Graph Minimum Factor Distance and Its Application to Large-Scale Graph Data Clustering

Jicong Fan

School of Data Science
The Chinese University of Hong Kong, Shenzhen, China

July 2025

Graph Comparison

- Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two undirected graphs in some space $\mathbb{G}$. We aim to provide a function $\text{dist} : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ to quantify the distance between G_1 and G_2.
- Graph comparison plays a crucial role in many graph analysis tasks such as graph search, classification, clustering, and generation.

Graph Comparison

- Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two undirected graphs in some space $\mathbb{G}$. We aim to provide a function dist : $\mathbb{G} \times \mathbb{G} \to \mathbb{R}$ to quantify the distance between G_1 and G_2.
- Graph comparison plays a crucial role in many graph analysis tasks such as graph search, classification, clustering, and generation.
- Popular methods for graph comparison include **graph kernels**, **graph edit distance**, **Gromov-Wasserstein distance**, etc. Most of them have high computational costs.

Motivation

- Assume that the adjacency matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ of G_1 and G_2 are generated by some kernel function k on two sets of data points denoted as matrices $\mathbf{Z}_1 \in \mathbb{R}^{m\times n_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{m\times n_2}$ respectively, i.e.,

$$[\mathbf{A}_i]_{uv} = k(\mathbf{z}_u^{(i)}, \mathbf{z}_v^{(i)}), \quad i = 1, 2,$$

where $\mathbf{z}_u^{(i)}$ denotes the u-th column of $\mathbf{Z}_i$.

Motivation

- Assume that the adjacency matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ of G_1 and G_2 are generated by some kernel function k on two sets of data points denoted as matrices $\mathbf{Z}_1 \in \mathbb{R}^{m \times n_1}$ and $\mathbf{Z}_2 \in \mathbb{R}^{m \times n_2}$ respectively, i.e.,

$$[\mathbf{A}_i]_{uv} = k(\mathbf{z}_u^{(i)}, \mathbf{z}_v^{(i)}), \quad i = 1, 2,$$

  where $\mathbf{z}_u^{(i)}$ denotes the u-th column of $\mathbf{Z}_i$.
- To quantify the distance between G_1 and G_2, we propose to calculate the distance between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and let

$$\text{dist}(G_1, G_2) := f(\mathbf{Z}_1, \mathbf{Z}_2)$$

  where $f : \mathbb{R}^{m \times n_1} \times \mathbb{R}^{m \times n_2} \to \mathbb{R}$ denotes a function to calculate the distance between two discrete distributions.

Graph Minimum Mean Factor Distance (MMFD)

- $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are unknown. But we know $\phi(\mathbf{z}_u^{(i)})^\top \phi(\mathbf{z}_v^{(i)})$, if $\mathbf{A}_i$ is PSD.
- Most graphs do not have PSD adjacency matrices. We then construct PSD proxy as

$$\mathcal{A}_i^{\phi} = \sum_{j=1}^{n_i} |\lambda_j^{(i)}| \mathbf{v}_j^{(i)} {\mathbf{v}_j^{(i)}}^\top, \quad i = 1, 2$$

where $\lambda_j^{(i)}$ and $\mathbf{v}_j^{(i)}$ are the j-th eigenvalue and eigenvector of $\mathbf{A}_i$, $i = 1, 2$. Then we have $\mathcal{A}_i^{\phi} = \mathbf{\Phi}_i^\top \mathbf{\Phi}_i$.

Graph Minimum Mean Factor Distance (MMFD)

- $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are unknown. But we know $\phi(\mathbf{z}_u^{(i)})^\top \phi(\mathbf{z}_v^{(i)})$, if $\mathbf{A}_i$ is PSD.
- Most graphs do not have PSD adjacency matrices. We then construct PSD proxy as

$$\mathcal{A}_i^\phi = \sum_{j=1}^{n_i} |\lambda_j^{(i)}| \mathbf{v}_j^{(i)} {\mathbf{v}_j^{(i)}}^\top, \quad i = 1, 2$$

  where $\lambda_j^{(i)}$ and $\mathbf{v}_j^{(i)}$ are the j-th eigenvalue and eigenvector of $\mathbf{A}_i$, $i = 1, 2$. Then we have $\mathcal{A}_i^\phi = \mathbf{\Phi}_i^\top \mathbf{\Phi}_i$.
- However, the difficulty is that $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ are usually not in the same space since they cannot be uniquely determined by $\mathcal{A}_1^\phi$ and $\mathcal{A}_2^\phi$ (or $\mathbf{A}_1$ and $\mathbf{A}_2$) respectively.

Graph Minimum Mean Factor Distance (MMFD)

- We introduce a rotation matrix $\mathbf{R}_{12}$ and let

$$f(\mathbf{Z}_1, \mathbf{Z}_2) = \min_{\mathbf{R}_{12} \in \mathcal{R}} \|\boldsymbol{\mu}_1 - \mathbf{R}_{12}\boldsymbol{\mu}_2\|$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of Φ_1 and Φ_2 respectively.

