MrMine: Multi-resolution Multi-network embedding

Presented by: Boxin Du

University of Illinois at Urbana-Champaign

Boxin Du

Hanghang Tong

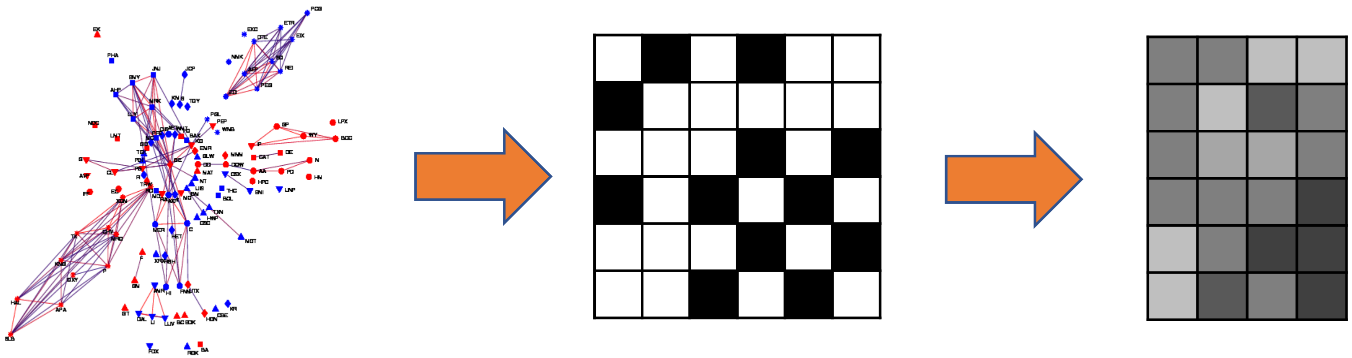
ILLINOIS

IDEA

Network Embedding: Why?

Goal: Map network object (node/edge/subgraph/network) into a low-dimensional space.

- Distributed representations for network objects
- Encode network characteristics into continuous vector space
- Automatic feature learning for downstream tasks on networks

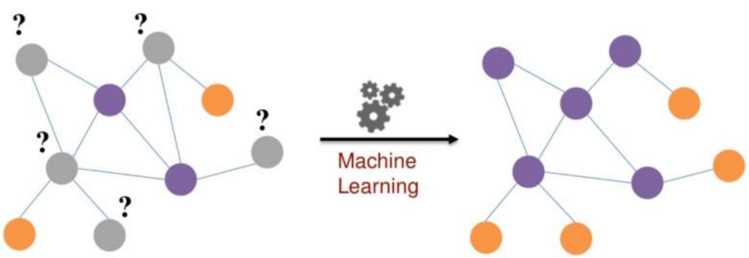


Input Network

Adjacency matrix

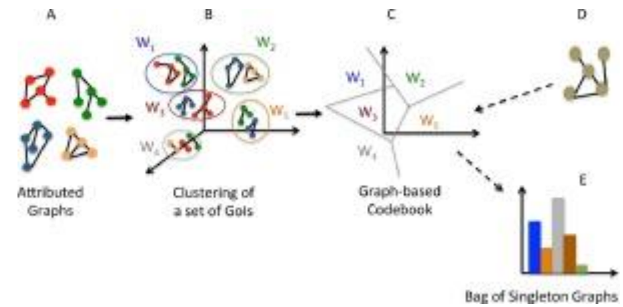
Low-dimensional representation

Network Embedding: Applications



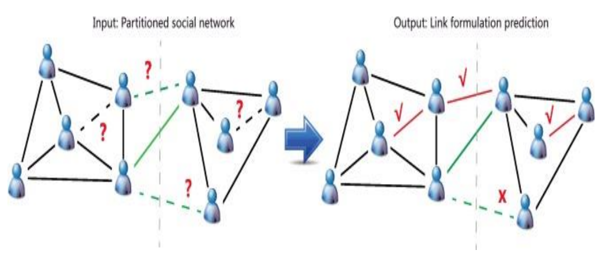
Node classification
[J Leskovec '17]

... ...



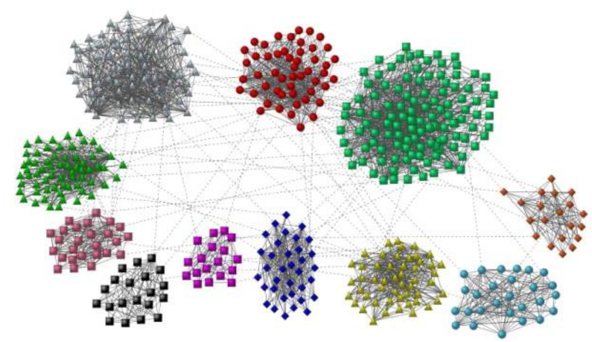
Network classification
[F Silva, et al '17]

Network Embedding



Link prediction
[J Wu, et al '17]

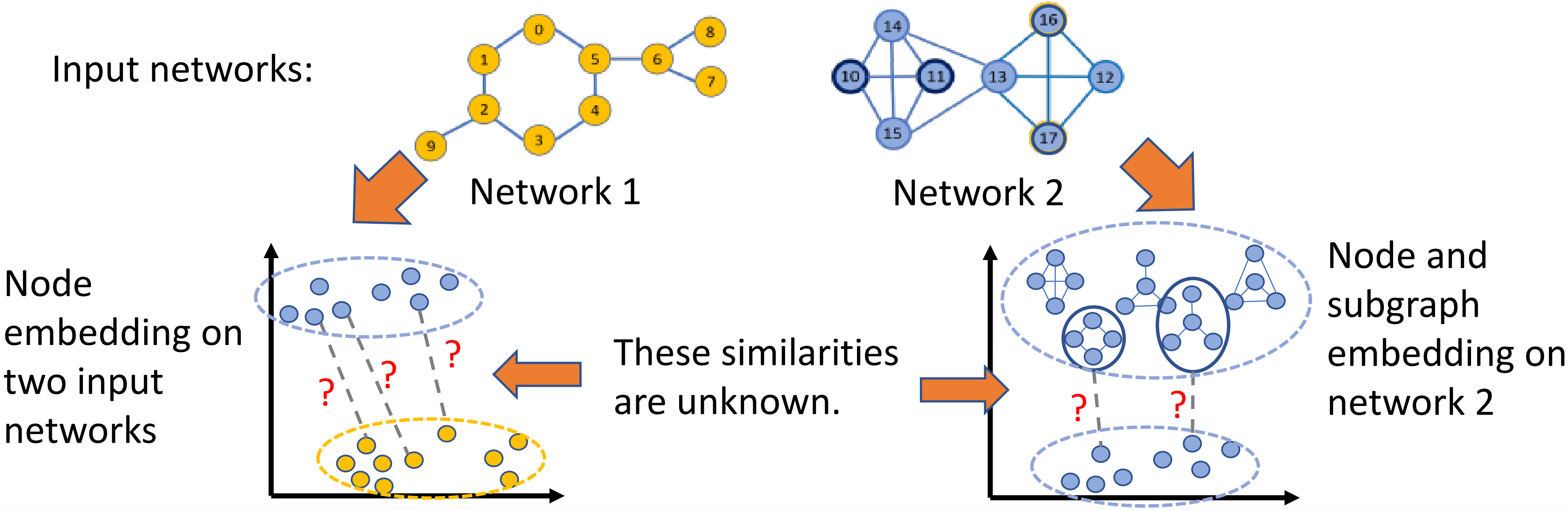
... ...



