The University of Hong Kong

Semi-supervised Clustering in Attributed Heterogeneous Information Networks

Xiang Li, Ben Kao, Yudian Zheng, The University of Hong Kong

Yao Wu, Martin Ester, Xin Wang, Simon Fraser University

April 7th, 2017

Introduction

- Attributed heterogeneous information network (AHIN)

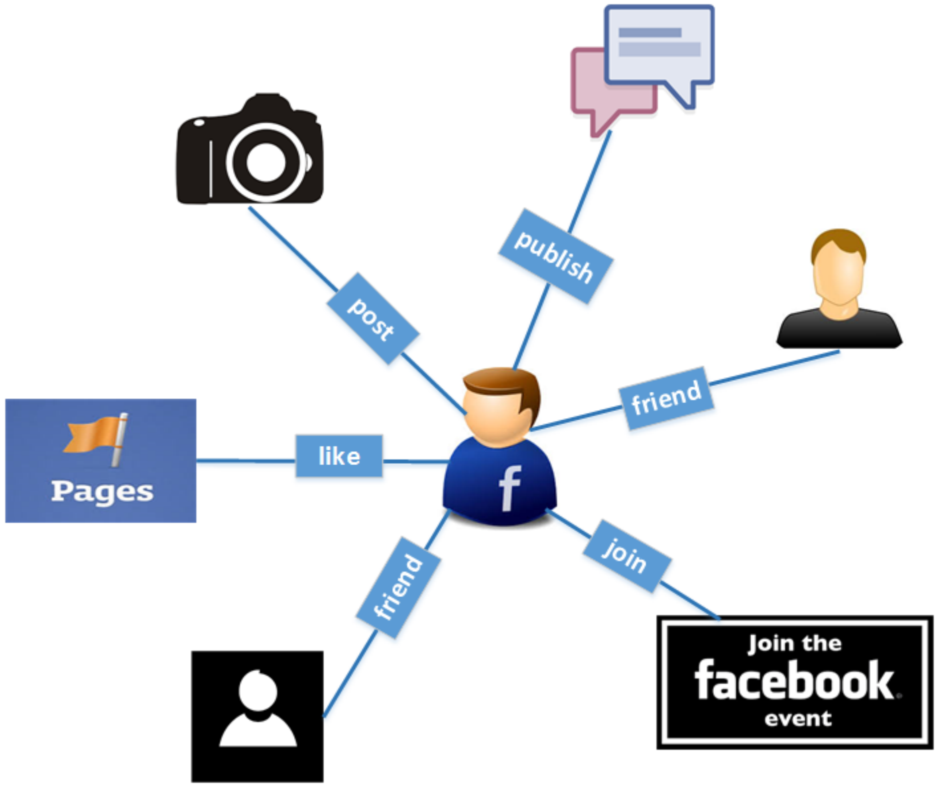
□ Heterogeneous information network
  - ○ multiple types of objects
  - ○ different types of links

□ Object attributes

□ Example: Facebook Open Graph

○ objects: users, pages, photos, events, etc.

○ attributes:
  - users (gender, age, school, etc.),
  - photo (lat-long, date/time)

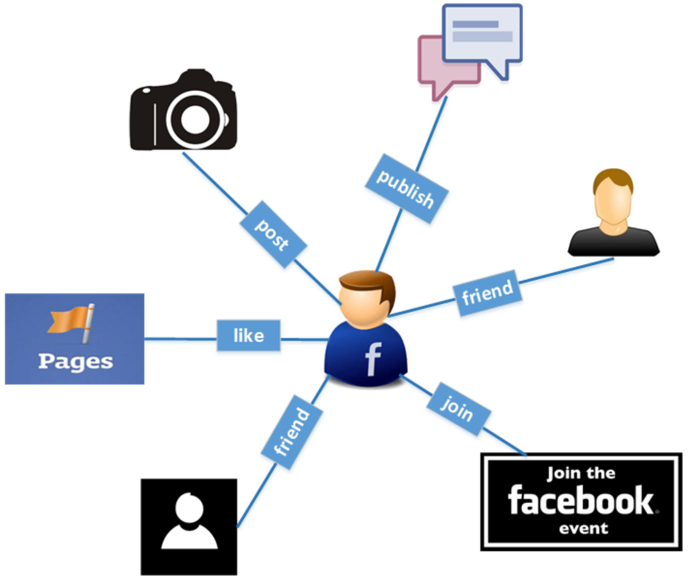


Meta-path

The University of Hong Kong

- A meta-path is a sequence of object types that expresses a relation between objects
- Example: Facebook Open Graph

□ objects: users (U), product pages (P), etc.



