Who Leads Whom: Topical Lead-Lag Analysis across Corpora

Xiaolin Shi, Ramesh Nallapati, Jure Leskovec, Dan McFarland, Dan Jurafsky
Stanford University
Stanford, CA 94305
{shixl,nmramesh,jure,dmcfarland,jurafsky}@stanford.edu

Abstract

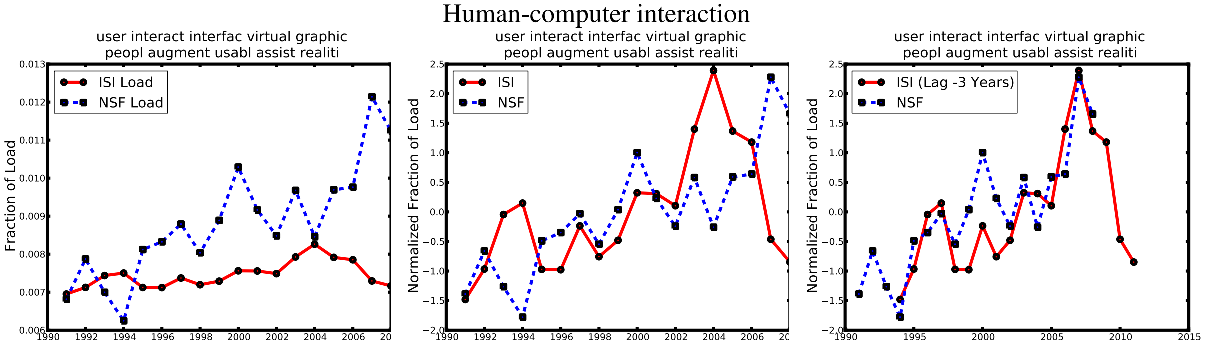
In this work, we study the problem of whether grant proposals lead academic publications in terms of generation of scientific ideas. This is an important computational social sciene question that can help us understand the dynamics of scientific innovation. We propose simple but scalable techniques for lead/lag estimation, based on LDA and time series analysis, that work on any unlabeled textual corpora with temporal information. We perform our analysis on about half a million Computer Science research paper abstracts and 20,000 successful NSF grant proposal abstracts that represent the entire field of Computer Science in the time span of 1991-2008. Our analysis, besides revealing interesting patterns, finds that the lead/lag of scientific papers with respect to grant proposals is highly topic specific.

1 Introduction

Massive records of academic publications and grant proposals that funded the corresponding research can inform our understanding of research policy making on a scale never possible before. We align research publications that represent scientific output and grant proposals that represent research initiatives that promise to move beyond the state of the art. This gives us a powerful framework for analyzing correlations between funding and scientific output. It also allows us to identify the areas where fund granting agencies such as NSF are merely lagging behind scientic innovation and the areas where NSF is clearly pushing the envelope and creating new and vibrant scientific areas. Understanding the temporal relation between scientific output and funding is one of the fundamental questions that shapes the scientific funding policy. It also has important implications on career trajectories of junior faculty.

Recent advances in modeling the mechanics of information propagation within and across these domains enables scholars to begin such large scale study of social dynamics in earnest. Early work on tracking trends within or across corpora made use of rich meta data such as hyperlinks, co-authorship, citations or use short and distinctive phrases, etc. [3, 5]. However, in many other types of textural corpora, they do not have such explicit evidences, through which we can trace the propagation of information. For example in the grants data, there are no incoming citations by the very nature of grants not being an evidence of accomplished research. Also, the references in grants are misleading because they do not capture future innovative research promised in them. On the other hand, matching authorship information between grants and publications will answer only questions about author-specific behavior but will not answer questions about the overall trends across the corpora. The same issues may also be relevant to other domains such as blogs and instant messages where citation information is either missing or unreliable.