Community detection
[S Fortunato '09]

Limitations of Existing Methods

- Single resolution:
  - Embedding on nodes/subgraphs/networks
  - Problem: Embeddings across resolutions are separate
- Single network:
  - Embedding on one single network
  - Problem: Embeddings on multiple networks are separate



Limitations of Existing Methods (cont'd)

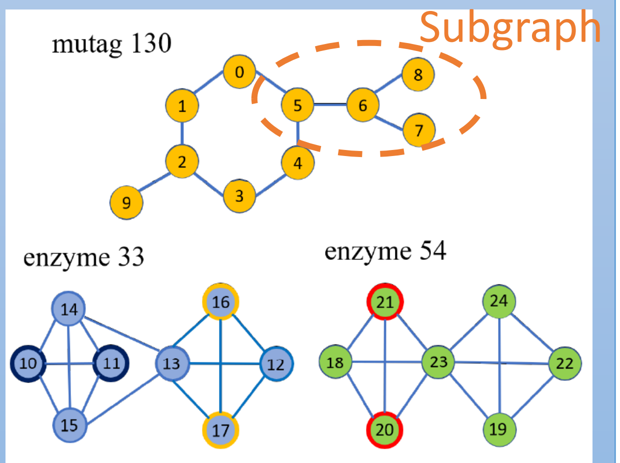
Name	Reference	Multi network	Multi resolution
deepwalk	B Perozzi, SIGKDD '14	✗	✗
LINE	J Tang, WWW '15	✗	✗
Node2vec	A Grover, SIGKDD '16	✗	✗
subgraph2vec	A Narayanan '16	✗	✓
Deep graph kernel	P Yanardag SIGKDD '15	✓	✗
struc2vec	LFR Ribeiro SIGKDD '17	✓	✗
…	…	…	…

Name	Reference	Multi network	Multi resolution
MrMine	This paper	✓	✓
MrMine+	This paper	✓	✓

Our methods

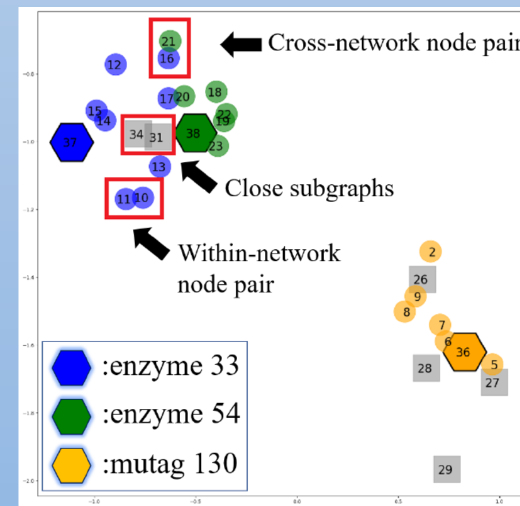
Multi-resolution Multi-network embedding

Three input networks



MrMine
Learning model

Learned embeddings of nodes, subgraphs, and networks



- Goals:
  - Unsupervised embedding learning.
  - Embeddings from different networks are comparable.
  - Embeddings from different resolutions are comparable.
  - Objects with close structural characteristics are close in embedding space.

Nodes:
Subgraphs:
Networks:

Roadmap

- Motivation
- Problem Definition
- Proposed Solution: ‘MrMine’
- Experiments
- Conclusions

Problem Definition

Given:

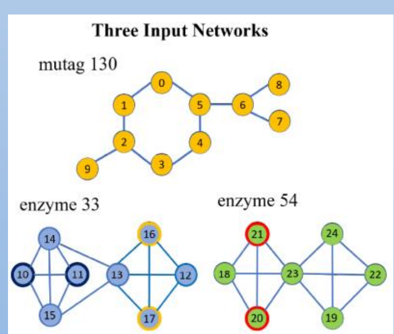
- The inputs for context building:
  - A set of networks **G**;
  - The dimension of embedding vectors **p**;
  - Subgraph constraints (e.g. the maximum height of WL subtrees H);
- Parameter set for language model for embedding learning.

We specify one subgraph type in proposed method.

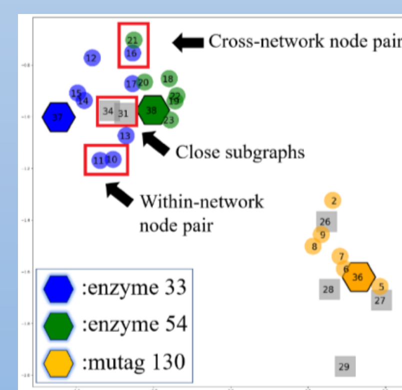
Find: Embedding matrices $\mathbf{Fg}$, $\mathbf{Fs}$, and $\mathbf{Fn}$ for:

- All input networks in **G**;
- All extracted subgraphs in **S**;
- All nodes in **G.**

With all embeddings in the same space.



