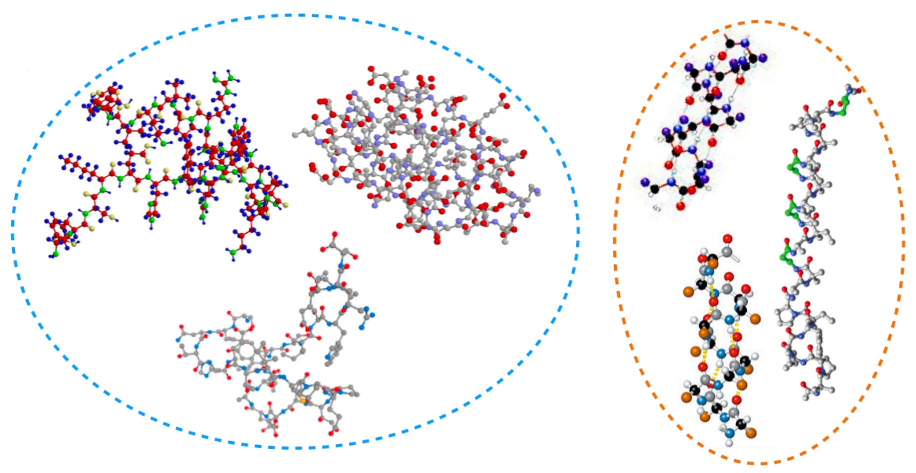
Graphon based Clustering and Testing of Networks: Algorithms and Theory

Mahalakshmi Sabanayagam,
Leena C. Vankadara,
Debarghya Ghoshdastidar

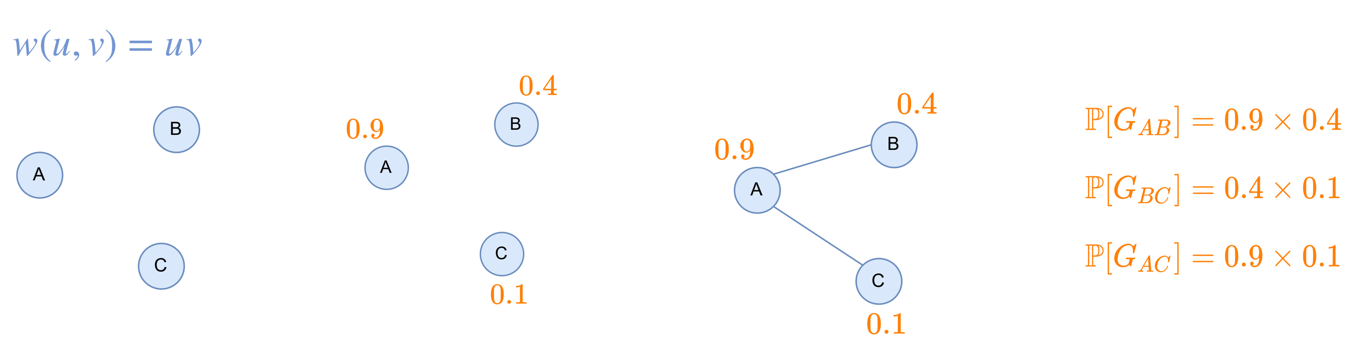
Problem Setting: Graph Clustering



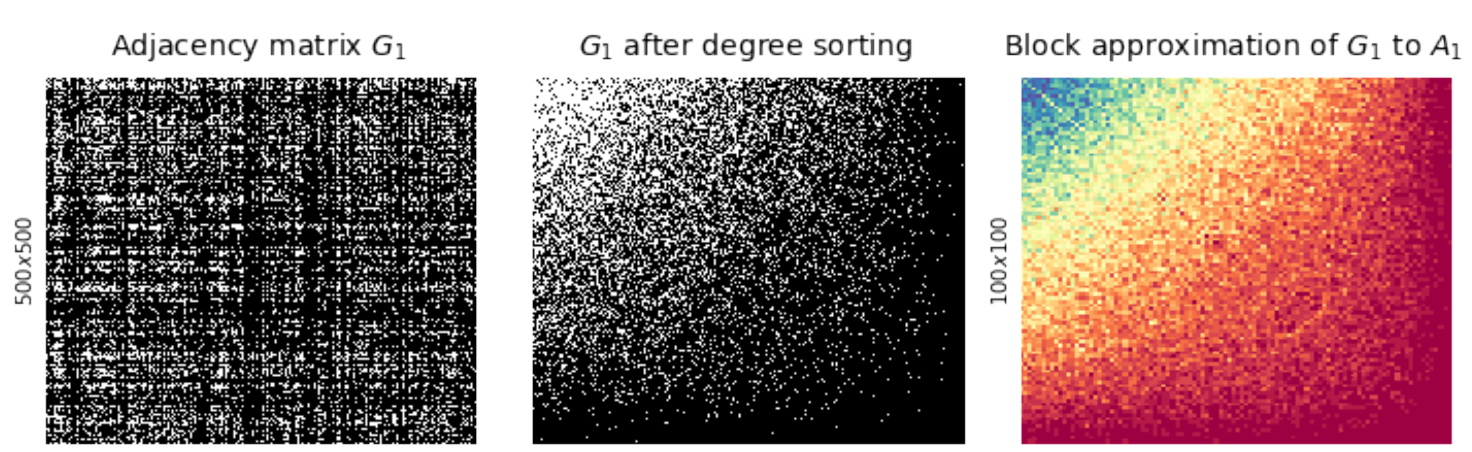
Cluster m graphs $G_1, \ldots, G_m$ of *different sizes* into K groups

