Flexible and Robust Multi-Network Clustering

Jingchao Ni[1], Hanghang Tong[2], Wei Fan[3], Xiang Zhang[1]

[1]Department of Electrical Engineering and Computer Science, Case Western Reserve University

[2]School of Computing, Informatics, Decision Systems Engineering, Arizona State University

[3]Baidu Research Big Data Lab

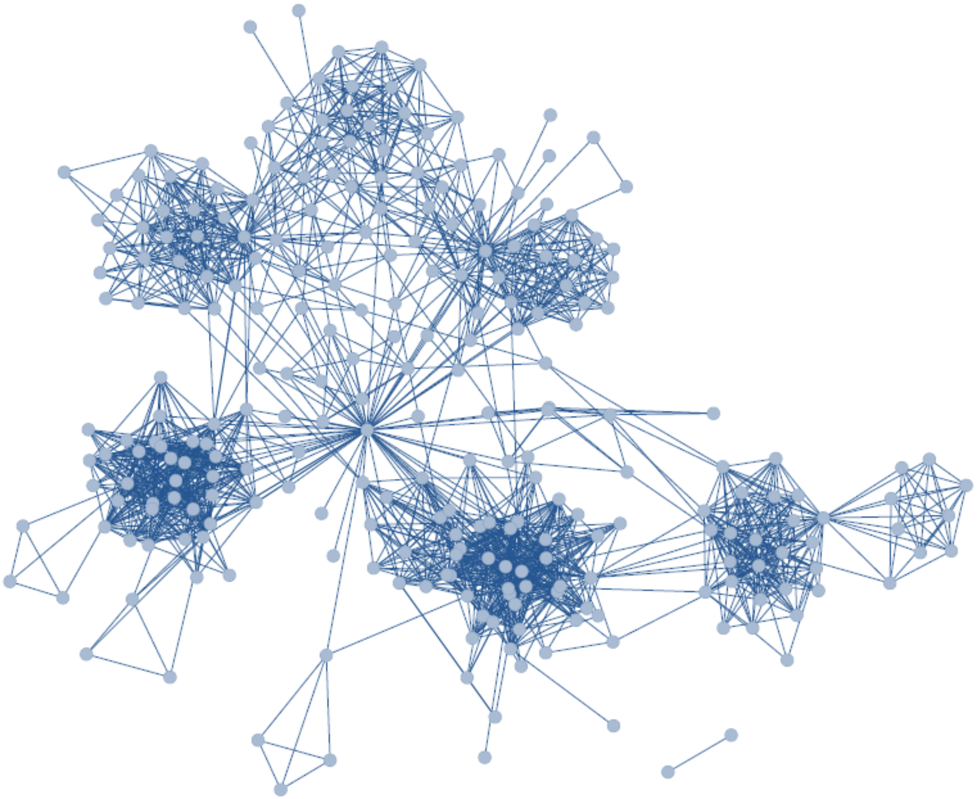
The 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

KDD2015
21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining
10 - 13 August 2015, Hilton, Sydney

Network Clustering

- ❑ **Network Data are ubiquitous**
  - ➢ Web networks
  - ➢ Social networks
  - ➢ Biological networks, etc.



- ❑ **Network Clustering**
  - ➢ Detect sub-networks that satisfy certain properties
  - ➢ Many connections within clusters and few connections across clusters

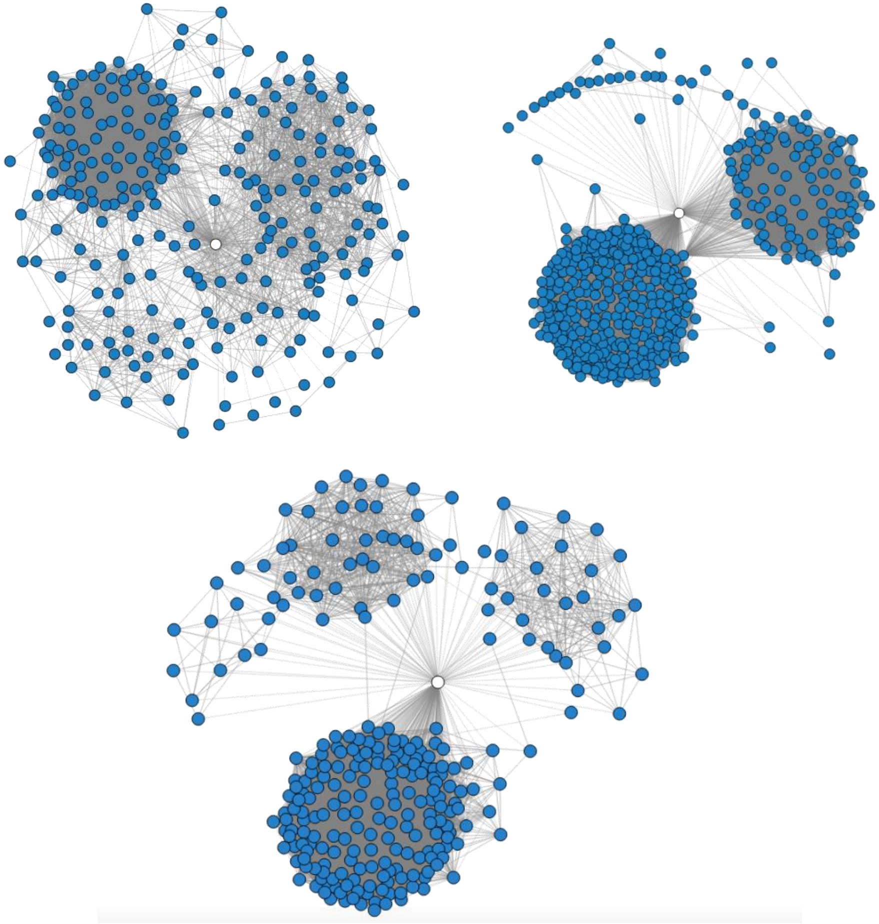
CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. ***Flexible and Robust Multi-Network Clustering.*** In KDD, 2015.

Network Clustering

- □ **Network Data are ubiquitous**
  - ➢ Web networks
  - ➢ Social networks
  - ➢ Biological networks, etc.

- □ **Network Clustering**
  - ➢ Detect sub-networks that satisfy certain properties
  - ➢ Many connections within clusters and few connections across clusters

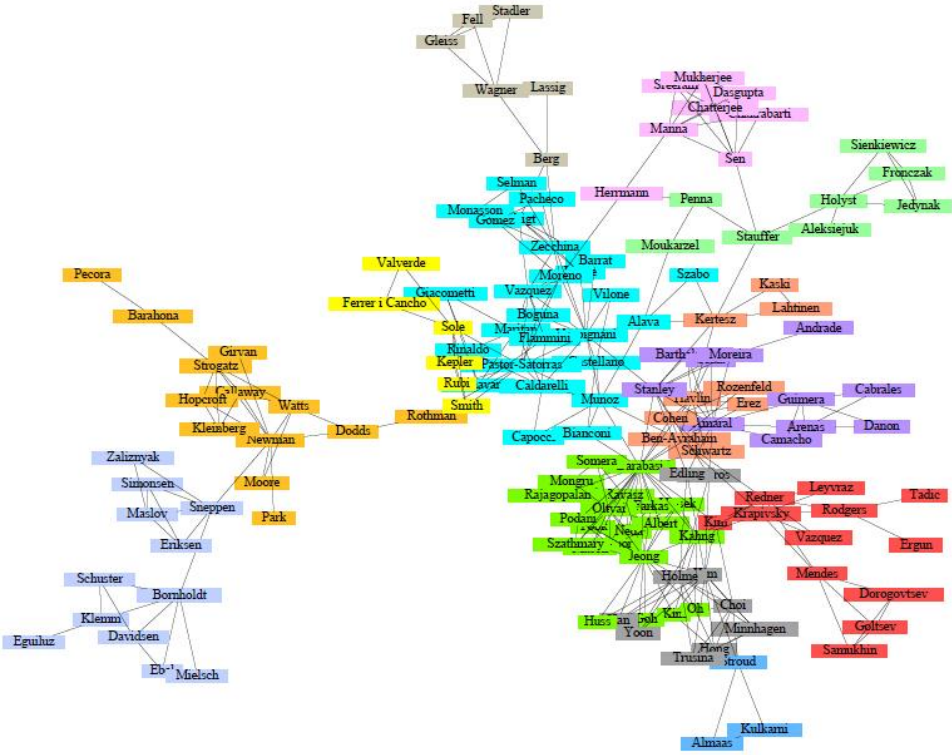


Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Network Clustering

- ❑ **Network Data are ubiquitous**
  - ➢ Web networks
  - ➢ Social networks
  - ➢ Biological networks, etc.

