

From Mutual Friends to Overlapping Community Detection: A Non-negative Matrix Factorization Approach

Xingyu Niu, Hongyi Zhang^(✉), Micheal R. Lyu, and Irwin King

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong
neoxyniu@gmail.com, {hyzhang,lyu,king}@cse.cuhk.edu.hk

Abstract. Community detection provides a way to unravel complicated structures in complex networks. Overlapping community detection allows nodes to be associated with multiple communities. *Matrix Factorization (MF)* is one of the standard tools to solve overlapping community detection problems from a global view. Existing MF-based methods only exploit link information revealed by the adjacency matrix, but ignore other critical information. In fact, compared with the existence of a link, the number of mutual friends between two nodes can better reflect their similarity regarding community membership. In this paper, based on the concept of mutual friend, we incorporate *Mutual Density* as a new indicator to infer the similarity of community membership between two nodes in the MF framework for overlapping community detection. We conduct data observation on real-world networks with ground-truth communities to validate an intuition that mutual density between two nodes is correlated with their community membership cosine similarity. According to this observation, we propose a *Mutual Density based Non-negative Matrix Factorization (MD-NMF)* model by maximizing the likelihood that node pairs with larger mutual density are more similar in community memberships. Our model employs stochastic gradient descent with sampling as the learning algorithm. We conduct experiments on various real-world networks and compare our model with other baseline methods. The results show that our MD-NMF model outperforms the other state-of-the-art models on multiple metrics in these benchmark datasets.

Keywords: Complex networks · Overlapping community detection · Matrix factorization

1 Introduction

In complex networks, there usually exist groups inside which nodes are connected more densely with one another than with the nodes outside. These groups of nodes are called *communities* [13]. In reality, these groups usually have physical meanings such as members of the same organization, scientists with publications in the same area, or proteins sharing the same function. Thus, uncovering

such latent communities in complex networks has attracted great research interests in the past decade [11]. Classic methods assume communities are mutually exclusive, i.e., each node of a network belongs to one and only one community. However, in real-world complex networks like social networks and biological networks, such community membership restriction does not apply because a node may have multiple characteristics and thus belongs to multiple communities. As a result, a more challenging problem named *overlapping community detection* has been introduced in recent years [31].

Matrix Factorization (MF), as one of the standard framework to solve the problem of overlapping community detection, detects communities from a global view [31]. Taking the adjacency matrix G of the given network as input, MF-based models assign the number of communities in advance, and seek out a node-community weight matrix F , which matches the information revealed by the input as accurately as possible. Early work [23, 29] simply aims to approximate G entry by entry with FF^T , which only makes use of the mathematical representation of adjacency matrix, but ignores its physical meaning. The most obvious information an adjacency matrix provides is the link information. Thus, recent work [32, 35] assumes that nodes sharing more communities have a higher probability to be linked and formulates the problem with a generative objective function. In other words, a link can be regarded as an indicator to reflect the similarity of community membership between two nodes.

However, a link is not a perfect indicator for two major reasons. First, it is common that two nodes sharing several communities do not have a link between them, or two nodes with no common community are connected. A survey conducted on Facebook [9] shows that edges between two individuals from different communities outnumber edges connecting users in the same community. For example, a salesperson may make connections with many strangers to sell his products, and the establishment of links between salespeople and customers does not indicate any similarity between their community memberships. In cases like these, links become noise instead of evidence. Second, a link is a binary indicator in an unweighted network. Given two linked node pairs with no other information at all, it is impossible to distinguish which one is more similar.

Inspired by the definition of tie strength [12], we incorporate a more powerful indicator, which is the number of mutual friends between two nodes, to reflect their community membership similarity. The definition of tie strength reveals that the stronger tie the two nodes own, the larger overlap in their friendship circles they will have. This idea can be incorporated into our matrix factorization framework for overlapping community detection, which meets the common sense that the more communities two nodes share, the more mutual friends they will have. For example, if two individuals attended the same class in high school, joined the same basketball team, and work in the same company now, they should know many mutual friends in different communities, i.e., their ego-networks (friend circles) are densely overlapped. Compared to a link, the number of mutual friends is no longer a binary indicator and it provides more confidence to predict the similarity of community membership between

two nodes. However, it still suffers from several issues: the lack of friends of two nodes may limit the number of mutual friends between them, and communities with different sizes may contribute different numbers of mutual friends to each node pair. To handle these limitations, we incorporate *Mutual Density* as a more consistent indicator, which is defined as the Jaccard similarity of two nodes’ ego-networks. Under the general description of “neighborhood similarity”, the concept of mutual density has been applied in community detection under different assumptions [1, 3, 21, 26, 28]. However, none of these methods are based on matrix factorization and none of them use mutual density to measure the similarity of community membership between two nodes.

