Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Incorporating Implicit Link Preference Into Overlapping Community Detection

Hongyi Zhang[1,2], Irwin King[1,2], Michael R. Lyu[1,2]
[1]Shenzhen Key Laboratory of Rich Media Big Data Analytics and Applications,
Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
[2]Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{hyzhang, king, lyu}@cse.cuhk.edu.hk

Abstract

Community detection is an important technique to understand structures and patterns in complex networks. Recently, overlapping community detection becomes a trend due to the ubiquity of overlapping and nested communities in real world. However, existing approaches have ignored the use of implicit link preference information, i.e., links can reflect a node's preference on the targets of connections it wants to build. This information has strong impact on community detection since a node prefers to build links with nodes inside its community than those outside its community. In this paper, we propose a preference-based nonnegative matrix factorization (PNMF) model to incorporate implicit link preference information. Unlike conventional matrix factorization approaches, which simply approximate the original adjacency matrix in value, our model maximizes the likelihood of the preference order for each node by following the intuition that a node prefers its neighbors than other nodes. Our model overcomes the indiscriminate penalty problem in which non-linked pairs inside one community are equally penalized in objective functions as those across two communities. We propose a learning algorithm which can learn a node-community membership matrix via stochastic gradient descent with bootstrap sampling. We evaluate our PNMF model on several real-world networks. Experimental results show that our model outperforms state-of-the-art approaches and can be applied to large datasets.

1 Introduction

Discovering the community structure in complex networks has been extensively investigated in the past decade (Fortunato 2010). A community is intuitively regarded as a group of nodes with more links inside the group than between its member and outside the group (Girvan and Newman 2002). In real world, communities can be social circles manually categorized by users in ego networks (McAuley and Leskovec 2012), authors from the same institution in collaboration networks (Newman 2001), proteins with the same functionality in biochemical networks (Gavin et al. 2002), etc. The research issue of finding such groups is known as the *community detection* problem.

Classic methods in community detection assume that one node belongs to exactly one community. However, many complex networks we encounter in daily life allow multiple memberships. For example, two colleagues in the same department are also in the same company (nested), one can join in several discussion groups in an online forum (overlapping), etc. Thus, the topic of overlapping community detection has attracted major attention recently (Xie, Kelley, and Szymanski 2013).

Existing overlapping community detection approaches can be categorized into two classes: one is based on dense subgraph extraction (Ahn, Bagrow, and Lehmann 2010; Palla et al. 2005; Kumpula et al. 2008), which uses certain criteria to find overlapping dense subgraphs or clusters in the network to be communities; the other is based on community affiliation model (Psorakis et al. 2011; Wang et al. 2011; Yang and Leskovec 2013; Zhang and Yeung 2012), which determines the number of communities in advance and assigns each node to multiple communities according to some optimization function. However, both classes of approaches only focus on links themselves but ignore the implicit preference information in links. In fact, a link can reflect the preferences of both sides to some extent. For example, in a social network, if user A wants to make friend with user B, a typical way for A is to send a friend invitation to B and wait for him to accept it. They cannot be friends if either step goes wrong. Thus, when we see the fact that A and B are friends, it is reasonable to argue that A prefers B than other strangers to be his friend. Assuming B also receives other people's invitations and only accepts a few of them (this is very likely to happen in real world), we can also argue that B prefers A than others who are still strangers to him. Following the intuition that a node are more likely to build links with other nodes in the same community than those outside its community, the implicit preference information can be helpful for community detection.

