Integrating Vertex-centric Clustering with Edge-centric Clustering for Meta Path Graph Analysis(KDD15)

From:Yang Zhou, Ling Liu, David Buttler

Reporter:Wenbao Li

Content

- Background
- Related works
- New challenges and Basic idea
- Model description
- Experiment

01 Background

Background

- Heterogeneous information network analysis,**especially meta path-based social network analysis** has attracted more and more attention.

Background

- What is heterogeneous information network
  - Multiple type nodes(objects).
  - Multiple type links between different type of nodes.

JUST LIKE THIS

Background

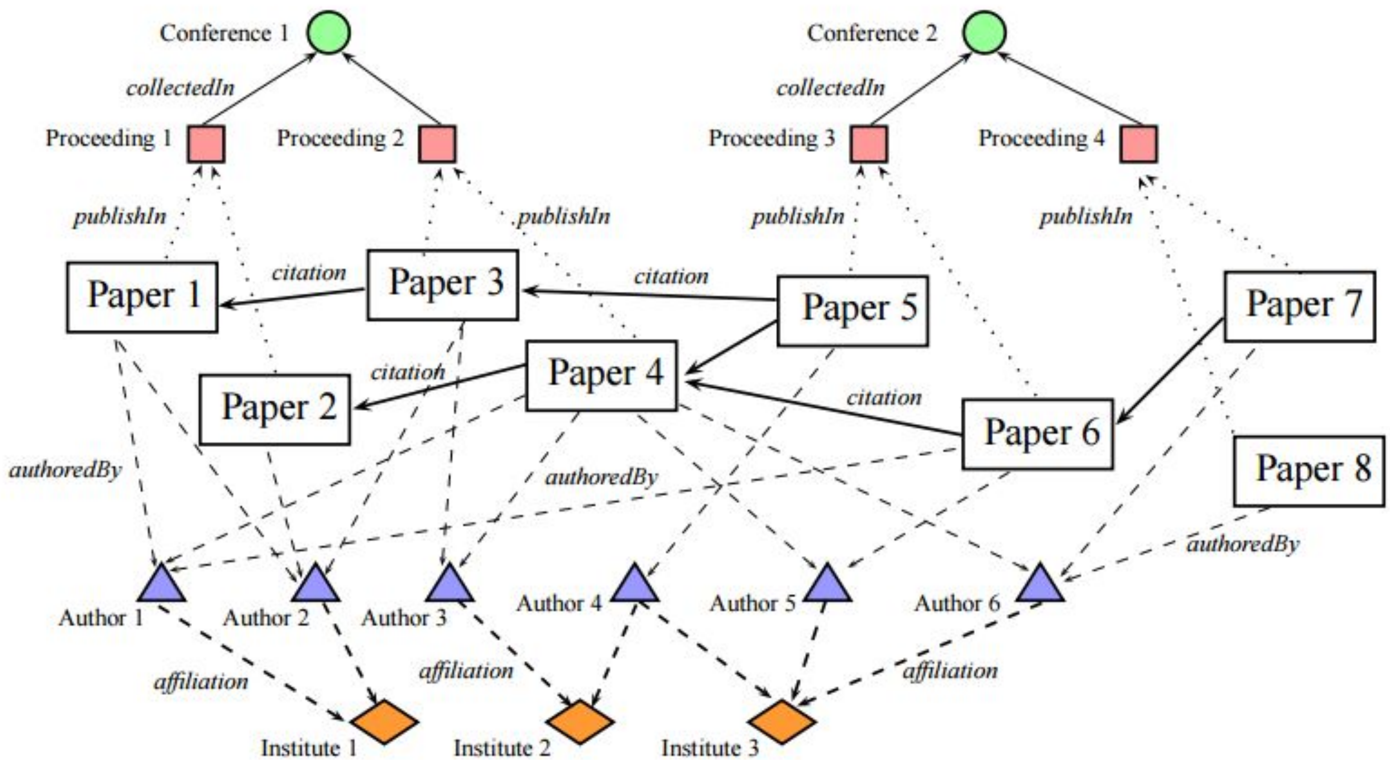


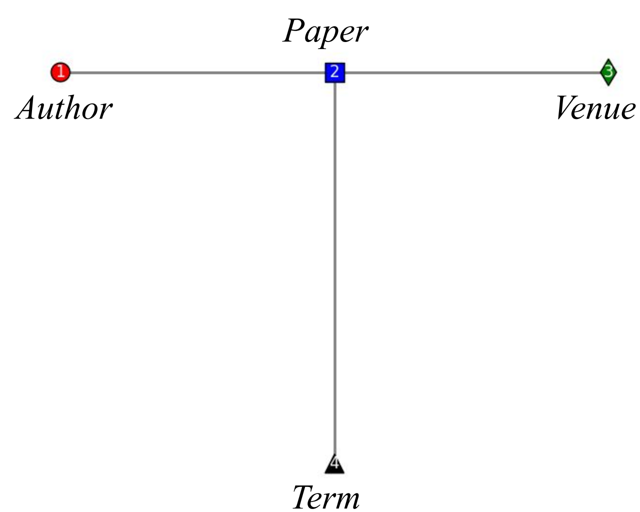
Figure 1: A Heterogeneous Information Network

Background

- What is meta path?

A — P — A
A — P — V — P — A
A — P — T — P — A
A — P — A — P — A
A — P — V — P — T — P — A
A — P — V — P — A — P — A
A — P — T — P — A — P — A
A — P — T — P — T — P — A
A — P — A — P — A — P — A
... ...

(a) Meta Paths between Authors



Background

- Utilizing meta-path to improving the quality of the following tasks.
  - Similarity search
  - Classification
  - Clustering(community detection)
  - Recommended system
  - Link prediction
  - ...

This work is focusing on the clustering task!

02 Related work

Related works

- Meta path-based
  - PathSim
    - presented a meta path-based **similarity measure** for Hete. gaph
  - User guided entity **similarity search** using meta-path selection in hete. information networks.
    - proposed a meta path-based ranking model to find entities with high similarity to a given query entity.
  - HCC
    - is a meta-path based heterogeneous collective **classification** method

Related works

- Meta path-based
  - PathSelClus
    - utilizes user guidance as seeds in some of the clusters to automatically learn the best weights for each meta-path in the **clustering**.
  - MLI
    - is a multi-network **link prediction** framework by extracting useful features from multiple meta paths.

Related works

- Graph **clustering**
- A spectral clustering approach to optimally combining numericalvectors with a modular network.
- presented a clustering method which integrates numerical vectors with modularity into a **spectral** relaxation problem.
- SCAN
- is a **structural** clustering algorithm to detect clusters, hubs and outliers in networks.

