Unsupervised Multiple Kernel Learning for Graphs via Ordinality Preservation

Yan Sun

yansun@comp.nus.edu.sg

Advisor: Stanley Kok

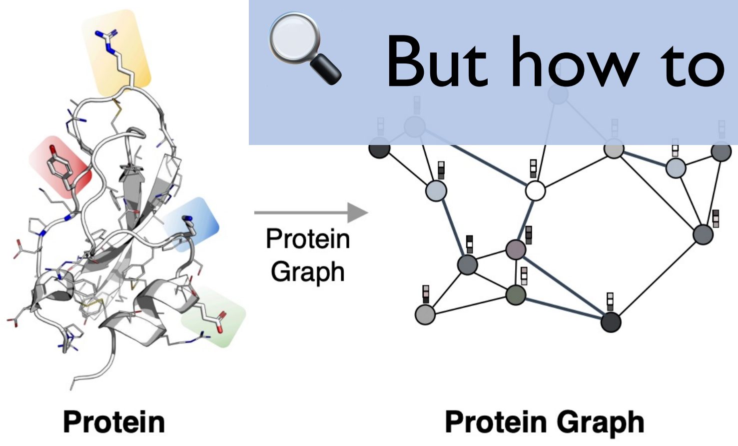
Introduction

NUS National University of Singapore
School of Computing

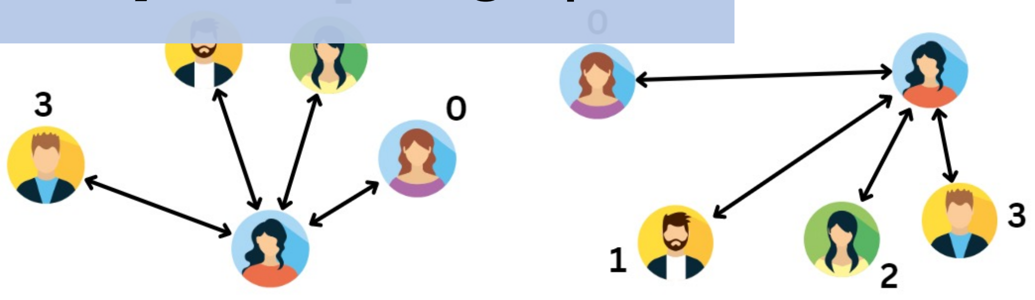
Graphs are fundamental in many fields, from *bioinformatics* to *social networks*, where graph-level* tasks receive increasing attention…

* A single graph represents one data point.

🔍 But how to **measure similarity** between graphs?



[Bioinformatics]
Protein-Protein interaction network analysis



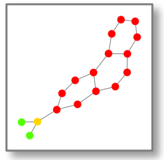
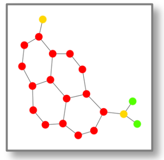
[Social Media]
Social Network Analysis

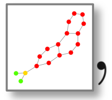
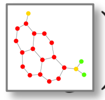
Source: ProteinGCN, Medium

Motivation

Measurements of graph similarity

- Graph Kernels $k^{(m)} : \mathcal{G} \times \mathcal{G} \rightarrow [0, 1]$

G_i  G_j 

$k^{(m)}($  ,  $)$

e.g., 0.4

- $\mathcal{R}$-convolution
- Optimal assignment
- Maximum mean discrepancy

……

Motivation

NUS National University of Singapore
School of Computing

Measurements of graph similarity

- Graph Kernels – A vast landscape
  - $\mathcal{R}$-convolution (Borgwardt & Kriegel, 2005; Shervashidze et al., 2009; Vishwanathan et al., 2010; Kriege et al., 2018)
  - Optimal assignment (Frohlich et al., 2005; Kriege et al., 2016; Togninalli et al., 2019; Chen et al., 2022)
  - Maximum mean discrepancy (Sun & Fan, 2023)
  - → Which kernel is best for a given task? Still an open question! 🧐

- Theory vs. Empirical Performance in Graph Kernels
  - ‘A’ *kernel* > ‘B’ *kernel* → More expressive in theory (Kriege et al., 2018; Oneto et al., 2017)
  - ‘A’ *kernel* < ‘B’ *kernel* → Suboptimal empirical performance (Kriege et al., 2020)
  - → Theoretical expressiveness does not guarantee the best kernel for downstream tasks.

Motivation

NUS National University of Singapore
School of Computing

Measurements of graph similarity

- Graph Kernels – A vast landscape
  - $\mathcal{R}$-convolution (Borgwardt & Kriegel, 2005; Shervashidze et al., 2009; Vishwanathan et al., 2010; Kriege et al., 2018)
  - Optimal assignment (Frohlich et al., 2005; Kriege et al., 2016; Togninalli et al., 2019; Chen et al., 2022)
  - Maximum mean discrepancy (Sun & Fan, 2023)
  - → Which kernel is best for a given task? Still an open question!

🤔 Can we ensemble multiple graph kernels for better performance in an unsupervised setting*?

* Graphs are not labeled. E.g., graph-level clustering

- Theory vs. Empirical Performan
  - *'A' kernel* > *'B' kernel* → More expressive in theory (Kriege et al., 2018; Oneto et al., 2017)
  - *'A' kernel* < *'B' kernel* → Suboptimal empirical performance (Kriege et al., 2020)
  - → Theoretical expressiveness does not guarantee the best kernel for downstream tasks.

