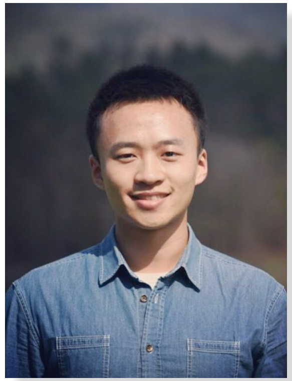
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# User-Guided Clustering in Heterogeneous Information Networks via Motif-Based Comprehensive Transcription


Yu Shi*, Xinwei He*, Naijing Zhang*, Carl Yang, Jiawei Han

University of Illinois at Urbana-Champaign (UIUC)

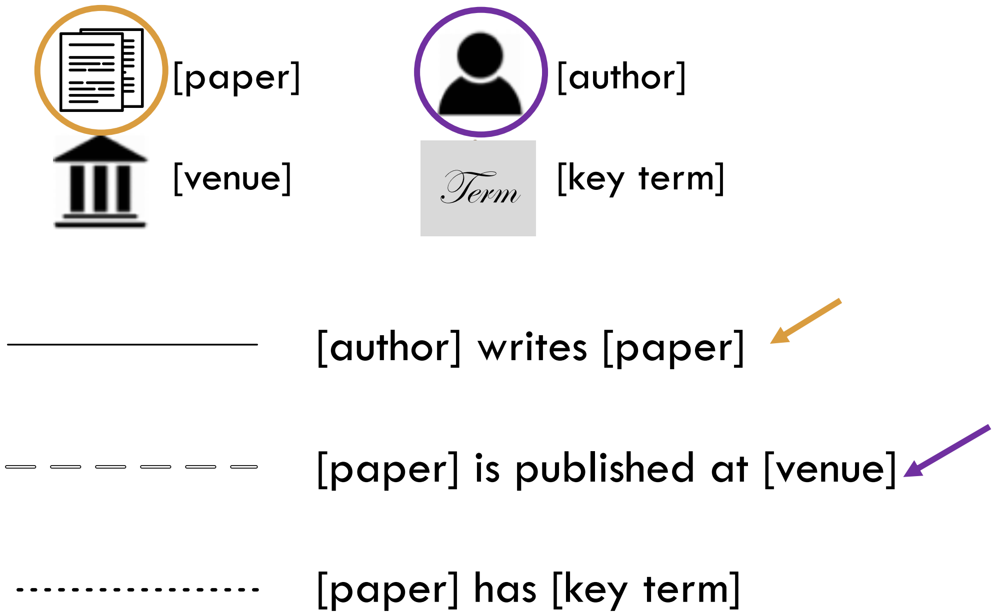



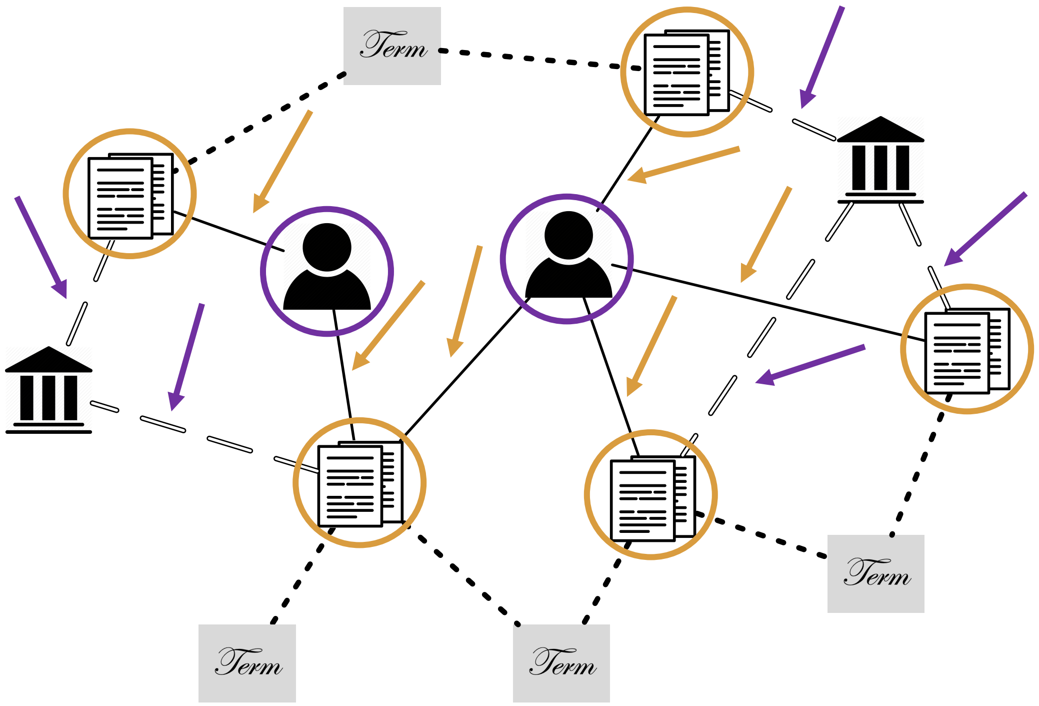


*These authors contributed equally to this work.

In real world applications, objects of different types can have different relations, which form heterogeneous information networks (**HINs**).

- **Typed** nodes: objects.
- **Typed** edges: relations.

