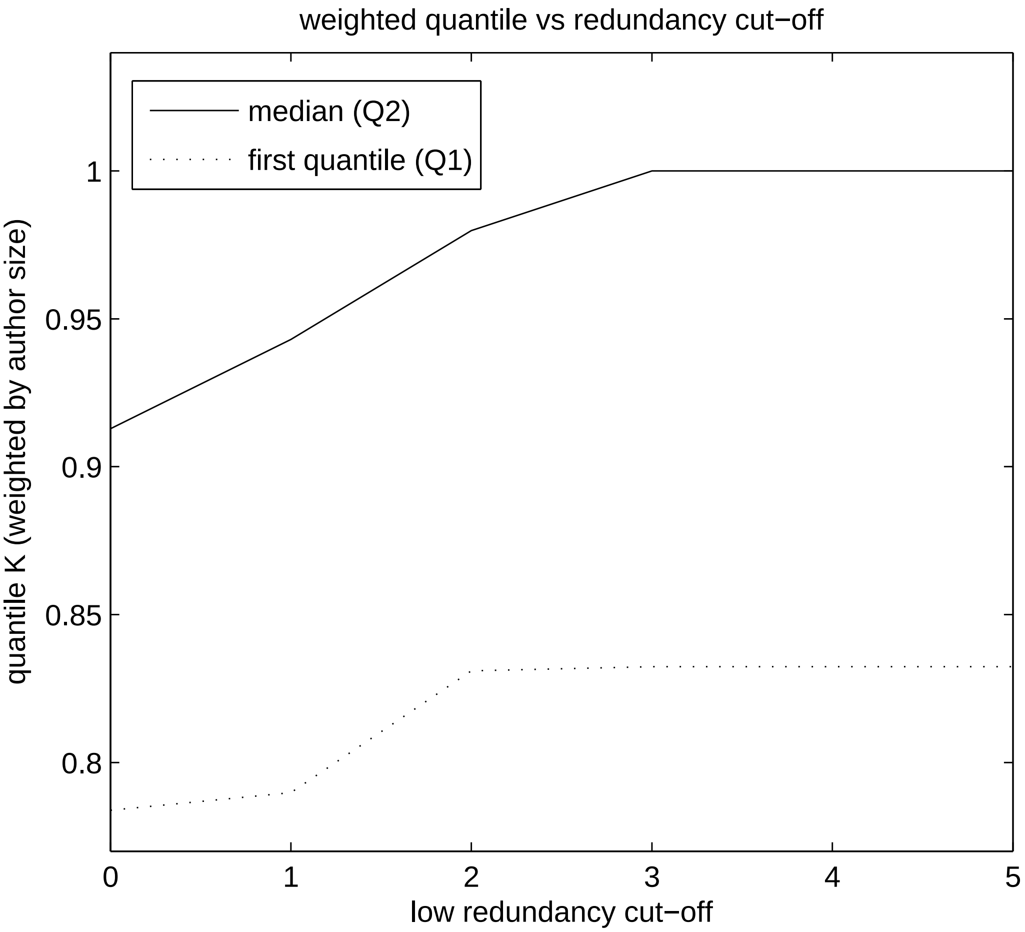


Figure 6.2: Quantiles of weighted K values for the training authors for each low redundancy cut-off.

The weighted median K curve in fig. 6.2 shows 3 to be the best low redundancy cut-off value. For the lower end of the K distribution, 3 is also the optimal cut-off as shown by the weighted first quantile in fig 6.2. This confirms our intuition that a name with such low raw redundancy is better disambiguated by merging all appearances of the name.

	R1	R2	R3	R4	R5	R6	R7
correct	98.0	80.4	51.5	22.5	88.9	72.7	32.1
reduce	0.0	7.8	11.9	16.9	6.9	10.4	28.6
split	1.0	3.9	10.9	11.2	4.2	13.0	17.9.
delete	1.0	7.8	25.7	49.4	0.0	3.9	21.4

Table 6.2: Percentage of role specific distortions of network by homonymy.

6.5 Results

6.5.1 Distortions in Undisambiguated Network

Based on the true identity of authors established for the author names in the groundtruth data set, we can derive estimates for the errors made by not disambiguating author names for homonymy. We distinguish three error types to reflect the different effects correcting them would have on the actual nodes in the network. "Split" means that the ground truth suggests that a node is split into at least two authors with a minimum of two papers each. "Reduce" means a node is to be reduced in size since additional authors were found each of which has no more than one paper, and hence does not survive initial filtering of data when building the network. Finally "delete" means a node is split into separate identities none of which has more than one paper, deleting the node entirely from the network, again due to filtering out of one-paper authors when building the network.

Table 6.2 shows, for the nodes in the ground truth data set, the different kinds of errors that were made by representing all instances of an author name by the same node, as if they all referred to the same individual. Based on these results we obtain the following estimates[3] of the proportion of correct nodes in the giant component of the non-disambiguated network: of the R1 non-hub nodes, almost all, 98% (± 0) correctly represent a single author, followed by the R5 hub nodes with 88.9% (± 1.1) correctly representing a single author. For R2 and R6 nodes the non-disambiguated network represents a large majority of nodes correctly, with 80.4% (± 1.7), and 72.7.% (± 2.9), respectively. Those rates go dramatically down for R3, R7 and R4 nodes, with 51.5.% (± 4.6), 32.1.%(± 12.5), and 22.5% (± 8.0) of nodes correctly representing a single author.