- ❑ **Network Clustering**
  - ➢ Detect sub-networks that satisfy certain properties
  - ➢ Many connections within clusters and few connections across clusters



Coauthorship network between physicisits

Figure from "Mark EJ Newman and Michelle Girvan. *Finding and evaluating community structure in networks*. Physical review E 69.2 (2004): 026113."

Network Clustering

- ❑ **Network Data are ubiquitous**
  - ➢ Web networks
  - ➢ Social networks
  - ➢ Biological networks, etc.

- ❑ **Network Clustering**
  - ➢ Detect sub-networks that satisfy certain properties
  - ➢ Many connections within clusters and few connections across clusters

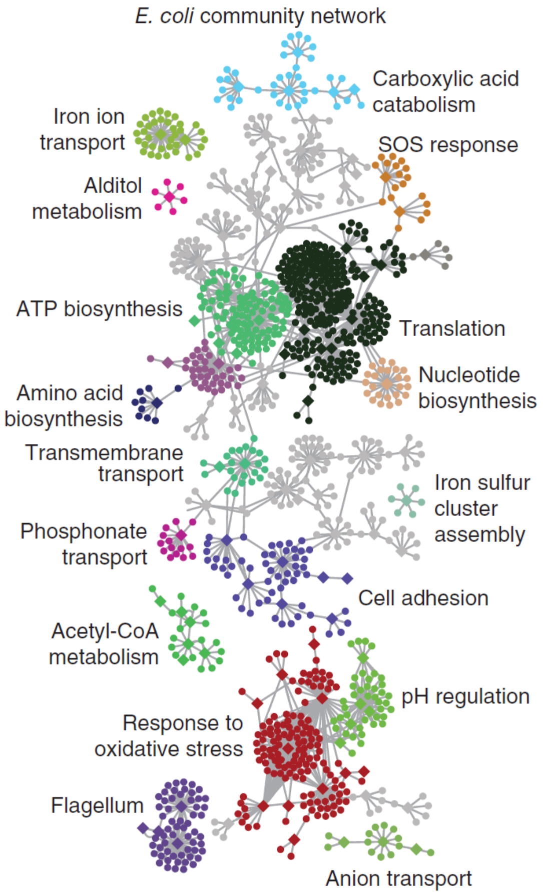


Figure from "Daniel Marbach, et al. *Wisdom of crowds for robust gene network inference*. Nature methods 9.8 (2012): 796-804."

Multi-Network Clustering

- ❑ **Networks collected from multiple conditions, sources or domains**
  - ➢ E.g., co-author networks from different research areas
  - ➢ E.g., gene co-expression networks from different tissues of model organisms

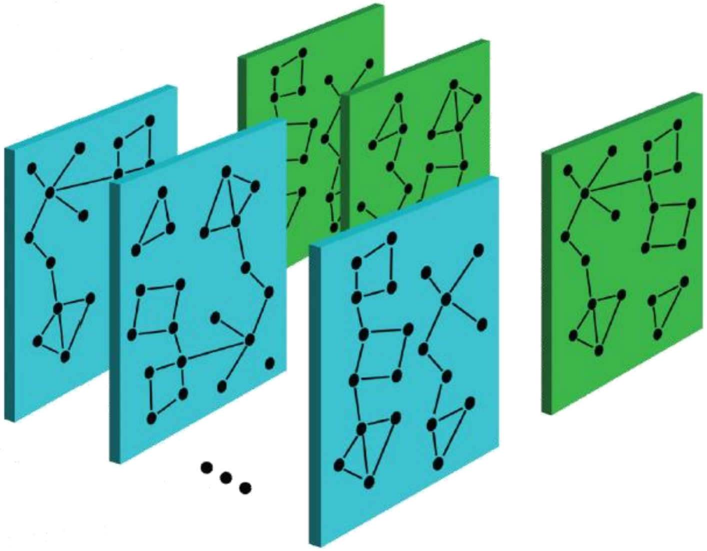


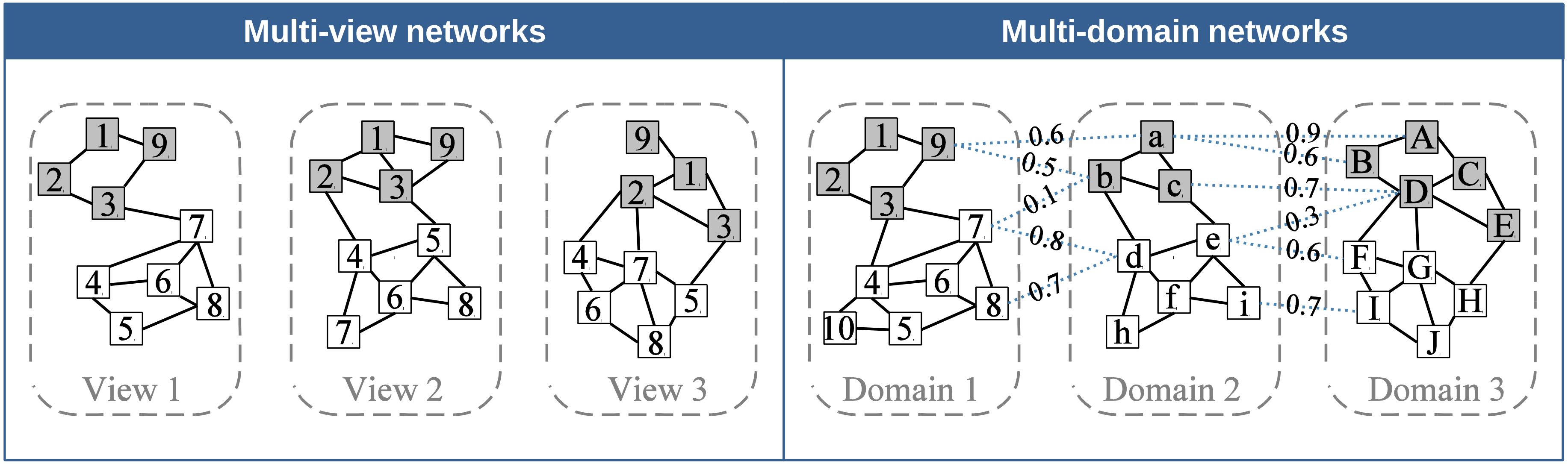
Figure from “Mikko Kivelä, et al. *Multilayer networks*. Journal of Complex Networks 2.3 (2014): 203-271.”

