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

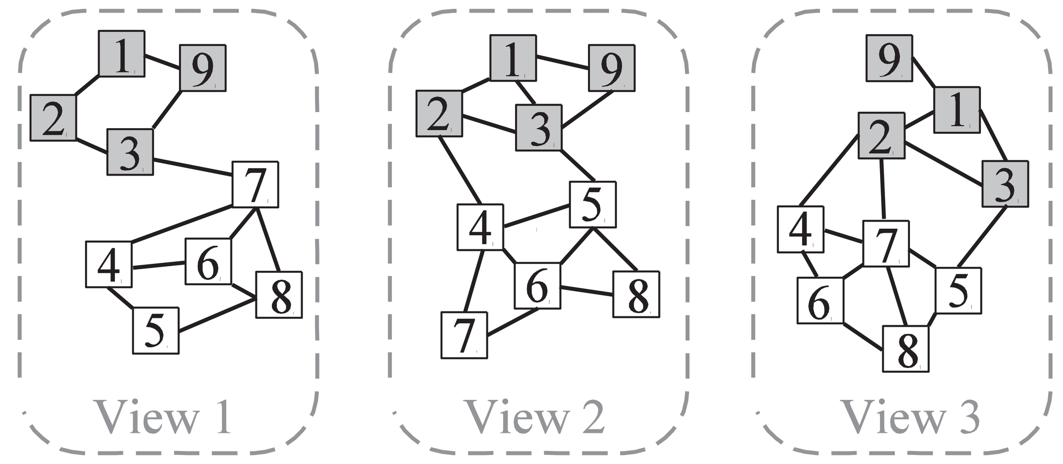
CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

Flexible and Robust Multi-Network Clustering

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

Multi-Network Clustering

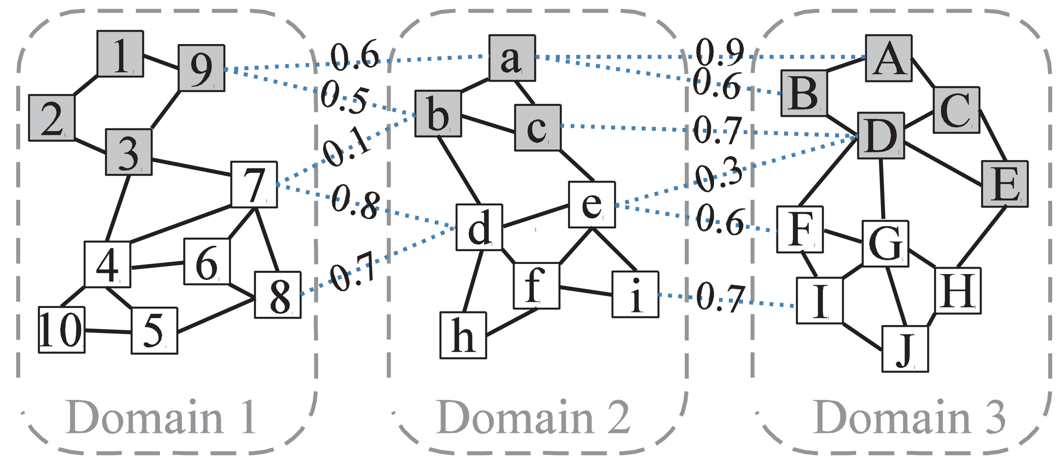
Multi-view Networks



Properties (of most works)

- All views have the *same* size
- *One-to-one* mapping across views
- *Full mapping* between nodes across views

Multi-domain Networks



Properties

- Domains can have *different* sizes
- *Many-to-many* mapping across domains
- *Partial mapping* across domains
- The mappings have weights