Related works

- Graph clustering
- MLR-MCL
- is a **multi-level graph** clustering algorithm using **flows** to deliver significant improvements in both quality and speed.
- TopGC
- is a fast algorithm to **probabilisticlly** search l**arge, edge weighted,directed graphs** for their best clusters in linear time.
- BAGC
- constructs a **Bayesian probabilistic model** to capture both structural and attribute aspects of graph.

Related works

- Graph clustering
- GenClus
- proposed a model-based method for clustering **hete. networks** with **different link types and different attribute types**.
- CGC
- is a **multi-domain graph** clustering model to utilize **cross-domain relationship** as co-regularizing penalty to guide the search of consensus clustering structure.
- FocusCO
- solves the problem of finding **focused clusters and outliers** in **large attributed graphs**.

03 New Challenges and Basic ideas

New Challenges——>Basic idea

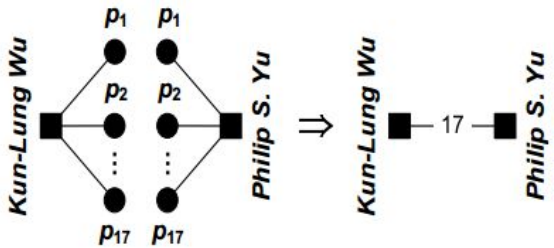
- Vertex-centric clustering *w.r.t* multiple path graphs
  - Different meta paths carry different semantics about the same type of entities.

A — P — A
A — P — V — P — A
A — P — T — P — A
A — P — A — P — A
A — P — V — P — T — P — A
A — P — V — P — A — P — A
A — P — T — P — A — P — A
A — P — T — P — T — P — A
A — P — A — P — A — P — A
... ...

(a) Meta Paths between Authors

Not identical

$AA_1, AA_2, \ldots$

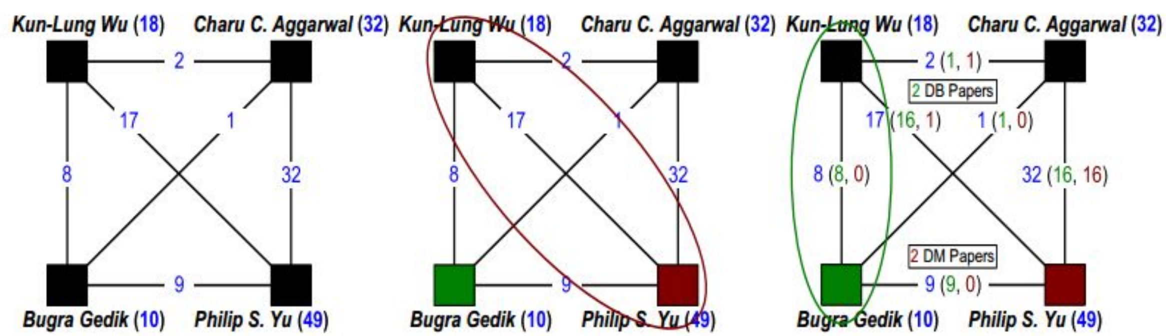


(b) Example Vertex-centric Path Graph

New Challenges——>Basic idea

- **Fine-grained vertex assignment** and clustering objective.
  - Kmeans,K-medoids cannot satisfy.

Kun-Lung Wu	*DB*
Bugra Gedik	*DB*
Charu C.Aggarwal	*DM*
Philip S.Yu	*DM*

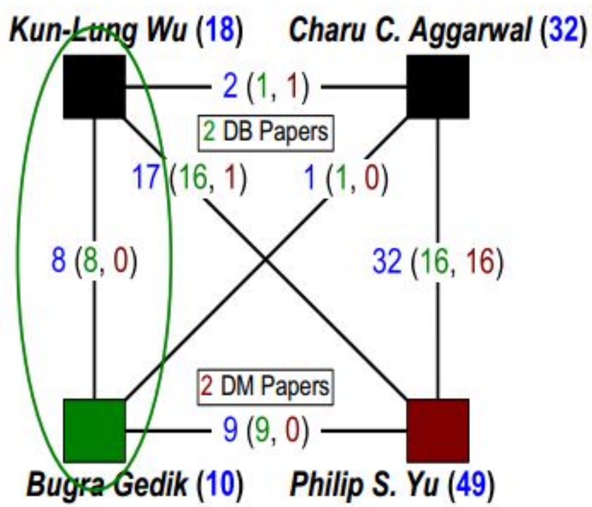


(a) A-P-A Vertex Graph (b) Clustering 1 (c) Clustering 2

Figure 2: Coarse Vertex Assignment/Clustering Objective

New Challenges——>Basic idea

- Edge-centric clustering *w.r.t* multiple path graphs
  - Vertex homophily without edge clustering is insufficient for meta-path graph analysis on hete. networks.



(c) Clustering 2

New Challenges——>Basic idea

- Integrating vertex-centric clustering and edge-centric clustering.

04 VEPathCluster

VEPathCluster

- Vertex/Edge-centric meta path graph clustering
- is to simultaneously perform two clustering **tasks**:
  - Edges soft clustering.
  - Vertex soft clustering.
- **Goals**:
  - Intra-cluster;
  - Inter-cluster.

VEPathCluster(1) Initialization

- Given heterogeneous network **G=(V,E)**,**M** meta paths,cluster number **K**.
- Construct M path graphs **VG_m** which have adjacent matrix **P_m(1≤m≤M) and unify**

$$\mathbf{P}^{(1)} = \omega_1^{(1)}\mathbf{P}_1 + \cdots + \omega_M^{(1)}\mathbf{P}_M \ s.t. \sum_{m=1}^{M} \omega_m^{(1)} = 1,\ \underline{\omega_1^{(1)}, \cdots, \omega_M^{(1)}} \geqslant 0 \quad (1)$$

How to initialize?

and how to update?

