Modeling the Homophily Effect between Links and Communities for Overlapping Community Detection

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

Abstract

Overlapping community detection has drawn much attention recently since it allows nodes in a network to have multiple community memberships. A standard framework to deal with overlapping community detection is *Matrix Factorization (MF)*. Although all existing MF-based approaches use links as input to identify communities, the relationship between links and communities is still under-investigated. Most of the approaches only view links as consequences of communities (community-to-link) but fail to explore how nodes' community memberships can be represented by their linked neighbors (link-to-community). In this paper, we propose a *Homophily-based Nonnegative Matrix Factorization (HNMF)* to model both-sided relationships between links and communities. From the community-to-link perspective, we apply a preference-based pairwise function by assuming that nodes with common communities have a higher probability to build links than those without common communities. From the link-to-community perspective, we propose a new community representation learning with network embedding by assuming that linked nodes have similar community representations. We conduct experiments on several real-world networks and the results show that our *HNMF* model is able to find communities with better quality compared with state-of-the-art baselines.

1 Introduction

Network is an abstraction representing relationships among real-world objects. A typical pattern of a network is that there are groups of nodes closely connected within the group but rarely making connections with nodes outside the group. Such groups are defined as *communities* [Girvan and Newman, 2002]. The task of finding such communities from complex networks is referred as *community detection*, an important research topic in web mining for more than a decade. Usually, the more complex a network is, the more challenging it will be to identify such communities. It is mainly due to the infeasibility of visualization and the variety of community structure. Classic graph-partition-based community detection approaches assume that a node belongs to one and only one community, which contradicts with the fact that a node often appears with multiple memberships. To relax this unrealistic constraint, several new algorithms for *overlapping community detection* have been proposed in recent years.

A majority of existing methods for overlapping community detection is based on *Matrix Factorization (MF)* [Psorakis *et al.*, 2011; Wang *et al.*, 2011; Zhang and Yeung, 2012; Zhang *et al.*, 2015], which has been a standard technique in other areas such as recommender systems and natural language processing. The basic idea of MF here is to use low-dimensional latent vectors to represent nodes' features in networks. MF naturally fits into overlapping community detection since the dimensions of factorized latent vectors of nodes can be interpreted as their community membership and hence are no longer latent. The MF-based overlapping community detection can be summarized into three steps: (1) assign the number of communities, (2) compute the node-community weight matrix F through a learning objective, and (3) obtain the final community set according to F. Here the most important part is the selection of learning objective. The simplest way is to recover the adjacency matrix of original network A by F with minimum error, i.e., to minimize $||A - FF^T||$ [Psorakis *et al.*, 2011; Wang *et al.*, 2011]. However, an entry in A is a label (either 0 or 1) whereas an entry in F is a real value. The mismatch between label and entry does not make sense. To fix it, recent approaches such as [Yang and Leskovec, 2013; Zhang *et al.*, 2015] adopt generative objectives, which are based on the intuition that a node is more likely to build a link with another node inside its community than outside.

When we look into this intuition, it implicitly reveals that links are the consequence of communities (community-to-link), i.e., if two nodes share common communities, they will have a higher probability to be linked. However, the investigation in reverse perspective (link-to-community) is largely ignored, i.e., whether a node's community membership can be represented by its neighbors' community membership. Taking MF as an example, the link-to-community

perspective can be interpreted as to learn a node's community representation via the community representations of its neighbors. Here we use the word **homophily**, the tendency of an individual node to associate with similar others [Tang *et al.*, 2013], to recapitulate both perspectives.

In this paper, we propose a *Homophily-based Non-negative MF (HNMF)* to explicitly model the effect of homophily from both perspectives. From the community-to-link perspective, we apply a pairwise objective function in the *Preference-based Non-negative MF (PNMF)* model [Zhang *et al.*, 2015]. From the link-to-community perspective, we develop a novel generative objective function based on unsupervised representation learning and network embedding. We combine both objective functions into a joint learning objective, in which parameter learning can be easily parallelized using asynchronous stochastic gradient descent. Through experiments on various real-world datasets, we demonstrate that our model can identify communities with better quality compared with state-of-the-art baselines and can be applied to large datasets.

