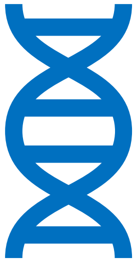
DPGS: Degree-Preserving Graph Summarization

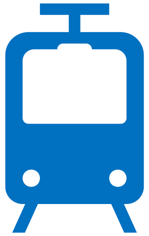
Houquan Zhou[1], Shenghua Liu[1], Kyuhan Lee[2], Kijung Shin[2], Huawei Shen[1] and Xueqi Cheng[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS)
[2]Korea Advanced Institute of Science and Technology (KAIST)

Graphs are useful

- Graph is a powerful tool to model the connection between objects.
- Many fields
  - Protein network
  - Social network
  - Transportation network
  - ...





Graphs grow larger

5.48B pages

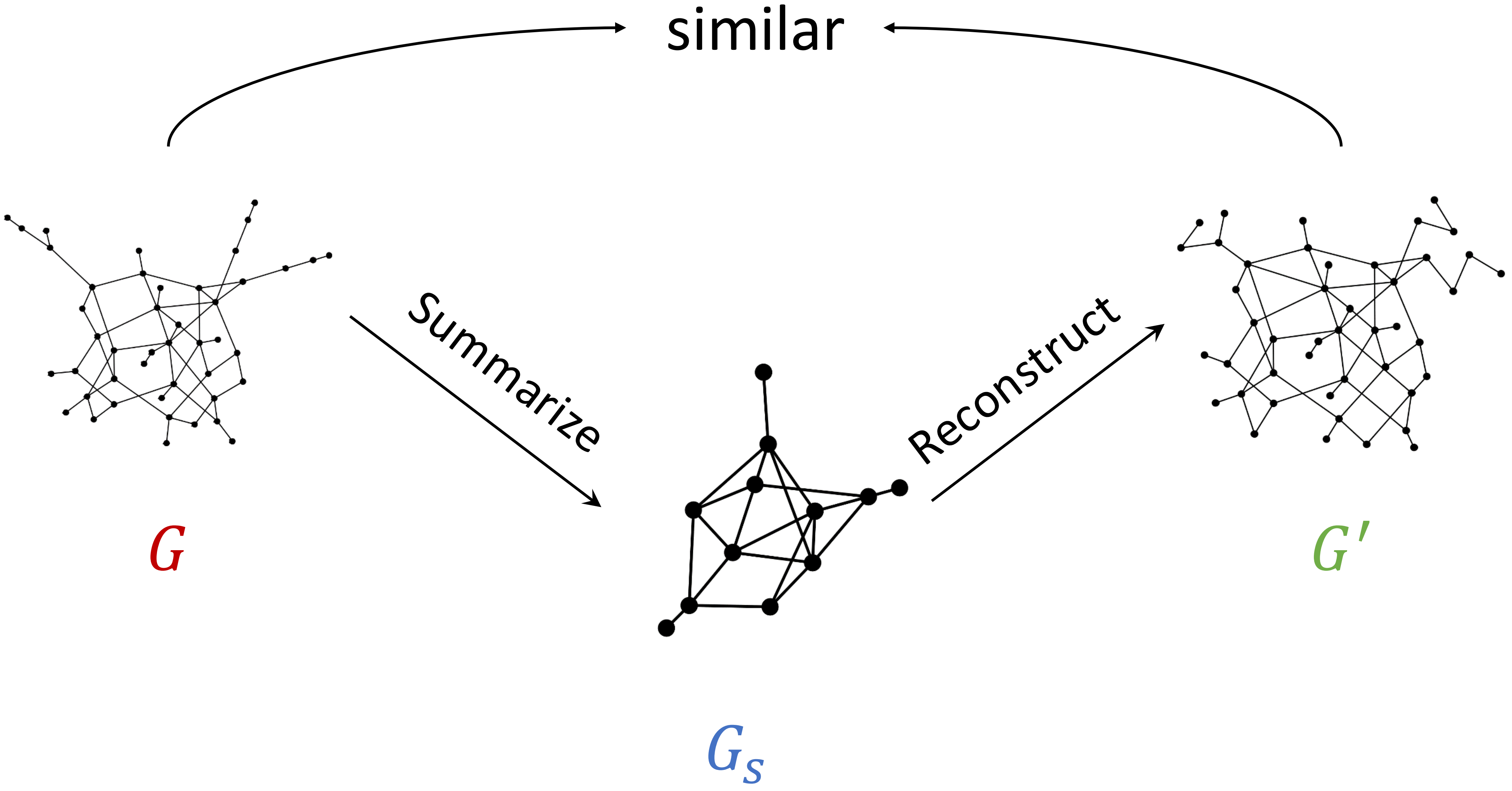
2.45B users

300M users
12M products

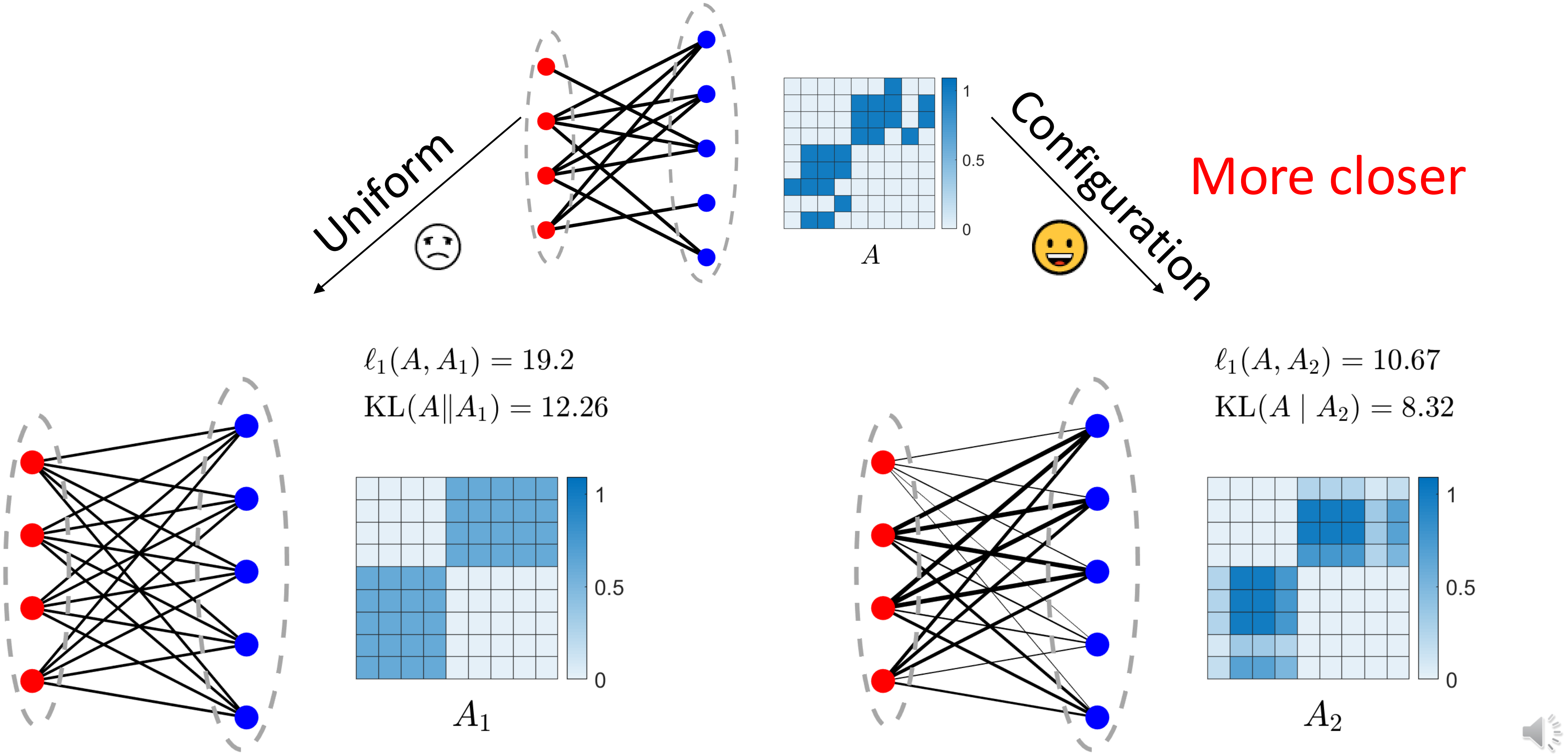
Hard to store, process and analyze.

Solution: **Graph Summarization**.

