GEOMETRIC SCATTERING FOR GRAPH DATA ANALYSIS

Feng Gao
Depts. of Comp. Math., Sci.
& Eng., and Plant, Soil & Microb.
Sci., Michigan State University
East Lansing, MI 48824, USA
gaofeng2@msu.edu

Guy Wolf
Dept. of Math. & Stat.
Université de Montréal
Montreal, QC, H3T 1J4, Canada
guy.wolf@umontreal.ca

Matthew Hirn
Depts. of Comp. Math., Sci.
& Eng., and Mathematics
Michigan State University
East Lansing, MI 48824, USA
mhirn@msu.edu

ABSTRACT

We explore the generalization of scattering transforms from traditional signals to graph data, analogous to the generalization of ConvNets in geometric deep learning, and the utility of extracted graph features in graph data analysis. In particular, we focus on the capacity of these features to retain informative variability and relations in the data (e.g., between individual graphs, or in aggregate). We demonstrate the application of our geometric scattering features in graph classification of social network data, and in data exploration of biochemistry data.

1 INTRODUCTION

Over the past decade, Convolutional Neural Networks (CNNs) have achieved great success in computer vision, where the utilization of 2D convolutions enable network designs that learn cascades of convolutional filters. Beyond their performances when applied to specific tasks, pretrained ConvNet layers have been explored as image feature extractors by freezing the first few pretrained convolutional layers and then retraining only the last few layers for specific datasets or applications (e.g., Yosinski et al., 2014; Oquab et al., 2014). Such transfer learning approaches provide evidence that suitably constructed deep filter banks should be able to extract task-agnostic semantic information from structured data, and in some sense mimic the operation of human visual and auditory cortices, thus supporting the neural terminology in deep learning.

An alternative approach towards such universal feature extraction was presented in Mallat (2012), where a deep filter bank, known as the scattering transform, is *designed*, rather than trained, based on predetermined families of distruptive patterns that should be eliminated to extract informative representations. The scattering transform is constructed as a cascade of linear wavelet transforms and nonlinear complex modulus operations that provides features with guaranteed invariance to a predetermined Lie group of operations such as rotations, translations, or scaling. Further, it also provides Lipschitz stability to small diffeomorphisms of the inputted signal. Recently several attempts have been made to generalize the scattering transform to graphs (Zou & Lerman, 2018; Gama et al., 2018) and manifolds (Perlmutter et al., 2018), which we will generally term "geometric scattering." While these works mostly focus on following the footsteps of Mallat (2012) in establishing the stability of their respective constructions to deformations of input signals or graphs, here we further explore the notion of geometric scattering features by considering the complimentary question of how much information is retained by them, since stability alone does not ensure useful features in practice (e.g., a constant all-zero map would be stable to any deformation, but would clearly be useless).

In this paper, we focus on empirical results on several data analysis tasks, and on two commonly used graph data types. Our results in Sec. 3.1 show that on social network data, geometric scattering features enable classic RBF-kernel SVM to match, if not outperform, leading graph kernel methods as well as most geometric deep learning ones. These experiments are augmented by additional results in Sec. 3.2 that show the geometric scattering SVM classification rate degrades only slightly when trained on far fewer graphs than is traditionally used in graph classification tasks. On biochemistry data, where graphs represent molecular structures of compounds (e.g., Enzymes or proteins), we show in Sec. 3.2 that scattering features enable significant dimensionality reduction. Finally, to establish their descriptive qualities, in Sec. 3.3 we use geometric scattering features extracted from enzyme data (Borgwardt et al., 2005a) to infer emergent patterns of enzyme commission (EC) exchange preferences in enzyme evolution, validated with established knowledge from Cuesta et al. (2015).

2 GEOMETRIC SCATTERING

Let $G = (V, E, W)$ be a weighted graph, consisting of n vertices $V = \{v_1, \ldots, v_n\}$, edges $E \subseteq \{(v_\ell, v_m) : 1 \leq \ell, m \leq n\}$, and weights $W = \{w(v_\ell, v_m) > 0 : (v_\ell, v_m) \in E\}$. Note that unweighted graphs are considered as a special case, by setting $w(v_\ell, v_m) = 1$ for each $(v_\ell, v_m) \in E$. Define the $n \times n$ (weighted) adjacency matrix $\mathbf{A}_G = \mathbf{A}$ of G by $\mathbf{A}(v_\ell, v_m) = w(v_\ell, v_m)$ if $(v_\ell, v_m) \in E$ and zero otherwise, where we use the notation $\mathbf{A}(v_\ell, v_m)$ to denote the (ℓ, m) entry of the matrix $\mathbf{A}$ so as to emphasize the correspondence with the vertices in the graph and to reserve sub-indices for enumerating objects. Define the (weighted) degree of vertex v_ℓ as $\deg(v_\ell) = \sum_m \mathbf{A}(v_\ell, v_m)$ and the corresponding diagonal $n \times n$ degree matrix $\mathbf{D}$ given by $\mathbf{D}(v_\ell, v_\ell) = \deg(v_\ell)$, $\mathbf{D}(v_\ell, v_m) = 0$, $\ell \neq m$. Define the $n \times n$ lazy random walk matrix as $\mathbf{P} = \frac{1}{2}\left(\mathbf{I} + \mathbf{A}\mathbf{D}^{-1}\right)$.