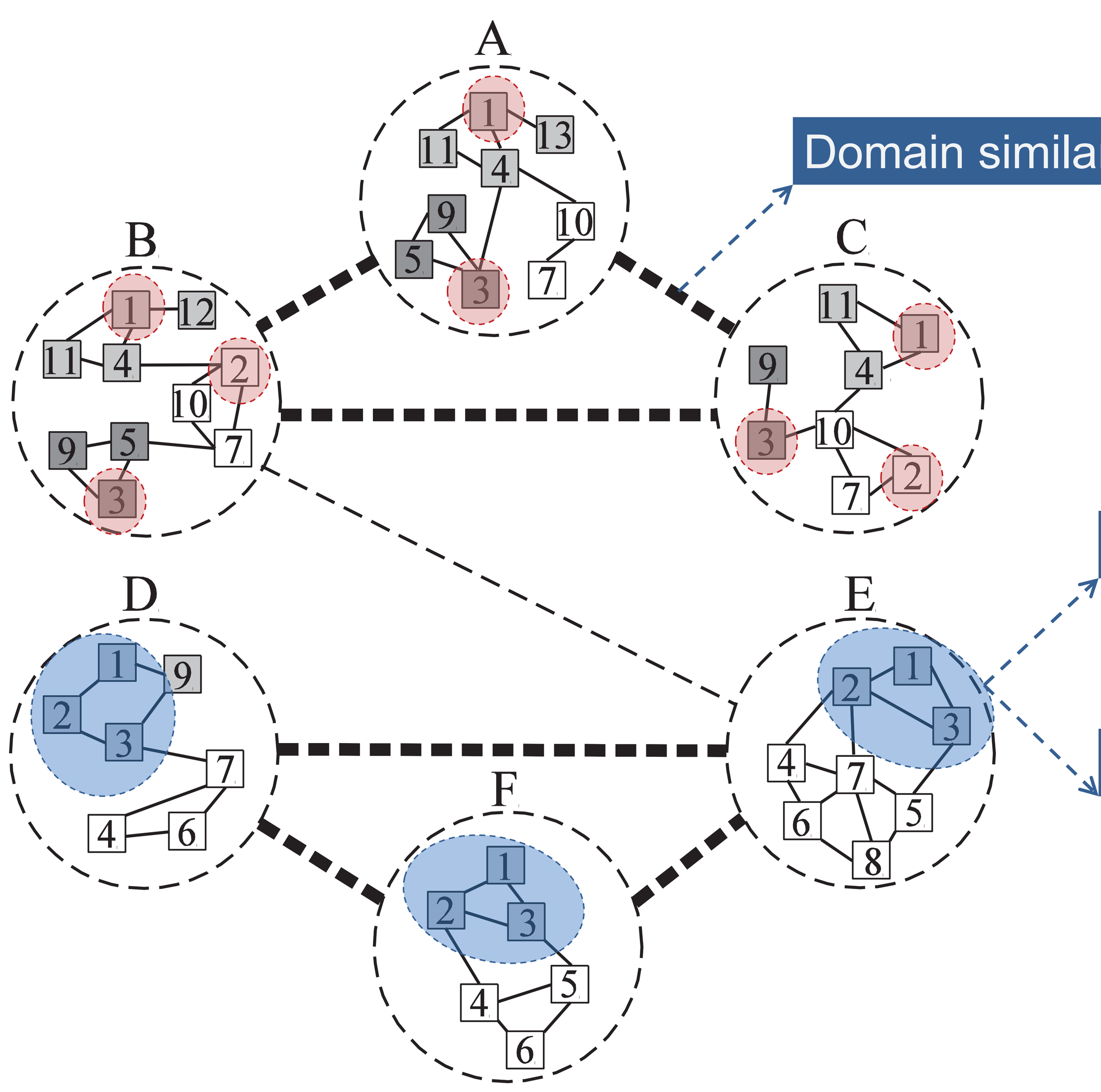
Key assumptions

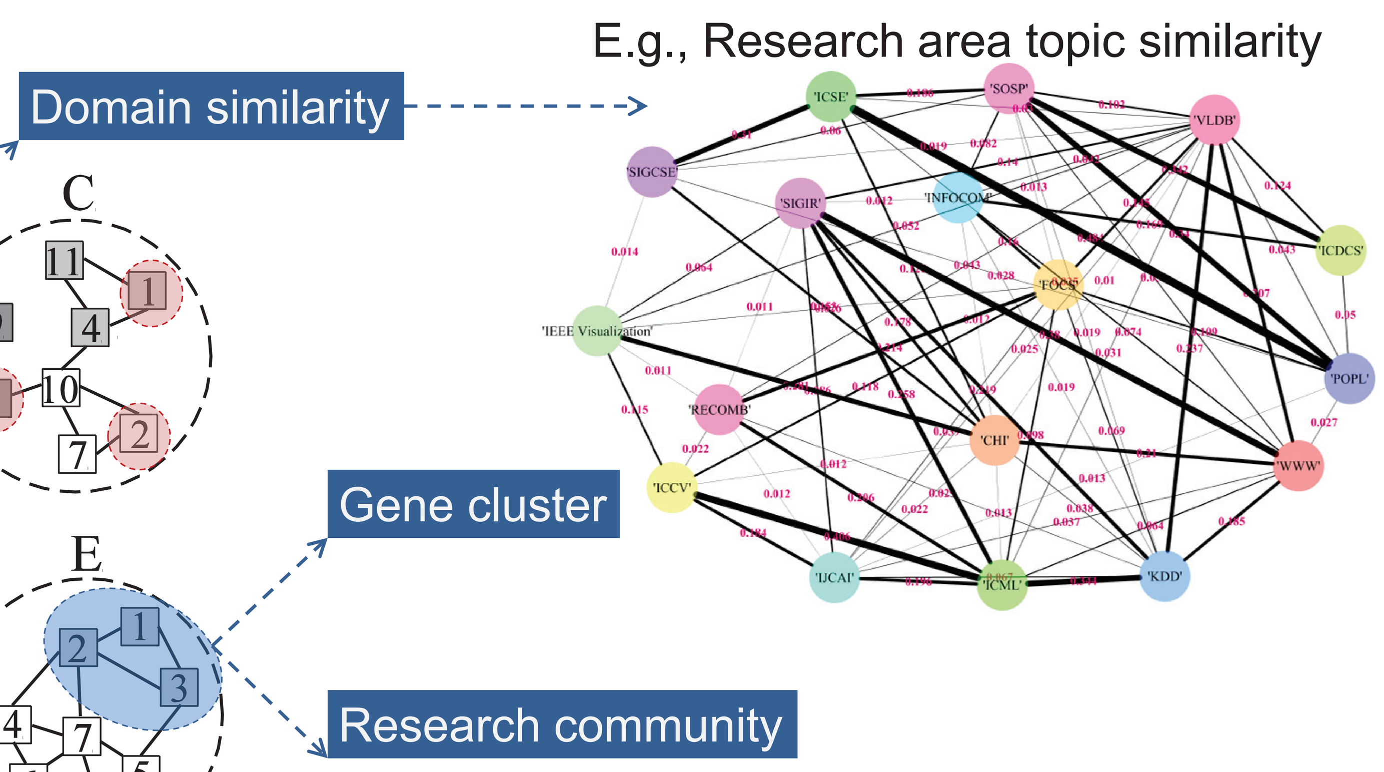
- Different views/domains share the same underlying clustering structure
- Methods are designed to identify consistent clustering structure across all views/domains

This basic assumption may not hold in some emerging applications.

Motivation

Network of Networks (NoN)





Examples

- Gene clusters in tissue-specific gene co-expression networks
- Research collaboration groups in co-author networks of different research areas

We can not assume networks {A, B, C, D, E, F} share a common underlying clustering structure.

- **This calls for a method simultaneously clustering multiple networks with multiple underlying clustering structures.**

Definitions

- We call the domain similarity network as the **main network** (the dashed line network).
- We call the network in each domain as the **domain-specific network** (the solid line networks).
- The adjacency matrix of the main network is $\mathbf{G}$. The adjacency matrices of the domain-specific networks are $\{\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(g)}\}$.