Contributions. We summarize our main contribution of this paper as follows,

1. Our work is the first to explore the link-to-community side of homophily effect between links and communities in overlapping community detection. We justify it via observation on real-world datasets with ground-truth communities;
2. We propose a new learning objective to model both perspectives of homophily within the non-negative MF framework. Experiments show that our HNMF model can detect overlapping communities with better quality.

2 Problem Definition and Data Observation

In this section, we first provide several definitions about community and community detection. Then we conduct a data observation on two large real-world networks to strengthen our motivation.

2.1 Problem Definition

Suppose we have a network $G(V, E)$, where V and E are node and edge sets respectively. A community in G is usually considered as a group of densely connected nodes with a common feature, e.g., students from a university, employees from a company, etc. The task of community detection can be defined as follows.

Definition 2.1 (**Community Detection**). *Community detection is a process that takes a network G as input and produce a set of communities S as output to maximize a particular objective function f, i.e.,*

$$\arg\max_{S} f(G, S), \tag{1}$$

where $S = \{C_i | C_i \neq \emptyset, C_i \neq C_j, 1 \leq i, j \leq |S|\}$.

While classic community detection requires that $C_i \neq C_j$ if $i \neq j$, overlapping community detection does not set any constraints on S. This relaxation matches the nature of real-world networks better but brings big challenges since classic garph-partition-based algorithms are no longer feasible.

Dataset	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	$\|\mathbf{S}\|$	$\mathbf{M}$	$\mathbf{A}$
Amazon	335k	926k	49k	100.0	14.83
DBLP	317k	1.0M	2.5k	429.8	2.57

Table 1: Data statistics. $|\mathbf{V}|$: number of nodes, $|\mathbf{E}|$: number of links, $|\mathbf{S}|$: number of ground-truth communities, $\mathbf{M}$: average number of nodes per community, $\mathbf{A}$: average community memberships per node.

Thus, new approaches are proposed to tackle this problem in recent years. In this paper, we mainly focus on matrix-factorization-based approaches for overlapping community detection.

Definition 2.2 (**Overlapping Community Detection via MF**). *Overlapping community detection via MF is a process that takes the adjacency matrix $A \in \{0, 1\}^{|V|\times|V|}$ of a network G as input and produces a node-community weight matrix $F \in \mathbb{R}^{|V|\times|S|}$ whose entry $F_{u,c}$ represents the weight of node $u \in V$ in community $c \in S$ to minimize a particular loss function l, i.e.,*

$$\arg\min_{F} l(A, FF^T), \tag{2}$$

where S is the set of communities. In the end, we obtain S according to F.

As we mentioned, the simplest l is in the form of $||A - FF^T||$. The main target of this paper is to seek for a better l that can capture the nature of communities more precisely.

2.2 Data Observation

In order to validate the link-to-community perspective, we observe two large network datasets with ground-truth communities[1] [Yang and Leskovec, 2012] to see whether linked node pairs have more similar community representations than non-linked ones. These two datasets are:

- **Amazon**: a products co-purchasing network based on Customers Who Bought This Item Also Bought feature of the Amazon website.
- **DBLP**: a collaboration network of research paper authors in computer science;

A simple statistics can be found in Table 1.

We exploit average number of shared communities (SC) and average Jaccard similarity of community memberships (JS) for all linked node pairs as our measurements. They are calculated by

$$SC = \frac{1}{2|E|} \sum_{i\in V} \sum_{j\in N^+(i)} |C_i \cap C_j|, \tag{3}$$

and

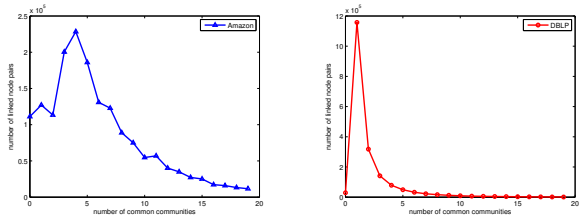
$$JS = \frac{1}{2|E|} \sum_{i\in V} \sum_{j\in N^+(i)} \frac{|C_i \cap C_j|}{|C_i \cup C_j|}, \tag{4}$$

respectively, where $N^+(i)$ is the set of i' neighbors and C_i represents the set of communities containing i. We also draw

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

Dataset	SC	$\mathbf{SC_r}$	JS	$\mathbf{JS_r}$
Amazon	**6.767**	0.178	**0.490**	0.010
DBLP	**2.078**	0.009	**0.347**	0.002

Table 2: Data observations. **SC**: average number of shared communities per linked node pair, $\mathbf{SC_r}$: average number of shared communities per random node pair, average Jaccard similarity of community memberships per linked node pair, **JS**: average Jaccard similarity of community memberships per linked node pair, $\mathbf{JS_r}$: average Jaccard similarity of community memberships per random node pair.



