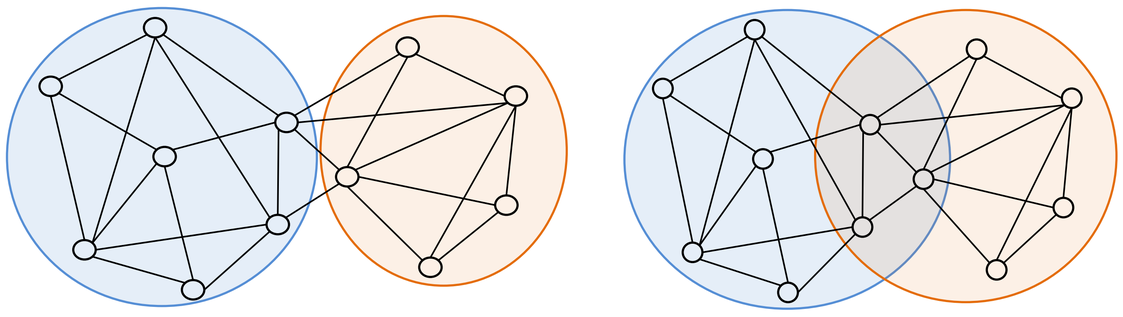
Incorporating Implicit Link Preference Into Overlapping Community Detection

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

Motivation



Why overlapping community detection (right)?

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

Previous methods are not good enough

- They only focus on a link itself, but ignore the preference information a link can carry.
- For community affiliation based approaches, their objective functions all penalize non-linked pairs inside one community equally as those between communities, which is not fair.

A nature idea:

- A node are more likely to build links with other nodes in the same community than those outside its community.

Setup and Notations

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

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

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

- $N^+(i)$ denotes the set of i's neighbors
- $N^-(i)$ denotes the set of i's "non-neighbors"
- Learning set $S = \{(\mathrm{i}, \mathrm{j}, \mathrm{k}) | \mathrm{j} \in N^+(i), k \in N^-(i)\}$

The PNMF Model

Model assumptions

- Communities exist before links and links are built according to each node's preference (recall our last motivation).
- The preferences can be observed from the given graph.
- Node independence, higher preference on neighbors

Model formulation

- For each node i, we denote the likelihood of the observed preference order given a node-community membership matrix as $p(>_i |F)$
- Then , the objective function is the product of all the nodes:

$$\max_{F} \prod_{i \in V} p(>_i |F) = \max_{F} \prod_{(i,j,k) \in S} p(j >_i k|F)$$

- Define $p(j >_i k|F) = \sigma(F_i \cdot F_j^T - F_i \cdot F_k^T)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.
- By denoting $F_i \cdot F_j^T$ as $\hat{x}(i,j)$ and adding a regularization term, the final objective function l is

$$\begin{aligned} l(F) &= ln \prod_{i \in V} p(j >_i k|F) - \frac{\lambda}{2} ||F||_F \\ &= \sum_{(i,j,k) \in S} \ln p(j >_i k|F) - \frac{\lambda}{2} ||F||_F \\ &= \sum_{(i,j,k) \in S} \ln \sigma(\hat{x}(i,j) - \hat{x}(i,k)) - \frac{\lambda}{2} ||F||_F \end{aligned}$$

Parameter Learning

Learning process

- Since F needs to be non-negative, we use projected gradient descent to solve our optimization problem:

$$\Theta \leftarrow \max(0, \Theta - \alpha \frac{\partial l}{\partial \Theta}),$$

  where Θ is any parameter and α is learning rate.
- Stochastic gradient descent is adopted to deal with large datasets.
- Sampling strategy for a triple (i, j, k):
  - Sample i from node set uniformly at random
  - Sample j from $\mathrm{N}^+(i)$ uniformly at random
  - Sample k from $N^-(i)$ uniformly at random
- Time complexity for each iteration is O(mp), where m is the number of links and p is the number of communities.