MrMine
Learning model

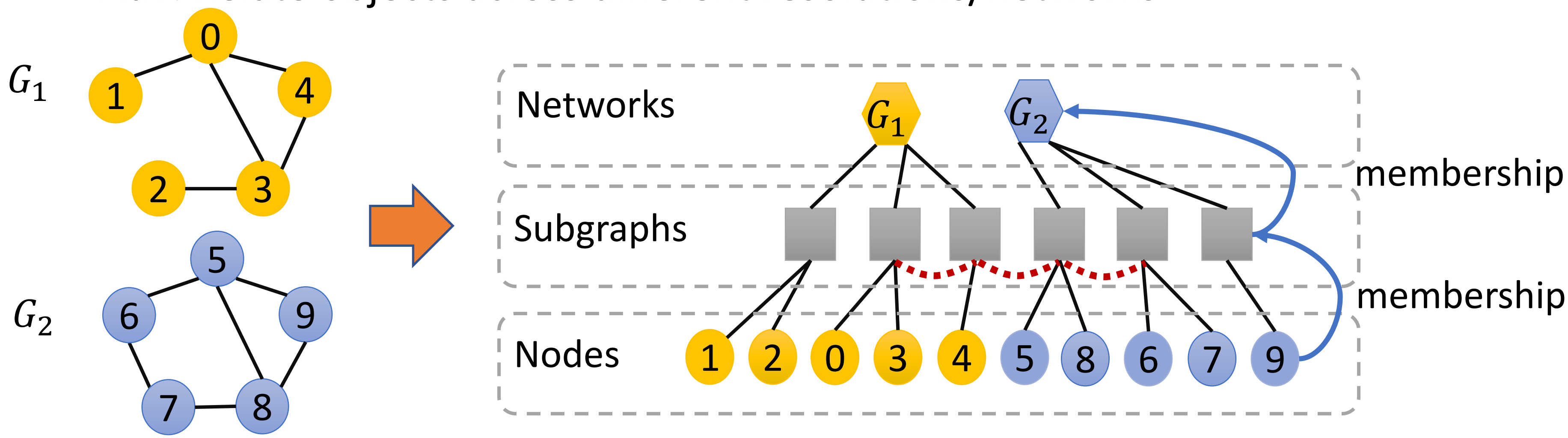


Roadmap

- Motivation
- Problem Definition
- Proposed Solution: ‘MrMine’
- Experiments
- Conclusions

MrMine: Challenges and Key Ideas

- C1: How to build context for cross-resolution cross-network objects?
  - Idea: Cross-Resolution Cross-Network (CRCN) relation network
  - Adv: Relate objects across different resolutions/networks



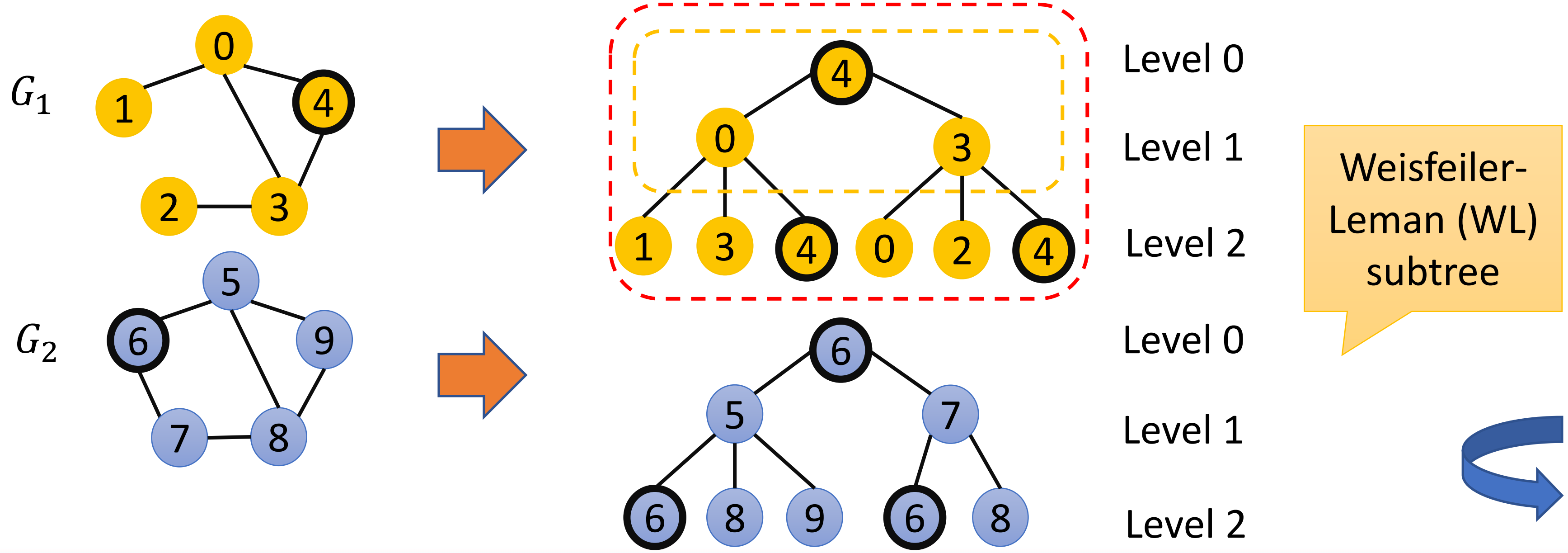
Input networks CRCN relation network

- Remarks:
  - Objects of multi-resolution and multi-network are vertices in the same network.
  - Links across different resolutions represent membership relation (e.g. blue arrow).
  - Links in the same layer on the subgraph resolution represents subgraph similarity.

MrMine: Challenges and Key Ideas (cont'd)

C2: How to construct the links of CRCN relation network?

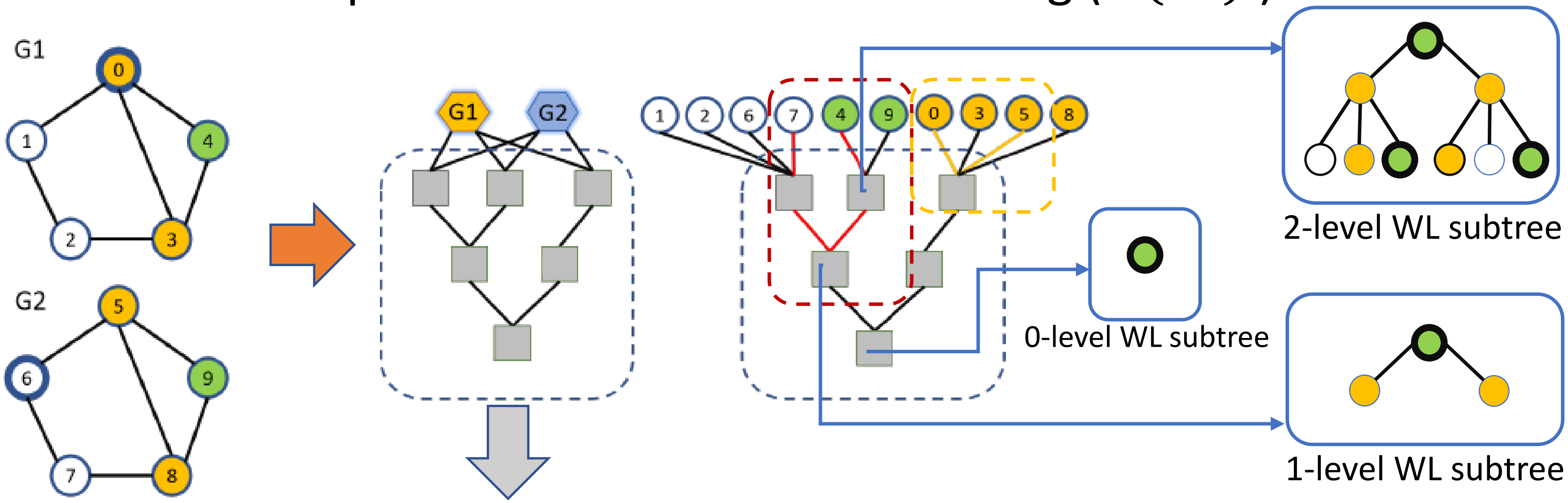
- Idea: WL subtree as the subgraph resolution
- Adv:
  - WL label transformation is efficient ($O(hm)$, h: constant; m: # of edges)
  - 'Borderless' across different networks (allow links across networks)
  - Bridge between node resolution and network resolution



[1] Shervashidze, Nino, et al. "Weisfeiler-lehman graph kernels." *Journal of Machine Learning Research* 12.Sep (2011): 2539-2561.