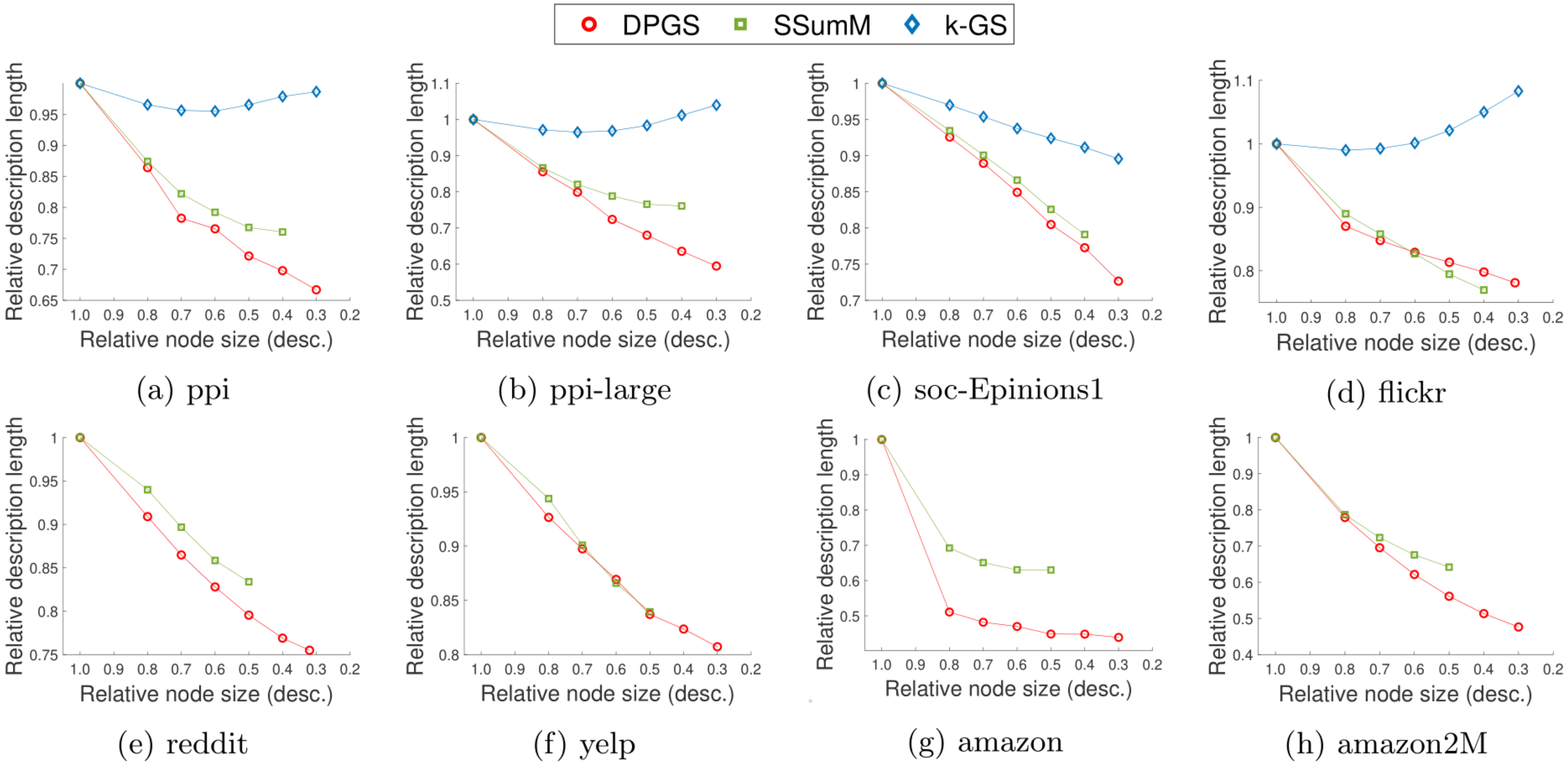
Graph Summarization



A novel reconstruction scheme



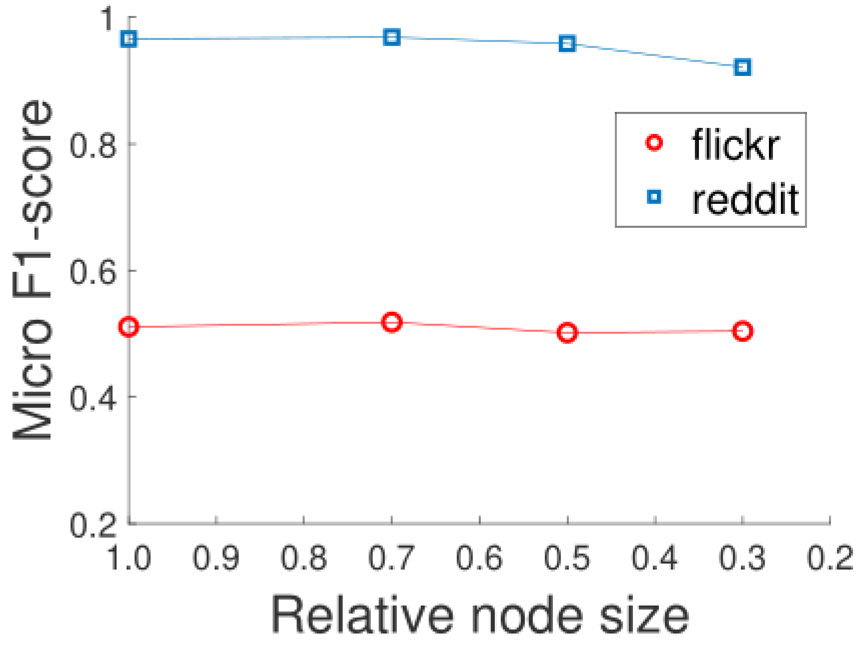
More compact summary graphs



Lower encoding length ✅

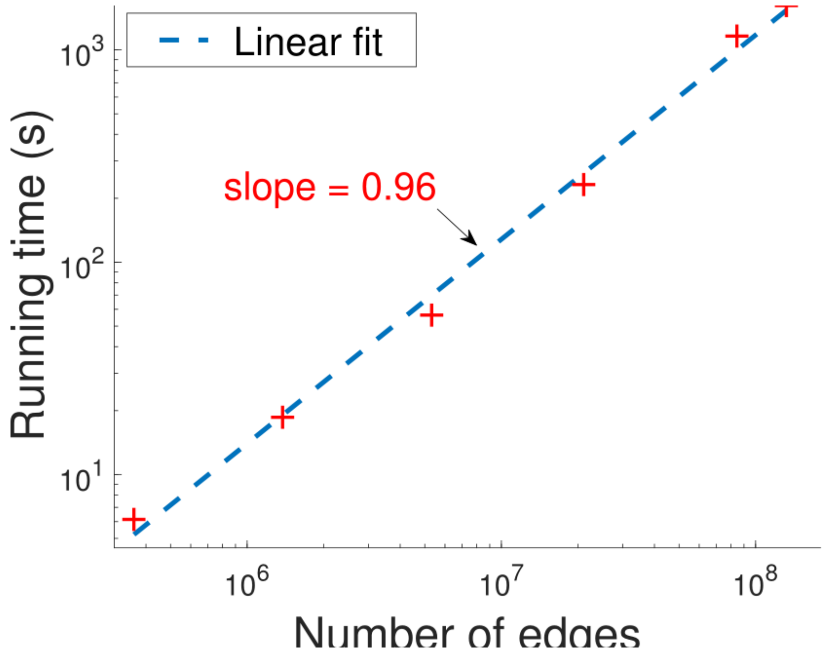
Save time and memory for GNN

- Save both time and memory
- Comparable performance



Fast and scalable

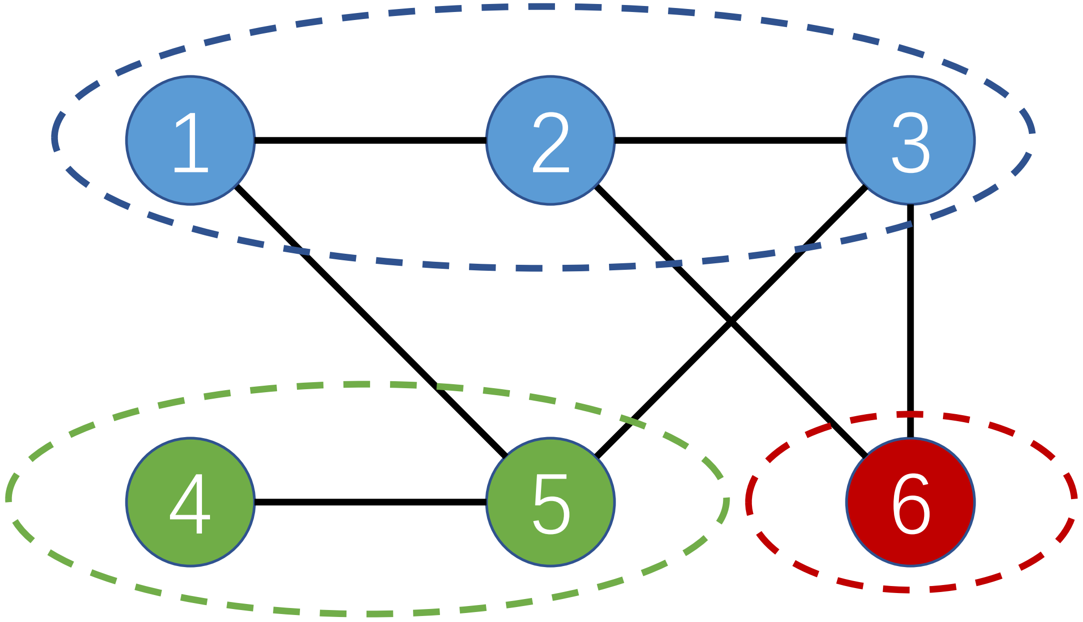
Scales linearly to number of edges (|E|).



