A Lead-Lag Analysis of the Topic Evolution Patterns for Preprints and Publications

Beibei Hu and Xianlei Dong

Beijing University of Technology

Chenwei Zhang, Timothy D. Bowman and Ying Ding

Indiana University

ErjiaYan

Drexel University

Staša Milojević and Chaoqun Ni

Indiana University

Vincent Larivière

Université de Montréal

Abstract

This paper applied LDA and regression analysis to conduct a lead-lag analysis to identify different topic evolution patterns between preprints and papers from arXiv and Web of Science (WoS) in astrophysics over the last twenty years (1992-2011). Fifty topics in arXiv and WoS were generated using an LDA algorithm and then regression models were used to explain four types of topic growth patterns. Based on the slopes of the fitted equation curves, the paper redefines the topic trends and popularity. Results show that arXiv and WoS share similar topics in a given domain, but differ in evolution trends. Topics in WoS lose their popularity much earlier and their durations of popularity are shorter than those in arXiv. This work demonstrates that open access preprints have stronger growth tendency as compared to traditional printed publications.

Keywords: lead-lag, topic evolution pattern, preprint, publication

A Lead-Lag Analysis of the Topic Evolution Patterns for Preprints and Publications

The “publish or perish” mantra reflects the fierce competition in scholarly communication. Being the first person to claim new ideas, new methods, or new discoveries is critical in science. Scholarly communication, especially through formal channels, plays a crucial role in this process. However, over the last decade the Internet and various Web 2.0 technologies have had an impact on the more traditional scholarly communication network of journals and conferences by enabling faster and broader dissemination through a variety of communicative channels. Researchers are sharing their initial and innovative thoughts through personal blogs, tweets, Facebook comments, online repositories, and online discussion groups. These ideas and drafts can be downloaded, discussed, tweeted/retweeted, forwarded, commented, and tagged through a variety of channels including arXiv, Twitter, Mendely, Citeseer, and CiteULike, to name just a few. This informal scholarly communication significantly speeds up the process of knowledge dissemination.

arXiv is an online repository of digital preprints from a number of fields – most notably from physics, mathematics, and computer science. Since its creation by Paul Ginsparg in 1991, arXiv has become central to the diffusion of research in these fields. arXiv is currently one of the largest open access self-archiving systems and hosts over 0.9 million digital preprints in science covering physics, mathematics, computer science, quantitative biology, statistics, and quantitative finance. This online archiving system, with the policy of allowing every author to submit his or her research output, offers the ideal platform to swiftly propagate knowledge. Before manuscripts enter lengthy peer-review processes that can take anywhere from 3 months to 1.5 years (depending on different journals or disciplines), they can be read, criticized, or even cited by other scholars. In addition, arXiv has become one of the major open access venues for an ever-growing number of researchers who want to reach a wider audience, but who do not have the means to pay extremely high open access fees to journals. Thus, in its role as a digital repository, arXiv allows for the dissemination of works in various stages of their life-cycle: from true preprints to post-prints.

Quite a bit of interest has been generated regarding the ways in which arXiv has potentially accelerated knowledge transfer and changed scholarly communication. While these works are useful, few studies have analyzed the difference between topic evolution in informal scholarly communication (i.e., preprints) and formal scholarly communication (i.e., publications). By analyzing the topic evolution (e.g., topic popularity and duration of topic popularity) patterns in formal and informal scholarly communication, this work can help us better understand whether informal scholarly communication (e.g. arXiv) can stimulate fast knowledge transfer. In this paper we applied LDA and regression analysis to conduct a lead-lag analysis and to identify different topic evolution patterns for preprints and papers in astrophysics for the last twenty years (1992-2011).

Literature Review

Preprint Analysis

Shuai, Pepe and Bollen (2012) analyzed the scientific community's responses to 4,606 preprints submitted to arXiv using the number of downloads, mentions on Twitter, and citations in scholarly articles. The authors studied the delay and time span of article downloads and Twitter mentions to understand the temporal difference. Through regression and correlation tests they found that Twitter mentions and arXiv downloads follow distinct temporal patterns, with Twitter mentions having shorter delays and narrower time spans.

A more recent large-scale analysis of the relationship between pre-prints and papers in terms of coverage and citation was investigated by Lariviere et al. (2014). This studied examined the overlap between papers found on arXiv and WoS (and vice versa), the elapsed time between submission of a preprint and publication of the matching journal article, and aging effects in citations for both genres. The study found that few disciplines are widely represented on arXiv, and even for those that are represented there remains a large share of publications in WoS that are not found on arXiv. Furthermore, the study found that the arXiv versions have lower citation rates than published papers and that citations to these versions decline precipitously following publication in the journal of record.

Temporal Pattern Analysis

Researchers have performed temporal analysis on a variety of corpora, such as news and email datasets, and have refined methods to identify topics in corpora using the combined approach of content analysis and time-series analysis (Chatfield, 2003). By analyzing when documents were written and by whom, Shaparenko, Caruana, Gehrke and Joachims (2005) developed a method to identify the most influential documents and authors in a collection. The authors applied k-means clustering to extract clusters of words from documents and identified the clusters using five significant words appearing nearest to the cluster centroid. They plotted the changes in term frequency of these five words along with different time durations and developed a lead/lag index based on the assumption that if one document spawns a great deal of follow-up work that uses similar vocabulary, then that document is very influential. The lead/lag index measures whether a document is a leader or a follower. The authors compared their results with citation counts and found that they were successful in identifying papers that contained new and influential ideas; in some cases they were able to identify works that were not identified by citation analysis alone.

Swan and Jensen (2000) developed TimeMines using robust, flexible techniques to determine significant keywords in documents and to judge their temporal significance in the context of a corpus. The authors assumed that a simple statistical ranking of term occurrence and co-occurrence can identify and group relevant documents into coherent time-dependent stories. Their approach determined whether the possibility of seeing a word/phrase/named entity varied during different time slots. They tested their approach on a corpus of CNN news documents and identified temporal topics.

Another statistical method labeled Temporal Text Mining (TTM) was introduced by Mei and Zhai (2005) and was used to identify temporal patterns through the discovery of latent topics from text, to construct an evolution graph of topics, and to analyze life cycles of topics. The authors constructed a topic evolution graph by calculating the KL-divergence of a word vector containing two pair of topics from different time spans. They applied their method to a document sets containing news about tsunami events and abstracts from the Knowledge Discovery and Data Mining (KDD) conference proceedings and were able to summarize the complete evolutionary theme patterns in a given text stream. Their evolution graph can reveal how topics change over time and how one topic in one time period has influenced other topics in later periods. They used the Hidden Markov Model (HMM) to analyze the life cycle of each topic and discovered globally interesting topics and computed the strength of a topic in each time period.

This analysis allowed the authors to identify the strength variation trends of various topics and to compare the relative strengths of different topics over time.

Kleinberg (2003) developed a method to robustly and efficiently identify bursts that are characterized by topics that appear, grow intensely, and then fade away. A burst, described as containing one or more sub-intervals signifying a dying and rebirth, can be represented as a tree in order to capture the hierarchical structures that are implicit in a corpus. The author performed an analysis of his own emails relating to proposal writing and identified inferred hierarchical structures that clearly identified an intent letter phase, pre-proposal submission phase, and full-proposal submission phase.

Shi, Nallapati, Leskovec, McFarland and Jurafsky (2010), on the other hand, conducted topical lead-lag analysis on papers and funding proposals to study whether research grants come before publications or vice versa. They proposed a general method for lead-lag estimation based on the LDA (Latent Dirichlet Allocation) model and time series analysis to determine topics discussed across time in 20,000 grant proposal abstracts and half a million computer science research paper abstracts. They found that the lead-lag of research papers, with respect to research grants, is topic specific. The authors also found that Security and Cryptography research papers lead grant proposals by two years, while grants related to the topic of Neural Network lead research papers by three years.

