Mieczyslaw A. Klopotek
ORCID: 0000-0003-4685-7045

Slawomir T. Wierzchon
ORCID: 0000-0001-8860-392X

Bartlomiej Starosta
ORCID: 0000-0002-5554-4596

Dariusz Czerski
ORCID: 0000-0002-3013-3483

Piotr Borkowski
ORCID: 0000-0001-9188-5147

Institute of Computer Science of Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

{klopotek,stw,barstar,dcz,p.borkowski}@ipipan.waw.pl

Dependence of Spectrogram from Graph Spectral Clustering in Text Document Domain on Word Distribution Models (Extended Abstract)

Abstract. Based on our earlier studies, we hypothesize that the shape of the spectrogram of a Laplacian of similarity matrix, used in Graph Spectral Clustering, could be attributed to writing style of the authors of the document group in the cluster. We investigate this hypothesis for a couple of models of word distributions.

Keywords: Explainable AI, Graph Spectral Clustering, Eigenvalue Spectrum of A Laplacian, Artificial Text Generation from Simple Language Model.

1 Introduction

This work aims to extend our previous studies on the specific shapes of combinatorial Laplace spectrograms, to unravel the hidden nature of graph spectral analysis (GSA) methods.

Despite its effectiveness, the "black box" nature of GSA makes companies reluctant to use it because the analysis results are expressed in terms of vectors and eigenvalues [6, 10]. This situation prompted the emergence of the so-called Explainable Artificial Intelligence (XAI) [1].

Previous research in GSA focused on exploring a few eigenvalues and eigenvectors [9]. However, we discovered that one could also use the full eigenvector spectrogram of eigenvalues [2]. This spectrum proved to be sufficient for classification [2], incremental clustering [4], hashtag explanation [8], and others. However, the question of why the aforementioned application areas benefited from eigenvalue spectrograms remained open. We hypothesize that the characterization of clusters/classes via spectrograms is possible due to the specific "style" of writing. The investigation, outlined in detail in Section 3 is an extension of the work in this direction presented in [5] by exploring another theory of word distribution in documents, namely the lognormal distribution. The experimental results are presented in Section 4. Section 2 overviews related work, while Section 5 presents our conclusions and outlines future resear ch.

2 Related Work

GSA in the clustering domain, explanation of which we want to contribute to here, is typically carried out using relaxations of ratio cut (RCut) graph clustering techniques. A similarity matrix is transformed to its Laplacian, for which the matrix of its eigenvectors is computed. A column submatrix linked to the k lowest eigenvalues of the related graph Laplacian is used as graph embedding, and rows of which are subjected to the k-means method [9]. For a similarity matrix S between pairs of items (e.g. documents), a combinatorial Laplacian L is defined as

$$L(S) = T(S) - S, \tag{1}$$

where $T(S)$ is the diagonal matrix with $t_{jj} = \sum_{k=1}^{n} s_{jk}$ for each $j \in [n]$.

The RCut clustering criterion itself means splitting a graph into parts in such a way that for each cluster, the average weight of links leading outside of a cluster is the lowest. Formally, RCut aims at finding the partition matrix $P_{RCut} \in \mathbb{R}^{n\times k}$ minimizing the formula $H'LH$ over the set of all partition matrices $H \in \mathbb{R}^{n\times k}$. This problem is NP-hard. GSC relaxes it by permitting that H is a column orthogonal matrix without further constraints. Then the solution is simple: the columns of P_{RCut} are eigenvectors of L corresponding to the k smallest eigenvalues of L. Further details can be found in e.g. [9].

The cosine similarity between the documents' bag-of-words representations is typically used to calculate the similarity matrix S between textual texts. (see e.g. [9]). Therefore, in this simulation study, we use models of word distribution in order to generate artificial documents. One of the earliest proposals of word distribution functions was so-called Zipf law [11], generalized in a number of ways, including the Mandelbrot version [7], where the

word distribution is proportional to:

$$Prob(w_i; \alpha, b) = \frac{\frac{1}{(i+b)^\alpha}}{\sum_{\ell=1}^{n_w} \frac{1}{(\ell+b)^\alpha}} \tag{2}$$

In this formula b plays the role of a distribution shift parameter, usually $\alpha \approx 1$ and $b \approx 2.7$. i ranges from 1 to n_w, where n_w is the number of words in the dictionary, and α is a parameter, usually set to 1. We investigated this distribution type in a previous paper [5].