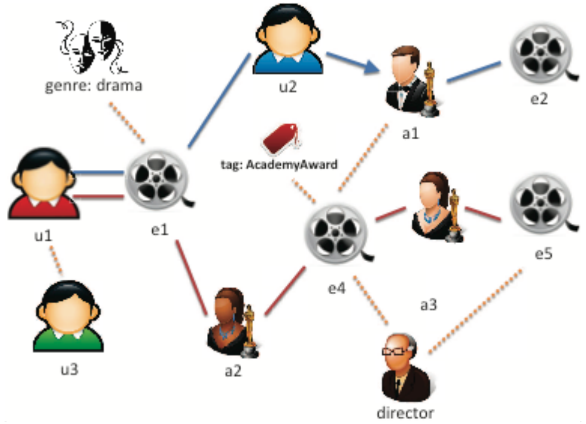


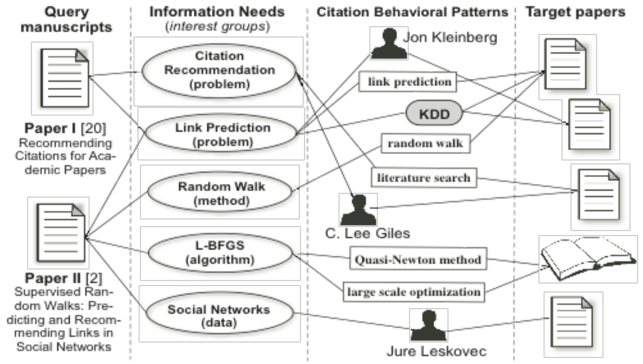

A toy bibliographic network

# Heterogeneous information networks (HINs) are ubiquitous.

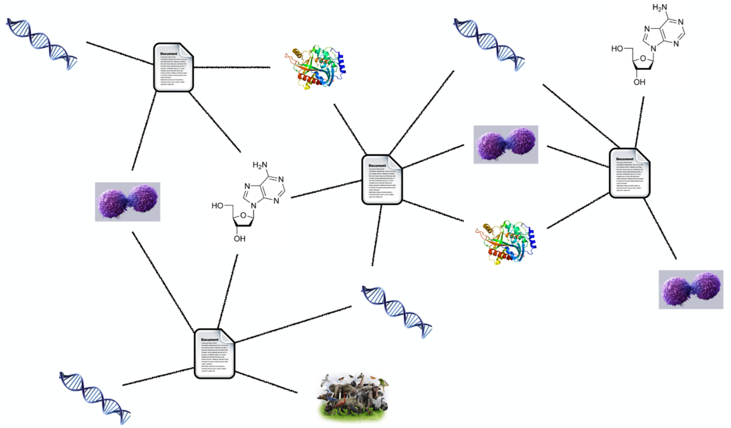
## Each embodies **rich semantics** due to the type information.




Movie Reviewing Network

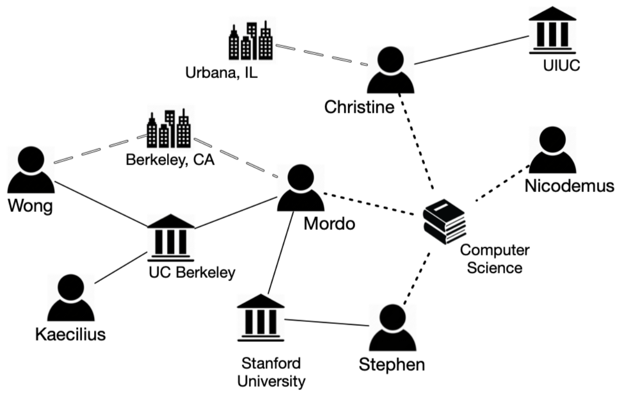



Bibliographic Network

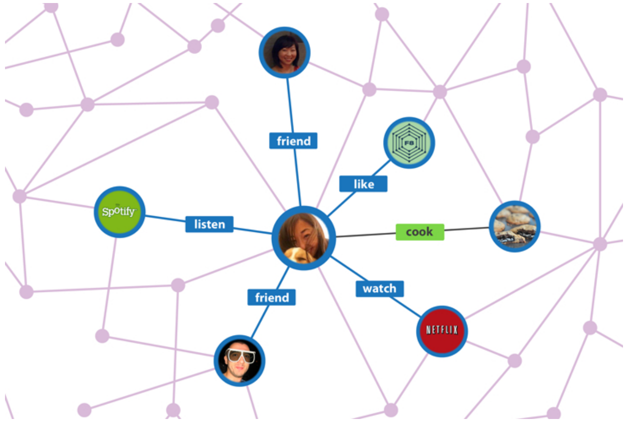


Biomedical Network

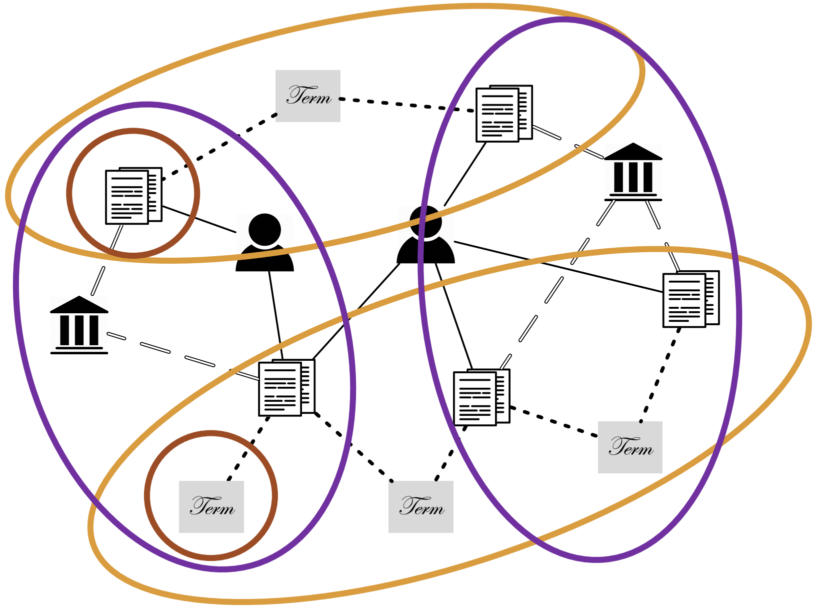

Economic Graph




Social Network




Facebook Open Graph




- Clustering is a fundamental task in network mining.
- With the rich semantics in HINs, there can be **multiple reasonable clustering results.**
- Result 1 – cluster by similar research topics (key terms).
- Result 2 – cluster by similar publishing venues.

User-guided clustering – User provides guidance on a very small portion of nodes as seeds.

- E.g., user guidance: the two nodes in the **Brown circles** are in the same cluster.

… likely the user want nodes related to **similar venues** to be clustered together.

… likely result 2 is preferable.

# Challenge in User-Guided Clustering

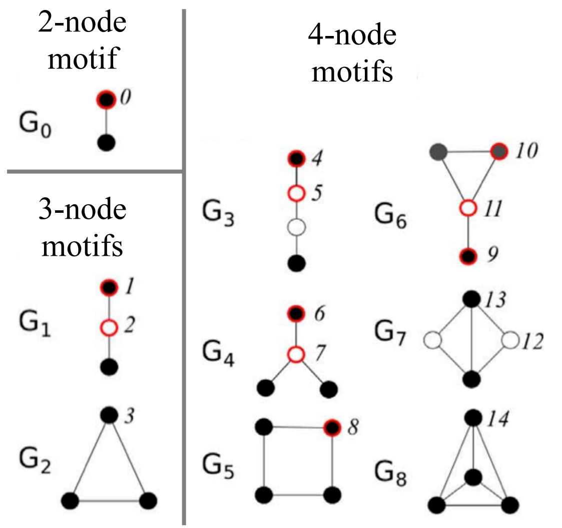
Since guidance is provided only on a very small portion of nodes…

- It is often necessary to exploit signals encoded in the **rich semantics** of heterogeneous information networks (HINs).
- We resort to network **motifs** to expose **higher-order interaction** signals.