Multiple Kernel Learning

NUS National University of Singapore
School of Computing

Definition of MKL

- *A supervised framework to learn the kernel directly from data* (Gonen & Alpaydın, 2011)

Unsupervised algorithm for MKL

Feature	UMKL (Zhuang et al., 2011)	sparse-UMKL (Mariette & Villa-Vialaneix, 2018)
Objective Function	$\min_{\mu,D} \frac{1}{2}\|X(I-K\circ D)\|_F^2 + \gamma_1 \operatorname{tr}(K\circ D\circ M) + \gamma_2\|D\|_{1,1}$	$\min_{\mathbf{b}} \operatorname{tr}(\mathbf{W}K) + \lambda\|\mathbf{b}\|_1,$ $K=\sum_{m=1}^{M} b_m K_m$
Beyond Euclidean	✗	✓
Global Topology	✗	✗
Theoretical Guarantees	✓	✗
Topology Preservation	Local reconstruction (D)	k-NN graph heuristics ($\mathbf{W}$)
Algorithm Complexity	Alternating minimization $O(I\cdot(MN^2+N^3))$	Quadratic programming solver $O(I\cdot(MN^2\log N+M^3))$

Insights:

1. Preserving data topology is essential ‼️
2. Yet still locally heuristic ⚠️
3. Poor empirical performance than individual kernels 👎 *(see later)*

Multiple Kernel Learning

NUS National University of Singapore
School of Computing

Definition of MKL

- *A supervised framework to learn the kernel directly from data* (Gonen & Alpaydın, 2011)

Unsupervised algorithm for MKL

🌟 Preserve **ordinal relationship** between graphs via *kernel values*

Feature		
Objective Function	$\min_{\mu,D} \frac{1}{2}\|X(I-K\circ D)\|_F^2$ $+\gamma_1 \operatorname{tr}(K\circ D\circ M)+\gamma_2\|D\|_{1,1}$	$\min_{\mathbf{b}} \operatorname{tr}(\mathbf{W}K)+\lambda\|\mathbf{b}\|_1,$ $K=\sum_{m=1}^{M} b_m K_m$
Beyond Euclidean	✗	✓
Global Topology	✗	✗
Theoretical Guarantees	✓	✗
Topology Preservation	Local reconstruction (D)	k-NN graph heuristics ($\mathbf{W}$)
Algorithm	Alternating minimization	Quadratic programming solver
Complexity	$O(I\cdot(MN^2+N^3))$	$O(I\cdot(MN^2\log N+M^3))$

Insights:

1. Preserving data topology is essential ‼️
2. Yet still locally heuristic ⚠️
3. Poor empirical performance than individual kernels 👎 *(see later)*