Topic Analysis

The goal of topic modeling algorithms is to capture topics from a corpus automatically by using the observed words in documents to infer the hidden topic structure (e.g., document topic distribution, and word topic distribution). The number of topics is usually decided by perplexity, which can be heuristically set ranging from 20 to 300 (Blei, 2012). The inference mechanics in topic models are independent of languages and contents; they capture the statistical structure of language used to represent thematic content. LDA approximates its posterior distribution by using inference (e.g., Gibbs sampling) or optimization (e.g., variational methods) (Asuncion, Welling, Smyth & Teh, 2009). Blei and Lafferty (2006) proposed a dynamic topic model to capture the evolution of topics in a sequentially organized set of documents based on an assumption that each year's articles arise from a set of topics that evolved from last year's topics. They extended classical state space models to specify a statistical model of topic evolution, and developed efficient approximate posterior inference techniques to determine the evolving topics from a sequential collection of documents.

In another work, Mann, Mimno and McCallum (2006) applied topic modeling methods to 300,000 computer science publications to provide topic based impact analysis. They extended journal impact factor measures to topics and introduced three topic impact measures: topical diversity (i.e., ranking papers based on citations from different topics), topical transfer (i.e., ranking papers based on citations from outside of their own topics), and topical precedence (ranking papers based on whether they are among the first to create a topic). They developed Topical N-Grams LDA using phrases rather than words to represent topics. Gerrish and Blei (2010) proposed the document influence model (DIM) based on the dynamic LDA model to identify influential articles without using citations. Their hypothesis was that the influence of an article on the future is corroborated by how the language of its field changes subsequent to its publication. In other words, an article with words that contribute to the change in word frequencies will have a higher influence score. They applied their model on three large corpora and their influence measurement significantly correlates with the article's citation counts.

Liu, Zhang and Guo (2012) applied Labelled LDA on full-text citation analysis to enhance traditional bibliometric analysis. Ding (2011b) combined topic modeling and path-finding algorithms to study scientific collaboration and endorsement in the field of information retrieval. The results show that productive authors tend to directly coauthor with, and closely cite, colleagues sharing the same research topics, but they do not generally collaborate directly with colleagues working on different research topics. Ding (2011c) proposed topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. The author applied the Author-Conference Topic (ACT) model to extract topic distribution for individual authors and conferences, and added this as the weighted vector to the PageRank algorithm. The results demonstrated that this method is able to identify representative authors with different topics at different time spans.

Later on, Ding (2011a) applied author topic model to detect communities of authors and compared this with traditional community detection methods (that are usually topology-based graph partitions) on coauthor networks. The results showed that communities detected by the topology-based community detection approach tend to contain different topics within each community, and communities detected by the author topic model tend to contain topologically diverse sub-communities within each community. Natale, Fiore and Hofherr (2012) examined the aquaculture literature using bibliometrics and computational semantic methods including latent semantic analysis, topic modeling, and co-citation analysis to identify main themes and trends. Song, Kim, Zhang, Ding and Chambers (2014) adopted Dirichlet Multinomial Regression (DMR)-based topic modeling technique to analyze the overall trends of bioinformatics during the periods between 2003 and 2011 and found that the field of bioinformatics had undergone a significant shift to co-evolve with other biomedical disciplines.

LDA can be applied in any field where texts are the main data format, but there are challenges with using LDA. First, labeling topics can be done in different ways (Mei, Shen & Zhai, 2007). LDA typically uses the top ranked keywords with high probabilities for each topic to label that topic, but such labels can be hard to interpret because they are sometimes contradictory. For example, because LDA uses soft clustering one keyword can appear in more than one topic and some topics can have very similar labels. The question that arises is how to provide a meaningful label for each topic automatically. Evaluating LDA remains another major challenge (Chang, Boyd-Graber, Gerrish, Wang & Blei, 2009) because LDA is an unsupervised probabilistic model and the generated latent topics are not necessarily semantically meaningful. LDA assumes that each document can be described as a set of latent topics, which are multi-nominal distributions of words. Chang et al. (2009) found that models which achieve better perplexity often generate less interpretable latent topics. By using Amazon Mechanic Turk, the authors found that humans appreciate the semantic coherence of topics. They recommended incorporating human judgments into the model-fitting process because it can increase the thematic meanings of topics. Evaluation of LDA is highly subjective to different applications.

Studies in temporal pattern analysis focused on developing methods or algorithms to identify topics in different corpora by considering their content and temporal features, while researches in topic analysis extended or revised the current topic modeling algorithms to capture the dynamic changes of topics. Few studies have compared topic evolution of research articles from the same domain but that were published in different (i.e., formal or informal) channels, especially popularity of topics and duration of popularity of topics between these formal and informal channels.

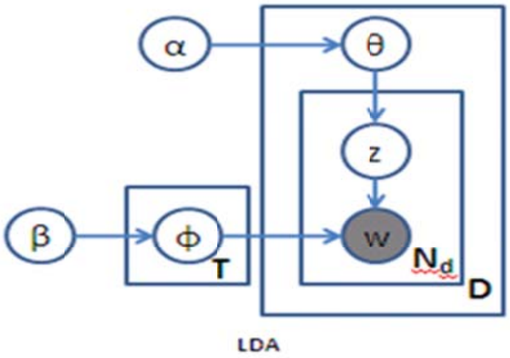
Method

Data

We used two major sources of data: arXiv and Web of Science (WoS). Data were crawled from arXiv using the astrophysics category. In arXiv the astrophysics category includes the following six subfields: cosmology and extragalactic astrophysics, earth and planetary astrophysics, galaxy astrophysics, high energy astrophysical phenomena, instrumentation and methods for astrophysics, and solar and stellar astrophysics. arXiv started to host astrophysics preprints in April 1992. We collected 117,913 astrophysics preprints from 1992 to 2011. For each preprint, year, title, author, abstract, category and DOI were collected. To distinguish the arXiv preprints from postprints, we searched these preprints in WoS and removed those which were published in arXiv later than WoS. In WoS, astrophysics was listed as one of the WoS subject categories. All papers in different document types were collected from this subject category for the period of 1992-2011, resulting in a total of 166,191 research articles total. For each document from WoS, year, title, and author information were collected.

Latent Dirichlet Allocation Modeling

We first used Latent Dirichlet Allocation (LDA) to capture the topical features of nodes by postulating a latent structure for a set of topics linking words and documents. The LDA method has been reliable for detecting multi-nominal word distribution of topics (Blei, Ng & Jordan, 2003). After the success of the LDA model, the basic model has been extended to various levels. The Author-Topic model proposed by Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) depicts the content of documents and the interests of authors simultaneously. Later, Tang, Jin and Zhang (2008) extended LDA to reveal the topic distribution of authors, conferences, and citations concurrently. LDA has been applied in scholarly communication to identify the topic distribution in dynamic research communities (Yan, Ding, Milojević & Sugimoto, 2012), to analyze disciplinary development using domain specific dissertations (Sugimoto, Li, Russell, Finlay and Ding, 2011), to study scientific collaboration and endorsement at the topic level (Ding, 2011c), and to calculate topic-based PageRank (Ding, 2011a).



Notations	Meaning
W	word
Z	topic
N_D	the number of words in the entire collection of documents
A	hyperparameter for generating Θ from Dirichlet Distribution
B	hyperparameter for generating φ from Dirichlet Distribution
Θ	a multinomial distribution over topics
Φ	a multinomial distribution over words
D	collection of documents
T	collection of topics

Figure 1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) provides a probabilistic model for the latent topic layer (Blei et al., 2003) (See Figure 1 for the graphical model representation of LDA). For each document *d*, a multinomial distribution θ_d over topics is sampled from a Dirichlet distribution with parameter α. For each wordw_{di}, a topic z_{di} is chosen from a topic-specific multinomial distribution $\phi_{z_{di}}$ sampled from a Dirichlet distribution with parameter β. The probability of generating a word *w* from a document *d* is:

$$P(w|d,\theta,\phi) = \sum_{z\in T} P(w|z,\phi_z)P(z|d,\theta_d)$$

Therefore, the likelihood of a document collection *D* is defined as:

$$P(Z,W|\Theta,\Phi) = \prod_{d\in D}\prod_{z\in T}\theta_{dz}^{n_{dz}} \times \prod_{z\in T}\prod_{v\in V}\phi_{zv}^{n_{zv}}$$

Where n_{dz} is the number of times that a topic z has been associated with a document d, and n_{zv} is the number of times that a word w_v has been generated by a topic z. The model can be explained as: to write a paper, an author first decides topics and then uses words that have a high probability of being associated with these topics to write the article.