The operator $\mathbf{P}$ can be considered as a low pass operator. Indeed, if $\mathbf{x} : V \to \mathbb{R}$ is a signal defined on G, then $\mathbf{P}\mathbf{x}$ will retain the low frequencies of $\mathbf{x}$ as defined by the eigenvalues of the (normalized) graph Laplacian of G. High frequency responses of $\mathbf{x}$ can be recovered in multiple different fashions, but we utilize multiscale wavelet transforms that group the non-zero frequencies of G into approximately dyadic bands. Following Coifman & Maggioni (2006), define the $n \times n$ wavelet matrix at the scale 2^j as

$$\mathbf{\Psi}_0 = \mathbf{I} - \mathbf{P}\,, \quad \mathbf{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j} = \mathbf{P}^{2^{j-1}}(\mathbf{I} - \mathbf{P}^{2^{j-1}})\,, \quad j \geq 1\,. \tag{1}$$

To compare and classify multiple graphs, we utilize signals $\mathbf{x}_G = \mathbf{x}$ that can be defined on any graph G, e.g., $\mathbf{x}(v_\ell) = \deg(v_\ell)$. The graph wavelet transform of $\mathbf{x}$, $\mathbf{W}\mathbf{x} = \{\mathbf{P}^{2^J}\mathbf{x}\,,\, \mathbf{\Psi}_j\mathbf{x} : 1 \leq j \leq J\}$, is a complete representation of $\mathbf{x}$. However, it is not invariant to index permutations of the vertices. Invariant graph features can be obtained via summation operators; the simplest of these computes the sum of the responses of the signal $\mathbf{x}$. For example, the unnormalized q^{th} moments of $\mathbf{x}$ yield the following "zero" order scattering moments:

$$S\mathbf{x}(q) = \sum_{\ell=1}^{n} \mathbf{x}(v_\ell)^q, \quad 1 \leq q \leq Q\,. \tag{2}$$

We can also replace (2) with normalized (i.e., standardized) moments of $\mathbf{x}$, in which case we store its mean ($q = 1$), variance ($q = 2$), skew ($q = 3$), kurtosis ($q = 4$), and so on.

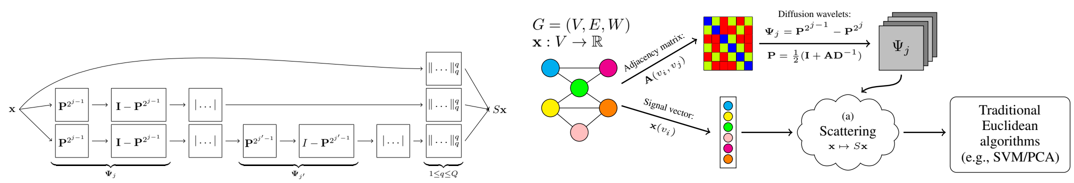
The invariants $S\mathbf{x}(q)$ do not capture the full variability of $\mathbf{x}$ and hence the graph G upon which the signal $\mathbf{x}$ is defined. We thus complement these moments with summary statistics derived from the absolute values of the high frequency wavelet coefficients of $\mathbf{x}$, which will lead naturally to the graph ConvNet structure of the geometric scattering transform:

$$S\mathbf{x}(j, q) = \sum_{\ell=1}^{n} |\mathbf{\Psi}_j\mathbf{x}(v_\ell)|^q,\ 1 \leq j \leq J,\ 1 \leq q \leq Q\,. \tag{3}$$

First order geometric scattering moments can be augmented with second order geometric scattering moments by iterating the graph wavelet and absolute value transforms. These moments are defined as:

$$S\mathbf{x}(j, j', q) = \sum_{\ell=1}^{n} |\mathbf{\Psi}_{j'}|\mathbf{\Psi}_j\mathbf{x}(v_\ell)||^q,\ \begin{array}{l} 1 \leq j < j' \leq J \\ 1 \leq q \leq Q\,, \end{array} \tag{4}$$