- UPU: user-page-user (two users like the same product page)
- UUU: user-user-user (two users have a common friend)

Challenge

- Why clustering in attributed heterogeneous information networks?
- Link-based similarity
  - simple network distance measure (eg: random walk)
  - meta-path based measure (eg: PathSim)
- Attribute-based similarity
- **Challenge1: how to aggregate various types of similarities?**

Challenge

- Not all the attributes and meta-paths are useful
- Automatic process to select best attributes and meta-paths
- User can provide guidance to supervise the clustering
- **Challenge2: how to automatically perform the selection?**

Related Work

	without supervision			supervision		
	attribute	link	both	attribute	link	both
homogeneous	k-means, Ncuts	METIS, AGM, BigClam	CODICIL, CESNA, SA-Cluster	Spectral-learning, SS-Kernel-kmeans	label propagation	FocusCO
heterogeneous		RankClus, NetClus, SI-Cluster	GenClus		PathSelClus, SemiRPClus	**SCHAIN**

Attribute-based similarity

- Suppose x_u has attribute vector f_u, x_v has attribute vector f_u

$$S_A(x_u, x_v) = \sum_{j=1}^{|A_i|} (\omega_j \cdot sim(f_{uj}, f_{vj})),$$

- sim() can be any standard similarity function defined over the j-th attribute

Link-based similarity

- An effective meta-path based measure: PathSim
- Each meta path P_j defines a similarity measure S_{P_j}
- Suppose we have m meta paths, then $$S_L = \sum_{j=1}^{m} \lambda_j S_{P_j}$$
- To combine attribute-based and link-based similarity, we have:

$$S = \alpha S_A + (1-\alpha) S_L$$

Supervision constraints

- Must-link set M and cannot-link set C (user supervision)
- To measure the clustering quality,

1. How similar intra-cluster and inter-cluster objects are?
2. How well the cluster agrees with the supervision constraints?

- we use normalized cuts to be the measure

- **reward** object pairs in M which are clustered in the same cluster
- **penalize** objects pairs in C which are clustered in the same cluster

$$\mathcal{J}(\boldsymbol{\lambda}, \boldsymbol{\omega}, \{C_r\}_{r=1}^k) = \sum_{r=1}^{k} \frac{links(C_r, \mathcal{X}_i \backslash C_r)}{links(C_r, \mathcal{X}_i)}$$

Ncuts

Reward

$$- \sum_{r=1}^{k} \sum_{\substack{(x_u, x_v) \in \mathcal{M} \\ L(x_u) = L(x_v) = r}} \frac{S(x_u, x_v)}{links(C_r, \mathcal{X}_i)} \quad (4)$$

penalty

$$+ \sum_{r=1}^{k} \sum_{\substack{(x_u, x_v) \in \mathcal{C} \\ L(x_u) = L(x_v) = r}} \frac{S(x_u, x_v)}{links(C_r, \mathcal{X}_i)},$$

- Our goal is to minimize J

$$\mathcal{J}(\boldsymbol{\lambda}, \boldsymbol{\omega}, \{\boldsymbol{z}_r\}_{r=1}^{k}) = \sum_{r=1}^{k} \frac{\boldsymbol{z}_r^T (D - S - \mathcal{W} \circ S) \boldsymbol{z}_r}{\boldsymbol{z}_r^T D \boldsymbol{z}_r} + \gamma(||\boldsymbol{\lambda}||^2 + ||\boldsymbol{\omega}||^2). \tag{6}$$

- Constraints:

$$\sum_{r=1}^{k} \boldsymbol{z}_r(u) = 1$$

$$\boldsymbol{z}_r(u) \in \{0, 1\}$$

$$\sum_{j=1}^{|\mathcal{PS}|} \lambda_j = 1$$

$$\sum_{l=1}^{|A_i|} \omega_l = 1$$

$$\lambda_j \geq 0$$

$$\omega_l \geq 0$$

Optimization

- An iterative method
  - Optimize $\{z_r\}_{r=1}^{k}$ given λ and ω
    - Transform into spectral clustering optimization problem
  - Optimize λ and ω given $\{z_r\}_{r=1}^{k}$
    - Transform into a non-linear fractional programming problem