(a) **Amazon** (b) **DBLP**

Figure 1: The number of linked node pairs sharing a particular number of communities.

ten thousand random node pairs that do not need to be linked and compute the same measurements for these pairs. The comparison results are shown in Table 2. The huge gap between linked ones (bold) and random ones (normal) reveals that two linked nodes share much more communities than two random nodes in average and thus strongly supports the necessity of link-to-community perspective.

Moreover, we count the number of linked node pairs that share a particular number of communities in Figure 1. In both networks, the number of linked node pairs reaches the peak near their average number of shared communities for linked node pairs and starts to decrease when this number continues to increase. This observation shows that average number of shared communities can be used to measure how strong the link-to-community side of homophily effect is. For example, we can claim that the link-to-community side of homophily effect in Amazon is much stronger than that in DBLP.

3 Related Work

A lot of efforts have been conducted on the research of community detection. An extensive survey can be found in [Fortunato, 2010]. As we have mentioned, overlapping community detection draws the attention because classic community detection declines multiple memberships. A recent survey of overlapping community detection algorithms can be found in [Xie *et al.*, 2013]. According to the key idea, we classify those algorithms into local approaches and global approaches. Local approaches explore a network from small components to the whole. For example, [Palla *et al.*, 2005] and [Kumpula *et al.*, 2008] search for all the k-cliques and combines those sharing $k-1$ nodes until no combinations can be made. [Coscia *et al.*, 2012] applies label propagation algorithm to detect small communities on ego network of each node, i.e., a node with its neighbors, and lastly merge communities with large overlap. [Whang *et al.*, 2013] picks several seed nodes and conducts PageRank on each node to find communities.

Global approaches, on the other side, assume that communities exist in the first place and aim to find the most suitable membership for each node. Some major models include game theory models [Chen *et al.*, 2010], stochastic block models [Airoldi *et al.*, 2009; Jin *et al.*, 2015], and matrix factorization (MF) models. Since our model is a MF-based model, we will introduce it in more details. [Psorakis *et al.*, 2011] and [Wang *et al.*, 2011] are the earliest ones applying MF into overlapping community detection. While the previous factorizes the adjacency matrix into two different matrices, the latter forces them to be the same and thus gives physical meaning to the factorized matrix. However, they both use the simplest squared loss as their learning objective. [Zhang and Yeung, 2012] adds a community interaction matrix to become a non-negative matrix tri-factorization model. [Yang and Leskovec, 2013] is among the first to exploit generative learning objective, which maximizes the likelihood of generating all the links in the original graph.

Our community-to-link perspective applies a pairwise learning objective proposed in the *Preference-based Non-negative Matrix Factorization* model [Zhang *et al.*, 2015]. The intuition is that two nodes are more likely to become friends if they share more common communities. Our link-to-community perspective borrows the idea from the *Skip-Gram* model [Mikolov *et al.*, 2013], which is originally designed for natural language processing tasks. The training objective of this model is to find word representations that can predict the surrounding words in a sentence or a document. Perozzi at el. [Perozzi *et al.*, 2014] are the first to extend *Skip-Gram* to represent nodes in a social network. Random walks passing a node is regarded as the context of this node.

Notation	Meaning
$G(V, E)$	graph G (node set V, edge set E)
$A \in \{0,1\}^{\lvert V\rvert \times \lvert V\rvert}$	adjacency matrix of G
S	the set of detected communities
C_u	the set of communities containing u
$F \in \mathbb{R}^{\lvert V\rvert \times \lvert S\rvert}$	node-community weight matrix
F_u	u's community representation
$N^+(u)$	node set of u's neighbors
$N^-(u)$	node set of u's non-neighbors

Table 3: A summary of notations.

4 A Homophily-based Non-negative Matrix Factorization (HNMF) Model

In this section, we first introduce our model assumptions. Then we formalize our *HNMF* model from both perspectives and combine them into a unified model. In the end, we exhibit our parameter learning algorithm and discuss some more detailed issues. All the notations are listed in Table 3.