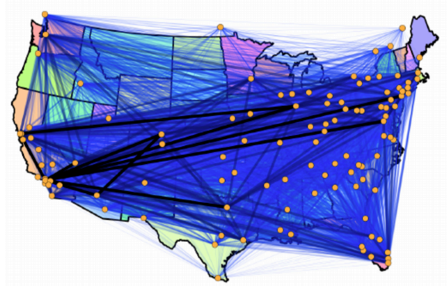
# Background – Network Motifs

Simpler case: in **homogeneous** networks (not typed)

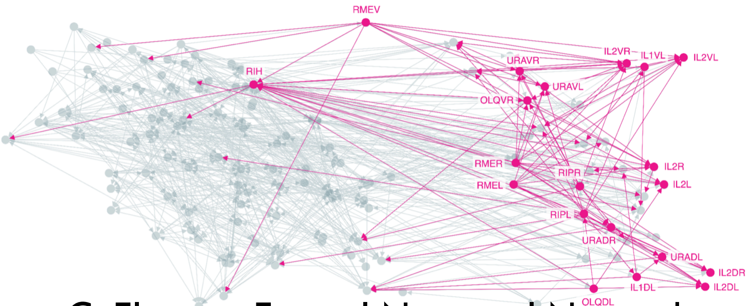
- Motifs are higher-order structures beyond nodes, edges, and paths (random walks).
- Motifs reveal **higher order interaction** in complex networks such as transportation networks and neuronal networks.
- Bringing performance boost in data mining/machine learning tasks such as **clustering, link prediction, ranking,** and **representation learning**.




Examples of Motifs



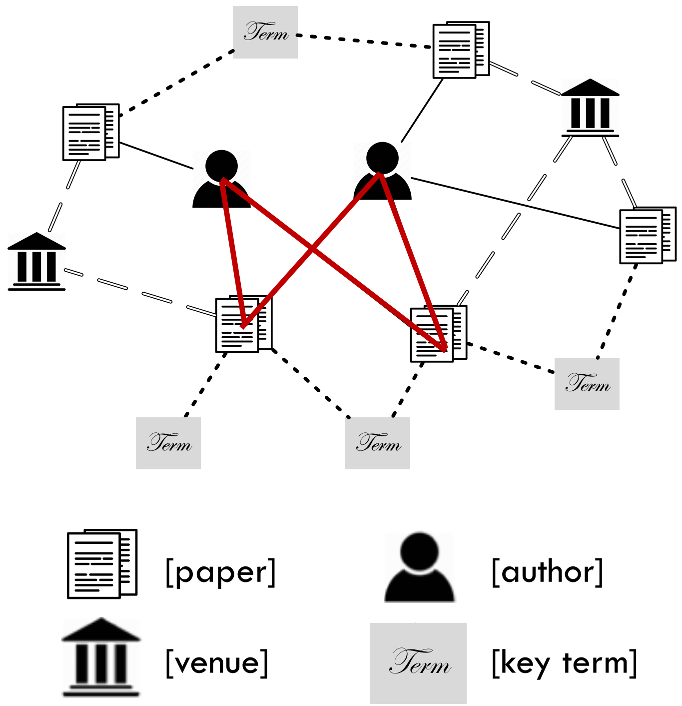
U.S. Airport Network




C. Elegans Frontal Neuronal Network

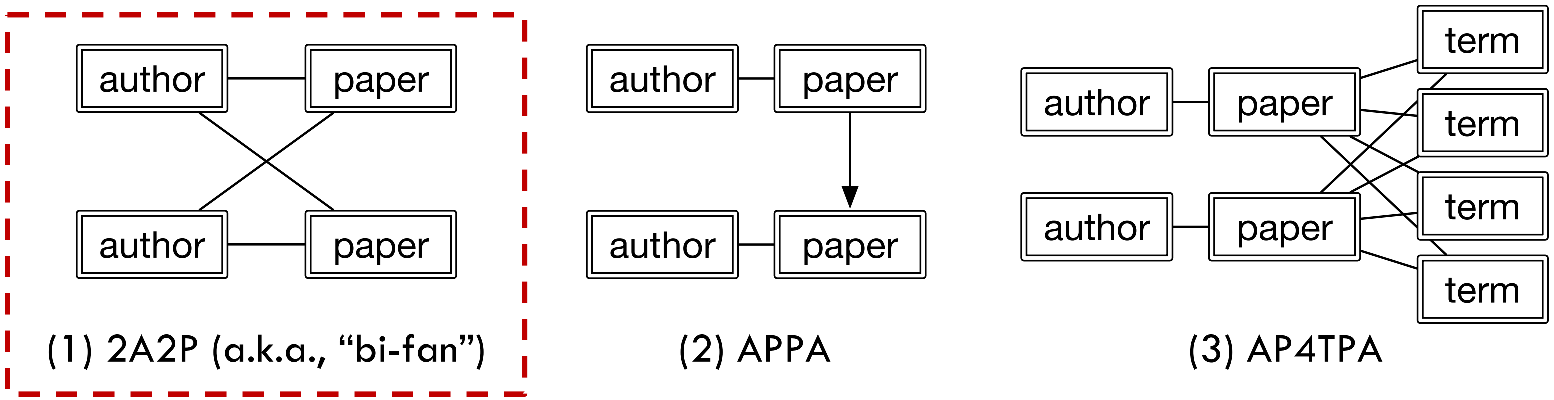
Figures adapted from Yaveroğlu, Ömer Nebil, et al. "Revealing the hidden language of complex networks." *Scientific reports* 4 (2014) and Benson, Austin R., David F. Gleich, and Jure Leskovec. "Higher-order organization of complex networks." Science 353.6295 (2016).

# Background – Motifs in HINs




In the context of **heterogeneous** network (**HINs**)…

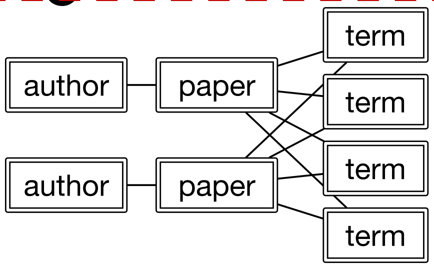
- Motifs additionally has type constraints.
- A.k.a. meta-graphs [1] or meta-structures [2].



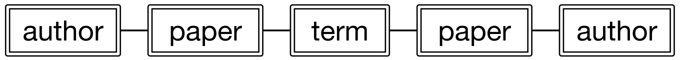

Examples of Motifs in HINs

[1] Zhao, Huan, et al. "Meta-graph based recommendation fusion over heterogeneous information networks." In KDD, 2017.
[2] Huang, Zhipeng, et al. "Meta structure: Computing relevance in large heterogeneous information networks." In KDD, 2016.