Detail in the later section

VEPathCluster(1) Initialization

- Initialize the weights $\omega_m^{(1)} (1 \le m \le M)$

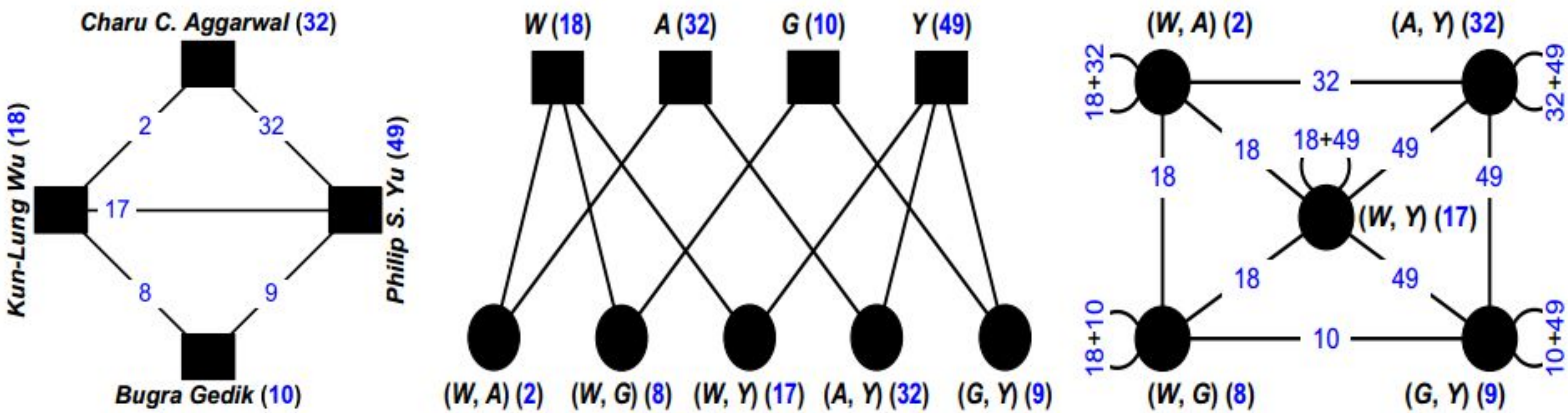
$$\omega_1^{(1)} = \frac{1/\max \mathbf{P}_1}{\sum_{m=1}^{M} 1/\max \mathbf{P}_m}, \ldots, \omega_M^{(1)} = \frac{1/\max \mathbf{P}_M}{\sum_{m=1}^{M} 1/\max \mathbf{P}_m}$$

- Then cluster using Fuzzy C-Means(FCM)(just for the first iteration)

$$X_k^{(1)}(i)(v_i \to c_k)$$

VEPathCluster(2)
Edge-centric random walk model

- Convert: $VG_m \rightarrow EG_m$



(a) Vertex-centric Graph (b) Vertex/Edge Bipartite Graph (c) Edge-centric Graph

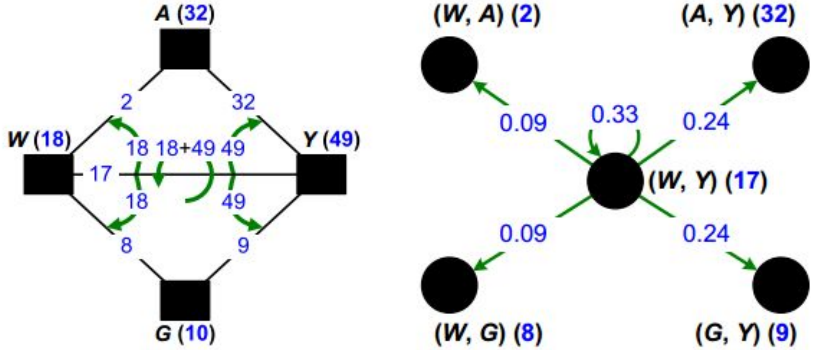
VEPathCluster(2)
Edge-centric random walk model

- Transition probability on edge-centric path graph.

$$\mathbf{T}_m(e_{mi}, e_{mj}) = \begin{cases} \dfrac{\mathbf{Q}_m(e_{mi}, e_{mj})}{\sum_{l=1}^{N_{E_m}} \mathbf{Q}_m(e_{ml}, e_{mj})}, (e_{mi}, e_{mj}) \in E_m \times E_m, \\ 0, \qquad otherwise. \end{cases}, \; 1 \le m \le M \tag{2}$$

- Matrix format:

$$\mathbf{T}_m = \mathbf{Q}_m \mathbf{D}^{-1}, \; 1 \le m \le M \tag{3}$$



(d) Transition between Edges (e) Transition Probability

VEPathCluster(3)
Clustering-based multigraph model

- Construct vertex multigraph **VMG_m** from **VG_m** based the edge clustering result **Y_m^{t-1}** of previous iteration

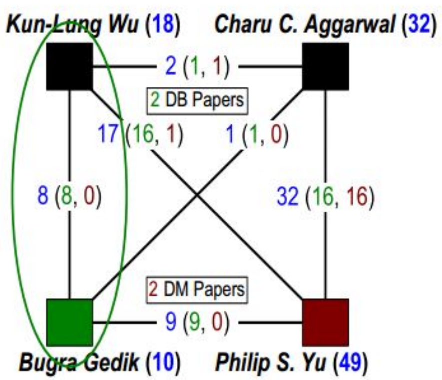
$$\mathbf{P}_{mk}^{(t)}(v_i, v_j) = \mathbf{P}_m(v_i, v_j) \times \mathbf{Y}_{mk}^{(t-1)}((v_i, v_j)),\ 1 \le m \le M,\ 1 \le k \le K \quad (4)$$

- The same as vertex,the edge multigraph

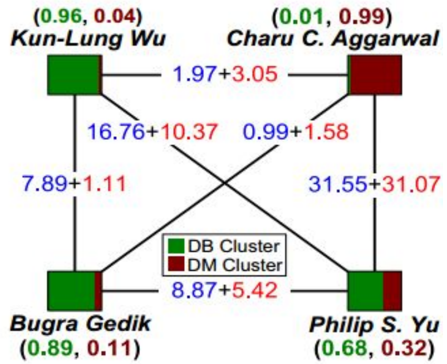
$$\mathbf{Q}_{mk}^{(t)}(e_{mi}, e_{mj}) = \begin{cases} \mathbf{Q}_m(e_{mi}, e_{mj}) \times \mathbf{X}_k^{(t)}(e_{mi} \wedge e_{mj}), & e_{mi} \neq e_{mj}, \\ \mathbf{R}_m(v_a) \times \mathbf{X}_k^{(t)}(v_a) + \mathbf{R}_m(v_b) \times \mathbf{X}_k^{(t)}(v_b), & e_{mi} = e_{mj}. \end{cases}, \quad (5)$$
$$1 \le m \le M,\ 1 \le k \le K$$

