Mieczysław A. Kłopotek[1]
ORCID: 0000-0003-4685-7045

Sławomir T. Wierzchoń[2]
ORCID: 0000-0001-8860-392X

Bartłomiej Starosta[3]
ORCID: 0000-0002-5554-4596

Dariusz Czerski[4]
ORCID: 0000-0002-3013-3483

Piotr Borkowski[5]
ORCID: 0000-0001-9188-5147

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

{[1]klopotek,[2]stw,[3]barstar,[4]dcz,[5]p.borkowski}@ipipan.waw.pl

Dependence of Spectrogram from Graph Spectral Clustering in Text Document Domain on Word Distribution Models

DOI: 10.34739/si.2024.31.01

Abstract. Analysis of the shape of a Laplacian spectrogram is a new line of research used in graph spectral clustering. More precisely, we observed that (properly normalized) plots of the eigenvalues of sub-Laplacians characterizing different groups of documents differ in their shape. Thus, by computing the distance between these plots, we can solve the problem of clustering and classifying new observations. This idea is developed in a

number of our papers and as such, can be considered a pioneering approach to cluster analysis. In an attempt to answer why it is so useful, in this paper we consider the hypothesis that the shape of a spectrogram could be attributed to the writing style of the authors of the document group in the cluster. We explore this hypothesis for several models of word distribution. In particular, we assume that the writing style is reflected in the word distribution of texts of an author or a group of them. We check if changing of distribution parameters of a widely accepted log-normal word distribution model changes in fact the Laplacian eigenvalue spectrogram in such a way as to distinguish between document groups. We found that in fact variation of each of the distribution parameters leads to distinct groups of documents. These findings justify the usage of Laplacian spectrograms to distinguish (cluster or classify) groups of documents.

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 investigations for reasons of specific shapes of Laplacian combinatorial spectrograms. The present work aims to expand our previous research due to the observed specific shapes of Laplacian combinatorial spectrograms. This should give an insight into understanding the results of Graph Spectral Clustering and other spectral clustering methods [13, 6, 7]. These are commonly treated as one of the most effective in grouping similar data points in graph data models. In the case of text document clustering, the underlying graph represents similarities between the vectors generated from texts or documents. As there is no unique spectral clustering algorithm, the authors of [18] compiled 16 such algorithms and compared their computational complexities. Our approach is a new and unique tool that fits into the framework of Graph Spectral Clustering. It is intended to contribute to the general field of Explainable Artificial Intelligence.

The "Black-box" nature of many AI methods, despite their efficacy, makes businesses reluctant to use them. This is especially affecting Graph Spectral Analysis (GSA) methods in which the analysis results are expressed in terms of eigenvectors and eigenvalues [13, 14, 28]. This situation led to the emergence of the so-called Explainable Artificial Intelligence (XAI) [2].

Earlier research in GSA concentrated on exploring a few eigenvalues and eigenvectors [13]. However, we discovered that one can also take advantage of the full eigenvector spectrogram of eigenvalues [4]. This spectrum proved to be sufficient when performing classification [4], incremental clustering [10], hashtag explanation [25], and other.

However the research question, of why the aforementioned application areas benefited from the eigenvalue spectrogram, remained open. We pursued the hypothesis that the possibility to characterize clusters/class via spectrograms can be attributed 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 [8] by exploring other theory of word distribution in documents, that is the lognormal distribution.

The experimental results are presented in Section 4. Section 2 summarizes related work, while Section 5 presents our conclusions and outlines future research.

2 Related Work

Graph spectral clustering (GSC) is typically performed 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 the graph embedding, and rows of which are subjected to the k-means method [13]. 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 to find the partition matrix $P_{RCut} \in \mathbb{R}^{n \times k}$ as the minimizer of 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. [13] or [27].

It is a mathematical property of combinatorial Laplacian that its eigenvalues are always non-negative, while the lowest one is always equal to zero.