# Why Would Motifs in HINs Provide Informative Signals?



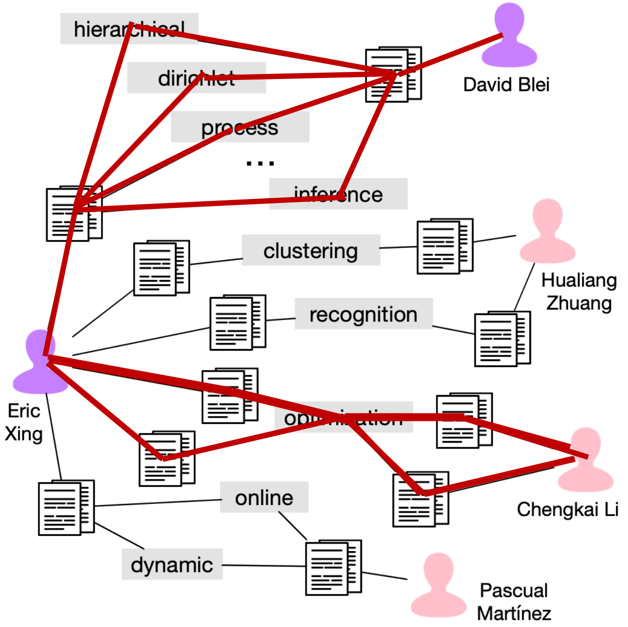

A motif: AP4TPA




A meta-path (a simpler motif): APTPA

Can we instead use simpler network structures?

- According to the ground truth labels: only Eric Xing and David Blei received their Ph.D. from the same research group.
- Even though every one is well connected with Eric Xing.
- Simpler structures such as paths (known as meta-paths in HINs) are not very discriminative.
- If the user's intention is to cluster by Ph.D. research group, the information from motif would be worth exploiting.




Part of the DBLP network

# How to Leverage Motifs?

Option 1: **Collapse** motifs to pairwise relation.

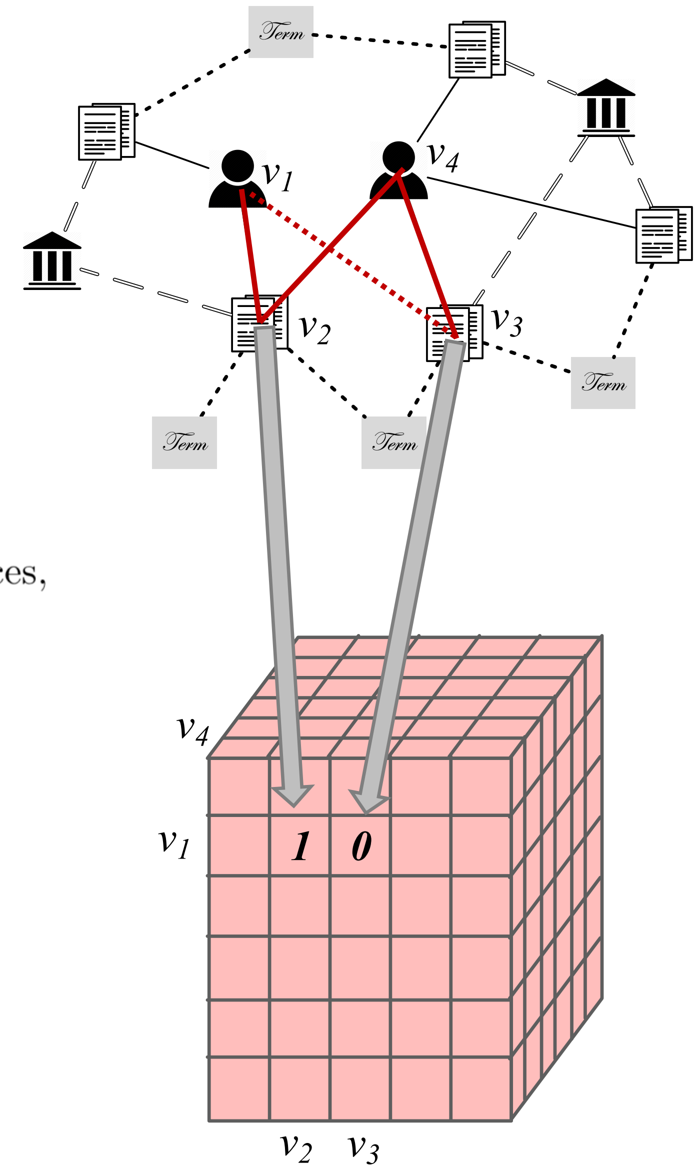
$$f : (G = (\mathcal{V}, \mathcal{E})) \times \mathcal{M} \to \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$$

- … so that traditional clustering methods ( $\mathcal{V} \to \mathcal{C}$ ) can be applied.
- Example:

$$f(v_1, v_2; m) = \begin{cases} \# \text{ motif instances}, & v_1 \text{ and } v_2 \text{ co-exist in some type-}m \text{ motif instances}, \\ 0, & \text{otherwise}. \end{cases}$$

- Will incur some information loss.

# How to Leverage Motifs?



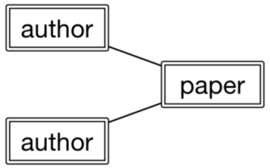

Option 2: Comprehensively transcribe signals revealed by motifs.

- A $k$-node motif instance $\rightarrow$ a $k$-tuple of nodes

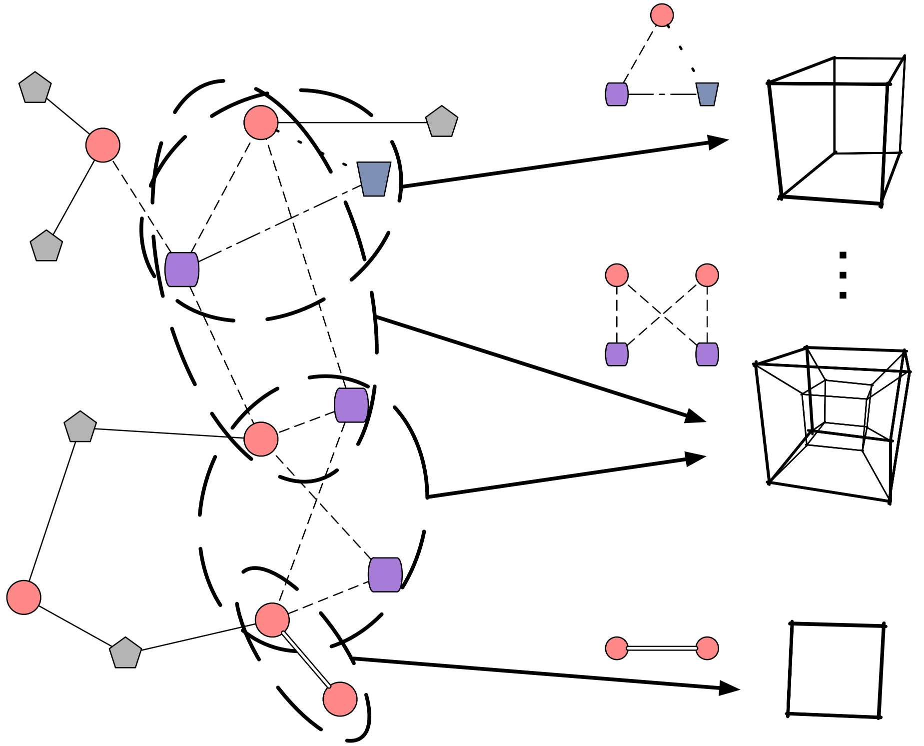
  $\rightarrow$ an element in a $k$-th-order tensor.

$$\mathcal{X}^{(m)}_{v_1,v_2,\ldots,v_k} = \begin{cases} \#\text{ motif instances}, & v_1, v_2, \ldots, v_k \text{ constitutes some type-}m\text{ motif instances}, \\ 0, & \text{otherwise.} \end{cases}$$

- Not causing information loss.
- Example: Motif APA




- We resort to this option 2 to preserve as much information as possible.