MrMine: Challenges and Key Ideas (cont'd)

- C3: How to reduce computation costs to build CRCN relation network?
  - Idea: Hierarchical structure of CRCN network (H-CRCN)
  - Adv:
    - Hierarchical structure of WL subtrees for finer resolutions
    - Avoid explicit cross-network link building ($O(n^2)$)

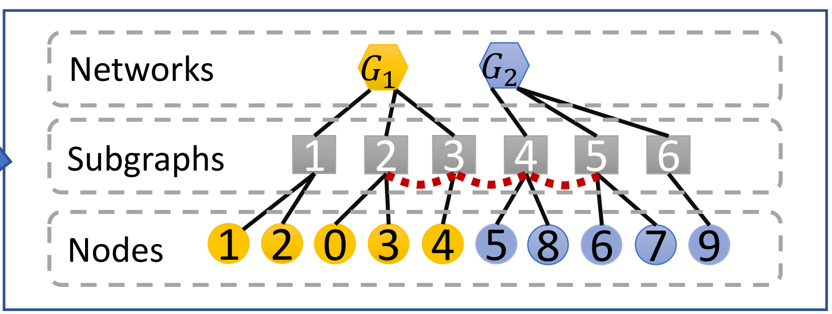


*Gray rectangle: hierarchical relation network of WL subtrees

MrMine: Method Details

- <u>Step 1</u>: H iterations of WL node label transformation:
  - Generate unique WL subtrees of height H as subgraphs
- <u>Step 2</u>: Function $f(S_i, S_j)$ to calculate the subgraph similarity:
  - Option 1: $f(S_i, S_j) = \sum_h DTW(Q^h_{S_i}, Q^h_{S_j})$
  - Option 2: $f(S_i, S_j) = \sum_h \sum_t |\tilde{Q}^h_{S_i}(t) - \tilde{Q}^h_{S_j}(t)|$
- <u>Step 3</u>: Construct CRCN relation network:
  - Add nodes, subgraphs, networks as **new vertices** in CRCN network
  - Add **cross resolution links** based on membership relation
  - Add **cross network links** based on $f(S_i, S_j)$ and threshold σ.
- <u>Step 4</u>: Apply truncated random walk for corpus generation and Skipgram model
- <u>Time complexity</u>:
  - $O(Hnlog(n))$ (n: # of nodes, H: small constant)

Graph kernel can also be used, but suffers from computational cost ($O(n^3)$) [1].



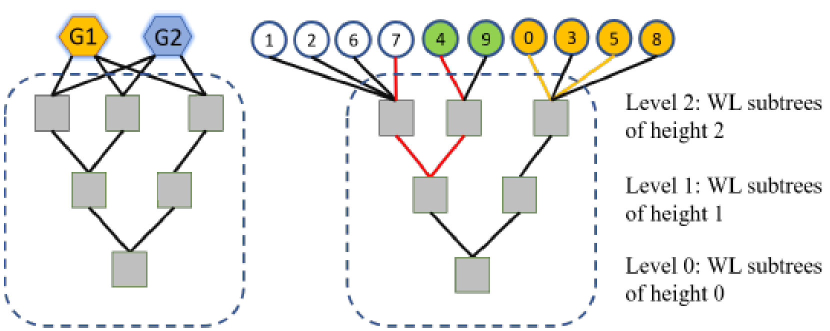
<u>**Notation**</u>:
$Q^h_{S_i}$: sorted degree sequence of S_i on level h
DTW: Dynamic Time Wrapping
$\tilde{Q}^h_{S_i}$: sorted degree sequence on level h with zero filling

[1] Vishwanathan, S. V. N., Karsten M. Borgwardt, and Nicol N. Schraudolph. "Fast computation of graph kernels." *NIPS*. Vol. 19. 2006.
[2] Ribeiro, Leonardo FR, Pedro HP Saverese, and Daniel R. Figueiredo. "struc2vec: Learning node representations from structural identity." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

MrMine+: Method Details

- Step 1: H iterations of WL node label transformation:
  - Generate unique WL subtrees of up to height H as subgraphs
- Step 2: Construct H-CRCN relation network:
  - Add **subtree vertices** to hierarchical relation network of WL subtrees (HRN)
  - Add **links** to HRN based on WL subtree generation relation
  - Attach **network vertices/node vertices** to the last level of HRN
- Step 3: Apply truncated random walk on two H-CRCN networks for corpus generation and Skipgram model
- Time complexity:
  - $O(Hm + cn)$ (n: # of nodes; m: # of edges; H, c: small constants)

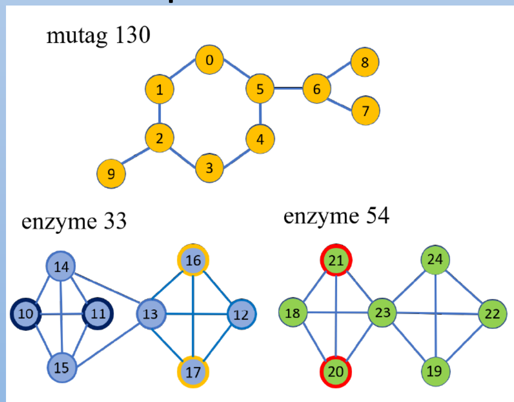
More in the paper



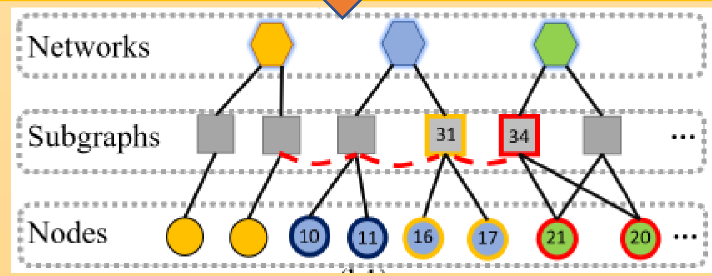
Nodes in the (H-)CRCN relation networks are called vertices