Graph Minimum Mean Factor Distance (MMFD)

- We introduce a rotation matrix $\mathbf{R}_{12}$ and let

$$f(\mathbf{Z}_1, \mathbf{Z}_2) = \min_{\mathbf{R}_{12} \in \mathcal{R}} \|\boldsymbol{\mu}_1 - \mathbf{R}_{12}\boldsymbol{\mu}_2\|$$

  where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors of Φ_1 and Φ_2 respectively.
- This leads to the following distance:

$$\begin{aligned}\mathrm{MMFD}(G_1, G_2) = & \min_{\mathbf{R}_{12} \in \mathcal{R}} \left\| \frac{1}{n_1} \sum_{j=1}^{n_1} \phi\left(\mathbf{z}_j^{(1)}\right) - \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{R}_{12} \phi\left(\mathbf{z}_j^{(2)}\right) \right\| \\ = & \left| \frac{1}{n_1} \sqrt{\sum_{uv} [\mathcal{A}_1^{\phi}]_{uv}} - \frac{1}{n_2} \sqrt{\sum_{uv} [\mathcal{A}_2^{\phi}]_{uv}} \right|\end{aligned}$$

* MMFD has a closed-form solution.

More Methods and Results

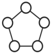
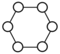
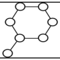
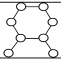
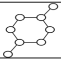
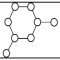
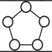
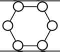
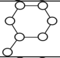
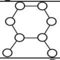
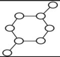
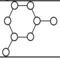
- **Extensions**
  - MMFD_{LR}: low-rank MMFD
  - MMFD-KM for large-scale clustering
  - MFD: beyond mean comparison
  - MFD-KD for large-scale clustering

More Methods and Results

- **Extensions**
  - MMFD_{LR}: low-rank MMFD
  - MMFD-KM for large-scale clustering
  - MFD: beyond mean comparison
  - MFD-KD for large-scale clustering
- **Theory**
  - Pseudo-metrics
  - Robustness
  - Low-rank approximation bound
  - Algorithmic convergence

Toy Examples of Graph Comparison

$G_1, G_2, \ldots, G_7$ from left to right:

							
	–	**0.0914**	0.1589	0.2097	0.2528	0.2505	0.2505
	0.0914	–	**0.0675**	0.1182	0.1614	0.1590	0.1591
	0.1589	0.0675	–	**0.0507**	0.0939	0.0915	0.0916
	0.2097	0.1182	0.0507	–	0.0432	**0.0408**	0.0409
	0.2528	0.1614	0.0939	0.0432	–	0.0024	**0.0023**
	0.2505	0.1590	0.0915	0.0408	0.0024	–	**0.0001**
	0.2505	0.1591	0.0916	0.0409	0.0023	**0.0001**	–

For instance, G_2 is more similar to G_3 than to G_1; G_7 lies between G_5 and G_6; the difference between G_6 and G_7 is less than the difference between G_5 and G_6.

Experiments of Graph Clustering

Method	AIDS ($N = 2000$)			PROTEINS ($N = 1113$)		
	ACC	NMI	ARI	ACC	NMI	ARI
SP kernel	79.49±0.84	0.39±0.62	-0.71±1.13	64.42±0.00	6.03±0.00	5.87±0.00
GK kernel	79.95±0.00	0.04±0.00	-0.07±0.00	59.61±0.22	0.24±0.18	0.10±0.19
RW kernel	79.90±0.00	0.09±0.00	-0.15±0.00	–	–	–
WL kernel	78.50±0.00	1.17±0.00	-2.09±0.00	60.38±0.00	1.55±0.00	0.81±0.00
LT kernel	79.95±0.00	0.04±0.00	-0.07±0.00	–	–	–
WL-OA kernel	80.40±0.00	2.46±0.00	2.38±0.00	60.38±0.00	1.55±0.00	0.81±0.00
InfoGraph+KM	92.21±0.81	54.49±3.53	63.78±3.84	59.22±0.21	3.22±1.94	0.00±0.00
InfoGraph+SC	95.65±1.55	72.21±9.20	80.17±7.19	64.02±2.31	5.17±1.87	7.06±2.65
GraphCL+KM	90.40±1.06	46.56±4.31	55.29±5.28	59.47±0.01	0.37±0.31	0.00±0.00
GraphCL+SC	96.08±1.96	72.97±10.86	81.65±8.51	59.96±0.10	2.81±0.07	3.88±0.08
JOAO+KM	88.25±0.00	38.02±0.00	44.62±0.00	59.48±0.00	0.64±0.05	-0.06±0.00
JOAO+SC	80.13±0.02	0.84±0.15	0.80±0.14	59.75±0.00	0.47±0.00	0.17±0.00
GWF+KM	96.43±1.71	74.48±9.15	84.71±7.02	66.87±2.36	9.07±1.21	11.43±3.19
GWF+SC	96.44±2.92	76.01±15.23	83.54±13.61	68.79±2.05	10.17±1.74	13.88±2.72
GLCC	79.02±0.62	4.18±2.01	5.05±2.13	60.65±2.69	2.08±1.43	4.16±2.28
DCGLC	96.77±0.33	73.51±2.30	85.74±1.45	68.89±2.04	10.90±1.35	14.32±2.88
MMD	50.10±0.00	0.00±0.00	0.03±0.00	52.56±0.00	0.08±0.00	0.14±0.00
GWD	88.30±0.00	49.73±0.00	56.45±0.00	68.82±0.00	12.42±0.00	12.37±0.00
GED	89.55±0.00	43.33±0.00	51.02±0.00	52.24±0.07	3.92±0.23	-0.23±0.03
MMFD	98.80±0.00	88.37±0.00	94.49±0.00	72.60±0.00	14.18±0.00	19.67±0.00
MMFD_{LR}	98.80±0.00	88.37±0.00	94.49±0.00	72.49±0.13	13.98±0.25	19.49±0.23
MMFD_{LR}-KM	98.96±0.02	89.62±0.18	95.25±0.11	71.87±0.18	12.74±0.34	18.51±0.28
MFD	99.45±0.00	93.82±0.00	97.47±0.00	72.60±0.00	14.18±0.00	19.67±0.00
MFD-KD	99.02±0.00	90.01±0.34	95.51±0.18	72.39±0.30	14.06±0.40	19.24±0.57