VEPathCluster(3)
Clustering-based multigraph model

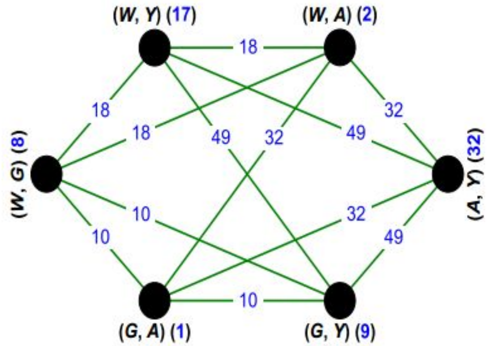
- For example



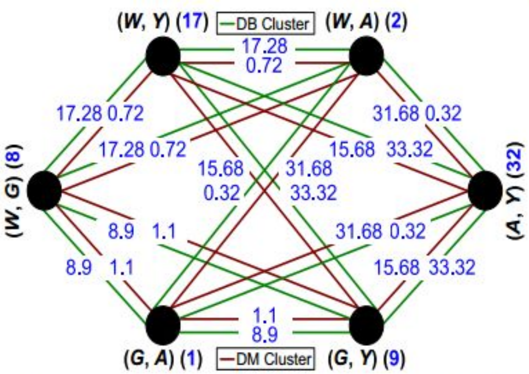
(c) Clustering 2



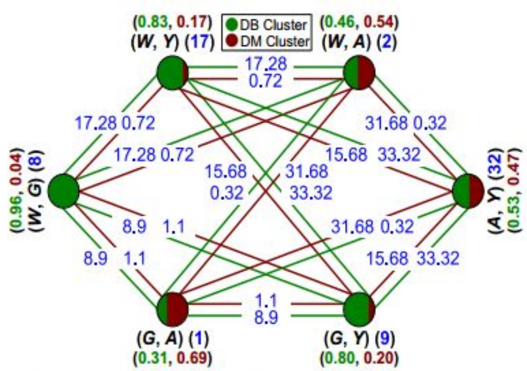
(a) Vertex Clustering w.r.t. Fig. 4 (c)



(b) A-P-A Edge-centric Path Graph w.r.t. Fig. 3 (a)



(c) A-P-A Edge-centric Path Multi-graph w.r.t. Fig. 6 (a) + Fig. 6 (b)



(d) A-P-A Edge Clustering w.r.t. Fig. 6 (a) + Fig. 6 (c)

VEPathCluster(4)
Edge-centric clustering

- Construct edge multigraph $EG_m \rightarrow EMG_m$
- Initialization of clustering result based on the vertex graph's initial clustering

$$\mathbf{Y}_{mk}^{(0)}((v_i, v_j)) = \frac{\sqrt{\mathbf{X}_k^{(1)}(v_i) \times \mathbf{X}_k^{(1)}(v_j)}}{\sum_{l=1}^{K} \sqrt{\mathbf{X}_l^{(1)}(v_i) \times \mathbf{X}_l^{(1)}(v_j)}},\ 1 \le m \le M,\ 1 \le k \le K \quad (6)$$

- N

$$\mathbf{T}_{mk}^{(t)}(e_{mi}, e_{mj}) = \begin{cases} \frac{\mathbf{Q}_{mk}^{(t)}(e_{mi}, e_{mj})}{\sum_{l=1}^{N_{E_m}} \mathbf{Q}_{mk}^{(t)}(e_{ml}, e_{mj})}, & \mathbf{Q}_{mk}^{(t)}(e_{ml}, e_{mj}) \neq 0, \\ 0, & \textit{otherwise}. \end{cases}, \quad (7)$$

$$1 \le m \le M,\ 1 \le k \le K$$

$$\mathbf{T}_{mk}^{(t)} = \mathbf{Q}_{mk}^{(t)}(\mathbf{D}_{mk}^{-1})^{(t)},\ 1 \le m \le M,\ 1 \le k \le K \quad (8)$$

VEPathCluster(4)
Edge-centric clustering

- The update of clustering membership matrix for each meta path

$$
\begin{aligned}
&Initilization:\ \mathbf{Y}_{mk} = \mathbf{Y}_{mk}^{(t-1)} \\
&Iteration:\ \mathbf{Y}_{mk} = \mathbf{T}_{mk}^{(t)} \mathbf{Y}_{mk}
\end{aligned}
\qquad (9)
$$

$$\downarrow converge$$

Normalize

$$\rightarrow \qquad \mathbf{Y}_{mk}^{(t)}(e_{mi}) = \frac{\boxed{\mathbf{Y}_{mk}}(e_{mi})}{\sum_{l=1}^{K} \mathbf{Y}_{ml}(e_{mi})} \qquad (10)$$

- The last updated edge clustering membership matrix

$$\mathbf{Y}_{m}^{(t)} = \left[\mathbf{Y}_{m1}^{(t)} \quad \mathbf{Y}_{m2}^{(t)} \quad \cdots \quad \mathbf{Y}_{mK}^{(t)}\right],\ 1 \leq m \leq M \qquad (11)$$

VEPathCluster(5) Vertex-centric clustering

- Construct vertex multigraph $VG_m \rightarrow VMG_m$

$$\mathbf{P}_{mk}^{(t)}(v_i, v_j) = \mathbf{P}_m(v_i, v_j) \times \mathbf{Y}_{mk}^{(t-1)}((v_i, v_j)),\ 1 \le m \le M,\ 1 \le k \le K \qquad (4)$$

- Cluster membership probability of the first iteration:
  - use FCM to get the $X^{(1)}$ (has mentioned in the secton 1)
- Unified Model:

$$\begin{aligned} \mathbf{P}_1^{(t)} &= \omega_1^{(t)}\mathbf{P}_{11}^{(t)} + \omega_2^{(t)}\mathbf{P}_{21}^{(t)} + \cdots + \omega_M^{(t)}\mathbf{P}_{M1}^{(t)} \\ &\cdots \\ \mathbf{P}_K^{(t)} &= \omega_1^{(t)}\mathbf{P}_{1K}^{(t)} + \omega_2^{(t)}\mathbf{P}_{2K}^{(t)} + \cdots + \omega_M^{(t)}\mathbf{P}_{MK}^{(t)} \end{aligned}, \qquad (13)$$

$$s.t. \sum_{m=1}^{M} \omega_m^{(t)} = 1,\ \omega_1^{(t)}, \cdots, \omega_M^{(t)} \geqslant 0$$

VEPathCluster(5)
Vertex-centric clustering

- New transition probability:

$$\mathbf{S}_k^{(t)}(v_i, v_j) = \begin{cases} \dfrac{\mathbf{P}_k^{(t)}(v_i, v_j)}{\sum_{l=1}^{N_{V_c}} \mathbf{P}_k^{(t)}(v_l, v_j)}, & \mathbf{P}_k^{(t)}(v_i, v_j) \neq 0, \\ 0, & \textit{otherwise}. \end{cases}, \ 1 \le k \le K \qquad (14)$$

$$\mathbf{S}_k^{(t)} = \mathbf{P}_k^{(t)}(\mathbf{D}_k^{-1})^{(t)}, \ 1 \le k \le K \qquad (15)$$