- ❑ **Multi-network clustering motivation**
  - ➢ Single network can be noisy, incomplete and provide partial knowledge
  - ➢ Multi-network can provide compatible and complementary information
  - ➢ Multi-network can be robust to noise in individual networks

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang. ***Flexible and Robust Multi-Network Clustering.*** In KDD, 2015.

Multi-Network Clustering

❑ Multi-view and multi-domain network clustering[1,2]



Multi-view networks

Multi-domain networks

❑ Key assumption

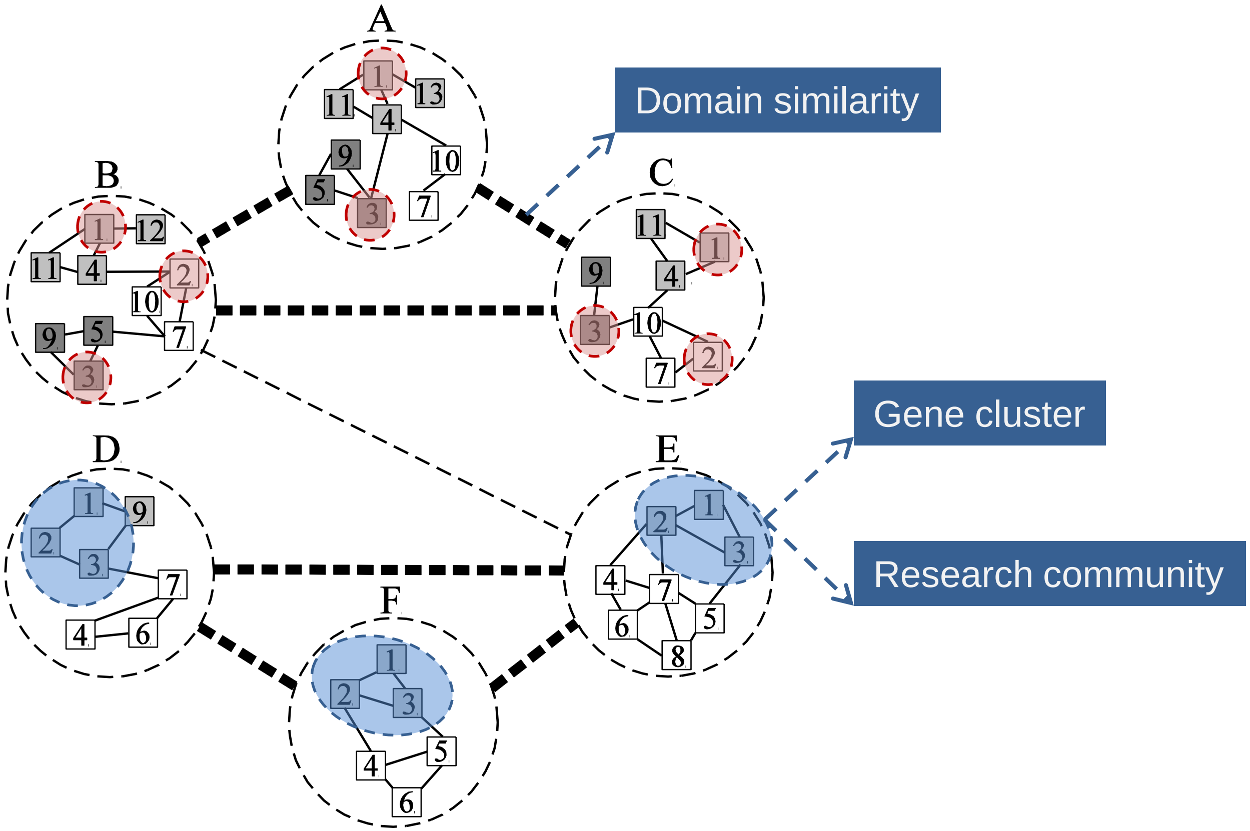
- ➢ Different views/domains share the same underlying clustering structure
- ➢ Methods are designed to identify consistent clustering structure across all views/domains

[1] Abhishek Kumar, et al., *Co-regularized multi-view spectral clustering*. In NIPS, 2011.

[2] Wei Cheng, et al., *Flexible and robust co-regularized multi-domain graph clustering*. In KDD, 2013.

Motivation

- ❑ **In many emerging applications, different networks have different data distributions**



Domain similarity

Gene cluster

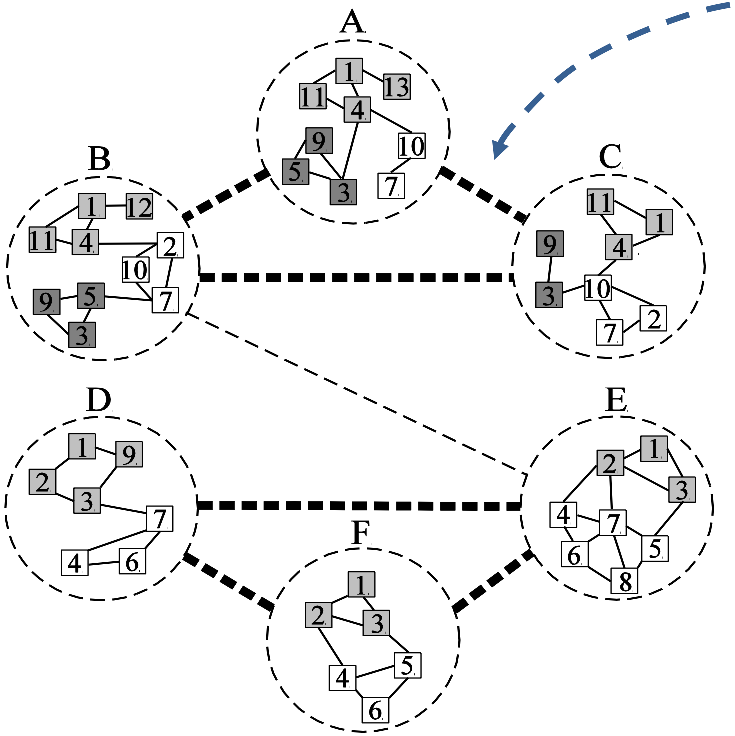
Research community

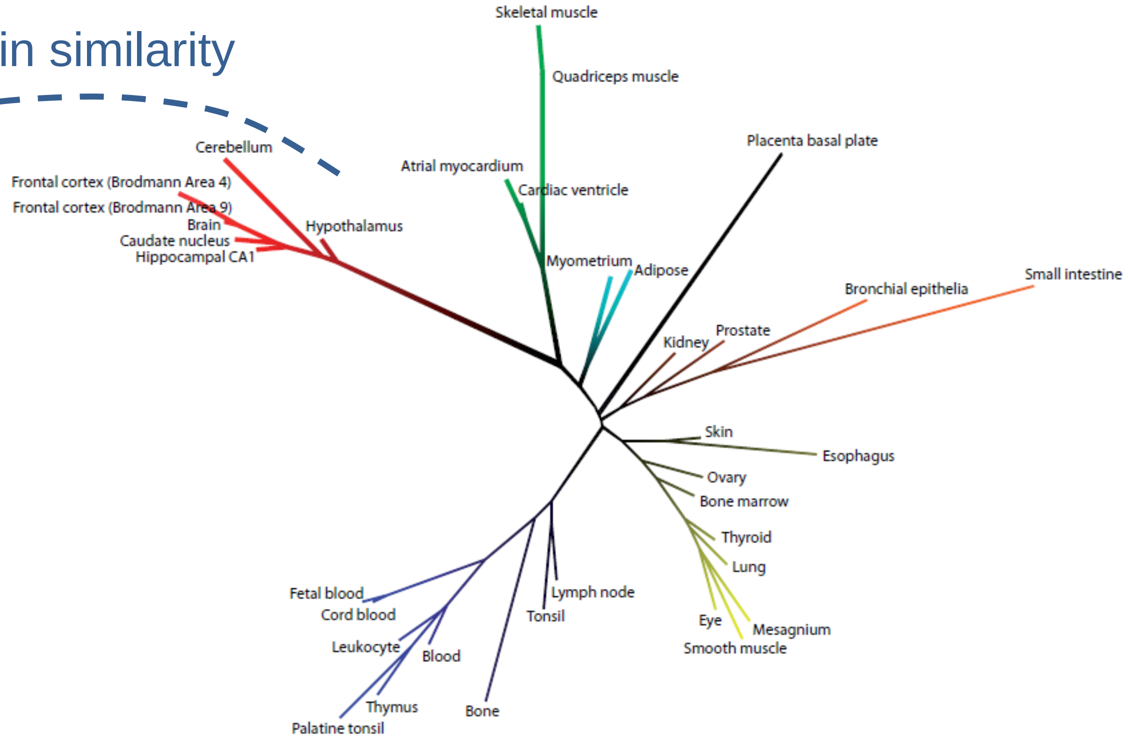
Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Motivation