In this paper, we focus on the quite popular competing lognormal model [3]. The log-normal distribution is defined as follows: Given a standard normal variable Z, and two real variables μ, σ,the latter being positive real, the distribution of the random variable

$$Prob(w_i; \mu, \sigma) = \frac{1}{w_i \sigma \sqrt{2\pi}} e^{-\frac{(\log(w_i)-\mu)^2}{2\sigma^2}} \tag{3}$$

is called log-normal distribution.

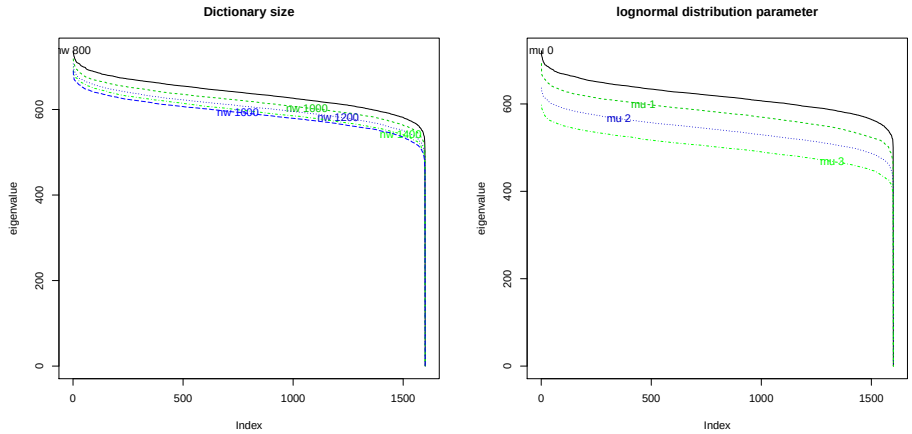

Figure 1. Spectrogram dependence, for artificial data generated, left: on number of words in the dictionary, right: on μ parameter

3 Experimental Settings

The goal of the study was to see if a generative model of synthetic texts may resemble actual ones using a predetermined parameterized word distribution. That is if changes in various text style elements impact shape of the spectrograms in such a way that spectrograms differentiate the style. A generator was developed that generates artificial papers using a bag of dictionary terms sampled based on certain word distribution criteria and other document attributes. The appropriate combinatorial Laplacian spectra are examined and document similarity matrices

are calculated for every set of created documents. To ascertain their impact, the parameters are changed one at a time. We investigated the log-normal model (formula (3)) and checked the following parameters: n_w - dictionary size, μ - log-normal distribution parameter (mean), σ - log-normal distribution parameter (standard deviation), $doclen0$ - document basic length.

In the experiments, one parameter was changed at a time, while the other ones were kept at default level. Default parameters were: $n_w = 1000$, $\mu = 0$, $\sigma = 4$, $doclen0 = 60$, Table 1 lists the parameter value ranges used in the experiment. In each run, 1600 documents were created.

Parameter	**Value Range**
n_w	{800,1000,1200,1400,1600}
μ	$\{0, 1, 2, 3\}$
σ	{4,5,6,7,8,9,10}
$doclen0$	{ 30,60,120,240,480}

Table 1. Ranges of parameters used in the experiments. %

4 Results of Experiments

The impacts of individual parameters on the spectrogram are presented in Figures 1 (n_w, μ) and 2 ($doclen0$, σ), An increase of n_w (dictionary size) appears to move the spectrogram

