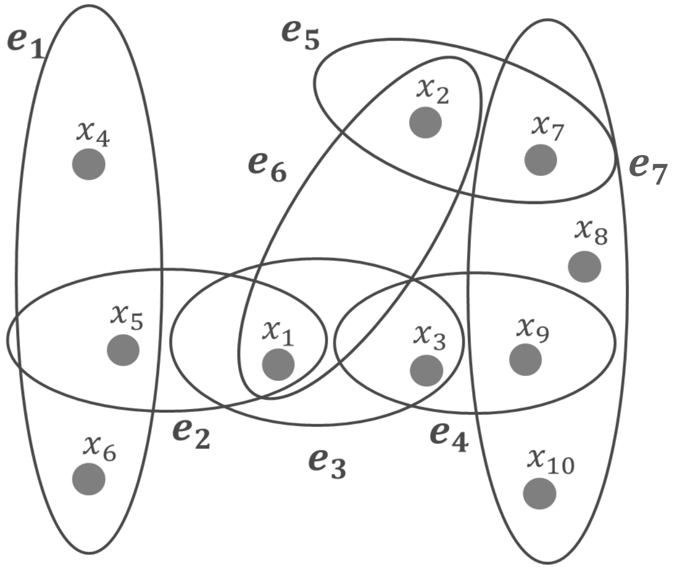
MEGA:
Multi-View Semi-Supervised Clustering of Hypergraphs

Joyce Jiyoung Whang
KAIST

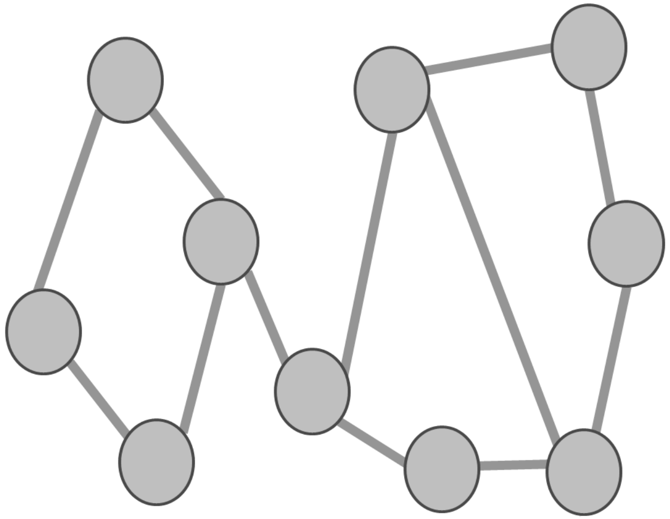
J. J. Whang, R. Du, S. Jung, G. Lee, B. Drake, Q. Liu, S. Kang, and H. Park
Proceedings of the VLDB Endowment (VLDB), 2020

Hypergraph

- A **hypergraph** is defined by a set of nodes and a set of hyperedges.
- A **hyperedge** connects two or more nodes.



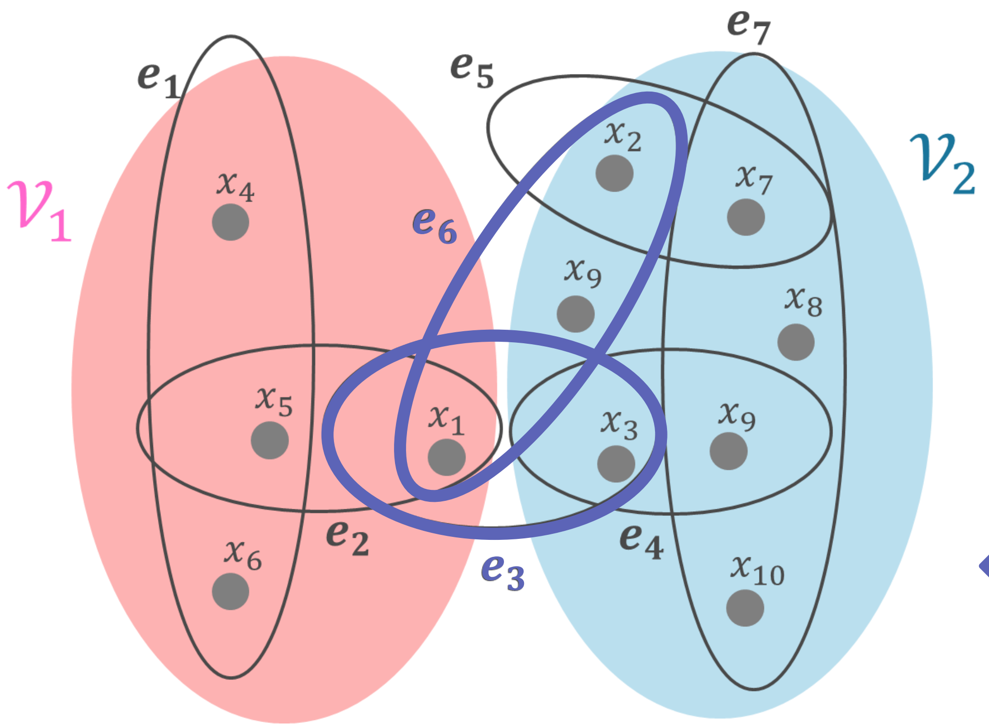
Hypergraph



Graph

Hypergraph Clustering

- **Hypergraph Normalized Cut**
  - Minimize the number of between-cluster hyperedges



$$\max_{\widetilde{\boldsymbol{Y}}\geq 0, \widetilde{\boldsymbol{Y}}^T\widetilde{\boldsymbol{Y}}=\boldsymbol{I}_k} \operatorname{trace}(\widetilde{\boldsymbol{Y}}^T \boldsymbol{D}_v{}^{-1/2}\boldsymbol{A}\boldsymbol{F}\boldsymbol{D}_e{}^{-1}\boldsymbol{A}^T\boldsymbol{D}_v{}^{-1/2}\widetilde{\boldsymbol{Y}})$$

Hypergraph normalized cut can be represented by the above trace maximization problem.

← e_3 and e_6 are between-cluster hyperedges.