- ❑ **In many emerging applications, different networks have different data distributions**

Domain similarity



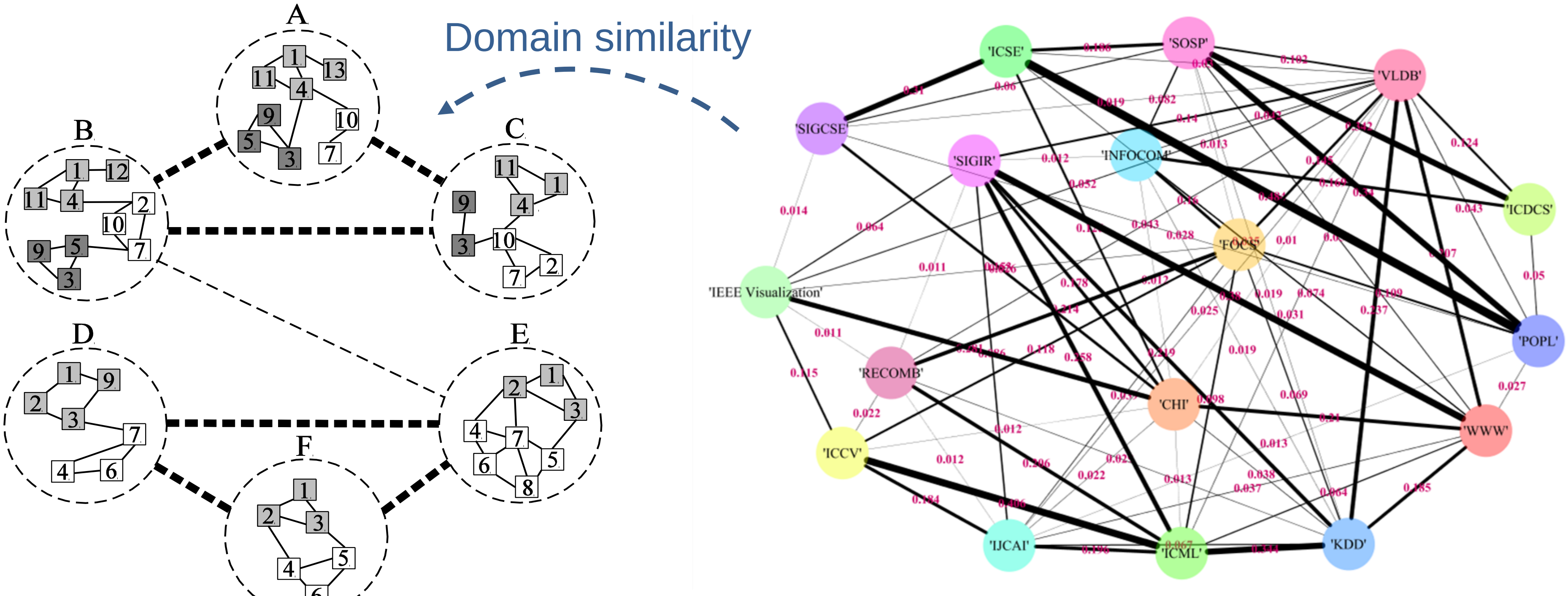


An unrooted tissue distance tree

Figure from "Daniela Börnigen, et al. *Concordance of gene expression in human protein complexes reveals tissue specificity and pathology*. Nucleic acids research 41.18 (2013): e171-e171."

Motivation

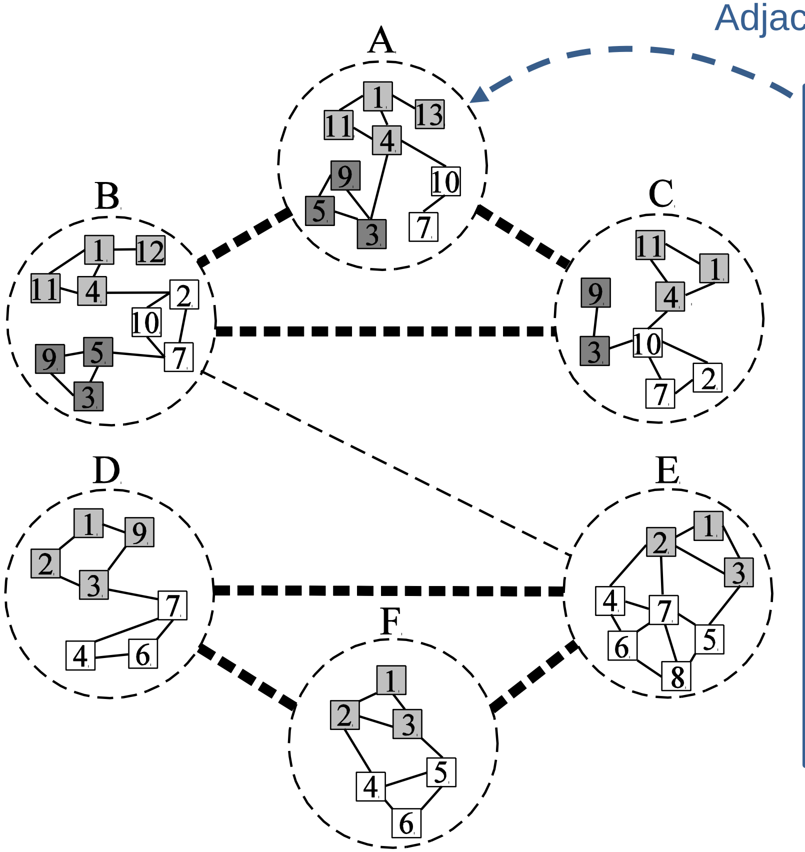
- ❑ **In many emerging applications, different networks have different data distributions**



A research conference similarity network

Motivation

❑ Network of Networks (NoN)