Summary of major steps

Three input networks

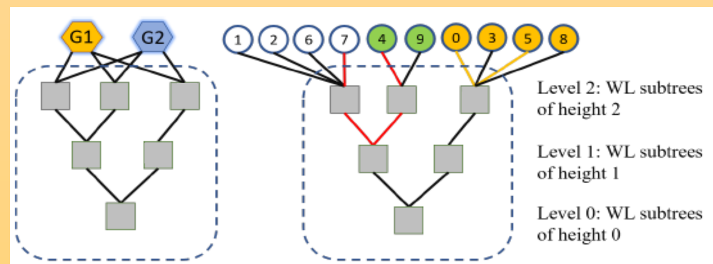


WL label transformation for subgraph generation



CRCN relation network

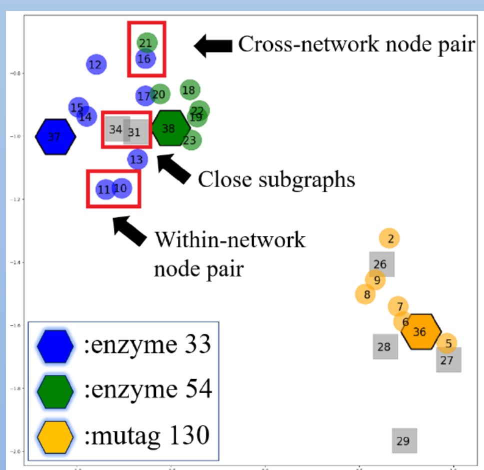
Or



H-CRCN relation network

Truncated random walk for corpus building and Skipgram model

Learned embeddings of nodes, subgraphs, and networks

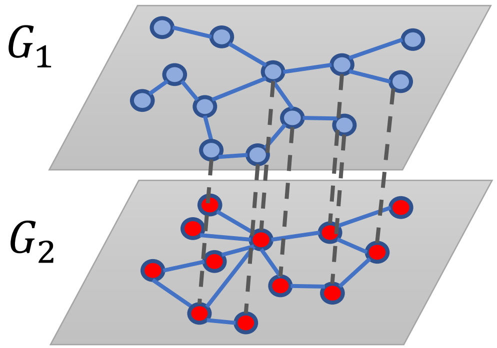


Roadmap

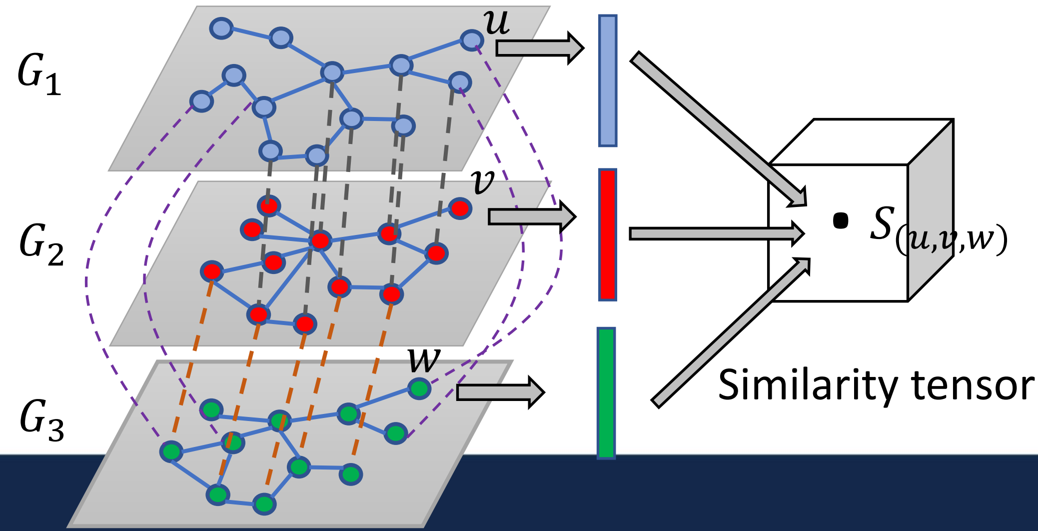
- Motivation
- Problem Definition
- Proposed Solution: ‘MrMine’
- Experiments
- Conclusions

Experimental Setup (downstream tasks)

- Query node retrieval:
  - Given a set of nodes from G_1, retrieve similar nodes from G_2.
- Network classification:
  - Classify networks into different categories.
- Two network alignment:
  - Align nodes from two input network G_1, G_2.
- Collective Network alignment:
  - Collectively align nodes in input networks G_1, G_2, G_3.



Two network alignment



Collective network alignment

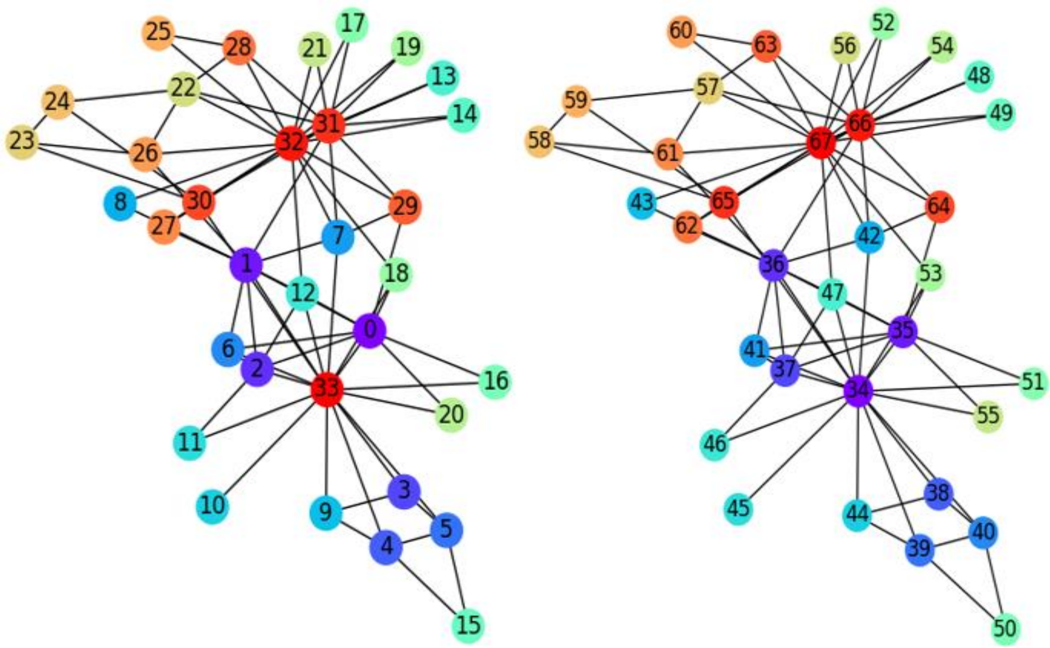
Real-world Datasets

Name	Category	# of Nodes	# of Edges
DBLP	Co-authorship	1,013	3,022
Flickr	User relationship	3,911	4,152
LastFm	User relationship	4,068	4,347
Douban	User relationship	1,118	3,022
MySpace	Social network	6,362	6,514
Aminer	Academic network	1,274,360	4,756,194