LDA was chosen because it clusters words, documents, authors and other related entities based on latent topics. It provides a mathematical grounding for latent topics so that it is not susceptible to severe overfitting, in contrast to other methods (e.g., Probabilistic Latent Semantic Indexing) that are more susceptible (Blei, Ng, & Jordan, 2003). Results from LDA are easy to interpret and do not require expert judgment to label clusters which is a norm for classic co-occurrence-based methods (e.g., co-citation or co-word analysis). Furthermore, LDA provides an opportunity to examine the research landscape of domains at a more granular level. This level of granularity is typically unattainable through co-occurrence-based methods because of the complexity of densely connected co-occurrence relations. Limitations of the LDA model are largely attributed to its assumptions (Blei, 2012). For instance, the model does not consider the orders of the words in a document and the orders of documents in a data set, and assumes that the number of topics is known and fixed. In reality, however, text corpora may not fit precisely into these assumptions and the performance may be compromised. Another limitation of LDA is that the number of topics to be extracted should be decided beforehand, which is usually based on perplexity. Labeling and judging the quality of topics can only be empirically evaluated. Solutions to relax these assumptions have been provided by Blei (2012).

The Stanford Topic Modeling Toolbox (Stanford TMT: http://nlp.stanford.edu/downloads/tmt/tmt-0.4/) was used to perform the topic modeling. Stanford TMT has a built-in feature to filter out uninformative words from the text corpus. The following data preprocessing was implemented: (1) title words that have less than three letters were removed from titles; (2) 30 most frequently occurred words were removed from titles; (3) those words that occurred in less than three publications were removed; (4) publications whose titles have less than three words were removed from the data set; and finally, (5) the number of topics was set as 50.

Regression Modeling

The dynamic patterns of lead-lag are analyzed by regression modeling. Time *T* was treated as the independent variable, while the number of publications in arXiv and WoS were treated as the dependent variables. Thus, 1992 was counted as the first year (i.e. t=1), 1993 as the second year (i.e. t=2), and so on.

Weierstrass approximation theorem is widely used to approximate the curve of any continuous function. It is defined that for any given closed interval [a,b], there must be a polynomial function which can be uniformly approximated (Stone, 1948). Uniform approximation is equal to uniform convergence that converges independent of x. For example, the number of publications related to one topic along different years can be viewed as a continuous function. The curve of this function can be approximated using Weierstrass approximation theorem stating that for any given time period [t-1, t] there exists a polynomial function that can be uniformly converged. Curvilinear regression is polynomial regression that fits for applications that do not have linear relations, meaning that they have curves and not straight lines. In curvilinear regression, usually an intrinsic linear model is assumed and data are fitted to this model using polynomial regression (Cohen, et al., 2013). In this article, the number of publications in WoS and the number of preprints in arXiv along the years can be viewed as curves which can be modeled by a continuous function. This function can be represented as a polynomial with independent variables $f(x, x^2, x^3, \dots)$. Curvilinear regression from Stata SE 11.2 was applied to fit the curves.

According to the Weierstrass Approximation Theorem, if *f* is a continuous real-valued function on [*a, b*] and if any ε>0 is given, then there exists a polynomial p on [*a, b*] such that $|f(x) - p(x)| < \varepsilon$ for all *x* in [*a, b*]. This can be interpreted as any continuous function on a closed and bounded interval can be uniformly approximated on that interval by polynomials to any degree of accuracy, If we make *x*=*t*, then the number of publications in WoS and arXiv would be the functions of t, i.e. y_{WoS}=*f*(t,t^2, ^), y_{arXiv}=*f*(t, t^2, ^)(*t*=1, 2, ^, 20). If we make that decision, however, then we run into a non-stationary problem (i.e., the data may follow some trends whose joint probability distribution will change when time changes and this might result in spurious or nonsensical regression). Because the differences of t^i(*i*>1) will always have the upward trend while WoS and arXiv are integrated of order 1 (i.e. data from WoS and arXiv will be stationary after difference for one time, notated as *I*(1)), then the difference between the publications of two consecutive years in each topic will remain stable. Meanwhile, t^i(*i*=1, 2, ^, 20) will grow rapidly and this might lead to the coefficients of t^ithat are very small, but important.

Considering the above scenario, we made *x*=ln(*t*). Then x^i would be *I*(1), which meant that there might be a cointegrated relationship between dependent variables and independent variables. If it was true, we could do the regression modeling if the cointegrating test proved our regressions to be correct. Meanwhile, we use $\ln^i$(t) to reduce the value of t^i.

After the selection of dependent and independent variables, we must decide the power *m* of the polynomial. Generally, the power of polynomial should be at least bigger than the number of extreme points of the curve. However, usually we cannot know the exact number of extreme points of the curve. Thus, the choice of m is partly subjective. Generally, most publications' fitted curves have no more than 4 extreme points, therefore we can make the power $m = 5$. Then, our original model would be:

$$Y = \alpha + \sum_{i=1}^{5} \beta_i \left(\ln^i(t)\right) + \varepsilon \tag{1}$$

Here, the Y-intercept α indicates the number of publications in the first year (t_1=1992), and β represents the slopes of the curve. Because arXiv's data starts from 1992 and WoS data begins earlier than 1992, then we can infer that α in equations of arXiv would be close to zero, while not in WoS.

By computing the fitting results using regression algorithm with the following steps, we were able to get the fitted equations and fitted curves for the data.

1. Set an appropriate significant level of 0.05.
2. Fit the equation (1) by ordinary least squares (OLS).
3. If all the variables pass the test, then stop, otherwise, do step4.
4. Select the variable which has the lowest significant level, drop it. Then fit the new equation by OLS again.
5. Repeat step 3 and step 4, until all variables pass the test.

Results

Overview

The LDA model was applied to articles in astrophysics collected from WoS (166,191) and arXiv (117,913) for the last 20 years (1992-2011). Fifty topics were extracted using the LDA model. The 50 topics of arXiv were matched to the 50 topics of WoS. Table 1 shows these 50 topics. Each topic was labeled using the top five ranked words (i.e., words with high probability in this topic). The extracted 50 topics demonstrate the major research topics in astrophysics including: astrophysical processes (e.g., blackhole radition, radiative transport, gravity, and star structure), stellar physics (e.g., stellar evolution, chemical dependency, white dwarfs, neutron stars, and black holes), interstellar medium (e.g., heating, gas dynamics, magnetic fields, and shocks), cosmology (e.g., models, dark matter, inflation, and accelerating), and galaxies (e.g., spiral, disk, surface, Milky Way, and density waves).