For the second class of approaches, i.e., community affiliation based approaches, nonnegative matrix factorization (NMF) has been applied as a standard technique. The basic idea of NMF technique is to find a node-community membership matrix F ($F_{u,c}$ represents the weight of node u in community c) and approximate the adjacency matrix G via

FF^T. Existing approaches use either the conventional least squares error or the generalized KL divergence as objective function (Lee and Seung 2001). However, both objective functions try to approximate the adjacency matrix G in value, which are inevitable to cause the indiscriminate penalty problem. Let us assume that there are two non-linked pairs (i, j) and (i, k), where i, j belong to the same community while i, k do not. Since i, j both have positive weights in some community c, $F_i F_j^T$ is positive. However, existing NMF based approaches will penalize $F_i F_j^T$ for being positive since $G_{i,j} = 0$. Thus, there is no difference between j and k with respect to i, which is against the intuition that for node i, node j in the same community is more preferable than node k outside i's community. In fact, it is reasonable that $F_i F_j^T$ is higher than $F_i F_k^T$, and indiscriminately penalizing the two pairs are problematic.

In this paper, we present a preference-based nonnegative matrix factorization (PNMF) model that not only fixes the indiscriminate penalty problem of previous NMF based models, but also incorporates the implicit link preference information into model formulation. Our model uses a new objective function, which maximizes the likelihood of a pairwise preference order for each node. In other words, from a node's perspective, we manage to ensure that the preference of any of its friends is higher than any of other nodes. When factorizing the adjacency matrix with node-community relationship matrix, our model gives no penalty to a non-zero value appearing in the position of a non-linked pair, as long as all the pair-wise preferences are preserved. Thus, this objective function can be regarded as a relaxation of previous approaches. We exploit stochastic gradient descent with bootstrap sampling to solve the optimization problem. We conduct experiments in several real world datasets including some with ground-truth communities. By comparing our model with several state-of-art approaches, we show that our model can detect overlapping communities with higher quality on widely-used metrics in community detection. It can also be applied to large datasets.

The remainder of this paper is organized as follows. Section 2 provides an overview of overlapping community detection and other works related to our model. Section 3 presents our PNMF model in mathematical form and propose a stochastic gradient descent method to learn model parameters, i.e., community weights for each node. Experimental results are shown in Section 4, followed by conclusions in Section 5.

2 Related Work

Traditional community detection methods are based on graph clustering/partitioning (Fortunato 2010; Newman 2004). To measure the quality of a particular partition, different functions are created. Among them, modularity is the most popular one (Newman 2006). However, the most severe drawback of these methods is that each node can only be assigned to a single community, which is against the nature of the real world. Thus, overlapping community detection has attracted lots of attention recently due to its practical assumption (Xie, Kelley, and Szymanski 2013).

Almost all the early work in overlapping community detection are based on dense subgraph extraction. Clique Percolation method (CPM) (Palla et al. 2005), which is still one of the most popular methods now, finds all the k-cliques at first and then combines cliques with $k-1$ common nodes to be a community. Kumpula et al. propose a sequential alternative of CPM, which significantly reduces the time of finding cliques (Kumpula et al. 2008). Link clustering (Ahn, Bagrow, and Lehmann 2010) uses links instead of nodes for clustering and consequently generates overlapping communities.

Community affiliation based approaches solve overlapping community detection from the perspective of communities. They determine the number of communities in advance, then assign nodes to each community based on some criterion. Some approaches apply *Markov Chain Monte Carlo (MCMC)* to iteratively change node membership until a local optimal is reached (Karrer and Newman 2011; Yang and Leskovec 2012). However, the performance is not good enough and they suffer from convergence problems. Recently, *nonnegative matrix factorization (NMF)* has become a standard technique since the latent factor can be naturally regarded as communities and it automatically generates overlapping communities. Psorakis et al. employ a Baysian generative model, which puts a half-normal prior β_i over each community and then maximizes the log-likelihood of generating original graph (Psorakis et al. 2011). Want et al. apply a multiplicative update rule to minimize $||G - FF^T||_F^2$, where G is the adjacency matrix and F is the targeted node-community membership matrix (Wang et al. 2011). Zhang and Yeung incorporate a community interaction matrix B, which represents the relations between communities and approximate G by a nonnegative matrix trifactorization FBF^T with bounds in F (Zhang and Yeung 2012). Yang and Leskovec develop a scalable model, which relaxes the graph fitting problem into a continuous optimization problem (Yang and Leskovec 2013). Though different objective functions have been engaged, the idea of these NMF-based methods is always to approximate the original adjacency matrix in value.