The cosine similarity between the documents' bag-of-words representations is typically used to calculate the similarity between textual texts. (see e.g. [27]). Therefore, in this simulation study, we use models of word distribution to generate artificial documents. One of the earliest proposals of word distribution functions was the so-called Zipf law [29]. It assumes that the probability of occurrence of a word w_i amounts to

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

where 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. It was generalized in many ways, including the Mandelbrot version [15], where the word distribution is proportional to:

$$Prob(w_i; \alpha, b) = \frac{\frac{1}{(i+b)^\alpha}}{\sum_{\ell=1}^{n_w} \frac{1}{(\ell+b)^\alpha}} \tag{3}$$

In this formula b plays the role of a distribution shift parameter, usually $\alpha \approx 1$ and $b \approx 2.7$. We investigated this distribution type in a previous paper [8]. Other proposals of Zipf law extensions include the one of Parker-Rhodes and Joyce [22] and of Orlov [20]. Sichel proposed a generalized inverse Gauss-Poisson law [24]. Its limiting distribution was studied by [3]. The fitting of various models to real distribution of words in natural language texts was studied e.g. in [1]. An interesting queuing-theory model of word frequency distributions was developed by Munro in [19]. Özbey et al. [21] proposed a framework to explain the nonlinear behavior of low frequencies on the log-log scale. Zipf law was widely studied and one of the interesting discoveries, according top Li, is that randomly generated texts exhibit Zipf-like distribution

of words [11]. Note that recently, there are some researchers questioning the Zipf law as such claiming that word frequency distribution is rather a result of mixture of distributions of various concept categories that produces the impression of the Zipf law [23].

In this paper, we focus on the quite popular competing lognormal model [5], which finds various applications in natural language processing [17, 26, 16]. 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

$$Prob(w_i; \mu, \sigma) = \frac{1}{i\sigma\sqrt{2\pi}} e^{-\frac{(\log(i)-\mu)^2}{2\sigma^2}} \tag{4}$$

is called log-normal distribution. It states what is the probability of occurrence of the i^{th} word w_i.

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 (4)) and checked the following parameters:

- n_w - dictionary size
- μ - log-normal distribution parameter,
- σ - log-normal distribution parameter,
- $doclen0$ - document basic length

In the experiments, one parameter was changed at a time, while the other ones were kept at the default level. Default parameters were: $n_w = 1000$, $\mu = 0$, $\sigma = 4$, $doclen0 = 60$, Table 3 lists the parameter value ranges used in the experiment.

Table 1. Ranges of parameters used in the experiments.

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}

In each run, 1600 documents were created.

4 Results of Experiments

The impacts of individual parameters on the Laplacian spectrogram are presented in Figures 1 - 4. In each case, the figure of the left hand side shows the spectrograms for various parameter values, while the right one depicts the behavior of the highest eigenvalue (the black line) and of the spectrum average (the blue line) of the spectrogram depending on parameter value (averaged over the experiments that were performed). The latter allows to see the impact of the distribution parameter - in which direction the spectrogram is shifted.

In particular, Figure 1 illustrates the dependence of the spectrogram on the number of words in the dictionary for artificially generated data. The larger the dictionary, the lower the obtained eigenvalues.

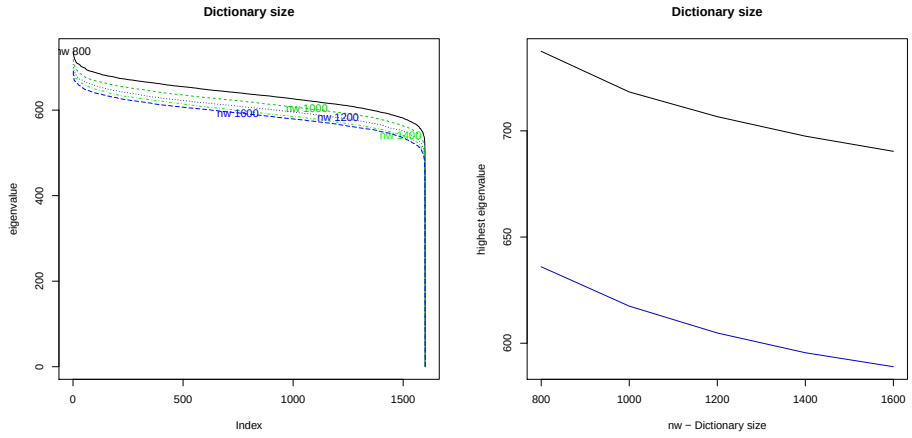

Figure 1. Spectrogram dependence on number of words in the dictionary for artificially generated data.