These results confirm our suspicion that the issue of name homonymy causes misrepresentation of individual authors especially for those nodes that determine the inter-cluster connectivity of the clustered network. So, whereas the most numerous node role types in the network, R1 and R2, have small error rates, and the overall estimated error rate across all node role types is about 20%, the error estimate for those nodes of role types that most crucially determine the mesoscopic structure of the collaboration network, those that link between clusters that represent research groups, rise to 68%, and 78% for R7, and R4 nodes, respectively.

6.5.2 Evaluation of Disambiguation Algorithm

Table 6.3 compares for author names in the groundtruth sample the weighted K quantiles before and after disambiguation. Results are reported for the node role specific strata of the sample. The median of weighted K shows notable improvements after disambiguation for node roles R4 and R7, further improvements at the lower 25% quantile level for R3, R4, and R7, and a slight decrease for R6

[3] Approximate error margins given for a 95% confidence interval

	median		**25%**		**minimum**	
	Before	After	Before	After	Before	After
R1	1.00	1.00	1.00	1.00	0.71	0.61
R2	1.00	1.00	1.00	1.00	0.44	**0.68**
R3	0.85	**1.00**	0.65	**0.89**	0.39	**0.56**
R4	0.50	**1.00**	0.40	**0.89**	0.28	**0.58**
R5	1.00	1.00	1.00	1.00	0.62	0.57
R6	1.00	1.00	1.00	0.98	0.41	**0.59**
R7	0.54	**0.93**	0.28	**0.89**	0.20	**0.69**

Table 6.3: Quantiles of weighted K for stratified samples before (i.e. all appearance of a name considered to be a single individual) and after disambiguation (our algorithm). Improvements are shown in bold.

nodes. There are also improvements of the minimum values of the weighted K distributions for all node role types, except R1 and R5. It is clear from table 6.3 that disambiguation is worth the effort for nodes with higher connectivity.

6.6 Conclusion

In this chapter we have shown that author name homonymy is a serious problem for the analysis of large-scale co-author networks. We have derived error estimates from a ground truth sample that is statistically representative of different types of nodes in the network distinguished by their role for the connectivity of the clustered network. Those estimates confirm that a large majority of those nodes that determine the interlinking between co-author clusters in the undisambiguated network include false merges of author identities due to name homonymy.

The disambiguation algorithm presented here deals effectively with those distortions. It rests on a co-author overlap feature that has been found to be very effective in previous work [54]. To increase performance we have added self-citation as a feature, and a cut-off parameter to protect last names of low name commonality from the negative effects of disambiguation. Applying this algorithm produces significant improvements, in particular for those nodes with a critical role in inter-cluster connectivity. The great advantage of this algorithm is its scalability for large data sets and its broad applicability as it uses only a minimal set of data features (co-authors and self-citation).

CHAPTER 7

GRAPH CALCULATION USING MAP-REDUCE

Map-Reduce programming has emerged as a new paradigm of computing useful for many tasks on large scale systems. In this simpler model of computation an application programmer writes Map and Reduce functions that deal with key and value pairs while the underlying system transparently performs the difficult task of parallelization. In this chapter we investigate the applicability of Map-Reduce to graph algorithms, namely Pagerank. We show that reformulation of Pagerank makes it very similar to many graph algorithms.

7.1 Introduction

During the last decade, increasing access to the World Wide Web has made the notion of scale more important than ever before, particularly with respect to system design with desirable features of fault tolerace, massive storage and fast computation. Nowadays the largest of the information systems have hundreds of millions of users from all around the globe. Continued decline in the cost of memory, both primary and secondary, has allowed these systems to store information in sizes unimaginable just 10 years ago. Unfortunately the increase of computational power on a single computer has been slower than this explosive growth of storage. As a result, distributed computing emerged as a necessity.

As the best engineers were working hard at building large scale reliable systems, it became obvious that writing application programs for these systems required non-trivial understanding of the underlying system. In addition, the programmer had the burden of breaking down the task into parts that could be run in parallel. These difficulties deterred rapid application development for large scale systems. The ideal solution to this problem would have been a compiler that could translate code written for a single computer to code for an underlying distributed system such that parallelization of computation and input/output is achieved. However this is an extremely difficult task and remains elusive.

A practical solution of making a distributed system transparent to an application programmer was proposed and implemented in the seminal paper [22] by Google[1] engineers Jeffrey Dean and Sanjay Ghemawat. In the article, authors put forward a simpler programming model called Map-Reduce where each task comprises a Map function and a Reduce function. The input and output to both these functions are <key, value> pairs, possibly from different domains of both keys and values. Map-Reduce programming is quite similar to the functional programming paradigm that has been around for a while.

In [22], application programs were shown to be written as pairs of Map and Reduce functions. Map function is described as the first step where a <key, value> pair is taken as input and <key, value> pairs are emitted as output. The transparent underlying system handles the responsibility of distributing keys to physical processors and aggregating the output of the Map functions to form lists of values for each key. These <key, {value}> pairs are then fed in to the Reduce function which, after performing necessary processing, emits <key, value> pairs to the system. In this restricted programming model of handling <key, value> pairs, the task for an application programmer is to write the necessary Map and Reduce functions.

[1] http://www.google.com/

In this chapter, we discuss how large scale graph computation can be parallelized using Map-Reduce. As a protoypical example, we undertake the task of computing Pageranks for networks by first reformulating the algorithm to a sequence of matrix-vector multiplications and then showing how these products can be performed efficiently. We conclude by commenting on formulations of graph algorithms as such sequences of matrix-vector multiplications.