Hypergraph Clustering

- Hypergraph Normalized Cut

$$\max_{\widetilde{\boldsymbol{Y}} \geq 0, \widetilde{\boldsymbol{Y}}^T \widetilde{\boldsymbol{Y}} = \boldsymbol{I}_k} \operatorname{trace}(\widetilde{\boldsymbol{Y}}^T \boldsymbol{D}_v{}^{-1/2} \boldsymbol{A}\boldsymbol{F}\boldsymbol{D}_e{}^{-1} \boldsymbol{A}^T \boldsymbol{D}_v{}^{-1/2} \widetilde{\boldsymbol{Y}})$$

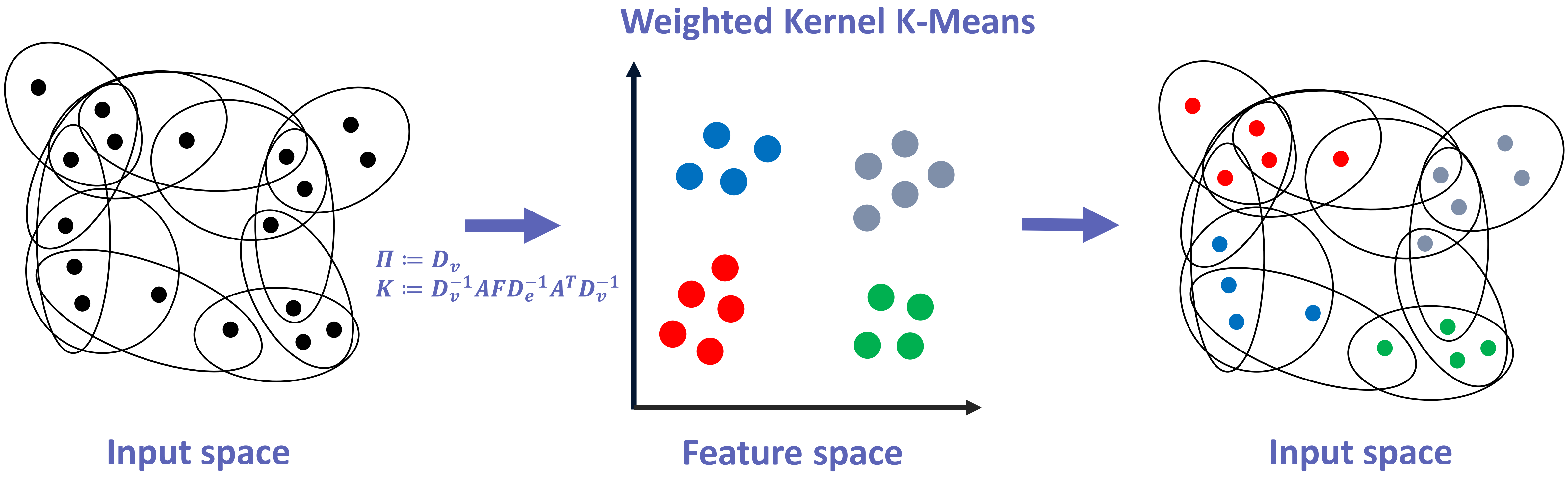
- **Weighted Kernel K-Means**

$$\max_{\widetilde{\boldsymbol{U}} \geq 0, \widetilde{\boldsymbol{U}}^T \widetilde{\boldsymbol{U}} = \boldsymbol{I}_k} \operatorname{trace}(\widetilde{\boldsymbol{U}}^T \boldsymbol{\Pi}^{1/2} \boldsymbol{K} \boldsymbol{\Pi}^{1/2} \widetilde{\boldsymbol{U}})$$

- Equivalence of the Objectives
  - $\Pi \coloneqq D_v,\ K \coloneqq D_v^{-1} A F D_e^{-1} A^T D_v^{-1}$ ← **Π : weight, K: kernel**
  - **Hypergraph normalized cut** is equivalent to **weighted kernel K-Means**

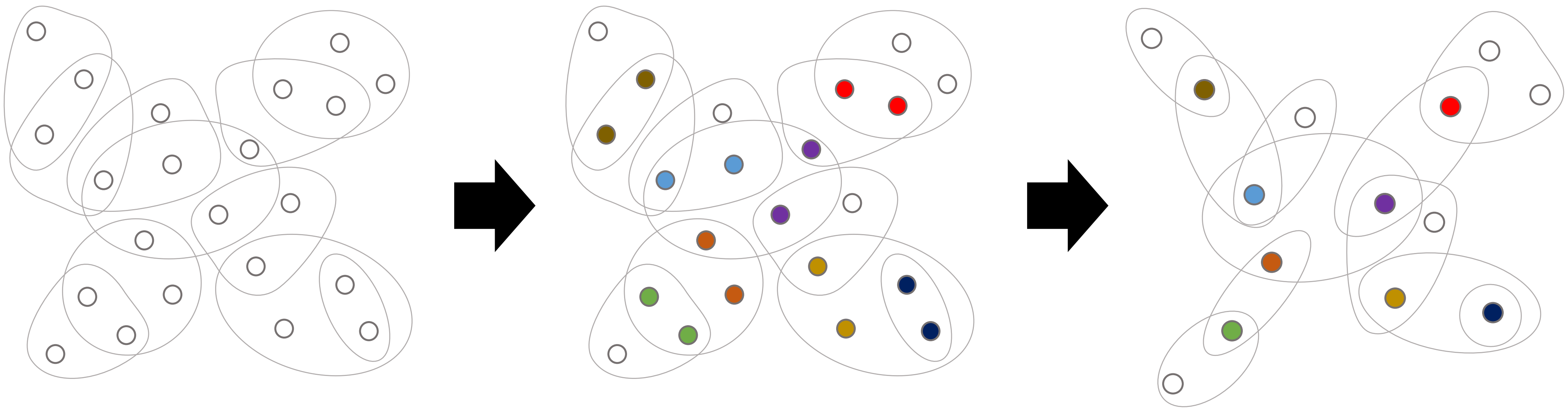
Hypergraph Clustering

- Weighted Kernel K-Means Algorithm (WKKM)
  - Optimize the **hypergraph normalized cut** using the WKKM algorithm.



Multilevel Hypergraph Clustering

- Multilevel Hypergraph Clustering Algorithm (**hGraclus**)
  - Coarsen the given hypergraph to get **a series of smaller hypergraphs**
  - Apply the WKKM algorithm multiple times at **different scales**.



Coarsening: create a smaller hypergraph by merging nodes.