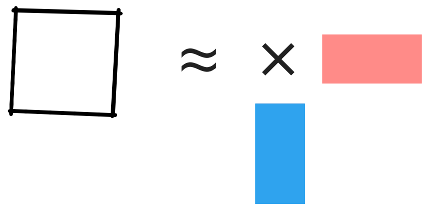
- As such, we can use motifs to transcribe the information in an HIN to a series of tensors.
- How to use a series of tenors in user-guided clustering?

# Revisit on Clustering by Non-Negative Matrix Factorization (NMF)

Given a matrix representing the pairwise relations between nodes

- … factorize the matrix into two non-negative matrices.

$$\min_{\mathbf{V}_1, \mathbf{V}_2 \geq 0} \left\| \mathbf{M} - \mathbf{V}_1^\top \mathbf{V}_2 \right\|_F^2$$



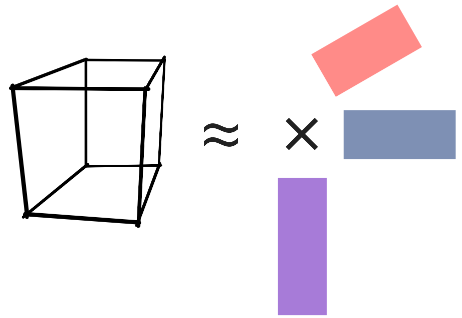
- Each row (or column) of the resulting matrix yields the clustering membership for the corresponding node.

# Single-Motif–Based Clustering in HINs

Similarly, given a $k$-th-order tensor

- … factorize the tensor into (the mode product of) $k$ non-negative matrices.

$$\min_{\mathbf{V}_1,\mathbf{V}_2\geq 0} \left\| \mathcal{X} - \mathcal{I} \times_{i=1}^{N} \mathbf{V}_i \right\|_F^2 + \underbrace{\lambda \sum_{i=1}^{N} \|\mathbf{V}_i\|_1}_{\text{Regularization}}$$



- This is essentially **CP decomposition** with additional non-negative constraints and $l_1$ regularization.

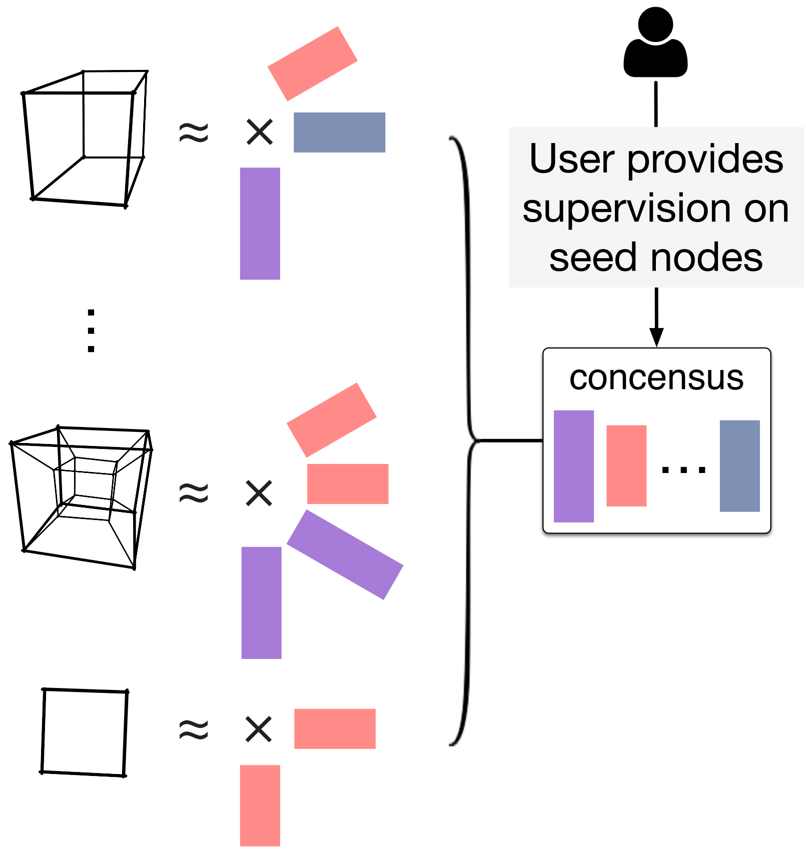
# Proposed Model "MoCHIN" for Motif-Based Clustering in HINs

MoCHIN – short for **Mo**tif-based **C**lustering in **HIN**s

$$\min_{\{\mathbf{V}_i^{(m)}\geq 0\},\boldsymbol{\mu}\in\Delta} \sum_{m\in\mathcal{M}} \left\|\mathcal{X}^{(m)} - \mathcal{I}^{(m)} \times_{i=1}^{o(m)} \mathbf{V}_i^{(m)}\right\|_F^2 + \lambda \sum_{m\in\mathcal{M}}\sum_{i=1}^{o(m)} \left\|\mathbf{V}_i^{(m)}\right\|_1$$