7.2 Map-Reduce Applications

In [22], it was shown that the Map-Reduce programming model could be used to efficiently perform word frequency counting, distributed grep, URL frequency counting, transposing a web-link graph, inverting an index, and sorting. In subsequent years, large scale systems adopted this paradigm of computing and widely used it for tasks that involved simple parallelization followed by aggregation (see [65] for examples and the new notion of "Map-Reduce design patterns"). One of the most effective uses was reported in [10], where very large scale language models were built for statistical machine translation using Map-Reduce. Ref. [27] builds on the work in [10] and shows how Expectation-Maximization could be performed on such a programming framework. Hadoop[2], an open source implementation, accelerated use of MapReduce for industrial and academic data mining. Database query systems such as Pig[3] and Hive[4] have been built on top of Map-Reduce systems.

Beyond simpler tasks of aggregating a large amount of data, the applicability

[2] http://hadoop.apache.org/

[3] http://pig.apache.org/

[4] http://hive.apache.org/

of Map-Reduce to various tasks have continuously been investigated. In [18], implementation and systematic evaluation of several machine learning algorithms were reported for the first time. Ref. [34] also reports results similar to [22, 18]. Application of Map-Reduce in collaborative filtering was provided in [19] and computation of document similarity was shown in [30]. Cornell University Web Lab[5] project produced a steady stream of experimental results of applying Map-Reduce programming to provide a transparent toolkit for researchers to analyze petabyte sized internet crawls captured once per month by the Internet Archive[6] project. Notable mention is the computation of similarity between Wikipedia entity pairs reported in [3].

7.3 Graph Computation and Pagerank

Although representation of graphs through <key, value> pairs is easy, implementing graph algorithms in Map-Reduce is non-trivial. The main difficulty is that in most algorithms information propagates through edges at each step. Finding connected components and computing single source shortest paths for example would require a number of Map-Reduce steps proportional to the number of edges along the longest path. Pararallelism is utilized at each such step. But in the worst case there is no asymptotic improvement, and in practice Map-Reduce runs slower than a single computer algorithm for pathological cases.

However all hope is not lost. Real world graphs often have small diameters. This puts a bound on the number of Map-Reduce (or propagation) steps that

[5] http://weblab.infosci.cornell.edu/

[6] hrefhttp://www.archive.org/http://www.archive.org/

are needed. Approximate, rather than exact, computation can also put bounds on the number of steps required to produce acceptable results. If we embrace these compromises, then graph computation can be expressed as a sequence of sparse matrix and dense vector multiplication where the dimensions of the matrix can be very large. In this chapter, we present an efficient Map-Reduce implementation of such sequences of matrix-vector multiplication. To demonstrate this method, we have reformulated the problem of computing Pagerank on a large sparse graph. Experimental results for this formulation have been reported in [17].

Pagerank is a network metric used to rank the importance of nodes in a graph. It was originally proposed in [72] and has been widely used for search engine algorithms [11]. The idea is to consider a graph as a Markov model such that the steady state probabilities of nodes is used for ranking. Steady state probabilities can be computed by running random walks on the network where each iteration of the algorithm corresponds to extending the random walk by one additional step. However since networks need not be strongly connected (or "irreducible", in the language of Markov models) sink nodes would hold all the probabilities in the steady state. A realistic fix is to uniformly distribute a fraction of the outgoing probability of each node to every node in the network. This models user behavior of browsing web pages through links from one page to another with a small probability that there would be a random jump to some other website not linked from the current page. The context of web page ranking was appropriate for such adaptation.

7.4 Mathematical Reformulation of Pagerank

Suppose M is the stochastic transition matrix of a directed graph $G = (V, E)$. So M is an $n \times n$ matrix with $n = |V|$ and $\sum_j M_{ij} = 1$ for all i. Suppose $A = M^T$ and $\mathbf{1}_n$ is the vector of all ones. If $0 < \epsilon < 1$ and $\mathbf{p}$ is the vector of pagerank values, then the Pagerank formulae is written as

$$\mathbf{p} = \epsilon A\mathbf{p} + \frac{1-\epsilon}{n}\mathbf{1}_n$$

If we define $\mathbf{q} = \frac{n}{1-\epsilon}\mathbf{p}$ then the formulae becomes

$$\mathbf{q} = \epsilon A\mathbf{q} + \mathbf{1}_n$$

Since $\mathbf{q}$ is proportional to $\mathbf{p}$, it can be used wherever pagerank is used except that $||\mathbf{q}||_1 = \frac{n}{1-\epsilon}$ whereas $||\mathbf{p}||_1 = 1$. The pagerank formulae can be further rewritten as

$$\mathbf{q} = (I_n - \epsilon A)^{-1}\mathbf{1}_n$$

For any vector $||\mathbf{x}||_2 = 1$

$$\underbrace{\mathbf{x}^T A\mathbf{x}}_{\text{scalar}} = (\mathbf{x}^T A\mathbf{x})^T = \mathbf{x}^T A^T\mathbf{x} = \mathbf{x}^T M\mathbf{x} \leq 1$$

So $\mathbf{x}^T \epsilon A\mathbf{x} < 1$ and $I_n - \epsilon A$ is positive definite. Thus $(I_n - \epsilon A)^{-1}$ exists. Expanding the inverse into an infinite series we get

$$\mathbf{q} = \sum_{k=0}^{\infty} \epsilon^k A^k \mathbf{1}_n$$

If we want δ precision in the values of $\mathbf{q}$, then we need to sum up the first N terms where $\epsilon^N \leq \delta$. In other words, $N \geq \frac{log(1/\delta)}{log(1/\epsilon)}$. Thus $N = \lceil\frac{log(1/\delta)}{log(1/\epsilon)}\rceil$ suffices and the formulae finally becomes

$$\mathbf{q} = \sum_{k=0}^{N} \epsilon^k A^k$$

The pagerank algorithm can be succintly written as a sequence of matrix-vector multiplications.

```
$\mathbf{q} = \mathbf{1}_n$, $\mathbf{r} = \mathbf{1}_n$
For $k = 1$ to $N$ do
    $\mathbf{r} = \epsilon A\mathbf{r}$
    $\mathbf{q} = \mathbf{q} + \mathbf{r}$
```

7.5 Map-Reduce Implementation

The sparse matrix M can be represented by tuples $< u, v, p_{uv} >$ where u, v are nodes of the graph and $p_{uv} = \frac{1}{\text{outdegree}(u)}$. And if we maintain tuples $< u, r_u, q_u >$ for every vertex u of the graph with r_u and q_u initially being 1, every iteration of the algorithm readily translates to two Map-Reduce programs.