Baysian Personalized Ranking (BPR) (Rendle et al. 2009b) is proposed to rank items for a specific user in recommender systems while only implicit feedback (e.g. clicks) is available. The basic assumption is that a user prefers labeled items than unlabeled ones. While traditional methods replace missing values with zeros or negative ones, BPR uses pairwise preference as training data to learn the model parameters. Technically, it maximizes a posterior probability $p(\Theta| >_u)$ where Θ is the parameter and $>_u$ is the latent preference structure for user u. We adopt this idea into our overlapping community detection task for a different learning goal of maximizing the probability $p(>_u |F)$, where F is a nonnegative matrix representing the latent node-community membership. BPR has become a classical model in one-class collaborative filtering and there are many further work on top of it. For example, Rendle et al. extend the original matrix factorization to a tensor factorization to recommend personalized tags for a user given an item (Rendle et al. 2009a). Zhao et al. leverage social connections to improve item rec-

ommendations by building a new preference system (Zhao, McAuley, and King 2014).

3 Community Detection via PNMF

In this section, we present our PNMF model in the context of overlapping community detection and propose a stochastic gradient descent method with bootstrap sampling to learn model parameters.

Preliminaries

Given an unweighted and undirected network $N(V, E)$, where V is the set of n nodes and E is the set of m edges, we can obtain its adjacency matrix $G \in \{0, 1\}^{n\times n}$ whose (i, j) entry $g_{i,j}$ is an indicator of whether node i and node j are connected. Since the network is undirected, G is a symmetric matrix.

We denote the set of communities by C and the number of communities by p. We use a nonnegative matrix $F \in \mathbb{R}_+^{n\times p}$ to denote the node-community membership for all the nodes. Each entry $F_{u,c}$ represents the weight between node $u \in V$ and community $c \in C$. The larger $F_{u,c}$ is, the more possible that u belongs to c. On the other hand, if $F_{u,c}$ is 0, u does not belong to c.

Given the information above, the objective is to recover G with some properties preserved by a nonnegative matrix factorization FF^T, i.e.,

$$G \approx FF^T. \tag{1}$$

Previous approaches simply approximate G in value. They expect $F_u F_v^t$ to be close to 1 if u, v are linked and to be 0 otherwise. In our model, we preserve the preference orders observed in G for all the nodes. We will discuss the details later.

The set of i's neighbors is denoted by $N^+(i)$. In addition, we define $N^-(i) := N^+(i)^c \backslash \{i\}$ to be "non-neighbors" of i, where $N^+(i)^c$ denotes the complement set of $N^+(i)$. By definition, $V = N^+(i) \cup N^-(i) \cup \{i\}$ for every i. Moreover, we define a learning set $S : V \times V \times V$ by

$$S = \{(i, j, k)|i \in V, j \in N^+(i), k \in N^-(i)\},$$

which consists of all the triples (i, j, k), where j is a neighbor of i while k is not.

In the end, we list three basic assumptions on implicit link preference in order to make model formulation clearer.

1. **Node independence**. Each node determines its preferences independently. The network can be regarded as a result after all the nodes make their decisions. Specifically, a link will be built between u and v if and only if u has a high preference on v and symmetrically v has a high preference on u.
2. **Higher preference on neighbors**. Let $u >_i v$ denote that node i prefers node u than node v. For a fixed node i, we have $j >_i k$ if $j \in N^+(i)$ and $k \in N^-(i)$, but no preference information between j and k is indicated if $j, k \in N^+(i)$ or $j, k \in N^-(i)$. Thus, the use of the learning set S is to record all the single triples (i, j, k) satisfying that i prefers to build a link with j than k.
3. **Pair independence**. For a fixed node i, its preference on j and k is independent with its preference on u and v when $j, u \in N^+(i)$ and $k, v \in N^-(i)$.