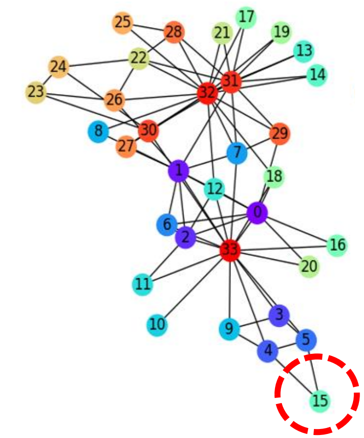
Bioinformatics	# of graphs	Classes	Avg. nodes
MUTAG	188	2	17.9
PTC	344	2	25.5
PROTEINS	1113	2	39.1
NCI1	4110	2	29.8
NCI109	4127	2	29.6

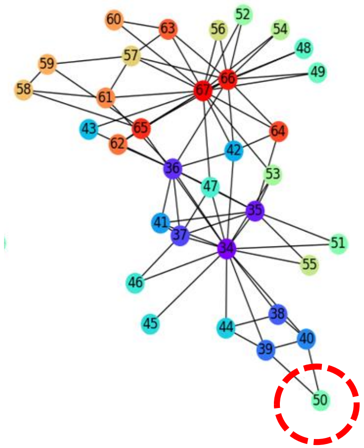
Visualization

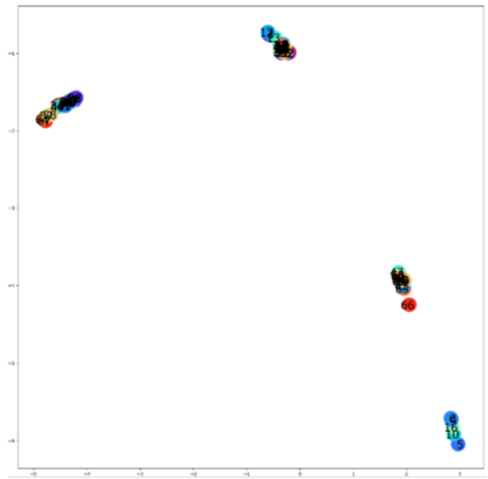
- Dataset: Mirrored Zachary's Karate Club data
- Corresponding nodes are colored the same.
- Learn the node embeddings first and project onto 2-D space.