The core of the iteration is a sparse matrix-vector multiplication which essentially computes summations of partial products of numbers. The first Map-Reduce program computes these products while the second Map-Reduce program sums them up. Both the Maps are identities, i.e., emitting input tuples as output. But in a real implementation, the maps perform a "join" between two kinds of tuples shown in the Map-Reduce programs listed below. The underlined variables are the keys.

1. - Map $\boxed{< \underline{u}, (v, p_{uv}) >, < \underline{u}, (r_u, q_u) >}$ => $\boxed{< \underline{u}, (v, p_{uv}) >, < \underline{u}, (r_u, q_u) >}$
   - Red $\boxed{< \underline{u}, \{(r_u, q_u), \{(v, p_{uv})\}\} >}$ => $\boxed{< \underline{v}, \epsilon r_u p_{uv} >}$

2.
- Map $\boxed{< \underline{v}, \epsilon r_u p_{uv} >, < \underline{v}, (r_v, q_v) >}$ => $\boxed{< \underline{v}, \epsilon r_u p_{uv} >, < \underline{v}, (r_v, q_v) >}$
- Red $\boxed{< \underline{v}, \{(r_v, q_v), \{\epsilon r_u p_{uv}\}\} >}$ => $\boxed{< \underline{v}, (\underbrace{\sum_u \epsilon r_u p_{uv}}_{\text{new } r_v}, \underbrace{q_v + \sum_u \epsilon r_u p_{uv}}_{\text{new } r_v}) >}$

7.6 Improved Implementation

In the above implementation, most work is being done by the two Reduce programs, whereas the two Maps are joining two kinds of tuples. So the obvious question is whether we can join the two tuples only once initially and use both Map and Reduce for computation, so that there is just one Map-Reduce task instead of two. On a real system, Map-Reduce programs take time to set up and reduction in the number of Map-Reduce tasks improves performance significantly.

Instead of keeping two kinds of tuples in the earlier Map-Reduce programs, we can keep tuples of the form $< u, v, r_u, \epsilon r_u p_{uv} >$. We will still need tuples of the form $< u, q_u >$. But joining these tuples is very easy, as shown in the following Map-Reduce program. Here $t_u = \epsilon r_u p_{uv}$ and $t = \sum_u \epsilon r_u p_{uv}$.

- Map $\boxed{< \underline{u}, (v, r_u, t_u) >, < \underline{v}, q_v >}$ => $\boxed{< \underline{v}, q_v >, < \underline{v}, t_u >, < \underline{v}, (w, r_v, \epsilon r_v p_{vw}) >}$
- Red $\boxed{< \underline{v}, \{q_v, \epsilon r_u p_{uv}, (w, r_v, \epsilon r_v p_{vw})\} >}$ => $\boxed{< \underline{v}, q_v + t >, < \underline{v}, (w, t, \frac{\epsilon r_v p_{vw}}{r_v} t >}$

The first tuple on the left hand side in the Map program is output exactly as the last tuple on the right hand side except that the variables u and v are renamed as v and w. The middle tuple on the right hand side is also obtained from the first tuple on the left hand side. The Reduce program sums up the values and adds the sum to the current sum **q**. It also updates the tuples $< u, v, r_u, \epsilon r_u p_{uv} >$

by appropriately dividing by the old value of r_u (r_v in the Reduce program) that is being carried around, and then multiplying by the new value of r_u (r_v in the Reduce program) just computed. Finally the old value is updated to the new value.

The important idea here is that the Map program cleverly splits tuples so that the Reduce program can perform summation for the current iteration and also multiplication to produce values to be summed up in the next iteration. Thus instead of multiplying and then summing up in two steps, we sum up and multiply at the same step. The tuples are maintained so that the output from the Map-Reduce program can be fed back to itself and the pagerank iterations, as we have formulated it, can be run.

The initial values for the tuples of the form $< u,\ v,\ r_u,\ \epsilon\ r_u\ p_{uv} >$ should be $< u,\ v,\ 1,\ \epsilon\ p_{uv} >$ and for the tuples of the form $< v,\ q_v >$ should be $< v,\ 1 >$. One interesting outcome of this improved Map-Reduce implementation is that the constant ϵ does not appear anywhere in the iteration! It appears only during initialization of the tuples.

7.7 Conclusion

In this chapter, we have reformulated the well known Pagerank formulae mathematically so that the computation can be expressed as a sequence of matrix-vector multiplications. Then we have shown how such sequences can be computed efficiently using Map-Reduce programming. Experimental results for this implementation is provided in [17].

Many graph algorithms can be formulated as a sequence of matrix-vector multiplications and thus can be parallelized using the approach presented. For example, if we redefine value multiplication in the matrix setting as the summation of graph edge weights (path weight) and summing up the values as taking the minimum of the set then, with appropriate initial vector, we can compute single source shortest paths by a sequence of matrix-vector multiplications. The number of iterations required for a real world graph with short diameter will be small. And it would be possible to handle large graphs that arise in various information systems.

