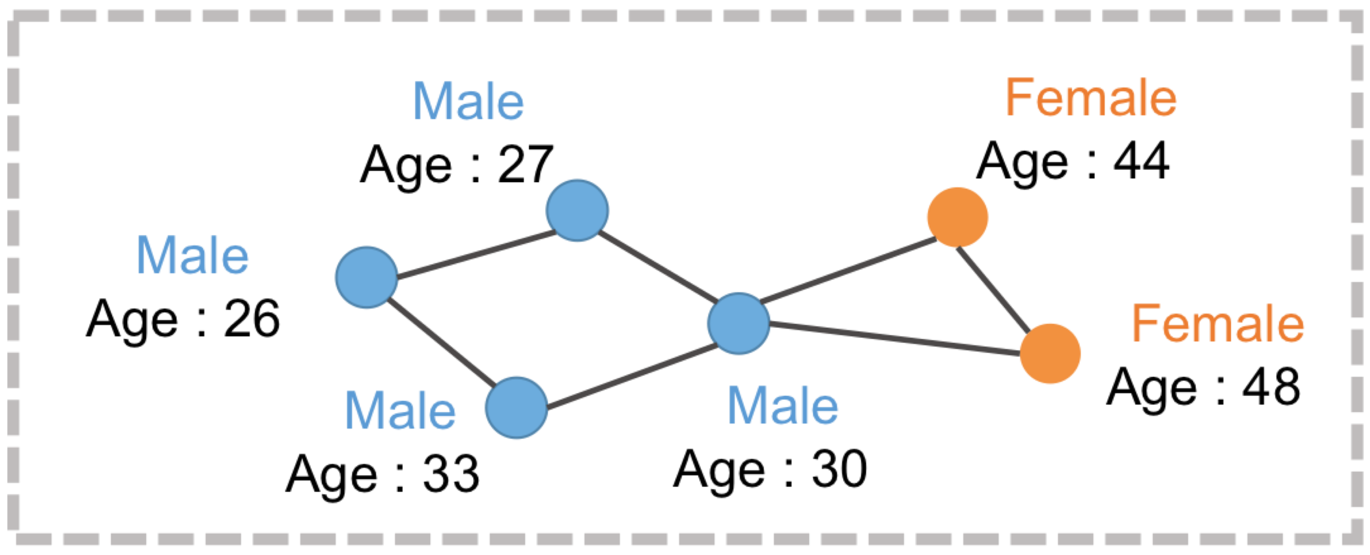
Weighted Micro-Clustering :

Application to Community Detection in Large-Scale Co-Purchasing Networks with User Attributes

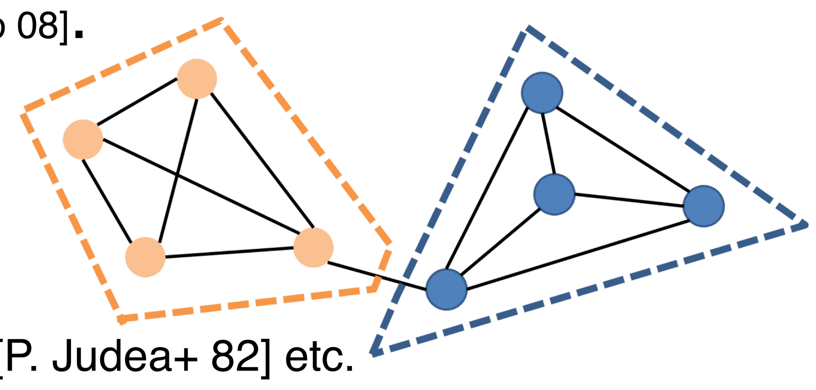
Tomoya Yamazaki, Nobuyuki Shimizu, Hayato Kobayashi, and Satoshi Yamauchi
Yahoo Japan Corporation

Our Goal

- Detecting communities from undirected graphs **with vertex attributes**.



- Communities are vertices which probably share common properties and/or play similar roles within the graph [S. Fortunato 08].



