OPEN ACCESS Freely available online

PLOS ONE

Uncovering Community Structures with Initialized Bayesian Nonnegative Matrix Factorization

Xianchao Tang[1]*, Tao Xu[2,3], Xia Feng[2,3], Guoqing Yang[1,3]

1 School of Computer Science and Technology, Tianjin University, Tianjin, China, **2** School of Computer Science and Technology, Civil Aviation University of China, Tianjin, China, **3** Information Technology Research Base of Civil Aviation Administration of China, Tianjin, China

Abstract

Uncovering community structures is important for understanding networks. Currently, several nonnegative matrix factorization algorithms have been proposed for discovering community structure in complex networks. However, these algorithms exhibit some drawbacks, such as unstable results and inefficient running times. In view of the problems, a novel approach that utilizes an initialized Bayesian nonnegative matrix factorization model for determining community membership is proposed. First, based on singular value decomposition, we obtain simple initialized matrix factorizations from approximate decompositions of the complex network's adjacency matrix. Then, within a few iterations, the final matrix factorizations are achieved by the Bayesian nonnegative matrix factorization method with the initialized matrix factorizations. Thus, the network's community structure can be determined by judging the classification of nodes with a final matrix factor. Experimental results show that the proposed method is highly accurate and offers competitive performance to that of the state-of-the-art methods even though it is not designed for the purpose of modularity maximization.

Citation: Tang X, Xu T, Feng X, Yang G (2014) Uncovering Community Structures with Initialized Bayesian Nonnegative Matrix Factorization. PLoS ONE 9(9): e107884. doi:10.1371/journal.pone.0107884

Editor: Peter Csermely, Semmelweis University, Hungary

Received March 30, 2014; **Accepted** August 19, 2014; **Published** September 30, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Data are uploaded to Figshare and are available at: (http://dx.doi.org/10.6084/m9.figshare.1149965).

Funding: This work is supported by the State Key Program of National Natural Science of China (61139002), National High Technology Research and Development Program of China (2012AA063301), the Fundamental Research Funds for the Central Universities (3122013P013, 3122013C005). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: riemannfly@163.com

Introduction

Many complex systems in the real world have the form of networks whose edges are linked by nodes or vertices. Examples include social systems such as personal relationships, collaborative networks of scientists, and networks that model the spread of epidemics; ecosystems such as neuron networks, genetic regulatory networks, and protein-protein interactions; and technology systems such as telephone networks, the Internet and the World Wide Web [1]. In these networks, there are many sub-graphs, called communities or modules, which have a high density of internal links. In contrast, the links between these sub-graphs have a fairly lower density [2]. In community networks, sub-graphs have their own functions and social roles. Furthermore, a community can be thought of as a general description of the whole network to gain more facile visualization and a better understanding of the complex systems. In some cases, a community can reveal the real world network's properties without releasing the group membership or compromising the members' privacy. Therefore, community detection has become a fundamental and important research topic in complex networks.

In recent decades, a number of methods have been developed for community detection in which an objective function is maximized or minimized. One of these community detection methods is nonnegative matrix factorization (NMF), which was proposed by Lee and Seung [3]. Using the matrix factorization method, one can find the community membership of each vertex in a network. Several improvements of the NMF have been proposed, such as the Bayesian nonnegative matrix factorization (BNMF) approach for identifying overlapping communities, which was presented by Psorakis et al. [4]; the symmetric nonnegative matrix factorization (SNMF) technique for detecting overlapping communities proposed by Wang et al. [5]; and the bounded NMF (BNMTF) technique for community detection proposed by Zhang and Yeung [6]. NMF is a nonconvex optimization problem with the inequality constraints shown in Eq. (1), and iterative methods are required to obtain the solution.

$$\min_{W,H}\|A-WH\| \qquad W,H\geq 0 \tag{1}$$

However, the current NMF methods converge slowly and at local minima [7]. Most of the algorithms in the literature randomly initialize W and H. The results of these algorithms are not unique when using different initializations, such as those obtained using BNMF to detect a karate network, which is shown in Figure 1. Therefore, several instances are needed to obtain a better solution; however, this process is expensive.

