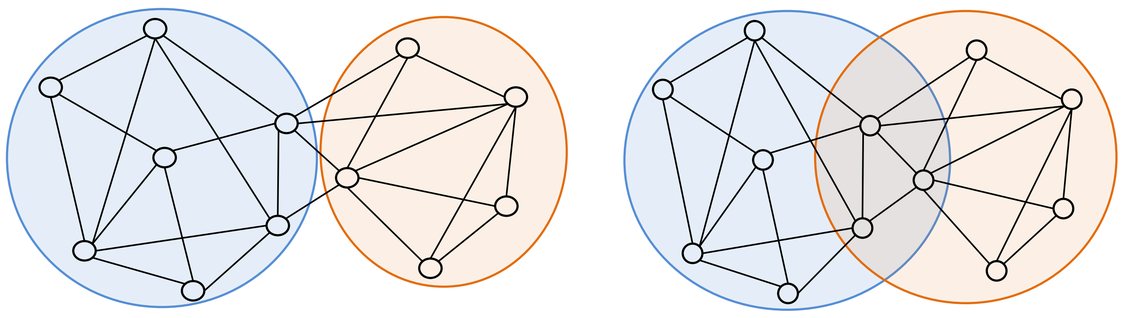
Exploiting k-Degree Locality to Improve Overlapping Community Detection

Hongyi Zhang, Michael R. Lyu, Irwin King
{hyzhang,lyu,king}@cse.cuhk.edu.hk

Introduction



Why overlapping community detection (right figure)?

- Classical community detection (left) assume that communities are mutually exclusive.
- However, many complex networks we encounter in daily life allow multiple memberships.

Community detection via matrix factorization (MF)

$$\min_{F} l(G, FF^T)$$

$G \in \{0,1\}^{n \times n}$ is the adjacency matrix

$F \in \mathbb{R}_+^{n \times p}$ is the node-community membership matrix, $F_{u,c}$ represents the weight of node u in community c

Motivation

- Local non-neighbors (e.g., my friend's friend but not my friend) are helpful when discovering communities.

Main contribution

- We propose a *Locality-based Non-negative MF (LNMF)* model to improve the *Preference-based NMF (PNMF)* model by enhancing the preference system with the help of locality.

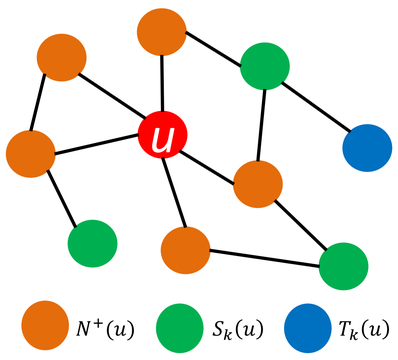
Related work: PNMF model (Zhang et al., AAAI'15)

- Motivation: value-approximation based objective function, e.g., squared loss, is problematic because entry in adjacency matrix is more like a label than a value.
- Intuition: two nodes are likely to build a link with each other if they share common communities.

Setup and Notations

k-Degree local network

- $L_k(u)$: the set of nodes whose length of shortest path to u is less than or equal to k.
- $S_k(u) \coloneqq L_k(u) \backslash L_1(u)$, where $k \geq 1$.
- $T_k(u) \coloneqq L_\infty(u) \backslash L_k(u)$, where $k \geq 1$.
- A k=2 case is shown on the right.



Other notations

- $N^+(u)$: u's neighbors
- $N^-(u)$: u's non-neighbors

Model Assumption

The notation

- $r_{u,i}$: the preference of node u on node i

Older assumption (PNMF)

- $r_{u,i} \geq r_{u,j}, i \in N^+(u), j \in N^-(u)$
  - neighbors are preferred to non-neighbors

New assumption (LNMF)

- $r_{u,i} \geq r_{u,j}, r_{u,j} \geq r_{u,d}, i \in N^+(u), j \in S_k(u), d \in T_k(u)$
  - neighbors are preferred to k-degree local non-neighbors
  - k-degree local non-neighbors are preferred to k-degree distant non-neighbors
- When $k = 1$, our new assumption degrades to the old one of PNMF.