Figure 2 shows the dependence of the spectrogram obtained for the data generated according to the log-normal model with varying values of the μ parameter.

Again, the increase of this parameter decreases the eigenvalues of the Laplacian spectrogram. In Figure 3 we illustrate how the spectrograms depend on the σ parameter in this model.

We see the same impact as previously. Lastly, Figure 4 shows how these spectrograms depend on the length of the artificially generated data. Here we see that the longer the documents, the larger are the eigenvalues of the Laplacian spectrogram.

To summarize, increasing n_w (dictionary size), μ and σ seems to move the spectrogram downwards. Only increasing the document length, $doclen0$, seems to move the spectrogram upwards.

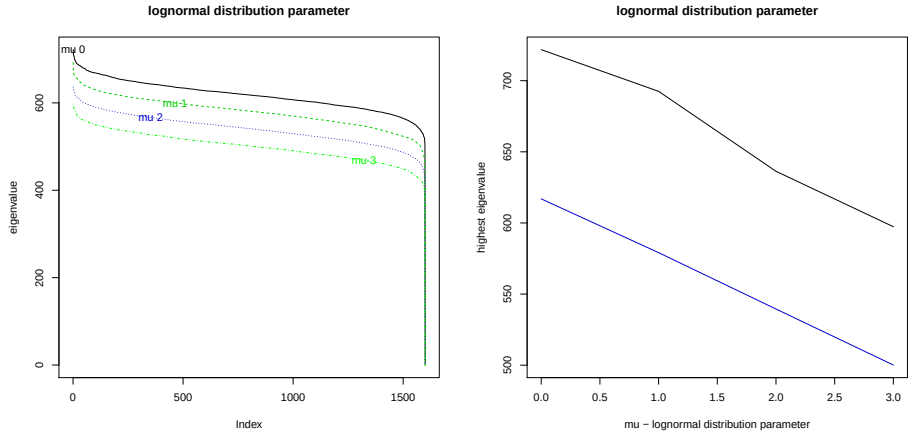

Figure 2. Spectrogram dependence on μ parameter for artificially generated data

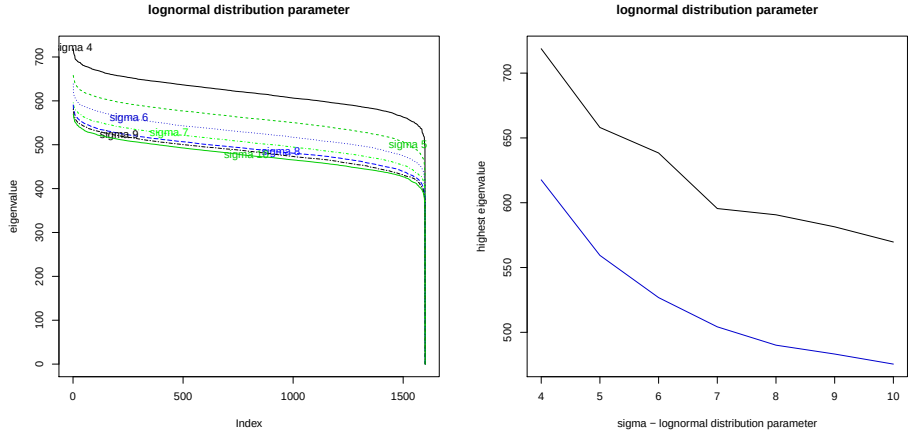

Figure 3. Spectrogram dependence on σ parameter for artificially generated data.