- The update of clustering membership matrix for each meta path

$$\begin{aligned} &\textit{Initilization}: \ \mathbf{X}_k = \mathbf{X}_k^{(t-1)} \\ &\textit{Iteration}: \ \mathbf{X}_k = \mathbf{S}_k^{(t)}\mathbf{X}_k \end{aligned} \Rightarrow \mathbf{X}_k^{(t)}(v_i) = \frac{\mathbf{X}_k(v_i)}{\sum_{l=1}^{K} \mathbf{X}_l(v_i)} \qquad (17)$$

$$\mathbf{X}^{(t)} = \left[\mathbf{X}_1^{(t)} \quad \mathbf{X}_2^{(t)} \quad \cdots \quad \mathbf{X}_K^{(t)}\right] \qquad (18)$$

VEPathCluster(6)
Clustering with weight learning

- Objective function
  - **maximize fuzzy intra-cluster similarity**[22,23].
  - format:

$$O(\mathbf{X},\mathbf{Y}_1,\cdots,\mathbf{Y}_M,\omega_1,\cdots,\omega_M)=\sum_{i=1}^{N_{V_c}}\sum_{j=1}^{N_{V_c}}\sum_{k=1}^{K}\mathbf{X}_k(v_i)\mathbf{X}_k(v_j)\mathbf{P}_k(v_i,v_j)$$
$$+\sum_{m=1}^{M}\sum_{i=1}^{N_{E_m}}\sum_{j=1}^{N_{E_m}}\sum_{k=1}^{K}\mathbf{Y}_{mk}(e_{mi})\mathbf{Y}_{mk}(e_{mj})\mathbf{Q}_{mk}(e_{mi},e_{mj})$$

$$\max_{\omega_1,\cdots,\omega_M} O(\mathbf{X},\mathbf{Y}_1,\cdots,\mathbf{Y}_M,\omega_1,\cdots,\omega_M),\ s.t.\ \sum_{m=1}^{M}\omega_m=1,\ \omega_1,\cdots,\omega_M\geqslant 0 \tag{19}$$

VEPathCluster(6)
Clustering with weight learning

- The above objective function is a fractional function which can be written in f

$$\max_{\omega_1,\cdots,\omega_M} O(\mathbf{X},\mathbf{Y}_1,\cdots,\mathbf{Y}_M,\omega_1,\cdots,\omega_M) = \max_{\omega_1,\cdots,\omega_M} \frac{\sum_{i=1}^{p} a_i \prod_{j=1}^{M} (\omega_j)^{b_{ij}}}{\sum_{i=1}^{q} o_i \prod_{j=1}^{M} (\omega_j)^{r_{ij}}}$$

$$a_i, b_{ij}, o_i, r_{ij} \geq 0, b_{ij}, r_{ij} \in \mathbb{Z}, \; s.t. \sum_{m=1}^{M} \omega_m = 1, \; \omega_1,\cdots,\omega_M \geqslant 0 \tag{20}$$

NFPP

$$\max_{\omega_1,\cdots,\omega_M} \frac{f(\omega_1,\cdots,\omega_M)}{g(\omega_1,\cdots,\omega_M)}, \; s.t. \sum_{m=1}^{M} \omega_m = 1, \; \omega_1,\cdots,\omega_M \geqslant 0 \tag{21}$$

Equivalent

NPPP

$$F(\gamma) = \max_{\omega_1,\cdots,\omega_M} f(\omega_1,\cdots,\omega_M) - \gamma g(\omega_1,\cdots,\omega_M), \; s.t. \sum_{m=1}^{M} \omega_m = 1, \; \omega_1,\cdots,\omega_M \geqslant 0 \tag{22}$$

VEPathCluster_Psedo code

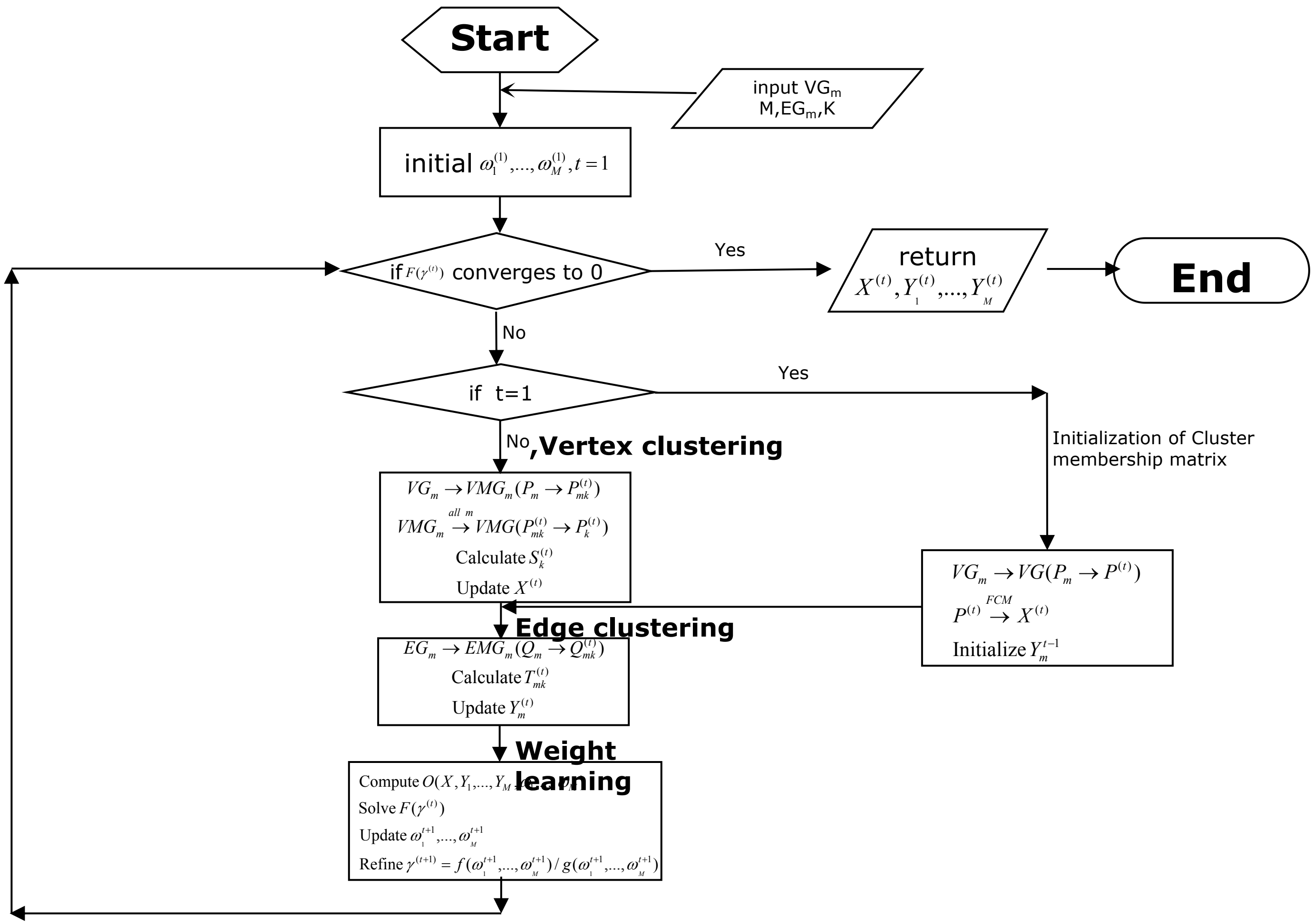
Algorithm 1 **<u>V</u>ertex/<u>E</u>dge-centric meta <u>PATH</u> graph <u>Cluster</u>ing**

Input: M vertex-centric path graphs VG_m, M edge-centric path graphs EG_m, a clustering number K, and a parameter $\gamma^{(1)}$=0.