Graphon based Graph Distance

Graphon: Random graph model $w : [0,1]^2 \to [0,1]$
Sample graph G with n nodes:
$U_1, \ldots, U_n \sim Uniform[0,1]$
$G_{ij} \mid U_i, U_j \sim Bernoulli(w(U_i, U_j))$ for all $i < j$



Graph transformation to fixed size representation $n_0 \times n_0$, $n_0 \ll$ #nodes



Proposed graph distance: $d(G_i, G_j) = \frac{1}{n_0}\|A_i - A_j\|_F$

Graph distance is consistent.
$G_1 \sim w_1$ and $G_2 \sim w_2$ then w.h.p.
$d(G_1, G_2) = \|w_1 - w_2\|_{L_2} + \mathcal{O}\left(\frac{1}{n_0}\right)$ for large graphs.

Graph Clustering Algorithms

Distance based Spectral Clustering (DSC)

- Use K smallest eigenvectors (in magnitude) of distance matrix $\widehat{D} \in \mathbb{R}^{m \times m}$ where $\widehat{D}_{ij} = d(G_i, G_j)$ to get K clusters

Similarity based Semi-definite Programming (SSDP)

- Similarity matrix $\widehat{S} \in \mathbb{R}^{m \times m}$ where $\widehat{S}_{ij} = \exp\left(-\frac{d(G_i, G_j)}{\sigma_i \sigma_j}\right)$
- SDP: $\widehat{X} = \max_X$ trace$(\widehat{S}X)$ s.t. $X \geq 0, X \succeq 0, X\mathbf{1} = \mathbf{1}, \mathrm{tr}(X) = K$
- Use K largest eigenvectors of $\widehat{X}$ to get K clusters

Consistency of DSC.

- $K = 2$ **clusters**

Equal number of large graphs generated from w_1 and w_2.
Then the **number of misclustered graphs** $\to 0$ w.h.p. when
$\|w_1 - w_2\|_{L_2} \geq C\frac{m}{n_0}$.

- $K > 2$ **clusters**

Large graphs $G_i \sim w_i$ then
the **number of misclustered graphs** depends on m, n_0 and K-th smallest eigenvalue of *ideal* distance matrix in magnitude.

Consistency of SSDP.

- $K \geq 2$ **clusters**

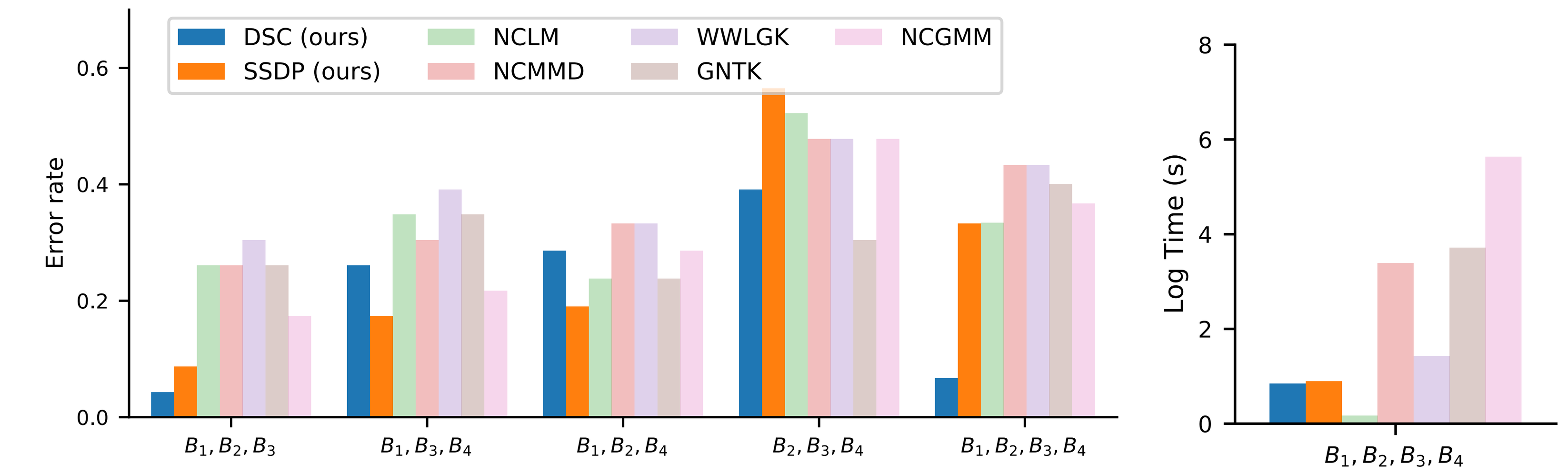
Large graphs $G_i \sim w_i$ and
$\min_{l \neq l'} \|w_l - w_{l'}\|_{L_2} = \Omega\left(\frac{m}{n_0}\right)$ then
the **number of misclustered graphs is** 0 w.h.p.