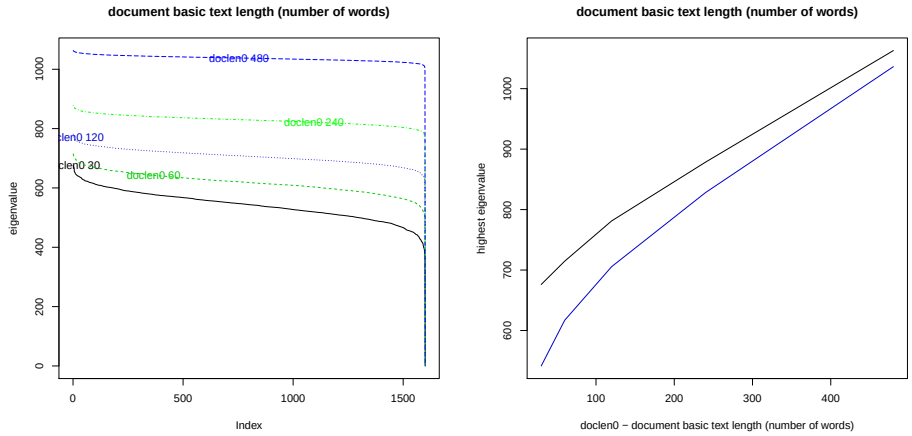

Figure 4. Spectrogram dependence on document length for artificially generated data

5 Conclusions

Our earlier research [4, 10, 25] indicated the capability of Laplacian spectrograms to distinguish between groups of documents. Notably, neither eigenvalues nor eigenvectors can be directly linked to document contents. Therefore we suspected that the discerning capability must be linked to different features of the document collections. We suspected that possibly the authorship/writing style of people engaged in posting messages related to a given hashtag are of importance. As the authorship information is not available for tweets we studied, we decided to take a different route. We have studied the dependence of spectrograms of the combinatorial Laplacian on several parameters of document collections artificially generated from widely accepted log-normal models of word distributions, assuming that the writing style would affect the distribution of words a group of people uses. All the generator parameters appear to impact the spectrogram shape, confirming our hypothesis that the writing style is 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[1], 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. If an interpretation of the log-normal distribution parameters can be found, it may also enrich explanations of the results of traditional spectral clustering.

One has to remark that though word distribution generator models are used in the NLP studies, these models do not quite reflect the real writing style behavior as e.g. interactions

[1] Some studies suggest [9] that word length and frequency affect readability and speed, which may somehow be due to writing style.

between words are neglected, while they may have an effect (see.g. [12]). This is surely also a limiting point iof this study.

Further research is required to determine the relative importance of individual generator parameters in terms of their effect on the shape of the spectrogram and the effects of their interaction.