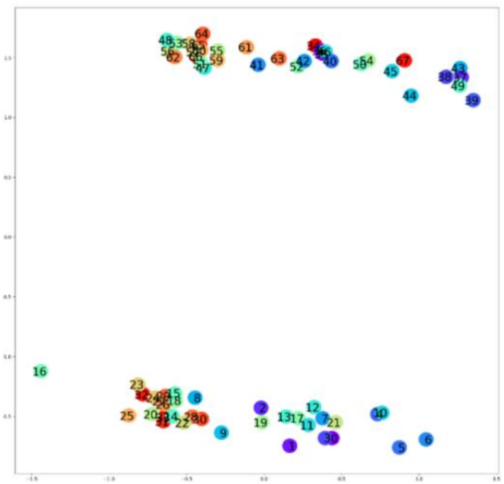
Visualization results



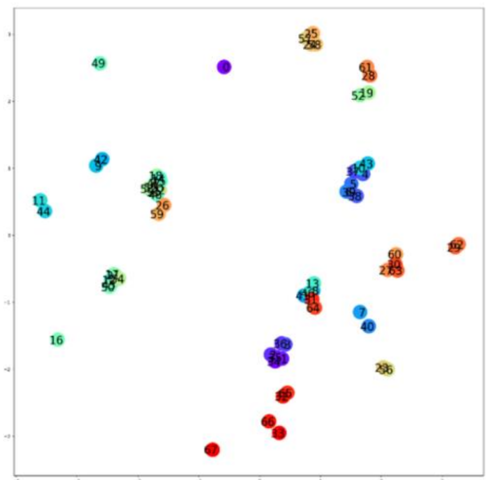




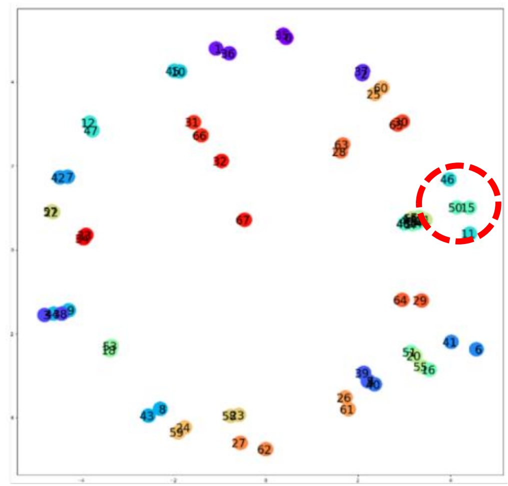
deepwalk



node2vec



struc2vec



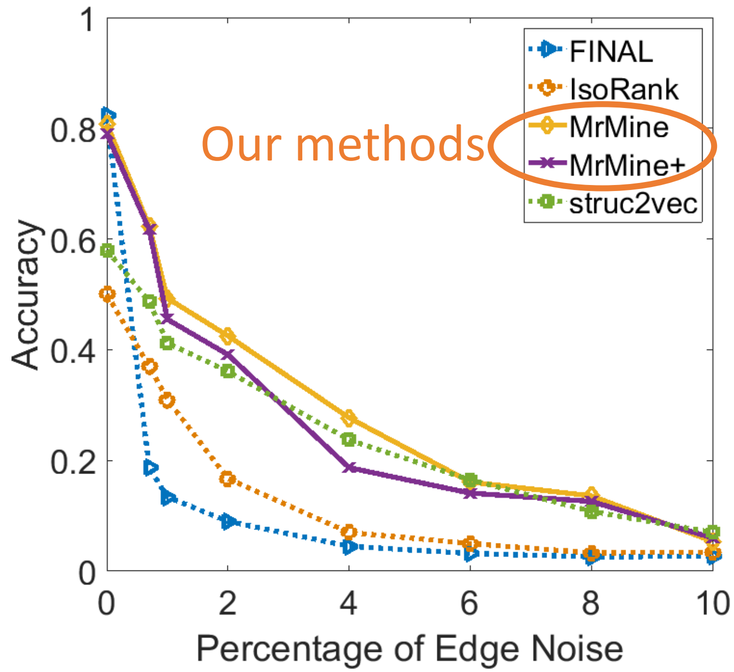
Paired

MrMine

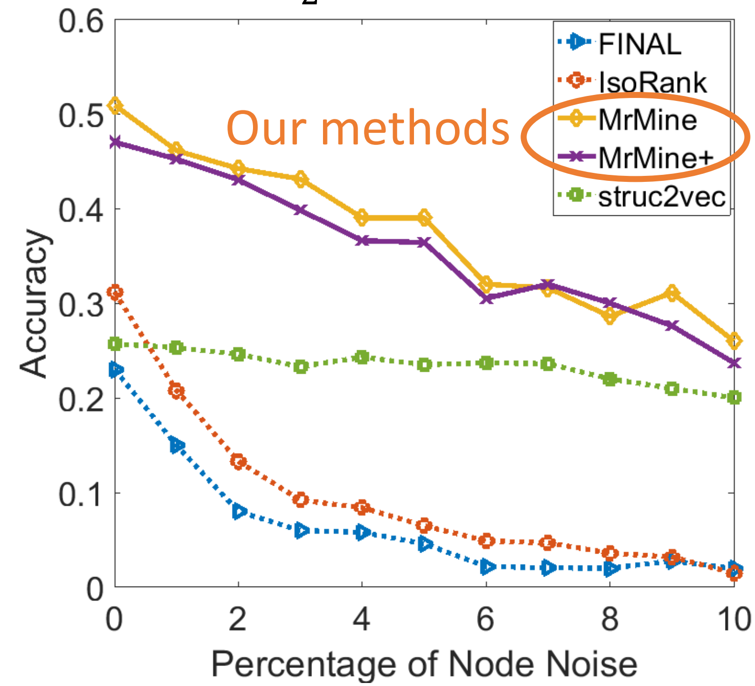
Identical nodes (same color) are supposed to be paired (close)

Network alignment

G_1: DBLP & G_2: DBLP with random noises



G_1: Douban online & G_2:Douban offine



- <u>Observation</u>:
  - Our methods are competitive in DBLP data; outperform baselines in douban data.
  - Our methods are more robust against noises.

Network classification

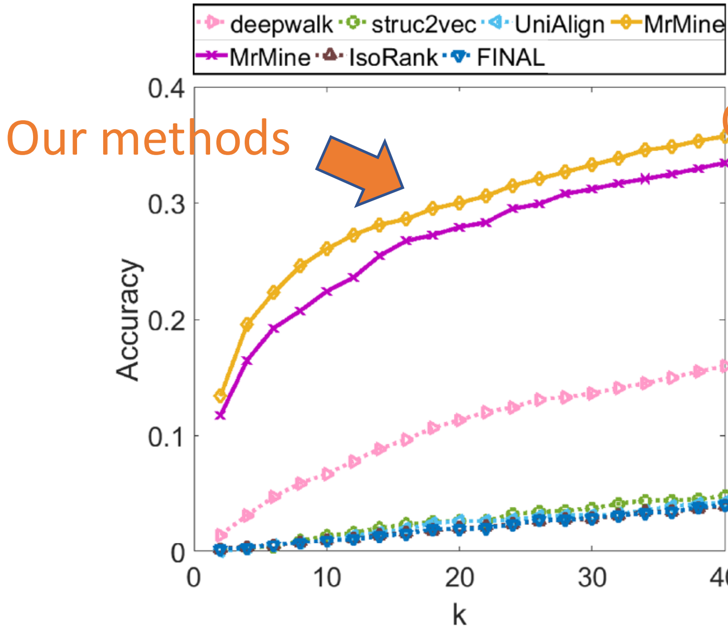
	MUTAG	PTC	PROTEINS	NCI1	NCI109
MrMine+	**83.47±2.01**	**62.00±0.07**	71.22±0.62	**68.50±0.03**	**65.57±0.02**
MrMine	82.19±1.58	55.41±2.52	70.88±0.38	66.90±0.05	64.53±0.01
WL Kernel	80.66±3.07	59.94±2.79	64.45±1.14	63.42±0.22	62.94±0.42
Deep WL Kernel	82.95±1.96	53.29±1.53	69.49±0.26	62.83±0.25	62.47±0.15
subgraph2vec	79.33±0.07	42.29±0.09	**73.04±0.04**	63.01±0.01	49.20±0.02

(± standard deviation)

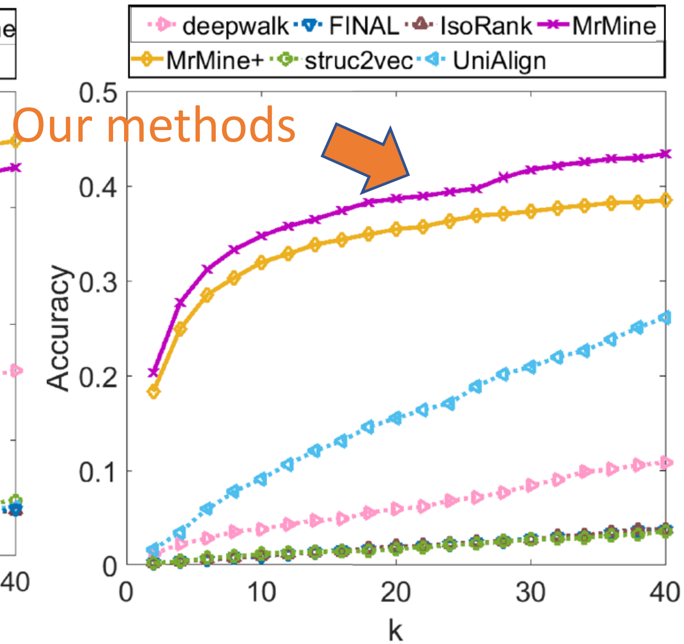
- Observation:
  - Our methods are competitive against baseline methods.
  - MrMine+ consistently outperform MrMine
  - (Cross-network relation captured by H-CRCN relation network is more effective than the basic CRCN relation network in network classification task)