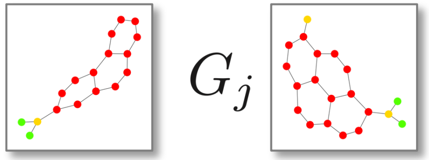
Our Model: UMKL-G

① Input

$$\mathcal{G} = \{G_i\}_{i=1}^{N}$$

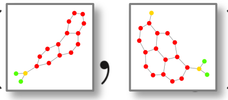
w/o labels

G_i G_j

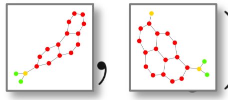


$$\mathcal{K} = \{k^{(m)}\}_{m=1}^{M}$$

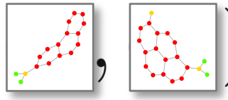
with hyperparameters

$k^{(1)}($  $)$

$\vdots$

$k^{(m)}($  $)$

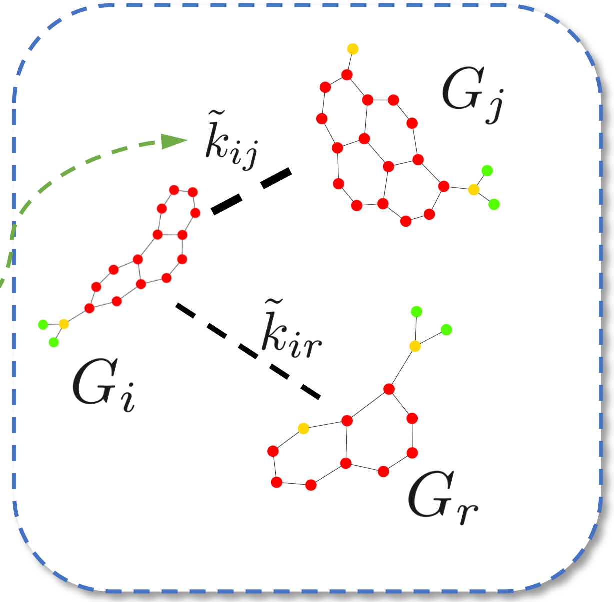
$\vdots$

$k^{(M)}($  $)$

Kernel Weights

$$\boxed{\boldsymbol{w} \in \mathbb{R}^M}$$

$$\tilde{k}_{ij} = \sum_{m=1}^{M} \boldsymbol{w}_m k^{(m)}(G_i, G_j)$$

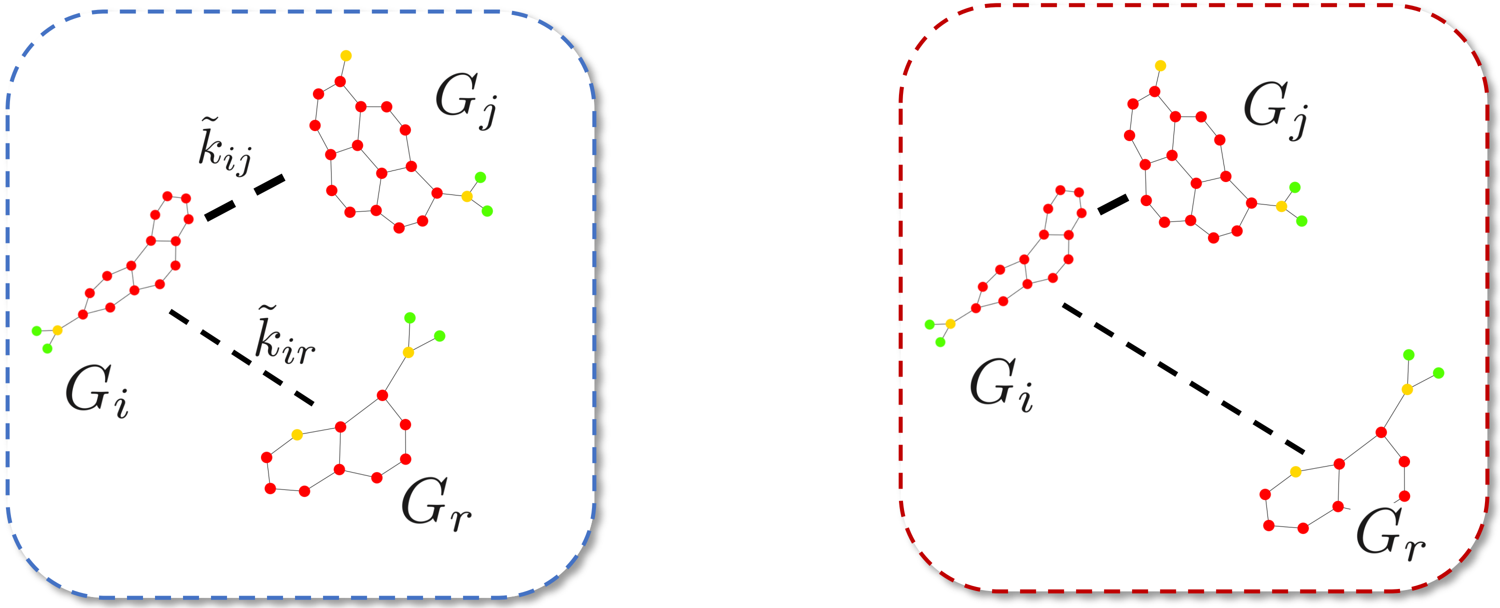


Ordinal Relationship

NUS National University of Singapore
School of Computing

Definition 1. *(Ordinal Relationship) Consider the graph G_i where its similarities to G_j and G_r are respectively given by the learned kernel values $\tilde{k}_{ij}(\boldsymbol{w})$ and $\tilde{k}_{ir}(\boldsymbol{w})$. The ordinal relationship between G_j and G_r with respect to G_i are preserved if, for any weights $\boldsymbol{w}$:*

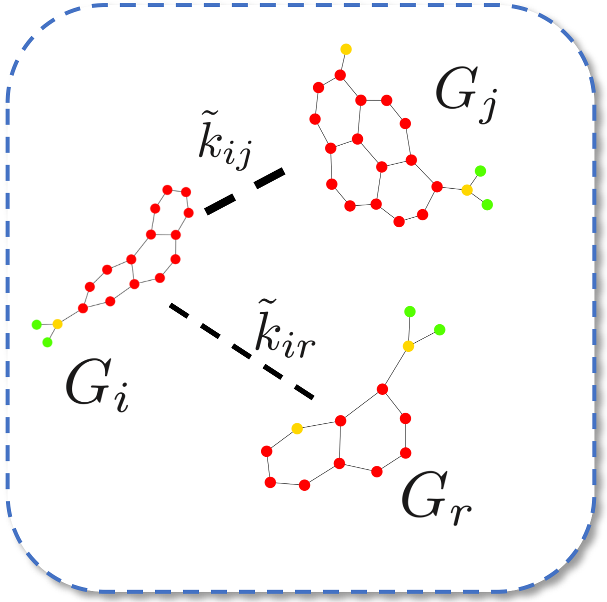
$$\tilde{k}_{ij}(\boldsymbol{w}) > \tilde{k}_{ir}(\boldsymbol{w})$$