In this work, we focus only on using textual and timestamp data to estimate the lead/lag of grants vs. publications. More specifically, we study the lead/lag of topics in Computer Science between two corpora of NSF grants and ISI publications from 1991 to 2008. An additional complexity in our dataset is that each document can discuss multiple topics, and therefore one needs to decompose each document into its topics before we analyze them. The techniques we propose are simple and intuitive but scalable, and we hope they generalize well to other corpora with similar attributes.



(a) Raw loading (b) Standardized loading (c) Standardized loading after alignment

Figure 1: Estimation of lead/lag by aligning topic loadings. Plot (a) shows the time series curves of topic loadings of the topic ”Human Computer Interaction” on ISI publications and NSF grants. In Plot (b), we use Gaussian normalization to make the variations in the both time-series more comparable. In Plot (c), we shift the ISI time series to the left and right until we find the optimal alignment in terms of cross-correlation.

2 Methodology

Since our corpus is unlabeled, we first need a model that learns the topics discussed in a corpus. For this purpose, we use LDA [2], a popular unsupervised model that learns from a corpus of documents, a prespecified number of topics, K, where each topic is represented as $\beta_k = (\beta_{k,1}, \cdots, \beta_{k,V})$, a multinomial distribution over the vocabulary of size V. In addition, for each document d, LDA also learns its soft labels in terms of these topics, represented as $\theta_d = (\theta_{d,1}, \cdots, \theta_{d,K})$, a multinomial distribution over topics.

We first train the LDA model on the union of both corpora. Then, for each corpus c, we compute its *topic-loading* with respect to year y on topic k, as:

$$l_c^{(k)}(y) = \frac{\sum_{d:t(d)=y \text{ and } d\in c)} \theta_{d,k}}{\sum_{d:t(d)=y \text{ and } d\in c)} 1} \tag{1}$$

where $t(d)$ is the timestamp of the document d. In other words, $l_c^{(k)}(y)$ represents the expected proportion of documents that discuss topic k in corpus c, in year y. Intuitively, this quantity is the popularity of the topic in the corpus in a given year.

In the academic environment, interests of the research community as well as those of the funding agencies on a specific topic tend to have life cycles that have patterns of ‘rises’ and ‘falls’. These broad patterns can be captured by the popularity of the topic in the respective corpora as a function of time. We postulate that the lead/lag one one corpus with respect to the other can be estimated by comparing the topic ‘popularity’ plots of the corpora. More formally, a corpus c_1 lags corpus c_2 on topic k by τ years, if the patterns of rise and fall of ‘popularity’ of the topic in corpus c_2 as we vary year y is similar to that in corpus c_1 $y + \tau$ years hence. Accordingly, we shift the time series of topic loadings of one corpus to the past and future until we find the best alignment, with the topic loadings of the other corpus. The optimal alignment value is the estimated lag of corpus c_1.

Mathematically, we compare the time-series data of year-wise topical loadings $l_{c_1}^{(k)}(y)$ and $l_{c_2}^{(k)}(y)$ from the two corpora c_1 and c_2 using cross-correlation as follows:

$$\text{Corr}_{c_1,c_2}^{(k)}(\tau) = \sum_y \hat{l}_{c_1}^{(k)}(y+\tau)\hat{l}_{c_2}^{(k)}(y) \tag{2}$$

where $\hat{l}_{c_2}^{(k)}(y)$ is the normalized $l_{c_2}^{(k)}(y)$ obtained after substracting its mean over all years and dividing by its standard deviation, and τ is the lag of corpus c_1 with respect to corpus c_2. The normalization is done to filter corpus specific characteristics[1] and make the two plots more comparable with each other.

[1] *e.g.*: a topic could have lower mean loading in one corpus compared to the other.

Thus, $\mathrm{Corr}^{(k)}_{c_1,c_2}(\tau)$ captures the normalized cross-correlation between the two time series data $l^{(k)}_{c_1}(y)$ and $l^{(k)}_{c2}(y)$ as a function of lag τ of the corpus c_1. We now compute the actual lag $\tau^*(k)$ of corpus c_1 on topic k as the value of lag that maximizes the normalized cross-correlation, as follows:

$$\tau^*(k) = \arg\max_{\tau} \mathrm{Corr}^{(k)}_{c_1,c_2}(\tau) \tag{3}$$

Thus, $\tau^*(k)$ is the time lag of corpus c_1 at which the two time series best align with each other. An example of the normalization of the topic loadings time series and their aligment process is illustrated in Figure 1.