Model Formulation

Based on our motivation, we aim to find the node-community membership matrix, which maximizes the likelihood of observed preference order for all the nodes. According to the "node independence" assumption, the overall likelihood can be denoted as a product of likelihood of each node. Thus, our objective function can be written as

$$\max_{F\in\mathbb{R}_+^{n\times p}} \prod_{i\in V} p(>_i |F), \tag{2}$$

where $>_i$ denotes the observed preferences for node i and F is the node-community membership matrix.

According to the "high preference on neighbors" assumption and the "pair independence" assumption, the probability of preference order for a single node i can be written as

$$\begin{aligned} p(>_i |F) = &\prod_{(j,k)\in V\times V} p(j >_i k|F)^{\delta(j\in N^+(i))\delta(k\in N^-(i))} \\ &\cdot (1 - p(j >_i k|F))^{1-\delta(j\in N^+(i))\delta(k\in N^-(i))} \\ = &\prod_{(j,k)\in V\times V} p(j >_i k|F)^{\delta((i,j,k)\in S)} \\ &\cdot (1 - p(j >_i k|F))^{\delta((i,j,k)\notin S)}, \end{aligned} \tag{3}$$

where S is the learning set mentioned in preliminaries and δ is the indicator function

$$\delta(a) = \begin{cases} 1 & \text{if } a \text{ is true,} \\ 0 & \text{else} \end{cases}.$$

For a triple (i, j, k), if $(i, j, k) \in S$, then $(i, k, j) \notin S$. Given $p(j >_i k|F) + p(k >_i j|F) = 1$, it is easy to see that $p(j >_i k|F)^{\delta((i,j,k)\in S)} = (1 - p(k >_i j|F)^{\delta((i,k,j)\notin S)}$. Applying this to Equation (3), maximizing $p(>_i |F)$ is equivalent to

$$\max_{F\in\mathbb{R}_+^{n\times p}} \prod_{(j,k)\in V\times V} p(j >_i k|F)^{\delta((i,j,k)\in S)}. \tag{4}$$

Combining Equation (2) and (4), our objective function can be rewritten as

$$\max_{F\in\mathbb{R}_+^{n\times p}} \prod_{(i,j,k)\in S} p(j >_i k|F). \tag{5}$$

Based on the intuition that two nodes have a higher probability to be linked if they share more communities, we define the probability that i prefers j than k given the node-community membership matrix as

$$p(j >_i k|F) = \sigma(F_i \cdot F_j^T - F_i \cdot F_k^T), \tag{6}$$

where σ is the sigmoid function $\sigma(x) := \frac{1}{1+e^{-x}}$.

The sigmoid function can map any real number into $(0, 1)$. We can see that the probability i prefers j than k is 0.5 when $F_i F_j^T = F_i F_k^T$. Also, this probability is close to

0 when $F_i F_j^T \ll F_i F_k^T$ and is close to 1 when $F_i F_j^T \gg F_i F_k^T$. These properties precisely characterize the requirements of our model.

For simplicity, we define $\hat{x}(i,j) := F_i \cdot F_j^T$. Equation (6) can be rewritten as

$$p(j >_i k|F) = \sigma(\hat{x}(i,j) - \hat{x}(i,k)). \tag{7}$$

Now combining Equation (5), (6), and (7), the final objective function of our PNMF model is

$$\begin{aligned} l(F) &:= \max_{F \in \mathbb{R}_+^{n \times p}} \ln \prod_{(i,j,k) \in S} p(j >_i k|F) - \lambda \cdot reg(F) \\ &= \max_{F \in \mathbb{R}_+^{n \times p}} \sum_{(i,j,k) \in S} \ln p(j >_i k|F) - \lambda \cdot reg(F) \\ &= \max_{F \in \mathbb{R}_+^{n \times p}} \sum_{(i,j,k) \in S} \ln \sigma(\hat{x}(i,j) - \hat{x}(i,k)) - \lambda \cdot reg(F), \end{aligned} \tag{8}$$

where $reg(F)$ is the regularization term we add to avoid overfitting and λ is the regularization parameter. We choose Frobenius norm as the regularization term, i.e., we set $reg(F) = ||F||_F^2$, since it is differentiable and fits our parameter learning process.