Consistency results → DSC and SSDP recover the clusters *exactly* when the underlying graphons are well separated.

Parameter free DSC and SSDP: from the consistency results
fix $n_0 = \mathcal{O}\left(\sqrt{\frac{n}{\log n}}\right)$

Empirical Performance

Compared with different embedding based, kernel based and neural network based methods.

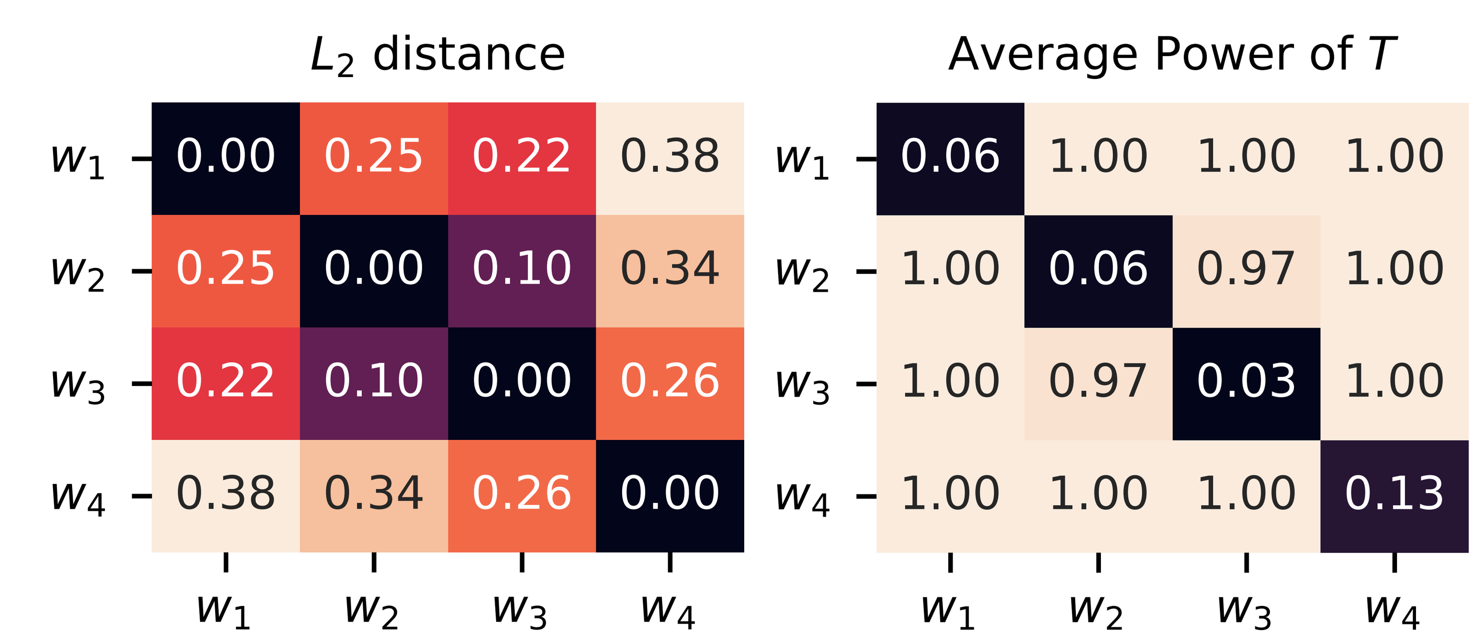


B_1 : Proteins	B_2 : KKI	B_3 : OHSU	B_4 : Peking_1
7 graphs, [273,620]	7 graphs, [62,90]	9 graphs, [140,171]	7 graphs, [86,134]

Two-sample Testing

Given graphs $G_1 \sim w_1$ and $G_2 \sim w_2$
Null Hypothesis $H_0 : \{w_1 = w_2\}$
Alternate Hypothesis $H_a : \{w_1 \neq w_2 : \|w_1 - w_2\|_{L_2} \geq \phi\}$ for some $\phi > 0$
Test $T : \mathbb{I}\{d(G_1, G_2) \geq \xi\}$ for some $\xi > 0$

Consistency of the test T.
$G_1 \sim w_1$ and $G_2 \sim w_2$ then $\exists\ \xi(\phi) > 0$ for which
Type-I and Type-II errors $\to 0$ if
$\phi \geq \frac{C}{n_0}$ for large graphs.



$W_1(u,v) = uv$
$W_2(u,v) = \exp\{-\max(u,v)^{0.75}\}$
$W_3(u,v) = \exp\{-0.5(\min(u,v) + u^{0.5} + v^{0.5})\}$
$W_4(u,v) = |u - v|$