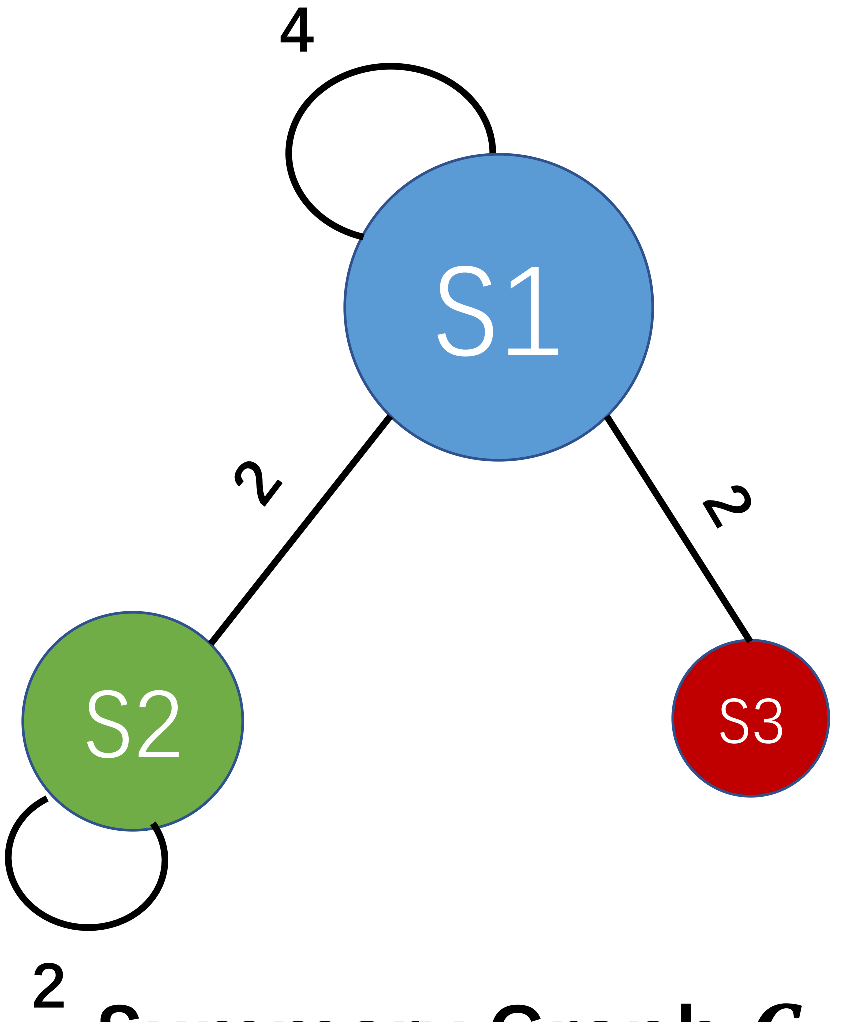
Outline

- Introduction
- **Our model**
- Our algorithm: DPGS
- Experiments
- Conclusion

Graph Summarization

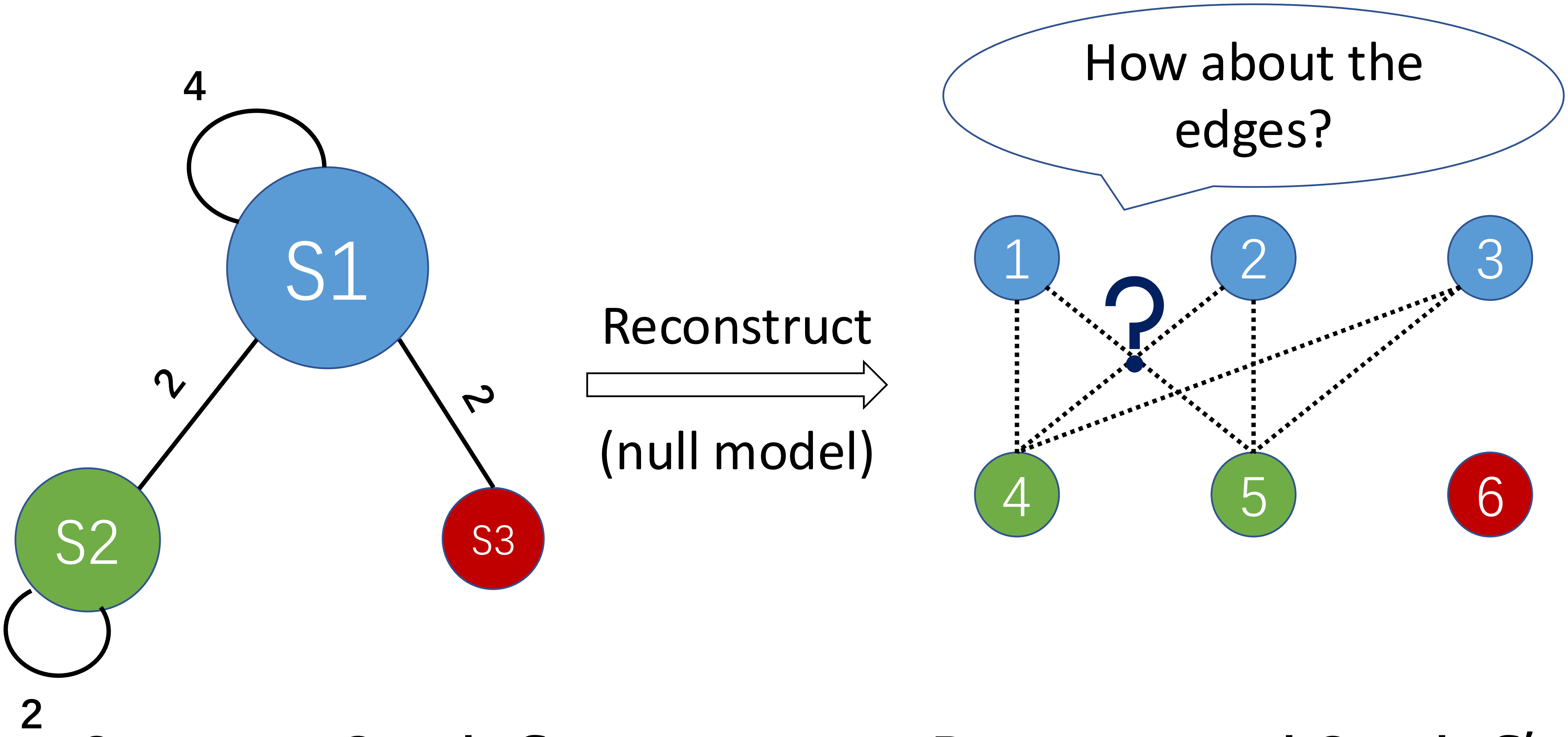


Original Graph G



Summary Graph G_s

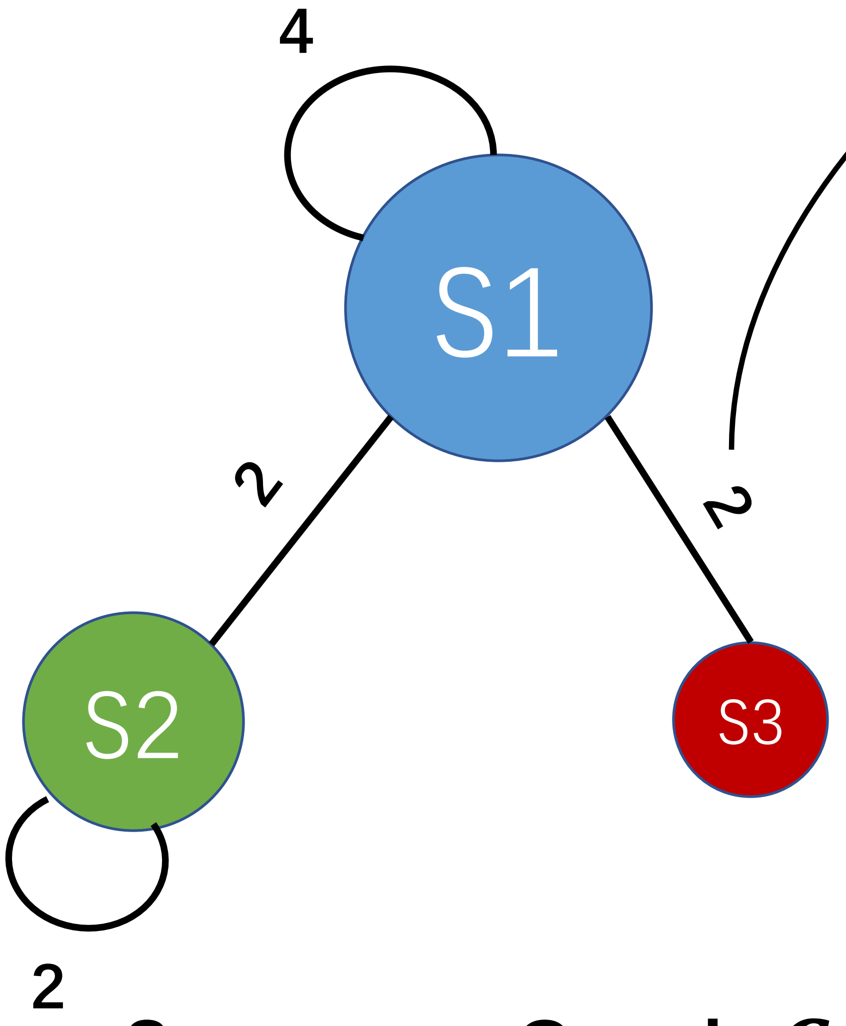
Graph Summarization



Summary Graph G_s

Reconstructed Graph G'

Graph Summarization



Summary Graph G_s

$$1 \times 3 = \boxed{\frac{2}{3}}$$

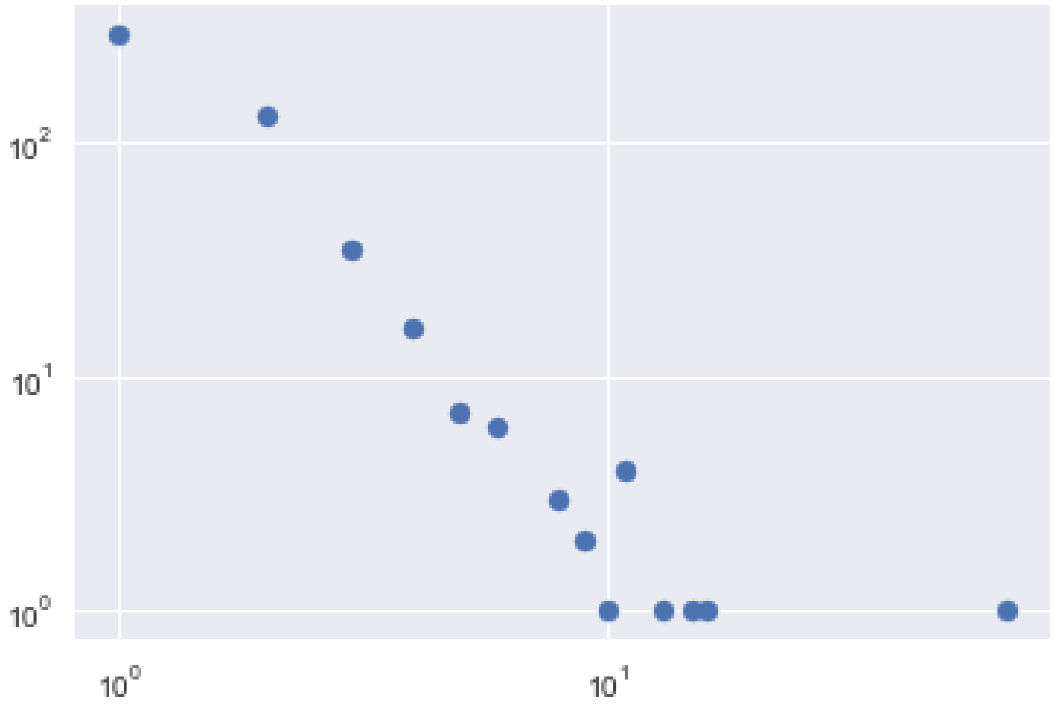
$$
\begin{array}{c|c}
 & \begin{array}{cccccc} & \text{S1} & & \text{S2} & & \text{S3} \end{array} \\
\begin{array}{c} \\ \text{S1} \\ \\ \\ \text{S2} \\ \\ \text{S3} \end{array} &
\left[\begin{array}{ccc|cc|c}
0 & 2/3 & 2/3 & 1/3 & 1/3 & 2/3 \\
2/3 & 0 & 2/3 & 1/3 & 1/3 & 2/3 \\
2/3 & 2/3 & 0 & 1/3 & 1/3 & 2/3 \\
\hline
1/3 & 1/3 & 1/3 & 0 & 1 & 0 \\
1/3 & 1/3 & 1/3 & 1 & 0 & 0 \\
\hline
2/3 & 2/3 & 2/3 & 0 & 0 & 0
\end{array}\right]
\end{array}
$$

Reconstructed Adjacency Matrix A'

Uniform Reconstruction Scheme

- Related works:
  - k-GS [LeFevre 2010]
  - SAA-Gs [Beg 2018]
  - SSumM [Lee 2020]
- Each node pair shares the same connect probability.
- Corresponding to **Erdos-Renyi** Random Graph Model.
- Is Erdos-Renyi Model a good null model?
  - Skewed-distributed
  - Power-law

Skewness of real-world graphs



Power-law

Erdos-Renyi

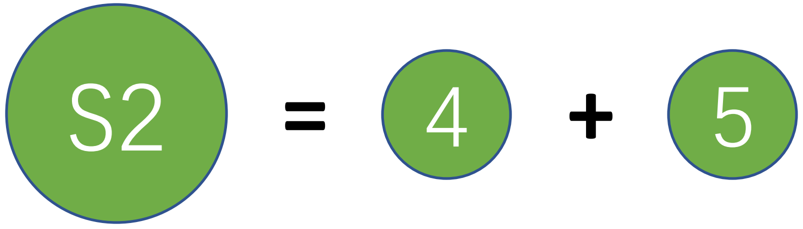
Configuration-based reconstruction

- Configuration model: $A'(i,j) \propto d_i d_j$.

- More specifically

$$\boxed{A'(i,j) = \frac{d_i}{D_p} A_s(p,q) \frac{d_j}{D_q}}$$

- d_i : degree of node i.
- $D_p = \sum_{i \in S_p} d_i$.
- $A_s(p,q)$: Weight of superedge (S_p, S_q).