4.1 Model Assumption

Since we model homophily from both community-to-link and link-to-community perspectives, our model assumption will be introduced in two separate parts as well.

For the community-to-link perspective, the basic assumption is that two nodes should have higher probability to build links with each other if they share more communities, i.e.,

$$\mathcal{P}(A_{u,i}=1) > \mathcal{P}(A_{u,j}=1), \text{if } |C_u \cap C_i| > |C_u \cap C_j|. \quad (5)$$

Since we apply the *PNMF* model [Zhang *et al.*, 2015] in this part, we also need to adopt the preference assumption, i.e.,

$$r_{u,i} > r_{u,j}, \text{if } i \in N^+(u) \text{ and } j \in N_u^-, \quad (6)$$

where $r_{u,i}$ is the preference of node u on node i. It indicates that a node prefers to build links with neighbors over non-neighbors.

For the link-to-community perspective, we assume that two linked nodes are more similar than two non-linked nodes. It is formally denoted as:

$$sim_{u,i} > sim_{u,j}, \text{if } i \in N^+(u) \text{ and } j \in N_u^-, \quad (7)$$

where $sim_{u,i}$ is the similarity between node u and node i.

4.2 Modeling Community-to-link Perspective

We demonstrate our learning objective of community-to-link perspective by following the formulation of the *PNMF* model [Zhang *et al.*, 2015]. For each node u, the objective of *PNMF* is to maximize the likelihood of a pairwise preference order, which can be denoted as $\mathcal{P}(>_u)$. According to the preference assumption, $\log \mathcal{P}(>_u)$ can be represented as:

$$\sum_{i \in N^+(u)} \sum_{j \in N^-(u)} \log \mathcal{P}(i >_u j). \quad (8)$$

Following the core idea of the community-to-link assumption, we use the community representations of node i, j, and k to model $\mathcal{P}(i >_u j)$. It can be written as

$$\mathcal{P}(i >_u j) = \sigma(F_u^T (F_i - F_j)), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function $\sigma(x) := \frac{1}{1+e^{-x}}$. We choose sigmoid function because it is a differentiable function which can map any real number into the range between 0 and 1.

Based on Eq. (8) and Eq. (9), the learning objective of the community-to-link perspective can be derived by summing up log-likelihoods of all the nodes, i.e.,

$$\begin{aligned} \mathcal{C}(F) &:= \sum_u \sum_{i \in N^+(u)} \sum_{j \in N^-(u)} \log \mathcal{P}(i >_u j) \\ &= \sum_u \sum_{i \in N^+(u)} \sum_{j \in N^-(u)} \log \sigma(F_u^T (F_i - F_j)). \end{aligned} \quad (10)$$

4.3 Modeling Link-to-community Perspective

Motivated by the success of *Skip-Gram* model [Mikolov *et al.*, 2013] where word representations are learned in terms of representations of surrounding words in the same context, we here adopt a similar idea to learn a node's community representation from other nodes in its local scope. In our case, for a node u, its local scope is constrained within u's' neighbors. According to our link-to-community assumption, u's neighbors should have similar community representations with u. Formally, given a node u and its neighbors, our learning objective for the link-to-community perspective is to maximize the sum of log-likelihoods for a node to represent its neighbors as follows,

$$\sum_{i \in N^+(u)} \log \mathcal{P}(i|u). \quad (11)$$

Following the formulation in *Skip-Gram*, we apply a softmax function to define $\mathcal{P}(i|u)$ as

$$\mathcal{P}(i|u) = \frac{\exp({F'}_i^T F_u)}{\sum_{i'=1}^{|V|} \exp({F'}_{i'}^T F_u)}. \quad (12)$$

Note that F' needs to be introduced into our model and should be regarded as the latent community representation matrix which is corresponding to the 'output' vector representations. Likewise, our learning target F is corresponding to the 'input' vector representations.

A computationally efficient approximation of the full softmax function is *Negative Sampling (NEG)*, which is simplified version of *Noise Contrastive Estimation (NCE)* [Gutmann and Hyvärinen, 2010]. It substitutes Eq. (12) with

$$\mathcal{P}(i|u) = \sigma({F'}_i^T F_u) + h\mathbb{E}_{i' \sim P_{N^-(u)}} \big[\sigma(-{F'}_{i'}^T F_u)\big], \quad (13)$$

where $\sigma(\cdot)$ is also the sigmoid function, h is the number of negative samples, and $P_{N^-(u)}$ is the unigram distribution raised to the power $\frac{3}{4}$.