Several methods have been adapted for initializing NMF. For example, Meyer et al. [8] use the *"random Acol"* method, which

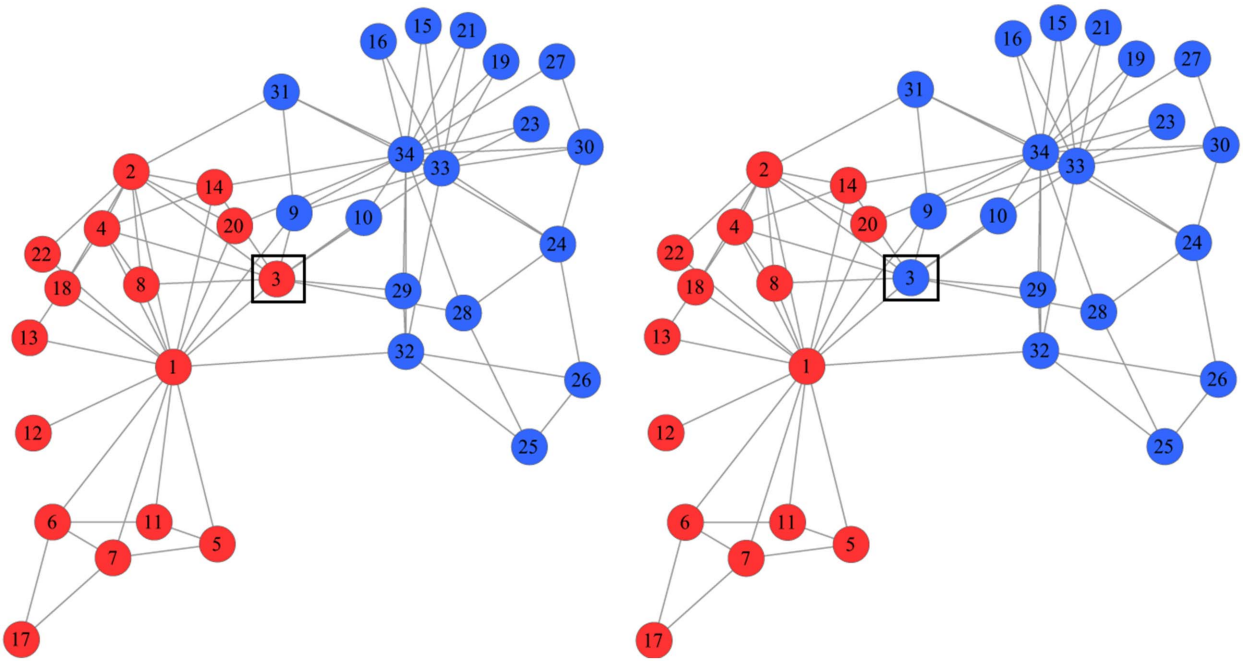

Figure 1. A comparison of BNMF with two random initializations.
doi:10.1371/journal.pone.0107884.g001

takes the average of p random rows as the initialization for NMF. Wild et al. [9] use *"Clustering Centroid"*, which uses the centroid vector for initialization. Another important initialization method is NNDSVD (nonnegative double singular value decomposition), which was proposed by C. Boutsidis and E. Gallopoulos [7]. NNDSVD uses the rank-2 matrix with the nearest positive approximation as its initialization and obtains better results than other initialization methods.

In this paper, we present a novel and running time efficient method for community detection based on BNMF with a simple NNDSVD approximation as the initialization, which we call IBNMF, to determine the community membership. The merits of this approach are as follows: *i*) computationally efficient and stable, *ii*) high accuracy in determining the membership of networks, and *iii*) overcoming the drawbacks of the maximum modularity criterion.

Methods

In this section, we introduce the community discovery framework of our method. Then, we test the performance of our approach on a range of synthetic networks and real-world benchmark examples and provide experimental evidence of the effectiveness of the proposed algorithm.