Parameter Learning

To make our model applicable to large datasets, we employ the widely used stochastic gradient descent (SGD) as our learning approach. In each update step, SGD randomly selects a triple in learning set S and updates the corresponding model parameters Θ by walking along the gradient direction,

$$\Theta^{t+1} = \Theta^t + \alpha \frac{\partial l}{\partial \Theta}, \tag{9}$$

where α is the learning rate. Specifically, the derivative of Equation (9) is calculated by

$$\begin{aligned} \frac{\partial l}{\partial \Theta} &= \frac{\partial}{\partial \Theta} \ln \sigma(\hat{x}(i,j) - \hat{x}(i,k)) - \lambda \frac{\partial}{\partial \Theta} reg(F) \\ &= \frac{-e^{\hat{x}(i,k) - \hat{x}(i,j)}}{1 + e^{\hat{x}(i,k) - \hat{x}(i,j)}} \cdot \frac{\partial}{\partial \Theta}(\hat{x}(i,j) - \hat{x}(i,k)) - \lambda \Theta \end{aligned} \tag{10}$$

and

$$\frac{\partial}{\partial \Theta}(\hat{x}(i,j) - \hat{x}(i,k)) = \begin{cases} F_{j,t} - F_{k,t} & \text{if } \Theta = F_{i,t} \\ F_{i,t} & \text{if } \Theta = F_{j,t} \\ -F_{i,t} & \text{if } \Theta = F_{k,t} \\ 0 & \text{else} \end{cases}, \tag{11}$$

where λ is the regularization parameter. Regarding the nonnegative constraints, we exploit the idea of projected gradient methods for NMF (Lin 2007), which maps the value of a parameter back to nonnegativity.

The whole process of parameter learning is described in Algorithm 1. As we can see, the time complexity of each iteration is $O(mp)$, where m is the number of links, and p is the number of community. The space complexity is $O(np)$, where n is the number of node, since we need to save the node-community membership matrix into memory.

Algorithm 1 Community detection using PNMF

Input: G, the adjacency matrix of original graph
Output: F, the node-community membership matrix

```
initialize F
compute initial loss
repeat
    for num_samples = 1 to |E| do
        sample node i from V uniformly at random
        sample node j from N+(i) uniformly at random
        sample node k from N−(i) uniformly at random
        for each entry Θ in F_i, F_j and F_k do
            update Θ according to Equation (9), (10), (11)
            Θ ← max(Θ, 0)
        end for
    end for
    compute loss
until convergence or max_iter is reached
```

Other Issues

Choosing the number of communities. Before learning the parameters, we need to set the number of communities p in advance. However, we have no prior knowledge about it. Here we adopt the approach in (Airoldi et al. 2008). We first reserve 10% of links as validation set. Then we vary p and learn model parameters with the remaining 90% of links for each p. After that, we use the node-community membership matrix F to generate the adjacency matrix G and use G to predict the links in validation set according to our motivation that linked pairs have higher value than non-linked pairs in G. Finally, we pick the p with the best prediction score as our pre-assigned number of communities.

