Understanding the evolution of science: analyzing evolving term co-occurrence graphs with spectral techniques

Zoltan Miklos, Mickaël Foursov, Franklin Lia, Ian Jeantet, David Gross-Amblard

Understanding the evolution of science: analyzing evolving term co-occurrence graphs with spectral techniques

Zoltán Miklós, Mickaël Foursov, Franklin Lia, Ian Jeantet, and David Gross-Amblard

Univ Rennes CNRS IRISA
zoltan.miklos@irisa.fr,
michael.foursov@irisa.fr,franklin.lia@etudiant.univ-rennes1.fr,
ian.jeantet@irisa.fr,david.gross-amblard@irisa.fr

Abstract. Given the high number of scientific papers that are published every year, it is a challenge to observe the evolution of scientific fields. While, for example, computer science and biology were considered rather unrelated 40 years ago, today, bioinformatics is a well-established field. One way to analyze these questions is to observe the evolution of the term co-occurrence graphs of the abstracts of scientific publications. In a term co-occurrence graph, two terms are connected if they appear together in an abstract of a publication. We weight the edges of this graph with the number of common occurrences. We analyze the evolution of this co-occurrence graph, that we constructed for each year, on the basis of a large collection of scientific articles, with the help of spectral techniques. We present our preliminary observations and discuss our ongoing work.

Keywords: term co-occurrence graph · evolving graph · spectral analysis

1 Introduction

Understanding the evolution of scientific fields, in particular in recent years, is a challenging task given the high number of publications. The availability of such large corpora allows one to analyze the articles with the help of text and data mining techniques. Our work also follows this line of research as we analyze the co-occurrence graph of the terms. On the basis of a corpus of scientific publications, we can construct a graph as follows. The nodes of the graph correspond to the terms and we consider that two terms are connected in the graph if they appear together in an abstract or in the title. We weight the graph edges with the number of common occurrences of the term pairs. We can construct a term co-occurrence graph for each year of publication and analyze this evolving graph.

Researchers have studied various aspects of the evolution. For example, Chavalarias et al. [5] describe the evolution of various scientific subfields. They obtain phylomemetic structures, which are directed acyclic graphs, that represent the

cartography of the evolution of different scientific fields. Dias et al. [7] define information theoretic similarity measures to describe and quantify the evolution of scientific fields.

In this short paper we try to understand the evolution from a different perspective. We try to characterize the evolution of the global structure of the (weighted) co-occurrence graph. Using graph signal processing techniques, we observe an interesting phenomenon: in various domains, the terms in scientific publications start with strongly connected clusters that are loosely connected among others. However, with time, these initially loosely-connected groups become more and more connected. Besides our preliminary results, we discuss some methodological aspects that need further elaboration.

The rest of the paper is organized as follows. In Section 2 we give some basic definitions of spectral graph theory. In Section 3 we elaborate on the concept of spectral plots, and on how to apply them to the analysis of evolving graphs. In Section 4 we describe our experiments. In Section 5 we discuss our observations and our plans for future work.

2 Preliminaries

One can associate matrices to graphs in different ways. The properties of these matrices are in close connection with the structure of the graph [6]. We are interested in analyzing the evolution of the edge weights and connections of a co-occurrence graph, so we based our analysis on the normalized Laplacian of the graph. The weighted adjacency matrix A of a graph $G(V, E)$ contains the edge weights that is $a_{ij} = w_{ij}$ where w_{ij} is the weight associated to the edge between the nodes i and j of the graph. The co-occurrence graphs are undirected and without self-loops. The degree matrix D of a graph is a diagonal matrix that contains the degree of for each vertex, that is $d_{i,j} = deg(v_i)$ if $i = j$ and $d_{i,j} = 0$ otherwise. The normalized Laplacian L of a graph is defined as follows.

$$L = I - D^{-1/2}AD^{-1/2} \tag{1}$$

The normalized Laplacian is a positive, semi-definite matrix, so all of its eigenvalues are positive. Moreover, for the normalized Laplacian (unlike for the algebraic Laplacian), all the eigenvalues are between 0 and 2, that is $\lambda_i \in [0, 2]$.