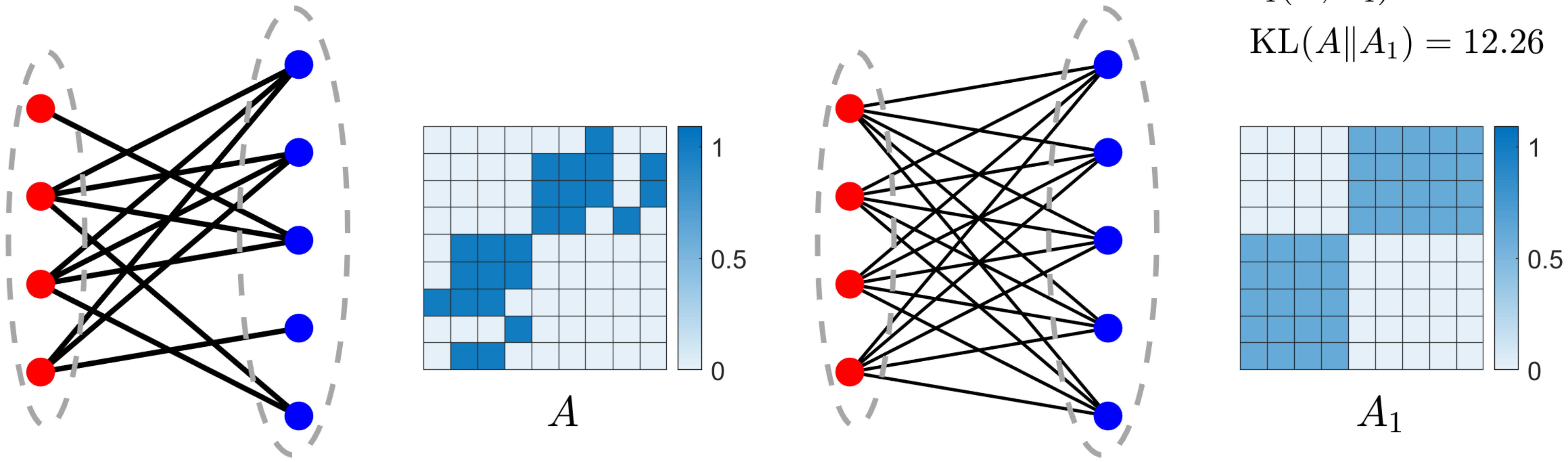
$$D_2 = d_4 + d_5$$

Uniform Scheme

$$A'(i,j) = \frac{1}{|S_p|} A_s(p,q) \frac{1}{|S_q|}$$

$$\ell_1(A, A_1) = 19.2$$

$$\mathrm{KL}(A \| A_1) = 12.26$$

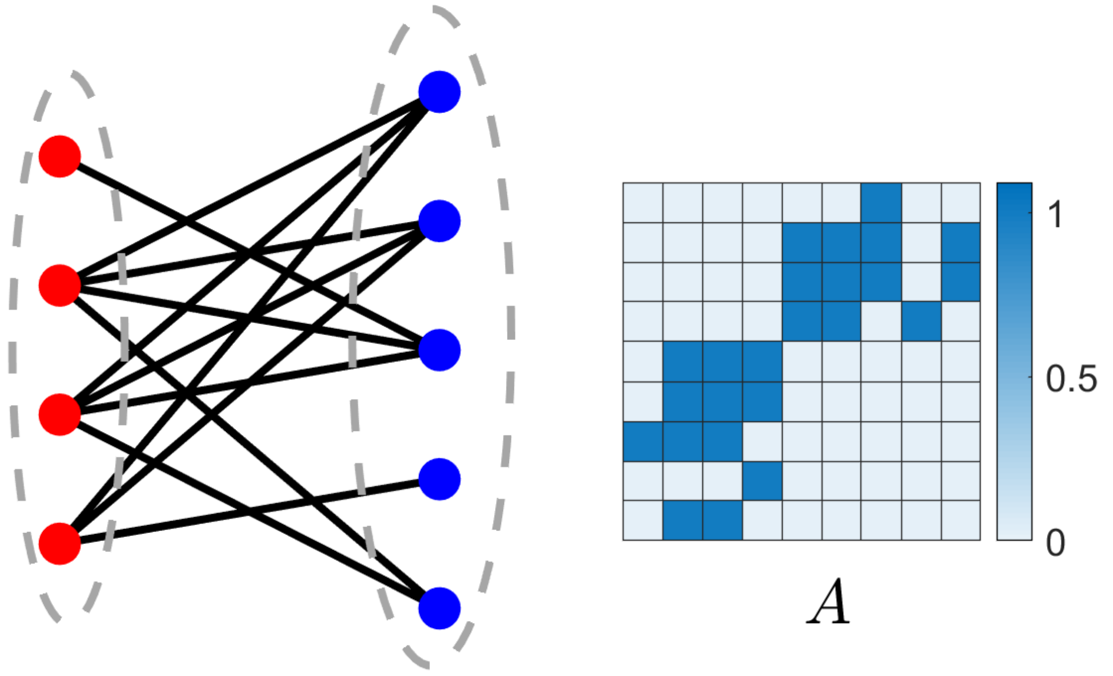


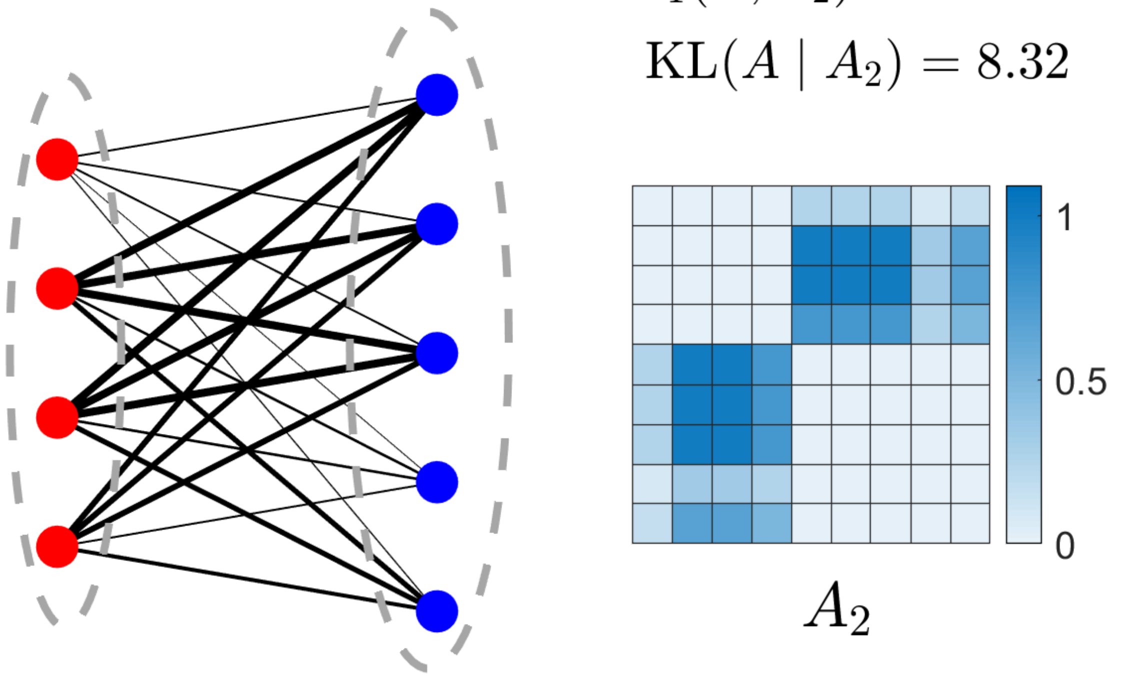
Configuration-based scheme

$$A'(i,j) = \frac{d_i}{D_p}\, A_s(p,q)\, \frac{d_j}{D_q}$$

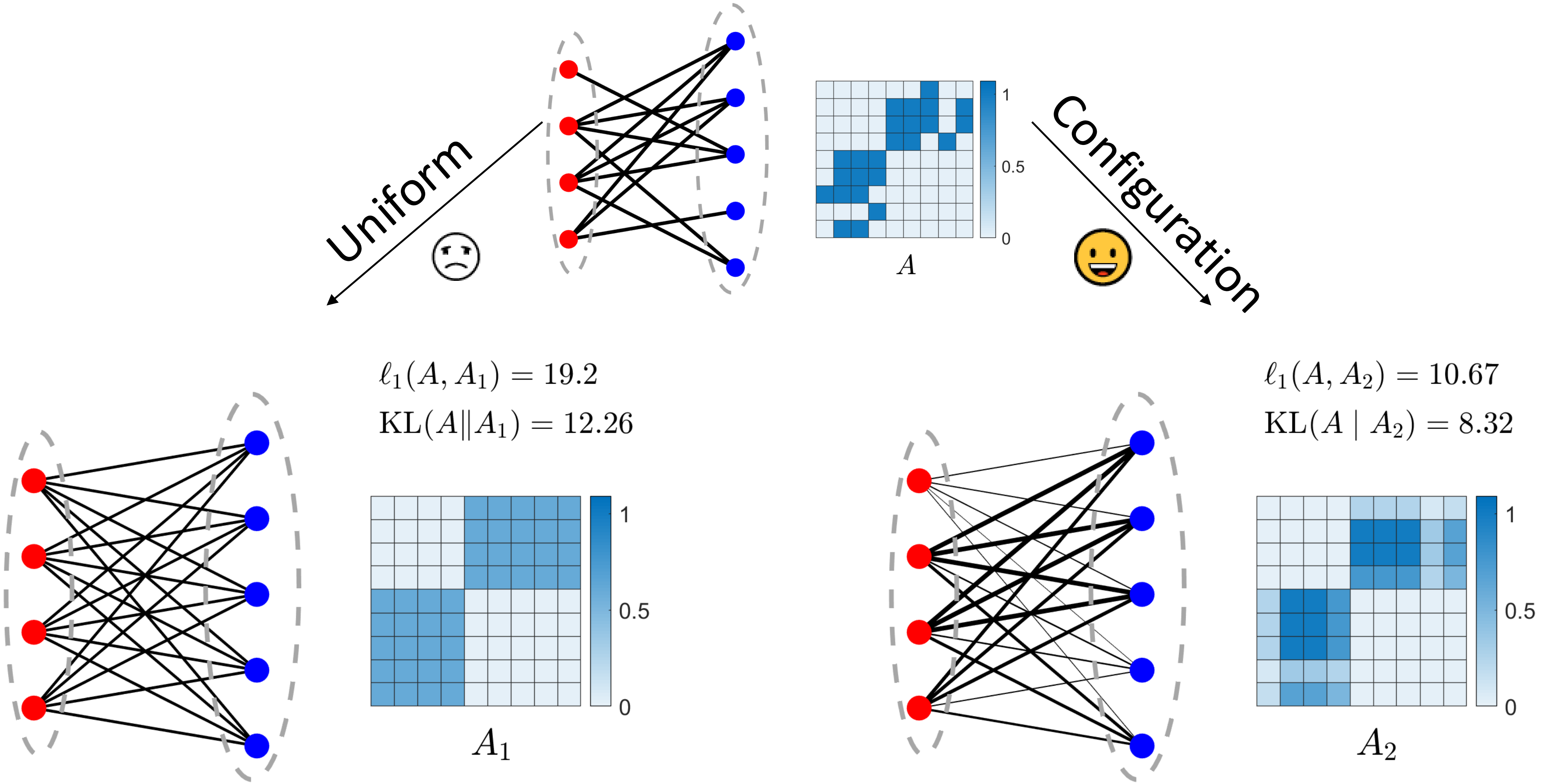
$$\ell_1(A, A_2) = 10.67$$

$$\mathrm{KL}(A \mid A_2) = 8.32$$

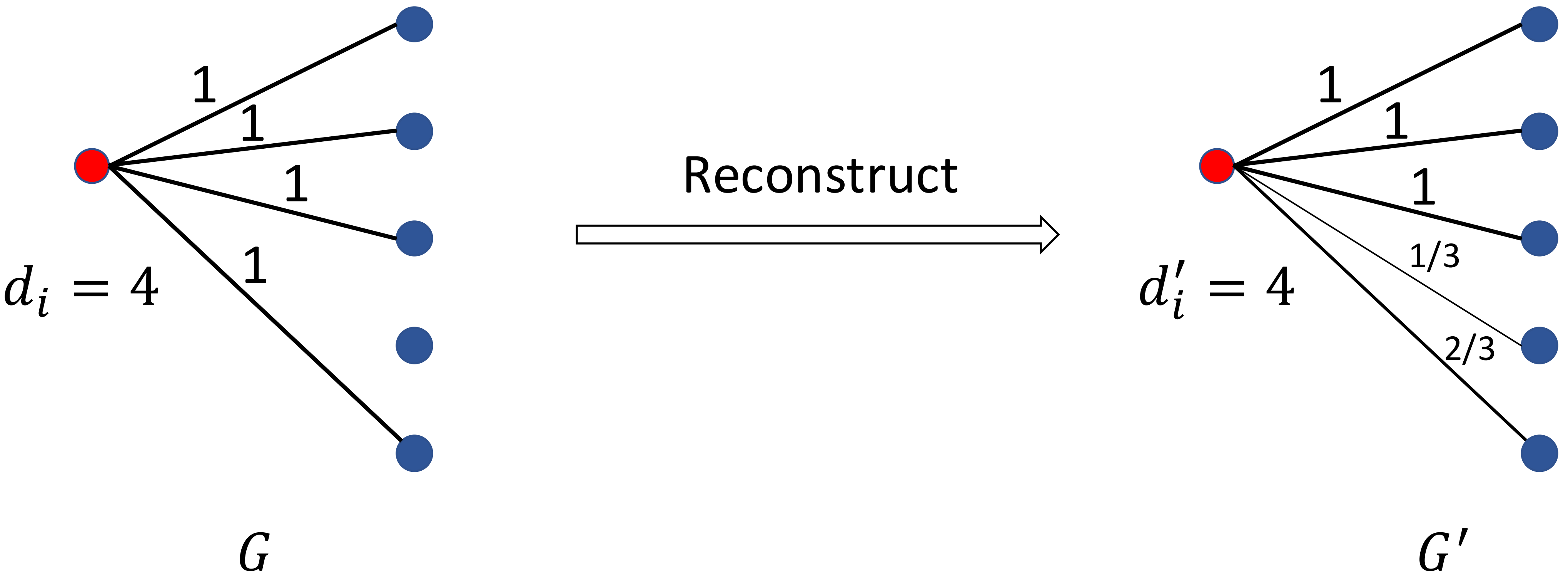




Our scheme is better



Degree-Preserving



$$d_i' = \sum_j A'(i,j) = \sum_j A(i,j) = d_i$$

Outline

- Introduction
- Our model
- **Our algorithm: DPGS**
- Experiments
- Conclusion

Main idea

- Basic operation: merging node together.

Model Part Error Part

- Criterion: Size of summary graph is small, Reconstruction error is small.
- MDL (Minimum Description Length) principle.
- MDL finds a model minimizing the total description length:

$$L(M, D) = \boxed{L(M)} + \boxed{L(D \mid M)}$$