Experiment

- Task1: Yelp-Business

❑ businesses (B), cities (C), users (U) and categories (T)

❑ business attributes: lat-long, review count, quality star and lot

❑ meta-paths = {BCB (two businesses are in the same city), BUB (two businesses have the same customer), BTB (two businesses are of the same category)}

❑ clustering objective: to cluster businesses by geographical state

Experiment

- Task2: Yelp-Restaurant

❑ restaurants (B), reviews (R), users (U) and keywords (K)

❑ restaurant attributes: service, reserve, review count, quality star and lot

❑ meta-paths = {BRURB (two restaurants have reviews written by the same customer), BRKRB (two restaurants have reviews with the same keyword)}

❑ clustering objective: to cluster restaurants by category

Experiment

- Task3: DBLP

❑ authors (A), papers (P) and terms (T)

❑ author attributes: published paper count to CIKM, KDD,VLDB and SIGIR

❑ meta-paths = {APA (co-authorship), APAPA (two authors publish papers with the same coauthor), APTPA (two authors publish papers with the same keyword)}

❑ clustering objective: to cluster authors by research interests

Clustering quality

The University of Hong Kong

Table 2: NMI comparison on Yelp-Restaurant

	Attribute-only		Link-only			Attribute+Link	SCHAIN Variants		
% seeds	SL	SNcuts	GNetMine	PathSelClus	SemiRPClus	FocusCO	SCHAIN-RWR	SCHAIN-NL	SCHAIN
5%	0.225	0.185	0.284	0.564	0.142	0.088	0.427	0.628	**0.689**
10%	0.258	0.188	0.332	0.610	0.134	0.087	0.429	0.635	**0.707**
15%	0.416	0.192	0.367	0.627	0.136	0.095	0.433	0.655	**0.725**
20%	0.425	0.198	0.379	0.635	0.132	0.087	0.426	0.678	**0.738**
25%	0.437	0.251	0.392	0.637	0.136	0.090	0.436	0.689	**0.744**