Thus, we can obtain the learning objective of the link-to-community perspective as follows,

$$\begin{aligned} \mathcal{L}(F, F') := \sum_u \sum_{i \in N^+(u)} \big(&\log \sigma({F'}_i^T F_u) \\ &+ h\mathbb{E}_{i' \sim P_{N^-(u)}} \big[\log \sigma(-{F'}_{i'}^T F_u)\big]\big). \end{aligned} \quad (14)$$

4.4 The Unified Model

Now we can combine the two perspectives, i.e., Eq. (10) and Eq. (14), into one unified model. The final learning objective of our *HNMF* model is to maximize

$$\begin{aligned} &\mathcal{U}(F, F') := \mathcal{C}(F) + \beta \mathcal{L}(F, F') - \lambda \mathcal{R}(F) \\ &= \sum_u \sum_{i \in N^+(u)} \big(\sum_{j \in N^-(u)} \log \sigma(F_u^T (F_i - F_j)) \\ &+ \beta \log \sigma({F'}_i^T F_u) + \beta h \mathbb{E}_{i' \sim P_{N^-(u)}} \big[\log \sigma(-{F'}_{i'}^T F_u)\big]\big) \\ &- \lambda ||F||_F, \end{aligned} \quad (15)$$

where $\mathcal{R}(F)$ is a regularization term, where we employ the Frobenius norm of F, β is the homophily coefficient used to adjust the importance of one perspective compared with the other, and λ is the regularization coefficient.

4.5 Parameter Learning

Considering time efficiency and the non-negativity constraint, we use projected stochastic gradient descent [Lee and Seung,

Input: A, the adjacency matrix of original graph.
Output: F, the node-community weight matrix.
Initialization:
Initialize F (uniformly at random);
for *each node* u **do**
  Construct $N^+(u)$;
end
Training:
Compute initial loss;
repeat
  for *each node* u **do**
    Uniformly sample node i from $N^+(u)$;
    Community-to-link:
    Uniformly sample node j from $N^-(u)$
    $F_u = F_u + \alpha \frac{\partial \mathcal{C}}{\partial F_u}$;
    $F_i = F_i + \alpha \frac{\partial \mathcal{C}}{\partial F_i}$;
    $F_j = F_j + \alpha \frac{\partial \mathcal{C}}{\partial F_j}$;
    Link-to-community:
    Sample h negative nodes $i' \sim P_{N^-(u)}$;
    $F_u = F_u + \alpha\beta \frac{\partial \mathcal{L}}{\partial F_u}$; $F'_i = F'_i + \alpha\beta \frac{\partial \mathcal{L}}{\partial F'_i}$;
    for *each node* i' **do**
      $F'_{i'} = F'_{i'} + \alpha\beta \frac{\partial \mathcal{L}}{\partial F'_{i'}}$;
    end
    Regularization and Projection:
    $F_u = \max\{F_u - \alpha\lambda \frac{\partial \mathcal{R}}{\partial F_u}, 0\}$;
    $F_i = \max\{F_i - \alpha\lambda \frac{\partial \mathcal{R}}{\partial F_i}, 0\}$;
    $F_j = \max\{F_j - \alpha\lambda \frac{\partial \mathcal{R}}{\partial F_j}, 0\}$;
  end
  Compute loss;
until *Convergence or max_iter is reached*;

Algorithm 1: Overlapping community detection using *HNMF*

2001; Lin, 2007] as our parameter learning method. It updates the corresponding parameters whenever a single sample or a small batch of samples arrive and maps the parameters back to the nearest point in the projected space, in our case, the non-negative space. The update rule for a parameter Θ is

$$\Theta^{t+1} = \max\{\Theta^t + \alpha \frac{\partial \mathcal{U}}{\partial \Theta}, 0\}, \tag{16}$$

where α is the learning rate.

The whole process of our learning method is shown in Algorithm 1. Here we discuss some of the steps in more detail.