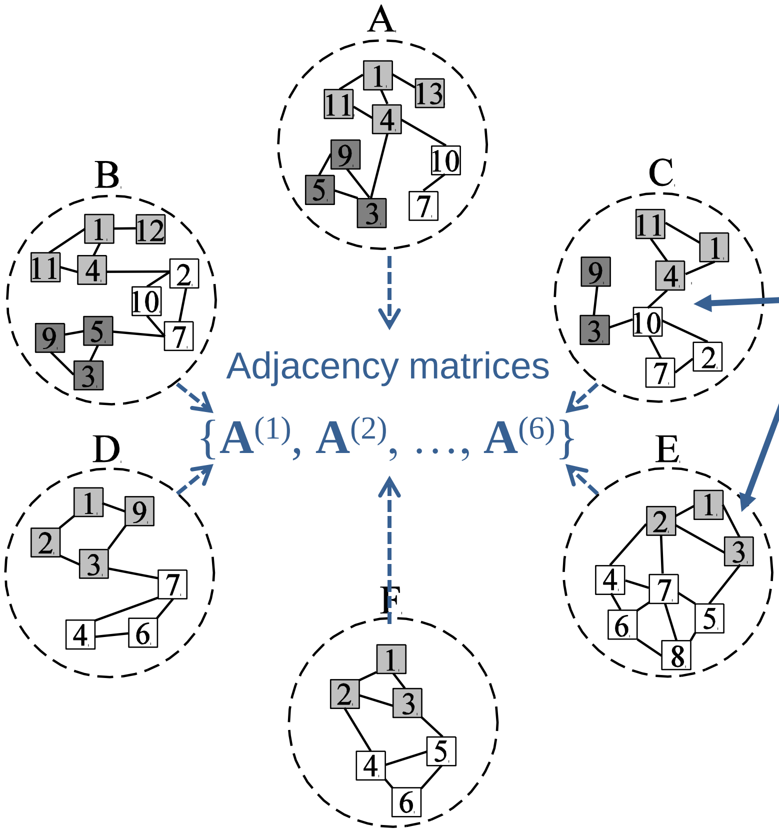
Adjacency matrix $\mathbf{G}$

- ❑ The dashed line network formed by A to F is called the **main network**. Denoted as $\mathbf{G}$.
- ❑ The solid line networks formed by 1 to 10 are called the **domain-specific networks**. Denoted as $\{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(g)}\}$.
- ❑ The goal of this work is to simultaneously clustering multi-network by using their multiple underlying clustering structures.

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Motivation

❑ Network of Networks (NoN)

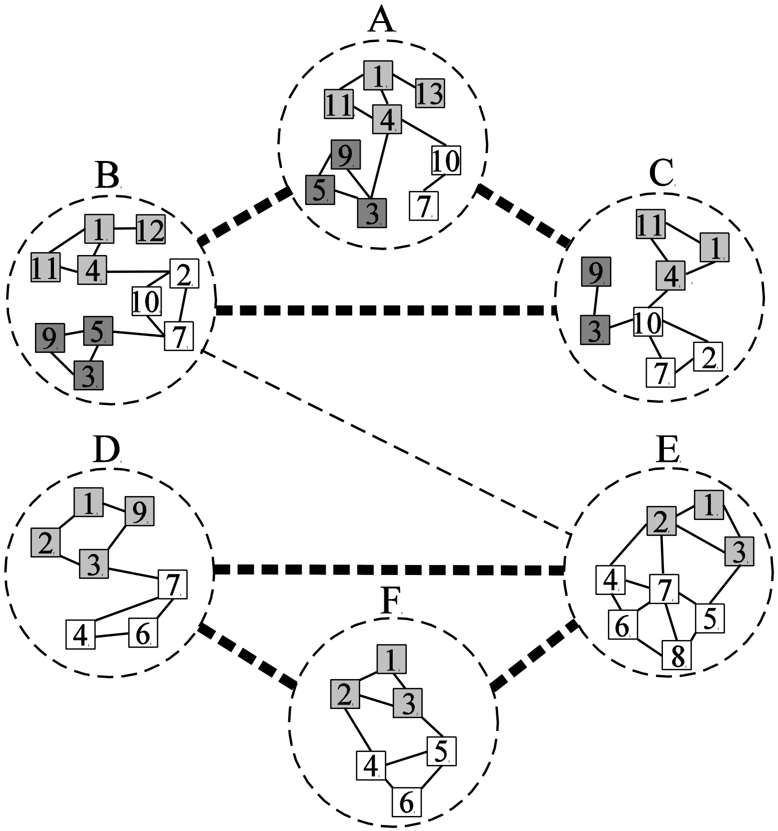


A
B
C
D
E
F
Adjacency matrices
$\{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(6)}\}$

- ❑ The dashed line network formed by A to F is called the **main network**. Denoted as **G**.
- ❑ The solid line networks formed by $\boxed{1}$ to $\boxed{10}$ are called the **domain-specific networks**. Denoted as $\{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(g)}\}$.
- ❑ The goal of this work is to simultaneously clustering multi-network by using their multiple underlying clustering structures.

Motivation

❑ Network of Networks (NoN)



- ❑ The dashed line network formed by A to F is called the **main network**. Denoted as $\mathbf{G}$.
- ❑ The solid line networks formed by $\boxed{1}$ to $\boxed{10}$ are called the **domain-specific networks**. Denoted as $\{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(g)}\}$.
- ❑ The goal of this work is to simultaneously clustering multi-network by using their multiple underlying clustering structures.

Problem Formulation

❑ **Phase I: Main Network Clustering**

- Symmetric Non-negative Matrix Factorization (SNMF)
- Minimizing

$$\boxed{J_M = \left\| \mathbf{G} - \mathbf{H}\mathbf{H}^T \right\|_F^2 \qquad s.t. \quad \mathbf{H} \geq 0}$$

- where $\mathbf{H} \in \Re_+^{g \times k}$ is the factor matrix of $\mathbf{G}$. k is the number of main clusters.
- Main cluster: the cluster in the main network
- h_{ij} indicates to which degree a main node i belongs to the j^{th} main cluster.

Problem Formulation

❑ Phase II: Domain-specific Network Clustering (A Simplified Case)

- ➢ Assumption: domain-specific networks in the same main cluster share a common underlying clustering structure, so we have k underlying clustering structures.

The number of main clusters

- ➢ A simplified case: all domains have n nodes and t clusters.
- ➢ Let the domain cluster assignment vector for node x in $\mathbf{A}^{(i)}$ be $u_{x*}^{(i)}$ $(i = 1, \ldots, g)$.
- ➢ Define k *hidden* domain cluster assignment vectors $v_{x*}^{(j)} \in \Re_{+}^{1\times t}$ $(j=1, \ldots, k)$ for each domain node x.

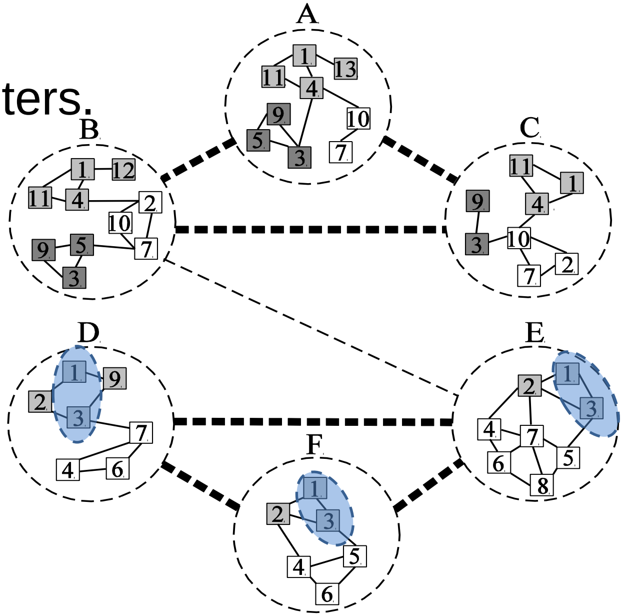
$$J_x = \sum_{i=1}^{g}\sum_{j=1}^{k} h_{ij} \left\| u_{x*}^{(i)} - v_{x*}^{(j)} \right\|_F^2 \Rightarrow J_D = \underbrace{\sum_{i=1}^{g} \left\| \mathbf{A}^{(i)} - \mathbf{U}^{(i)}(\mathbf{U}^{(i)})^T \right\|_F^2}_{\text{Domain-specific network clustering}} + \underbrace{a \sum_{i=1}^{g}\sum_{j=1}^{k} h_{ij} \left\| \mathbf{U}^{(i)} - \mathbf{V}^{(j)} \right\|_F^2}_{\text{Main cluster guided regularization}}$$

Recall h_{ij} represents main cluster membership

Problem Formulation

❑ Phase II: Domain-specific Network Clustering (The General Case)

- ➢ Different domains can have different set of nodes thus different sizes.
- ➢ Define two mapping matrices $\mathbf{O}^{(ij)} \in \{0,1\}^{n_i \times \tilde{n}_j}$, $\mathbf{D}^{(ij)} \in \{0,1\}^{n_i \times n_i}$ such that the same rows of $\mathbf{D}^{(ij)}\mathbf{U}^{(i)}$ and $\mathbf{O}^{(ij)}\mathbf{V}^{(j)}$ represent the same instances.
- ➢ Different domains can have different number of clusters.



