

SGR: Self-Supervised Spectral Graph Representation Learning

Anton Tsitsulin
Hasso Plattner Institute
Potsdam, Germany

Davide Mottin
Hasso Plattner Institute
Potsdam, Germany

Panagiotis Karras
Aarhus University
Aarhus, Denmark

Alexander Bronstein
Technion
Haifa, Israel

Emmanuel Müller
Hasso Plattner Institute
Potsdam, Germany

ABSTRACT

Representing a graph as a vector is a challenging task; ideally, the representation should be easily computable and conducive to efficient comparisons among graphs, tailored to the particular data and analytical task at hand. Unfortunately, a “one-size-fits-all” solution is unattainable, as different analytical tasks may require different attention to global or local graph features. We develop SGR, the first, to our knowledge, method for learning graph representations in a *self-supervised* manner. Grounded on spectral graph analysis, SGR seamlessly combines all aforementioned desirable properties. In extensive experiments, we show how our approach works on large graph collections, facilitates self-supervised representation learning across a variety of application domains, and performs competitively to state-of-the-art methods without re-training.

1 INTRODUCTION

A multitude of data in various domains, from natural sciences to sociology, is represented as collections of graphs. For example, on a small scale molecules are modeled by atoms and their atomic bonds as nodes and edges in large graph collections. While on a larger scale collections of social networks are analyzed by their community structures within the networks. Analytical tasks run on such collections to classify, for instance, which drugs can be used for the treatment of a disease or how molecules cluster together in functional groups. To fully discern a graph’s properties, representation learning for such analysis requires a *multiscale* view of a graph. Representations have incorporated properties ranging from *local* (e.g., atomic bonds) to *global* (e.g., community structures).

Kernel methods [6, 14, 18, 19, 21], dominate the field of graph analytics, as they define functional similarities among pairs of graphs and can perform tasks such as graph classification. Among a plethora of graph kernels, to our knowledge, only the Multiscale Laplacian Graph kernel (MLG) [14] preserves multiscale properties. Still, such graph kernels require *direct comparisons* among pairs of graphs, hence scale quadratic in the size of any graph collection. As more and more data is stored as graph collections, such all-pairs direct comparisons methods are rendered infeasible.

In another vein, *graph representations* extract a vector signatures from a graph and perform comparison among those vectors in lieu of the graphs. Initial approaches derived such representations by hand-crafted statistics on the graph structure (e.g., the average node degree [4]). Recently, supervised neural approaches for graph representation [1, 16] attained competitive performance in supervised classification among graphs of a few tens of nodes. However, such neural methods are applicable to particular datasets only, as they require labels to be available; besides, they fail to scale to graphs of a few thousands of nodes. Most recently, Verma et al. [20] proposed the *Family of Graph Spectral Distances* (FGSD): a graph representation method based on histograms of the biharmonic kernel. While FGSD representations are designed for classification tasks, these are neither expressive nor scalable enough to be used in both classification tasks.

In this paper we propose SGR, a self-supervised method for learning graph representations that is at the same time efficient to compute and customizable to multiple scales, analytical tasks, and datasets. SGR leverages a graph’s Laplacian spectrum to learn a mapping a collection of graphs to their vector representation, by learning a simple single-layer perceptron on global structure recognition. The perceptron learns to distinguish synthetic graphs with community structure (i.e., sampled from a stochastic block model [11]) from random graphs by the Erdős-Rényi model. SGR representation is self-learning in the sense that it requires no real training data. We conduct an experimental study with several real datasets, using the ensuing graph signature representations on tasks such as graph classification by logistic regression. The results attest the superiority of our approach on classification tasks with real data.

2 RELATED WORK

Previous work on learning graph representations falls into three groups, outlined in Table 1.