Community Discovery Framework of IBNMF

Our community discovery framework for complex networks is shown in Figure 2. First, we construct the networks' adjacency matrix from the original data. Then, using the simple NNDSVD method, the initialization of W and H can be obtained. Thereafter, we combine the initialized W and H and BNMF to acquire the final matrix factor W after several iterations. Lastly, the matrix factor is used to determine the community membership.

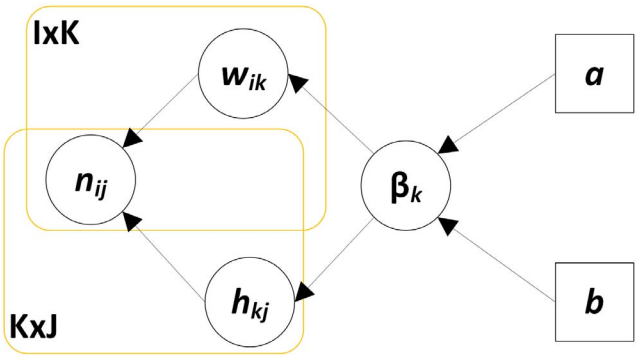

Figure 3. A directed graphical model illustrates BNMF. This graphical model describes the generation of N from W and H with the components' scale hyperparameters β_k; a and b are fixed parameters.
doi:10.1371/journal.pone.0107884.g003

Adjacency matrix. For a given non-weighted undirected network $G(V,E)$ whose vertex set is V and whose edge set is E, we use an adjacency matrix N to describe the network. When nodes i and j are connected by an edge, the element n_{ij} is set to 1; otherwise, this element is set to 0. The diagonal elements are

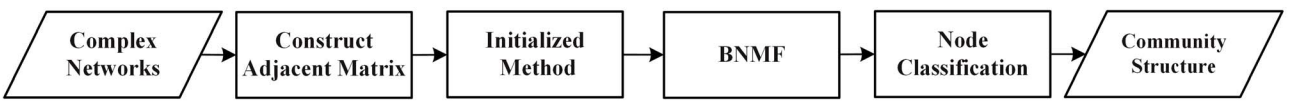

Figure 2. The community discovery framework of IBNMF.
doi:10.1371/journal.pone.0107884.g002

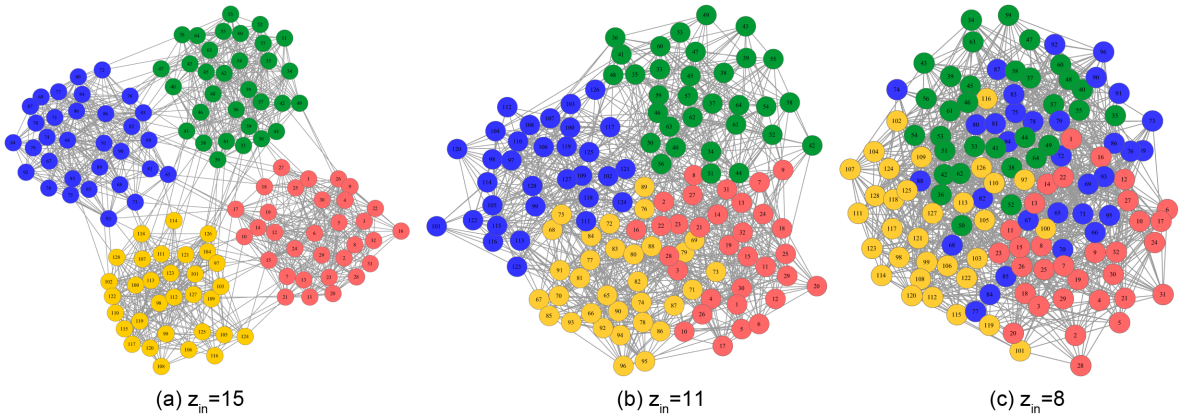

Figure 4. Three different parameter GN benchmark networks. Each graph contains 128 vertices, and each vertex is connected to exactly $z=16$ others. These vertices are divided into four separate communities: z_{in} is the number of intra-community connections made to members of its own community whereas $z_{out}=z-z_{in}$ is the number of an inter-community connections with other communities.
doi:10.1371/journal.pone.0107884.g004

usually 0, but by considering the difference between the zero elements on the diagonal and off-diagonal, we set the degree of node i as the value for each diagonal element n_{ii}.