- Model M: Summary graph; Data D: Original Graph.

MDL encoding

- Error part: (generalized) KL-divergence:

$$L(D \mid M) = \mathrm{KL}(A||A') = \sum_{ij} A(i,j) \ln \frac{A(i,j)}{A'(i,j)} - A(i,j) + A'(i,j)$$

- Extra bits to encode A given A'.

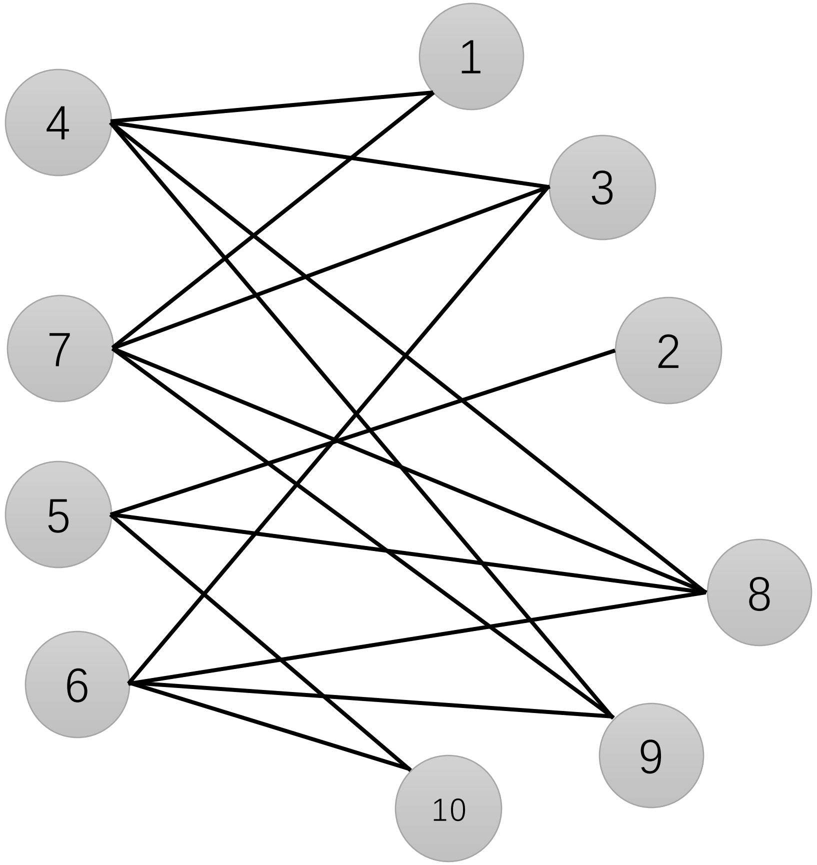
Algorithm Procedure

Main Procedure:

- Initialize each node as a supernode.
- Iteration (T turns):
  - Group supernodes using LSH
  - For each group:
    - Sample supernode pairs and merge supernodes in each group
- Return summary graph

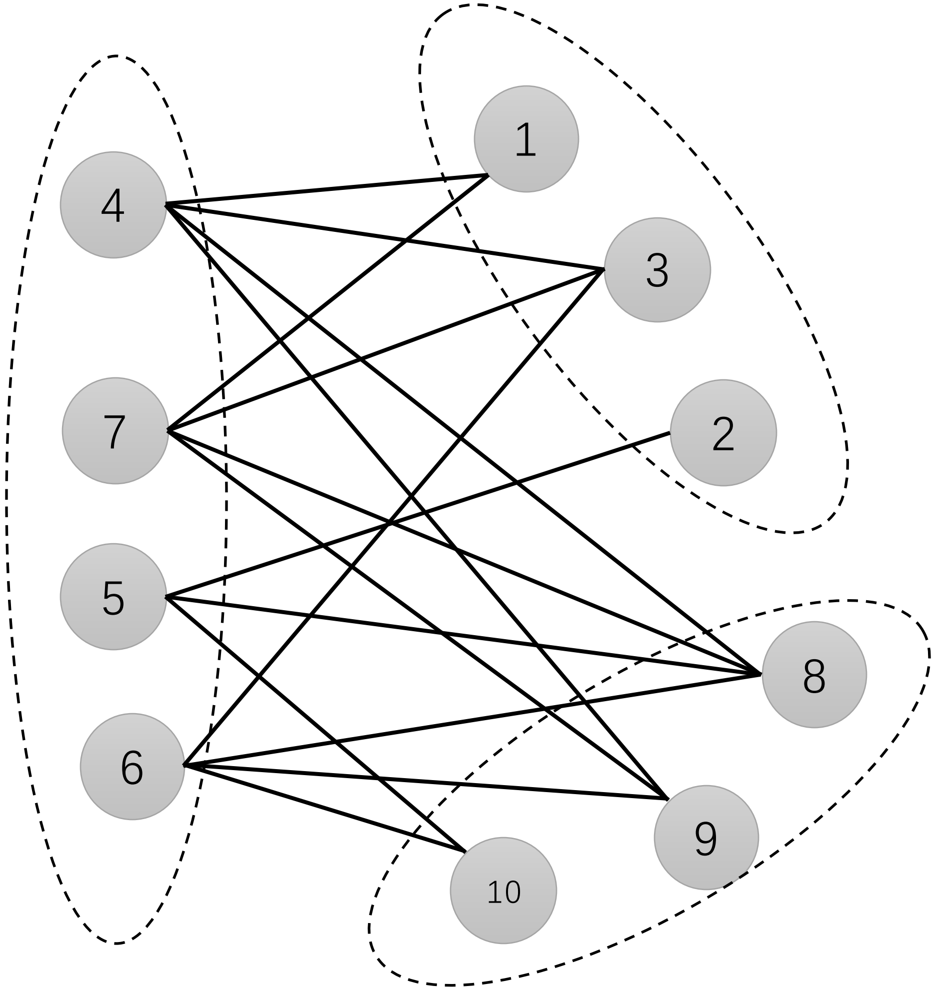
- Tips:
  - Merge nodes with similar neighborhood yield greater decrease to total description length.
  - Use LSH (Locality Sensitive Hashing) to group nodes.

Initialization



$$\begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 10 & 0 & 0 & 0 & 0
\end{bmatrix}$$

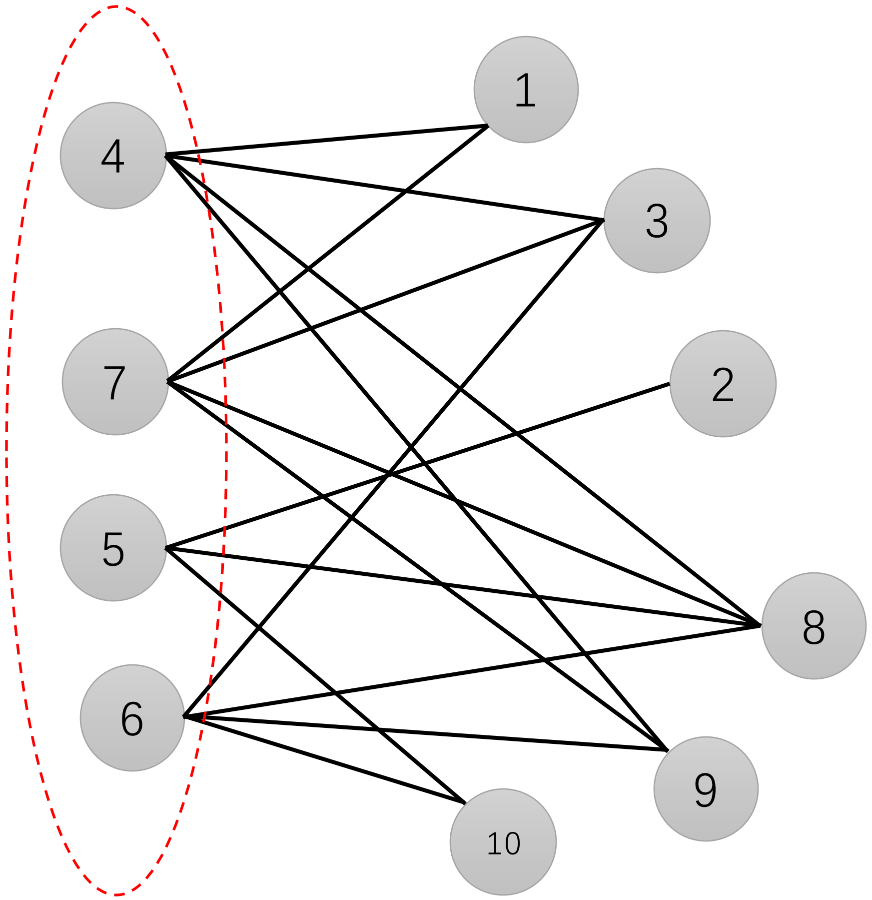
LSH Grouping



$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 10 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Sample and Merge