Output: vertex clustering membership matrix $\mathbf{X}$, M edge clustering membership matrices $\mathbf{Y}_1, \cdots, \mathbf{Y}_M$.

Initialize weights $\omega_1^{(1)}, \cdots, \omega_M^{(1)}$ in terms of the scales of edge values in each VG_m;
for t=1 **to** $F(\gamma^{(t)})$ converges to 0
  if $t = 1$
    Combine $\mathbf{P}_m$ of each VG_m into $\mathbf{P}^{(t)}$ of VG with Eq.(1);
    Invoke FCM to cluster vertices V_o in VG to generate $\mathbf{X}^{(t)}$ of VG;
  else
    Convert $\mathbf{P}_m$ of each VG_m into $\mathbf{P}_{mk}^{(t)}$ of each VMG_m with Eq.(4);
    Combine each VMG_m into VMG by computing all $\mathbf{P}_k^{(t)}$ in Eq.(13);
    Calculate $\mathbf{S}_k^{(t)}$ of VMG for each cluster c_k in Eqs.(14)-(15);
    Update $\mathbf{X}^{(t)}$ of VG with Eqs.(16)-(18);
  if $t = 1$
    Initialize $\mathbf{Y}_m^{(t-1)}$ of each EG_m with Eq.(6);
  Convert $\mathbf{Q}_m$ of each EG_m into $\mathbf{Q}_{mk}^{(t)}$ of each EMG_m with Eq.(5);
  Calculate $\mathbf{T}_{mk}^{(t)}$ of each EMG_m for each cluster c_k in Eqs.(7)-(8);
  Update $\mathbf{Y}_m^{(t)}$ of each EG_m with Eqs.(9)-(11);
  Compute $O(\mathbf{X}, \mathbf{Y}_1, \cdots, \mathbf{Y}_M, \omega_1, \cdots, \omega_M)$ in Eq.(19);
  Solve $F(\gamma^{(t)})$ in Eq.(22);
  Update $\omega_1^{(t+1)}, \cdots, \omega_M^{(t+1)}$;
  Refine $\gamma^{(t+1)}=f(\omega_1^{(t+1)}, \cdots, \omega_M^{(t+1)})/g(\omega_1^{(t+1)}, \cdots, \omega_M^{(t+1)})$;
Return $\mathbf{X}^{(t)}$ and $\mathbf{Y}_1^{(t)}, \cdots, \mathbf{Y}_M^{(t)}$

VEPathCluster_Algorithm Flow



05 Experiment

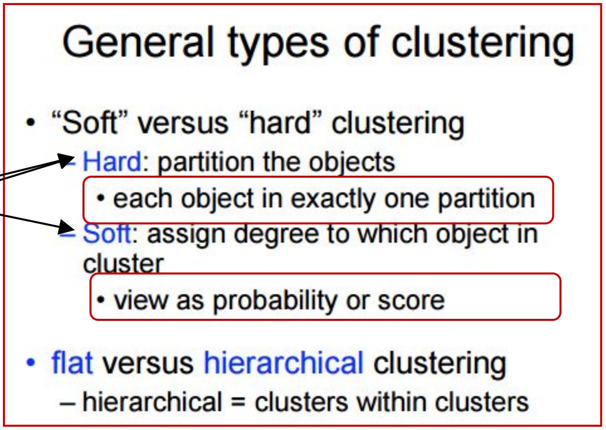
Experiment

- Datasets
- DBLP,IMDb,Yelp

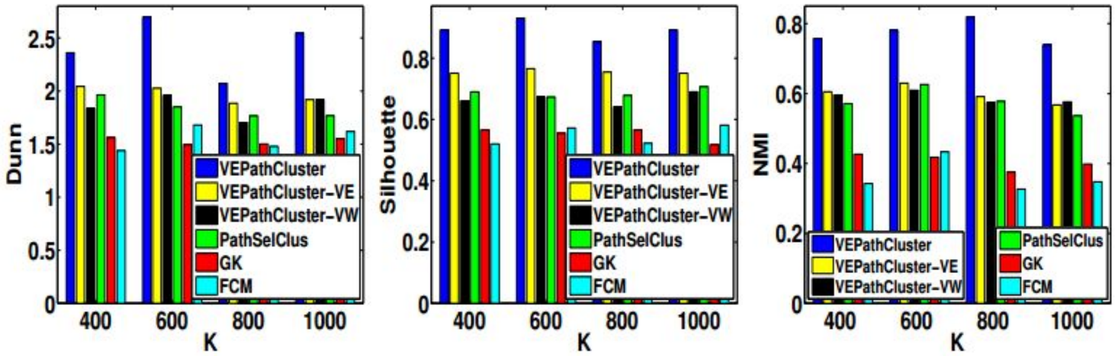
Dataset	#NT	#MP	#Type 1	#Type 2	#Type 3	#Type 4	Meta path
DBLP	4(**A**,P,V,T)	3	112483	728497	2633	45968	A-P-A; A-P-V-P-A; A-P-T-P-A.
IMDb	4(**A**,M,D,G)	3	48975	31188	4774	28	A-M-A; A-M-D-M-A; A-M-G-M-A.
Yelp	4(**B**,R,U,T)	2	15715	470212	138969	30475	B-R-U-R-B; B-R-T-R-B.