Traditional methods

1. Propagation-based methods [C.D. Manning+ 08], [P. Judea+ 82] etc.
2. Modularity-based methods [V. Blondel+ 08], [M. Girvan+ 02] etc.
3. Clique-based methods [T. Uno+ 15], [G. Palla+ 05] etc.

Our Results

- Propose the new algorithm by extending Micro-clustering [T. Uno+ 15] to accept weighted graphs
- Apply our method to large-scale co-purchasing networks of real online auction data (YAHUOKU!)

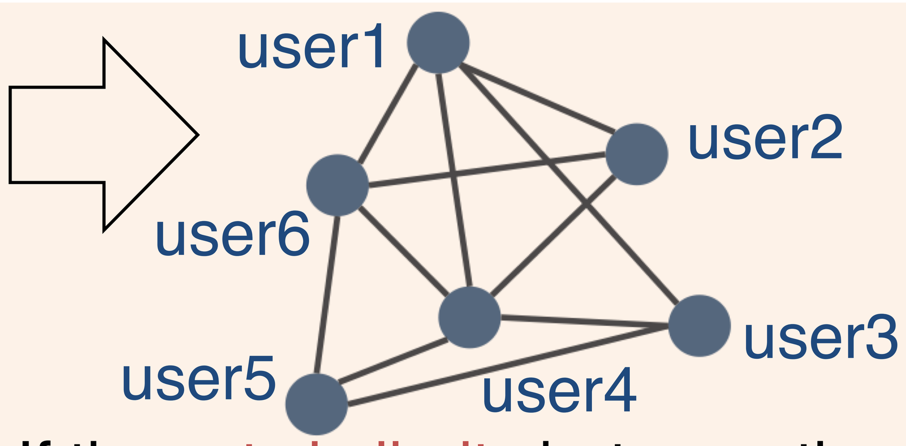
Our Method : Weighted Micro-Clustering

- Simple and scalable soft-clustering method based on Micro-Clustering [T.Uno+ 15] for detecting communities from undirected graphs with vertex attributes

Algorithm

1. Compose a co-purchasing graph $G=(V,E)$ from data.

User	Purchased items
user1	{itemA, itemB, itemC, itemD}
user2	{itemA, itemB, itemD}
user3	{itemB,itemC, itemE}
…	…



If the set similarity between the purchased items of users $> \phi$: Connecting between the users.

2. Compose a polished graph $G_w=(V,E_w)$ from G where

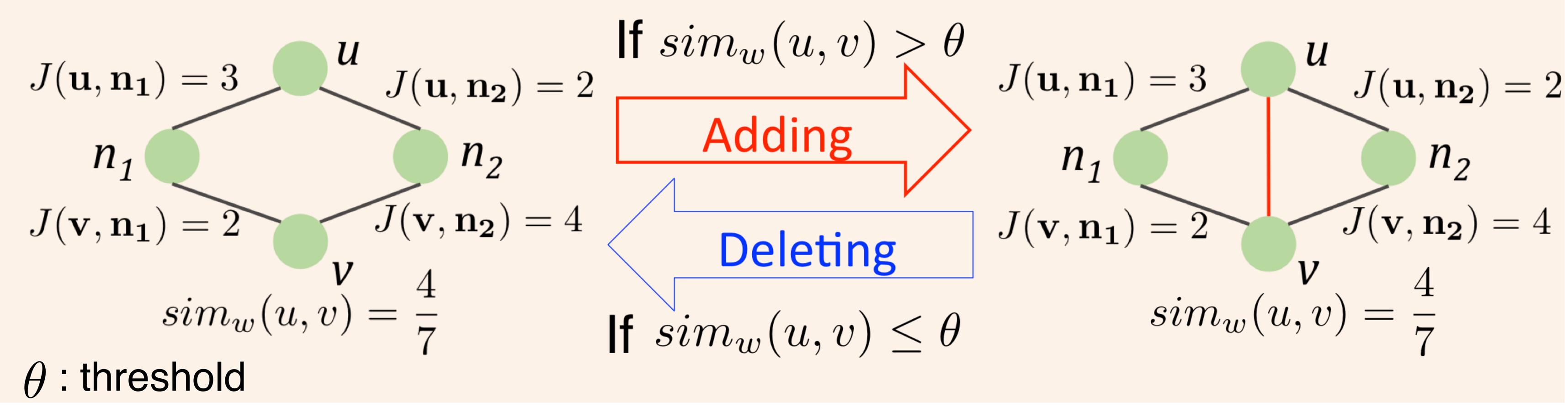
$$E_w = \{(u,v) \mid sim_w(u,v) > \theta, |N(u)\cap N(v)| > 0, (u,v)\in E\},$$