Multilevel Hypergraph Clustering

- Clustering Performance
  - **hGraclus** shows the best performance

		SWS	SPC	hMetis	hGraclus
QUERY	hNCut	1.276	3.286	0.659	**0.550**
	Run Time	51.3	0.131	1.230	**0.005**
GENE	hNCut	0.720	2.361	0.512	**0.496**
	Run Time	193.5	0.267	0.519	**0.009**
CORA	hNCut	2.163	4.542	0.588	**0.512**
	Run Time	871.7	0.090	0.432	**0.008**
DBLP5	hNCut	0.937	2.920	0.206	**0.131**
	Run Time	2331.0	4.628	1.387	**0.057**
DBLP10	hNCut	2.149	6.289	0.435	**0.321**
	Run Time	8068.3	20.7	3.394	**0.114**

Revisit Hypergraph Normalized Cut

- Hypergraph Normalized Cut

$$\max_{\widetilde{\boldsymbol{Y}}\geq 0, \widetilde{\boldsymbol{Y}}^T\widetilde{\boldsymbol{Y}}=\boldsymbol{I}_k} \operatorname{trace}(\widetilde{\boldsymbol{Y}}^T \underline{\boldsymbol{D}_v{}^{-1/2}\boldsymbol{A}\boldsymbol{F}\boldsymbol{D}_e{}^{-1}\boldsymbol{A}^T\boldsymbol{D}_v{}^{-1/2}}\widetilde{\boldsymbol{Y}})$$

$$\max_{\widetilde{\boldsymbol{Y}}\geq 0, \widetilde{\boldsymbol{Y}}^T\widetilde{\boldsymbol{Y}}=\boldsymbol{I}_k} \operatorname{trace}(\widetilde{\boldsymbol{Y}}^T\boldsymbol{B}\widetilde{\boldsymbol{Y}}) \equiv \underline{\min_{\widetilde{\boldsymbol{Y}}\geq 0, \widetilde{\boldsymbol{Y}}^T\widetilde{\boldsymbol{Y}}=\boldsymbol{I}_k} \|\boldsymbol{B}-\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^T\|_F^2}.$$

- Symmetric Nonnegative Matrix Factorization (**SymNMF**)

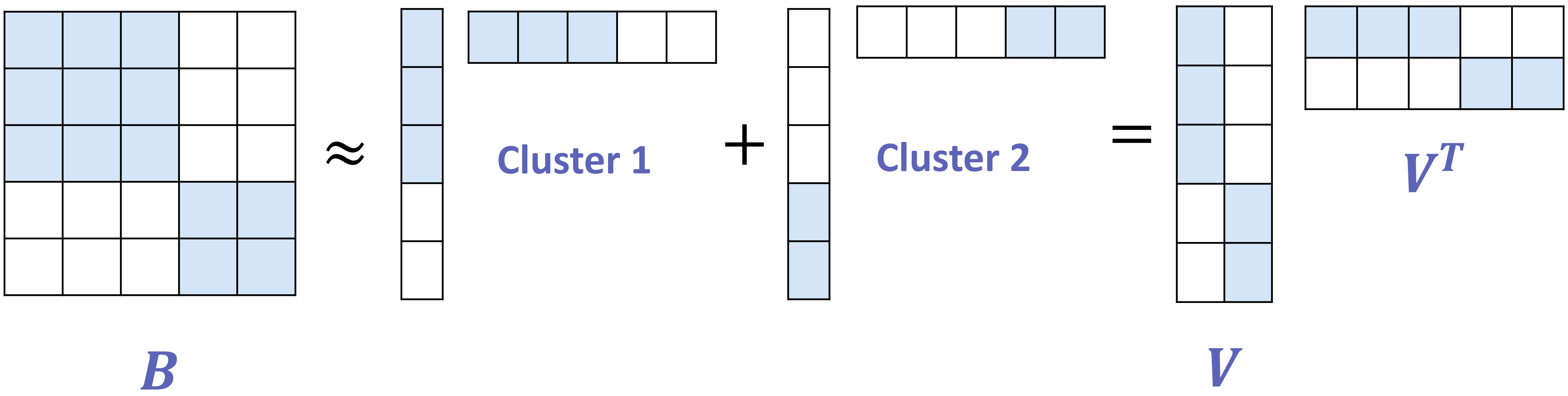
$$\underline{\min_{V\geq 0}\|B-VV^T\|_F^2} \qquad (V\in R_+^{n\times k})$$

Although the constraints on $\widetilde{Y}$ and V are different, the function to minimize is the same. The hypergraph normalized cut can be reformulated as a SymNMF problem.

Hypergraph Clustering via SymNMF

- V can be interpreted as **a clustering assignment matrix**.

$$\min_{V\geq 0} \|B - VV^T\|_F^2 \quad (V \in R_+^{n\times k}) \quad \boldsymbol{n} \textbf{ nodes and } \boldsymbol{k} \textbf{ clusters}$$

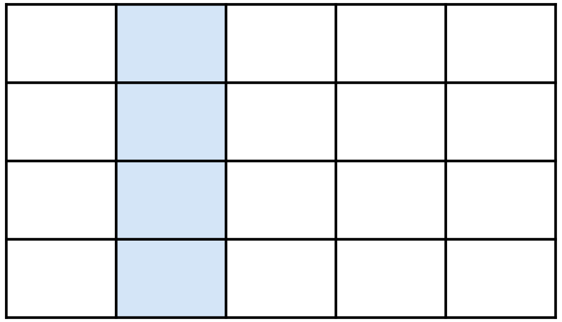


Nonnegative Matrix Factorization (NMF)

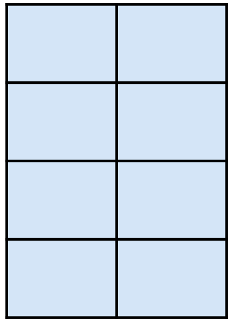
- NMF as a Clustering Method: $\min_{(W,H)\geq 0} \|X - WH\|_F^2$
  - $\boldsymbol{H}$ matrix can be interpreted as a **clustering assignment matrix**

Data matrix ($\boldsymbol{n}$ data points, $\boldsymbol{l}$ features)

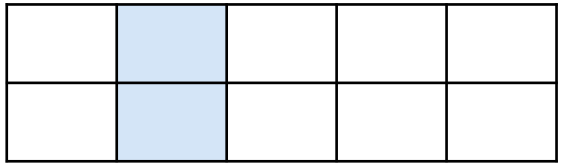
Low-rank representation ($\boldsymbol{n}$ data points, $\boldsymbol{k}$ clusters)



$\approx$



$\times$



$\boldsymbol{H} \in R_+^{k\times n}$

$\boldsymbol{X} \in R_+^{l\times n}$

$\boldsymbol{W} \in R_+^{l\times k}$

Columns of $\boldsymbol{W}$: Basis vectors

Multi-View Clustering of Hypergraphs

- **Multi-View Clustering**
  - **Hypergraph**: higher-order relationships among the objects

    → optimize the hypergraph normalized cut
  - **Auxiliary relationships** among the objects
  - **Similarity** between the objects
  - **Multiple features** or attributes of the objects

Multi-View Clustering of Hypergraphs

- **Multi-View Clustering**
  - Hypergraph: higher-order relationships among the objects
    → optimize the hypergraph normalized cut
  - Auxiliary relationships among the objects
  - Similarity between the objects