Table 1. *Fifty Topics of Astrophysics in arXiv and WoS*

Topic 00	Blackhole-massive-accretion-binaries-disk	Topic 25	dark-matter-universe-cosmology-milky
Topic 01	rotating-model-stability-theory-cosmology	Topic 26	how-what-meteor-astronomy-why
Topic 02	impact-meteorite-origin-chondrite-lunar	Topic 27	type-supernovae-neutron-core-nucleosynthesis
Topic 03	region-maser-source-infrared-line	Topic 28	alpha-redshift-field-quasar-absorption
Topic 04	line-excitation-transition-atomic-irons	Topic 29	coronal-region-loop-flux-heating
Topic 05	binaries-spectroscopy-eclipsing-light-photometric	Topic 30	spiral-surface-disk-brightness-gas
Topic 06	variable-period-cataclysmic-majoris-variation	Topic 31	active-nuclei-seyfert-variability-line
Topic 07	gammaRay-burst-GRB-afterglow-blazar	Topic 32	magellanic-cloud-globular-cloud-photometry
Topic 08	motion-observation-photograph-determination-reference	Topic 33	dwarf-group-globular-compact-elliptical
Topic 09	line-profile-polarization-spectrum-absorption	Topic 34	dwarf-lowMass-open-sequence-binaries
Topic 10	motion-orbit-theory-satellite-peridic	Topic 35	supernova-comet-remnant-cygus-coma
Topic 11	planet-system-extrasolar-satellite-jupiter	Topic 36	sky-source-sample-catalog-digital-rosat
Topic 12	background-cosmic-microwave-power-spectrum	Topic 37	dust-circumstence-disk-tauri-envelope
Topic 13	cosmic-energy-ray-gammaRay-highEnergy	Topic 38	spectral-distance-distribution-determination
Topic 14	oscillation-model-pulsation-mode-delta	Topic 39	wave-convection-dynamo-flow-rotating
Topic 15	nova-outburst-spectrum-cygni-dwarf	Topic 40	accretion-disk-simulation-wind-jet
Topic 16	photometry-cepheids-open-photoelectric-distance	Topic 41	acceleration-plasma-wave-shock-radiation
Topic 17	telescope-space-hubble-imaging-observation	Topic 42	data-analysis-method-astronomy-application
Topic 18	spectra-ultraviolet-analysis-atmosphere-wolfRayed	Topic 43	measurement-atmosphere-mars-satellite-venus
Topic 19	sunspot-rotation-activity-cycle-variation	Topic 44	flare-hard-burst-observed-coronal
Topic 20	abundance-giant-chemical-red-branch	Topic 45	interstellar-could-dust-grain-diffuse
Topic 21	gas-interstellar-hydrogen-neutral-cloud	Topic 46	molecular-cloud-core-region-dense
Topic 22	nebula-planetary-central-orion-bipolar	Topic 47	gravitation-lensing-microlens-quasar-weak
Topic 23	pulsar-X1-PSR-source-transient	Topic 48	luminosity-function-relation-distribution-redshift
Topic 24	source-radio-polarization-object-compact	Topic 49	transfer-radiative-method-equation-radiation

Figure 2 shows the topic distribution of these 50 topics for arXiv and WoS over these 20 years. In arXiv (the blue line), topics were distributed with topic 25 (dark matter research) as the one sharing the most proportion throughout the 20-year time span, then followed by topic 12 (Cosmic Microwave Background (CMB)), topic 28 (Lyman-alpha systems and cosmology), topic 7 (Gamma-ray burst (GRB)), and topic 33 (dwarf-group-globular-compact-elliptical). In WoS (the red line), the topic distribution was comparably stable with lower amplitude oscillations. The proportion of topic 25 (dark-matter-universe-cosmology-milky) remained high during the later years, followed by topics: 31 (active-nuclei-seyfert-variability-line), 28 (alpha-redshift-field-quasar-absorption), 23 (pulsar-X1-PSR-source-transient) and 40 (accretion-disk-simulation-wind-jet).

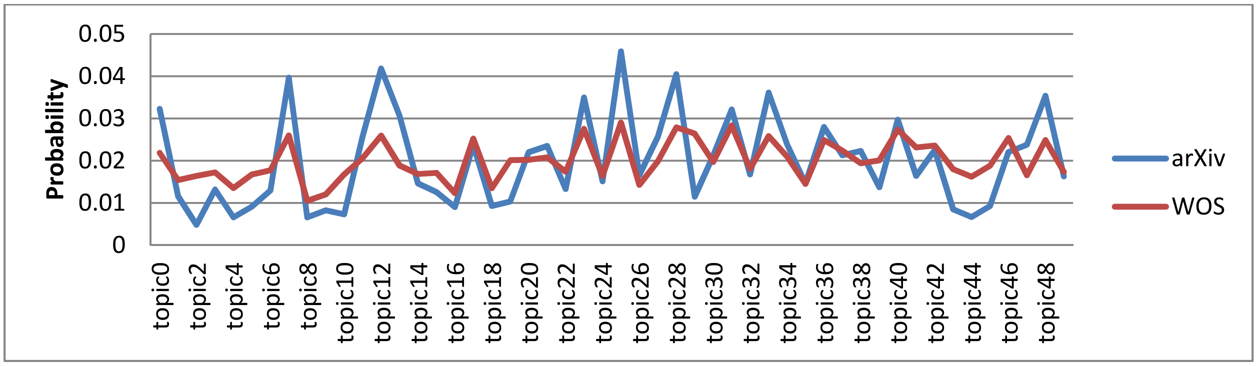


Figure 2. Overview of the topic distribution for astrophysics paper in WoS and arXiv (1992-2011)

(Note: horizontal axis represents topics, and vertical axis represents the topic distribution probability)

Topic growth patterns

The regression algorithm was applied to articles in astrophysics collected from WoS (166,191) and arXiv (117,913) for the last 20 years (1992-2011). 50 pairs of fitted equations for the 50 topics (listed above in Table 1) were extracted. Based on the p-values and R squares, we are able to fit most of the equations except for topic 3 in WoS. Since the number of publications on topic 3 (region-maser-source-infrared-line) is stationary, we can use its mean as its fitting value. Thus, all the data has been fitted. All the regression equations have been evaluated and passed the three statistical tests including: the White Test for the heteroscedasticity (White, 1980), the Bartlett test for the autocorrelation (Bartlett, 1937), and the cointegrating test for the long-run equllibrium (Granger, 1986). From the fitted equation curves, four types of topic growth patterns have been identified.

Both in upward trend. Figure 3 shows two representative curves for 29 topics in both WoS and arXiv that show the upward trends. 58.6% of these 29 topics (0, 7, 11, 12, 13, 14, 25, 27, 28, 33, 35, 37, 38, 40, 42, 46, 48) have the curves similar to topic 0 (blackhole-massive-accretion-binaries-disk) in which both WoS and arXiv display upward trends such that WoS led first and was overtaken by arXiv. 41.4% of 29 topics (2, 6, 8, 10, 19, 29, 39, 41, 43, 44, 45, 49) have the curves similar to topic 2 (impact-meteorite-origin-chondrite-lunar) in which WoS and arXiv both grow, but at different rates such that WoS is in the dominant leading position throughout the whole 20 years. However, arXiv has higher growth rate than that of WoS.

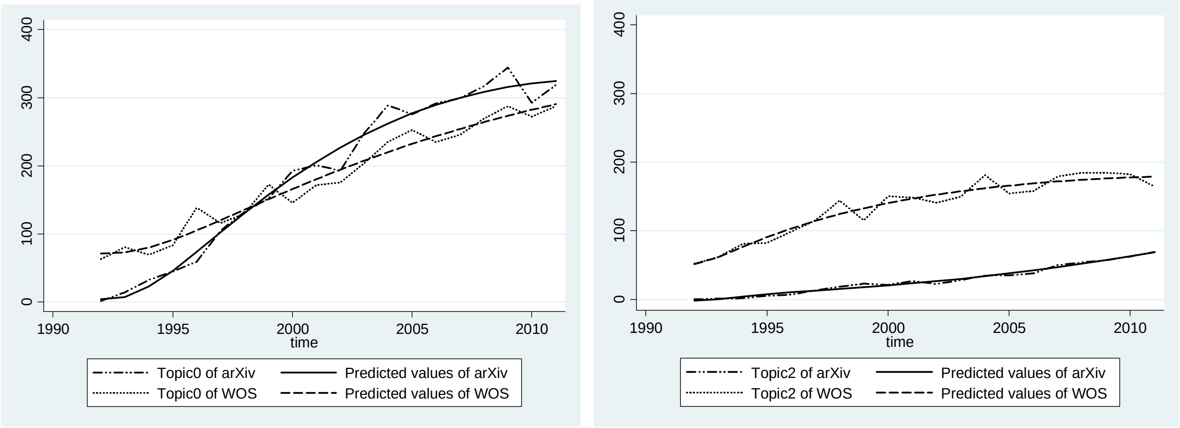


Figure 3. Fitted curves for topic 0 and topic 2

(Note: Y axis shows the publications in arXiv and WoS while X axis shows the time from 1992 to 2011)

WoS in upward trend vs. arXiv in downward trend. Figure 4 shows that topic 1 (rotating-model-stability-theory-cosmology) in WoS has a trend of rising but that it has the opposite direction in arXiv, which indicates that the publications of this topic in WoS are increasing but the corresponding publications in arXiv are decreasing. This is the only instance of such a behavior and it is perhaps due to the spurious identification of topic 1.

