Entropic Graph Spectrum

Diego Granziol, Binxin Ru, Xiaowen Dong, Stefan Zohren, Michael Osborne, Stephen Roberts
diego@robots.ox.ac.uk, robin@robots.ox.ac.uk
Department of Engineering Science, University of Oxford

Abstract

Graph spectral techniques for measuring graph similarity, or for learning the cluster number, require kernel smoothing. The choice of kernel function and bandwidth are typically chosen in an ad-hoc manner and heavily affect the resulting output. We prove that kernel smoothing biases the moments of the spectral density. We propose an information theoretically optimal approach to learn a smooth graph spectral density, Entropic Graph Spectrum(EGS), which fully respects the moment information. Our method's computational cost is linear in the number of edges, and hence can be applied to large networks, with millions of nodes. We apply our method to the problems to graph similarity and cluster number learning, where we outperform comparable iterative spectral approaches on synthetic and real graphs.

1 Introduction

Many systems of interest can be naturally characterised by complex networks; examples include social networks [12, 8, 10], biological networks [15] and technological networks. Of crucial importance to the understanding of the properties of a network or graph is its spectrum, which is defined as the eigenvalues of its adjacency or Laplacian matrix [6, 4]. The spectrum of a graph can be considered as a natural set of graph invariants and has been extensively studied in the fields of chemistry, physics and mathematics [3]. Spectral techniques have been extensively used to characterise the global network structure [14] and in practical applications thereof, such as clustering [18] and measuring graph similarity [16]. A major limitation in utilizing graph spectra to solve problems such as graph similarity and estimating the number of clusters is the inability to automatically and consistently learn an everywhere-positive non-singular approximation to the spectral density. Our main contributions of this paper are: (1) We prove that the method of kernel smoothing, commonly used in methods to visualize and compare graph spectral densities, biases moment information; (2) We propose a computationally efficient and information theoretically optimal spectral density approximation, Entropic Graph Spectrum(EGS), which fully respects the moment information; (3) We demonstrate the effectiveness of EGS on graph similarity comparison and clusters number estimation.

2 Motivations for A New Graph Spectral Learning Method

For large sparse graphs, with millions or billions of nodes, learning the exact spectrum using eigen-decomposition is unfeasible due to the $\mathcal{O}(n^3)$ cost. Powerful iterative methods, such as the Lanczos algorithm, are then proposed as cost-efficient alternatives to approximate the graph spectrum with a sum of weighted Dirac delta functions closely matching the first m moments (explained in Appendix B) of the spectral density [17]:

$$p(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\delta(\lambda-\lambda_i) \approx \sum_{i=1}^{m} w_i\delta(\lambda-\lambda_i), \tag{1}$$

where $\sum_{i=1}^{m} w_i = 1$, and λ_i denotes the i-th eigenvalue in the spectrum. However, such an approximation is undesirable because natural divergence measures between densities, such

Preprint. Under review.

as the relative entropy $\mathcal{D}_{\mathrm{KL}}(p||q) \in (0,\infty)$ from the fields of information theory [5] [1] (i.e. $\mathcal{D}_{\mathrm{KL}}(p||q) = \int p(\lambda)\log\frac{p(\lambda)}{q(\lambda)}d\lambda$) is infinite for densities that are mutually singular. The use of the Jensen-Shannon divergence simply re-scales the divergence into $\mathcal{D}_{\mathrm{JS}}(p||q) \in (0,1)$. This puts us in the counter-intuitive scenarios, such as, an infinite (or maximal) divergence between two graphs with the removal of a single edge/node or two graphs generated using the same random graph model with identical hyper-parameters.

To alleviate the above-mentioed limitations, practitioners typically generate a smoothed spectral density by convolving the Dirac mixture with a smooth kernel [16, 2], typically Gaussian or Cauchy, which heavily affects the resolution and moment information of the underlying spectra. We show this with the following theorem, which we prove in Appendix C .