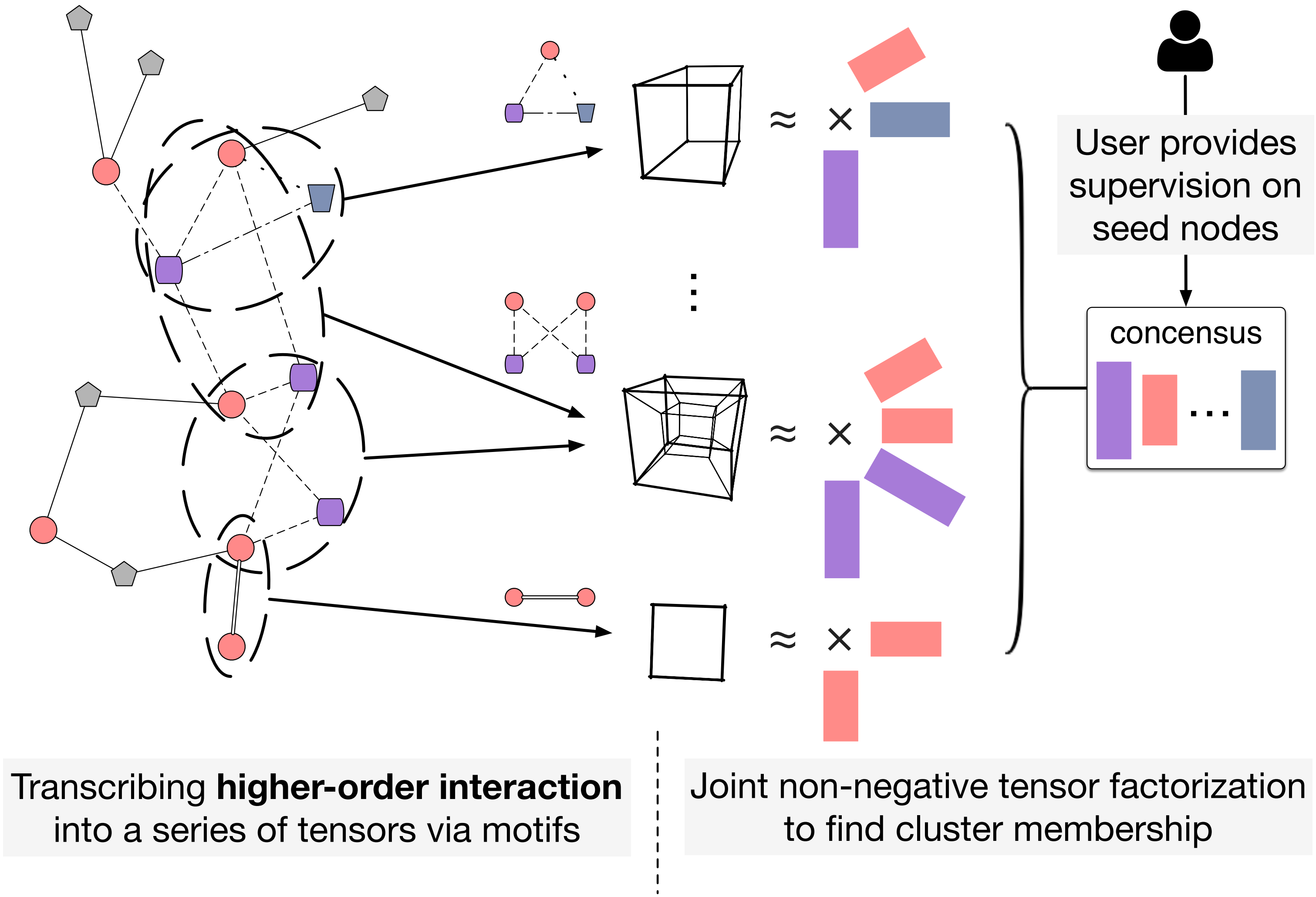
$$\underbrace{+\,\theta \sum_{m\in\mathcal{M}}\sum_{i=1}^{o(m)} \left\|\mathbf{V}_i^{(m)} - \mathbf{V}^*_{\varphi(m,i)}\right\|_F^2}_{\text{Consensus}} \underbrace{+\,\rho\sum_{t\in\mathcal{T}} \left\|\mathbf{M}^{(t)}\circ\mathbf{V}_t^*\right\|_F^2}_{\text{User guidance}},$$

where $\mathbf{V}_t^* := \underbrace{\sum_{\varphi(m,i)=t}}_{\text{All matrices for the same node type}} \frac{\mu_m \mathbf{V}_i^{(m)}}{\sum_{i'=1}^{o(m)} \mathbb{1}_{[\varphi(m,i')=\varphi(m,i)]}}$




Joint non-negative tensor factorization to find cluster membership

# Proposed Model "MoCHIN" for Motif-Based Clustering in HINs



# Inference Algorithm and Speed-Up Tricks

**Update rule:**

$$\mathbf{V}_k^{(l)} \leftarrow \mathbf{V}_k^{(l)} \circ \left[ \frac{\mathcal{X}_{(k)}^{(l)}[\otimes_{i=1}^{o(l)\backslash k}\mathbf{V}_i^{(l)}]\mathcal{I}_{(k)}^{(l)\top} + \theta(1-\eta_k^l)(\mathbf{V}_{\varphi(l,k)}^* - \eta_k^l\mathbf{V}_k^{(l)})}{\mathbf{V}_k^{(l)}\mathcal{I}_{(k)}^{(l)}[\otimes_{i=1}^{o(l)\backslash k}\mathbf{V}_i^{(l)}]^\top[\otimes_{i=1}^{o(l)\backslash k}\mathbf{V}_i^{(l)}]\mathcal{I}_{(k)}^{(l)\top} + \rho\eta_k^l\mathbf{M}^{\varphi(l,k)} \circ \mathbf{V}_{\varphi(l,k)}^*} \right.$$
$$\left. \frac{+\theta\eta_k^l\sum_{\varphi(m,i)=\varphi(l,k)}^{(m,i)\neq(l,k)}[\mathbf{V}_i^{(m)} - \mathbf{V}_{\varphi(l,k)}^* + \eta_k^l\mathbf{V}_k^{(l)}]^+}{+\theta\eta_k^l\sum_{\varphi(m,i)=\varphi(l,k)}^{(m,i)\neq(l,k)}([\mathbf{V}_i^{(m)} - \mathbf{V}_{\varphi(l,k)}^* + \eta_k^l\mathbf{V}_k^{(l)}]^- + \eta_k^l\mathbf{V}_k^{(l)}) + \theta(1-\eta_k^l)^2\mathbf{V}_k^{(l)} + \lambda} \right]^{\frac{1}{2}}$$

**Challenge:** Tensor size grows exponentially with the number of nodes in a motif.

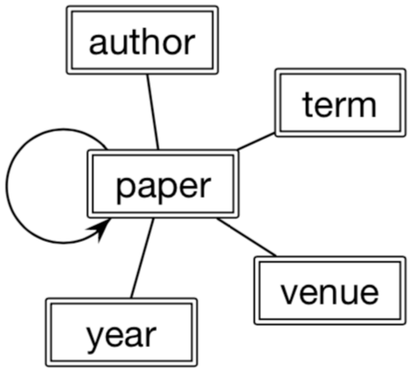
- Multiple speed-up tricks leveraging the **sparsity** of the motif instances and the **compositionality** of dense tensors in the model.
- … so that the complexity is bounded by **the number of motif instances** instead of the tensor size

$$\left\|\mathcal{X}^{(m)} - \mathcal{I}^{(m)} \times_{i=1}^{o(m)} \mathbf{V}_i^{(m)}\right\|_F^2 = \left\|\mathcal{X}^{(m)}\right\|_F^2 - 2\left\|\mathcal{X}^{(m)} \circ \mathcal{I}^{(m)} \times_{i=1}^{o(m)} \mathbf{V}_i^{(m)}\right\|_1 + \left\|\mathcal{I}^{(m)} \times_{i=1}^{o(m)} \mathbf{V}_i^{(m)}\right\|_F^2$$
$$= \left\|\mathcal{X}^{(m)}\right\|_F^2 - 2\sum_{j_1,\ldots,j_{o(m)}} (\mathcal{X}^{(m)})_{j_1,\ldots,j_{o(m)}} \sum_{c=1}^{C}\prod_{i=1}^{o(m)}(\mathbf{V}_i^{(m)})_{j_i,c} + \sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\prod_{i=1}^{o(m)}(\mathbf{V}_i^{(m)})_{:,c_1}^\top(\mathbf{V}_i^{(m)})_{:,c_2}.$$