References

1. Baayen, R.H.: Statistical models for word frequency distributions: A linguistic evaluation. Comput. Humanit. **26**(5-6), 347–363 (1992). https://doi.org/10.1007/BF00136980, https://doi.org/10.1007/BF00136980
2. 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
3. Bogachev, L.V., Nuermaimaiti, R., Voss, J.: Limit shape of the generalized inverse Gaussian-Poisson distribution. arXiv 2303.08139 (2023), https://arxiv.org/abs/2303.08139
4. 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
5. 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)
6. Chaudhuri, K., Chung, F., Tsiatas, A.: Spectral clustering of graphs with general degrees in the extended planted partition model. In: Mannor, S., Srebro, N., Williamson, R.C. (eds.) Proceedings of the 25th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 23, pp. 35.1 – 35.23. PMLR, Edinburgh, Scotland (25 - 27 Jun 2012), https://proceedings.mlr.press/v23/chaudhuri12.html
7. Dudek, A.: Classification via spectral clustering. Acta Universitas Lodziensis **135**, 121–130 (2021), https://dspace.uni.lodz.pl/xmlui/bitstream/handle/11089/344/121-130.pdf?sequence=1
8. Kłopotek, M.A., Wierzchoń, S.T., Starosta, B., Czerski, D., Borkowski, P.: Towards explaining the spectrogram of graph spectral clustering in text document domain. In: Computer Information Systems and Industrial Management: 23rd International Conference, CISIM 2024, Bialystok, Poland, September 27–29, 2024, Proceedings. p. 372–386. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-71115-2_26, https://doi.org/10.1007/978-3-031-71115-2_26
9. Kuperman, V., Schroeder, S., Gnetov, D.: Word length and frequency effects on text reading are highly similar in 12 alphabetic languages. Journal of Memory and Language **135**, 104497 (2024). https://doi.org/https://doi.org/10.1016/j.jml.2023.104497, https://www.sciencedirect.com/science/article/pii/S0749596X23000967
10. Kłopotek, M.A., Starosta, B., Wierzchoń, S.T.: Eigenvalue-based incremental spectral clustering. Journal of Artificial Intelligence and Soft Computing Research **14**(2), 157–169 (2024). https://doi.org/10.2478/jaiscr-2024-0009, https://doi.org/10.2478/jaiscr-2024-0009
11. Li, W.: Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory **38**(6), 1842–1845 (1992). https://doi.org/10.1109/18.165464
12. Linke, M., Ramscar, M.: How the probabilistic structure of grammatical context shapes speech. Entropy **22** (2020), https://api.semanticscholar.org/CorpusID:211089876
13. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)

14. 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
15. Mandelbrot, B.: An informational theory of the statistical structure of languages. In: Jackson, W. (ed.) Communication Theory, pp. 486–502. Academic Press, Princeton (1953)
16. Matricciani, E.: Multi–dimensional data analysis of deep language in j.r.r. tolkien and c.s. lewis reveals tight mathematical connections. AppliedMath **4**(3), 927–949 (2024). https://doi.org/10.3390/appliedmath4030050, https://www.mdpi.com/2673-9909/4/3/50
17. Matsubara, Y.: Fluctuations in the email size modeled by a log-normal-like distribution (2025), https://arxiv.org/abs/2501.04042
18. Mondal, R., Ignatova, E., Walke, D., Broneske, D., Saake, G., Heyer, R.: Clustering graph data: the roadmap to spectral techniques. Discov Artif Intell **4**(7) (2024), https://doi.org/10.1007/s44163-024-00102-x
19. Munro, R.: A queueing-theory model of word frequency distributions. In: Bow, C., Hughes, B. (eds.) Proceedings of the Australasian Language Technology Workshop, ALTA 2003, Melbourne, Australia, December 8-12, 2003. pp. 70–77. Australasian Language Technology Association (2003), https://aclanthology.org/U03-1009/
20. Orlov, J., Chitashvili, R.: On the distribution of frequency spectrum in small samples from populations with a large number of events. Bulletin of the Academy of Sciences, Georgia **108.2**, 297–300 (1982)
21. Özbey, C., Çolakoğlu, T., Bilici, M.c., Erkuş;, E.C.: A unified formulation for the frequency distribution of word frequencies using the inverse Zipf's law. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1776–1780. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591942, https://doi.org/10.1145/3539618.3591942
22. Parker-Rhodes, A.F., Joyce, T.: A theory of word-frequency distribution. Nature **178**, 1308 (1956). https://doi.org/10.1038/1781308a0
23. Ramscar, M.: The empirical structure of word frequency distributions. arXiv 2001.05292 (2020), https://arxiv.org/abs/2001.05292
24. Sichel, H.: On a distribution law for word frequencies. Journal of the American Statistical Association **70**, 542–547 (1975)
25. 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)
26. Torre, I., Luque, B., Lacasa, L., Kello, C., Hernández-Fernández: On the physical origin of linguistic laws and lognormality in speech. R. Soc. Open Sci. **6, 191023** (2019). https://doi.org/10.1098/rsos.191023
27. Wierzchoń, S., Kłopotek, M.: Modern Clustering Algorithms, Studies in Big Data, vol. 34. Springer Verlag (2018)
28. 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
29. Zipf, G.: Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA (1932)