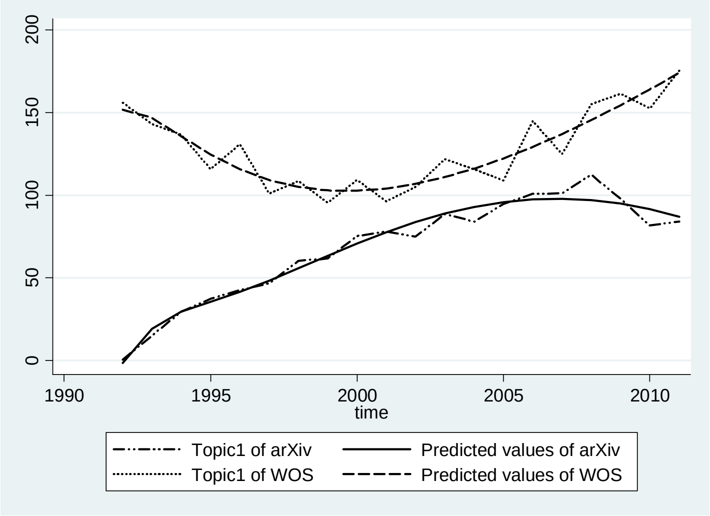


Figure 4. Fitted curves for topic 1

(Note: Y axis shows the publications in arXiv and WoS while X axis shows the time from 1992 to 2011)

WoS in downward trend vs. arXiv in upward trend. Figure 5 shows examples of topics that exhibit downward trend in WoS and an upward trend in arXiv. There are 12 topics that belong to this type. Among them, 11 topics (3, 9, 15, 17, 20, 23, 26, 31, 32, 34, 36) have curves similar to that of topic 15 (telescope-space-hubble-imaging-observation). The WoS curve demonstrates a declining trend, while the arXiv curve displays a growing trend even though WoS leads temporarily. Topic 5 (binaries-spectroscopy-eclipsing-light-photometric) has no intersection between WoS and arXiv during the studied period.

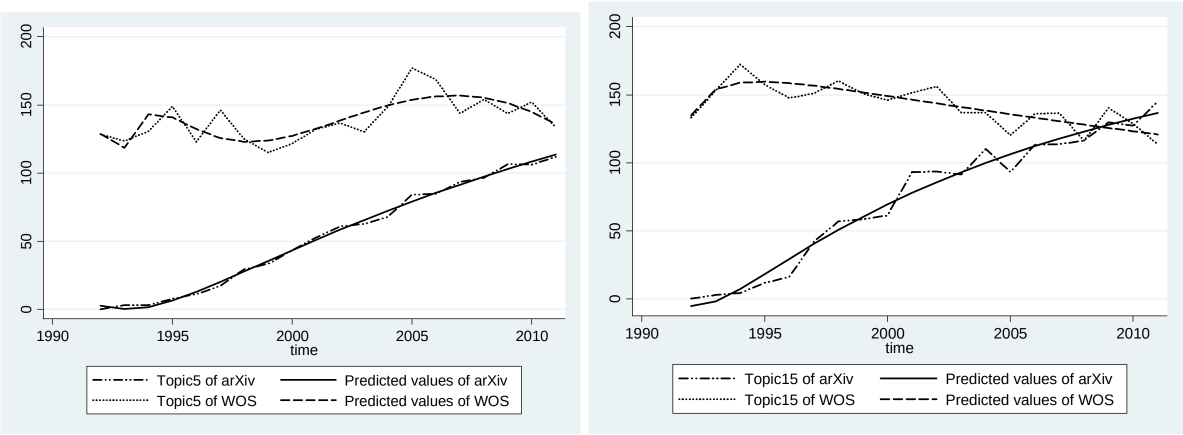


Figure 5. Fitted curves for topic 15 and topic 5

(Note: Y axis shows the publications in arXiv and WoS while X axis shows the time from 1992 to 2011)

Both in downward trend. Figure 6 shows two representative curves for eight topics that exhibit declining trends in both WoS and arXiv. Seven of these eight topics (16, 18, 21, 22, 24, 30, 47) have curves similar to topic 21 (gas-interstellar-hydrogen-neutral-cloud) such that WoS starts to decline before arXiv. Topic 4 (line-excitation-transition-atomic-irons) displays no interaction between WoS and arXiv, which WoS leads.

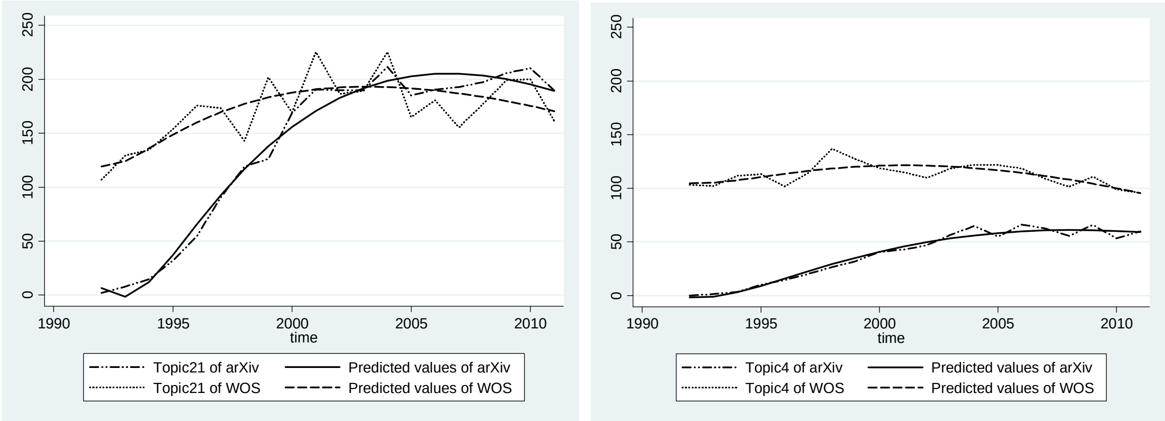


Figure 6. Fitted curves for topic 21 and topic 4

(Note: Y axis shows the publications in arXiv and WoS while X axis shows the time from 1992 to 2011)

Topic Trends and Popularity

Because a derivative indicates the slopes of a curve, it is applied here to represent the trend of the topics. Different from the publication number, a derivative value (i.e., the slope of a curve) indicates the growing/declining rate according to time. The absolute value of the derivative measures the "steepness" of the growth/decline, while the minus or plus sign of the derivative indicates the direction of "steepness". For example, a positive derivative means the publication number in the next period will be higher than that in the current period, while a negative one implies the opposite. From the definition of derivatives, $f'(x) = [f(x+\delta) - f(x)]/\delta$, when $\delta \rightarrow 0$. By approximating the derivative by the equation as follows, we could get the corresponding derivative of the fitted equations, i.e. the slope of the fitted curves.

$f'(time) = [f(time+1) - f(time)]/1$ (i.e. the difference of order 1 for the prediction data)

A common property of trends

Almost all the topics (41 topics) have a common characteristic: as time goes on, the topics in arXiv have much higher growth rates than those in WoS (for example, see topic 4 in Figure 7) even though in the beginning the number of publications in WoS is higher than that in arXiv (for example, see topic 2 in Figure 7). This indicates that publishing papers in arXiv is becoming popular. This phenomenon also means that arXiv has much higher potential growth power as more people tend to publish their articles in arXiv.