Node classification. For an $n\times k$ factor matrix W, n is the number of total nodes in the network, k is the number of the sub-networks in the social network, and the element w_{ij} represents the probability of the $i-th$ node being in the $j-th$ community. In this work, we select the principle of probability maximization to determine the community to which the node belongs: if w_{ij} is the largest element in the $i-th$ row $w_{i:}$, then node i is part of community j.

In the following section, we give the theoretical foundations of the singular value decomposition (SVD) initialization method and the IBNMF algorithm.

Simple NNDSVD Initialization

The SVD [10] of an $m\times n$ matrix A involves the factorization of A into three matrices $A=U\Sigma V^T$, where both U and V are orthogonal matrices, and Σ is an diagonal matrix with following form:

$$\Sigma=\begin{pmatrix}\sigma_1 & & & & \\ & \ddots & & 0 & \\ & & \sigma_r & & \\ & & & 0 & \\ & 0 & & & \ddots \\ & & & & 0\end{pmatrix} \quad (2)$$

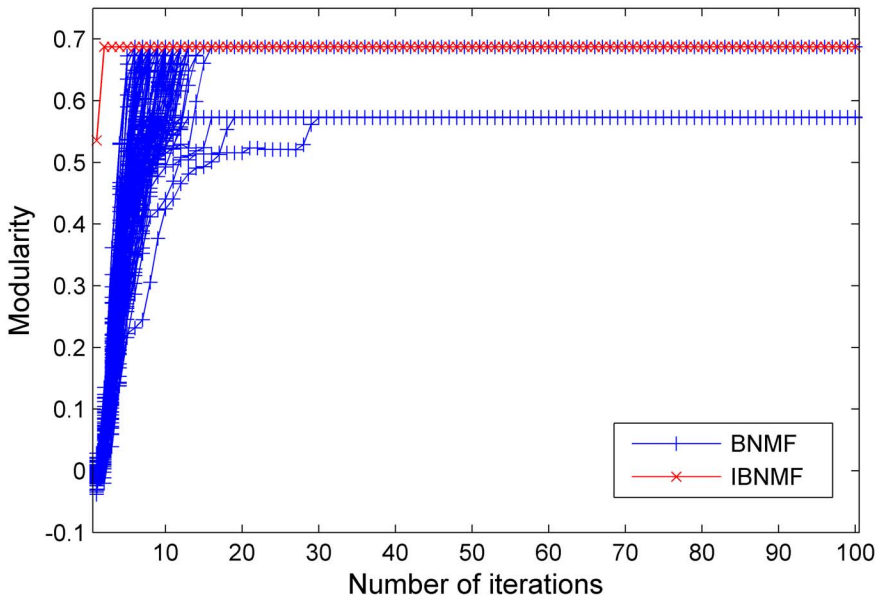

Figure 5. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for a GN benchmark network with $z_{in}=15$.
doi:10.1371/journal.pone.0107884.g005