2.1 Kernel methods

Graph kernels [6, 10, 14, 17, 19, 21] are similarity functions among graphs, which perform an implicit transformation of graph structure to compare two graphs (e.g. Shortest-path (SP) kernel [19]). However, kernel methods are limited due to (i) high on-demand computational complexity at comparison time, which renders them inapplicable to large-scale graph comparisons, and (ii) rigidity: once a kernel is chosen, it cannot be tailored to the analytical task or dataset at hand. The Multi-scale Laplacian Graph kernel (MLG) [14] is a mature work on this domain, as it adapts to different scales via

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD’18 Deep Learning Day, August 2018, London, UK

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Method	Type	Properties			Complexity	
		Learned	Unsupervised	Multi-scale	Precomputation	Comparison
SP [6]	kernel	✗	✗	✗	$O(1)$	$O(n^3)$
WL [19]	kernel	✗	✗	✗	$O(1)$	$O(m \log(n))$
MLG [14]	kernel	✗	✓	✓	$O(km + k^2n)$	$O(k^3)$
PSCN [16]	supervised NN	✓	✗	✗	$O(dn \log(n))$	$O(dn \log(n))$
DCNN [1]	supervised NN	✓	✗	✗	$O(dn^2)$	$O(dn^2)$
NETSIMILE [4]	fixed representation	✗	✓	✗	$O(n \log n)$	$O(d)$
FGSD [20]	fixed representation	✗	✓	✗	$O(n^2)$	$O(d)$
SGR	self-learned representation	✓	✓	✓	$O(km + k^2n)$	$O(d)$

Table 1: Related work allowing graph comparison in terms of fulfilled (✓) and missing (✗) characteristics and complexity (n nodes, m edges, k eigenvalues, d representation dimensions).

an iterative information-propagation method. Yet it also raises a computational overhead cubic in Laplacian matrix eigenvalues.

2.2 Supervised neural methods

Advances in neural learning have led to the application of *supervised neural* approaches to classify collections of graphs. The Diffusion Convolutional Neural Network (DCNN) [1] learns graph representations by averaging values after a diffusion process on a graph’s nodes. Similarly, PATCHY-SAN [16] learns a representation through a CNN filter after imposing a sampling order on nodes. Yet such approaches share the drawbacks of kernel methods: high computational overhead at comparison time and lack of variable adaptability to local or global structures. Besides, the learning component in these neural methods is *supervised* by means of node and edge labels, raising an additional resource requirement.

2.3 Fixed representation methods

Another class of approaches eschew the supervised learning component of neural approaches. Such works started out using features engineered by aggregating local graph properties such as node degree and neighbors’ degrees [4, 5, 7]. However, in eschewing supervision, such works eschew learning altogether. We call them *fixed representation* methods. The Family of Spectral Distances (FSGD) [20] produces a high-dimensional sparse representation as a histogram on the dense biharmonic graph kernel; however, FGSD does not capture graph features at different scales of resolution or graph sizes, and is also inapplicable to reasonably large graphs, due to its quadratic time complexity.

By contrast to the above, we devise a lightweight *self-learned* representation method, which is extracted directly from the graph structure and can be used across graph analysis tasks.

3 PROBLEM STATEMENT

An undirected graph is a pair $G = (V, E)$, where $V = (v_1, \dots, v_n)$, $n = |V|$ is the set of vertices and $E \subseteq (V \times V)$ the set of edges. Since the vertex set is isomorphic to $\{1, \dots, n\}$, we will henceforth use the latter notation. We assume the graph is unweighted, yet our method readily applies to the weighted case. A *representation* is a function $\sigma : \mathcal{G} \rightarrow (\mathbb{R}^N, \ell_2)$ from a graph G in a collection of graphs \mathcal{G} to the N -dimensional space equipped with the Euclidean metric; the element j of the representation is denoted as $\sigma_j(G)$. Notably,

once a graph’s representation is computed, comparisons between representations (e.g., for retrieval or classification) is independent of graph size.

The first and foremost property a representation has to satisfy is *permutation-invariance*, implying that if two graphs have the same structure (i.e., are isomorphic) the distance between their representations is zero. In other words, we demand that for every graph G , the representation is invariant to every permutation π of the graph vertices, $\sigma \circ \pi(G) = \sigma(G)$. In the sequel, we propose representations based on the Laplacian spectrum, which are permutation-invariant by construction.