  Using SymNMF

  - Multiple features or attributes of the objects

  Using NMF

Multi-View Clustering of Hypergraphs

- Given p **feature sets** and q **relationships**
  - $X_i \in R_+^{l_i \times n}, i = 1, 2, \cdots, p,\ S_j \in R_+^{n \times n}, j = 1, 2, \cdots, q$
  - $S_j \coloneqq D_v^{-1/2} A F D_e^{-1} A^T D_v^{-1/2}$ → **Hypergraph normalized cut**
  - α_i and β_j weigh the relative importance

$$\min_{(\boldsymbol{W}_i, \boldsymbol{H}, \widehat{\boldsymbol{H}}_j) \geq 0} \sum_{i=1}^{p} \alpha_i \|\boldsymbol{X}_i - \boldsymbol{W}_i \boldsymbol{H}\|_F^2 \quad \rightarrow \text{NMF of } X_i$$

$$+ \sum_{j=1}^{q} \beta_j \left\|\boldsymbol{S}_j - \widehat{\boldsymbol{H}}_j^T \boldsymbol{H}\right\|_F^2 + \sum_{j=1}^{q} \gamma_j \left\|\widehat{\boldsymbol{H}}_j - \boldsymbol{H}\right\|_F^2 \quad \rightarrow \text{SymNMF of } S_j$$

Multi-View Clustering of Hypergraphs

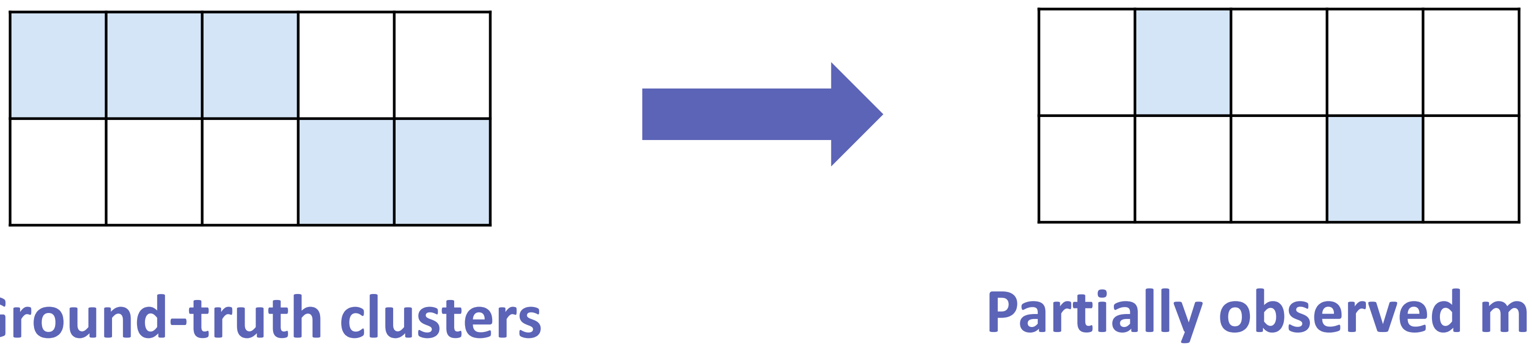
- **Multi-view Clustering Objective Function**
  - $\boldsymbol{H}$ is the shared factor: captures all the signals given by $\boldsymbol{X_i}$ and $\boldsymbol{S_j}$
  - $\boldsymbol{H}$ can be used as a **clustering assignment** matrix

$$\min_{(\boldsymbol{W}_i, \boldsymbol{H}, \widehat{\boldsymbol{H}}_j) \geq 0} \sum_{i=1}^{p} \alpha_i \left\| \boldsymbol{X}_i - \boldsymbol{W}_i \boldsymbol{H} \right\|_F^2 + \sum_{j=1}^{q} \beta_j \left\| \boldsymbol{S}_j - \widehat{\boldsymbol{H}}_j^T \boldsymbol{H} \right\|_F^2 + \sum_{j=1}^{q} \gamma_j \left\| \widehat{\boldsymbol{H}}_j - \boldsymbol{H} \right\|_F^2$$

Semi-Supervised Learning

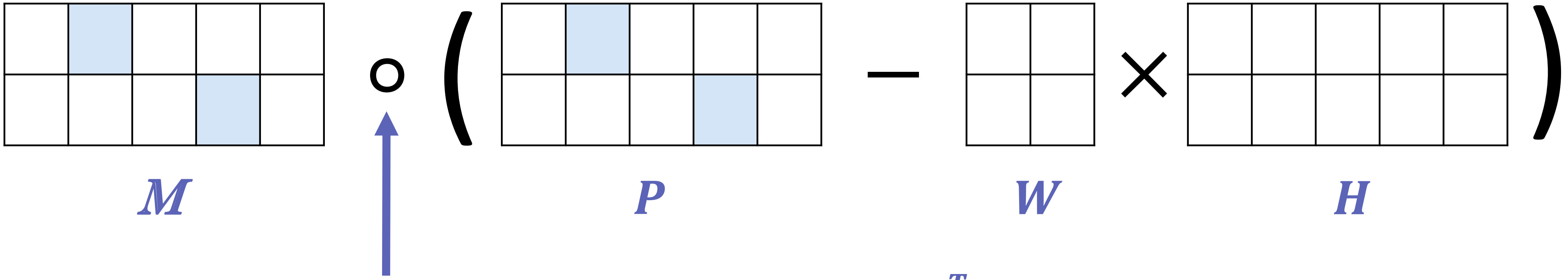
- Incorporating **partially observed labels**
  - $P \in \{0,1\}^{k\times n}$, $p_{ij} = 1$ if the j-th object belongs to the i-th cluster
  - $p_{ij} = 0$: **(i)** the j-th object does not belong to the i-th cluster **(ii)** not observed

To distinguish these two cases, we introduce a masking matrix $\boldsymbol{M}$



Semi-Supervised Learning

- **Masking Matrix $\boldsymbol{M}$**
  - $m_{ij} = 1$ if p_{ij} is observed, 0 otherwise.
  - $P \approx WH, W \in R_{+}^{k \times k}$, **$\boldsymbol{H}$ is the shared factor**.