Weight learning

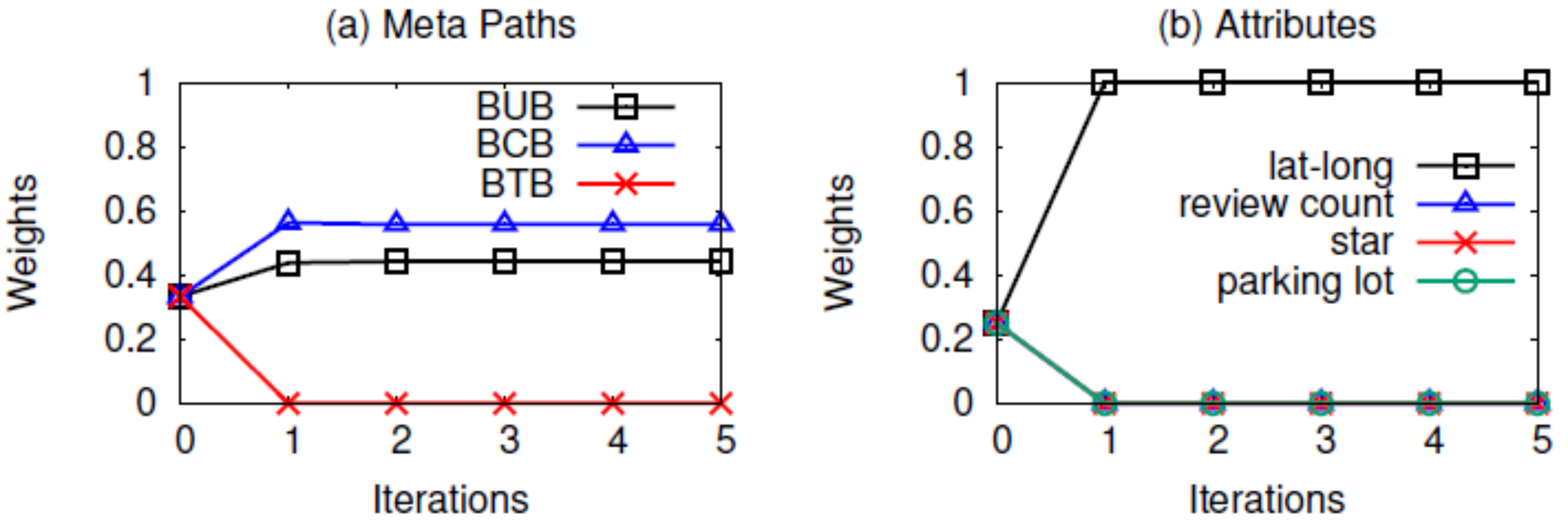


Figure 2: Weight learning on Yelp-Business

Weight learning

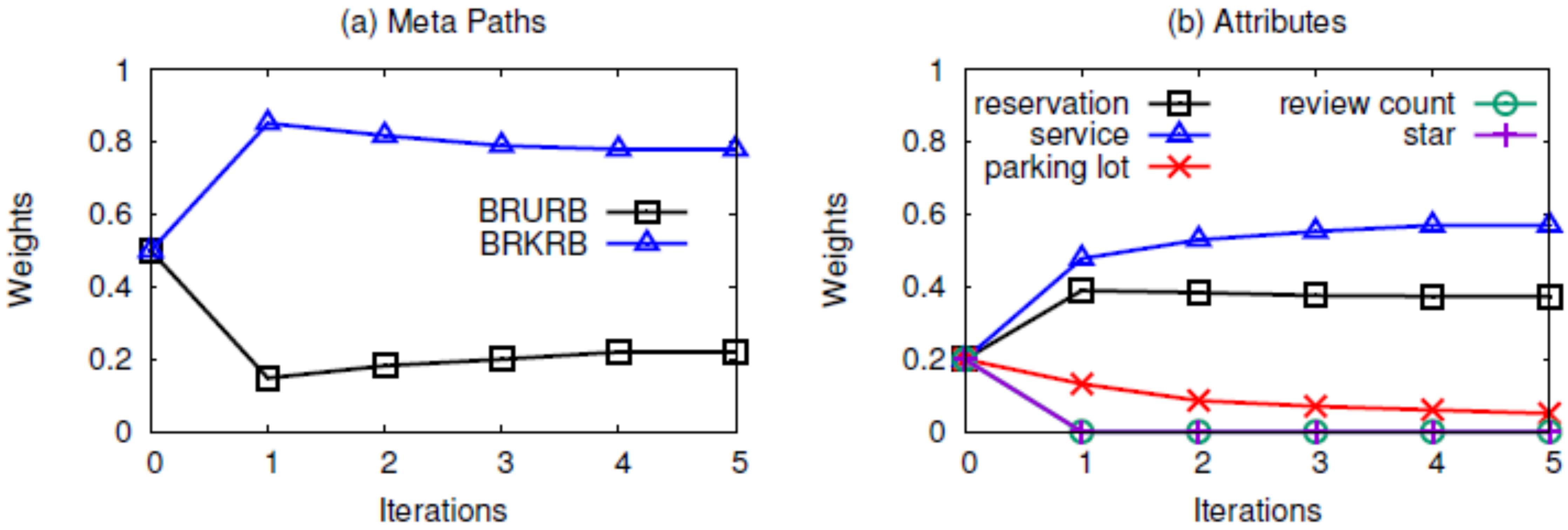


Figure 3: Weight learning on Yelp-Restaurant

Weight learning

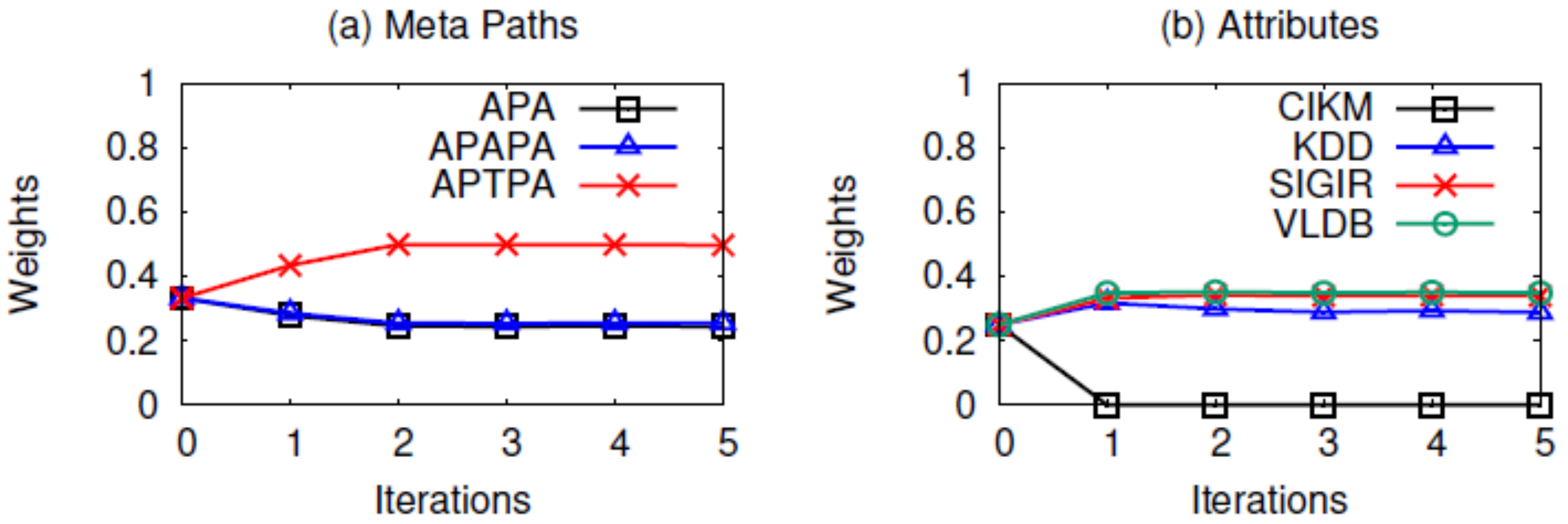


Figure 4: Weight learning on DBLP