Spectral graph theory [6] is the field of research that tries to understand the connections between the spectrum of the Laplacian (and other related matrices) and the structure of the graph. The most well-known results state that the multiplicity of the 0 eigenvalue corresponds to the number of connected components of the graph. There are also a number of attempts to characterize graphs through spectral properties [8], [9], [10]: the i-th eigenvalue of the Laplacian gives insights about how well the graph admits a partitioning into i components.

3 Spectral plots for evolving graphs

Graph signal processing [14], [12] is an emerging field that analyzes the properties of signals defined on graphs. Analyzing the spectral density of a signal can give important insights and allow one to describe the frequency composition of this signal.

Spectral plots were introduced by Banerjee et al. [1], [2], [3]. They give a simple intuitive summary of the distribution of the eigenvalues of the normalized Laplacian. One can consider them as a possible way to define spectral densities over a graph Laplacian. As we recalled in Section 2, the eigenvalues of the normalized Laplacian fall in the interval $[0, 2]$. Banerjee et al [1] defined the following function $f : [0, 2] \to \mathbb{R}$ that they refer to as the spectral plot of the graph. We have chosen the parameter σ as the standard deviation of the elements of the normalized Laplacian.

$$f(x) = \sum_{\lambda_j} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{|x - \lambda_j|^2}{2\sigma^2}) \tag{2}$$

They also analyzed the generating processes that could potentially lead to a given graph that they wish to analyze [3]. A randomized generative process that applies node duplication can lead to spectral plots with a high peak at 1. The presence of many small eigenvalues indicates that there are a number of densely connected regions in the graph that are only loosely connected to the rest of the graph [2].

We propose to analyze the spectral plots for evolving graphs. Let $G_t = \{G_1, G_2, \ldots, \}$ be a time-varying graph: we assume that the graphs G_i, are defined on the same vertex set (that is $V(G_1) = V(G_2) = \cdots = V$) and the edges and their weights change over time. We can thus compute the normalized Laplacian L_t of the graph G_t and obtain the spectral plot $f_t(x)$, where $f_t(x)$ is a function over the interval $[0, 2]$ (that is $f_t(x) : [0, 2] \to \mathbb{R}$). The changes of the graph of the function $f_t(x)$ (as a function of t) can provide insights about the nature of evolution of the underlying evolving graph.

4 Experimental results

4.1 Co-occurrence graphs

To analyze to evolution of the co-occurrence graphs, we have collected data from the site arXiv [1]. In particular, we have collected the title, the abstract, the category (as defined on the site) and the creation date (of the first version) of all computer-science-related articles. If an article appears in multiple versions, we only consider the first version. The dataset contains a total of 203638 (unique) articles, published between 1990 and 2019, and it is grouped into 40 categories. Some categories contain more articles, the maximum is 31761 articles for the

[1] https://arxiv.org/

category cs.LG (Machine learning), while the smallest category (in terms of the number of articles) is cs.GL (General literature). The average size par category is 6947.

To construct the word co-occurrence graph, we applied standard NLP techniques, such as lemmatization (to associate the same graph node to the conjugated forms of the same word) and elimination of frequent stop-words (such as "the", "a", "and", etc).

4.2 Analysis of the evolution of spectral plots

We have constructed spectral plots for the co-occurrence graph for each category and for each year. We give some examples of the obtained spectral plots in the appendix. Figure 2 represents the evolution of spectral plots for the category Artificial Intelligence, while Figure 3 for category Software Engineering. For a better understanding, we present the logarithm of the spectral plots $log(1 + f_t(x))$, since the values of $f_t(x)$ have a high peak. The range of x axis is $[0, 2]$, as the eigenvalues of the normalized Laplacian fall to this interval. On the y axis we have different ranges for the values of the spectral plot, such that the figures have the same size.