Pairwise multiplication

Approximate p_{ij} by $w_i^T h_j$ only if p_{ij} is an observed entry

Semi-Supervised Multi-View Clustering

- The Objective Function
  - **Hypergraph Clustering**, Multi-View Clustering, Semi-Supervised Learning

    → **Optimize the hypergraph normalized cut**

$$S_j := D_v^{-1/2} A F D_e^{-1} A^T D_v^{-1/2}$$

$$\min_{(\widetilde{\boldsymbol{W}}, \boldsymbol{W}, \boldsymbol{H}, \widehat{\boldsymbol{H}}_j) \geq 0} \left\| \widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{W}} \boldsymbol{H} \right\|_F^2 + \underline{\sum_{j=1}^{q} \beta_j \left\| \boldsymbol{S}_j - \widehat{\boldsymbol{H}}_j{}^T \boldsymbol{H} \right\|_F^2}$$
$$+ \sum_{j=1}^{q} \gamma_j \left\| \widehat{\boldsymbol{H}}_j - \boldsymbol{H} \right\|_F^2 + \|\mathbb{M} \circ (\boldsymbol{P} - \boldsymbol{W}\boldsymbol{H})\|_F^2$$

Semi-Supervised Multi-View Clustering

- The Objective Function
  - Hypergraph Clustering, **Multi-View Clustering**, Semi-Supervised Learning

→ X_i and S_j are incorporated

$$\widetilde{\boldsymbol{X}} = [\sqrt{\alpha_1}\boldsymbol{X}_1; \sqrt{\alpha_2}\boldsymbol{X}_2; \cdots; \sqrt{\alpha_p}\boldsymbol{X}_p]$$

$$\min_{(\widetilde{\boldsymbol{W}}, \boldsymbol{W}, \boldsymbol{H}, \widehat{\boldsymbol{H}}_j) \geq 0} \left\| \widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{W}} \boldsymbol{H} \right\|_F^2 + \sum_{j=1}^{q} \beta_j \left\| \boldsymbol{S}_j - \widehat{\boldsymbol{H}}_j{}^T \boldsymbol{H} \right\|_F^2$$

$$+ \sum_{j=1}^{q} \gamma_j \left\| \widehat{\boldsymbol{H}}_j - \boldsymbol{H} \right\|_F^2 + \| \mathbb{M} \circ (\boldsymbol{P} - \boldsymbol{W}\boldsymbol{H}) \|_F^2$$

Semi-Supervised Multi-View Clustering

- The Objective Function
  - Hypergraph Clustering, Multi-View Clustering, **Semi-Supervised Learning**

→ Partially observed labels $\boldsymbol{P}$

$$\min_{(\widetilde{\boldsymbol{W}},\boldsymbol{W},\boldsymbol{H},\widehat{\boldsymbol{H}}_j)\geq 0} \left\|\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{W}}\boldsymbol{H}\right\|_F^2 + \sum_{j=1}^{q} \beta_j \left\|\boldsymbol{S}_j - \widehat{\boldsymbol{H}}_j{}^T \boldsymbol{H}\right\|_F^2$$

$$+ \sum_{j=1}^{q} \gamma_j \left\|\widehat{\boldsymbol{H}}_j - \boldsymbol{H}\right\|_F^2 + \|\mathbb{M} \circ (\boldsymbol{P} - \boldsymbol{W}\boldsymbol{H})\|_F^2$$

MEGA Algorithm

- **M**ulti-view s**E**mi-supervised hyper**G**r**A**ph clustering
  - An alternating minimization scheme of **block coordinate descent** (BCD)
  - Example: a 4-block coordinate descent where $p = 2, q = 1$

$$\min_{(\boldsymbol{W}_1,\boldsymbol{W}_2,\widehat{\boldsymbol{H}}_1,\boldsymbol{W},\boldsymbol{H})\geq 0} \alpha_1 \|\boldsymbol{X}_1 - \boldsymbol{W}_1\boldsymbol{H}\|_F^2 + \alpha_2 \|\boldsymbol{X}_2 - \boldsymbol{W}_2\boldsymbol{H}\|_F^2$$
$$+\beta_1 \left\|\boldsymbol{S}_1 - \widehat{\boldsymbol{H}}_1{}^T\boldsymbol{H}\right\|_F^2 + \gamma_1 \left\|\widehat{\boldsymbol{H}}_1 - \boldsymbol{H}\right\|_F^2 + \|\mathbb{M}\circ(\boldsymbol{P} - \boldsymbol{W}\boldsymbol{H})\|_F^2$$

$$\min_{\widetilde{\boldsymbol{W}}\geq 0} \left\|\boldsymbol{H}^T\widetilde{\boldsymbol{W}}^T - \widetilde{\boldsymbol{X}}^T\right\|_F^2$$

where $\widetilde{\boldsymbol{W}}^T = [\sqrt{\alpha_1}\boldsymbol{W}_1^T \sqrt{\alpha_2}\boldsymbol{W}_2^T]$, and $\widetilde{\boldsymbol{X}}^T = [\sqrt{\alpha_1}\boldsymbol{X}_1^T \sqrt{\alpha_2}\boldsymbol{X}_2^T]$,

$$\min_{\widehat{\boldsymbol{H}}_1\geq 0} \left\| \begin{bmatrix} \sqrt{\beta_1}\boldsymbol{H}^T \\ \sqrt{\gamma_1}\boldsymbol{I}_k \end{bmatrix} \widehat{\boldsymbol{H}}_1 - \begin{bmatrix} \sqrt{\beta_1}\boldsymbol{S}_1 \\ \sqrt{\gamma_1}\boldsymbol{H} \end{bmatrix} \right\|_F^2,$$

$$\min_{\boldsymbol{W}^T\geq 0} \left\|\mathbb{M}^T \circ (\boldsymbol{H}^T\boldsymbol{W}^T - \boldsymbol{P}^T)\right\|_F^2,$$

$$\min_{\boldsymbol{H}\geq 0} \left\| \begin{bmatrix} \widetilde{\boldsymbol{W}}\boldsymbol{H} \\ \sqrt{\beta_1}\widehat{\boldsymbol{H}}_1{}^T\boldsymbol{H} \\ \sqrt{\gamma_1}\boldsymbol{H} \\ \mathbb{M}\circ(\boldsymbol{W}\boldsymbol{H}) \end{bmatrix} - \begin{bmatrix} \widetilde{\boldsymbol{X}} \\ \sqrt{\beta_1}\boldsymbol{S}_1 \\ \sqrt{\gamma_1}\widehat{\boldsymbol{H}}_1 \\ \mathbb{M}\circ\boldsymbol{P} \end{bmatrix} \right\|_F^2$$

Initialization of MEGA using hGraclus

- When **MEGA** is initialized by **hGraclus**, the performance of MEGA is improved.
  - hGraclus optimizes the **hypergraph normalized cut** → SymNMF term in MEGA

