SGR: Self-Supervised Spectral Graph Representation Learning

Anton Tsitsulin [1] **Davide Mottin** [1] Panagiotis Karras [2]
Alex Bronstein [3] Emmanuel Müller [1]

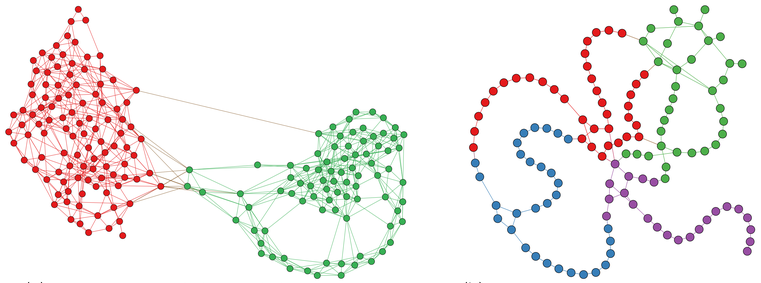
[1] Hasso Plattner Institute [2] Aarhus University [3] Technion

Expressive graph comparison is hard

Four key properties for expressive graph similarity:

- **Permutation invariance**: reordering nodes does not change similarity
- **Scale-adaptivity**: structure is captured on **both** local and global scales
- **Size-invariance**: structure of the graph **may not** depend on its size
- **Scalability**: able to deal with **both** many and big graphs

Scalability is possible with a suitable **representation** to **analyze** large graph collections (e.g., classification, clustering).



(a) Protein network. $|V| = 190, |E| = 744, d = 9$.

(b) Enzyme. $|V| = 125, |E| = 141, d = 32$.

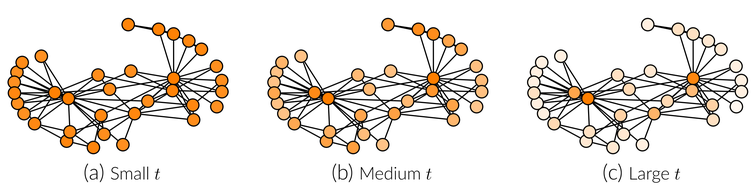
Figure 1: How can we compare these two graphs?

Graph structure at different scales

In different applications, graphs are analyzed at **different scales**:

- **Local interactions**, e. g. molecular bonds in computational biology
- **Medium-scale structure**, e. g. core-periphery in economic networks
- **Global connectivity**, e. g. community structure in social networks

We argue that **scale is a continuum**, as in Kronecker graphs.



(a) Small t (b) Medium t (c) Large t

Figure 2: Heat distribution (diagonal of H_t) at different scales on the Karate club graph.

Enjoy SGR? Try NetLSD!

poster #39 oral tomorrow @ 4PM

`anton.tsitsulin@hpi.de`

Rushing to dinner? Read this!

SGR is a **learned** graph descriptor that allows to compare graphs:

- **Fast:** in $\mathcal{O}(1)$, with $\mathcal{O}(m)$ precomputation;
- **On multiple scales:** capturing both local and global information;
- **Focused:** SGR learns which part of the spectrum to pay attention to

We take a **geometric approach** to graph comparison. We start from the optimal transport of the **heat kernel** and adapt a powerful **lower bound** introduced for manifolds.

We propose a **self-supervised** framework to learn representations, and show that our approach achieves **state-of-the-art** for classification.

Learning is better than heating

Heat kernel can be defined in terms of the Laplacian matrix $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$ that has eigenvalue decomposition $\mathcal{L} = \Phi\Lambda\Phi^\top$. Then, the heat equation is

$$\frac{\partial u_t}{\partial t} = -\mathcal{L}u_t,$$

where the timescale t encodes an **explicit notion of scale**. Solution to the heat equation is given by the **heat kernel** matrix:

$$H_t = e^{-t\mathcal{L}} = \Phi e^{-t\Lambda}\Phi^\top = \sum_{j=1}^{n} e^{-t\lambda_j}\phi_j\,\phi_j^\top,$$

Heat kernel matrix involves pairs of nodes, so we use its **trace**:

$$h_t = \operatorname{tr}(H_t) = \sum_j e^{-t\lambda_j}$$

We view h_t as a family of low-pass filters on the spectrum. We can **learn** a filter with a neural network if we **re-sample** the spectrum $\tilde{\lambda}$ as

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

where ψ is SeLU non-linearity. However, NN parameters $\Theta = (\mathbf{W}, \mathbf{b})$ still need to be trained.

We propose a **self-supervised** training that does not require any real-world data nor labels. We use a **surrogate task** of distinguishing between Erdős-Rényi and stochastic block model graphs.

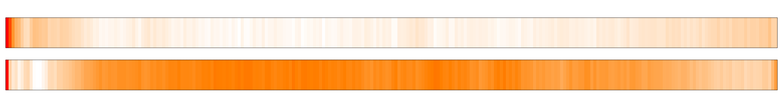


Figure 3: Gradient magnitude of the classifier highlights the learned attention to the spectrum $\tilde{\lambda}$, averaged on 600 graphs. Top: SBM; bottom: Erdős-Rényi.

Theory: computational geometry

Mémoli [1] suggests a spectral definition of the Gromov-Wasserstein distance between Riemannian manifolds. Matching a pair of points (x, x') on manifold $\mathcal{M}$ to a pair of points (y, y') on manifold $\mathcal{N}$ at scale t costs

$$\Gamma(x, y, x', y', t) = |H_t^{\mathcal{M}}(x, x') - H_t^{\mathcal{N}}(y, y')|.$$

The distance $d(\mathcal{M}, \mathcal{N})$ is defined as the infimal measure coupling

$$d(\mathcal{M}, \mathcal{N}) = \inf_{\mu} \sup_{t>0} e^{-2(t+t^{-1})} \|\Gamma\|_{L^2(\mu\times\mu)},$$

where the infimum is sought over all measures on $\mathcal{M} \times \mathcal{N}$ marginalizing to the standard measures on $\mathcal{M}$ and $\mathcal{N}$. This distance can be lower bounded by

$$d(\mathcal{M}, \mathcal{N}) \geq \sup_{t>0} e^{-2(t+t^{-1})} |h_t^{\mathcal{M}} - h_t^{\mathcal{N}}|.$$

We adopt this result *mutatis mutandis* to graphs, substituting the Laplace-Beltrami operator of the manifold with the normalized graph Laplacian.

Experiments

SGR is both versatile and expressive. Table 1 shows that SGR captures **natural properties** of real graphs. We do not use node and edge labels of graphs, and do not tune the classifiers.

dataset	WL	NetSimile	FGSD	Λ	SGR
D&D	68.27	70.02	64.88	64.54	**76.12**
ENZYMES	25.11	28.06	28.85	25.28	**33.67**
MUTAG	81.16	83.66	85.23	82.07	**86.97**
PROTEINS	72.33	70.59	63.27	71.32	**73.83**
COLLAB	**78.52**	74.26	70.66	66.15	71.98
IMDB-B	**72.26**	70.96	69.20	63.16	70.38
IMDB-M	**50.75**	46.80	48.88	41.14	47.97
REDDIT-B	71.97	86.84	87.12	76.25	**87.45**
REDDIT-M	48.57	44.96	48.51	48.02	**53.22**

Table 1: Accuracy of a logistic regression classifier.

References

[1] Facundo Mémoli.
A Spectral Notion of Gromov-Wasserstein Distance and Related Methods.
Applied and Computational Harmonic Analysis, 30(3):363--401, 2011.