- ➢ Indirect regularization:

Example: if nodes $\boxed{1}$ *and* $\boxed{3}$ *have similar cluster assignments in* D *, their cluster assignments in the underlying clustering structure shared by* {D, E, F} *should be similar as well.*

- ➢ Minimize $\boxed{h_{ij}\left(\hat{\mathbf{u}}_{x^*}^{(ij)}(\hat{\mathbf{u}}_{y^*}^{(ij)})^T - \hat{\mathbf{v}}_{x^*}^{(ij)}(\hat{\mathbf{v}}_{y^*}^{(ij)})^T\right)^2}$

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Problem Formulation

❑ Phase II: Domain-specific Network Clustering (The General Case)

- ➢ Optimization problem

$$\min_{\substack{\mathbf{U}^{(i)}\geq 0,(i=1,\ldots,g)\\ \mathbf{V}^{(j)}\geq 0,(j=1,\ldots,k)}} J_D = \sum_{i=1}^{g} J_A + a\sum_{i=1}^{g}\sum_{j=1}^{k} h_{ij} J_R$$

Domain-specific network clustering

Main cluster guided regularization

Where

$$J_A = \left\|\mathbf{A}^{(i)} - \mathbf{U}^{(i)}(\mathbf{U}^{(i)})^T\right\|_F^2$$

$$J_R = \sum_{x,y=1}^{n_i} (\hat{\mathbf{u}}_{x*}^{(ij)}(\hat{\mathbf{u}}_{y*}^{(ij)})^T - \hat{\mathbf{v}}_{x*}^{(ij)}(\hat{\mathbf{v}}_{y*}^{(ij)}))^2 = \left\|(\mathbf{D}^{(ij)}\mathbf{U}^{(i)})(\mathbf{D}^{(ij)}\mathbf{U}^{(i)})^T - (\mathbf{O}^{(ij)}\mathbf{V}^{(j)})(\mathbf{O}^{(ij)}\mathbf{V}^{(j)})^T\right\|_F^2$$

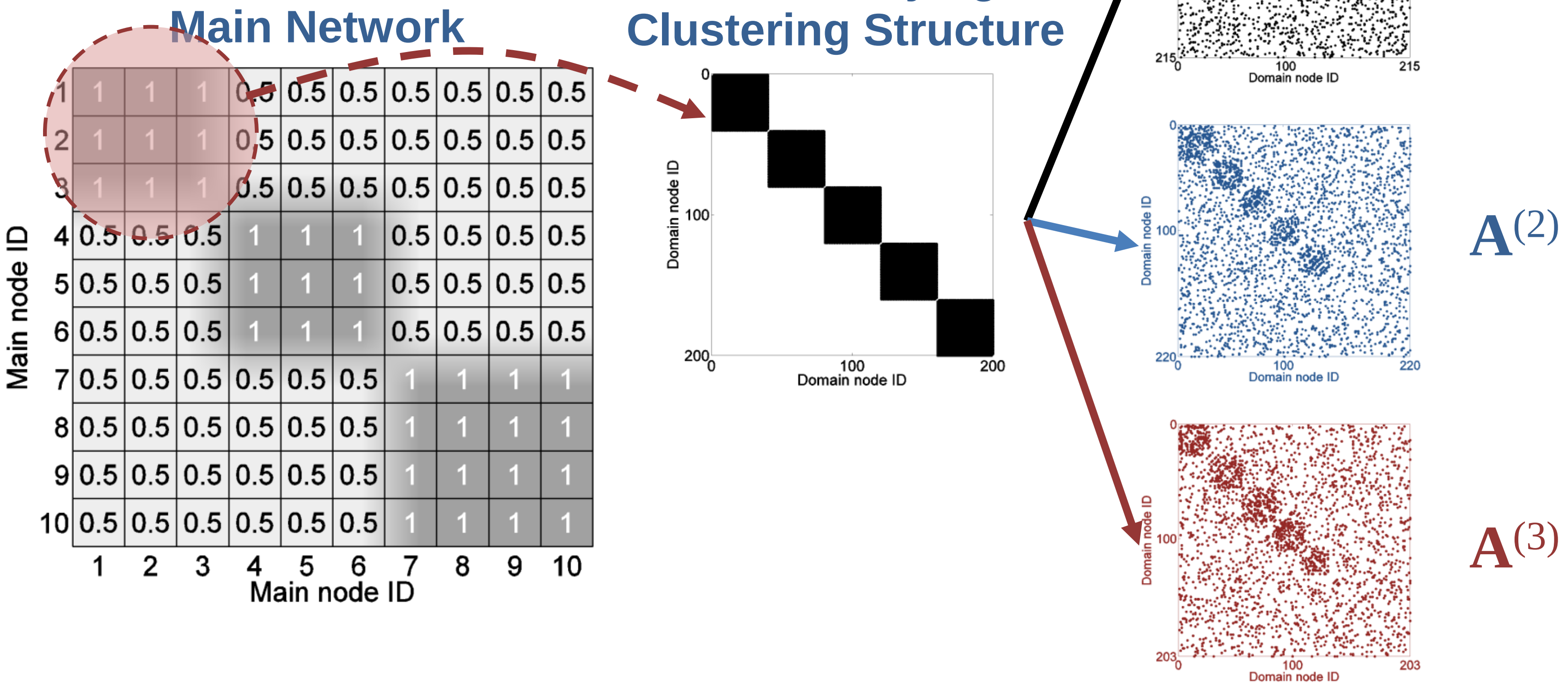
- ➢ Learning algorithm: an alternating minimization approach. $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(j)}$ are alternately solved by multiplicative updating rules with convergence guarantee.