In this paper, we introduce mutual density and the number of mutual friends as the new indicators instead of links themselves for inferring community membership similarity in the matrix factorization framework. We conduct data observation on real-world networks with ground-truth communities to validate that mutual density is more consistent with community memberships similarity than the other two indicators. Thus, we formulate our *Mutual Density based Non-negative Matrix Factorization (MD-NMF)* model, which incorporates mutual density as the community similarity indicator and employs a novel objective function to ensure that a node pair with higher mutual density is more likely to have a higher community membership similarity. From a node’s perspective, we ensure that it is more likely to join the same communities with its acquaintances than with its strangers. To solve the optimization problem, we apply projected stochastic gradient descent with sampling. By applying our model to real-world and open-source network datasets, we find that our new MD-NMF model outperforms several state-of-the-art methods on either modularity or F_1 score.

The main contributions of this paper are:

1. We incorporate *Mutual Density* as a new indicator to reflect the community membership similarity between two nodes in substitution for a link within the matrix factorization framework for overlapping community detection.
2. We find that there is consistency between the mutual density of two nodes and their community memberships similarity by empirically studying real-world networks with ground-truth communities.
3. We propose a novel *Mutual Density based Non-negative Matrix Factorization (MD-NMF)* model for overlapping community detection by formulating mutual density properly in the matrix factorization framework. Our model outperforms state-of-the-art baselines.

2 Definition and Data Observation

2.1 Problem Definition

Definition 1 (Community Detection). *Given an unweighted and undirected graph $G(V, E)$, community detection aims to find a communities set $S = \{C_i | C_i \neq \emptyset, C_i \neq C_j, 1 \leq i, j \leq p\}$ where C_i represents a community consisting a set of nodes, to maximizes a particular objective function f , i.e.,*

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

where p is the number of communities.

Different from the traditional community detection problem, an overlapping community detection problem allows communities to overlap with each other. This relaxation enables the *Matrix Factorization* approach to be employed. In MF-based methods, the graph is represented by its adjacency matrix $G \in \{0, 1\}^{n \times n}$, whose (i, j) entry indicates whether node i and node j are connected or not. The goal is to find a node-community weight matrix F , with its entry $F_{u,c}$ representing the weight of node u in community c , and apply F to approximate the adjacency matrix.

2.2 Indicator Definitions

To infer the community membership similarity between two nodes, we have mentioned three indicators in Introduction. They are link existence $l(u, v)$, the number of mutual friends $m(u, v)$ and mutual density $d(u, v)$, where u and v are both nodes in V . We formally define each of them as follows.

Definition 2 (Link Existence). *Given a graph $G(V, E)$ and two nodes $u, v \in V$, the link existence between u and v is*

$$l(u, v) = \begin{cases} 1 & \text{if } G_{uv} = 1, \\ 0 & \text{else} \end{cases}. \quad (2)$$

Definition 3 (The Number of Mutual Friends). *Given a graph $G(V, E)$ and two nodes $u, v \in V$, the number of mutual friends between u and v is*

$$m(u, v) = |\{i | (u, i) \in E \text{ and } (v, i) \in E\}|. \quad (3)$$

Definition 4 (Mutual Density). *Given a graph $G(V, E)$ and two nodes $u, v \in V$, the mutual density between u and v is*

$$d(u, v) = \frac{|\{i | (u, i) \in E \text{ and } (v, i) \in E\}|}{|\{j | (u, j) \in E \text{ or } (v, j) \in E\}|}. \quad (4)$$

2.3 Data Observation

To validate (1) the number of mutual friends is better than a link in inferring community membership similarity, and (2) mutual density is more stable compared with the number of mutual friends, we conduct two experiments on two

Table 1. Dataset statistics. $|V|$: number of nodes, $|E|$: number of edges, $|C|$: number of ground-truth communities, D : average degree of nodes, M : average number of nodes per community, A average number of joined communities per node.