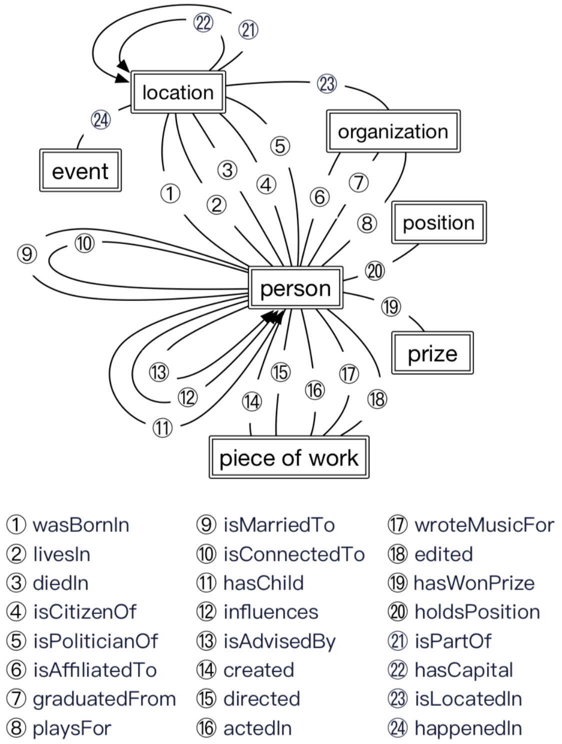
# Experiments

## Datasets and evaluation tasks

- DBLP: a bibliographic network in the computer science domain.
  - **Task 1**: cluster 7,165 authors into 14 **research areas.** 1% authors as seeds; 16,100 nodes and 30,239 edges.
  - **Task 2**: cluster 250 authors into 5 **research groups.** 5% authors as seeds; 19,500 nodes and 108,500 edges.
  - **Motifs used**: AP4TPA, APPA, and all edge types (2-node motifs).
- YAGO: a knowledge graph.
  - **Task**: cluster 11,368 people to 10 countries. 1% people as seeds; 17,109 nodes and 70,251 edges.
  - **Motifs used**: $P^6O^{23}L$, $P^7O^{23}L$, $P^8O^{23}L$, $^2P^2W$, $^3PW$, and all edge types.




Schema of DBLP




Schema of DBLP

# Experiments

## Baselines

No motifs

- KNN: k-nearest neighbors.
- GNetMine [3]: A graph-based regularization framework for the transudative classification problem in HINs. Only leverages edge-level information.
- PathSelClus [4]: A probabilistic graphical model for HIN clustering by integrating meta-path selection with user-guidance.

Collapse motifs into pairwise relation

- KNN+Motifs: Construct a new network for each motif with an edge for two nodes matched to a motif instance, apply KNN on new network, and linear combine results.
- TGS [5]: A motif-based spectral clustering algorithm for HINs.

[3] Ji, Ming, et al. “Graph regularized transductive classification on heterogeneous information networks.” In ECMLPKDD, 2010.
[4] Sun, Yizhou, et al. “Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks.” In KDD, 2012.
[5] Carranza, Aldo G., et al. "Higher-order Spectral Clustering for Heterogeneous Graphs." arXiv:1810.02959 (2018).

# Experiments

No motifs

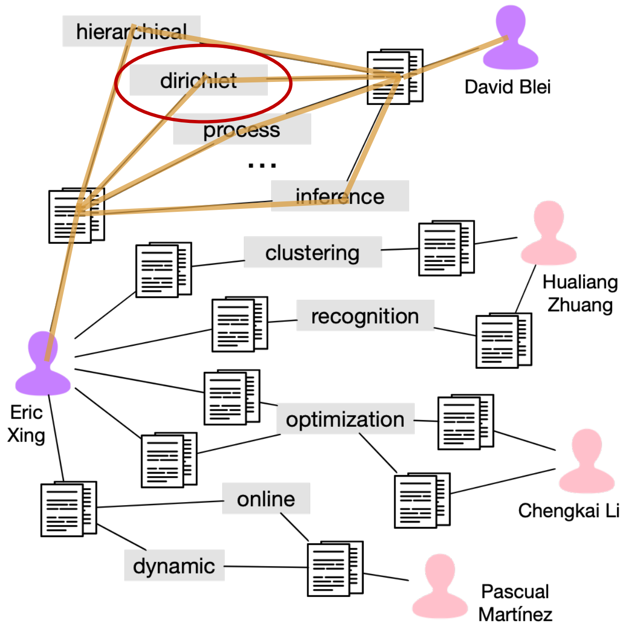
Collapse motifs into pairwise relation

| Task | DBLP-group | | | DBLP-area | | | YAGO | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | Acc./Micro-F1 | Macro-F1 | NMI | Acc./Micro-F1 | Macro-F1 | NMI | Acc./Micro-F1 | Macro-F1 | NMI |
| KNN | 0.4249 | 0.2566 | 0.1254 | 0.4107 | 0.4167 | 0.2537 | 0.3268 | 0.0921 | 0.0810 |
| GNetMine [3] | 0.5880 | 0.6122 | 0.3325 | 0.4847 | 0.4881 | 0.3469 | 0.3832 | 0.2879 | 0.1772 |
| PathSelClus [4] | 0.5622 | 0.5535 | 0.3246 | 0.4361 | 0.4520 | 0.3967 | 0.3856 | 0.3405 | 0.2864 |
| KNN+Motifs | 0.4549 | 0.2769 | 0.1527 | 0.4811 | 0.4905 | 0.3296 | 0.3951 | 0.1885 | 0.1660 |
| TGS [5] | 0.6609 | 0.6513 | 0.3958 | 0.4391 | 0.4365 | 0.2790 | 0.6058 | 0.3564 | 0.4406 |
| MoCHIN | **0.7382** | **0.7387** | **0.5797** | **0.5318** | **0.5464** | **0.4396** | **0.6134** | **0.5563** | **0.4607** |

- MoCHIN uniformly outperformed all five baselines in all three tasks.
- MoCHIN prevailed by **comprehensively transcribing signals to tensors.**
  - KNN+Motifs and TGS also leverage signals from motifs.
  - TGS can generally outperform other baselines, but is still worse than MoCHIN.