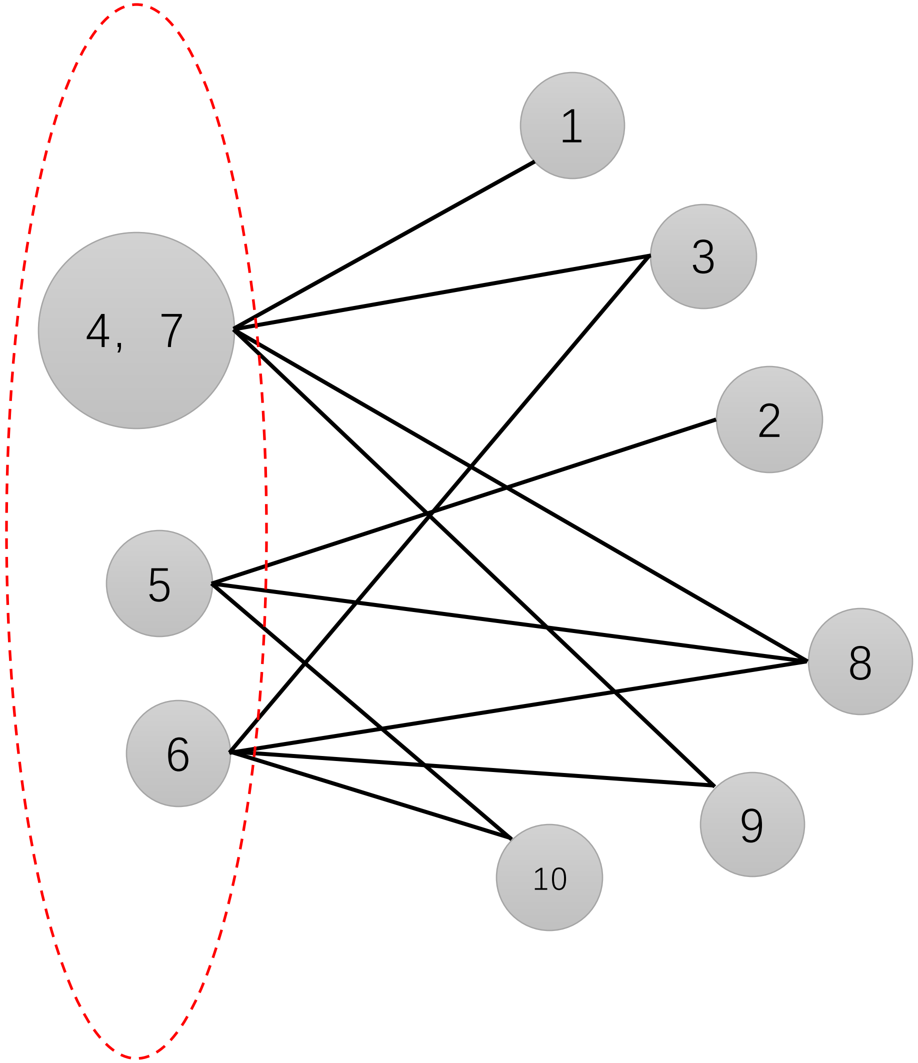
Sample pairs: (4, 7), (5, 6) $\quad \arg\max \Delta L(M, D)$



$$
\begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 10 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Sample and Merge

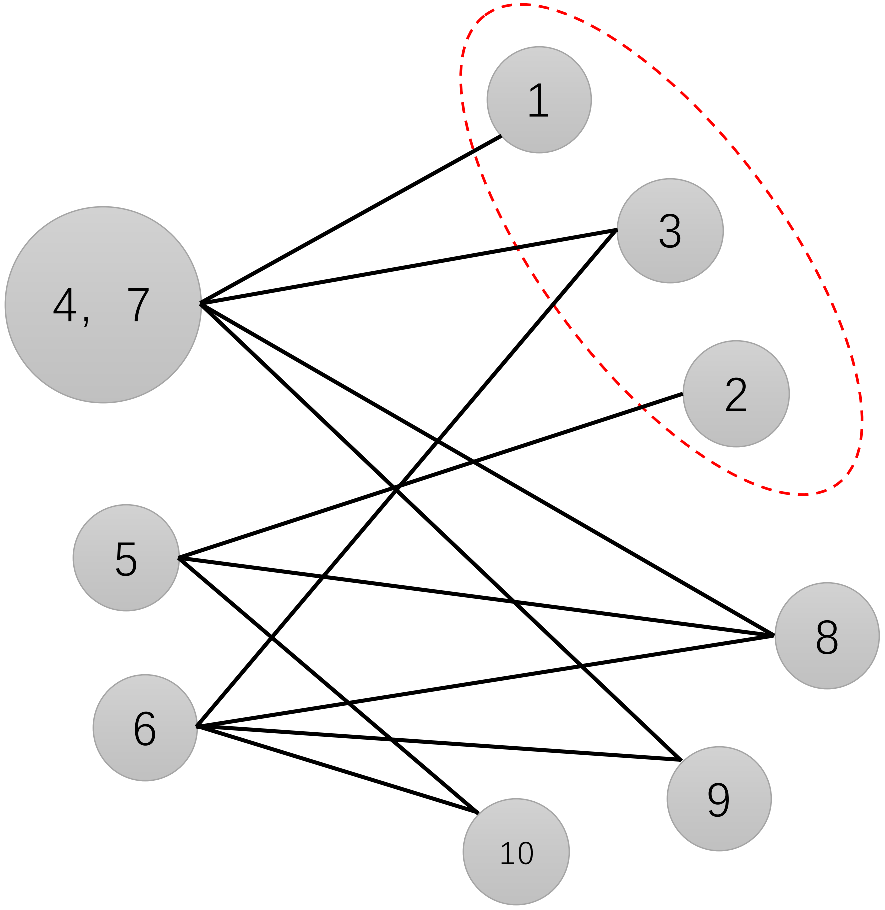
Merge (4, 7)



$$\begin{bmatrix} 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

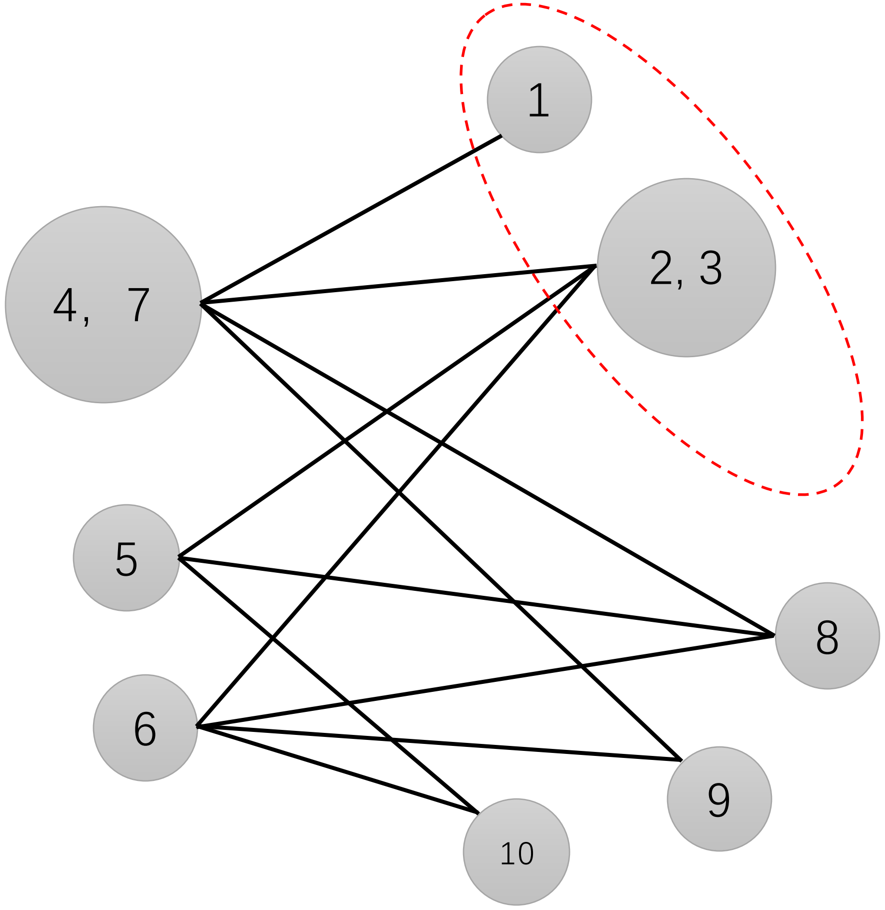
Sample and Merge

Merge (2, 3)



$$\begin{bmatrix}
0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\
2 & 0 & 2 & 0 & 0 & 0 & 2 & 2 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0
\end{bmatrix}$$