Hyper-parameter we need to tune:

- p – number of communities
- δ – community membership threshold

Experiments and Results

Dataset

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

Baseline Methods

- Dense subgraph extraction based methods
  - SCP , LC
- Matrix Factorization based methods
  - BNMF, BNMTF, BigCLAM

Evaluation

- Modified modularity (M):

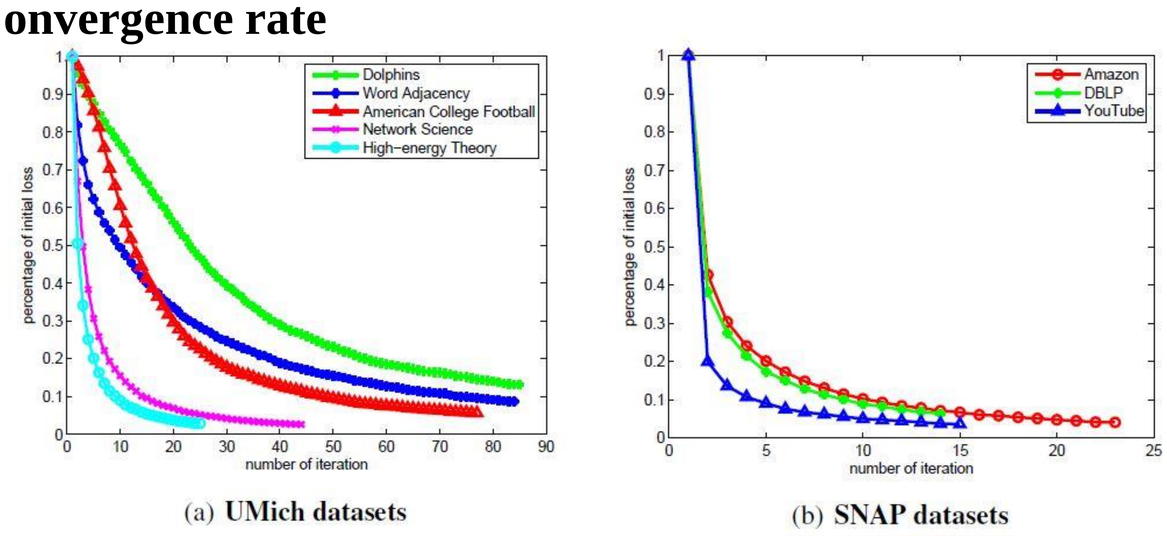
$$\frac{1}{2m} \sum_{u,v \in V} \left(g_{u,v} - \frac{d(u)d(v)}{2m} \right) |C_u \cap C_v|$$

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

Results

Dataset	Metric	SCP	LC	BNMF	BNMTF	BigCLAM	PNMF
Dolphins	M	0.3049	0.6538	0.5067	0.5067	0.4226	**0.9787**
Les Misérables	M	0.3066	0.7730	0.1247	0.1031	0.5395	**1.1028**
Books about US politics	M	0.4955	0.8507	0.4613	0.4924	0.5290	**0.8640**
Word adjacencies	M	0.0707	0.2705	0.2539	0.2677	0.2312	**0.6680**
American college football	M	0.6050	0.8907	0.5584	0.5733	0.5175	**1.0492**
Jazz musicians	M	0.0114	1.1424	0.1133	0.1118	**1.1438**	0.9357
Network science	M	0.7286	0.9558	0.6607	0.7413	0.5026	**1.6570**
Power grid	M	0.0439	0.3713	0.3417	0.3682	1.0097	**1.1051**
High-energy theory	M	0.5427	**0.9965**	0.5648	0.6004	0.9636	0.9725
DBLP	F_1	0.0967	0.0402	-	-	0.0390	**0.0985**
Amazon	F_1	0.0315	0.0070	-	-	**0.0441**	0.0419
YouTube	F_1	0.0445	-	-	-	0.0194	**0.0605**

Convergence rate



(a) **UMich datasets** (b) **SNAP datasets**

References

- SCP: Kumpula et al. Sequential algorithm for fast clique percolation. Physical Review E, 2008
- LC: Ahn et al. Link communities reveal multiscale complexity in networks. Nature, 2010.
- BNMF: Psorakis et al. Overlapping community detection using Bayesian non-negative matrix factorization. Physical Review E, 2011.
- BNMTF: Zhang and Yeung. Overlapping community detection via bounded nonnegative matrix tri-factorization. KDD 2012.
- BigCLAM: Yang and Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. WSDM 2013.