- Initialization. We initialize each entry of F to be a random real value between 0 and 1 divided by the number of communities, i.e., the number of columns in F.
- Negative sampling. For the negative sample j from $N^-(u)$, we keep sampling j from V until $j \notin N^+(u)$.
- Convergence criterion. We randomly sample a number of triples (u, i, j) and use them to compute the initial loss on according to Eq. (15) without considering the regularization term. After each iteration, we repeat the same process with a different set of samples and stop when the difference between current loss and previous loss is less than a very small value, say ϵ, of the initial loss.
- Setting the number of communities. We adopt a cross-validation paradigm by reserving 10% of nodes as a validation set. Since the computational cost on the validation set is still huge, sampling will be used as well.

Dataset	$\lvert\mathbf{V}\rvert$	$\lvert\mathbf{E}\rvert$
Dolphins	62	159
Les Misérables	77	254
Books about US politics	105	441
Word adjacencies	112	425
American college football	115	613
High-energy theory	8,361	15,751

Table 4: Statistics of six Newman's datasets. $|\mathbf{V}|$: number of nodes, $|\mathbf{E}|$: number of links.

4.6 Other Issues

Scalability. To scale up our *HNMF* model on large networks, we employ an asynchronous version of stochastic gradient descent to update the parameters. Since most updates only modify a small part of all the parameters, the chance that a parameter is simultaneously being updated by more than one worker is very small. Thus, a lock-free approach [Recht *et al.*, 2011] can be adopted to parallelize our parameter learning process. We will show in the experiments that the convergence speed is satisfactory.

Community membership threshold. After we obtain the node-community weight matrix F, we still need to figure out community memberships for each node. A standard solution is to set a threshold and discard all the nodes whose weights are below the threshold. We employ the approach in [Zhang *et al.*, 2015] and omit the details here due to space limit.

5 Experiments

In this section, we compare our *HNMF* model with six baselines on various real-world datasets, including large networks with ground-truth communities. We measure the quality of communities with two metrics, modularity and F_1 score. Our experimental procedures and results are described as follows.

5.1 Data Description

Apart from the two large networks with ground-truth communities introduced in Section 2, we also use six benchmark networks collected by Newman[2] as our datasets. These networks are relatively small and have no ground-truth communities. We list the basic information of these datasets in Table 4.

5.2 Experimental Setup

Comparison methods. We select two local approaches, namely *Sequential Clique Percolation (SCP)* [Kumpula *et al.*, 2008] and *Demon* [Coscia *et al.*, 2012], and four state-of-the-art global approaches, namely *BNMF* [Psorakis *et al.*,

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

Dataset	SCP	Demon	BNMF	BNMTF	BigCLAM	PNMF	HNMF
Dolphins	0.305	0.680	0.507	0.507	0.423	0.979	**1.021**
Books about US politics	0.496	0.432	0.461	0.492	0.529	0.864	**0.988**
Word adjacencies	0.071	0.032	0.254	0.268	0.231	0.668	**0.699**
American college football	0.605	0.540	0.558	0.573	0.518	1.049	**1.113**
Power grid	0.044	0.195	0.342	0.368	1.010	1.105	**1.135**
High-energy theory	0.543	0.962	0.565	0.600	0.964	0.973	**1.060**

Table 5: Experimental results on Newman's networks in terms of modularity.

Dataset	Demon	BigCLAM	PNMF	HNMF
Amazon	0.082	0.044	0.042	**0.122**
DBLP	0.102	0.039	0.098	**0.104**

Table 6: Experimental results on two large networks in terms of F_1 score.

2011], *BNMTF* [Zhang and Yeung, 2012], *BigCLAM* [Yang and Leskovec, 2013], and *PNMF* [Zhang *et al.*, 2015], to compare with our *HNMF* model.

Evaluation metrics. We use modularity for datasets without ground-truth communities and F_1 score for datasets with ground-truth communities.

- **Modularity.** The well-known modularity [Newman, 2006] Q is defined as

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

  where $d(u)$ is u's degree. We can see that a node pair (u, v) positively contributes to modularity if they are linked and negatively contributes otherwise.

- **$\mathbf{F_1}$ score.** F_1 score of a detected community S_i is defined as the harmonic mean of

$$\text{precision}(S_i) = \max_j \frac{\hat{S}_j \bigcap S_i}{|\hat{S}_j|}$$

  and

$$\text{recall}(S_i) = \max_j \frac{\hat{S}_j \bigcap S_i}{|S_i|},$$

  where $\hat{S}_j$ is one of ground-truth communities. The overall F_1 score of the set of detected communities S is the average F_1 score of all communities in S.