Part V

Conclusion and Appendices

CHAPTER 8

CONCLUSION

Social interaction on information systems is both interesting to examine and necessary to analyze. In this dissertation, we explored scholarly communication systems as examples of information systems that allow social interaction and presented several methods of its measurement. The algorithms proposed are aimed to be as generalizeable as possible, yet simple enough to accommodate large scale processing that exploits availability of data. At the core of our methods is novel integration of various sources of information, such as metadata, full-text, networks and log data, through applied machine learning and non-parametric statistical analysis. For our experimentation, we have used the arXiv preprint system that has been serving scholarly communities for two decades now.

In chapter 2, we showed that articles in the top few positions of daily listings in several arXiv subject areas on average receive higher downloads during the first several weeks and higher citations years afterwards. Although visiblity has an effect on citations, possibly through a stronger effect on early downloads, self-promotion of better articles is the dominant cause for higher citations. Since both self-promotion and visibility effects are correlated with early and long term downloads, we presented a method of predicting citations by supervised machine learning using downloads and other features extracted from metadata related to articles. Our results show that we can predict citations with significant accuracy.

In chapter 3, we showed that articles positioned near the end of daily listings of certain subject areas in arXiv also receive higher citations and early downloads on average. This reverse-visiblity effect could be due to declining reader's attention as she goes through a daily listing top to bottom, but increased attention near the very end of the list. It could also be due to procrastination by deadline-aware authors submitting right before the daily deadline. The possibility that higher citations are due to geographical effects was thoroughly investigated and eliminated.

In chapter 4, we demonstrated how phrases representing subtopical concepts can be extracted from scholarly text. We characterized phrases by variable length *n*-grams of vocabulary words. Vocabulary selection involved computing the distance between word frequency distributions of the desired corpus and a contrasting corpus, and selecting the words most discriminative in a statistical sense. Phrases were then extracted and ranked using a quantity computed from the citation network of articles. This network measure, **CNLC** (compensated normalized link count), signifies the density of citation links for each phrase and thus the importance of the phrase. We showed both CNLC and a combination of word frequency and CNLC (tf-CNLC) perform better than metrics computed solely from citation network or text. Our evaluation method used human annotated topic keywords and search queries submitted to arXiv's search engine. Log data is easy to collect on information systems and such sources of implicit information was advocated as valuable in algorithm evaluation.

In chapter 5, we aggregated user clicks on articles to understand topic trends in high-energy physics. To balance sparseness and bursts of high activity, coarse binning was employed and hierarchical clustering on binned normalized patterns produced broad overviews of different areas of scholarly interest. Such an overview may be useful not only to individual researchers but also social scientists and policy makers.

In chapter 6, we proposed an algorithm to tackle the problem of name homonymy in the context of network analysis. Our simple algorithm uses co-authorship and self-citation in an unsupervised setting. A novel use of network node roles to sample names allowed stratified evaluation of our disambiguation algorithm and quantification of its effect on network structure.

In chapter 7, adaptation of graph computation to Map-Reduce programming was discussed. As many graph algorithms can be reduced to a sequence of matrix-vector multiplications we provided an efficient Map-Reduce program to compute a sequence of such matrix-vector multiplications. Finally we showed how the Pagerank algorithm can be reformulated and easily implemented.

APPENDIX A

SUPPLEMENTARY MATERIAL

A.1 Power Law Fitting

To fit data to a power law, often the method of maximum likelihood estimation is used to compute the power law exponent, followed by a least squares of a straight line with the computed slope in a log-log plot.

For a power law distribution with $p(x) \propto x^{-\alpha}$, the cumulative distribution $F(X > x) \propto x^{-(\alpha-1)}$, is also a power law. The cumulative distribution $F(X > x)$ is smoother so is customarily used for power law fitting, and is often plotted as a Rank-Frequency (RF) plot [71]. Swapping the axes of an RF plot gives the Zipf plot, which follows a power law behavior with the inverse of the RF exponent.

[26] gives the Zipf plots of citations for the top 10 positions and the remaining positions, binned appropriately. These curves give the cumulative distribution function $F(X > x)$ for different positions, and are useful in comparing two distributions for stochastic dominance. A cumulative distribution $F(X > x)$ is said to *stochastically dominate (first order)* [5] another cumulative distribution $G(X > x)$ iff for all x we have

$$F(X > x) \geq G(X > x) \ .$$

In risk analysis, it is always safer to gamble according to the dominating distribution, since it is expected to produce higher values. If one RF curve is always above another, then there is stochastic dominance, although the statistical significance of the dominance needs to be verified.

In [26], the citation distribution of the top position is found to be higher than the lower positions. The power law exponent of the Zipf plot in [26] is $\beta = 0.48$, in accord with [77]. The power law exponent of the citation distribution is thus $\alpha = 1 + \frac{1}{\beta} = 3.0833$. At this value of α, the mean [71] citation is

$$\langle x \rangle = \frac{\alpha - 1}{\alpha - 2} x_{\min} \, .$$

If there is an upper limit, $x_{\max}$, then the mean becomes

$$\langle x \rangle = \frac{\alpha - 1}{\alpha - 2} x_{\min} \left[1 - \left(\frac{x_{\min}}{x_{\max}} \right)^{\alpha - 2} \right] \, .$$

To restrict to the region where the power law is valid, the small and large rank regions of the Zipf plots are excluded in [26], introducing (a) a normalization bias, and (b) a potential bias of eliminating a large fraction of the data; as we now describe:

• (a) Given two curves, say citations corresponding to position 1, and to positions 10–40, the restriction to the power law region introduces cutoffs $x^1_{\min} > x^{10-40}_{\min}$ and $x^1_{\max} > x^{10-40}_{\max}$, where

$$\log x^1_{\min} - \log x^{10-40}_{\min} = \log x^1_{\max} - \log x^{10-40}_{\max}$$