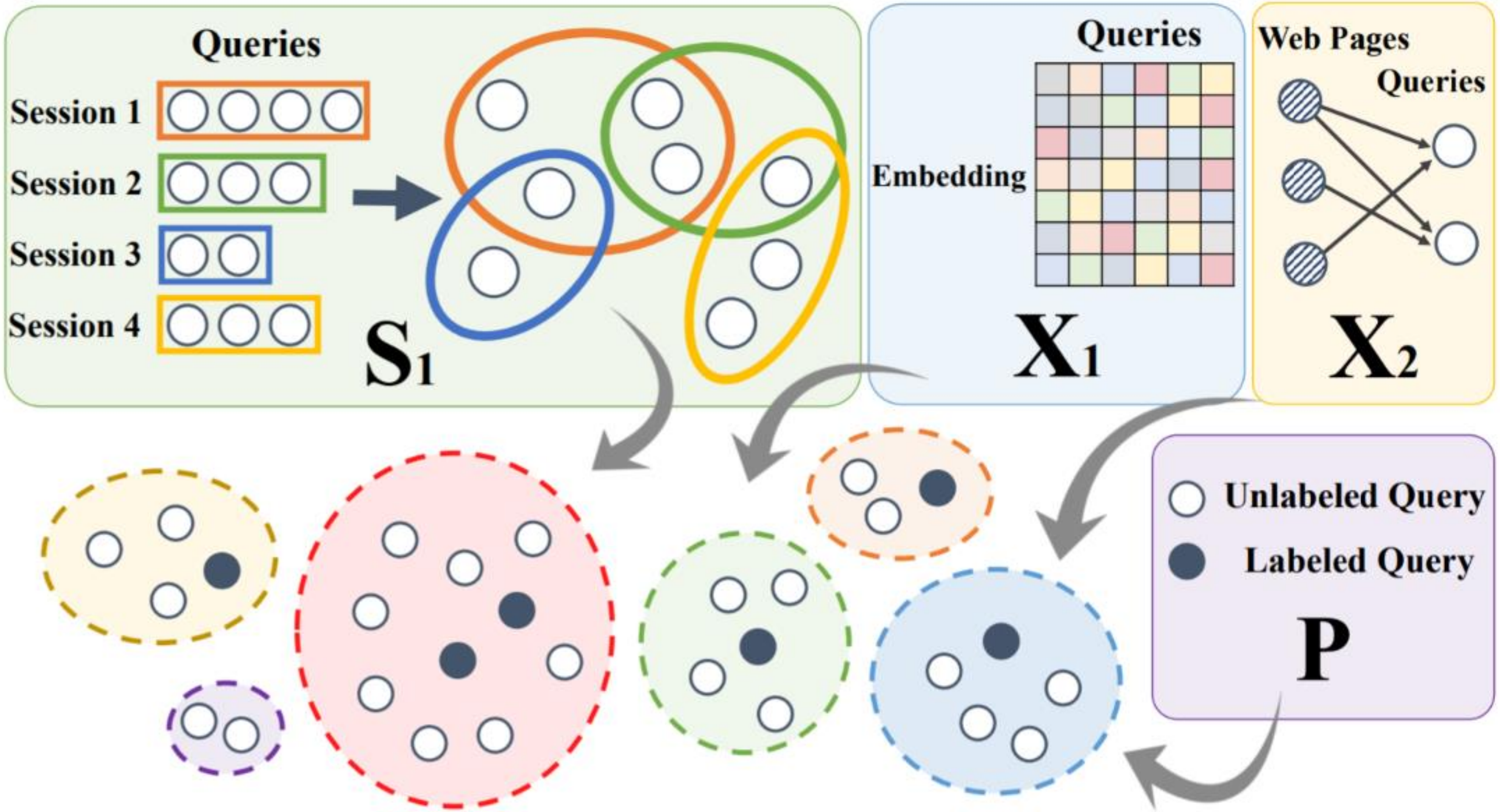
	F1 (↑)		
	random	hGraclus	Gain (%)
SYN1	87.19%	**98.06%**	12.47
SYN3	93.17%	**100.0%**	7.33
SYN5	94.84%	**100.0%**	5.44
SYN6	94.22%	**100.0%**	6.13
SYN7	78.19%	**100.0%**	27.89
CORA	65.48%	**68.58%**	4.73
DBLP5	84.41%	**86.89%**	2.94
GENE	57.16%	**58.50%**	2.34
DBLP10	**70.67%**	69.51%	-1.64
QUERY	**57.97%**	57.55%	-0.72
		Average Gain	6.69

Performance of MEGA with two different initializations: random and hGraclus

→ 5 synthetic datasets, 5 real-world datasets
→ Gain: (hGraclus-random)/random*100
→ By initializing MEGA with hGraclus, we get more accurate results.

Multi-View Semi-Supervised Clustering of Web Queries

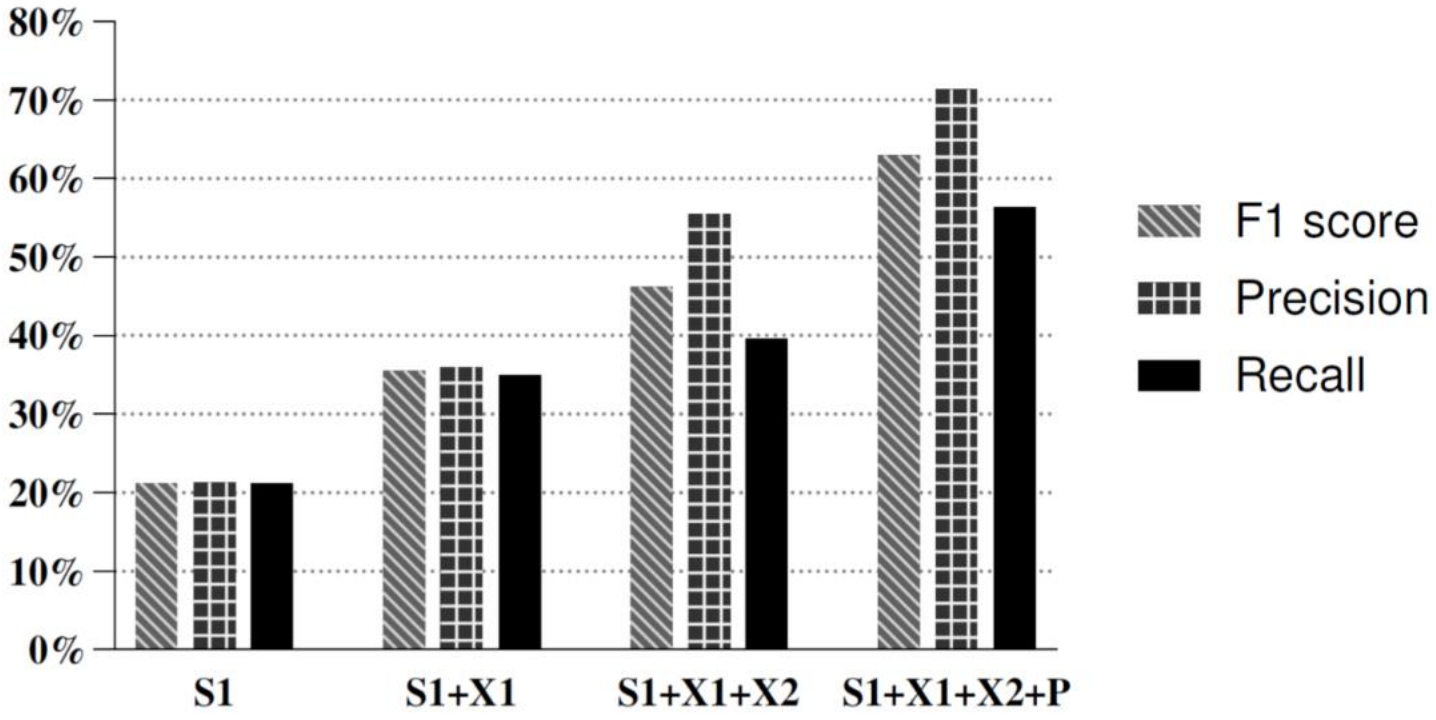
S_1: Query sessions → Hypergraph X_1: Embedding X_2: Clicked documents