5.3 Results

Results on Newman's networks in terms of modularity are shown in Table 5. Despite that *PNMF* already has large improvement over other baselines, our *HNMF* model further outperforms *PNMF* on all datasets, which reflects the significance of the link-to-community perspective in overlapping community detection.

Results on two large networks in terms of F_1 score are shown in Table 6. We notice that the improvement on Amazon is much larger than that of DBLP. Recall our claim in data observation that the link-to-community side of homophily effect in Amazon is much stronger than that in DBLP. This explains such difference of improvement between these two datasets. With asynchronous stochastic gradient descent, the running time of our learning algorithm is about 4 hours for Amazon and about 6 hours for DBLP on a computer with a Xeon 24-core 2.60GHz CPU and 128GB memory.

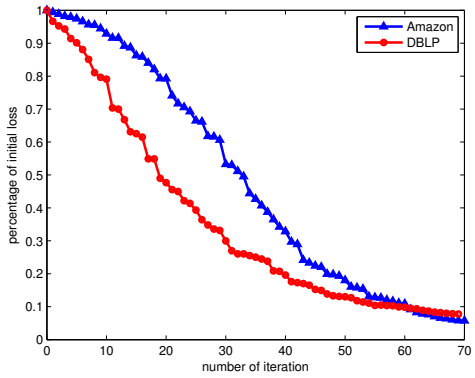


Figure 2: Convergence speed of our learning algorithm.

Figure 2 illustrates the convergence speed of our learning algorithm on Amazon and DBLP. Since our computation of loss employs a global sampling strategy, we can directly sort the losses from all workers according to time sequence. We set the ϵ in Section 4.5 to be 0.001. We can see that our learning algorithm is able to converge within a small number of iterations.

6 Conclusion

In this paper, we propose a *Homophily-based Non-negative Matrix Factorization* model to capture both sides of homophily effect for overlapping community detection. Our unified learning objective is a combination of a preference-based pair-wise learning objective for the community-to-link perspective and a generative community representation learning with network embedding for the link-to-community perspective. We adopt an asynchronous stochastic gradient descent to learn model parameters efficiently. Experiments on real-world networks show that this model can indeed improve the quality of detected overlapping communities.

Acknowledgement

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14203314 and

No. CUHK 14205214 of the General Research Fund), and 2015 Microsoft Research Asia Collaborative Research Program (Project No. FY16-RES-THEME-005).

References

[Airoldi *et al.*, 2009] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40, 2009.

[Chen *et al.*, 2010] Wei Chen, Zhenming Liu, Xiaorui Sun, and Yajun Wang. A game-theoretic framework to identify overlapping communities in social networks. *Data Mining and Knowledge Discovery*, 21(2):224–240, 2010.

[Coscia *et al.*, 2012] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. Demon: a local-first discovery method for overlapping communities. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 615–623. ACM, 2012.

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

[Girvan and Newman, 2002] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[Gutmann and Hyvärinen, 2010] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[Jin *et al.*, 2015] Di Jin, Zheng Chen, Dongxiao He, and Weixiong Zhang. Modeling with node degree preservation can accurately find communities. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Kumpula *et al.*, 2008] Jussi M Kumpula, Mikko Kivelä, Kimmo Kaski, and Jari Saramäki. Sequential algorithm for fast clique percolation. *Physical Review E*, 78(2):026109, 2008.

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

[Lin, 2007] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[Palla *et al.*, 2005] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[Psorakis *et al.*, 2011] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011.

[Recht *et al.*, 2011] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[Tang *et al.*, 2013] Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 53–62. ACM, 2013.

[Wang *et al.*, 2011] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521, 2011.

[Whang *et al.*, 2013] Joyce Jiyoung Whang, David F Gleich, and Inderjit S Dhillon. Overlapping community detection using seed set expansion. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2099–2108. ACM, 2013.

[Xie *et al.*, 2013] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43, 2013.

[Yang and Leskovec, 2012] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 3. ACM, 2012.

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

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

[Zhang *et al.*, 2015] Hongyi Zhang, Irwin King, and Michael R. Lyu. Incorporating implicit link preference into overlapping community detection. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 396–402. ACM, 2015.