Theorem 1. *The m-th moment of a Dirac mixture $\sum_{i=1}^{n} w_i\delta(\lambda-\lambda_i)$, which is smoothed by a kernel k_σ supported on the real line with bounded moments, is perturbed from its unsmoothed counterpart by an amount $\sum_{i=1}^{n} w_i \sum_{j=1}^{r/2}\binom{r}{2j}\mathbb{E}_{k_\sigma(\lambda)}(\lambda^{2j})\lambda_i^{m-2j}$, where $r=m$ if m is even and $m-1$ otherwise. $\mathbb{E}_{k_\sigma(\lambda)}(\lambda^{2j})$ denotes the $2j$-th central moment of the kernel function $k_\sigma(\lambda)$*

Remark. This proves that kernel smoothing alters moment information and that this process gets more pronounced for higher moments. Furthermore, given that $w_i > 0$, $\mathbb{E}_{k_\sigma(\lambda)}(\lambda^{2j}) > 0$ and, for the normalised Laplacian $\lambda_i > 0$, the corrective term is manifestly positive and so the smoothed moment estimates are biased. For large random graphs, the moments of a generated instance converge to those averaged over many instances [7], hence by biasing our moment information we limit our ability to learn about the underlying stochastic process. We include a fuller discussion about the relationship between the moments of the graph and the underlying stochastic process in the Appendix D.

3 Entropic Graph Spectrum (EGS)

We present an information theoretic approach to learning a maxmially smooth spectral approximation for large graphs, EGS, by combining stochastic trace estimation (the multiplication of a matrix by a vector to learn the moment information of the underlying spectral density, explained in Appendix E), with the method of Maximum Entropy (MaxEnt) [9]. Our algorithm is summarized in Algorithm 1 and the modelling power of our method is visualized in Appendix G where we test our algorithm on examples with known limiting spectral density. We empirically show the superiority of our EGS approach in two example applications in Section 4 and 5 with more results in Appendix H.

4 EGS for Measuring Graph Similarity

In this section, we test the use of our EGS in combination with symmetric KL divergence [1] to measure similarity between different types of synthetic and real world graphs.

Inferring parameters of random graph models We generate a random graph of a given size and parameter value (e.g., $n=50, p=0.6$) and learn its EGS characterisation. Then, we generate another graph of the same size but learn its parameter value by minimising the symmetric-KL divergence between its EGS and that of the original graph. We repeat the above procedures for different random graph models: Erdős-Rényi (ER), Watts-Strogatz (WS) and Barabási-Albert (BA), and different graph sizes ($n = 50, 100, 150$), and the results are shown in Table 1. It can be seen that given the approximate EGS, we can learn rather well the parameters of the graph producing that.

Comparing different real world networks We now consider the feasibility of comparing real world graphs using their EGSs. Specifically, we take 3 biological networks, 5 citation networks and 3 road networks [11], and compute the symmetric KL divergences between their EGS densities with $m = 100$ moments. The results in the form of a heat map in FIG 2 show clearly that the intra-class divergences among the biological, citation and road networks are much smaller than their inter-class divergences. This strongly suggests that the combination of our EGS approach and the symmetric KL divergence can be used to identify similarity in networks. Furthermore, as can be seen in the divergence between the human (column/row (0) and (1)) and mouse (column/row (2) index 2)

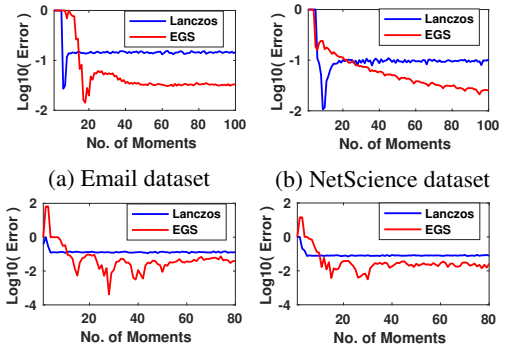
[1] Note that our proposed EGS which is based on MaxEnt distribution enables the symmetric KL divergence to be computed analytically, as shown the formulae in Appendix F

Algorithm 1 Entropic Graph Spectrum (EGS) Learner

1: **Input:** Normalized Laplacian $\mathbf{L}_{\text{norm}}$, Number of probe vectors d, Number of moments used m
2: **Output:** EGS $p(\lambda)$
3: Moments $\{\mu_i\}_{i=1}^{m} \leftarrow$ STE $(\mathbf{L}_{\text{norm}}, d, m)$
4: $\{\alpha_i\}_{i=1}^{m} \leftarrow$ MaxEnt algorithm $(\{\mu_i\}_{i=1}^{m})$
5: EGS $p(\lambda) = \exp[-(1 + \sum_i \alpha_i \lambda^i)]$

Table 1: Average parameters estimated by EGS for the 3 types of network. n denotes the number of nodes in the network.

n	100	150
ER ($p = 0.6$)	0.598	0.604
WS ($p = 0.4$)	0.454	0.414
BA ($r = 0.4n$)	40.239	58.428



(a) Email dataset (b) NetScience dataset

(c) 305 clusters (d) 1,355 clusters

Figure 1: Log error of cluster number detection using EGS and Lanczos methods on real networks (a) (b) and large synthetic (c) (d).