NoNClus

Phase I: Main Network Clustering

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

Phase II: Domain-specific Network Clustering

Individual domain-specific network clustering

$$\text{minimize} \quad J_A = \left\|\mathbf{A}^{(i)} - \mathbf{U}^{(i)}(\mathbf{U}^{(i)})^T\right\|_F^2 \qquad s.t. \quad \mathbf{U}^{(i)} \geq 0$$

Main cluster guided regularization

$$J_R = \underbrace{h_{ij}}_{\text{Main cluster membership}} \underbrace{\left\|\underbrace{(\mathbf{D}^{(ij)}\mathbf{U}^{(i)})(\mathbf{D}^{(ij)}\mathbf{U}^{(i)})^T}_{\text{Inner product similarity}} - \underbrace{(\mathbf{O}^{(ij)}\mathbf{V}^{(j)})(\mathbf{O}^{(ij)}\mathbf{V}^{(j)})^T}_{\text{Inner product similarity}}\right\|_F^2}_{\text{Clustering Inconsistency}}$$

$\mathbf{D}^{(ij)}$, $\mathbf{O}^{(ij)}$ are Mapping matrices such that the same rows of $\mathbf{D}^{(ij)}\mathbf{U}^{(i)}$ and $\mathbf{O}^{(ij)}\mathbf{V}^{(j)}$ represent the same instances where we allow different domains to have different sizes.

The unified objective function

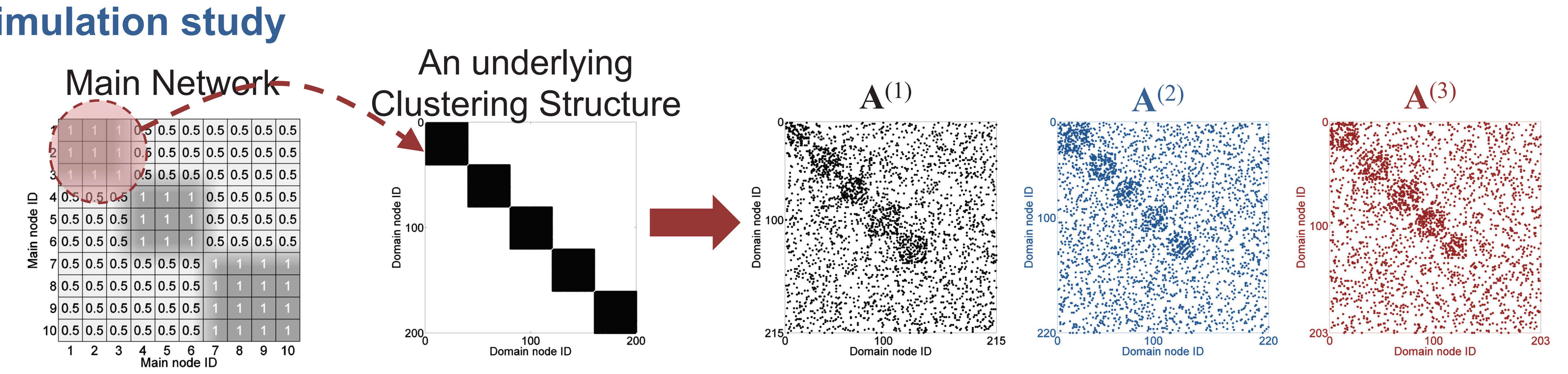
$$\min_{\substack{\mathbf{U}^{(i)} \geq 0, (i=1,\ldots,g) \\ \mathbf{V}^{(j)} \geq 0, (j=1,\ldots,k)}} \quad J_D = \sum_{i=1}^{g} J_A + a \sum_{i=1}^{g} \sum_{j=1}^{k} J_R$$

This joint optimization problem can be solved by an alternating minimization approach where $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(j)}$ are alternately solved by multiplicative updating rules with convergence guarantee.