Another desirable property is *scale-adaptivity*, implying that the representation shall account for both local (edge and node) and global (community) graph features. A global feature is such that cannot be captured by *any* combination of features on nodes at distance $r < D(G) - 1$, where $D(G)$ is the diameter (longest shortest-path length) of G . Let the set of all subgraphs of G be $\xi(G) = \{g \subseteq G : D(g) < D(G)\}$. We define scale-adaptivity as the property of a representation σ having at least one local feature (i.e., derived solely from information encoded in subgraphs $\xi(G)$), and at least one global feature (i.e., derived by strictly more than the information encoded in any $\xi(G)$). Using local features only, a similarity measure would deem two graphs sharing local patterns to have near-zero distance although their global properties (e.g., page-rank features) may differ; in reverse, relying on global features only would miss local ones (e.g., degree distribution).

We construct a parametric family of graph representations $\sigma^\theta : \mathcal{G} \rightarrow (\mathbb{R}^N, \ell_2)$, with parameter set θ , such that σ^θ captures global and local features to different extents, depending on θ . Further, we adapt θ to fit a purpose by means of unsupervised self-learning.

4 SPECTRAL GRAPH REPRESENTATIONS

The *adjacency matrix* of a graph G is a $n \times n$ matrix A having $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise. The *normalized graph Laplacian* is defined as the matrix $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is the diagonal matrix with the degree of node i as entry D_{ii} , i.e. $D_{ii} = \sum_{j=1}^n A_{ij}$. Since the Laplacian is a symmetric matrix, its eigenvectors ϕ_1, \dots, ϕ_n , are real and orthonormal. Thus, it is factorized as $L = \Phi \Lambda \Phi^T$, where Λ is a diagonal matrix with the sorted eigenvalues $\lambda_1 \leq \dots \leq \lambda_n$, and Φ is the orthonormal matrix $\Phi = (\phi_1, \dots, \phi_n)$ having the corresponding eigenvectors as its columns. Belkin and Niyogi [2]

Figure 1: Different regions of the spectrum have a different impact on the classifier co-trained with the SGR. The color map shows the gradient magnitude of the classifier output with respect to the input spectrum visualized in increasing order from left to right, averaged on 600 graphs. Top: SBM; bottom: Erdős-Rényi.

showed that the eigenvectors of the normalized Laplacian of a point cloud graph converge to the eigenfunction of the Laplace-Beltrami operator [3] on the underlying Riemannian manifold.

The set of eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ is called the *spectrum* of a graph and is bounded in the range $\lambda_i \in [0, 2]$. Being an algebraic invariant of the Laplacian, its spectrum is independent of the choice of the basis, which, in particular means that it is permutation-invariant.

4.1 Heat propagation

A useful metaphor for studying the graph structure at different scales is that of a system of connected heat-conducting rods corresponding to the graph edges. Heat propagation in such a system is governed by the *heat equation* associated with the Laplacian,

$$\frac{\partial \mathbf{u}_t}{\partial t} = -\mathbf{L}\mathbf{u}_t, \quad (1)$$

where \mathbf{u}_t is a scalar field on the graph vertices representing the amount of heat at each vertex at time t . The solution to the heat equation provides the heat at each vertex at time t , when the initial heat \mathbf{u}_0 is initialized with a fixed value on one of the vertices. It has a closed-form fundamental solution in the form of the $n \times n$ *heat kernel matrix*,

$$\mathbf{H}_t = e^{-t\mathbf{L}} = \Phi e^{-t\Lambda} \Phi^\top = \sum_k e^{-t\lambda_k} \phi_k \phi_k^\top, \quad (2)$$

where $(\mathbf{H}_t)_{ij}$ represents the amount of heat transferred from vertex i to vertex j in time t . The diagonal entries of \mathbf{H}_t are called the auto-diffusivity function, representing the amount of heat remaining at each graph vertex after time t . This auto-diffusivity function is affected by increasingly global structures of the graph’s topology as the time parameter t grows.