P: Partial supervision
(Semi-Supervised Learning)

Multi-View Semi-Supervised Clustering of Web Queries

- **Clustering Performance** with Different Numbers of Views
  - **As an additional view is incorporated, the clustering performance is improved.**
  - Incorporating **multiple views** as well as **partial supervision** is important.



Experimental Results

- Baselines: **13 different state-of-the-art methods**
  - **Hypergraph structure only**: hGraclus, hMetis, SPC, SWS
  - **Multi-view clustering**: PCLDC, JNMF, SEC
  - **Semi-supervised clustering**: CMMC, MCCC, LGC, PLCC
  - **Multi-view semi-supervised clustering**: SMACD, MLAN

- In MEGA, all the parameters (α_i and β_j) are set to be ones.
- Initialize MEGA, PCLDC, JNMF with hGraclus.

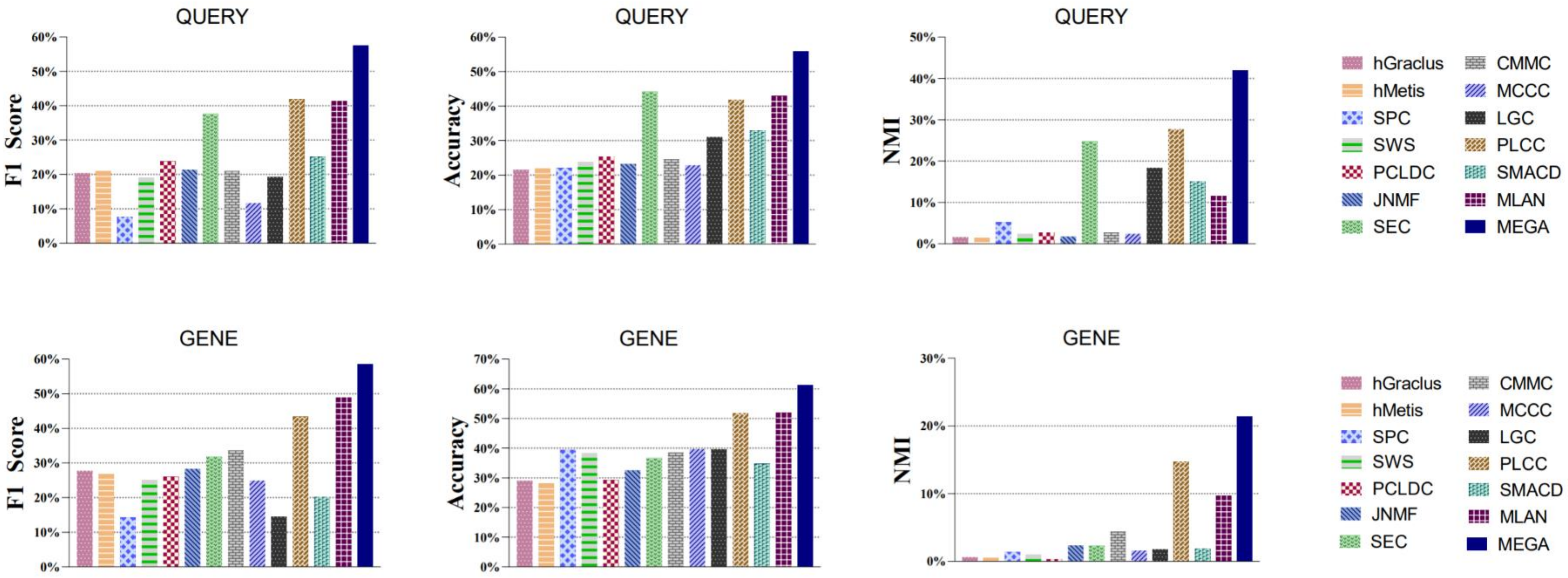
Experimental Results

- Real-World Datasets

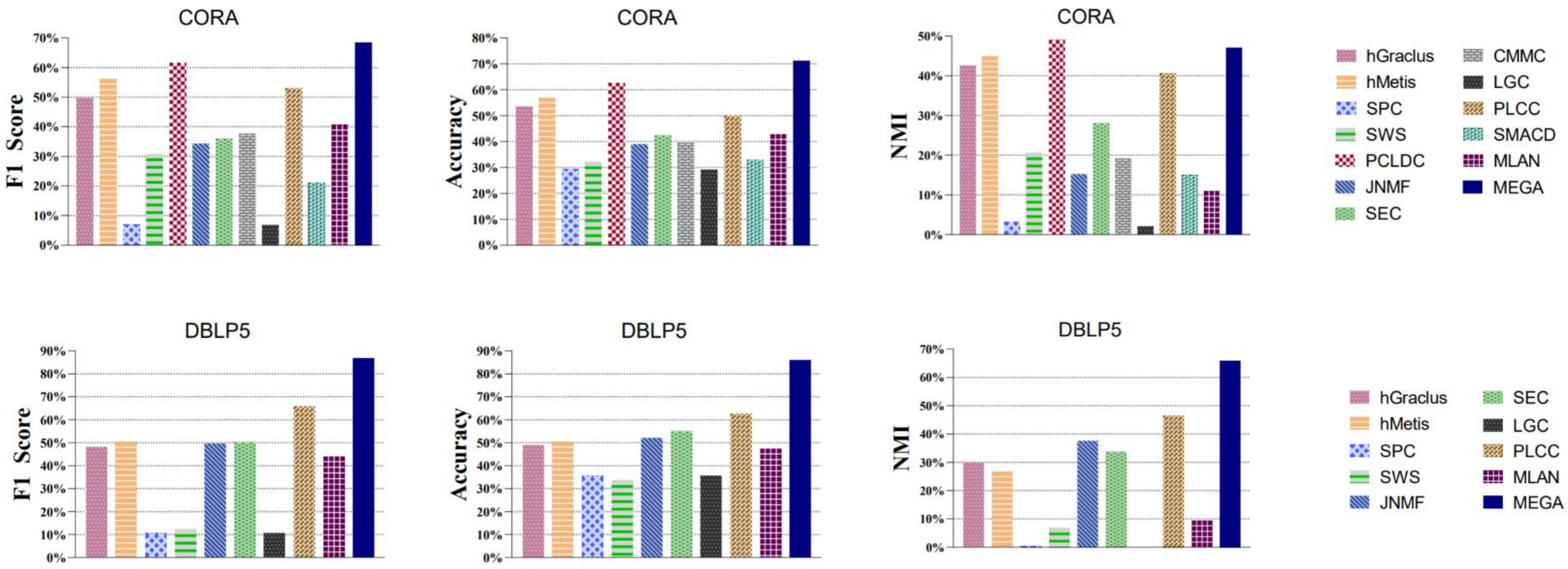
	No. of nodes	No. of hyperedges	k	Views
QUERY	481	15,762	6	$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{S}_1$, $\boldsymbol{P}$
GENE	2,014	2,023	4	$\boldsymbol{X}_1$, $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, $\boldsymbol{P}$
CORA	2,485	2,485	7	$\boldsymbol{X}_1$, $\boldsymbol{S}_1$, $\boldsymbol{P}$
DBLP5	19,756	21,492	5	$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{S}_1$, $\boldsymbol{P}$
DBLP10	42,889	34,834	10	$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{S}_1$, $\boldsymbol{P}$