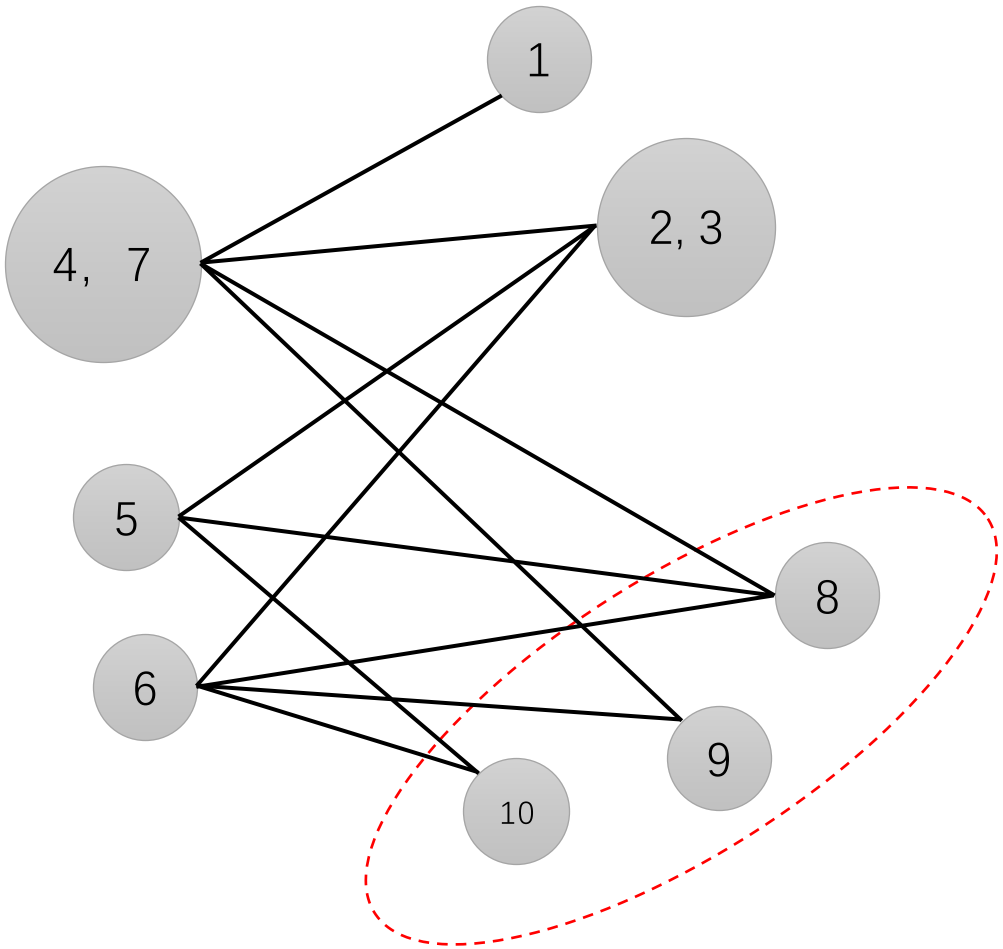
Sample and Merge



$$\begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Sample and Merge

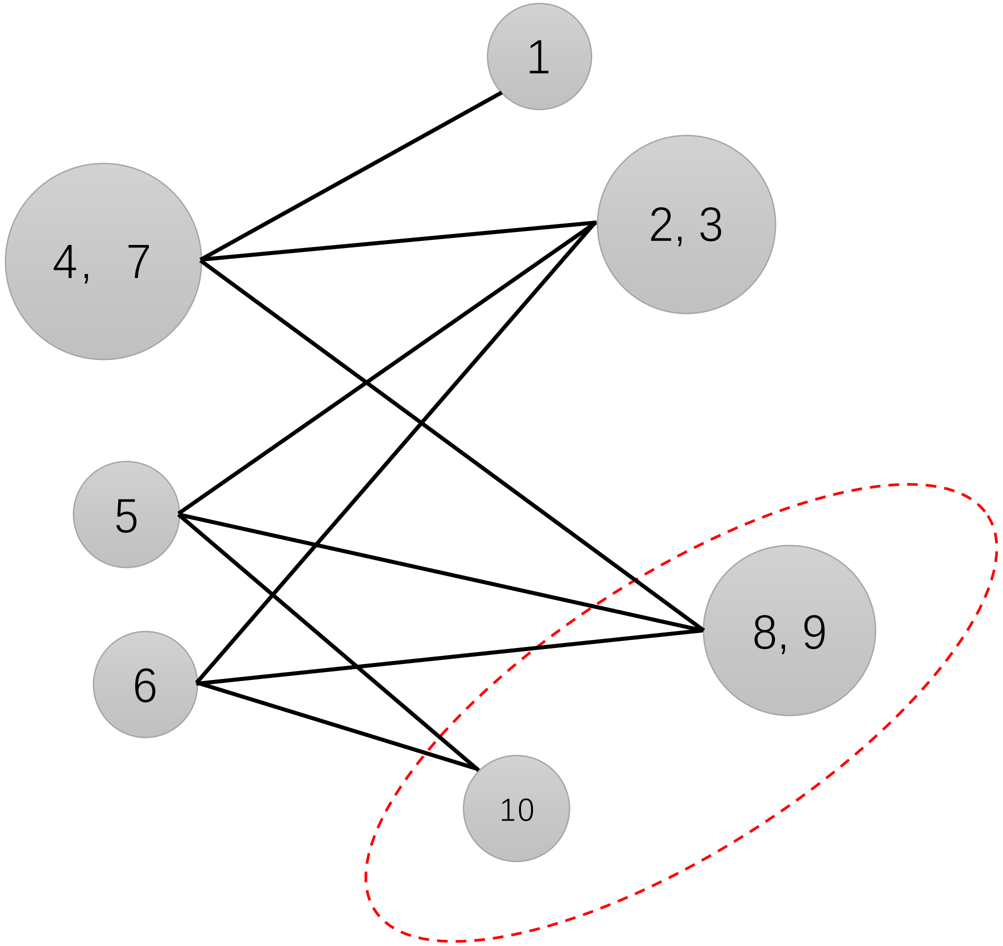
Merge (8, 9)



$$\begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 2 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Sample and Merge

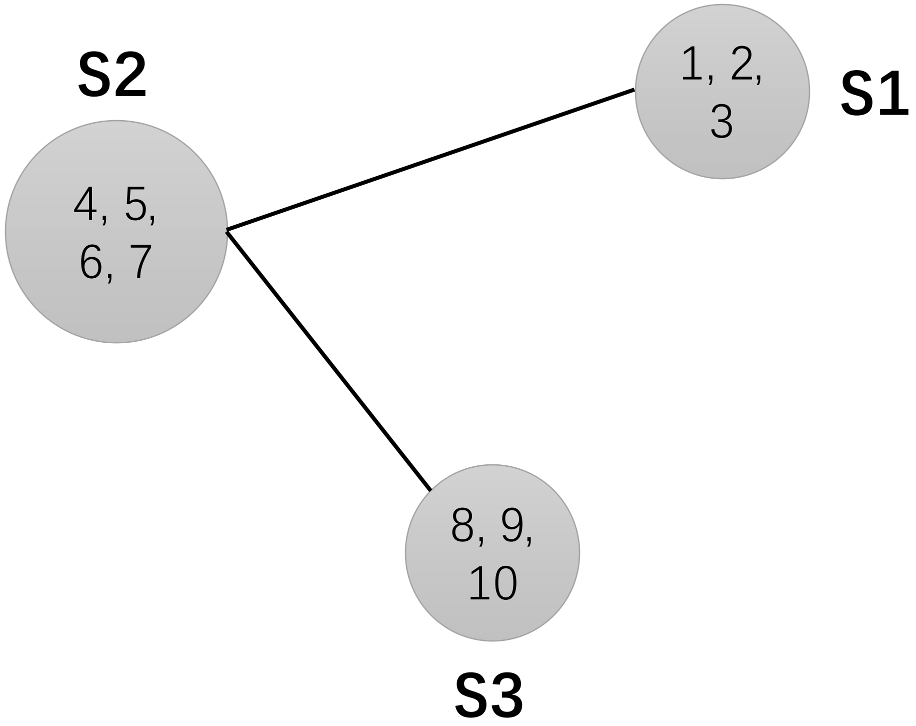
Merge (8, 9)



$$\begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 2 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 2 & 1 \\ 0 & 1 & 4 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Return summary graph

After T iterations



$$\begin{array}{c c} & \begin{array}{ccc} S1 & S2 & S3 \end{array} \\ \begin{array}{c} S1 \\ S2 \\ S3 \end{array} & \begin{bmatrix} 0 & 6 & 0 \\ 6 & 0 & 9 \\ 0 & 9 & 0 \end{bmatrix} \end{array}$$

Spectral Preservation

- **Theorem (Eigenvalue Perturbation)**

$$\sum_i (\lambda_i - \lambda_i')^2 \le 2 \cdot L(\, D \mid M \,)$$

Eigenvalue of G
(normalized Laplacian)

Eigenvalue of G'
(normalized Laplacian)

Outline

- Introduction
- Our model
- Our algorithm: DPGS
- **Experiments**
- Conclusion

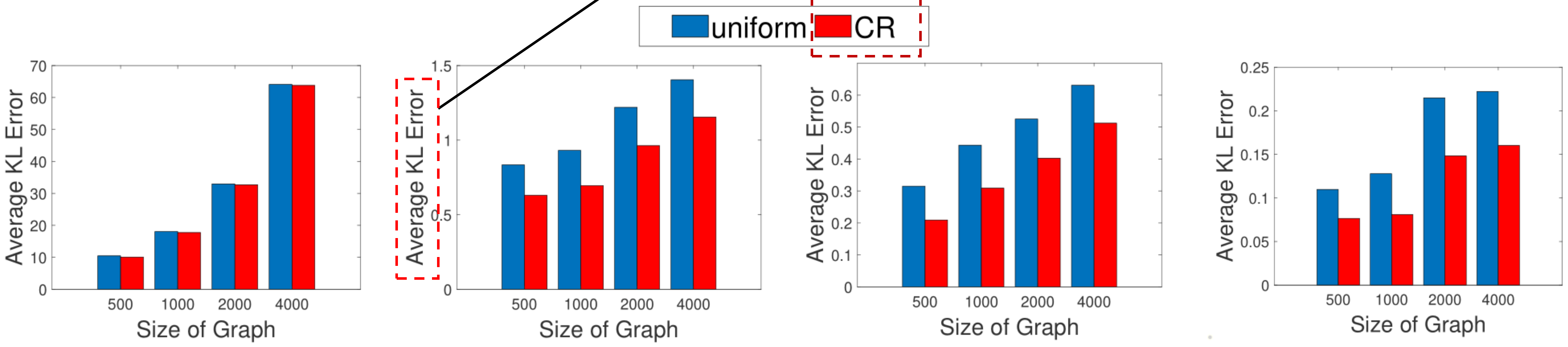
Dataset

- Synthetic graphs using different random graph.
- 8 real-world networks (up to 100M edges).
  - Protein network.
  - Social Network.
  - Co-purchase network.