$$sim_w(u,v) = \frac{\sum_{n\in N(u)\cap N(v)} \min(J(\mathbf{u},\mathbf{n}), J(\mathbf{v},\mathbf{n}))}{\sum_{n\in N(u)\cup N(v)} \max(J(\mathbf{u},\mathbf{n}), J(\mathbf{v},\mathbf{n}))}$$

$J(\mathbf{u},\mathbf{v})$ is a similarity between vectors. For example

$$J(\mathbf{u},\mathbf{v}) = \frac{\sum_{i=1}^{n}\min(\mathbf{u_i},\mathbf{v_i})}{\sum_{i=1}^{n}\max(\mathbf{u_i},\mathbf{v_i})} \quad \text{(Generalized jaccard index)}$$

$\mathbf{u},\mathbf{v}$: the attribute vectors of u, v, respectively.
$\mathbf{u_i},\mathbf{v_i}$: the ith element of the vector $\mathbf{u},\mathbf{v}$, respectively.



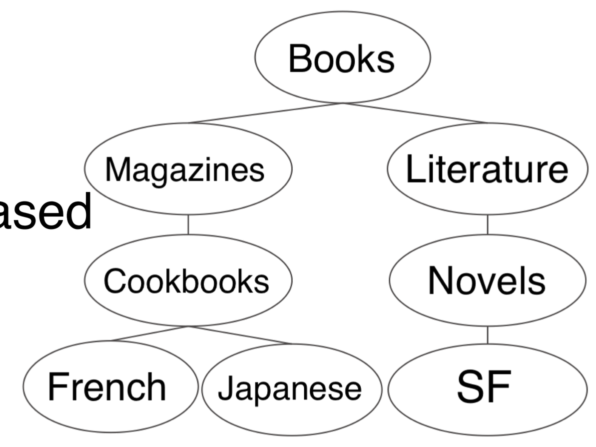
3. Enumerate maximal cliques from the polished graph G_w

Experiments

Data : Co-purchasing graph data from YAHUOKU!

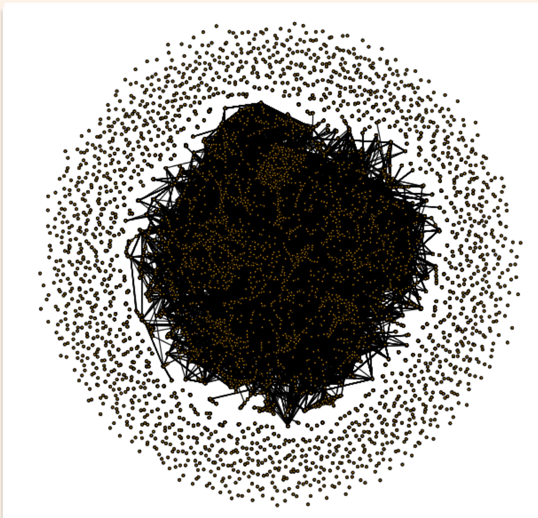
Data set	CHARITY	PET	FOOD	BOOKS	SPORTS
Vertices	99	4,249	22,398	39,469	120,741
Edges	1,645	60,110	492,158	525,873	2,185,242
True Class	2	12	119	151	141