We can observe that the spectral plots for consecutive years are rather similar. The co-occurrence graph constructed from the abstracts of consecutive years show rather similar structural properties. In this preliminary work we have not quantified this similarity, but we recall that there are several known methods for this purpose, for example, on e could use the spectral distance of graphs defined in [11].

The other observation is that in the course of the years, the changes of the spectral plot show a rather systematic direction. We can realize that the small eigenvalue components are slowly disappearing and the various eigenvalues close to 1 gain in importance. That is, the number of eigenvalues close to one is becoming higher and higher. That means that the use of the terminology of the considered fields is changing: from the previously disconnected use of terms we move towards a use where the terms are used in a more homogenous way. We also observe that the spectral plots are slightly asymmetric around 1, there are more values slightly below 1 than above.

We have also obtained the spectral plots for a number of synthetically generated graphs. For example, Figure 1 of the appendix represents the spectral plots that we constructed for Erdős-Rényi graphs that we generated with various parameters. The parameter p corresponds to the probability of presence of edges in the graph. High values of p (that is, close to 1), indicates that the graph is close to a clique.

5 Discussion and future work

We analyzed the evolution of scientific publications, on the basis of the evolving co-occurrence graphs of scientific publications. We presented some preliminary results and we plan to complete this initial analysis in a number of ways.

- We observed that the spectral plots of graphs of consecutive years are similar, however we did not quantify this similarity. We plan to compute the spectral distances of these graphs.
- One can observe a direction of the evolution of the spectral plots. We would like to describe mathematically these changes.
- We would like to give a more qualitative analysis of the evolution. On the basis of our preliminary results, we cannot say whether the evolution of the spectral plots is due to the higher number of papers or whether we can observe that the sub-fields of the scientific domains are more and more connected. To understand this question we also need to analyze the eigenvectors and not only the eigenvalues of the Laplacian. The eigenvectors are closely related to the clusters of the graph: they serve as the basic tool for spectral clustering [13] and graph embedding techniques, such as the Laplacian eigenmaps method [4]. Comparing the eigenvectors of consecutive years, and the corresponding spectral clusters of consecutive years could explain the merging or splitting of scientific sub-fields, at different scales.
- We also need specific techniques to select keywords of the domain. Such techniques are important to separate the changes related to the evolution of the scientific fields from other changes in language use.
- We plan to collect more data and also realize the experiments on other datasets. In particular, we plan to collect data from ISTEX [2] and from the SCOPUS [3] database.

Acknowledgements

The work was supported by CominLabs (https://www.cominlabs.u-bretagneloire.fr), through the ISNLP project (19YM310-03D) and also by the French National Research Agency, through the project ANR Epique (ANR-16-CE38-0002-01).

References

1. Anirban Banerjee and Jürgen Jost. Spectral plots and the representation and interpretation of biological data. *Theory in Biosciences*, 126(1):15–21, Mar 2007.
2. Anirban Banerjee and Jürgen Jost. *Spectral Characterization of Network Structures and Dynamics*, pages 117–132. Birkhäuser Boston, Boston, MA, 2009.
3. Anirban Banerjee and Jürgen Jost. On the spectrum of the normalized graph laplacian. *Linear Algebra and its Applications*, 428(11):3015 – 3022, 2008.
4. Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
5. David Chavalarias and Jean-Philippe Cointet. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLOS ONE*, 8(2):1–11, 02 2013.
6. F R K Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.
7. Laercio Dias, Martin Gerlach, Joachim Scharloth, and Eduardo Altmann. Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, 5, 06 2017.