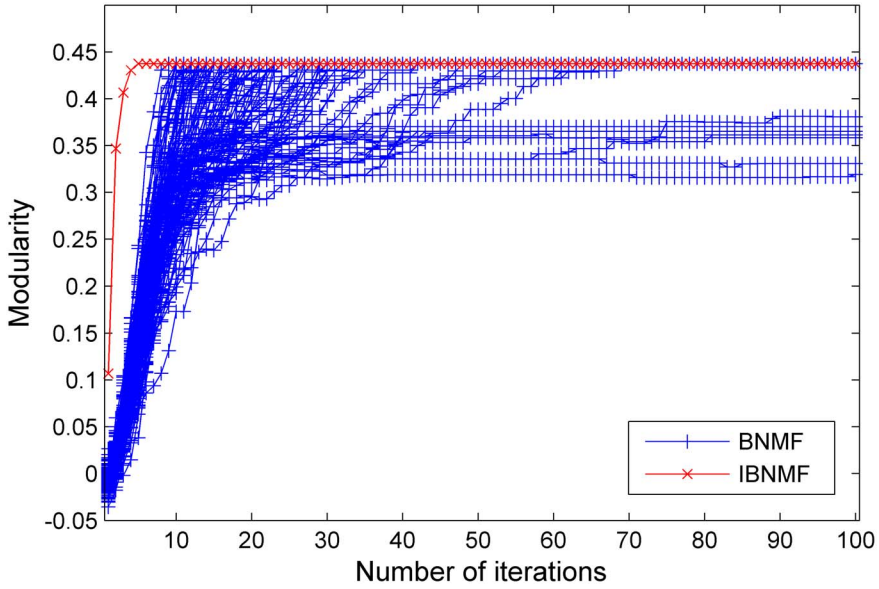

Figure 6. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for a GN benchmark network with $z_{in} = 11$.
doi:10.1371/journal.pone.0107884.g006

In the above matrix, $\sigma_1,\sigma_1,\ldots,\sigma_r$ are the singular values of A. For each $k \leq r$, the rank-k approximation of the matrix A based on Frobenius norm can be written as [7]:

$$A^{(k)} := \sum_{j=1}^{k} \sigma_j C^{(j)} = \sum_{j=1}^{k} \sigma_j u_j v_j^T, \sigma_1 > \sigma_2 > \ldots > \sigma_k \qquad (3)$$

In the F-norm, each $C^{(j)}$ can be best approximated by the nonnegative section $C_+^{(j)}$. We use the modification shown in expansion (3) to produce the nonnegative approximation of A and to obtain effective initial values for W and H to determine the community membership.

To reduce the running time, the following two steps are used in this paper to obtain a quick approximation of the network's adjacency matrix: first, the maximum rank of $C_+^{(j)}$ is set to 1. We use the main component $\lambda_1 e_1 f_1$ of $C_+^{(j)}$ as an approximation of $C_+^{(j)}$ because this component contains most of the information in the networks. Secondly, because $C_+^{(j)}$ is the nearest positive approximation of $C^{(j)}$, we can use $C_+^{(j)}$ as the approximation of $C^{(j)}$. Hence, if $A = U\Sigma V^T$ is the decomposition of A by SVD, then we have $u = U(:, j)$ and $v = V(j, :)$. We initialize the column and

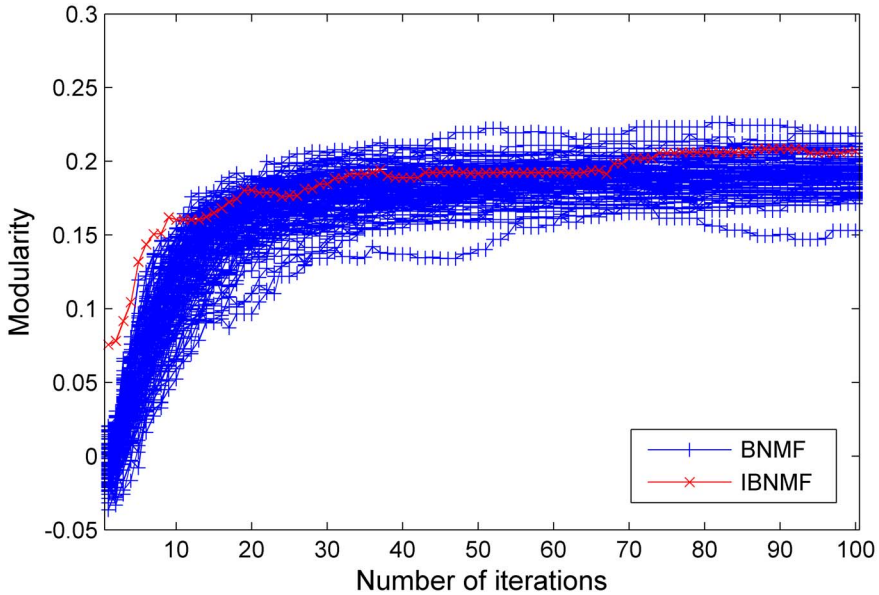

Figure 7. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for a GN benchmark network with $z_{in} = 8$.
doi:10.1371/journal.pone.0107884.g007