Experiment

- Comprison methods
  - Fuzzy C-Means
  - Gustafson-Kessel
  - PathSelClus
  - VEPathCluster-VE,VEPathCluster-VW,VEPathCluster-EW
- Measures
  - Fuzzy dunn index[0,+Inf]
  - Silhouette[-1,1]
  - NMI[0,1]

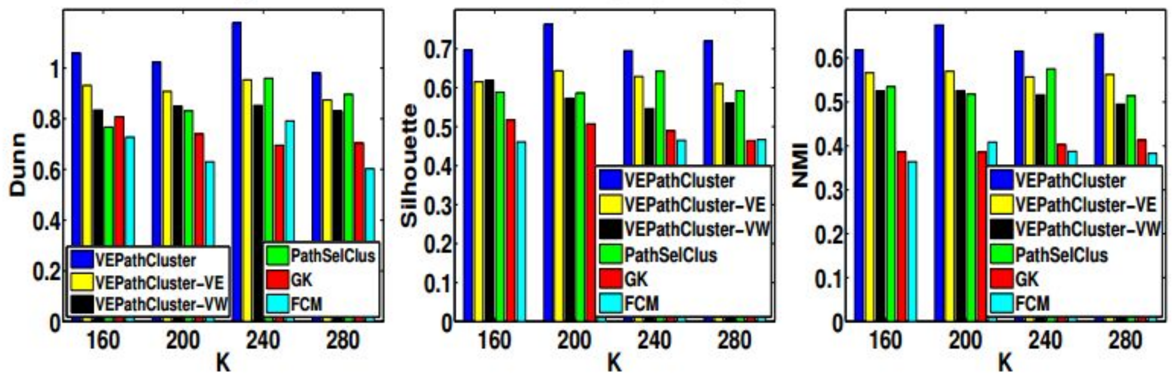


Vertex clustering quality



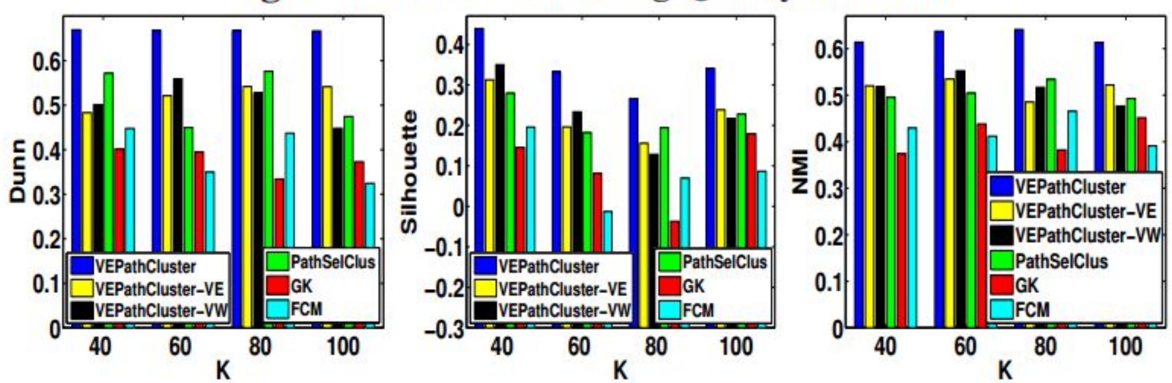
(a) Dunn (b) Silhouette (c) NMI

Figure 7: Vertex Clustering Quality on DBLP



(a) Dunn (b) Silhouette (c) NMI

Figure 8: Vertex Clustering Quality on IMDb

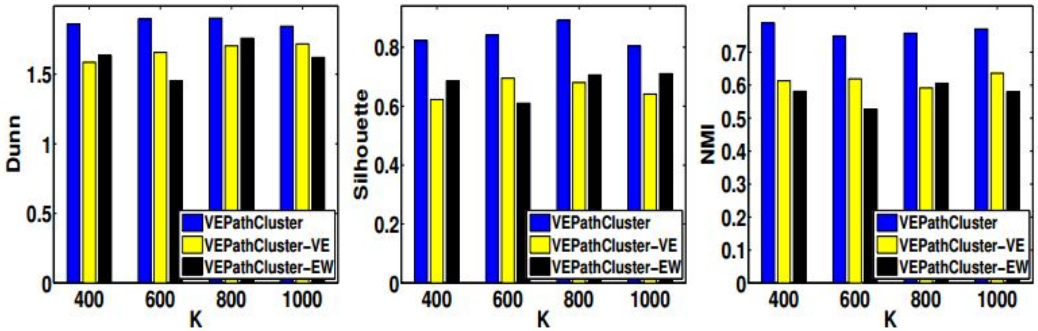


(a) Dunn (b) Silhouette (c) NMI

Figure 9: Vertex Clustering Quality on Yelp

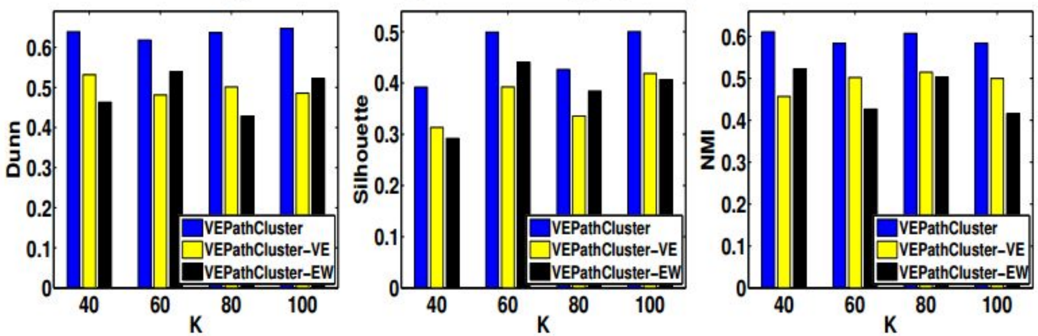
Experiment

- Edge Clustering Quality



(a) Dunn (b) Silhouette (c) NMI

Figure 10: Edge Clustering Quality on DBLP

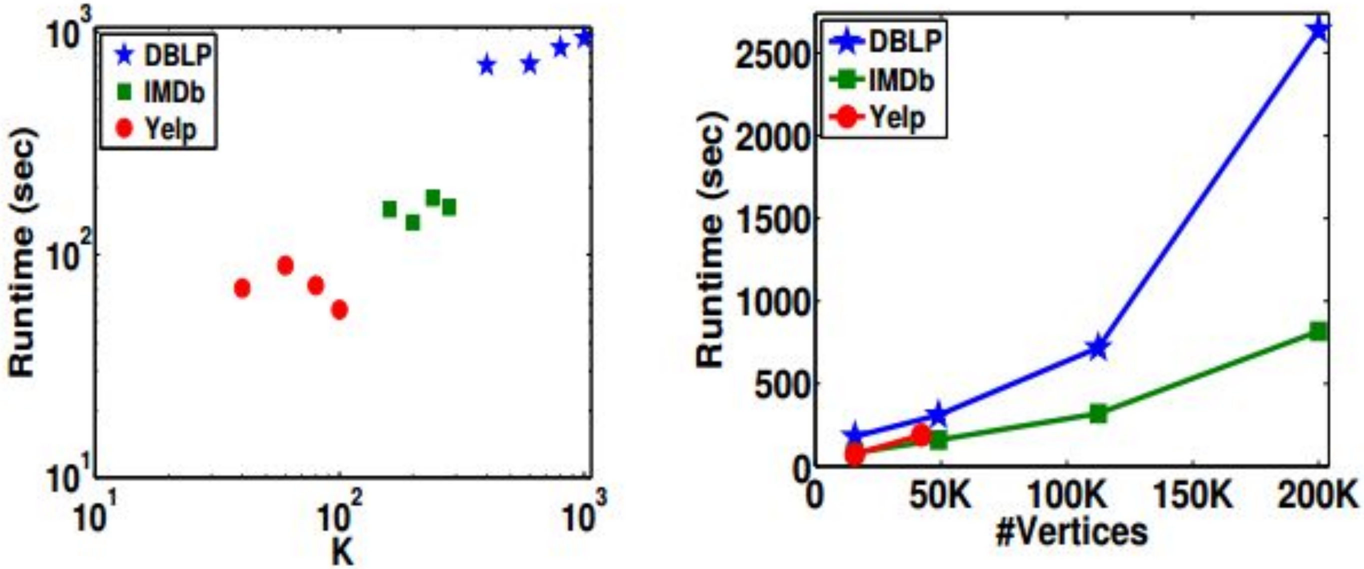


(a) Dunn (b) Silhouette (c) NMI

Figure 11: Edge Clustering Quality on Yelp

Experiment

- Clustering efficiency

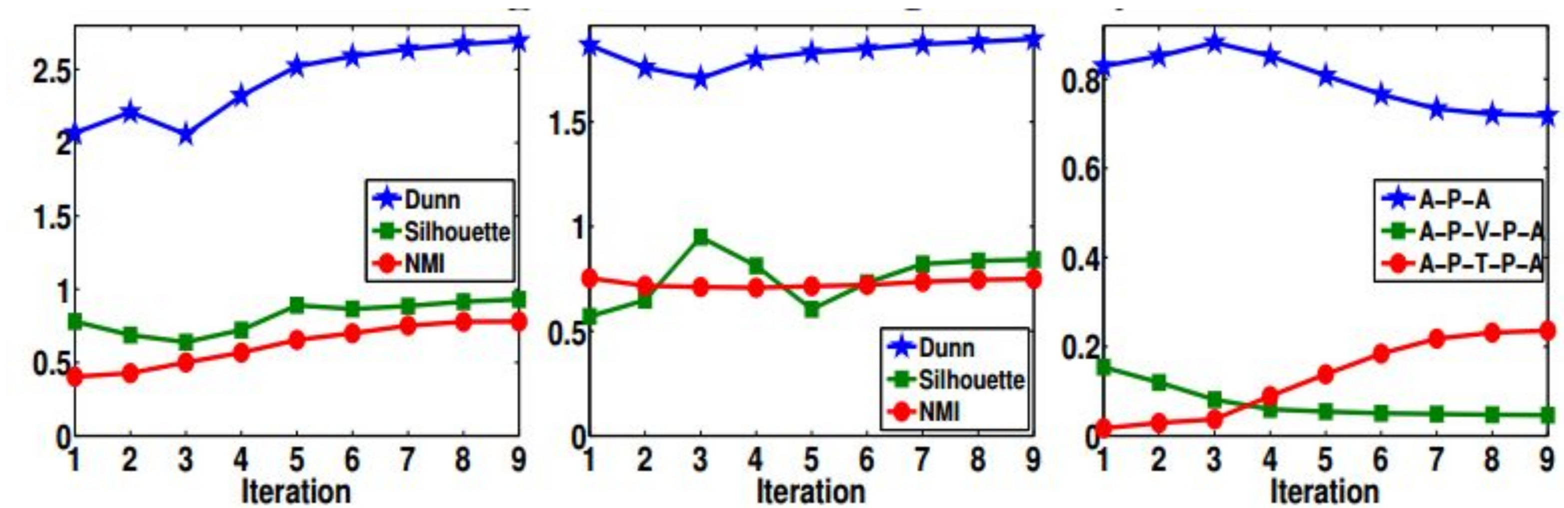


(a) Varying *K* (b) Varying #Vertices

Figure 12: Clustering Efficiency

Experiment

- Clustering convergence



(a) Vertex Clustering (b) Edge Clustering (c) Weight Assignment

Figure 13: Clustering Convergence

Experiment

- Case study

DB,DM,SN,MMN
IR, DL
DM
DM
Big data,data science,DB
DB
DB
machine learning,statistic(AI)
AI,DM
DM
DM

Author/Cluster	DB	DM	AI	IR
Ming-Syan Chen	0.258	0.588	0.021	0.134
W. Bruce Croft	0.058	0.006	0.026	0.909