The collection of graph scattering moments $S\mathbf{x} = \{S\mathbf{x}(q),\ S\mathbf{x}(j,q),\ S\mathbf{x}(j,j',q)\}$ (illustrated in Fig. 1(a)) provides a rich set of multiscale invariants of the graph G. Similar constructions were introduced in Zou & Lerman (2018) and Gama et al. (2018), which studied the theoretical stability of the representation with respect to certain metrics between graphs. In this paper we observe that geometric scattering features can be used in supervised settings as input to graph classification or regression models, or in unsupervised settings to embed graphs into a Euclidean feature space for further exploration, as demonstrated in Sec. 3.



(a) Representative zeroth-, first-, and second-order cascades of the geometric scattering transform for an input graph signal $\mathbf{x}$.

(b) Architecture for using geometric scattering of graph G and signal $\mathbf{x}$ in graph data analysis, as demonstrated in Sec. 3.

Figure 1: Illustration of (a) the proposed scattering feature extraction (see eqs. 2, 3, and 4), and (b) its application for graph data analysis.

3 Application & Results

3.1 Graph classification on social networks

As a first application of geometric scattering, we apply it to graph classification of social network data taken from Yanardag & Vishwanathan (2015). In particular, this work introduced six social network data sets extracted from scientific collaborations (COLLAB), movie collaborations (IMDB-B & IMDB-M), and Reddit discussion threads (REDDIT-B, REDDIT-5K, REDDIT-12K).

These social network datasets contain graph structures but no associated graph signals. Therefore we compute the eccentricity (for connected graphs) and clustering coefficient of each vertex, and use these as input signals to the geometric scattering transform. We then use the normalized geometric scattering features as inputs to the SVM classifier with an RBF kernel for classification; see also Figure 1(b). Utilizing 10-fold cross validation, we compare our results to 10 prominent methods that report results for most, if not all, of the considered datasets in Table 1.

Table 1: Comparison of the proposed GS-SVM classifier with leading graph kernel and deep learning methods on social graph datasets.

	COLLAB	IMDB-B	IMDB-M	REDDIT-B	REDDIT-5K	REDDIT-12K	
WL (Shervashidze et al., 2011)	77.82 ± 1.45	71.60 ± 5.16	N/A	78.52 ± 2.01	50.77 ± 2.02	34.57 ± 1.32	Graph kernel
Graphlet (Shervashidze et al., 2009)	73.42 ± 2.43	65.40 ± 5.95	N/A	77.26 ± 2.34	39.75 ± 1.36	25.98 ± 1.29	
WL-OA (Kriege et al., 2016)	80.70 ± 0.10	N/A	N/A	89.30 ± 0.30	N/A	N/A	
DGK (Yanardag & Vishwanathan, 2015)	73.00 ± 0.20	66.90 ± 0.50	44.50 ± 0.50	78.00 ± 0.30	41.20 ± 0.10	32.20 ± 0.10	
DGCNN (Zhang et al., 2018)	73.76 ± 0.49	70.03 ± 0.86	47.83 ± 0.85	N/A	48.70 ± 4.54	N/A	Deep learning
2D CNN (Tixier et al., 2017)	71.33 ± 1.96	70.40 ± 3.85	N/A	89.12 ± 1.70	52.21 ± 2.44	48.13 ± 1.47	
PSCN (Niepert et al., 2016, with $k = 10$)	72.60 ± 2.15	71.00 ± 2.29	45.23 ± 2.84	86.30 ± 1.58	49.10 ± 0.70	41.32 ± 0.42	
GCAPS-CNN (Verma & Zhang, 2018)	77.71 ± 2.51	71.69 ± 3.40	48.50 ± 4.10	87.61 ± 2.51	50.10 ± 1.72	N/A	
S2S-P2P-NN (Taheri et al., 2018)	81.75 ± 0.80	73.80 ± 0.70	51.19 ± 0.50	86.50 ± 0.80	52.28 ± 0.50	42.47 ± 0.10	
GIN-0 (MLP-SUM) (Xu et al., 2019)	80.20 ± 1.90	75.10 ± 5.10	52.30 ± 2.80	92.40 ± 2.50	57.50 ± 1.50	N/A	
GS-SVM	79.94 ± 1.61	71.20 ± 3.25	48.73 ± 2.32	89.65 ± 1.94	53.33 ± 1.37	45.23 ± 1.25	

Examining Table 1 one can see that the geometric scattering SVM (GS-SVM) classifier generally matches the performance of, and sometimes outperforms, all but the two most recent methods, i.e., S2S-N2N-PP (Taheri et al., 2018) and GIN (Xu et al., 2019). With regards to these two approaches, the GS-SVM outperforms S2S-N2N-PP (Taheri et al., 2018) on $^3/_6$ datasets. Finally, while GIN (Xu et al., 2019) outperforms geometric scattering on $^5/_6$ datasets, the results on COLLAB and IMDB-B are not statistically significant, and on the REDDIT datasets the geometric scattering approach trails only GIN (Xu et al., 2019). We thus conclude that the geometric scattering transform yields a rich set of invariant statistical moments, which have nearly the same capacity as the current state of the art in graph neural networks.