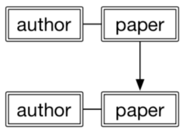
[3] Ji, Ming, et al. “Graph regularized transductive classification on heterogeneous information networks.” In ECMLPKDD, 2010.
[4] Sun, Yizhou, et al. “Integrating Meta-Path Selection with User-Guided Object Clustering in Heterogeneous Information Networks.” In KDD, 2012.
[5] Carranza, Aldo G., et al. "Higher-order Spectral Clustering for Heterogeneous Graphs." arXiv:1810.02959 (2018).

# Impact of Candidate Motif Choice

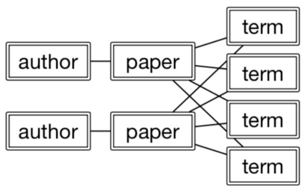



In the DBLP-group task, we optionally remove APPA and/or AP4TPA.

- AP4TPA is crucial for clustering <u>Eric Xing</u> correctly.
- Each node (e.g., "dirichlet") contributes to the semantic meaning of an motif instance – **comprehensive transcription** via motifs is **helpful**.

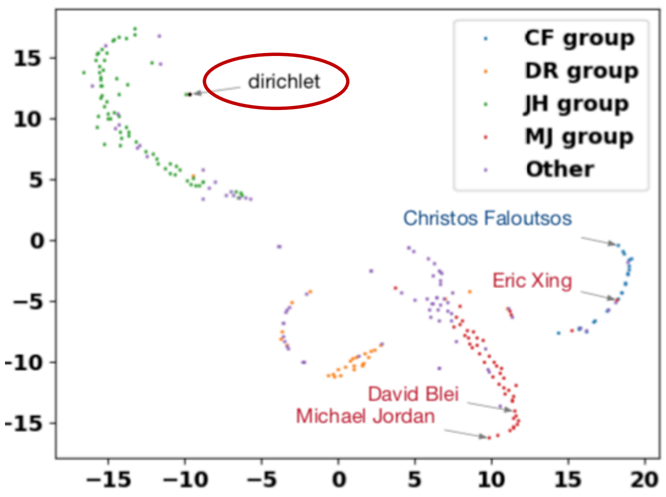



The APPA motif.

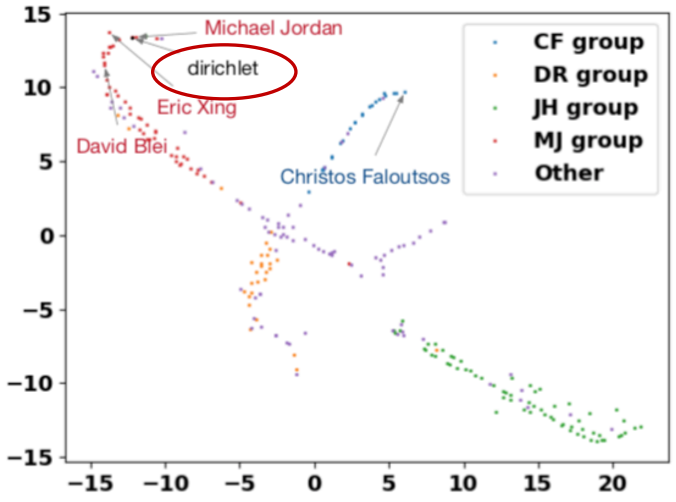



The AP4TPA motif.

| Metric | Acc./Micro-F1 | Macro-F1 | NMI | Result for Eric Xing |
|---|---|---|---|---|
| W/o both | 0.6567 | 0.6411 | 0.5157 | ✗ |
| W/ APPA | 0.7039 | 0.7062 | 0.5166 | ✗ |
| W/ AP4TPA | 0.6781 | 0.6589 | 0.5502 | ✓ |
| Full model | **0.7382** | **0.7387** | **0.5797** | ✓ |

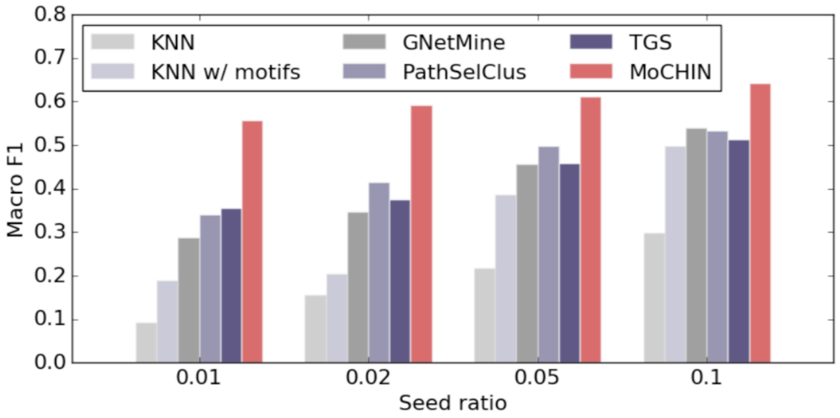



(a) W/o both.

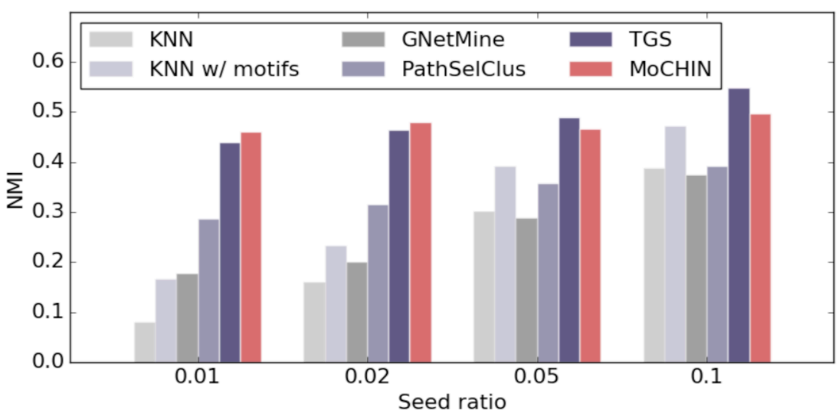



(b) Full model

# Varied Seed Ratio




(a) Macro-F1.




(b) NMI.

- For all methods, the performance increased as the seed ratio increased.
- MoCHIN outperformed most baselines, **especially when seed ratio is small.**
    - MoCHIN is particularly useful when users provide less guidance – the most common scenario for user guided clustering – because it can better exploit subtle information from limited data.

# Summary

- We identify the utility of motifs without collapsing it into pairwise interactions in user-guided clustering.
- We propose the MoCHIN model that captures higher-order interaction via motif-based comprehensive transcription and develop an inference algorithm and speed-up methods for MoCHIN.
- In experiments, we demonstrate that the proposed approach can avoid losing the rich and subtle information captured by HIN motifs.
- Code available at https://github.com/NoSegfault/MoCHIN.