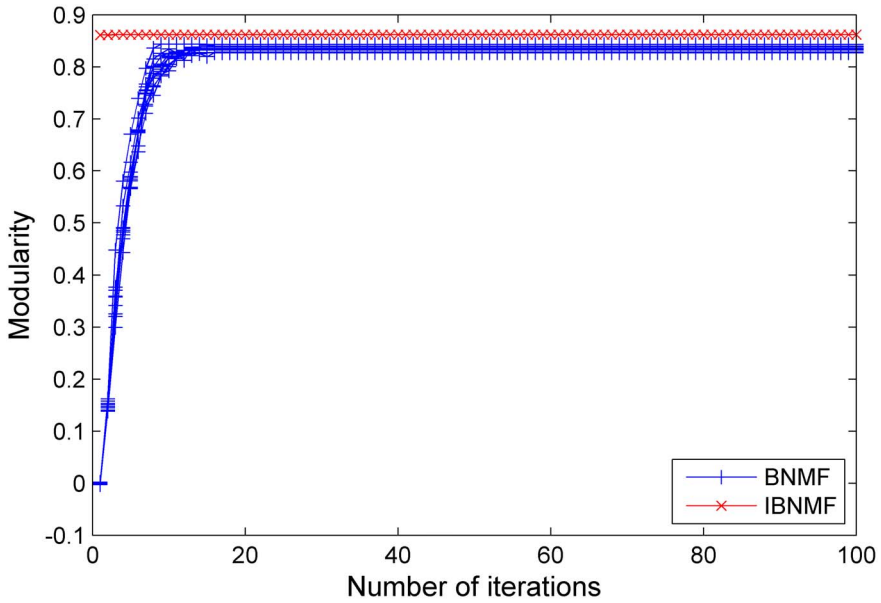

Figure 8. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for an LFR benchmark network with $\mu = 0.1$.
doi:10.1371/journal.pone.0107884.g008

row vectors in W and H, respectively, using the equations below.

$$
\begin{aligned}
W(:,1) &= \sqrt{\sigma_1}u_1 \\
H(1,:) &= \sqrt{\sigma_1}v_1^T \\
W(:,j) &= \begin{cases} \sqrt{\sigma_j}u_+ & if \|u_+\|_1 \geq \|u_-\|_1 \\ \sqrt{\sigma_j}u_- & if \|u_+\|_1 < \|u_-\|_1 \end{cases} \\
H(j,:) &= \begin{cases} \sqrt{\sigma_j}v_+^T & if \|v_+\|_1 \geq \|v_-\|_1 \\ \sqrt{\sigma_j}v_-^T & if \|v_+\|_1 < \|v_-\|_1 \end{cases}
\end{aligned}
\qquad (4)
$$

From the preceding results, it is possible to approximate the factors (W, H) as follows: *i*) perform a SVD of A with descending eigenvalues, *ii*) compute the first column and row vectors in W and H with Eq.(4), *iii*) compute the subsequent column and row vectors in W and H with Eq. (4), and *iv*) use the results as an initialization of the network's adjacency matrix.