Dataset	$ V $	$ E $	$ C $	D	M	A
Amazon	335k	926k	49k	3.38	100.0	14.83
DBLP	317k	1.0M	2.5k	4.93	429.8	2.57

large real-world networks with ground-truth communities [33]. Table 1 shows the statistics of these two networks, which are **Amazon** and **DBLP**.¹

To quantify the community membership similarity between two nodes, we use *cosine similarity* as our measurement, which is defined as follows.

Definition 5 (Cosine Similarity of Community Membership). *Given a graph with p ground-truth communities $\{C_i | i = 1, 2, \dots, p\}$, the cosine similarity of community membership $s(u, v)$ between u and v is*

$$s(u, v) = \frac{\mathbf{u} \cdot \mathbf{v}^T}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}, \quad (5)$$

where $\mathbf{u} \in \mathcal{R}^p$ is the community membership vector of node u and u_i represents the weight u belongs to community C_i .

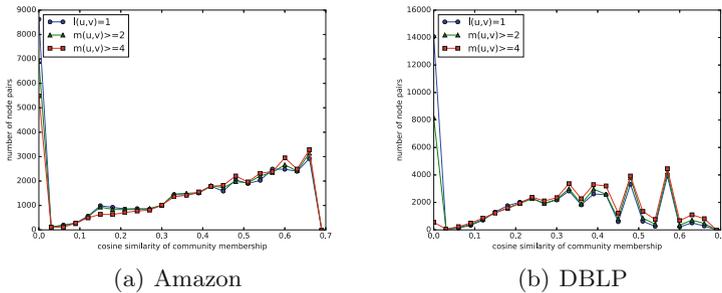


Fig. 1. The number of sampled node pairs having a same value of cosine similarity

First, we randomly sample 100,000 node pairs with links as well as 100,000 node pairs with at least two or four mutual friends and compute the cosine similarity of community membership for each node pair. Figure 1 plots the number of 3 different types of node pairs with the same value of cosine similarity. We expect all three types of node pairs to share at least one community and thus to have non-zero cosine similarity. However, nearly 14,000 node pairs with links do not share any communities. The error rate is about 14%. On the other side, less than 8% of the node pairs with at least two mutual friends and only about 1% of the node pairs with at least four mutual friends are out of our expectation. When the value of cosine similarity is non-zero, all three types are pretty similar, and the number of node pairs with four mutual friends is slightly greater than the other types. Thus, the number of mutual friends is a more accurate and more flexible indicator compared to the existence of links.

Second, we compare the stability of indicator between the number of mutual friends and mutual density. A stable indicator is expected to be monotonic while community membership similarity increases. We sample 10,000 node pairs each

¹ <http://snap.stanford.edu/data/>.

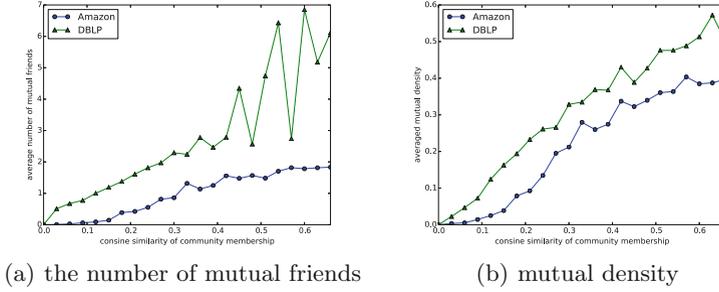


Fig. 2. Averaged value of each indicator as a function of cosine similarity in community membership

time with a certain value of cosine similarity and calculate the average number of mutual friends and average mutual density of these node pairs. The result is shown in Fig. 2. We can see that on the DBLP data, the average number of mutual friends vibrates up and down while average mutual density is almost monotonic as cosine similarity increases. Thus, mutual density is a more stable indicator than the number of mutual friends to infer community membership similarity.

In summary, mutual density is the best indicator among all three indicators we mentioned with highest accuracy and stability.