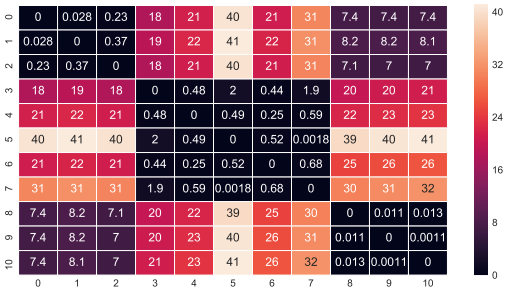


Figure 2: Symmetric KL heatmap between 9 graphs from the SNAP dataset: (0) bio-human-gene1, (1) bio-human-gene2, (2) bio-mouse-gene, (3) ca-AstroPh, (4) ca-CondMat, (5) ca-GrQc, (6) ca-HepPh, (7) ca-HepTh, (8) roadNet-CA, (9) roadNet-PA, (10) roadNet-TX.

networks, the spectra of the human genes are more closely aligned with each other than they are with the spectra of mouse genes. This suggests a reasonable amount of intra-class distinguishability.

5 EGS for learning the number of clusters

Algorithm 2 Cluster Number Estimation

1: **Input:** Normalized graph Laplacian $\mathbf{L}_{\text{norm}}$, graph dimension n, tolerance η
2: **Output:** Number of clusters N_c
3: Compute EGS $p(\lambda) \leftarrow$ Algorithm 1$(\mathbf{L}_{\text{norm}})$
4: Minimize λ_* s.t. $\frac{dp(\lambda)}{d\lambda}|_{\lambda=\lambda_*} \leq \eta$ and $\frac{d^2p(\lambda)}{d\lambda^2}_{\lambda=\lambda_*} > 0$
5: Calculate $N_c = n \int_0^{\lambda*} p(\lambda)d\lambda$

The number of zero eigenvalues in the normalized graph Laplacian represents the number of disconnected components [18]. Hence for a small number of inter-cluster links, we would expect a smoothed spectral density plot to have a spike near 0. We compute the number of clusters using EGS as shown in Algorithm 2. In this section, we compare EGS against Lanczos algorithm, a powerful iterative method, with kernel smoothing.

Cluster number estimation for real and large synthetic networks We experiment with median-size real world networks, such as the Email network with $n = 1,003$ nodes and a co-authorship network among network scientiest (NetScience) with $n = 1,589$ nodes [13]. For such network, we can still calculate the ground-truth number of clusters by computing the eigenvalues explicitly and measure the spectral gap near 0. We present the fractional error in estimating the true number of clusters in the networks over number of moments in Fig. 1b and 1a. It is clear that our EGS approach quickly outperforms the Lanczos method after around 20 moments.

To test the performance of our approach for networks that are too large to apply eigen-decomposition, we generate two large networks by mixing three types of random graph models(ER, WS, BA). The first large network (Fig 1c) has a size of 201,600 nodes and comprises 305 interconnected clusters whose size varies from 500 to 1000 nodes. The second large network (Fig 1d) has a size of 404,420 nodes and comprises interconnected 1355 clusters whose size varies from 200 to 400 nodes. The results in Fig 1c and 1d show that our EGS approach again outperforms the Lanczos method for both large synthetic networks.

References

[1] S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[2] A. Banerjee. *The spectrum of the graph Laplacian as a tool for analyzing structure and evolution of networks*. PhD thesis, 2008.

[3] N. Biggs, E. Lloyd, and R. Wilson. Graph theory 1736-1936, 1976, 1976.

[4] D. Cohen-Steiner, W. Kong, C. Sohler, and G. Valiant. Approximating the spectrum of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1263–1271. ACM, 2018.

[5] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek. Spectra of "real-world" graphs: Beyond the semicircle law. *Physical Review E*, 64(2):026704, 2001.

[7] A. R. Feier. *Methods of proof in random matrix theory*. PhD thesis, Harvard University, 2012.

[8] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–160. ACM, 2000.

[9] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[10] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[11] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[12] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[13] M. E. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[14] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[15] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814, 2005.

[16] D. Y. Takahashi, J. R. Sato, C. E. Ferreira, and A. Fujita. Discriminating different classes of biological networks by analyzing the graphs spectra distribution. *PLoS One*, 7(12):e49949, 2012.

[17] S. Ubaru, J. Chen, and Y. Saad. Fast Estimation of tr (f (a)) via Stochastic Lanczos Quadrature. 2016.

[18] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.