(since log-log plots of two parallel straight lines are equidistant at the endpoints). This gives

$$\begin{aligned} \frac{x^1_{\min}}{x^1_{\max}} &= \frac{x^{10-40}_{\min}}{x^{10-40}_{\max}} \\ \implies 1 - \left(\frac{x^1_{\min}}{x^1_{\max}} \right)^{\alpha-2} &= 1 - \left(\frac{x^{10-40}_{\min}}{x^{10-40}_{\max}} \right)^{\alpha-2} \\ \implies \frac{\langle x^1 \rangle}{\langle x^{10-40} \rangle} &= \frac{x^1_{\min}}{x^{10-40}_{\min}} \, . \end{aligned}$$

The cut-off in [26] was such that $\frac{x^1_{\min}}{x^{10-40}_{\min}} \approx 2$, so it is not clear whether the factor of 2 advantage in the average was due to the cut-off having given $\langle x^1 \rangle$ the benefit of higher $x_{\min}$.

- (b) Our analysis of the same data gives a median citation for position 1 of 10, and for positions 10–40 of 4. A large lower cutoff will thus ignore a large fraction of the data. Ref. [26] used $x^1_{\min} \approx 50$, whereas the 75th percentile of the citations for position 1 is 22. This means at least $\frac{3}{4}$ of the data was ignored to compute the aggregate values.

A.2 Statistical Significance

To test the statistical significance of the difference in median citations, we use the Mann-Whitney U (MWU) test, with the null hypothesis that the medians are equal, and the two-sided alternative that the medians are not equal, at 1% significance level. Table A.1 shows that for astro-ph the medians of the top 5 positions are significantly different from the medians of the positions 10 and beyond.

Position from	Position onwards
1	2
2	5
3	5
4	7
5	11
6	11

Table A.1: *Mann-Whitney U test for astro-ph*. Left column is the position whose median we are assessing for significant difference (1% significance level) with a two-sided alternative (either median the greater). The right column is the position whose median (and that of positions beyond) is significantly different from the corresponding position on the left column. For example the median number of citations for position 2 is greater than that of positions 5 and beyond, at 1% significance level.

A significant difference in median does not necessarily mean a distribution

Position from	Position onwards
1	4
2	5
3	6
4	7
5	11

Table A.2: *Kolmogorov-Smirnov test for astro-ph.* Left column is the position whose distribution we are assessing for stochastic domination (1% significance level) with a one-sided alternative. The right column is the position whose distribution (and the positions beyond) is stochastically dominated by the corresponding position on the left column. For example the median number of citations for position 2 is greater than that of positions 5 and beyond, at all levels, at 1% significance level.

is better at all levels. To test stochastic domination, we used the Kolmogorov-Smirnov (KS) test with the null hypothesis that the two distributions are the same, and the one-sided alternative that the first distribution dominates the second, at 1% significance level. Table A.2 shows that for astro-ph the first 5 positions are indeed better than all other positions, for all values.

A.3 SVM Regression

SVM regression [84] is different from the standard regression task in two ways. Firstly, SVM uses the ε-insensitive loss function where for individual sample points only an error of greater than ε counts as "error", and the total error is the sum of the samplewise errors. Secondly, the minimization function is a combination of the ε-insensitive loss function as well as the squared norm of the vector of regression coefficients. The tradeoff between this norm and the loss function is controlled by a parameter C. The algorithm takes both ε and C as parameters, and setting small ε and large C gives a form of least squares result. SVM regression uses state of the art constrained optimization techniques to find a solution. Its real power, however, is the ease with which nonlinearity can be incorporated by higher order kernels. The efficiency and accuracy of this approach has already been established firmly in the realm of machine learning through numerous principled applications.

To explore the predictive capacity of readership and other features, we treated it as a standard supervised prediction task in machine learning. Some past attempts to correlate citations with article and author features [87, 2, 93] used samples that were several orders of magnitude smaller and hence allowed manual extraction of features.[1] In such a setting regression is used for the entire dataset and the total error is reported. A potential problem with this approach is that it may simply validate the regression model used, rather than result in learning and prediction. Use of the full dataset may also be vulnerable to overfitting through extraneous features. In machine learning, the standard approach is to cross-validate through random training and test splits of the data, and report the average accuracy on the test sets. This puts less emphasis on human verification of the model being learned, especially when higher order kernels are used.

[1] While manual extraction of features is not as feasible on the larger datasets currently in use, modern text-mining tools together with the increased availability of the full-texts in digital form should ultimately permit automated extraction of a comparable set of features.

BIBLIOGRAPHY

[1] Eytan Adar, Jaime Teevan, and Susan T. Dumais. Large scale analysis of web revisitation patterns. In *ACM Special Interest Group on Computer-Human Interaction*, 2008.

[2] S. Baldi. Normative versus social constructivist processes in the allocation of citaions: A network-analytic model. *American Sociological Review*, 63:829–846, 1998.

[3] Jacob Bank and Benjamin Cole. Calculating the jaccard similarity coefficient with map reduce for entity pairs in wikipedia. Technical report, Cornell University, 2008.

[4] K. Barker and N. Cornacchia. Using noun phrase heads to extract document keyphrases. In *13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, 2000.

[5] V. S. Bawa. Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, 2:95–121, 1975.

[6] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2000.

[7] Indrajit Bhattacharya and Lise Getoor. A latent dirichlet model for unsupervised entity resolution. In *Society for Industrial and Applied Mathematics Conference on Data Mining*, 2006.

[8] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:1:17–35, 2007.