3 Mutual Density Based NMF Model

3.1 Model Assumption

From the data observation, we can see that the cosine similarity of community membership between two nodes is correlated with their mutual density. It leads to the intuition of our model that two nodes with larger mutual density are more likely to have higher cosine similarity of community membership.

To formally illustrate our model assumption, we need to define two relationships between two nodes in the first place: α -acquaintance and β -stranger.

Definition 6 (α -acquaintance). Given $\alpha \in [0, 1]$, for two nodes $u, v \in V$, v is u 's α -acquaintance if and only if

$$d(u, v) \geq \alpha.$$

By the symmetry of $d(u, v)$, u is also v 's α -acquaintance.

Definition 7 (β -stranger). Given $\beta \in [0, 1]$, for two nodes $u, v \in V$, v is u 's β -stranger if and only if

$$d(u, v) \leq \beta.$$

By the symmetry of $d(u, v)$, u is also v 's β -stranger.

In both definitions, $d(u, v)$ is the mutual density between u and v defined in Eq. (4). Moreover, for a node u , we define its set of α -acquaintances as $A(u, \alpha) = \{i | d(u, i) \geq \alpha\}$ and its set of β -strangers as $B(u, \beta) = \{j | d(u, j) \leq \beta\}$.

Following our intuition, our model assumption can be formally defined as

$$\begin{aligned} s(u, i) &> s(u, j), \\ \text{if } i &\in A(u, \alpha), j \in B(u, \beta), \text{ and } \alpha > \beta, \end{aligned} \quad (6)$$

where $s(u, i)$ is the cosine similarity of community memberships between u and i .

In other words, we expect that the cosine similarity between u and any of its α -acquaintances should be greater than the cosine similarity between u and any of its β -strangers. Adjusting α and β for different graphs enables us to make sure that the difference of cosine similarity is significant. If α is only slightly greater than β , we are not confident enough to make such assumption.

3.2 Model Formulation

In the MD-NMF model, we aim to find the node-community weight matrix F which maximizes the likelihood that every node in the graph has higher cosine similarity in community membership with all its α -acquaintances than with all its β -strangers. For each node u , we want to maximize

$$\mathcal{P}(>_u | F, \alpha, \beta) = \prod_{i \in A(u, \alpha)} \prod_{j \in B(u, \beta)} \mathcal{P}(s(u, i) > s(u, j) | F). \quad (7)$$

Given any two nodes $u, v \in V$, we can obtain their node-community weight vectors F_u, F_v from F . From the observation that the higher cosine similarity of community membership vectors between two nodes, the greater mutual density they will have, we define the probability that $s(u, i) > s(u, j)$ given the node-community membership matrix as

$$\mathcal{P}(s(u, i) > s(u, j) | F) = \sigma\left(\frac{F_u F_i^T}{\|F_u\|_2 \|F_i\|_2} - \frac{F_u F_j^T}{\|F_u\|_2 \|F_j\|_2}\right), \quad (8)$$

where σ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. For simplicity, we define $\phi(i, j) = \frac{F_i F_j^T}{\|F_i\| \|F_j\|}$, so we have

$$\mathcal{P}(s(u, i) > s(u, j) | F) = \sigma(\phi(u, i) - \phi(u, j)). \quad (9)$$

Since the sigmoid function maps any real value into $(0, 1)$, this probability approaches to 1 when $\phi(u, i) \gg \phi(u, j)$ and approaches to 0 when $\phi(u, i) \ll \phi(u, j)$.

By multiplying Eq. (7) for each node and combining Eqs. (8) and (9), we can derive the final learning objective of the MD-NMF model, which is

$$\begin{aligned}
l(F) &= \max_{F \in \mathbb{R}_+^{n \times p}} \log \prod_{u \in V} \mathcal{P}(s(u, \cdot) | F, \alpha, \beta) - \lambda \cdot \text{reg}(F) \\
&= \max_{F \in \mathbb{R}_+^{n \times p}} \sum_{u \in V} \sum_{i \in A(u, \alpha)} \sum_{j \in B(u, \beta)} \log \mathcal{P}(s(u, i) > s(u, j) | F) \\
&\quad - \lambda \cdot \text{reg}(F) \\
&= \max_{F \in \mathbb{R}_+^{n \times p}} \sum_{u \in V} \sum_{i \in A(u, \alpha)} \sum_{j \in B(u, \beta)} \log \sigma(\phi(u, i) - \phi(u, j)) \\
&\quad - \lambda \cdot \text{reg}(F),
\end{aligned} \tag{10}$$

where $\text{reg}(F)$ is a regularization term in order to prevent overfitting of F , and λ is the regularization parameter. For the simplicity of differentiation, we set $\text{reg}(F) = \|F\|_F^2$, which is the Frobenius norm of F .