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Experimental Results

❑ Simulation Study

- ➢ Synthetic data generation



Main Network

Main node ID	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	1	1	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5
3	1	1	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5
4	0.5	0.5	0.5	1	1	1	0.5	0.5	0.5	0.5
5	0.5	0.5	0.5	1	1	1	0.5	0.5	0.5	0.5
6	0.5	0.5	0.5	1	1	1	0.5	0.5	0.5	0.5
7	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1
8	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1
9	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1
10	0.5	0.5	0.5	0.5	0.5	0.5	1	1	1	1

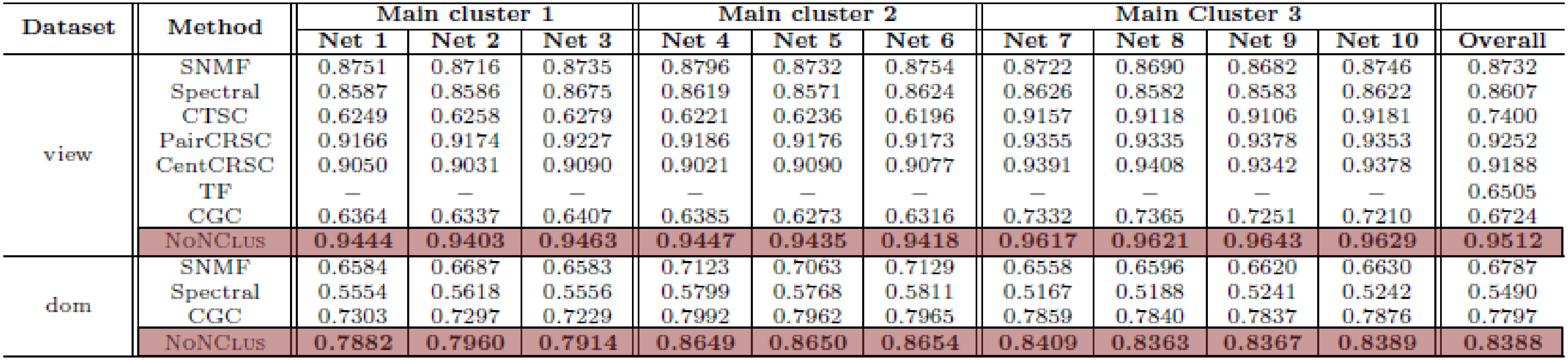
An underlying Clustering Structure

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Experimental Results

❑ Simulation Study

- ➢ Accuracy of different methods on synthetic datasets



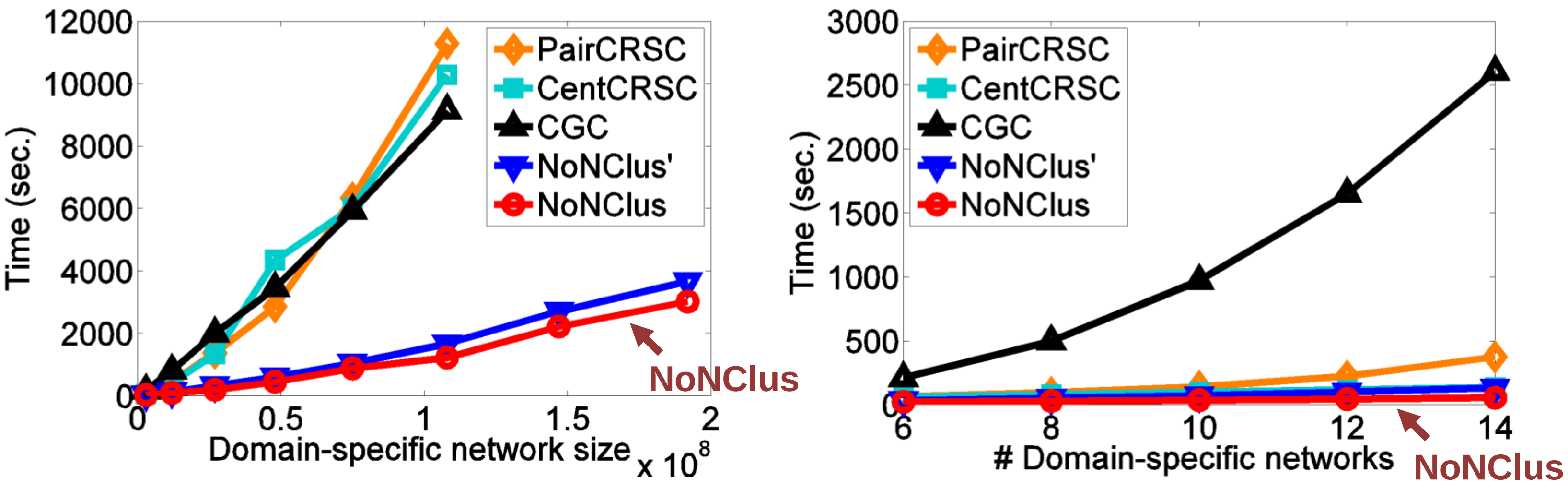
Dataset	Method	Main cluster 1			Main cluster 2			Main Cluster 3				
		Net 1	Net 2	Net 3	Net 4	Net 5	Net 6	Net 7	Net 8	Net 9	Net 10	Overall
view	SNMF	0.8751	0.8716	0.8735	0.8796	0.8732	0.8754	0.8722	0.8690	0.8682	0.8746	0.8732
	Spectral	0.8587	0.8586	0.8675	0.8619	0.8571	0.8624	0.8626	0.8582	0.8583	0.8622	0.8607
	CTSC	0.6249	0.6258	0.6279	0.6221	0.6236	0.6196	0.9157	0.9118	0.9106	0.9181	0.7400
	PairCRSC	0.9166	0.9174	0.9227	0.9186	0.9176	0.9173	0.9355	0.9335	0.9378	0.9353	0.9252
	CentCRSC	0.9050	0.9031	0.9090	0.9021	0.9090	0.9077	0.9391	0.9408	0.9342	0.9378	0.9188
	TF	–	–	–	–	–	–	–	–	–	–	0.6505
	CGC	0.6364	0.6337	0.6407	0.6385	0.6273	0.6316	0.7332	0.7365	0.7251	0.7210	0.6724
	NoNClus	**0.9444**	**0.9403**	**0.9463**	**0.9447**	**0.9435**	**0.9418**	**0.9617**	**0.9621**	**0.9643**	**0.9629**	**0.9512**
dom	SNMF	0.6584	0.6687	0.6583	0.7123	0.7063	0.7129	0.6558	0.6596	0.6620	0.6630	0.6787
	Spectral	0.5554	0.5618	0.5556	0.5799	0.5768	0.5811	0.5167	0.5188	0.5241	0.5242	0.5490
	CGC	0.7303	0.7297	0.7229	0.7992	0.7962	0.7965	0.7859	0.7840	0.7837	0.7876	0.7797
	NoNClus	**0.7882**	**0.7960**	**0.7914**	**0.8649**	**0.8650**	**0.8654**	**0.8409**	**0.8363**	**0.8367**	**0.8389**	**0.8388**

- ➢ In `view` dataset, all $\mathbf{A}^{(i)}$ have the same size. In `dom` dataset, different $\mathbf{A}^{(i)}$ have different sizes.

- ➢ CTSC, PairCRSC, CentCRSC are multi-view graph clustering methods. TF is the tensor factorization. CGC is a multi-domain graph clustering method.

Experimental Results

❑ Scalability Evaluation on Synthetic Dataset



(a) Varying network size **(a) Varying number of networks**

Running time evaluation

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

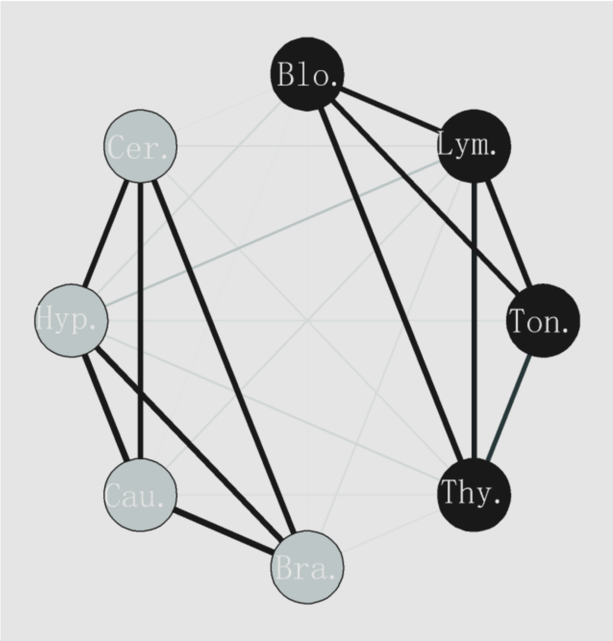
Experimental Results

❑ Functional Module Detection in Tissue-specific Gene Co-Expression Networks

Tissue-specific gene co-expression networks[1]

Tissue-specific Network	# nodes	# edges
Blood	633	2,573
Lymph node	648	2,256
Tonsil	682	2,480
Thymus	786	2,939
Brain	746	3,135
Caudate nucleus	640	2,578
Hypothalamus	641	2,500
Cerebellum	679	2,636
Total	**5,455**	**21,097**

*5372 samples for 128 different tissues in four different cell types, i.e., normal, disease, neoplasm and cell line. We select 8 tissues to construct gene co-expression networks.