[9] D. Blei and J. Lafferty. Topic models. *Text Mining: Theory and Applications*, 2009.

[10] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[11] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *International World Wide Web Conference*, 1998.

[12] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *ACM Special Interest Group on Information Retrieval*, 2007.

[13] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citatin impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

[14] B. Cao, J. Sun, E. Xiang, D. Hu, Q. Yang, and Z. Chen. Pqc: Personal query classification. In *ACM Conference on Information and Knowledge Management*, 2009.

[15] J. Chang and D. Blei. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[16] J. Chang, J. Boyd-Graber, and D. Blei. Connections between the lines: Augmenting social networks with text. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2009.

[17] V. Chokkapu and A. Haque. Pagerank calculation using map reduce. Technical report, Cornell University Web Lab, 2008.

[18] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In *Neural Information Processing Systems Conference*, 2006.

[19] Abhinandan Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: Scalable online collaborative filtering. In *International World Wide Web Conference*, 2007.

[20] P. M. Davis and M. J. Fromerth. Does the arxiv lead to higher citations and reduced publisher downloads for mathematics articles. *Journal of the American Society for Information Science and Technology*, 71:203–215, 2007.

[21] P. M. Davis, B. V. Lewenstein, D. H. Simon, J. G. Booth, and M. J. L. Connolly. Open access publishing, article downloads, and citations: Randomised controlled trial. *British Medical Journal*, 337:a568, 2008.

[22] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Symposium on Operating Systems Design and Implementation*, 2004.

[23] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[24] E. Diemart and G. Vandelle. Unsupervised query categorization using automatically-built concept graphs. In *International World Wide Web Conference*, 2009.

[25] J. P. Dietrich. Disentangling visibility and self-promotion bias in the arxiv:astro-ph positional citation effect. *Publications of the Astronomical Society of the Pacific*, 120:801–804, 2008.

[26] J. P. Dietrich. The importance of being first: Position dependent citation rates on arxiv:astro-ph. *Publications of the Astronomical Society of the Pacific*, 120:224–228, 2008.

[27] Christopher Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. Fast, easy and cheap: Construction of statistical machine translation models with mapreduce. In *Third Workshop on Statistical Machine Translation*, 2008.

[28] H. Ebbinghaus. On memory: A contribution to experimental psychology. *New York: Teachers College, Columbia University*, 1913 (Original work published 1885).

[29] Editor. Deciphering citation statistics. *Nature Neuroscience*, 11(619), 2008.

[30] Tamer Elsayed, Jimmy Lin, and Douglas W. Oard. Pairwise document similarity in large collections with mapreduce. In *Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 2008.

[31] Anderson Ferreira, Adriano Veloso, Marcos Goncalves, and Alberto Laender. Effective self-training author name disambiguation in scholarly digital libraries. In *Joint Conference on Digital Libraries*, 2010.

[32] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences*, 103:12648–12689, 2006.

[33] J. D. Gibbons. *Nonparametric Methods for Quantitative Analysis*. American Science Press, 1997.

[34] Dan Gillick, Arlo Faria, and John Denero. Mapreduce: Distributed computing for machine learning, 2006.

[35] P. Ginsparg. Next-generation implications of open access. *CTWatch Quarterly*, 3, 2007.

[36] Paul Ginsparg. First steps towards electronic research communication. *Computers in Physics*, 8(4):390–396, 1994.

[37] Paul Ginsparg. Winners and losers in the global research. In *Electronic Publishing in Science*, UNESCO Headquarter, Paris, 1996.

[38] Paul Ginsparg. Creating a global knowledge network. In *Electronic Publishing in Science II*, UNESCO Heqadquater, Paris, 2001.

[39] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *ACM Special Interest Group on Information Retrieval*, 2004.

[40] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics*, 3(1):63–69, 2007.

[41] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Joint Conference on Digital Libraries*, 2004.

[42] Hui Han, Hongyuan Zha, and Lee Giles. Name disambiguation in author citations using k-way spectral clustering method. In *Joint Conference on Digital Libraries*, 2005.

[43] A. Haque and P. Ginsparg. Positional effects on citation and readership in arxiv. *Journal of the American Society for Information Science and Technology*, 60(11):2203–2218, 2009.

[44] A. Haque and P. Ginsparg. Last but not least: Additional positional effects on citation and readership in arxiv. *Journal of the American Society for Information Science and Technology*, 61(12):2381–2388, 2010.

[45] A. Haque and P. Ginsparg. Phrases as subtopical concepts in scholarly text. In *Joint Conference on Digital Libraries*, 2011.

[46] Edwin Henneken, Micahel Kurtz, Alberto Accomazzi, Carolyn Grant, Donna Thompson, Elizabeth Bohlen, Giovanni Di Milia, Jay Luker, and Stephen Murray. Finding your literature match – a recommender system. In *Future Professional Communication in Astronomy II*, 2010.

[47] T. Hofmann. Probabilisitic latent semantic indexing. In *ACM Special Interest Group on Information Retrieval*, 1999.

[48] Jian Huang, Seyda Ertekin, and C. Lee Giles. Efficient name disambiguation for large-scale databases. In *Principles and Practice of Knowledge Discovery in Databases*, 2006.

[49] X. Jiang, Y. Hu, and H. Li. A ranking approach to keyphrase extraction. In *Microsoft Research Technical Report*, 2009.

[50] Y. Jo, C. Lagoze, and C. Giles. Detecting reseach topics via the correlation between graphs and texts. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2007.

[51] T. Joachims. Making large-scale svm learning practical. In *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, 1999.

[52] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, 25(2), April 2007.

[53] T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *IEEE Computer*, 40(8):34–40, August 2007.