3.3 Parameter Learning

To make our model scalable to large datasets, we employ the widely used paradigm of *Stochastic Gradient Descent (SGD)* as our learning algorithm. Also considering the non-negativity constraint, we apply a projected gradient method [18] which maps the vector with negative parameters back to the nearest point in the projected space. Following the learning objective l , we update the matrix F by

$$\Theta_{t+1} = \max\{\Theta_t + \delta \frac{\partial l}{\partial \Theta}, 0\}, \tag{11}$$

where δ is the learning rate and Θ can be any entry of matrix F .

Algorithm 1 describes the whole iterative process of parameter learning. In each iteration, the time complexity is $O(|E|p)$, where $|E|$ is the number of edges and p the number of communities. Because we need to save the whole node-community weight matrix F in memory, the space complexity of the algorithm is $O(|V|p)$, where V is the number of nodes. When V becomes too large, the algorithm needs huge memory to store the whole matrix F , which is the limitation of the algorithm. To scale this algorithm to billions of nodes, distributed storage and update of F should be considered.

Choosing the Number of Communities. Before running Algorithm 1, we need to set the number of communities p in advance. After conducting some experiments on small datasets, we find that if we set p to be larger than the intended p and learn the parameters accordingly, our detected communities contain the results we obtain with the intended p as well as some duplicated communities or trivial communities with few nodes. Thus, our strategy is to pick a relatively large p based on the number of nodes and edges in the network and further refine our results via merging or deletion.

Algorithm 1. Overlapping community detection using *MD-NMF*

Require: G , the adjacency matrix of original graph; α , the acquaintance threshold; β , the stranger threshold**Ensure:** F , the node-community weight matrix

```

1: initialize  $F$ 
2: compute initial loss
3: repeat
4:   for  $num\_samples = 1$  to  $|E|$  do
5:     sample node  $u$  from  $V$  uniformly at random
6:     sample node  $i$  from  $u$ 's  $\alpha$ -acquaintances set  $A(u, \alpha)$  uniformly at random
7:     sample node  $j$  from  $u$ 's  $\beta$ -strangers set  $B(u, \beta)$  uniformly at random
8:     for each entry  $\Theta$  in  $F_i, F_j$  and  $F_k$  do
9:       update  $\Theta$  according to Equation (11)
10:    end for
11:  end for
12:  compute loss
13: until convergence or  $max\_iter$  is reached

```

Acquaintances and Strangers Sampling. For node $u \in V$ and any of its α -acquaintances i , if $\alpha > 0$, it is guaranteed that u and i have mutual friends. To find i , we first do a breadth-first search and group all u 's neighbors as well as friends of these neighbors into a set. Then we filter out any node k with $d(u, k) < \alpha$ in this set and sample i from the remaining nodes uniformly at random. If u does not have any α -acquaintance, we sample another u and repeat the above process until we get a valid u . To sample the β -stranger of u , we simply sample a random node from graph until we get the β -stranger. From Table 1 we can see that in each graph, the average degree of nodes is much smaller than the number of edges. Thus the time complexity sampling acquaintances and strangers of a node remains constant.

Setting Membership Threshold. For each node, to determine whether it belongs to a particular community, our strategy is to set a membership threshold t for the node-community membership matrix, i.e., if $F_{u,k} \geq t$, we say that node u is associated with community k . t is a hyper-parameter which is tuned via experimental results.

4 Experiments

4.1 Dataset

The real-world datasets we use include the two large networks we have described in the data observation section, as well as six benchmark networks collected by Newman². Table 2 lists the basic information of the six benchmark datasets. They are relatively small compared to the two large networks and have no ground-truth communities.