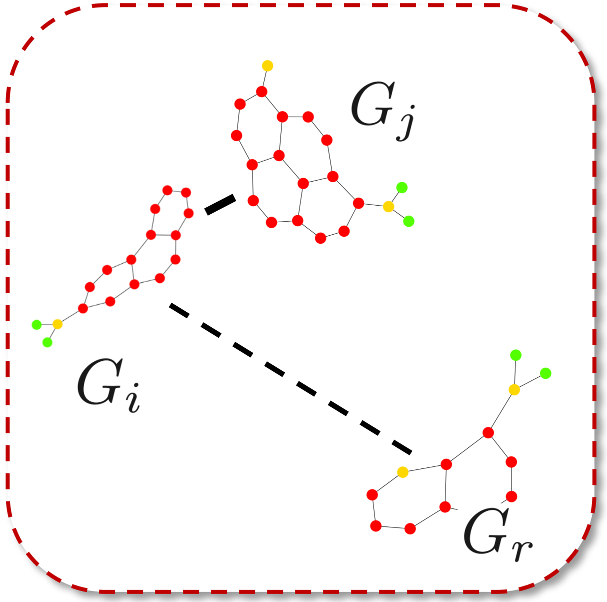


Our Model: UMKL-G

$$\boldsymbol{q}_i = (q_{i_1}, \cdots, q_{i_j}, \cdots, q_{i_N}) \in \mathbb{R}^N \quad q_{i_j} := q_{i_j}(\boldsymbol{w}) = \frac{\tilde{k}_{ij}(\boldsymbol{w})}{\sum_{j'=1}^{N} \tilde{k}_{ij'}(\boldsymbol{w})}$$

$$Q = \{\mathbf{q}_i\} \in \Delta_N$$

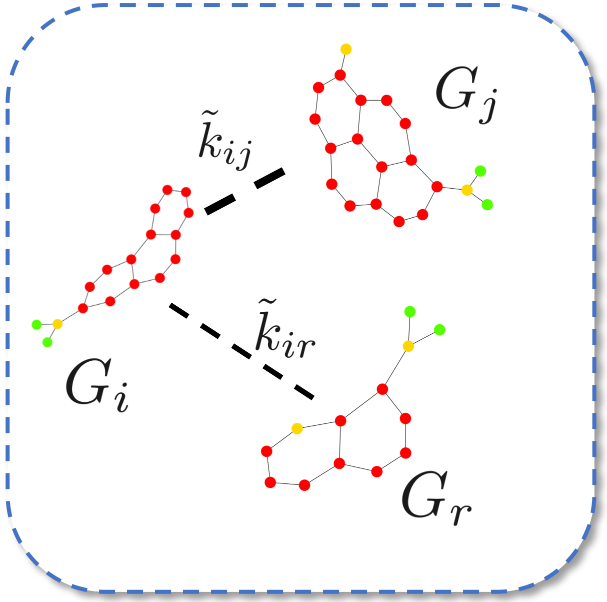


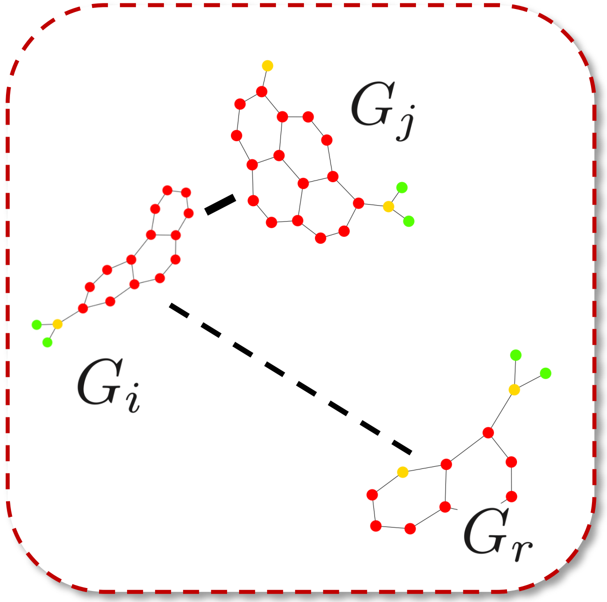


Our Model: UMKL-G

$$\boldsymbol{p}_i^{(o)} = (p_{i_1}^{(o)}, \cdots, p_{i_N}^{(o)}) \in \mathbb{R}^N \qquad p_{i_j}^{(o)} = \frac{\tilde{k}_{ij}^{o}}{\sum_{j'} \tilde{k}_{ij'}^{o}}$$