[54] In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing and Management*, 45:84–97, 2009.

[55] Jon Kleinberg. Bursty and hierarchical structure in streams. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2002.

[56] N. Kumar and Srinathan K. Automatic keyphrase extraction from scientific documents using n-gram filtration technique. In *ACM Symposium on Document Engineering*, 2008.

[57] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C Grant, M. Demleitner, E. Henneken, and S. Murray. The effect of use and access on citations. *Information Processing and Management*, 41:1395–1402, 2005.

[58] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. Grant, M. Demleitner, S. Murray, N. Martimbeau, and B. Elwell. The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56:111–128, 2005.

[59] Michael Kurtz, Alberto Accomazzi, Edwin Henneken, Giovanni Di Milia, and Carolyn Grant. Using multipartite graphs for recommendation and discovery. In *Astronomical Society of the Pacific Conference Series*, 2009.

[60] Michael Kurtz, Guenther Eichhorn, Alberto Accomazzi, Carolyn Grant, Markus Demleitner, Stephen Murray, Nathalie Martimbeau, and Barbara Elwell. The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2):111–128, 2005.

[61] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2009.

[62] X. Li, Y. Wang, and A. Acero. Learning query intent from regularized click graphs. In *ACM Special Interest Group on Information Retrieval*, 2008.

[63] X. Li, Y. Wang, D. Shen, and A. Acero. Learning with click graph for query intent classification. *ACM Transactions on Information Systems*, 28(3), 2010.

[64] Y. Li, Z. Zheng, and H. Dai. Kdd cup-2005 report: Facing a great challenge. *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter*, 7(2), 2005.

[65] Jimmy Lin and Chris Dyer. *Data-Intensive Text Processing with MapReduce*. Morgan and Claypool Publishers, 2010.

[66] Y. Lui. Extraction of significant phrases from text. *International Journal of Computer Science*, 2(2):101–109, 2007.

[67] Bradley Malin. Unsupervised name disambiguation via social network similarity. In *Society for Industrial and Applied Mathematics Industrial Conference on Data Mining*, 2005.

[68] Duncan M. McRae-Spencer and Nigel R. Shadbolt. Also by the same author: Aktiveauthor, a citation graph approach to name disambiguation. In *Joint Conference on Digital Libraries*, 2006.

[69] H. F. Moed. The effect of ”open access” upon citation impact: An analysis of arxiv’s condensed matter section. *Journal of the American Society for Information Science and Technology*, 58:2047–2054, 2007.

[70] Henk Moed. Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10):1088–1097, 2005.

[71] M. E. J. Newman. Power laws, pareto distributions and zipf’s law. *Contemporary Physics*, 46:323–351, 2005.

[72] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[73] T. V. Perneger. Relation between online ”hit counts” and subsequent citations: Prospective study of research papers in the bmj. *British Medical Journal*, 329:546–547, 2004.

[74] D. P. Phillips, E. J. Kanter, B. Bednarczyk, and P. L. Tastad. Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine*, 325(16):1180–1183, 1991.

[75] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2005.

[76] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *ACM Conference on Information and Knowledge Management*, 2008.

[77] S. Redner. How popular is your paper? an empirical study of citation distribution. *European Physical Journal*, B4:131–134, 1998.

[78] Sidney Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58(6):49–54, 2005.

[79] Martin Rosvall and Carl T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104:7327–7331, 2007.

[80] M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854–856, 2006.

[81] Benyah Shaparenko, Rich Caruana, Johannes Gehrke, and Thorsten Joachims. Identifying temporal patterns and key players in document collections. In *IEEE International Conference on Data Mining: Temporal Data Mining*, 2005.

[82] D. Shen, J. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *ACM Special Interest Group on Information Retrieval*, 2006.

[83] Neil Smallheiser and Vetle Torvik. Author name disambiguation. *Annual Review of Information Science and Technology*, 43:287–313, 2009.

[84] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[85] F. Song and W. Croft. A general language model for information retrieval. In *ACM Conference on Information and Knowledge Management*, 1999.

[86] Yang Song, Jian Huang, Issac Councill, Jia Li, and C. Lee Giles. Efficient topic-based unsupervised name disambiguation. In *Joint Conference on Digital Libraries*, 2007.

[87] J. A. Stewart. Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces*, 62:166–189, 1983.

[88] Andreas Strotmann, Dangzhi Zhao, and Tania Bubela. Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology*, 46:1–20, 2010.

[89] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[90] L. Tang and J. P. Walsh. Bibliometric fingerprints: Name disambiguation

based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3):763–784, 2010.

[91] Vetle Torvik, Marc Weeber, Don Swanson, and Neil Smallheiser. A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, 56(2):140–158, 2005.

[92] P. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336, 2000.

[93] H. P. van Dalen and K. Henkens. What makes a scientific article influential? the case of demographers. *Scientometrics*, 50:455–482, 2001.

[94] T. Velden, A. Haque, and C. Lagoze. A new approach to analyzing patterns of collaboration in co-authorship networks: Mesoscopic analysis and interpretation. *Scientometrics*, 85(1):219–242, 2010.

[95] T. Velden, A. Haque, and C. Lagoze. Resolving author name homonymy to improve resolution of structures in co-author networks. In *Joint Conference on Digital Libraries*, 2011.

[96] K. Wang, T. Walker, and Z. Zheng. Pskip: Estimating relevance ranking quality from web search clickthrough data. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2009.

[97] J. Wen, J. Nie, and H. Zhang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1), 2002.

[98] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-manning. Kea: Practical automatic keyphrase extraction. In *ACM International Conference on Digital Libraries*, 1999.

[99] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *ACM Special Interest Group on Information Retrieval*, 2003.