² <http://www-personal.umich.edu/mejn/netdata>.

Table 2. Statistics of six Newman’s datasets. $|V|$: number of nodes, $|E|$: number of edges.

Dataset	$ V $	$ E $
Dolphins	62	159
Books about US politics (Books)	105	441
American college football (Football)	115	613
Network science	1,589	2742
Power grid	4,941	6,594
High-energy theory (High-energy)	8,361	15,751

4.2 Comparison Methods

For comparison, we select the following six baseline approaches, namely *Sequential Clique Percolation (SCP)* [16], *Demon* [8], *Bayesian Non-negative Matrix Factorization (BNMF)* [23], *Bounded Non-negative Matrix Tri-Factorization (BNMTF)* [36], *BigCLAM* [34], and *Preference-based Non-negative Matrix Factorization (PNMF)* [35]. Notice that the latter four approaches are also based on matrix factorization.

4.3 Evaluation Metrics

We use modularity as the evaluation metric for small datasets without ground-truth communities and F_1 score for large datasets with ground truth communities.

Modularity. The classic modularity is defined as

$$Q = \frac{1}{2|E|} \sum_{u,v \in V} (G_{u,v} - \frac{d(u)d(v)}{2|E|}) I_{u,v},$$

where $d(u)$ is the degree of node u , $G_{u,v}$ is the (u, v) entry of the adjacency matrix G , and $I_{u,v} = 1$ if u, v are in the same community otherwise 0 [20].

In the overlapping scenario, since a node pair may share more than one communities, a minor modification has been made by replacing $I_{u,v}$ with $|C_u \cap C_v|$, i.e., the number of overlapped community between u and v :

$$\hat{Q} = \frac{1}{2|E|} \sum_{u,v \in V} (G_{u,v} - \frac{d(u)d(v)}{2|E|}) |C_u \cap C_v|.$$

From the definition, we can see that greater value of modularity reveals denser connectivity within the detected communities because only linked node pairs sharing common communities contribute positively to the value. This metric has also been frequently used in previous MF-based works [34, 35].

F_1 Score. The F_1 score of a detected community S_i is defined as the harmonic mean of

$$precision(S_i) = \max_j \frac{S'_j \cap S_i}{|S_i|}$$

and

$$recall(S_i) = \max_j \frac{S'_j \cap S_i}{|S'_j|},$$

i.e.,

$$F_1 = \frac{precision(S_i) \cdot recall(S_i)}{precision(S_i) + recall(S_i)},$$

where S'_j is one of the given ground-truth communities. The overall F_1 score of the result of detected communities is the average F_1 score of all communities in the detected communities set.

4.4 Results

For the small networks, we set the learning rate θ as 0.5 and p ranging from 10 to 50. We assume each node joins at most 3 to 10 communities and set the threshold based on this assumption. For the large network datasets, we set θ much greater because the normalized term in cosine similarity limits the altered amount of weight in each gradient descent iteration. We set p ranging from 1,000 to 5,000 and assume that each node joins at most 100 communities. The maximum number of iteration is set to be 100, while in most cases F converges before reaching the iteration limit.

Table 3. Comparison of experiment results in terms of modularity.

Dataset	SCP	Demon	BNMF	BNMTF	BigCLAM	PNMF	MD-NMF
Dolphins	0.305	0.680	0.507	0.507	0.423	0.979	1.019
Books	0.496	0.432	0.461	0.492	0.592	0.864	0.987
Football	0.605	0.540	0.558	0.573	0.518	1.049	1.163
Network science	0.729	0.642	0.661	0.741	0.503	1.657	1.695
Power grid	0.044	0.195	0.342	0.368	1.010	1.105	1.228
High-energy	0.543	0.962	0.565	0.600	0.964	0.973	1.031

Table 3 shows the results in terms of modularity on six small benchmark networks without ground-truth communities. We can see that our MD-NMF model outperforms all baseline methods on all datasets on modularity, including *LC* that leverages the general concept of “neighborhood similarity” as well and *PNMF* that is also based on a pairwise objective function but employs links as the indicator.

Table 4. Comparison of experiment results in terms of F_1 score.