3.2 Classification with limited training-data and dimensionality reduction

We performed graph classification under four training/validation/test splits: 80%/10%/10%, 70%/10%/20%, 40%/10%/50% and 20%/10%/70%. Following Sec. 3.1, we discuss the classification accuracy on six social network datasets under these splits. When the training data is reduced from 80% to 70%, the classification accuracy in fact increased by 0.047%, which shows the GS-SVM classification accuracy is not affected by the decrease in training size. Further reducing the training size to 50% results in an average decrease of classification accuracy of 1.40% while from 90% to 20% causes an average decrease of 3.00%. Fig. 2(a) gives a more nuanced statistical description of these results.

Relatedly, we also considered the viability of scattering-based embedding for dimensionality reduction of graph data. As a representative example, we consider here the ENZYMES dataset introduced in Borgwardt et al. (2005b), which contains 600 enzymes evenly split into six enzyme classes (i.e., 100 enzymes from each class).

We applied PCA to geometric scattering features extracted from input enzyme graphs in the data, while choosing the number of principal components to capture 99%, 90%, 80% and 50% explained variance. For each of these thresholds, we computed the mean classification accuracy (with ten-fold cross validation) of SVM applied to the GS-PCA low dimensional space, as well as the dimensionality of this space. The relation between dimensionality, explained variance, and SVM accuracy is shown in Fig. 2(b), where we can observe that indeed geometric scattering combined with PCA enables significant dimensionality reduction (e.g., to $\mathbb{R}^{16}$ with 90% exp. variance) with only a small impact on classification accuracy.

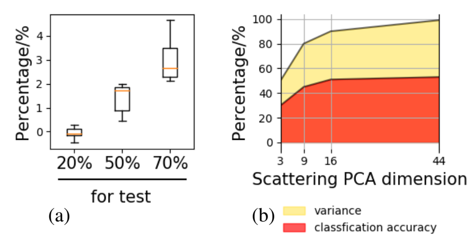


Figure 2: (a) Box plot showing the drop in SVM classification accuracy over social graph datasets when reducing training set size (horizontal axis marks portion of data used for testing); (b) Relation between explained variance, SVM classification accuracy, and PCA dimensions over scattering features in ENZYMES dataset.

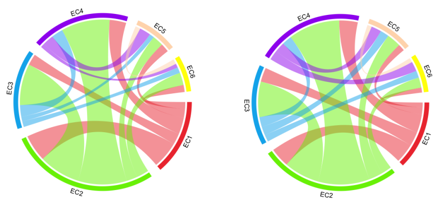
3.3 Data exploration: Enzyme class exchange preferences

In this section, we focus our discussion on the ENZYMES dataset described in the previous section. Here, geometric scattering features can be considered as providing "signature" vectors for individual enzymes, which can be used to explore interactions between the six top level enzyme classes, labelled by their Enzyme Commission (EC) numbers (Borgwardt et al., 2005a). In order to emphasize the properties of scattering-based feature extraction, rather than downstream processing, we mostly limit our analysis of the scattering feature space to linear operations such as principal component analysis (PCA).

To explore the scattering feature space, and the richness of information captured by it, we use it to infer relations between EC classes. First, for each enzyme e, with scattering feature vector $\boldsymbol{b}_e$ (i.e., with $S\mathbf{x}$ for all vertex features $\mathbf{x}$), we compute its distance from class EC-j, with PCA subspace $\mathcal{C}_j$, as the projection distance: $\text{dist}(e, \text{EC-}j) = \|\boldsymbol{b}_e - \text{proj}_{\mathcal{S}_j} \boldsymbol{b}_e\|$. Then, for each enzyme class EC-i, we compute the mean distance of enzymes in it from the subspace of each EC-j class as $D(i,j) = \text{mean}\{\text{dist}(e, \text{EC-}j) : e \in \text{EC-}i\}$.