Setting membership threshold. After we learn F, we need to set a threshold δ in order to determine whether a node belongs to a community or not. If $F_{u,c} \geq \delta$, we say that node u belongs to community c. According to Equation (6), we need $p(j >_i k|F)$ to be closer to 1 than 0 if i prefer j than k. We assume that i, j share exactly one community and i, k do not share any communities. Thus $F_i F_k^T = 0$. Due to the symmetry of i and j, we have

$$\sigma(F_i F_j^T - F_i F_k^T) = \sigma(\delta^2 - 0) = \frac{1}{1 + e^{-\delta^2}} = \beta,$$

where β is in the range of $(0.5, 1)$. When β is given, we can compute δ by

$$\delta = \sqrt{-\ln(\frac{1}{\beta} - 1)}. \tag{12}$$

4 Experiments

In this section, we conduct several experiments to compare our PNMF model with other state-of-the-art overlapping community detection approaches in terms of community quality and scalability.

Datasets

We examine our model with several benchmark datasets available on the Internet. We separate them into two categories, one without ground-truth communities and the

Dataset	GT	V	E
Dolphins	N	62	159
Les Misérables	N	77	254
Books about US politics	N	105	441
Word adjacencies	N	112	425
American college football	N	115	613
Jazz musicians	N	198	2,742
Network science	N	1,589	2,742
Power grid	N	4,941	6,594
High-energy theory	N	8,361	15,751
DBLP	Y	317,080	1,049,866
Amazon	Y	334,863	925,872
YouTube	Y	1,134,890	2,987,624

Table 1: **Statistics of twelve datasets (nine without ground-truth and three with ground-truth). GT: whether has ground-truth communities or not, V: number of nodes, E: number of links.**

other with ground-truth communities. For the first category, we choose nine undirected networks collected by Newman[1] as our datasets. For the second category, three large datasets from SNAP[2] are used. Among them, *DBLP* is a co-authorship network in computer science, *Amazon* is a product co-purchase network, *YouTube* is an online social network with communities of various video interests. Simple statistics for all the datasets can be found in Table 1.

Baseline methods

We select five state-of-the-art algorithms to be our baseline methods. The latter three are nonnegative matrix factorization based models, thus are highly comparable with our PNMF model.

SCP (Sequential Clique Percolation) (Kumpula et al. 2008). Since the original Clique Percolation method (Palla et al. 2005) is slow when dealing with large datasets, we choose a sequential alternative, which obtains the same performance but is much faster. For the choice of k-clique, we set k to be 4 or 5.

LC (Link Clustering) (Ahn, Bagrow, and Lehmann 2010). We do not manually set the threshold at which the dendrogram is cut. The algorithm automatically chooses the threshold where the maximum partition density is found. Among the detected communities, we get rid of all the communities whose size is smaller than 3 since these communities make no sense.

BNMF (Baysian NMF) (Psorakis et al. 2011). We use the classic squared loss $||G - WH^T||_F^2$ as the loss function, where G is the adjacency matrix, W and H are the results of nonnegative matrix factorization.

BNMTF (Bounded NM Tri-Factorization) (Zhang and Yeung 2012). To be consistent with BNMF, we also use squared loss $||G - FBF^T||_F^2$ as our loss function, where F and B are the results of nonnegative matrix tri-factorization.

BigCLAM (Yang and Leskovec 2013). For the number of communities, we set a minimum value and a maximum value and let the algorithm find the best choice between these two numbers based on cross-validation.

Metrics

We choose two well-known metrics to measure the performance of our model. The choice of metric depends on whether the specific dataset has ground-truth communities.

Modularity. We employ the most widely used modularity (Newman 2006) as our measure for datasets without ground-truth communities. Since communities are overlapping in our case, we need to modify the original definition of modularity a bit. The new modularity Q is defined as

$$Q = \frac{1}{2m} \sum_{u,v \in V} (g_{u,v} - \frac{d(u)d(v)}{2m}) |C_u \cup C_v|,$$

where m denotes the number of links, V denotes the set of nodes, $g_{i,j}$ denotes the (i, j) entry of adjacency matrix G, $d(i)$ denotes the degree of node i, and C_u denotes the set of communities including u.

As we mentioned, two nodes are more possible to link each other if they have more common communities. Modularity matches our intuition very well in the way that more common communities two nodes have, more penalty they will receive if they do not build a link between them. $\frac{d(u)d(v)}{2m}$ can be regarded as the link probability between u and v.