$$Q = \{\mathbf{q}_i\} \in \Delta_N \qquad P = \{\mathbf{p}_i\} \in \Delta_N$$





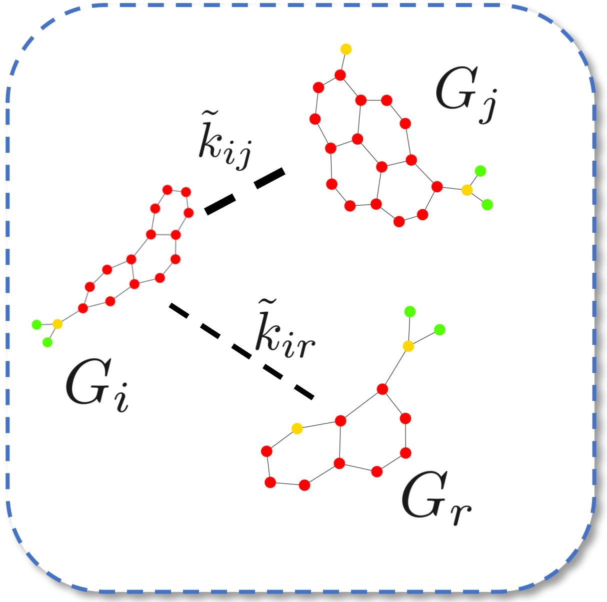
Our Model: UMKL-G

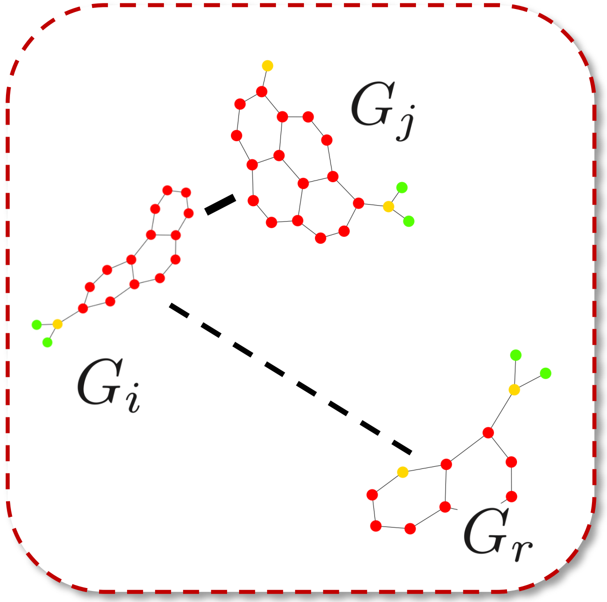
NUS National University of Singapore
School of Computing

$$\boldsymbol{w}^* = \arg\min \mathbb{KL}(Q \| P)$$

$$Q = \{\mathbf{q}_i\} \in \Delta_N$$

$$P = \{\mathbf{p}_i\} \in \Delta_N$$





Theoretical Guarantees

NUS National University of Singapore
School of Computing

Ordinality Preservation

Theorem 1. *(Ordinality Preservation) Let $\tilde{k}_{ij}$ and $\tilde{k}_{ir}$ represent the kernel values between graph G_i and graphs G_j and G_r, respectively. If the ordinal relationship $\tilde{k}_{ij} > \tilde{k}_{ir}$ holds, then for any power $o > 1$, the corresponding probabilities in the powered kernel distribution satisfy $p_{i_j}^{(o)} > p_{i_r}^{(o)}$.*

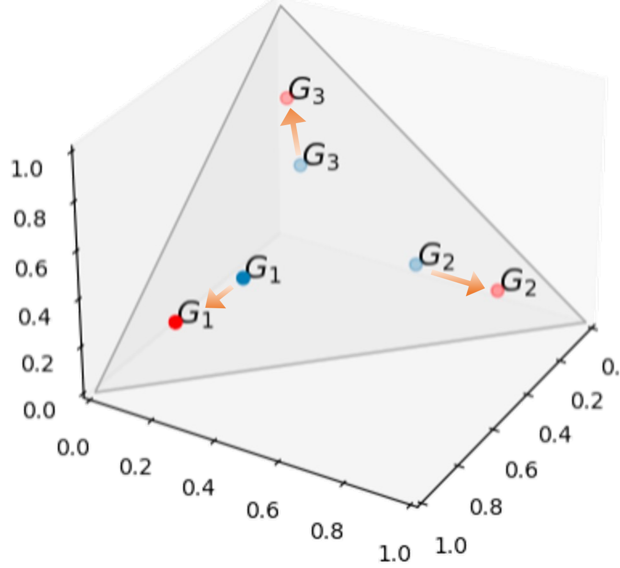
Concentration Effect

"Push to the edges of the simplex"

Theorem 2. *(Concentration Effect) For any graph G_i and any $o > 1$, the entropy of the powered kernel distribution $\mathbf{p}_i^{(o)}$ is strictly less than the entropy of the original distribution $\mathbf{q}_i$, i.e.,*