Dataset	SCP	BigCLAM	PNMF	MD-NMF
Amazon	0.0315	0.0441	0.0419	0.0961
DBLP	0.0967	0.0390	0.0985	0.1013

Table 4 shows the results on two large benchmark networks with ground-truth communities. We can see that only three of our comparison methods are able to scale to networks of such size. On both Amazon and DBLP dataset, our MD-NMF model prevails on the metric F_1 score.

5 Related Work

5.1 Community Detection

Community detection has been an important line of research in physics and computer science for a long period, and many different classes of approaches are proposed to solve this problem [11]. Apart from tradition graph partitioning/clustering approaches [15], modularity-based methods are particularly designed for community detection tasks [20]. As the most well-known quality function by far, modularity can be directly optimized. Since optimizing modularity has been proven to be an NP-complete problem [6], many heuristics are proposed to solve it in polynomial time [7, 10, 19]. However, these classic community detection algorithms have a severe limitation that a node belongs to one and only one community.

Until recently, major attention has been focused on the case where communities are allowed to be overlapped [31]. According to the general strategy, overlapping community detection methods can be classified into local methods and global methods. Local methods adopt divide-and-conquer which discovers communities in small subgraphs before merging small communities into larger ones based on some criteria [8, 17, 30]. Global methods employ stochastic block models [2, 14] or community affiliation models [32] which aim to figure out the relationship between nodes and communities in a macro view. As one of the major frameworks, *Matrix Factorization (MF)* introduces a node-community membership matrix to match the adjacency matrix according to some optimization function [23, 29, 34, 36].

5.2 Mutual Friends

Mutual friend as a strong factor to indicate the closeness between two nodes has been investigated in many social-related tasks. Friend recommender systems provide the potential friends list through discovering the latent information behind network topology and friends in common [4, 25]. Link prediction models in complex networks use common neighbors to evaluate the probabilities of link establishments [5]. Online social rating networks make use of the co-commenting and

co-rating behaviors of users to recommend products and predict new rating [27]. In community detection problem, mutual friends have also been employed to measure the strength of connections between nodes. Newman defines connection strength as the normalized term of mutual friends and uses it to cluster nodes [21]. Tang and Liu directly interpret Jaccard similarity as node similarity to fit into K-means algorithm for community detection [28]. Steinhäuser and Chawla exam Jaccard coefficient as an edge weighting method and employ it in community detection. However, this algorithm fails to detect any community structure without the addition of node attribute [26]. Alvari et al. regard neighborhood similarity, i.e., the number of common neighbors, as a similarity measure and incorporate it into a game theory framework [3]. Ahn et al. explicitly give the definition of link similarity and hierarchically cluster links accordingly [1].

In this paper, mutual density has the same mathematical form as Jaccard similarity or link similarity but is used for measuring the community membership similarity. Thus, we can still calculate mutual density between two nodes even if they are not linked. Also, our model is built on the matrix factorization framework instead of link clustering.

5.3 Bayesian Personalized Ranking

The pairwise objective function of our model is based on the Bayesian Personalized Ranking [24]. This method and its extensions are originally proposed to solve the ranking problem in recommender systems [22,37]. Zhang et al. employ this model on the overlapping community detection problem [35]. They focus on the link indicator and assume that each node shares more common communities with its neighbors than its non-neighbors, which is more realistic both conceptually and experimentally.