$\mathbf{F_1}$ score. For datasets with ground-truth communities, we employ another criterion F_1 score to measure the quality of detected communities. We denote the set of ground-truth communities as C and the set of detected communities as $\hat{C}$. C_i represents the i-th community in C and $\hat{C}_i$ represents the i-th community in $\hat{C}$. We define F_1 score to be the average of the F_1 score of the best-matching ground-truth community to each detected community, i.e,

$$F_1 = \frac{1}{|\hat{C}|} \sum_{\hat{C}_i \in \hat{C}} F_1(C_{b(i)}, \hat{C}_i),$$

where the best matching function $b(i)$ is defined as

$$b(i) = \arg\max_j F_1(C_j, \hat{C}_i),$$

and $F_1(\cdot, \cdot)$ is the harmonic mean of precision and recall.

Results

We compare our PNMF model with all the baseline methods listed above and show the results in Table 2. For the first nine datasets without ground-truth communities, we use modularity as our measurement. The results show that our model performs best on seven out of nine datasets. Especially, our model dominates other nonnegative matrix factorization based models (BNMF, BNMTF, BigCLAM) on all the datasets except "Jazz musicians".

For the last three datasets with ground-truth communities, we use F_1 score as our measurement. We can see that our model significantly outperforms LC and is comparable with the other two methods with a fair overall advantage. Another advantage of our model is scalability. Some results are

[1] http://www-personal.umich.edu/mejn/netdata

[2] http://http://snap.stanford.edu/data/

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**

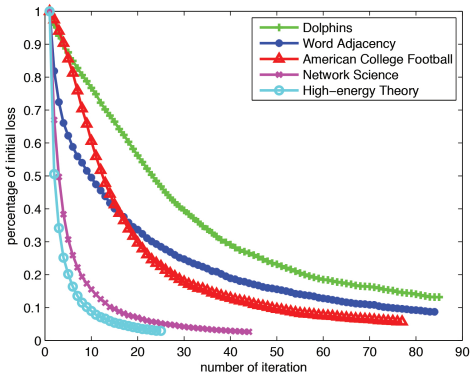
Table 2: **Experimental results in terms of modularity (M) and $\mathbf{F_1}$ score (F_1).**

not shown because the corresponding baseline methods cannot scale to networks with such size. Only SCP, BigCLAM and our PNMF model can deal with the largest dataset, i.e. "YouTube", which consists of more than one million nodes.

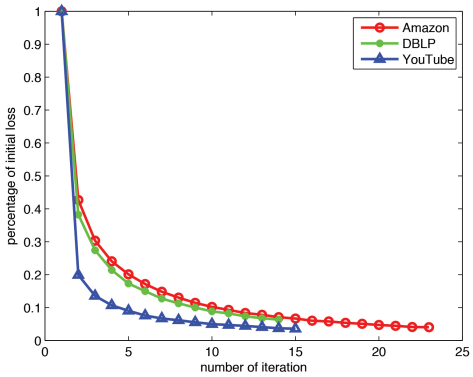
For the choice of membership threshold β, we examine different values from 0.5 to 1 to find a reasonable range. It is clear that a community will contain less nodes if we set a higher value to β. According to our experiments, $[0.7, 0.8]$ appears to be a suitable range for candidates since a community may contain nearly half of the nodes when β is less than 0.7, while many nodes may not belong to any communities when β is larger than 0.8. To determine the final value of β, we again use the cross-validation paradigm with several candidates in this range and pick the one with the best performance on validation data.