Node retrieval

G_1: DBLP & G_2: DBLP with random noises

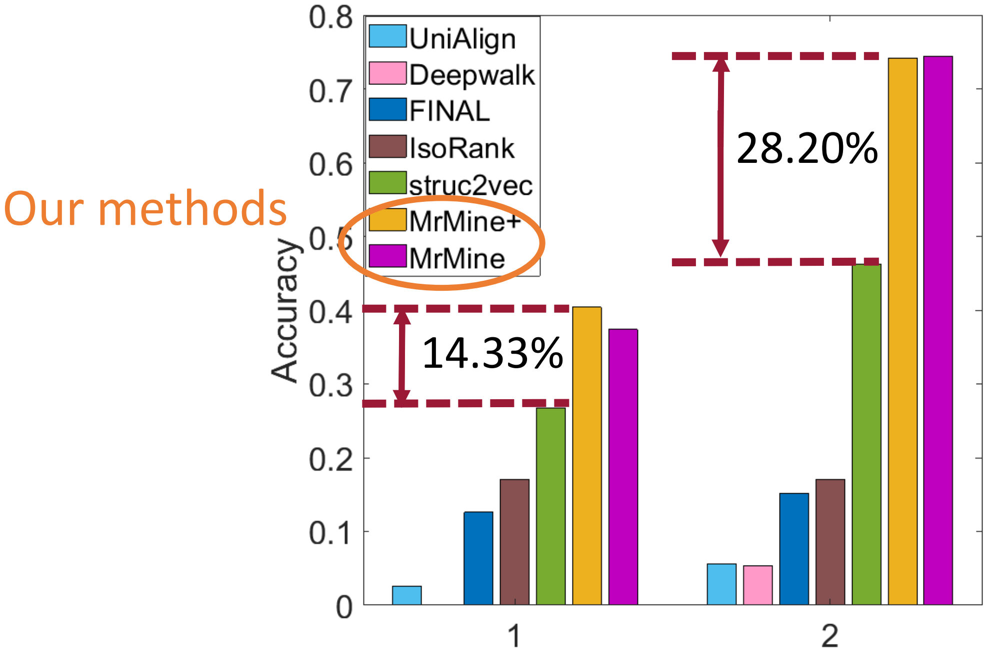


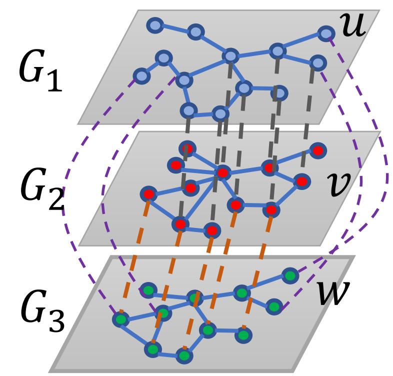
G_1: Douban online & G_2: Douban offline



- K: For one query node of G_1, top-k node list are retrieved from G_2.
- Accuracy = # of hits/# of query nodes (hit: correct node appears in top-k list)
- Observation: Our methods outperform baselines.

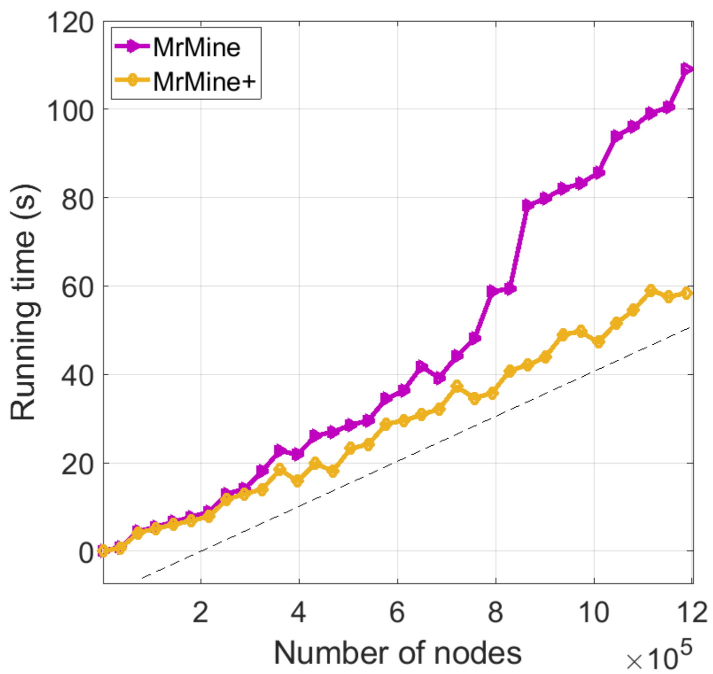
Collective network alignment





- Dataset: Douban-online, douban-offline, douban-online with random noises
- Metrics: For each pair of three-node alignment:
  - Metric 1: successfully alignment when all nodes are aligned correctly.
  - Metric 2: successfully alignment when two of the three nodes are aligned correctly.
- Obs.: our methods outperform all baselines (embedding-based and non embedding-based)

Scalability



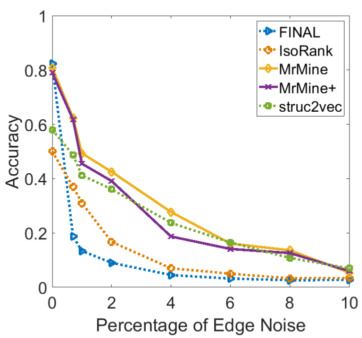
- <u>Observation</u>:
  - MrMine scales super-linearly w.r.t. # of nodes of input networks
  - MrMine+ scales linearly w.r.t. # of nodes of input networks

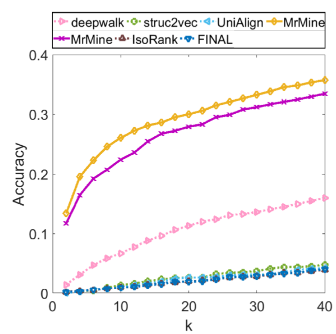
Roadmap

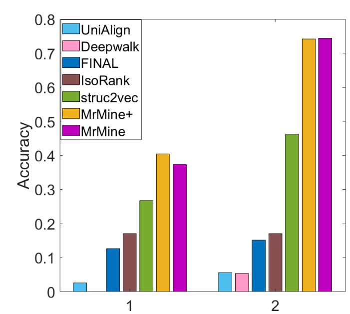
- Motivation
- Problem Definition
- Proposed Solution: ‘MrMine’
- Experiments
- Conclusions

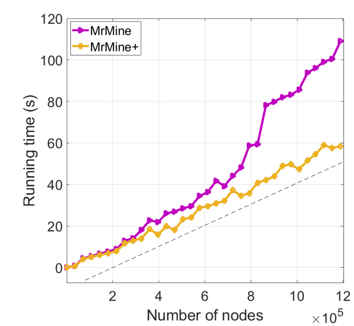
Conclusion

- Goal: *Unsupervised* **multi-resolution multi-network** embedding.
- Solution: MrMine, MrMine+
  - Key idea 1: Cross-Resolution Cross-Network (CRCN) relation network
  - Key idea 2: WL subtree as the subgraph resolution
  - Key idea 3: Hierarchical structure of CRCN network
- Results:
  - Boost traditional network mining tasks (e.g. network classification)
  - Enable novel network mining tasks (e.g. collective network alignment)
  - Accelerated method has *linear* time complexity









Thank you!

- Q & A Session