The sum of the auto-diffusivity function, known as the *heat trace*

$$h_t = \text{tr}(\mathbf{H}_t) = \sum_j e^{-t\lambda_j} \quad (3)$$

is an algebraic invariant of the heat kernel and can be, therefore, expressed only in terms of the invariant graph spectrum.

Theoretical results by Mémoli [15] subscribe the expressiveness of heat traces, suggesting a spectral definition of the Gromov-Wasserstein distance between Riemannian manifolds based on matching the heat kernels at all scales. In what follows, we briefly review this construction, adapting it to graphs. Let us set the cost of matching a pair of vertices (i, i') in a graph G_1 to a pair of points (j, j') in a graph G_2 at a scale t to be the discrepancy of the corresponding heat kernels,

$$\Gamma_t(i, j, i', j') = e^{-2(t+t^{-1})} |(\mathbf{H}_t^{G_1})_{ii'} - (\mathbf{H}_t^{G_2})_{jj'}|,$$

where the factor $e^{-2(t+t^{-1})}$ scales the kernels. A distance between the graphs can then be defined in terms of the minimal measure coupling

$$d^2(G_1, G_2) = \min_{\mathbf{M}} \sup_{t>0} \|\Gamma_t\|_{\ell^2(\mathbf{M} \times \mathbf{M})}^2,$$

where the minimum is sought over all doubly-stochastic matrices \mathbf{M} representing a discrete measure on $G_1 \times G_2$ that marginalizes to the uniform measures on G_1 and G_2 . This distance can be thought of as a “soft” version of the standard graph edit distance and has the useful property that $d(G_1, G_2) = 0$ iff G_1 and G_2 are isomorphic.

Mémoli [15] showed that the spectral Gromov-Wasserstein distance can be lower bounded by

$$d(G_1, G_2) \geq \sup_{t>0} e^{-2(t+t^{-1})} |h_t^{G_1} - h_t^{G_2}|,$$

which is the scaled L_∞ distance between heat traces of the graphs.

4.2 Learned spectral representations

The heat traces can be viewed as a nonlinear transformation of the graph spectrum of the form $\sum_k f_t(\lambda_k)$ with $f_t(\lambda) = e^{-t\lambda}$. Sampling the time parameter on some grid $\{t_1, \dots, t_N\}$ yields the following N -dimensional representation of the graph:

$$\sigma = \left(\sum_k f_{t_1}(\lambda_k), \dots, \sum_k f_{t_N}(\lambda_k) \right).$$

We propose to *extend this view* to a more general parametric family of spectrum transformations. Given a graph G with n vertices, we first compute its spectrum or a part thereof $\{\lambda_k\}$, and interpolate it producing $\lambda(x)$ on the interval $[0, 1]$ such that $\lambda(k/n) = \lambda_k$. The spectrum is then sampled on a fixed grid (x_1, \dots, x_M) with M points, producing an M -dimensional vector $\tilde{\lambda}$ with the entries $\tilde{\lambda}_k = \lambda(x_k)$. Note that $\tilde{\lambda}$ is insensitive to a graph’s size and invariant to the ordering of its vertices.

The interpolated and sampled spectrum $\tilde{\lambda}$ undergoes next a parametric non-linear transformation implemented as a single-layered perceptron,

$$\sigma = \psi(\mathbf{W}\tilde{\lambda} + \mathbf{b})$$

where \mathbf{W} is an $N \times M$ weight matrix, \mathbf{b} is an N -dimensional bias vector, and ψ is an element-wise SeLU non-linearity [13]. The resulting N -dimensional spectral graph representation (SGR) is parameterized by $\theta = (\mathbf{W}, \mathbf{b})$.