Bayesian Nonnegative Matrix Factorization

BNMF follows the generative model in Figure 3 [11], where the detected nonnegative value n_{ij} denotes interactions occurring between two nodes i and j in the network with adjacency matrix

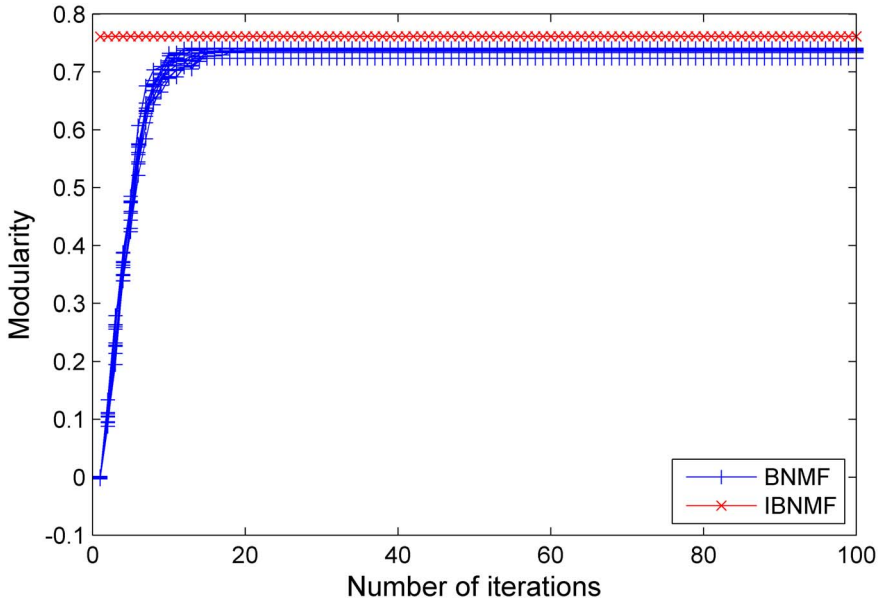

Figure 9. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for an LFR benchmark network with $\mu = 0.2$.
doi:10.1371/journal.pone.0107884.g009

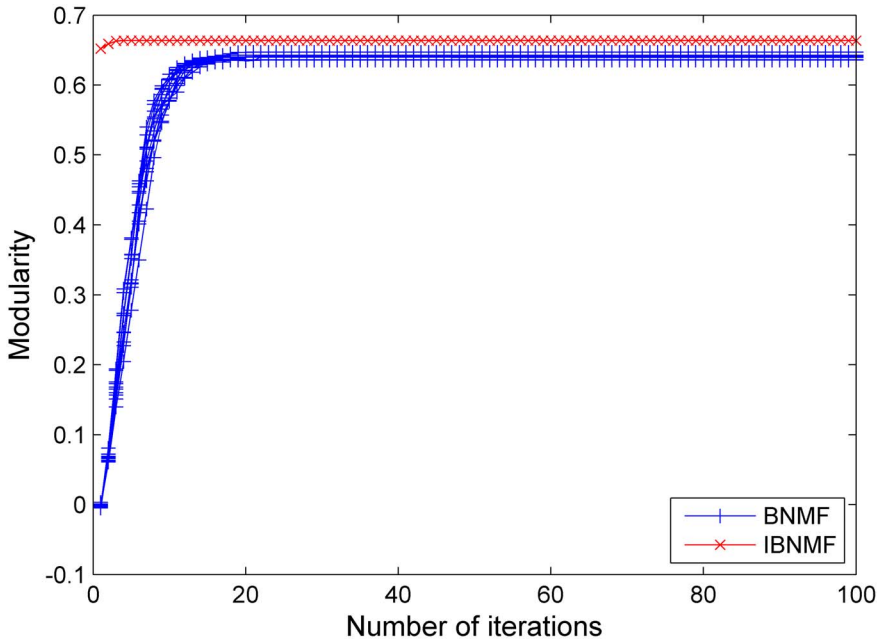

Figure 10. A comparison between our simple NNDSVD initialization method and a random initialization method. The results are given in terms of modularity for an LFR benchmark network with $\mu = 0.3$.
doi:10.1371/journal.pone.0107884.g010