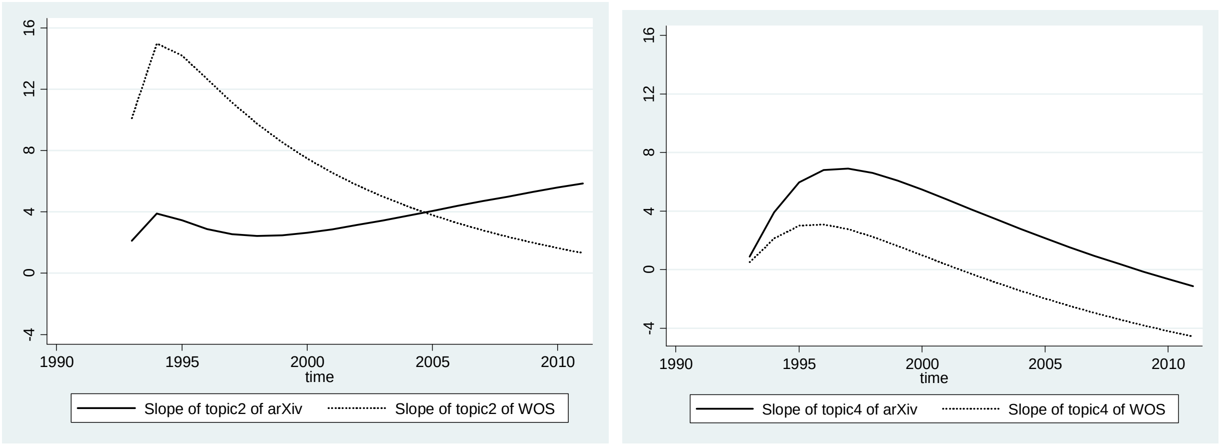


Figure 7. Comparison of slopes from arXiv curves and WoS curves

(Note: Y axis shows the publication slopes in arXiv and WoS while X axis shows the time from 1992 to 2011)

Analyses of Topic Popularity. Each topic has its own cycle of popularity: becoming popular, losing popularity, or reganining popularity. Each popularity lasts a period of time. We define the measures of popularity as following:

- Gaining popularity: In period [*t, t*+*n*], the time with the largest addition of new elements, i.e. having maximum $f'(t_m)$, is defined such that the topic is gaining popularity.
- Losing popularity: If after the maximum point, $f'(t)$ becomes negative, the topic is considered to be losing popularity.
- Regaining popularity: If the curve of topic has a minimum point, then the topic might become popular again in few years, so we view this case as becoming re-popular.
- Duration of popularity: If exists a *t* which is a maximum point to $f(t)$, then t-t_m is the period during which the topic was popular.

Let us examine the fitted and sloped curves of topic 22 to illustrate the above definitions (Figure 8). Point B is a gaining popularity point because it has the biggest slope, i.e., the growth rate of publications covering topic 22 in arXiv in 1997 is the highest. In other words, that is the time in which topic 22 is gaining fastest in popularity. Point C is a losing popularity point (i.e., maximum value point) because after that point in 2009 the number of publications of topic 22 in arXiv decreases (i.e., the slope of the curve changes to negative from positive), which means this topic is losing its attraction. Point A is a regaining popularity point (i.e., minimum value point) because before 1994, the number of publications of this topic decreases and after this year the number of publications increases, and the topic becomes popular again in 1997 (we assume that no topic is unpopular at the very start). From Figure 8, we can compute the duration of popularity by deducting 1997 from 2009 (the period between losing popularity and gaining popularity, i.e., between points C and B).

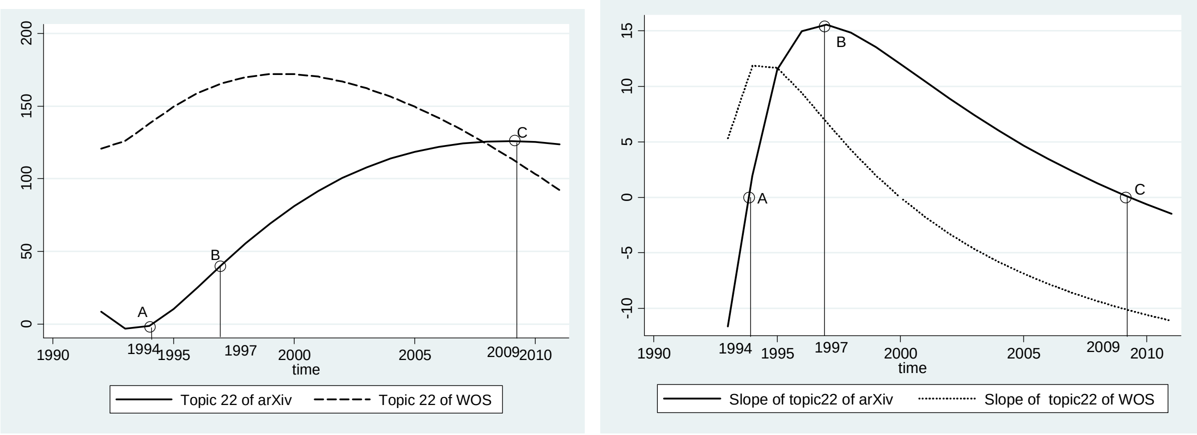


Figure 8. Example of the mathematical measures of topic popularity

(Note: Left graph shows topic 22 fitted values which Y axis indicates the publications in arXiv and WoS while X axis shows the time from 1992 to 2011; right graph shows the slopes of topic 22 in arXiv and WoS which Y axis indicates the left curves' slopes while X axis shows the same time period.)

The derivative curves show that topic 1 in arXiv, topics 5, 6, 9, 15, 16, 17, 18, 20, 23, 26, 31, 32, 34 in WoS, and topics 4, 21, 22, 24, 30, 47 in both WoS and arXiv are losing popularity in recent years. Topic 5 in WoS and topic 22 in arXiv have regained popularity early on, only to lose popularity. Topics 1, 8, 10, 39, and 49 in WoS have regained popularity at first and have kept growing in popularity. The rest of topics (0, 7, 11, 12, 13, 14, 25, 27, 28, 33, 35, 37, 38, 40, 42, 46, 48) have consistently been gaining in popularity. Figure 9 displays the examples of topics gaining, losing, and regaining popularity. For example, topic 1 in WoS has a minimum point in 2000, and after that only a positive slope, so it is a topic that has regained popularity; topic 1 in arXiv has a maximum point in 2007, and after that has only a negative slope, so it is an example of a topic losing popularity; topic 2 in WoS and arXiv always has a positive slope and is therefore gaining popularity throughout the whole period.

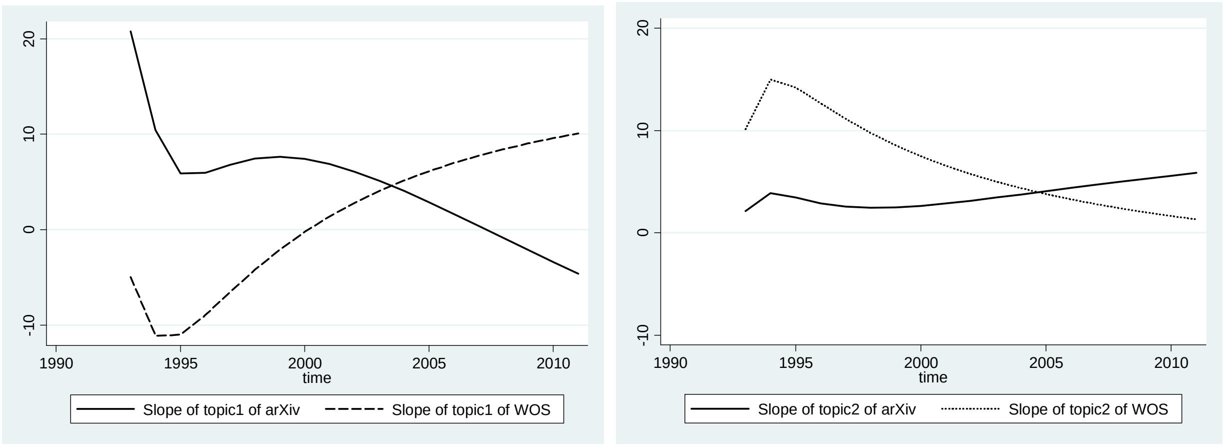


Figure 9. Examples of topic popularity

(Note: Y axis shows the publication slopes in arXiv and WoS while X axis shows the time from 1992 to 2011)

Popularity duration and leading time (i.e., the duration of which arXiv or WoS leads the other for the same topic) can be calculated using the derivative curves. We listed information on popularity for all the 50 topics in Table 2 using the above definitions. For example, topic 0 in arXiv becomes popular in 1997 while in WoS in 1998, so arXiv leads WoS and the lead time is one year. The maximum value points of the topic 0 in both arXiv and WoS appears in 2011, meanwhile the slopes of the both curves are not 0, so the popular duration of the topic in arXiv is 15 years and in WoS is 14 years. For topic 3 in WoS, the maximum value of slope appears in the first year, so the topic gains popularity before 1993. The same topic in arXiv becomes popular in 1999, so the WoS leads arXiv 6 years. The topic 20 in both WoS and arXiv gains popularity in 1995, so the lead time is 0. The topic's losing popularity point in WoS is 2007 while the maximum value point with nonzero slope is in 2011, so the duration of popularity in WoS is 13 years and in arXiv is 17 years.