$$H(\boldsymbol{p}_i^{(o)}) < H(\boldsymbol{q}_i). \tag{4}$$

Blue Q **Red** P $\in \Delta_3$



Experiment Results

NUS National University of Singapore

School of Computing

UMKL-G consistently outperforms the baseline methods across all datasets

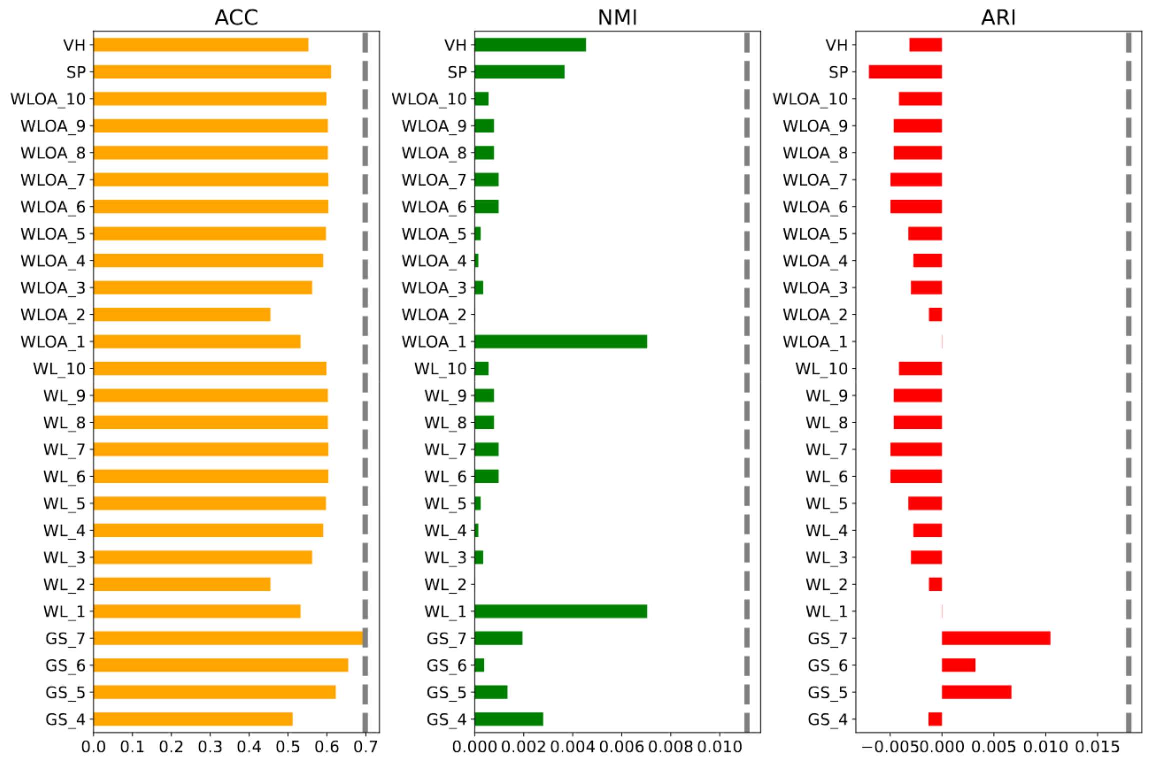
Table 2: Comparison with Baseline Methods. *The best score is in bold. The second best is underlined.*

Method	BZR			COX2			DD			DHFR		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
AverageMKL	0.7341	0.0041	0.0307	0.6167	0.0000	-0.0016	0.5764	0.0060	0.0172	0.6495	0.0000	-0.0021
UMKL	0.7341	0.0041	0.0307	0.6167	0.0000	-0.0016	0.5764	0.0060	0.0172	0.6495	0.0000	-0.0021
sparse-UMKL ($k = 10$)	0.7400	0.0040	0.0299	0.6200	0.0001	-0.0010	0.5750	0.0059	0.0170	0.6480	0.0001	-0.0020
sparse-UMKL ($k = 50$)	0.7415	0.0042	0.0305	0.6180	0.0000	-0.0015	0.5770	0.0061	0.0175	0.6498	0.0000	-0.0022
sparse-UMKL ($k = 100$)	0.7420	0.0041	0.0306	0.6175	0.0000	-0.0016	0.5768	0.0060	0.0172	0.6592	0.0000	-0.0021
UMKL-G	**0.9432**	**0.0279**	**0.0812**	**0.8009**	**0.0045**	**0.0247**	**0.5815**	**0.0098**	**0.0224**	**0.6984**	**0.0111**	**0.0180**

Method	ENZYMES			IMDB-BINARY			MUTAG			PTC_FM		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
AverageMKL	0.2617	0.0539	0.0220	0.5470	0.0152	0.0083	0.5585	0.1468	0.1946	0.8825	0.0208	0.0343
UMKL	0.2567	0.0517	0.0199	0.5470	0.0152	0.0083	0.5585	0.1469	0.1947	0.8729	0.0208	0.0343
sparse-UMKL ($k = 10$)	0.2570	0.0520	0.0201	0.5485	0.0153	0.0084	0.5590	0.1475	0.1950	0.8320	0.0210	0.0345
sparse-UMKL ($k = 50$)	0.2580	0.0518	0.0200	0.5475	0.0154	0.0085	0.5595	0.1470	0.1948	0.8373	0.0211	0.0344
sparse-UMKL ($k = 100$)	0.2575	0.0521	0.0198	0.5480	0.0151	0.0082	0.5588	0.1468	0.1946	0.8528	0.0209	0.0342
UMKL-G	**0.2983**	**0.0648**	**0.0399**	**0.5590**	**0.0159**	**0.0132**	**0.8455**	**0.2950**	**0.3389**	**0.8825**	**0.0394**	**0.0637**