Our scheme is better than uniform scheme

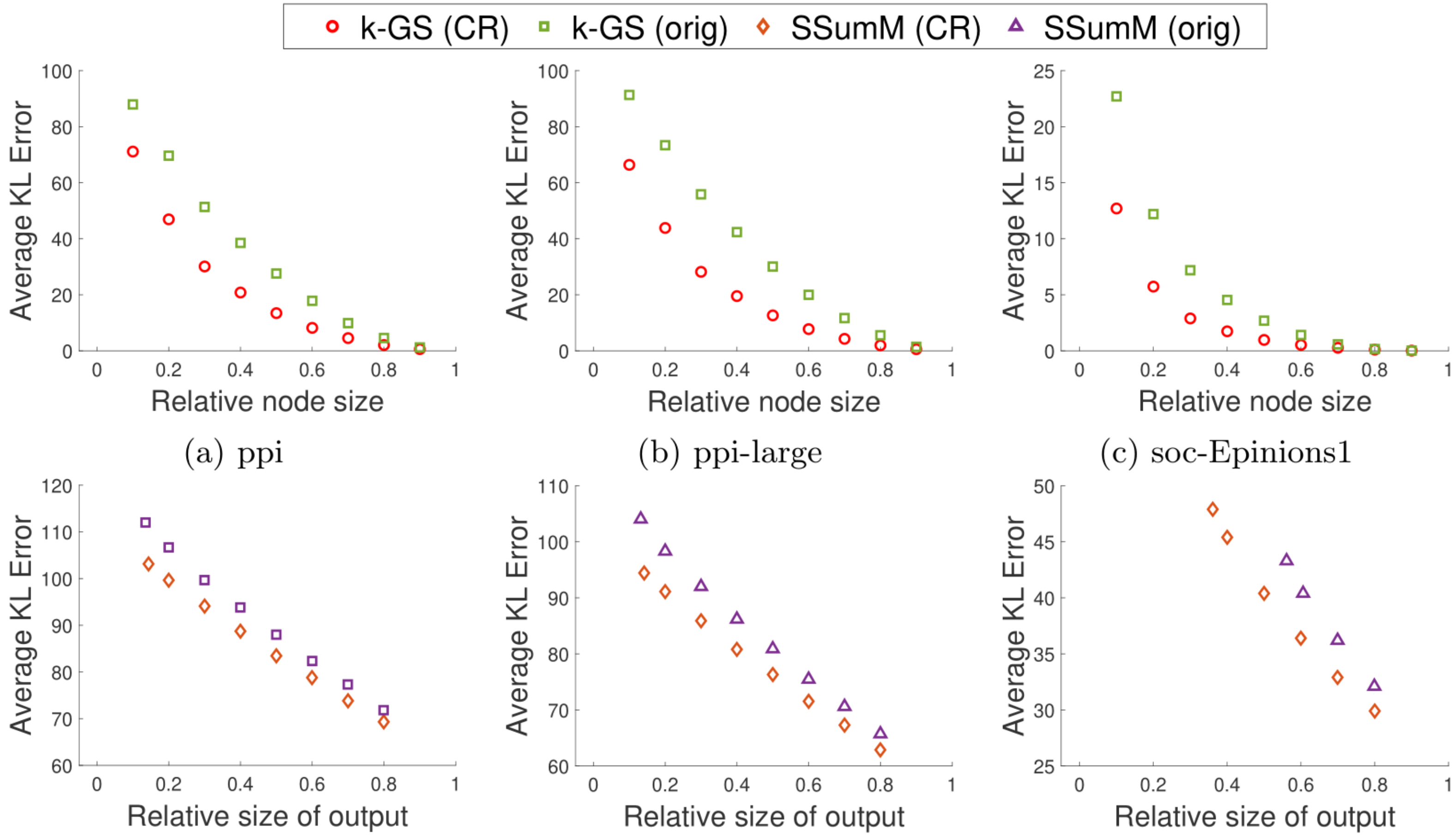
- Two synthetic data: E-R model and power-law model.
- Compare encoding error L(D|M).



(a) Erdős-Rényi ($p = 0.02$) (b) Power-law ($\alpha = 3.0$) (c) Power-law ($\alpha = 3.5$) (d) Power-law ($\alpha = 4.0$)

Uniform (blue) v.s. Configuration (red)

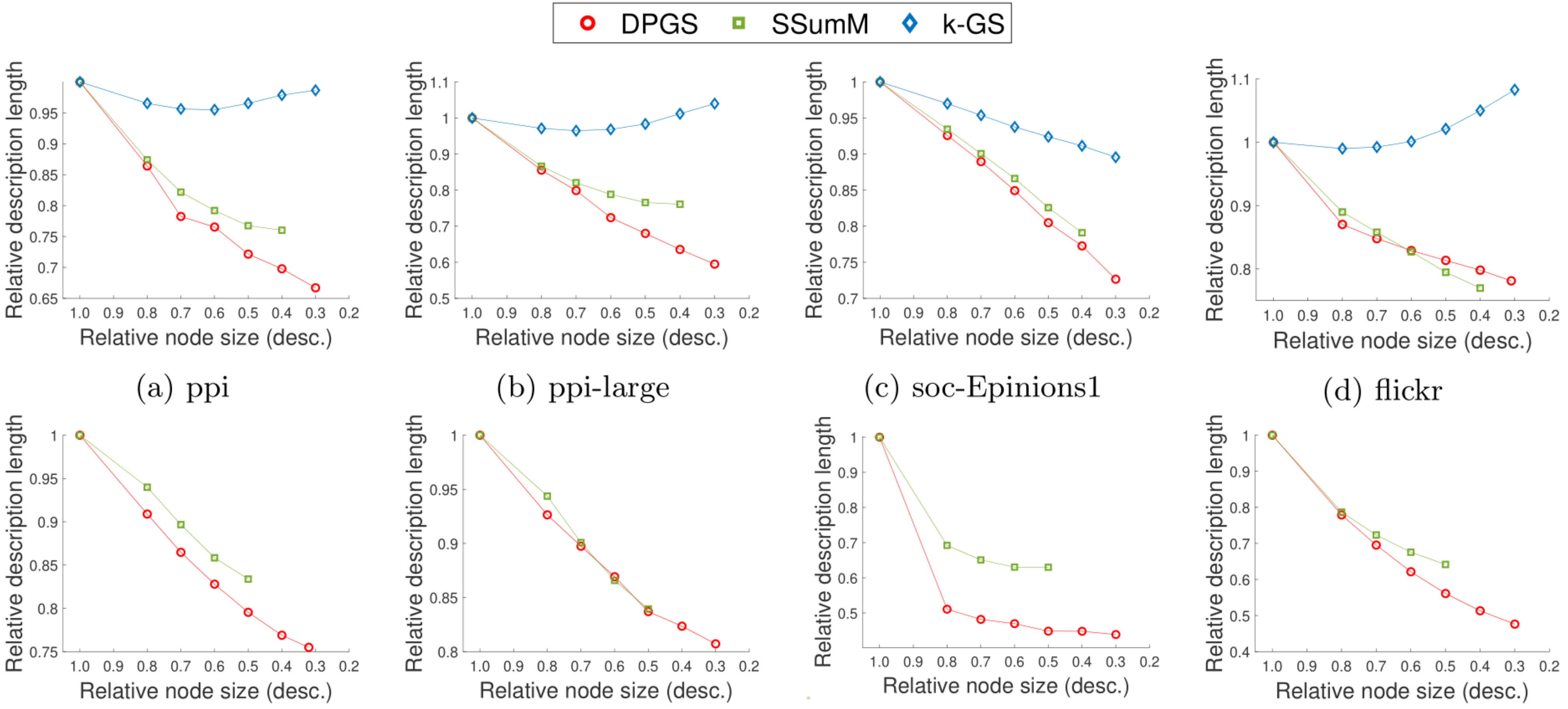
Our scheme can improve existing methods.



(a) ppi (b) ppi-large (c) soc-Epinions1

(d) ppi (e) ppi-large (f) soc-Epinions1

DPGS yields the most compact summary graphs



(a) ppi (b) ppi-large (c) soc-Epinions1 (d) flickr

(e) reddit (f) yelp (g) amazon (h) amazon2M

Lower encoding length ✅

Save time and memory for GNN

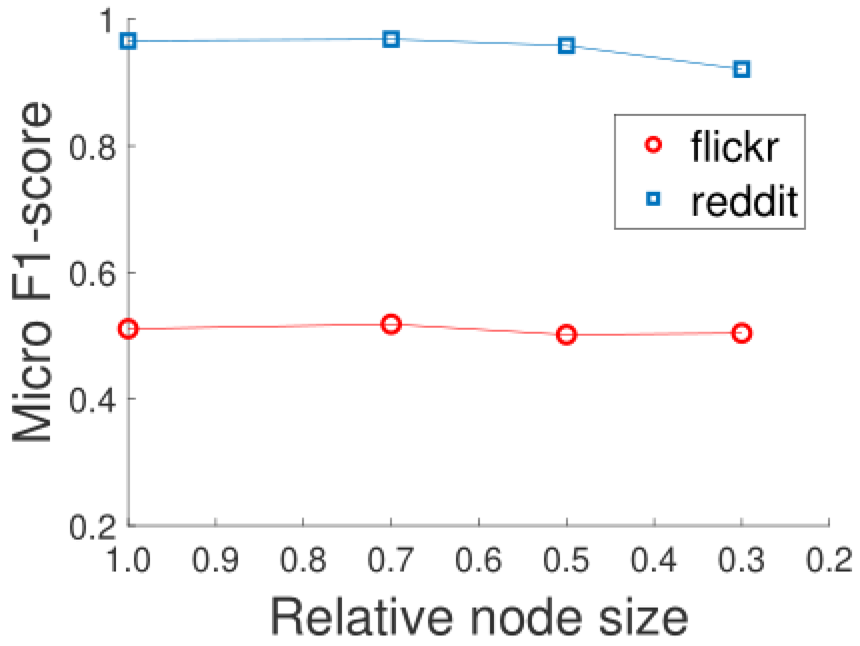
Amazon2M (2.4 M nodes, 61 M edges)

Original graph (❌)

Summary graph (✅)

F1 score: 0.890[1] (orig) vs 0.870 (summ)

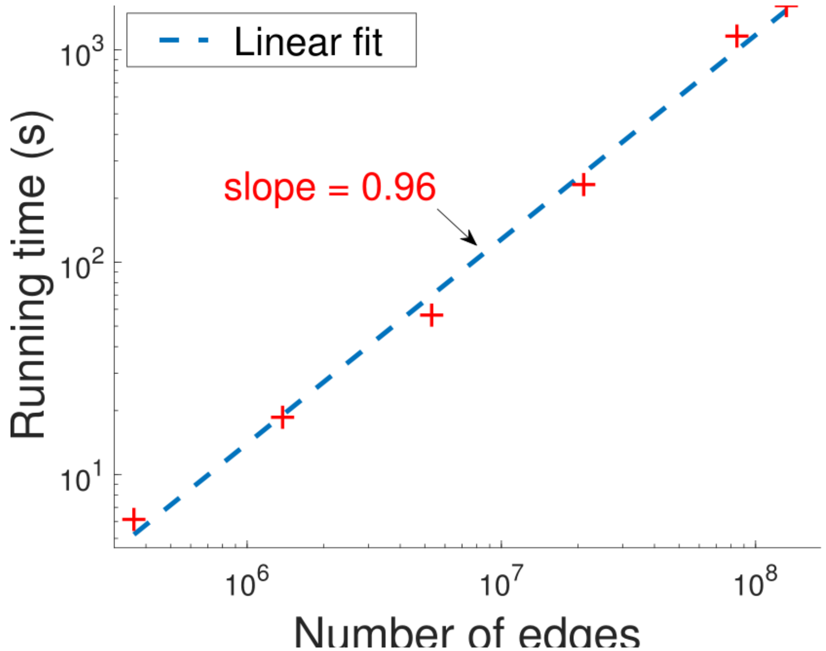
- Save both time and memory
- Comparable performance



[1]Cluster-GCN (2 layer)

Fast and scalable

Scales linearly to number of edges (|E|).



Outline

- Introduction
- Our model
- Our algorithm: DPGS
- Experiments
- **Conclusion**

Conclusion

- Introduce the configuration-based reconstruction scheme.
- Propose a novel degree-preserving graph summarization algorithm.
- Our algorithm yields more compact summary graphs.
- Our algorithms runs fast, scales linearly, and helps to train large GNN model.


Houquan Zhou


Shenghua Liu


Kyuhan Lee


Kijung Shin


Huawei Shen


Xueqi Cheng

Thank you!

Contact us

zhouhouquan18@mails.ucas.edu.cn

liushenghua@ict.ac.cn

kyuhan.lee@kaist.ac.kr

https://github.com/BGT-M/DPGS