We propose a regime to train this representation. To obtain a representation capturing predominantly the *global* structure of the graph (manifested in the lower part of the spectrum), we co-train σ jointly with a binary classifier attempting to distinguish between Erdős-Rényi random graphs and stochastic block model [11] graphs of various degrees and sizes, which have very different community

dataset	Kernels				Fixed Representations			Self-sup. Repr.
	SP	GK-4	WL	MLG	NETSIMILE	FGSD	Λ	SGR
D&D	>1D	73.39	68.27	>1D	70.02	64.88	64.54	76.12
ENZYMES	22.57	19.11	25.11	31.40	28.06	28.85	25.28	33.67
MUTAG	80.30	80.76	81.16	86.54	83.66	85.23	82.07	86.97
PROTEINS	72.04	71.48	72.33	73.10	70.59	63.27	71.32	73.83
COLLAB	>1D	>1D	78.52	>1D	74.26	70.66	66.15	71.98
IMDB-B	57.10	61.79	72.26	59.18	70.96	69.20	63.16	70.38
IMDB-M	39.39	39.80	50.75	34.31	46.80	48.88	41.14	47.97
REDDIT-B	>1D	72.30	71.97	>1D	86.84	87.12	76.25	87.45
REDDIT-M-5k	>1D	23.39	48.57	>1D	44.96	48.51	48.02	53.22

Table 2: Graph classification accuracy on bio-chemical (top) and social (bottom) graph collections. Best results are highlighted.

structures. The binary classifier is embodied as a single linear layer on top of the output of σ followed by softmax, and is trained using the regular cross-entropy loss.

The classifier is tossed away, leaving an appropriately trained graph representation. This approach is inspired by the versatility of image embeddings obtained from deep neural networks trained on visual recognition tasks. We henceforth denote the representation as SGR. Figure 1 depicts the saliency map for the interpolated spectrum. Perhaps surprisingly, the neural network leaned to utilize not only the global information, but also very local part of the spectrum.

Full eigendecomposition takes $O(n^3)$ time and $O(n^2)$ space. While for graphs with $\Delta(G) \ll n$ the sparse structure of the Laplacian allows to reduce the complexity to $O(n^2)$, it is still prohibitive for large graphs. Instead, we compute $k \ll n$ top and bottom eigenvalues, and use interpolation in between. This reduces complexity to $O(n^2k)$ in the general case and to $O(nk)$ in the case of bounded degree graphs.

5 EXPERIMENTS

We evaluate SGR on classification and clustering tasks on a variety of real graph collections. We compare against state-of-the-art kernels and graph representation methods, in terms of accuracy and running time. In order to ensure experimental repeatability we provide data, parameters, and source code¹.

Experimental setup. We ran experiments on a 20-core Intel Xeon CPU E5-2640v4, 3.20GHz machine with 256Gb RAM. Unless otherwise stated, we repeat each experiment 100 times and report the average across all trials. SGR interpolates the spectrum of the normalized Laplacian of each graph in the collection through cubic splines; we use 256 values uniformly sampled in the interpolated spectrum.

We compare SGR against representative graph kernel methods: the Shortest-Path (SP) [6] kernel, the Graphlet kernel (GK-4) [18], the Weisfeiler-Lehman kernel (WL), and the state-of-the-art Multi-scale Laplacian Graph kernel (MLG) [14], using default parameters for each method. We also compare SGR against NETSIMILE [4] and FGSD [20] graph representations. We additionally report the results of a naïve baseline spectral representation (Λ) that represents the graphs with a 256-dimensional vector sampled uniformly from a cubic spline-interpolated [9] spectrum of the normalized Laplacian.

¹Will be available upon publication

Datasets. We use 9 graph collections from the standard benchmark for Graph Kernels [12]. Such collections describe either social interactions (e.g., REDDIT-B from messages in the Reddit platform) or biological connections (e.g., protein-protein interactions in PROTEINS). The number of graphs in each collection varies from 200 (MUTAG) to 5000 (REDDIT-M-5k), while the average graph size varies from 18 (MUTAG) to 500 (REDDIT).