[2] https://www.istex.fr/
[3] https://dev.elsevier.com/

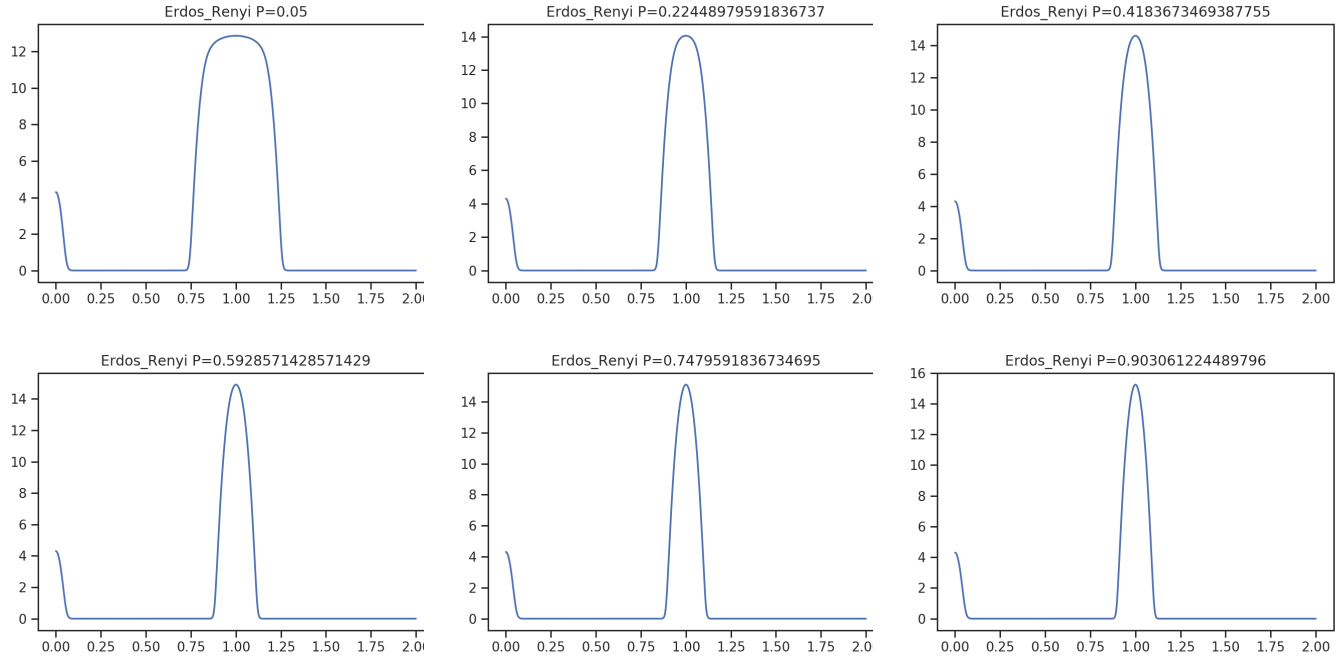

Fig. 1. Random Erdős-Rényi graphs with different parameters

8. Tsz Chiu Kwok, Lap Chi Lau, and Yin Tat Lee. Improved cheeger's inequality and analysis of local graph partitioning using vertex expansion and expansion profile. *SIAM J. Comput.*, 46(3):890–910, 2017.
9. James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *J. ACM*, 61(6):37:1–37:30, December 2014.
10. Anand Louis, Prasad Raghavendra, Prasad Tetali, and Santosh Vempala. Many sparse cuts via higher eigenvalues. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 1131–1140, New York, NY, USA, 2012. ACM.
11. Nicole Pearcy, Jonathan J. Crofts, and Nadia Chuzhanova. Network motif frequency vectors reveal evolving metabolic network organisation. *Mol. BioSyst.*, 11:77–85, 2015.
12. A. Sandryhaila and J. M. F. Moura. Discrete signal processing on graphs. *IEEE Transactions on Signal Processing*, 61(7):1644–1656, April 2013.
13. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.
14. D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013.