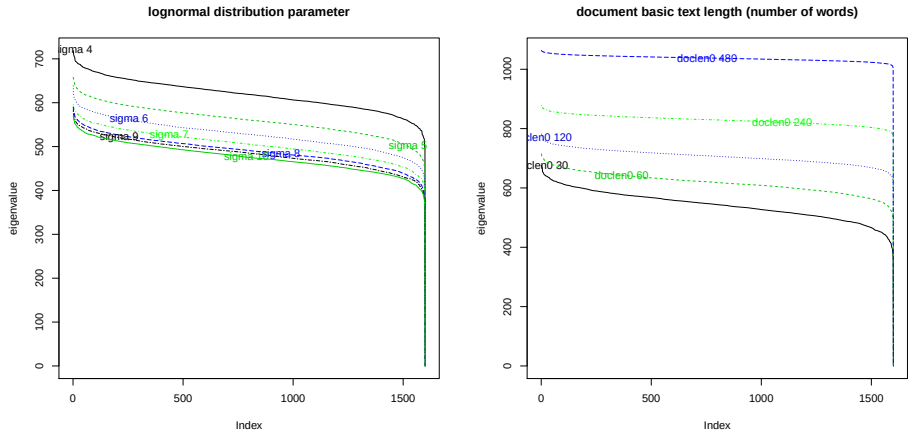

Figure 2. Spectrogram dependence, for artificial data generated, left: on σ parameter, right: on document length.

downwards. An increase of μ (distribution parameter) appears to move the spectrogram downwards. An increase of σ (distribution parameter) appears to move the spectrogram downwards. Only n increase of $doclen0$ (document length) seems to move the spectrogram upwards.

5 Conclusions

We have studied the dependence of spectrograms of combinatorial Laplacian on several parameters of document collections generated artificially from widely accepted log-normal models of word distributions. All the generator parameters appear to impact the spectrogram shape, confirming our hypothesis that the writing style io responsible for the capability to discern between clusters/classes of textual documents via Graph Spectral Analysis.

The presented research results can be used as a basis for studies on document group similarity or collective authorship, as well as experiments with synthetic data on the utility of Laplacian eigenvalue spectra for Graph Spectral Analysis based clustering, incremental clustering, and document classification. They can also enrich explanations of the results of traditional spectral clustering, if an interpretation of log-normal distribution parameters is found.

References

1. Barredo Arrieta, A.e.a.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82 – 115 (2020), https://doi.org/10.1016/j.inffus.2019.12.012
2. Borkowski, P., Kłopotek, M., Starosta, B., Wierzchoń, S., Sydow, M.: Eigenvalue based spectral classification. PLOS ONE **18**(4), e0283413 (2023), https://doi.org/10.1371/journal.pone.0283413
3. Carroll, J.: On sampling from a lognormal model of word frequency distribution. In: Kurera, H., Francis, W. (eds.) Computational Analysis of Present-Day American English, pp. 406–424. Providence: Brown University Press (1967)
4. Kłopotek, M.A., Starosta, B., Wierzchoń, S.T.: Eigenvalue-based incremental spectral clustering. arXive 2308.10999 (2023)
5. Kłopotek, M.A., Wierzchoń, S.T., Starosta, B., Czerski, D., Borkowski, P.: Towards explaining the spectrogram of graph spectral clustering in text document domain, to appear in Proc. CISIM2024 Conference, Białystok, 27-29.9.2024ad. (2024)
6. Macgregor, P., Sun, H.: A tighter analysis of spectral clustering, and beyond. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 14717–14742. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/macgregor22a.html
7. Mandelbrot, B.: An informational theory of the statistical structure of languages. In: Jackson, W. (ed.) Communication Theory, pp. 486–502. Academic Press, Princeton (1953)
8. Starosta, B., Kłopotek, M., Wierzchoń, S.: Hashtag similarity based on laplacian eigenvalue spectrum. In: Proc. PP-RAI'2023 - 4th Polish Conference on Artificial Intelligence , Progress in Polish Artificial Intelligence Research 4, Łódź, Poland 2023 (2023)
9. Wierzchoń, S., Kłopotek, M.: Modern Clustering Algorithms, Studies in Big Data, vol. 34. Springer Verlag (2018)
10. Xu, Y., Srinivasan, A., Xue, L.: A Selective Overview of Recent Advances in Spectral Clustering and Their Applications, pp. 247–277. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72437-5_12
11. Zipf, G.: Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA (1932)