These distances are summarized in the supplement, as well as the proportion of points from each class that have their true EC as their nearest (or second nearest) subspace in the scattering feature space. In general, 48% of enzymes select their true EC as the nearest subspace (with additional 19% as second nearest), but these proportions vary between individual EC classes. Finally, we use these scattering-based distances to infer EC exchange preferences during enzyme evolution, which are presented in Fig. 3 and validated with respect to established preferences observed and reported in Cuesta et al. (2015). The details of calculation of EC exchange preferences can be found in the supplement.

We note that the result there is observed independently from the ENZYMES dataset. Our results in Fig. 3 demonstrate that scattering features are sufficiently rich to capture relations between enzyme classes, and indicate that geometric scattering has the capacity to uncover descriptive and exploratory insights in graph data analysis.



(a) Observed (b) Inferred

Figure 3: Comparison of EC exchange preferences in enzyme evolution: (a) observed in Cuesta et al. (2015), and (b) inferred from scattering features. Our inference (b) mainly recovers (a).

3.4 Ablation Study

To fully understand the power of our geometric scattering coefficients, especially to demonstrate the usefulness of those generated from second order scattering moments defined in equation 4, we perform an ablation study using the REDDIT-B and ENZYME datasets as representative examples. We used normalized features in this section to be consistent with the results of previous sections. We conducted classification tasks with only zero order features and zero order + first order features. Our results show that indeed scattering coefficients from higher order moments contribute to the capacity of the geometric scattering SVM classifier. The classification accuracy only using zero order features on the ENZYME dataset drops sharply to 34.33%. Adding first order features increases the accuracy to 54.17%, which is still lower than 56.83% using zero order, first order and second order features. For the REDDIT-B dataset, the classification accuracy dropped to 85.75% with only zero order features, and 87.30% with zero order + first order features, compared to 89.65% using all three orders of features.

Table 2: Ablation study on REDDIT-B and ENZYME datasets using only zero order and zero order + first order features.

	zero order only	zero order + first order
REDDIT-B	85.75 ± 2.80	87.30 ± 1.76
ENZYME	34.33 ± 5.64	54.17 ± 4.73

4 Conclusion

We presented the geometric scattering transform as a deep filter bank for feature extraction on graphs, which generalizes the Euclidean scattering transform. Our evaluation results on graph classification and data exploration show the potential of the produced scattering features to serve as universal representations of graphs. They raise the possibility of embedding entire graphs in Euclidean space and computing meaningful distances between graphs, which can be used for both supervised and unsupervised learning.

5 Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. F.G. is supported by the grant P42 ES004911 through the National Institute of Environmental Health Sciences of the National Institutes of Health. M.H. is supported by the Alfred P. Sloan Fellowship (grant FG-2016-6607), the DARPA YFA (grant D16AP00117), and NSF grant 1620216.

REFERENCES

Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005a.

Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005b.

Ronald R. Coifman and Mauro Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

Sergio Martínez Cuesta, Syed Asad Rahman, Nicholas Furnham, and Janet M. Thornton. The classification and evolution of enzyme function. *Biophysical Journal*, 109(6):1082–1086, 2015.

Fernando Gama, Alejandro Ribeiro, and Joan Bruna. Diffusion scattering transforms on graphs. arXiv:1806.08829, 2018.

Nils M. Kriege, Pierre-Louis Giscard, and Richard Wilson. On valid optimal assignment kernels and applications to graph classification. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1623–1631. Curran Associates, Inc., 2016.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65 (10):1331–1398, October 2012.

Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pp. 2014–2023, 2016.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724, 2014.

Michael Perlmutter, Guy Wolf, and Matthew Hirn. Geometric scattering on manifolds. In *NeurIPS Workshop on Integration of Deep Learning Theories*, pp. arXiv:1812.06968, 2018.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In David van Dyk and Max Welling (eds.), *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 488–495, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009. PMLR.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep): 2539–2561, 2011.

Aynaz Taheri, Kevin Gimpel, and Tanya Berger-Wolf. Learning graph representations with recurrent neural network autoencoders. In *KDD Deep Learning Day*, 2018.

Antoine Jean-Pierre Tixier, Giannis Nikolentzos, Polykarpos Meladianos, and Michalis Vazirgiannis. Classifying graphs as images with convolutional neural networks. *arXiv preprint*, pp. arXiv:1708.02218, 2017.

Saurabh Verma and Zhi-Li Zhang. Graph capsule convolutional neural networks. *arXiv preprint*, pp. arXiv:1805.08090, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374. ACM, 2015.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pp. 3320–3328, 2014.

Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *AAAI Conference on Artificial Intelligence*, pp. 4438–4445, 2018.

Dongmian Zou and Gilad Lerman. Graph convolutional neural networks via scattering. arXiv:1804:00099, 2018.