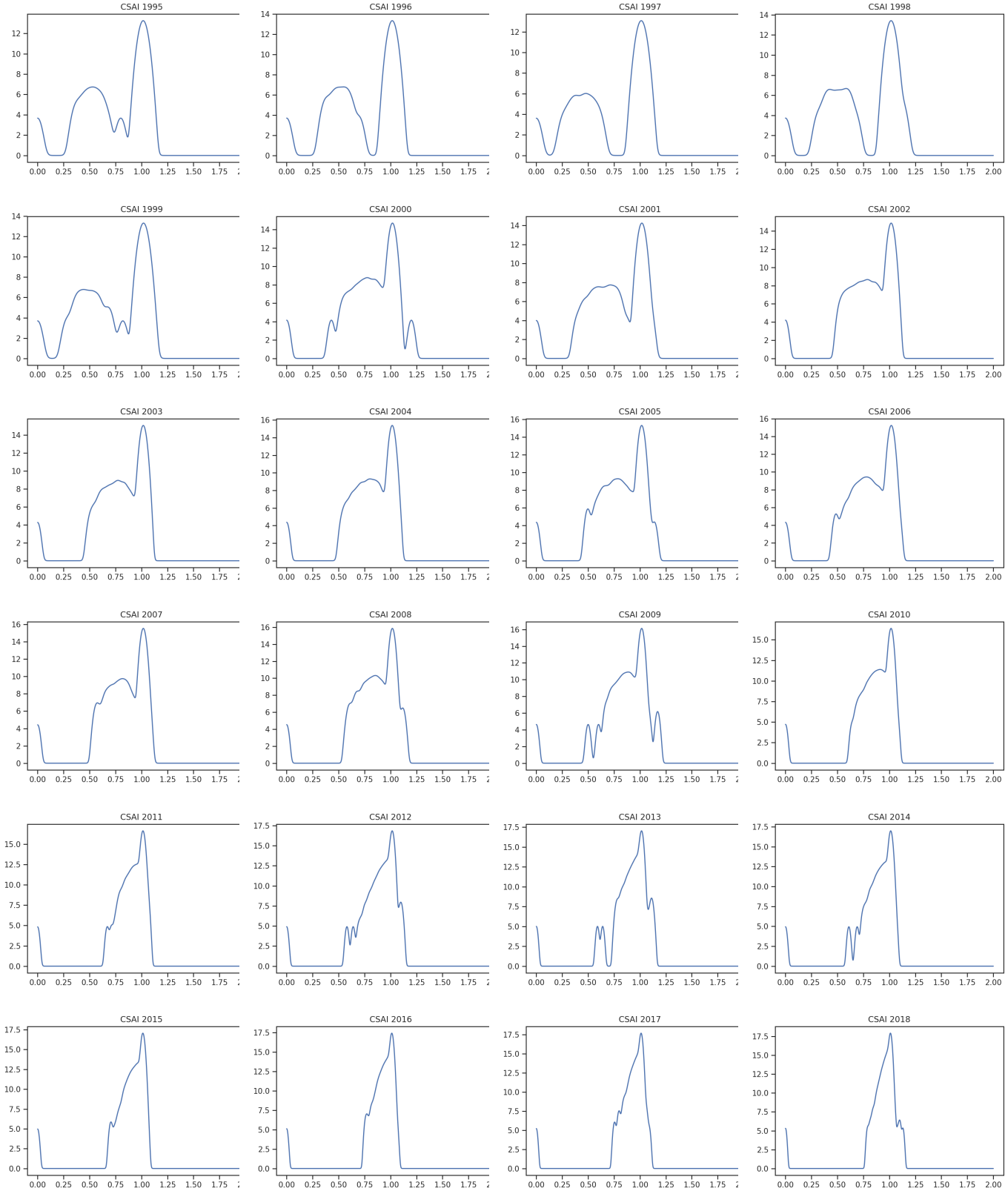

Fig. 2. Spectral plots for the arXiv category cs.AI