5.1 Classification

In our classification experiment, on each of the datasets we randomly select 80% of the data for training, and 20% for testing. We train an SVM using LibSVM [8] with default parameter $C=1$ and each kernel. For all graph representations, including SGR, we train a logistic regression classifier with default $C=1$ and L_2 regularization. Table 2 reports the classification accuracy averaged over 100 runs.

Our method attains good quality in almost all datasets except for IMDB datasets, for which FGSD outperforms SGR. Due to the small average graph size and density of these datasets, the task becomes harder for our self-supervised approach that relies on local and global graph structures. At the same time, while state-of-the-art kernels (MLG) outperform SGR, they fail to deliver results on medium and large collections in less than one day.

6 CONCLUSIONS

We introduced SGR, a lightweight and concise graph representation that is self-learned by means of a single-layer perceptron over a collection of synthetically generated graphs. In particular, SGR learns a single-layer perceptron encoding global and local graph properties as nonlinear transformations of the graphs’ Laplacian spectra; thus, it can adapt to a multitude of analytical tasks and application domains. Through extensive experimentation, we established that SGR achieves accuracy matching (or negligibly below) that of the most computationally demanding kernel methods on graph classification and clustering. In the future, we want to investigate more advanced architectures, and ways to incorporate both node and edge labels into the learning task.

REFERENCES

- [1] James Atwood and Don Towsley. 2016. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- [2] Mikhail Belkin and Partha Niyogi. 2007. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems*. 129–136.
- [3] Marcel Berger. 2012. *A panoramic view of Riemannian geometry*. Springer Science & Business Media.

- [4] Michele Berlingerio, Danai Koutra, Tina Eliassi-Rad, and Christos Faloutsos. 2013. Network similarity via multiple social theories. In *ASONAM*. 1439–1440.
- [5] Stephen Bonner, John Brennan, Ibad Kureshi, G Theodoropoulos, and AS McGough. 2016. Efficient Comparison of Massive Graphs Through The Use Of Graph Fingerprints. In *Twelfth Workshop on Mining and Learning with Graphs (MLG) Workshop at KDD'16*.
- [6] Karsten M Borgwardt and Hans-Peter Kriegel. 2005. Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*. IEEE, 8–pp.
- [7] Alexander M Bronstein, Michael M Bronstein, Leonidas J Guibas, and Maks Ovsjanikov. 2011. Shape google: Geometric words and expressions for invariant shape retrieval. *TOG* 30, 1 (2011), 1.
- [8] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *TIST* 2, 3 (2011), 27.
- [9] Paul Dierckx. 1995. *Curve and surface fitting with splines*. Oxford University Press.
- [10] Thomas Gärtner, Peter Flach, and Stefan Wrobel. 2003. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*. Springer, 129–143.
- [11] Brian Karrer and Mark EJ Newman. 2011. Stochastic blockmodels and community structure in networks. *Physical review E* (2011).
- [12] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. 2016. Benchmark Data Sets for Graph Kernels. (2016). <http://graphkernels.cs.tu-dortmund.de> <http://graphkernels.cs.tu-dortmund.de>.
- [13] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*. 972–981.
- [14] Risi Kondor and Horace Pan. 2016. The multiscale laplacian graph kernel. In *Advances in Neural Information Processing Systems*. 2990–2998.
- [15] Facundo Mémoli. 2011. A spectral notion of Gromov–Wasserstein distance and related methods. *Applied and Computational Harmonic Analysis* 30, 3 (2011), 363–401.
- [16] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*.
- [17] Giannis Nikolentzos, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Matching Node Embeddings for Graph Similarity. In *AAAI*. 2429–2435.
- [18] Nino Shervashidze and Karsten M Borgwardt. 2009. Fast subtree kernels on graphs. In *Advances in neural information processing systems*. 1660–1668.
- [19] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. 2011. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research* 12, Sep (2011), 2539–2561.
- [20] Saurabh Verma and Zhi-Li Zhang. 2017. Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs. In *Advances in Neural Information Processing Systems*. 87–97.
- [21] Pinar Yanardag and SVN Vishwanathan. 2015. Deep graph kernels. In *KDD*. 1365–1374.