Tissue-tissue similarity network (the main network in NoN)

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

[1] Margus Lukk, et al. *A global map of human gene expression*. Nature biotechnology 28.4 (2010): 322-324.

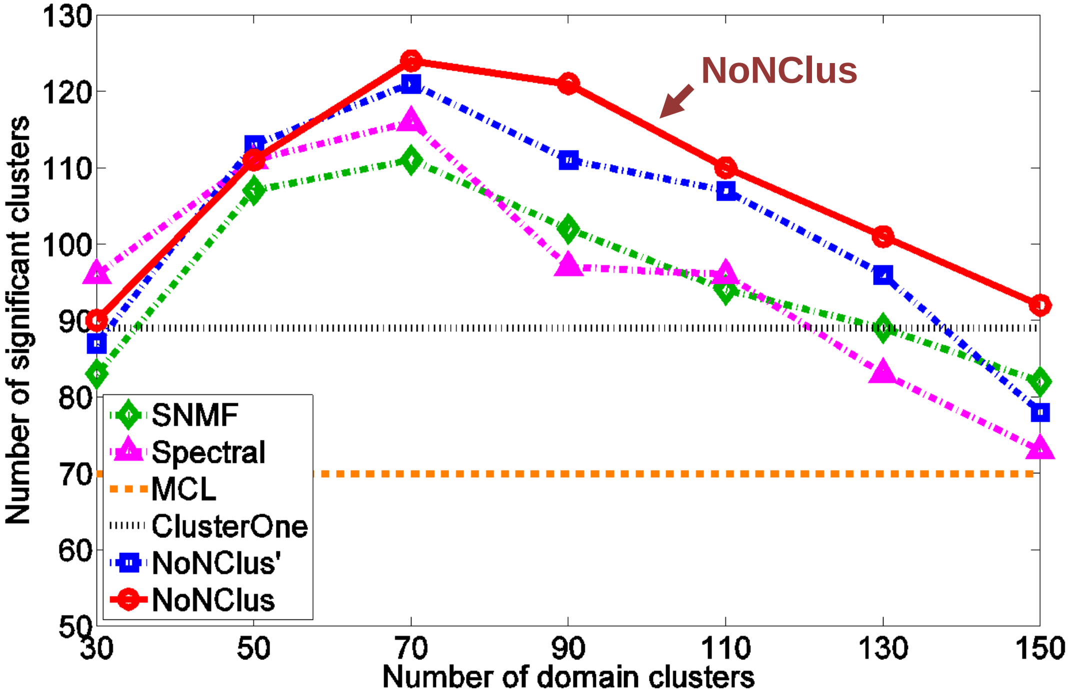
Experimental Results

❑ Functional Module Detection in Tissue-specific Gene Co-Expression Networks

- Evaluation method: standard Gene Set Enrichment Analysis (GSEA).
- The most significant Gene Ontology (GO) term in the biological process category is assigned to each identified gene cluster.
- The significance is assessed by Hypergeometric distribution.
- Raw p-values are adjusted for multiple testing problem by False Discovery Rate (FDR).

Experimental Results

❑ **Functional Module Detection in Tissue-specific Gene Co-Expression Networks**



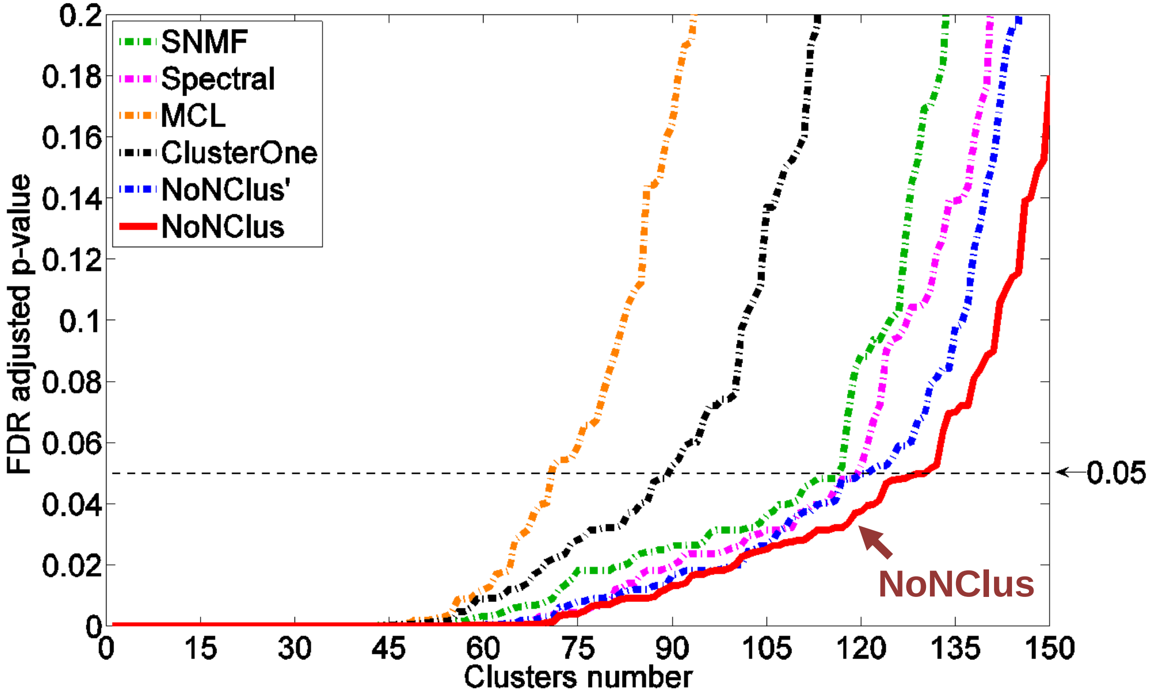
Number of detected significant clusters with various input number of clusters

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Experimental Results

❑ **Functional Module Detection in Tissue-specific Gene Co-Expression Networks**



Comparison of FDR adjusted p-values of detected clusters

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Experimental Results

❑ **Functional Module Detection in Tissue-specific Gene Co-Expression Networks**

Comparison of number of detected significant clusters

Method	**# significant clusters**	***p*-values**
SNMF	116	4.64e^{-5}
Spectral clustering	119	6.66e^{-3}
Markov clustering	70	6.45e^{-17}
ClusterOne	89	1.43e^{-10}
NoNClus’	121	4.87e^{-2}
NoNClus	**130**	1

Conclusion

- □ **A novel multi-network clustering problem**
  - ➢ Multi-network with multi-underlying clustering structures
- □ **A new clustering framework based on new network model**
  - ➢ NoNClus on a Network of Networks (NoN)
- □ **Comprehensive experiments**
  - ➢ Results on both synthetic and real datasets demonstrate the effectiveness of NoNClus

Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang.
Flexible and Robust Multi-Network Clustering. In KDD, 2015.

Thank you!

Questions?