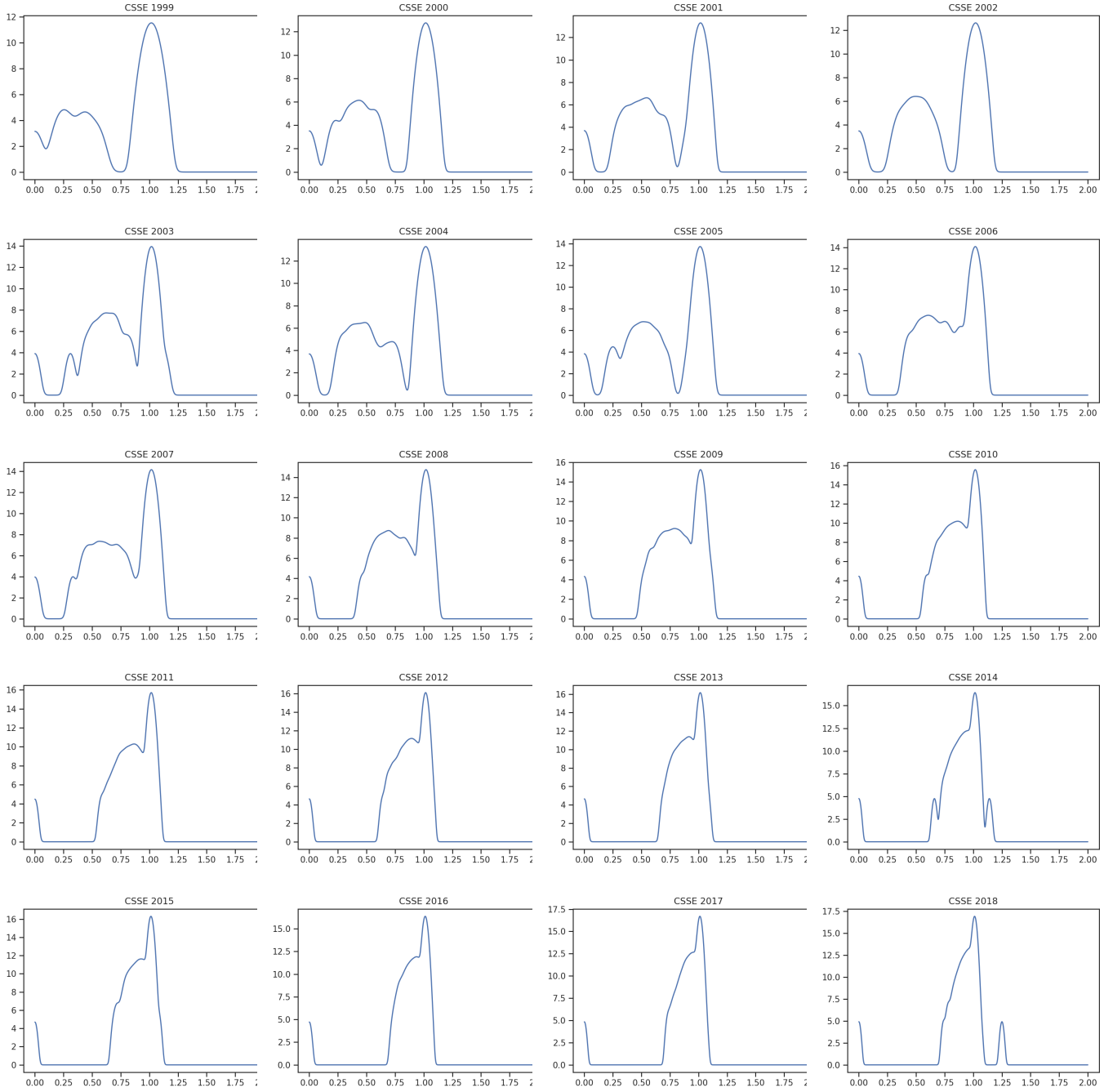

Fig. 3. Spectral plots for the arXiv category cs.SE