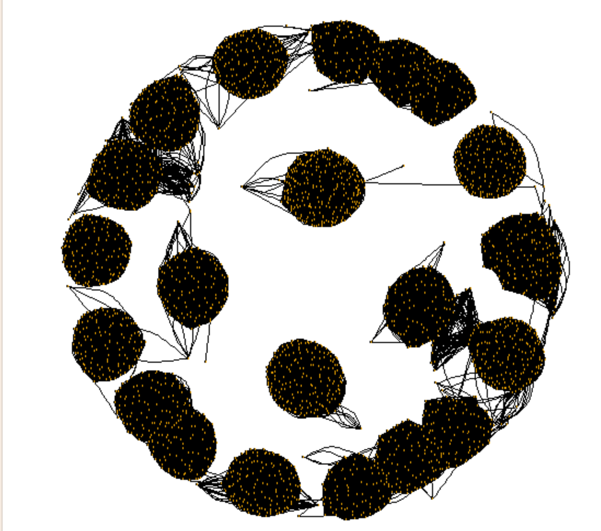
Vertices represent buyers, and if two buyers buy products from almost the same sellers, their vertices are adjacent.
True class : 3rd level categories of items user purchased
User attributes : 2nd level categories of items user purchased

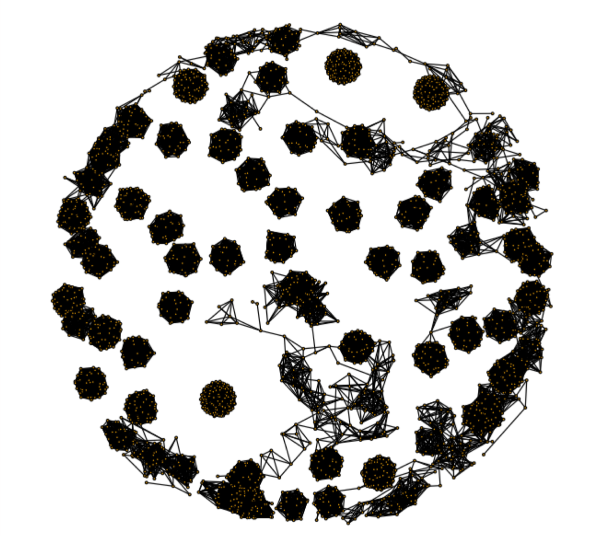


Results

- Results of BOOKS data


Similarity graph


Polished graph by Micro-clustering


Polished graph by our method

- Evaluation

1. Partition vertices by associating users with their most frequently purchased item categories.
2. Measure NMI and Purity scores between the set of users in communities $\Omega = \{w_1,\dots,w_M\}$ and in the true class $C=\{c_1,\dots,c_L\}$

$$NMI(\Omega,C) = \frac{-2\sum_{m=1}^{M}\sum_{l=1}^{L}\frac{|w_m\cap c_l|}{S}\log\frac{S|w_m\cap c_l|}{|w_m||c_l|}}{\sum_{m=1}^{M}\frac{|w_m|}{S}\log\frac{|w_m|}{S}+\sum_{l=1}^{L}\frac{|c_l|}{S}\log\frac{|c_l|}{S}}$$

$$purity(\Omega,C) = \frac{1}{S}\sum_{m}\max_{l}|w_m\cap c_l| \quad \text{where } S = |\bigcup_{\omega\in\Omega} w|$$

NMI (Normalized Mutual Information)

Data set	CHARITY	PET	FOOD	BOOKS	SPORTS
Our method	0.6723	0.5368	0.5899	0.6463	0.5251
Micro-Clustering [T. Uno+ 15]	0.6689	0.5272	0.5779	0.6020	0.5212
Louvain method [V. Blondel+ 08]	0.6083	0.6614	0.5707	0.5110	0.3213
k-means method	0.7782	0.2898	0.4008	T.O.	T.O.

Purity Scores

Data set	CHARITY	PET	FOOD	BOOKS	SPORTS
Our method	1.0	0.9582	0.8953	0.7917	0.8101
Micro-Clustering [T. Uno+ 15]	0.9914	0.9482	0.8286	0.7619	0.4298
Louvain method [V. Blondel+ 08]	0.9759	0.8409	0.6688	0.5635	0.3740
k-means method	0.8989	0.3656	0.6845	T.O.	T.O.

Conclusion

- Our method takes account of both the graph structure and their vertex attributes.
- Our method outperforms previous methods.