Cross-correlation is only one of several metrics available to align time series data. We also experimented with the L2 norm between the two normalized time series as an alternative, but we found that it had very high correlation with cross-correlation. Hence we do not report it in this paper.

3 Datasets

We focused our analysis in the area of Computer Science. From the ISI Dataset[2] consisting of most academic journal publications since 1960's, we extracted abstracts from Computer Science publications based on the "Field" labels, which resulted in about 471,000 documents. A vast majority of the these documents are uniformly distributed in the timespan between 1991 and 2008. We also have successful grant proposals data from NSF[3] whose awards are mostly from year 1990 to 2009. We extracted all Computer Science abstracts from this dataset using the NSF program names, which resulted in about 17,000 abstracts.

The ISI dataset is much larger than the NSF dataset, but we decided not to subsample the dataset because subsampling could introduce artifacts that might distort the time series plots. For the purposes of our experiments, we assume that the two datasets represent the true state of the research and grants worlds in Computer Science.

4 Experiments and Results

We performed all our experiments using a parallelized implementation of David Blei's LDA code[4], as described in [4]. We ran LDA models for various number of topics on an 8-core Intel Xeon 2.4GHz machine with Linux kernel and 72G RAM. For the union of ISI and NSF datasets consisting of about half a million documents, a 150-topic model finished in under 24 hours, wall clock time.

Figure 2 shows the aligned time series plots for four representative Computer Science topics discovered by the LDA model. The plots reveal some interesting trends. In both ISI and NSF datasets, topics such as "Security and cryptography" and "Mobile networks" show increase in popularity in recent times, while topics such as "Data structures" and "Neural networks" show decrease, which agree with our general understanding of the Computer Science field. We speculate that the more recent smaller spurt (the second peak in Figure 2(d)) in popularity of "Neural networks" may have to do with their reincarnation as "Deep belief networks".

Although not displayed in this paper, we found some interesting topics such as "Kernel machines"[5] where ISI and NSF exhibit opposite trends of popularity. In addition, there are some topics that contain loadings in one of the corpora but not the other[6].

On the topic of "Security and Cryptography", ISI leads by 2 years, which effectively means that on an average, research papers are published 2 years earlier than the award of grants on this topic. On the topics of "Mobile computing" and "Data structures", award of grants and research publications happen almost simultaneously. On the topic of "Neural networks", on an average, NSF awards grants 3 years before papers are published, which perhaps means that NSF has had high confidence on this field. Among the 150 topics we get from LDA, there are 49 topics in which ISI leads NSF, 67 topics in which ISI lags NSF, and 34 topics having about the same pace in ISI and NSF.