Convergence analysis

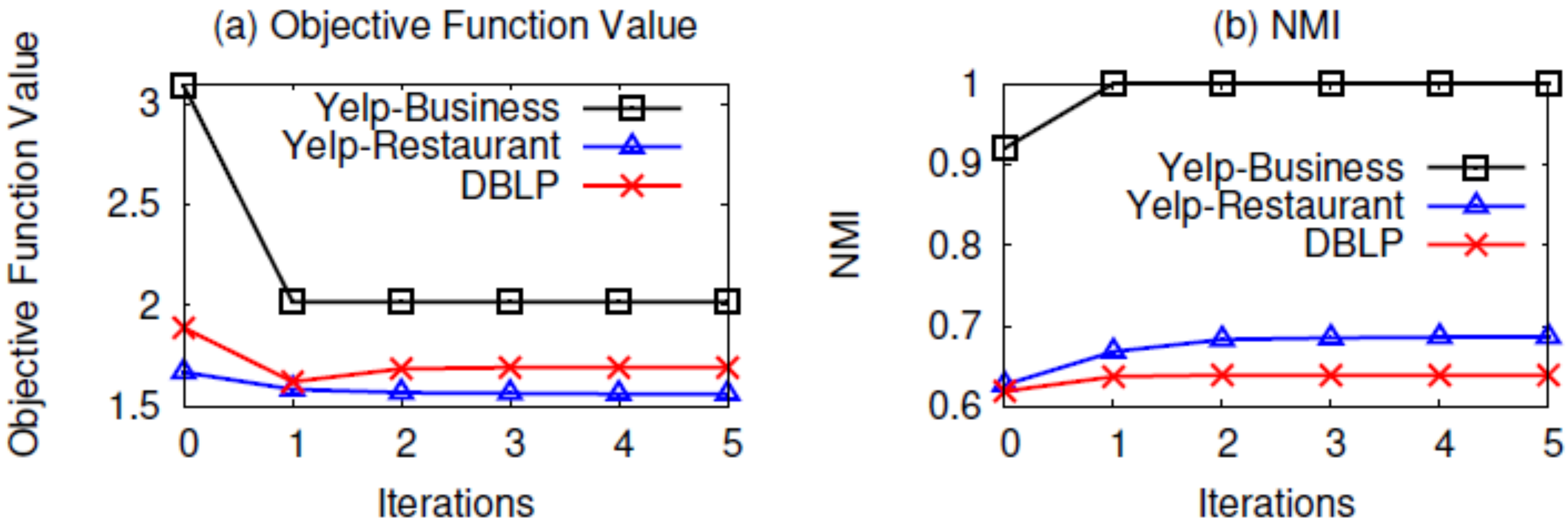


Figure 5: Convergence analysis

Conclusion

- We studied semi-supervised clustering in AHINs
- We proposed a novel algorithm SCHAIN which considers both object attributes and meta-paths
- We experimentally proves the usefulness of SCHAIN

Thank you!