Table 2. *Topic Popularity Information of the 50 Topics*

Topics	Lead	Lag	Time of becoming Popular(Lead)	Lead Time (year)	Popularity duration (Lead) (year)	Popularity duration (Lag)(year)
0	arXiv	WoS	1997	1	>= 15	>=14
1	WoS	arXiv	before 1993	-	-	>=15
2	WoS	arXiv	1994	>=17	>=18	>=1
3	WoS	arXiv	before 1993	>=6	>=19	>=13
4	WoS	arXiv	1996	1	6	12
5	WoS	arXiv	before 1993	>=6	-	>=13
6	WoS	arXiv	1994	1	5	>=17
7	WoS	arXiv	1994	2	>=18	>=16
8	WoS	arXiv	1994	0	3(re-popular)	>=18
9	WoS	arXiv	before 1993	>=6	>=3	>=14
10	WoS	arXiv	before 1993	>=7	>=2(re-popular)	>=13
11	WoS	arXiv	2011	0	1	1
12	arXiv	WoS	before 1993	>=1	>=19	>=18
13	WoS	arXiv	1994	3	>=18	>=15
14	WoS	arXiv	1999	>=12	>=12	>=1
15	WoS	arXiv	before 1993	>=3	>=4	>=16
16	WoS	arXiv	before 1993	>=3	-	16
17	WoS	arXiv	1996	3	11	>=13
18	WoS	arXiv	before 1993	>=4	>=2	12
19	WoS	arXiv	before 1993	>=18	>=19	>=1
20	WoS	arXiv	1995	0	13	>=17
21	WoS	arXiv	1995	1	9	12
22	WoS	arXiv	1994	3	6	13
23	WoS	arXiv	1995	1	11	10
24	WoS	arXiv	1994	1	6	14
25	arXiv	WoS	before 1993	>=18	>=19	>=1
26	arXiv	WoS	1996	1	>=16	12
27	arXiv	WoS	1997	>=14	>=15	>=1
28	arXiv	WoS	1995	2	16	>=15
29	WoS	arXiv	before 1993	>=18	>=19	>=1
30	WoS	arXiv	before 1993	>=2	>=4	14
31	WoS	arXiv	1995	1	10	8
32	WoS	arXiv	1994	1	6	>=17
33	WoS	arXiv	1995	1	14	>=16
34	WoS	arXiv	1997	2	11	>=13
35	WoS	arXiv	before 1993	>=3	>=19	>=16
36	WoS	arXiv	1994	1	14	>=17
37	WoS	arXiv	1994	7	>=18	>=11
38	WoS	arXiv	1994	0	>=18	>=18
39	WoS	arXiv	2011	0	>=1	>=1
40	WoS	arXiv	before 1993	>=3	>=19	>=16
41	arXiv	WoS	1998	>=13	>=14	>=1
42	WoS	arXiv	1994	>=17	>=18	>=1
43	WoS	arXiv	1997	>=14	>=15	>=1
44	WoS	arXiv	2011	0	>=1	>=1
45	WoS	arXiv	1995	4	4(Re-popular)	>=13
46	arXiv	WoS	2002	>=9	>=10	>=1
47	WoS	arXiv	1994	0	9	11
48	WoS	arXiv	1994	1	>=18	>=17
49	arXiv	WoS	1994	>=17	>=18	>=1

(Note: 1. "-" means that the accurate value cannot be calculated during the period of 1992 to 2011.

2. in the fourth column, "before 1993" means that the maximum $f'(t_m)$ appears in the first year (since we use *I(1)* of prediction data as the derivative values, the value of 1992 is missed). However, it is not a certainty that the topic becomes popular in or before 1993, so we use "before 1993" indicating this condition.)

Domain Evaluation

The analytical results have been evaluated by a domain expert from astrophysics and are summarized as follows. The topics that were identified as having upward trends in both arXiv and WoS are topics 0, 7, 11, 12, 13, 14, 25, 27, 28, 33, 35, 37, 38, 40, 42, 46, 48. All of these are "hot topics" -- sources of enormous attention in recent years, and the site of many of the most important discoveries. Among other things, (1) young astronomers seeking to build a career would tend to work in these fields, (2) many groups would be likely to work on similar problems, leading to a desire to use to arXiv's instant publishing model to establish priority, (3) papers in some of these fields involve "big data" gathered by large collaborations who may wish to use arXiv to rapidly present results to funders. The Gamma ray bursts (Topic 7) topic is an interesting phenomenon that has had a rapid onset. The community has often tried to synchronize observations; where time is of the essence, use of arXiv to disseminate results seems natural.

The topics which both have upward growth in both WoS and arXiv (but at different rates such that WoS is in the dominant leading position throughout the whole 20 years) are topics 2, 6, 8, 10, 19, 29, 39, 41, 43, 44, 45, 49. These topics are more "traditional" topics. For example, Topic 6 (cataclysmic variables) involves the study of a phenomenon that, while unlikely to have sensational consequences for the hot topics mentioned above, is perennially popular. Topic 8 appears to involve a more "traditional" topic that is focused on astrometry; this is a topic that is useful to many fields but not of direct interest (though this might change with the increasing role of astrometry in extra-solar planets) to astrophysicists. Topic 45 (diffusion on interstellar dust grains) and Topic 49 (problems in radiation transfer) also appear to fall in this “traditional” category. Solar and solar system phenomena, including Topic 10 (solar system dynamics), topic 19 (solar phenomena, sunspots), Topic 29 (solar phenomena, magnetic heating of plasma), Topic 39 (solar dynamo), Topic 43 (other planets in our solar system), and Topic 44 (solar flares), appear here suggesting that a subgroup of astronomers have not turned to pre-prints as much.

The topics in WoS are increasing but the corresponding publications in arXiv are decreasing are Topic 1. Topic 1 may include a somewhat unfashionable field involving rotating universe models. If this identification is correct, it is a theoretical field with some senior researchers, but any early interest among the younger generation may have died out in the mid-1990s when cosmological observations began to seriously rule out these speculative models.

The topics in WoS that have shown downward trend while trending upward in arXiv are topics 3, 5, 9, 15, 17, 20, 23, 26, 31, 32, 34, 36. These topics are difficult to identify coherently because they appear to be associated with (what some would denigrate as) "stamp collecting" fields -- observations of a particular object or small number of objects (globular clusters, radio masers (topic 3), polarization of spectral lines (topic 9), variable stars (topic 15), Hubble images (topic 17), pulsars (topic 23), meteors (topic 26), Active Galactic Nuclei (topic 31), the magellanic cloud (topic 32); or ROSAT images (topic 36); or studies of stellar evolution (topic 20 and topic 34). It may be the case that authors are using arXiv to publish results examining these topics because there is insufficient interest to push through to peer review.