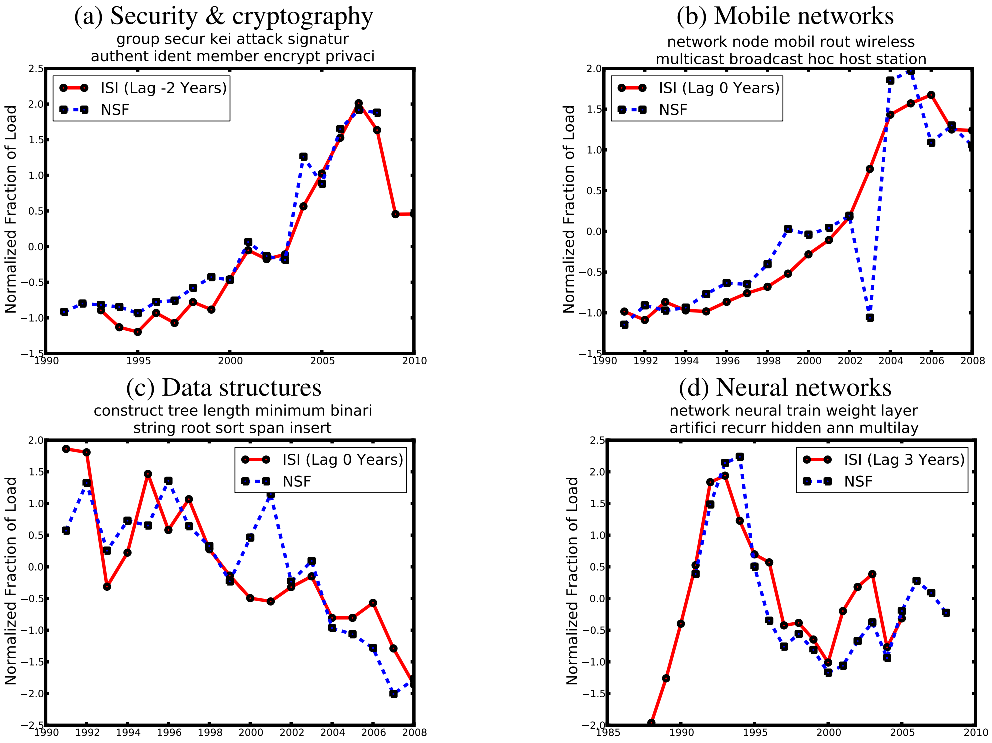
[2] http://www.isiknowledge.com

[3] http://www.nsf.gov

[4] http://www.cs.princeton.edu/ blei/lda-c/

[5] In this topic, we found that there is increasing loading with time in publications, but the opposite in grants!

[6] Topics that discuss past experience and post doctoral fellowships, etc. are specific to NSF.



(a) Security & cryptography

(b) Mobile networks

(c) Data structures

(d) Neural networks

Figure 2: Aligned time series plots and lead/lag estimates for four different CS topics. The words listed on the top of the curve are the top rankings words in the respective topics as found by LDA. The titles for the topics are manually assigned. The values next to the ISI legend in all the plots indicates the number of years ISI lags NSF by. Negative values indicate lead of ISI.

5 Conclusions and Future Work

In this paper, we present an LDA based approach for analyzing the topical lead-lag relationships across corpora with only textural and temporal information. There are other LDA extensions for temporal data such as [1, 6], but we decided to use basic LDA mainly because parallelization of LDA is well understood [4], and is therefore scalable, while it is less clear for the other models. Apart from scalability, the methodologies we adopted are standard and intuitive. Moreover, our approach is very robust: we find that even when we vary the number of topics in the LDA model, the lead/lag values for the same topic (which we matched by inspection) tally across the models.

In the future, we plan to evaluate the accuracy of our model by running it on corpora with known lead/lag patterns such as journal articles vs. conference proceedings. In addition, to verify the robustness of our results, we intend to use an alternative approach for lead/lag estimation such as similarity of word usage patterns within topics. Finally, we will also build a predictive model for lead/lag patterns between grants and publications based on the features of the corpora and the topics.

Acknowledgments

This research was supported in part by NSF grant NSF-0835614 CDI-Type II: What drives the dynamic creation of science? and in part by the Office of President, Stanford University. We wish to thank our anonymous reviewers and the members of the Stanford Mimir Project team for their insights and engagement.

References

[1] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[2] David Blei and Andrew Ng and Michael Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003.

[3] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, New York, NY, USA, 2009. ACM.

[4] Ramesh Nallapati, William Cohen, and John Lafferty. Parallelized variational EM for latent dirichlet allocation: An experimental evaluation of speed and scalability. In *ICDM workshop on high performance data mining*, 2007.

[5] Jie Tang and Jing Zhang. Modeling the evolution of associated data. *Data & Knowledge Engineering*, 69(9):965 – 978, 2010.

[6] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Conference on Knowledge Discovery and Data Mining*, 2006.