Experiment Results

UMKL-G can beat the base graph kernels across all metrics



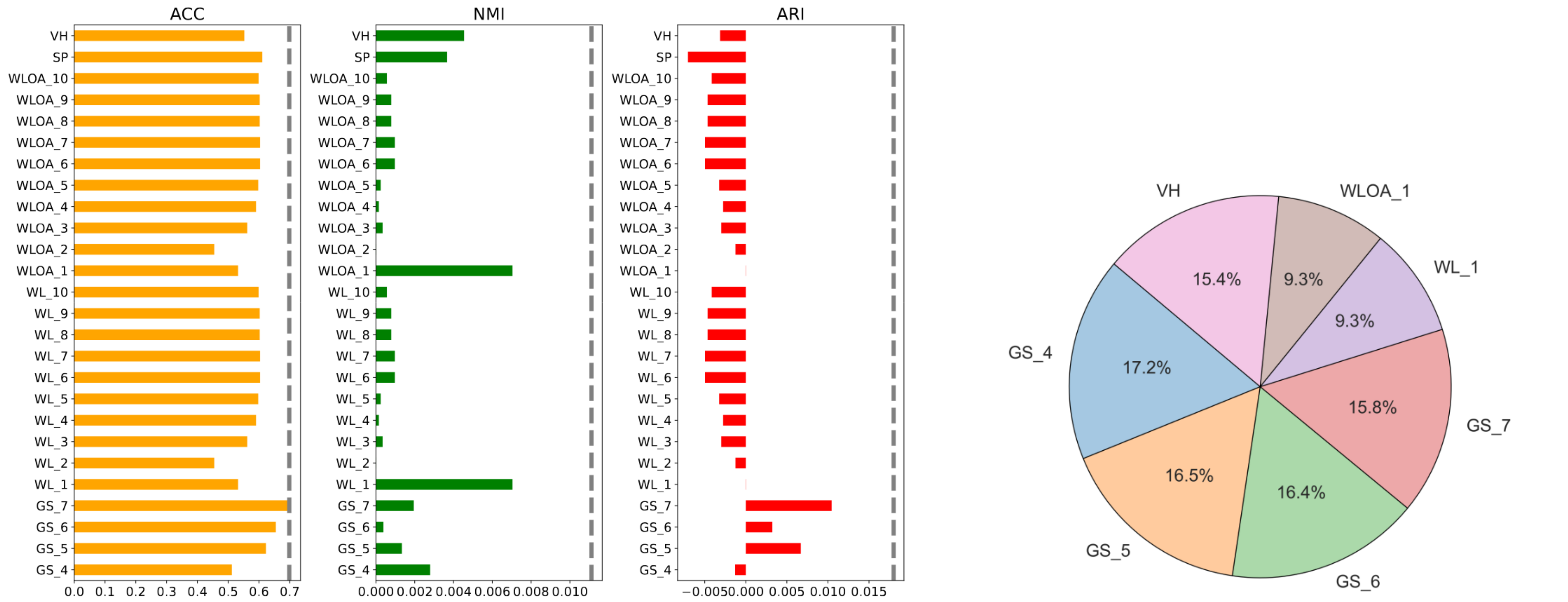
(a) Comparison with Base Graph Kernels. *The bar plots represent the performance metrics for different kernels. The dashed grey lines indicate the performances of UMKL-G.*

Figure 2: Performance on the DHFR dataset. *Kernel names are shown with hyperparameters.*

Experiment Results

NUS National University of Singapore
School of Computing

UMKL-G can automatically select graph kernels and their hyperparameters



(a) Comparison with Base Graph Kernels. *The bar plots represent the performance metrics for different kernels. The dashed grey lines indicate the performances of UMKL-G.*

(b) Learned Kernel Weights of UMKL-G.

Figure 2: Performance on the DHFR dataset. *Kernel names are shown with hyperparameters.*

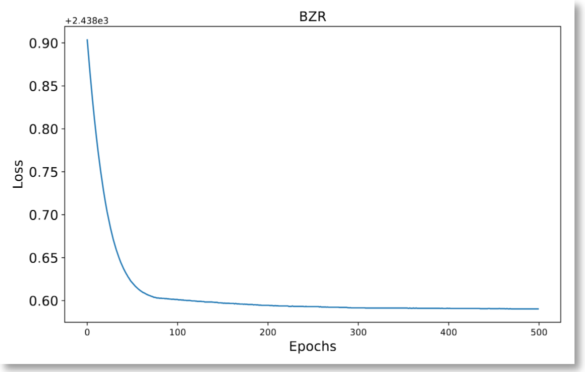
Theoretical Analysis

Smooth convergence

Theorem 3. *For the set of graphs $\mathcal{G}$ with $N = |\mathcal{G}|$ and the graphs $G_i, G_j \in \mathcal{G}$, let $\|\boldsymbol{k}_{ij}\| \leq K_{\max}$ ($\boldsymbol{k}_{ij}$ is defined for Eq. 1), $0 < \alpha \leq \sum_j \tilde{k}_{ij} \leq \beta$, and $0 < \delta \leq q_{i_j} \leq \gamma < 1$. Denote ψ_1 as $\frac{N}{\alpha^2}$, ψ_2 as $\frac{\beta+N}{\alpha^3}$, and ψ_3 as $\frac{\gamma}{\delta}$. The gradient of the objective function $\mathcal{L}^{(o)}$ is Lipschitz continuous with a constant L, such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^M$: $\|\nabla_{\boldsymbol{w}} \mathcal{L}^{(o)}(\boldsymbol{w}) - \nabla_{\boldsymbol{w}} \mathcal{L}^{(o)}(\boldsymbol{w}')\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|$ with*

$$L = C_1 \cdot N^2(1 + \gamma N) \cdot K_{\max}^2, \tag{6}$$

where the constant $C_1 = \left(1 + (o-1)\log \delta^{-1} + \log(N\delta^{-o}) + \gamma\right) \cdot \psi_1 + (1 + (o-1)\delta^{-1} + (o + (2o-1)\psi_3^o)\psi_3^{o-1}\delta^{-1}) \cdot \psi_2$.