Model Formulation and Parameter Learning

Model formulation

- For each node u, the objective is to maximize a product of pairwise preference:

$$\prod_{i \in N^+(u), j \in S_k(u)} P(r_{u,i} \geq r_{u,j} | F) \prod_{j \in S_k(u), d \in N^-(u)} P(r_{u,j} \geq r_{u,d} | F).$$

- $P(r_{u,i} \geq r_{u,j} | F) = \sigma(F_u \cdot F_i^T - F_u \cdot F_j^T)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, F_i is the i-th row of F.
- The final objective function is:

$$\sum_u \lambda(u) [\sum_{i \in N^+(u), j \in S_k(u)} lnP(r_{u,i} \geq r_{u,j} | F) + \sum_{j \in S_k(u), d \in N^-(u)} lnP(r_{u,j} \geq r_{u,d} | F)] + \lambda_r ||F||_F.$$

relative influence coefficient

regularization term

Parameter learning

- Projected stochastic gradient descent is used for parameter learning
  - $\Theta \leftarrow \max(0, \Theta - \alpha \frac{\partial l}{\partial \Theta})$, Θ is any parameter and α is learning rate
- In each iteration, we need to sample a set of quadruples (u, i, j, d):
  - pre-process the whole graph to record $S_k(u)$ for each u
  - sample node u from V uniformly at random
  - sample node i from $N^+(u)$ uniformly at random
  - sample node j from $S_k(u)$ uniformly at random
  - sample node d from $N^-(u)$ uniformly at random until $d \notin S_k(u)$

Experiments and Results

Datasets

- 6 UMich datasets (10^1 to 10^3 nodes, without ground-truth)
- 3 SNAP datasets (10^5 to 10^6 nodes, with ground-truth)

Metrics

- Modified modularity (M):

$$\frac{1}{2m} \sum_{u,v \in V} \left(g_{u,v} - \frac{d(u)d(v)}{2m} \right) |C_u \cap C_v|$$

- F-1 score (F_1): datasets with ground-truth communities only

Results

Dataset	SCP	LC	BNMF	BNMTF	BigCLAM	PNMF	LNMF(RI)
Dolphins	0.305	0.654	0.507	0.507	0.423	0.979	**1.086(10.9%)**
Les Misérables	0.307	0.773	0.125	0.103	0.540	1.103	**1.184(7.3%)**
Books about US politics	0.496	0.851	0.461	0.492	0.529	0.864	**1.270(47.0%)**
Word adjacencies	0.071	0.271	0.254	0.268	0.231	0.668	**0.701(4.9%)**
American College football	0.605	0.891	0.558	0.573	0.518	1.049	**1.235(17.7%)**
Coauthorships in network science	0.729	0.956	0.661	0.741	0.503	1.657	**2.310(39.4%)**

Table 4: **Comparison in terms of modularity. RI: Relative Improvement over PNMF.**

Dataset	BigCLAM	PNMF	LNMF(RI)
DBLP	0.039	0.098	**0.107(9.2%)**
Amazon	0.044	0.042	**0.048(11.4%)**
YouTube	0.019	**0.060**	0.057(0.0%)

Table 5: **Experimental results on SNAP datasets in terms of F_1 score. RI: Relative Improvement over PNMF.**

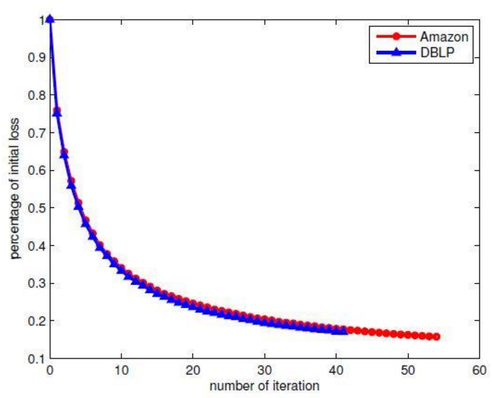


Figure 2: **Convergence speed of learning algorithm**