Convergence Issues

Since our PNMF model applies stochastic gradient descent as learning technique, we also observe convergence rate and convergence speed while conducting experiments. For convergence rate, as long as the learning rate and the regularization parameter are appropriate, all the datasets can converge before reaching maximum number of iteration. For convergence speed, Figure 1(a) show the results on five UMich datasets and Figure 1(b) show the results of three SNAP datasets. Here the y-axis represents the ratio of current loss to initial loss. From both figures we can see that loss drops quickly in the beginning and starts to slow down significantly after it reaches 20% of initial loss. Comparing these two figures, we can also find that, although SNAP datasets need more time for one iteration than UMich datasets, the total number of iteration is smaller, which proves the scalability of our model from another perspective.



(a) **UMich datasets**



(b) **SNAP datasets**

Figure 1: **Convergence speed of learning algorithm**

5 Conclusion and Future Work

In this paper, we have presented a preference-based nonnegative matrix factorization (PNMF) model for overlapping community detection. The most important contribution of our model is to incorporate implicit link preference information into model formulation. By following the intuition that a node prefers any of its neighbors than any of its "non-neighbors", we maximize the likelihood of a preference order for each node instead of simply approximating the original adjacency matrix in value. Our model can eliminate the unreasonable indiscriminate penalty on pairs inside and between communities. In the learning process, we choose stochastic gradient descent with bootstrap sampling to learn the parameters of node-community member-

ship matrix. We apply our PNMF model on several real-world datasets both with and without ground-truth communities. Our results show that our PNMF model outperforms state-of-art approaches in two metrics, namely modularity and F_1 score, and is scalable for large datasets.

Our current work only focuses on the difference between neighbors and "non-neighbors". We assume that all the "non-neighbors" have the same preference. However, this assumption may not hold in real-world networks. Considering two nodes A and B with no link between them, if there are other nodes which are neighbors of both A and B, from the perspective of A it is reasonable to assign higher preference on B than nodes which have no common neighbors with A. We plan to employ the concept of common neighbors to enhance our preference system in our future work.

Acknowledgement

The work described in this paper was fully supported by the National Grand Fundamental Research 973 Program of China (No. 2014CB340401 and No. 2014CB340405), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413212 and CUHK 415113), and Microsoft Research Asia Regional Seed Fund in Big Data Research (Grant No. FY13-RES-SPONSOR-036).

References

Ahn, Y.-Y.; Bagrow, J. P.; and Lehmann, S. 2010. Link communities reveal multiscale complexity in networks. *Nature* 466(7307):761–764.

Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; and Xing, E. P. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9(1981-2014):3.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486(3):75–174.

Gavin, A.-C.; Bösche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A.-M.; Cruciat, C.-M.; et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868):141–147.

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12):7821–7826.

Karrer, B., and Newman, M. E. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1):016107.

Kumpula, J. M.; Kivelä, M.; Kaski, K.; and Saramäki, J. 2008. Sequential algorithm for fast clique percolation. *Physical Review E* 78(2):026109.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562.

Lin, C.-J. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19(10):2756–2779.

McAuley, J., and Leskovec, J. 2012. Learning to discover social circles in ego networks. In *NIPS*, volume 272, 548–556.

Newman, M. E. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* 98(2):404–409.

Newman, M. E. 2004. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems* 38(2):321–330.

Newman, M. E. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23):8577–8582.

Palla, G.; Derényi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435(7043):814–818.

Psorakis, I.; Roberts, S.; Ebden, M.; and Sheldon, B. 2011. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E* 83(6):066114.

Rendle, S.; Balby Marinho, L.; Nanopoulos, A.; and Schmidt-Thieme, L. 2009a. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 727–736. ACM.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009b. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461. AUAI Press.

Wang, F.; Li, T.; Wang, X.; Zhu, S.; and Ding, C. 2011. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery* 22(3):493–521.

Xie, J.; Kelley, S.; and Szymanski, B. K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)* 45(4):43.

Yang, J., and Leskovec, J. 2012. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 1170–1175. IEEE.

Yang, J., and Leskovec, J. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 587–596. ACM.

Zhang, Y., and Yeung, D.-Y. 2012. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 606–614. ACM.

Zhao, T.; McAuley, J.; and King, I. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 261–270. ACM.