Robustness to small perturbations in kernel

Theorem 4. *Let the perturbed kernel values be $\boldsymbol{k}'_{ij} = \boldsymbol{k}_{ij} + \Delta k_{ij}$, where $\|\Delta k_{ij}\| \leq \eta$ for any graphs G_i and G_j. Assume $\sum_{j'} \tilde{k}_{ij'} \geq \alpha$, $\delta \leq q_{i_j} \leq \gamma$ and $\|\boldsymbol{w}\| \leq \sigma$. Denote $\mathcal{O}(\boldsymbol{w}) = 0$ as the optimal condition. The magnitude of its change $\Delta\mathcal{O}$ due to the kernel perturbations is bounded by*

$$|\Delta\mathcal{O}| \leq C_2 \cdot \eta, \tag{7}$$

where the constant $C_2 = \left((o-1)\delta + o\gamma^{o-1}\delta^o + o\right) \alpha\sigma(1 + \gamma N)$.

Table 16: Evaluation of Perturbation $\mathcal{N}(0, \sigma^2)$ in Base Kernels on DD Dataset.

σ	ACC	NMI	ARI
0.01	0.5823	0.0100	0.0215
0.001	0.5815	0.0099	0.0224
–	0.5815	0.0098	0.0224

Theoretical Analysis

NUS National University of Singapore
School of Computing

Generalization to unseen data

Theorem 5. *Denote $A_{\mathcal{G}}$ as the output of our unsupervised learning algorithm UMKL-G after training on $\mathcal{G}$. UMKL-G is uniformly ω-stable with respect to the loss function $\mathcal{L}^{(o)}$ if for any $G_i \in \mathcal{X}$, the following holds:*

$$\forall \mathcal{G} \in \mathcal{X}^N, \max_{i=1,\cdots,N} \left| \mathcal{L}^{(o)}(G_i, A_{\mathcal{G}}) - \mathcal{L}^{(o)}(G_i, A_{\mathcal{G}^{\backslash r}}) \right| \leq \omega. \tag{8}$$

Theorem 6. *Denote A as the algorithm UMKL-G, which is uniformly ω-stable, $\forall G \in \mathcal{X}$, and $\forall \mathcal{G} \in \mathcal{G}^N$. Then, for any $N \geq 1$, and any $\delta \in (0,1)$, the following bounds hold with probability at least $1-\delta$ over any $\mathcal{G}$,*

$$\text{(i)}\ \ R(A_{\mathcal{G}}) \leq \hat{R}_{EMP}(A_{\mathcal{G}}) + 2\omega + (4N\omega + c)\sqrt{\frac{\log(1/\delta)}{2N}}, \tag{9}$$

$$\text{(ii)}\ \ R(A_{\mathcal{G}}) \leq \hat{R}_{LOO}(A_{\mathcal{G}}) + \omega + (4N\omega + c)\sqrt{\frac{\log(1/\delta)}{2N}}, \tag{10}$$

where $\hat{R}_{LOO}(A_{\mathcal{G}}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^{(o)}(G_i, A_{\mathcal{G}^{\backslash i}}(G_i))$, is the leave-one-out (LOO) error estimate.

Table 22: Generalization Evaluation on BZR Dataset.

Dataset	ACC	NMI	ARI
Test	0.9407	0.0329	0.0886
All	0.9432	0.0279	0.0812

Table 23: Generalization Evaluation on DD Dataset.

Dataset	ACC	NMI	ARI
Test	0.5658	0.0076	0.0148
All	0.5815	0.0098	0.0224

Table 24: Generalization Evaluation on COX2 Dataset.

Dataset	ACC	NMI	ARI
Test	0.8043	0.0048	0.0258
All	0.8009	0.0045	0.0247

Other Robustness Checks

Alternative initial weights

- Equal ✓
- (Inverse) eigenvalues ✓
- Random from Dirichlet distribution ✓

Varying hyperparameter o

- Robust performance: $o = 2\ /3\ /4$ ✓

Conclusion

NUS National University of Singapore
School of Computing

We propose ***UMKL-G****, an unsupervised algorithm to measure similarity between graphs*

- *combining multiple* ***graph kernels***
- *focusing on* ***ordinal relationships*** *to preserve the topological structure between graphs*

Feature	UMKL	sparse-UMKL	UMKL-G (Ours)
Objective Function	$\min_{\mu,D} \frac{1}{2}\|X(I - K \circ D)\|_F^2$ $+\gamma_1 \operatorname{tr}(K \circ D \circ M) + \gamma_2 \|D\|_{1,1}$	$\min_{\mathbf{b}} \operatorname{tr}(\mathbf{W}K) + \lambda\|\mathbf{b}\|_1,$ $K = \sum_{m=1}^{M} b_m K_m$	$\min_{\mathbf{w}} L^{(o)} = \mathrm{KL}(Q\|P),$ $Q_{ij} = \frac{\tilde{k}_{ij}}{\sum_{j'} \tilde{k}_{ij'}},\quad P_{ij} = \frac{\tilde{k}^o_{ij}}{\sum_{j'} \tilde{k}^o_{ij'}}$
Beyond Euclidean	✗	✓	✓
Global Topology	✗	✗	✓
Theoretical Guarantees	✓	✗	✓
Topology Preservation	Local reconstruction (D)	k-NN graph heuristics ($\mathbf{W}$)	Ordinal relationships
Algorithm	Alternating minimization	Quadratic programming solver	KL divergence
Complexity	$O(I \cdot (MN^2 + N^3))$	$O(I \cdot (MN^2 \log N + M^3))$	$O(I \cdot (MN^2 + M \log M))$

Table 3: Comparison of UMKL, sparse-UMKL, and UMKL-G.

Thank you!

Scan to Find More 👉

PAPER

CODE