Experiments

Effectiveness Evaluation

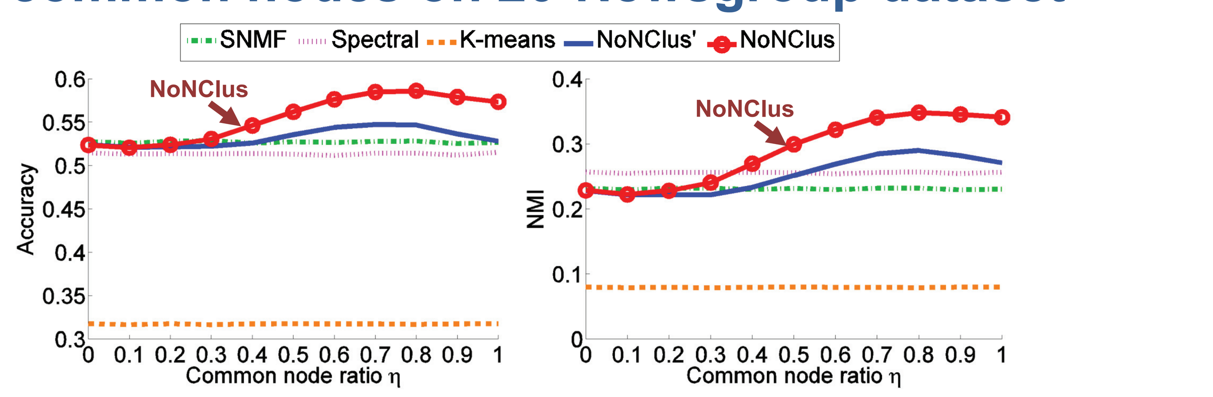
Simulation study



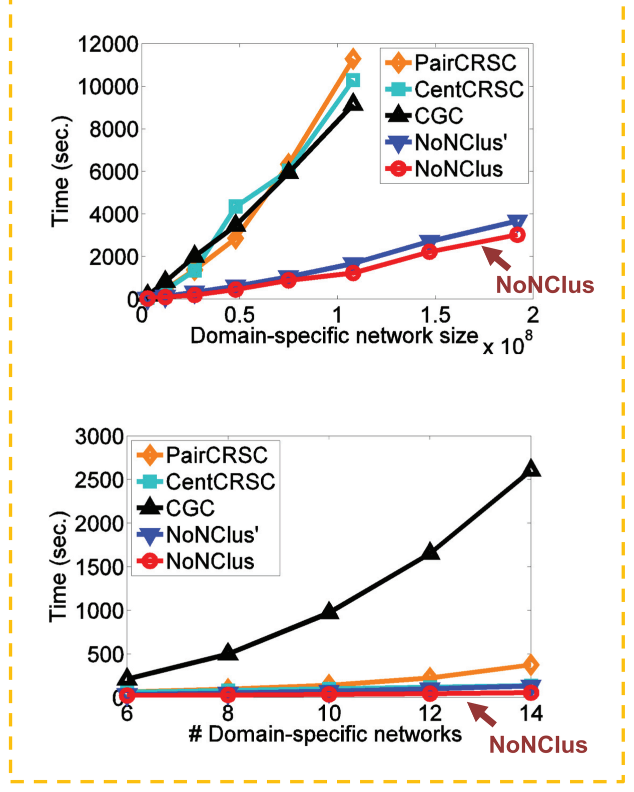
Accuracy of different methods on synthetic datasets

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**

Effects of common nodes on 20-Newsgroup dataset



Scalability

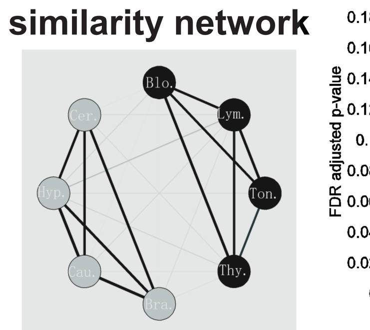


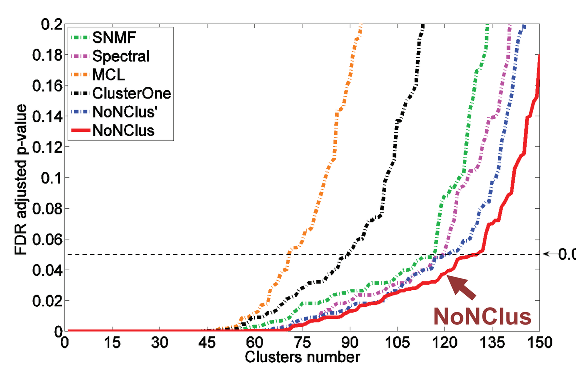
A Case Study of Tissue-specific Gene Co-expression Networks

Tissue-specific gene co-expression networks

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**

Tissue-tissue similarity network





Comparison of number of detected significant clusters

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