GENE → S_1: gene-disease association (hypergraph),
S_2: similarity between diseases, X_1: tf-idf representations of the diseases
CORA → S_1: citation information (hypergraph), X_1: predefined keywords of papers
DBLP → S_1: collaboration information (hypergraph),
X_1: tf-idf representations of papers, X_2: citation information

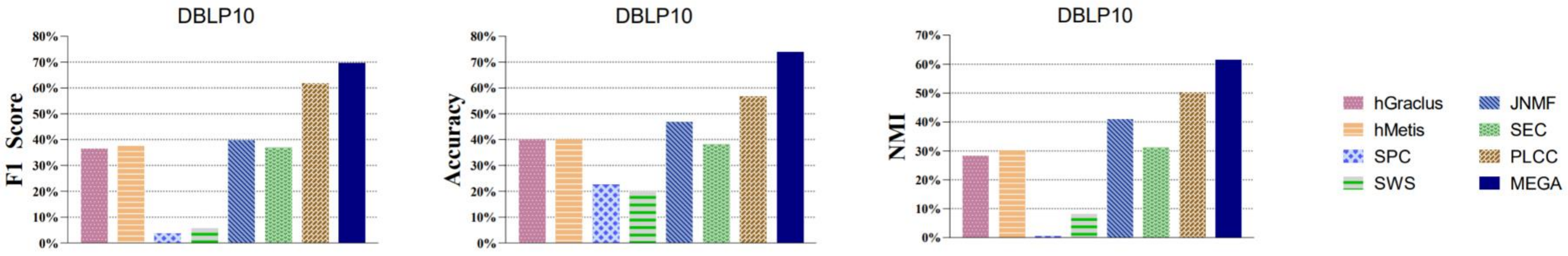
Experimental Results



Experimental Results

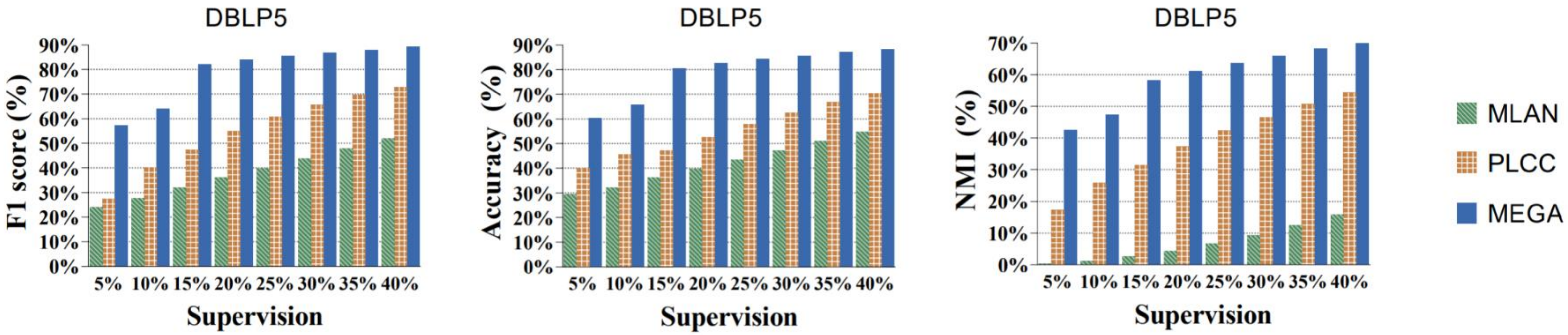


Experimental Results



Higher F1, accuracy, and NMI scores indicate better clustering results.
In terms of identifying the ground-truth clusters, MEGA outperforms the 13 different state-of-the-art methods.

Experimental Results



Performance of MLAN, PLCC, and MEGA with different levels of supervision on DBLP5. MEGA achieves better clustering performance than the other semi-supervised methods at all different levels of supervision.

Experimental Results

	MEGA						MLAN	PLCC
X_1 (α_1)	0.3	0.3	0.3	1.0	1.0	1.0		
S_1 (β_1)	0.3	0.3	1.0	0.3	1.0	1.0		
S_2 (β_2)	0.3	1.0	1.0	1.0	0.3	1.0		
F1 (%)	57.65	56.95	56.32	56.33	57.27	58.50	48.93	43.33
ACC (%)	60.10	59.47	58.41	59.92	60.41	61.22	51.94	51.77
NMI (%)	23.70	22.07	18.90	19.76	20.41	21.36	9.72	14.74

Performance of MEGA with different parameters and the two most competitive baseline methods on GENE.
The performance of MEGA does not largely fluctuate depending on the parameters, and MEGA consistently outperforms the baseline methods.

Summary

- Multilevel Hypergraph Clustering (**hGraclus**)
  - Mathematical equivalence between the **hypergraph normalized cut** and the weighted kernel K-Means objective
- Multi-view Semi-supervised Clustering of Hypergraphs (**MEGA**)
  - Optimize the **hypergraph normalized cut**
  - Incorporate **multiple attributes/features**
  - **Semi-supervised** learning
  - Initialized by hGraclus
  - Effective in identifying **the ground-truth clusters** in real-world datasets

Big Data Intelligence Lab @ KAIST

http://bdi-lab.kaist.ac.kr/

jjwhang@kiast.ac.kr