Christos Faloutsos	0.346	0.539	0.012	0.102
Jiawei Han	0.373	0.459	0.057	0.111
H. V. Jagadish	0.904	0.048	0.014	0.034
Laks V. S. Lakshmanan	0.809	0.128	0.011	0.053
Hector Garcia-Molina	0.810	0.028	0.021	0.141
Eric P. Xing	0.009	0.123	0.830	0.038
Qiang Yang	0.012	0.265	0.512	0.210
Philip S. Yu	0.358	0.507	0.027	0.108
Chengqi Zhang	0.023	0.744	0.140	0.093

Table 1: Cluster Membership Probabilities of Authors Based on Three Meta Paths from DBLP

Experiment

- Case study

Path Edge/Cluster	DB	DM	AI	IR
(Ming-Syan Chen, Philip S. Yu)	0.630	0.284	0.023	0.063
(W. Bruce Croft, Hector Garcia-Molina)	0.702	0.035	0.065	0.199
(Christos Faloutsos, H. V. Jagadish)	0.547	0.365	0.017	0.072
(Christos Faloutsos, Eric P. Xing)	0.238	0.713	0.015	0.034
(Jiawei Han, Laks V. S. Lakshmanan)	0.624	0.356	0.006	0.013
(Jiawei Han, Philip S. Yu)	0.518	0.424	0.013	0.045
(Qiang Yang, Philip S. Yu)	0.083	0.785	0.131	0.001
(Qiang Yang, Chengqi Zhang)	0.023	0.684	0.228	0.065

Table 2: Cluster Membership Probabilities of A-P-A Path Edges from DBLP

06 Our plan and exsiting questions

Our plan and exsiting questions

- First, based on the meta-path decomposition method.
  - Question:
- Then , use a new clustering method such as **sync.**
  - Question1:cluster a whole homogeneous network(how to integrate different networks?)
    - how to decide the weights?
  - Question2:cluster different networks seperately?(how to integrate the clustering results?)