6 Conclusion

In this paper, we propose a *Mutual Density based Non-negative Matrix Factorization* model for overlapping community detection. We introduce mutual density as a more consistent indicator of community membership similarity than links in traditional methods. The formulation of our model is based on empirical findings that mutual density correlates with the cosine similarity of community membership. Our learning objective maximizes the likelihood that each node has a more similar community membership with its acquaintances than its strangers. Experiment results show that our new model outperforms the other baseline methods as well as the link-based *PNMF* model in real-world datasets.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. *Nature* **466**(7307), 761–764 (2010)
2. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9**(1981–2014), 3 (2008)
3. Alvari, H., Hashemi, S., Hamzeh, A.: Detecting overlapping communities in social networks by game theory and structural equivalence concept. In: Deng, H., Miao, D., Lei, J., Wang, F.L. (eds.) *AICI 2011. LNCS*, vol. 7003, pp. 620–630. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-23887-1_79](https://doi.org/10.1007/978-3-642-23887-1_79)
4. Armentano, M.G., Godoy, D.L., Amandi, A.A.: A topology-based approach for followees recommendation in twitter. In: *Workshop Chairs*, p. 22 (2011)
5. Backstrom, L., Leskovec, J.: Supervised random walks: predicting and recommending links in social networks. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 635–644. ACM (2011)
6. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hofer, M., Nikoloski, Z., Wagner, D.: *On Modularity- η -completeness and Beyond*. Univ., Fak. für Informatik, Bibliothek (2006)
7. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* **70**(6), 066111 (2004)
8. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Demon: a local-first discovery method for overlapping communities. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 615–623. ACM (2012)
9. De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: On facebook, most ties are weak. *Commun. ACM* **57**(11), 78–84 (2014)
10. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. *Phys. Rev. E* **72**(2), 027104 (2005)
11. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**(3), 75–174 (2010)
12. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 211–220. ACM (2009)
13. Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**(12), 7821–7826 (2002)
14. Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**(1), 016107 (2011)
15. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.* **49**(2), 291–307 (1970)
16. Kumpula, J.M., Kivelä, M., Kaski, K., Saramäki, J.: Sequential algorithm for fast clique percolation. *Phys. Rev. E* **78**(2), 026109 (2008)
17. Li, Y., He, K., Bindel, D., Hopcroft, J.E.: Uncovering the small community structure in large networks: a local spectral approach. In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 658–668. ACM (2015)
18. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **19**(10), 2756–2779 (2007)
19. Newman, M.E.: Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**(6), 066133 (2004)
20. Newman, M.E.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**(23), 8577–8582 (2006)

21. Newman, M.: Communities, modules and large-scale structure in networks. *Nat. Phys.* **8**(1), 25–31 (2012)
22. Pan, W., Chen, L.: Gbpr: group preference based bayesian personalized ranking for one-class collaborative filtering. *IJCAI* **13**, 2691–2697 (2013)
23. Psorakis, I., Roberts, S., Ebdem, M., Sheldon, B.: Overlapping community detection using bayesian non-negative matrix factorization. *Phys. Rev. E* **83**(6), 066114 (2011)
24. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461. AUAI Press (2009)
25. Silva, N.B., Tsang, R., Cavalcanti, G.D., Tsang, J.: A graph-based friend recommendation system using genetic algorithm. In: *IEEE Congress on Evolutionary Computation*, pp. 1–7. IEEE (2010)
26. Steinhäuser, K., Chawla, N.V.: Community detection in a large real-world social network. In: *Social Computing, Behavioral Modeling, and Prediction*, pp. 168–175. Springer, Boston (2008)
27. Symeonidis, P., Tiakas, E., Manolopoulos, Y.: Product recommendation and rating prediction based on multi-modal social networks. In: *Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 61–68. ACM (2011)
28. Tang, L., Liu, H.: Community detection and mining in social media. *Synth. Lect. Data Min. Knowl. Discov.* **2**(1), 1–137 (2010)
29. Wang, F., Li, T., Wang, X., Zhu, S., Ding, C.: Community discovery using non-negative matrix factorization. *Data Min. Knowl. Discov.* **22**(3), 493–521 (2011)
30. Whang, J.J., Gleich, D.F., Dhillon, I.S.: Overlapping community detection using seed set expansion. In: *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, pp. 2099–2108. ACM (2013)
31. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. *ACM Comput. Surv. (CSUR)* **45**(4), 43 (2013)
32. Yang, J., Leskovec, J.: Community-affiliation graph model for overlapping network community detection. In: *2012 IEEE 12th International Conference on Data Mining (ICDM)*, pp. 1170–1175. IEEE (2012)
33. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, p. 3. ACM (2012)
34. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 587–596. ACM (2013)
35. Zhang, H., King, I., L.M.: Incorporating implicit link preference into overlapping community detection. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. ACM (2015)
36. Zhang, Y., Yeung, D.Y.: Overlapping community detection via bounded nonnegative matrix tri-factorization. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 606–614. ACM (2012)
37. Zhao, T., McAuley, J., King, I.: 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*, pp. 261–270. ACM (2014)