Time Cost Comparison

	G_1, G_2	$G_1, G_2, \ldots, G_N$
Shortest path kernel (Borgwardt & Kriegel, 2005)	$\mathcal{O}(n^4)$	$\mathcal{O}(N^2 n^4)$
Random walk kernel (Vishwanathan et al., 2010)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
Weisfeiler-Lehman subtree kernel (Shervashidze et al., 2011)	$\mathcal{O}(hl)$	$\mathcal{O}(Nhl + N^2 hn)$
Graph Edit Distance (Serratosa, 2014)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
(Entropic) Gromov–Wasserstein (Peyré et al., 2016)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
Sampled Gromov–Wasserstein (Kerdoncuff et al., 2021)	$\mathcal{O}(n^2)$	$\mathcal{O}(N^2 n^2)$
MMFD_{LR}	$\mathcal{O}(n^2 \log(d) + d^2 n)$	$\mathcal{O}(N(n^2 \log(d) + d^2 n) + N^2)$
MMFD_{LR}-KM	$\mathcal{O}(n^2 \log(d) + d^2 n)$	$\mathcal{O}(N(n^2 \log(d) + d^2 n) + NKT)$

Table 1: Time complexity comparison between MMFD (with $d \ll n$) and a few representative graph distances or similarities on two graphs or a set of N graphs, each with n nodes. See Appendix C.4 for the running time comparison.

Time Cost Comparison

	G_1, G_2	$G_1, G_2, \ldots, G_N$
Shortest path kernel (Borgwardt & Kriegel, 2005)	$\mathcal{O}(n^4)$	$\mathcal{O}(N^2 n^4)$
Random walk kernel (Vishwanathan et al., 2010)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
Weisfeiler-Lehman subtree kernel (Shervashidze et al., 2011)	$\mathcal{O}(hl)$	$\mathcal{O}(Nhl + N^2 hn)$
Graph Edit Distance (Serratosa, 2014)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
(Entropic) Gromov–Wasserstein (Peyré et al., 2016)	$\mathcal{O}(n^3)$	$\mathcal{O}(N^2 n^3)$
Sampled Gromov–Wasserstein (Kerdoncuff et al., 2021)	$\mathcal{O}(n^2)$	$\mathcal{O}(N^2 n^2)$
MMFD_{LR}	$\mathcal{O}(n^2 \log(d) + d^2 n)$	$\mathcal{O}(N(n^2 \log(d) + d^2 n) + N^2)$
MMFD_{LR}-KM	$\mathcal{O}(n^2 \log(d) + d^2 n)$	$\mathcal{O}(N(n^2 \log(d) + d^2 n) + NKT)$

Table 1: Time complexity comparison between MMFD (with $d \ll n$) and a few representative graph distances or similarities on two graphs or a set of N graphs, each with n nodes. See Appendix C.4 for the running time comparison.

	AIDS (N=2000)	PROTEINS (N=1113)	ENZYMES (N=600)
Shortest-path kernel	1.51	7.55	1.34
WL subtree kernel	0.81	0.90	0.38
Gromov-Wasserstein	25544.26	4549.31	1600.87
MMFD_{LR}	**0.26**	**0.61**	**0.14**

The End

Thanks for your attention!

Paper: https://openreview.net/pdf?id=hyPWP38j5k

Code: https://github.com/jicongfan/Graph-Minimum-Factor-Distance