The topics that show a downward trend in both WoS and arXiv are topics 4, 16, 18, 21, 22, 24, 30. Some of these topics are clearly superseded by other advances; Topic 16 (Cepheid variables) was a crucial part of cosmology until Supernovae and other "standard candles" took over. Topic 47 (Gravitational microlensing) is still a very interesting topic (and may well be again), but it had high popularity early on when astrophysicists thought it was possible that dark matter was baryonic, or at least in compact object form. After the year 2000, astrophysicists believe this to be impossible, so microlensing studies became much less interesting to the wider community. Others are "sleepy" in a similar way to the previous section -- e.g., Topic 4 (atomic lines), Topic 18 (Wolf-Rayet stars), Topic 22 (planetary nebulae), and Topic 24 (radio polarization of compact objects). Topic 21(neutral interstellar hydrogen) is *so* sleepy that it might be said that we know everything there is to know.

Conclusion

This paper analyzes the topic evolution patterns for preprints and traditionally published articles in the area of astrophysics. Both arXiv and WoS share similar topics in astrophysics but have diverse evolution trends. First, LDA was applied to arXiv and WoS for the time period of 1992-2011 to identify 50 topics. The regression model has been set up for each of 50 topics to calculate their trend curves; four topic development patterns were identified. The first pattern (containing 29 topics) signifies that both arXiv and WoS have upward growth tendency, with arXiv having stronger growth tendency than WoS. Even though WoS leads at the beginning, arXiv surpasses or will surpass eventually. The second pattern (containing 1 topic) shows that WoS is in upward trend while arXiv is in a downward trend; this only happens for topic 1 (rotating-model-stability-theory-cosmology), which remains an isolated case. The third pattern (containing 12 topics) found that WoS is in a downward trend while arXiv is in an upward trend. The fourth and final pattern (containing 8 topics) identified that both WoS and arXiv are in downward tendencies, with WoS having a stronger downward tendency than arXiv. In summary, arXiv is getting more popular than WoS and arXiv is leading or will lead in the near future.

Table 2 summarizes the lead-lag situation of WoS and arXiv. WoS leads almost all of the topic popularity (41 topics) ranging from one year to six years except with topics 0, 12, 25-28, 41, 46 and 49. In summary, topics 14, 43, 2, 42, 19, and 29 (6 topics) in WoS and topics 41, 27, 49 and 25 (4 topics) in arXiv lead the other for more than 10 years; topic 46 in arXiv leads WoS by 9 years; topics 3, 5, 9, 10 and 37 (5 topics) in WoS lead these topics in arXiv by 6-7 years; topics 13, 15, 16, 17, 22, 35, 40, 18 and 45 (9 topics) in WoS lead these topics in arXiv by 3-4 years; topics 0, 12, 26 and 28 (4 topics) in arXiv and topics 4, 6, 21, 23, 24, 31, 32, 33, 36, 48, 7, 30 and 34 (13 topics) in WoS lead the other by 1-2 years; and finally topics 8, 11, 20, 38, 39, 44, 47 (7 topics) do not lead each other. Compared with arXiv, however, WoS carries much shorter popularity duration. The average topic popularity duration of arXiv is about 11 years while that of WoS is about 10 years. Only 17 topics' popularity durations of WoS are larger than those of arXiv. This also suggests that most topics in WoS start losing popularity sooner than those in arXiv. It can be clearly seen in our results that the strength of arXiv is growing; this verifies our conclusion that arXiv has more potential growth power as more people tend to publish their articles in arXiv. It also indicates that arXiv tends to carry popular topics which attract wider attention.

From the lead time we find that most of the topics (total 34) found in both arXiv and WoS affect the popularity of each other because when a topic becomes popular in either arXiv or WoS, it will become popular in the other in less than 4 years. Very few topics (such as topics 19, 25, 42 and 49) become popular in one channel without becoming popular in the other. Only 10 topics' lead time in arXiv and WoS are longer than 10 years (i.e. meaning that a topic becomes popular in arXiv or WoS but is not popular in the other channel in the next 10 years). In summary, although WoS leads the majority of the topics (i.e. 41 out of 50 topics), arXiv has much stronger growth tendency than WoS. Open access has recently become a trend in scientific publishing that provides free and easy access to everyone, accelerates knowledge transfer and discovery, and enables better opportunities for learning; today's open access becomes tomorrow's open knowledge. Open access goes beyond articles to enable a link between research datasets and reference data entities. Domain-specific research data repositories have been built up to ensure sustainable access, provide better data integration, and empower analytics and impact assessment.

This work clearly demonstrates that open access preprints will have stronger growth tendency as compared to traditional printed publications in astrophysics. One of the limitations of this paper is that the main findings in this paper might be difficult to generalize toward other domains. To address this limitation, future work should include: 1) applying author-topic model or author-conference topic model to extract topic distributions for authors and journals in order to compare the topic evolution for authors and journals; 2) comparing the evolution patterns in astrophysics with other domains in science (e.g. Chemistry), social science and humanities; 3) studying the usage of datasets in open access articles to identify best practices for sharing and integrating publically available research datasets; 4) extending the lead-lag study to co-author networks or author citation networks to identify the scholarly communication behavior differences in formal vs. informal publishing channels; and 5) investigating the author credit distribution differences in these two publishing channels, especially to pay attention to first authors or corresponding authors and co-author's contribution.

Acknowledgement

This project is supported by Digging Into Data Project funded by NSF (NSF grant SMA-1208804). We thank Cassidy R. Sugimoto for comments and Simon DeDeo for evaluating the results from the perspective of a domain expert in astrophysics.

References

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009, June). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (pp. 27-34). AUAI Press.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. *Series A-Mathematical and Physical Sciences, 160*(901), 268-282.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120). ACM.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009, December). Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS* (Vol. 22, pp. 288-296).

Chatfield, C. (2003). The Analysis of Time Series: An Introduction, (Chapman & Hall/CRC Texts in Statistical Science).

Cohen, J., Cohen, P., West, S.G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge Publisher.

Ding, Y. (2011a). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498-514.

Ding, Y. (2011b). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics, 5*(1), 187-203.

Ding, Y. (2011c). Topic ‐ based PageRank on author cocitation networks. *Journal of the American Society for Information Science and Technology, 62*(3), 449-466.

Gerrish, S., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 375-382).

Granger, C. W. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of economics and statistics, 48*(3), 213-228.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery, 7*(4), 373-397.

Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXiv E‐prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*.

Liu, X., Zhang, J., & Guo, C. (2012, October). Full-text citation analysis: enhancing bibliometric and scientific publication ranking. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1975-1979). ACM.

Mann, G. S., Mimno, D., & McCallum, A. (2006, June). Bibliometric impact measures leveraging topic analysis. In *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on* (pp. 65-74). IEEE.

Mei, Q., & Zhai, C. (2005, August). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198-207). ACM.

Mei, Q., Shen, X., & Zhai, C. (2007, August). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 490-499). ACM.

Natale, F., Fiore, G., & Hofherr, J. (2012). Mapping the research on aquaculture. A bibliometric analysis of aquaculture literature. *Scientometrics, 90*(3), 983-999.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487-494). AUAI Press.

Shaparenko, B., Caruana, R., Gehrke, J., & Joachims, T. (2005). Identifying temporal patterns and key players in document collections. In *Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05)* (pp. 165-174).

Shi, X., Nallapati, R., Leskovec, J., McFarland, D., & Jurafsky, D. (2010). Who leads whom: Topical lead-lag analysis across corpora. In *NIPS Workshop*.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations. *PloS one, 7*(11), e47523.

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the Association for Information Science and Technology, 65*(2), 352-371.

Stone, M. H. (1948). The Generalized Weierstrass Approximation Theorem. *Mathematics Magazine*, 21(4): 167-184.

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. *Journal of the American Society for Information Science and Technology, 62*(1), 185-204.

Swan, R., & Jensen, D. (2000, August). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining* (pp. 73-80).

Tang, J., Jin, R., & Zhang, J. (2008, December). A topic modeling approach and its integration into the random walk framework for academic search. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 1055-1060).

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817-838.

Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics, 6*(1), 140-153.