$N \in \mathrm{R}_{+}^{I \times J}$. In the process of interactions, two nonnegative matrices $W \in \mathrm{R}_{+}^{I \times \mathrm{K}}$ and $H \in \mathrm{R}_{+}^{\mathrm{K} \times J}$ are found such that $N \approx \hat{N} = WH$. BNMF assumes that each single element n_{ij} of N obeys Poisson distribution at a rate $\hat{n}_{ij} = \sum w_{ik} h_{kj} (k \in \{1,...,K\})$. In the nonnegative matrices W and H, rank K is the number of groups or communities in the networks, whose initial value is unknown. By using scale hyperparameters β that control the importance of the community in both the columns of W and the rows of H [12], the values of these hyperparameters and the values of W and H can be iteratively inferred by maximizing the posterior of the parameters given by the data [13]. To be specific, the precise values of W, H and β can be obtained by optimizing the maximum a posteriori criterion:

$$\max_{W,H,\beta} p(W,H,\beta|N) \tag{5}$$

Maximizing the posterior criterion is equivalent to minimizing a cost function F in (6).

$$\begin{array}{c} \max\limits_{W,H,\beta} p(W,H,\beta|N) \\ \Updownarrow \\ \max\limits_{W,H,\beta} p(\beta)\, p(H|\beta) p(W|\beta) p(N|W,H) \\ \Updownarrow \\ \min\limits_{W,H,\beta} F = -\log p(\beta) - \log p(H|\beta) - \log p(W|\beta) - \log p(N|W,H) \end{array} \tag{6}$$

Considering the priors for W and H and the parameters' probability distribution (standard Gamma distribution over β [13], half-normal probability distribution of W and H parameterized by precision β [13–17], and Poisson distribution of N over $\hat{N}$ [11,13]), the optimization model is.

$$\min_{W,H,\beta} F = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}} + \hat{n}_{ij} - n_{ij} + \frac{1}{2} \log(2\pi n_{ij})\right) + const \tag{7}$$

According to the expression for F, the object function can be minimized by optimizing the sum of W, H, and β's log-likelihoods. Considering [2,13,18,19] and adopting the update algorithm

Table 1. Iteration times for GN benchmarks.

	GN(z_{out})							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
IBNMF	3	4	5	3	5	7	19	19
BNMF	9	11	10	11	19	16	30	39

doi:10.1371/journal.pone.0107884.t001

Table 2. Iteration times for LFR benchmarks.

	μ					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
IBNMF	1	2	2	10	14	20
BNMF	11	17	21	20	37	40

doi:10.1371/journal.pone.0107884.t002

proposed in [13], we obtain the update steps in Algorithm 1 with an algorithmic complexity of $O(NK)$.

Algorithm 1 Community Detection using IBNMF

Input:

Nonnegative matrix N, initial k, fixed Gamma hyper-parameters a, b;

Output:

Nonnegative matrices W_*,H_*;

1: [m, n] = size(N); W = zeros(m, k); H = zeros(k, n);

2: [U,S,V] = psvd (N, k);

$$3: W(:,1) = \sqrt{s_1}u_1$$

$$4: H(1,:) = \sqrt{s_1}v_1^T$$

$$5: W(:,j) = \begin{cases} \sqrt{s_j}u_+ & if \|u_+\|_1 \geq \|u_-\|_1 \\ \sqrt{s_j}u_- & if \|u_+\|_1 < \|u_-\|_1 \end{cases}$$

$$6: H(j,:) = \begin{cases} \sqrt{s_j}v_+^T & if \|v_+\|_1 \geq \|v_-\|_1 \\ \sqrt{s_j}v_-^T & if \|v_+\|_1 < \|v_-\|_1 \end{cases}$$

7: **for** each i in n_{iter} **do**

$$8: H \leftarrow \frac{H}{W^T 1_{I\times J} + diag(\beta)H} W^T \frac{N}{WH}$$

$$9: W \leftarrow \frac{H}{1_{I\times J}H^T + Wdiag(\beta)} \frac{N}{WH} H^T$$

$$10: \beta \leftarrow \frac{I+J+2(a-1)}{1_{I\times J}W^2 + H^2 1_{I\times J} + 2b}$$

11: check termination criterion: $\|W_{new} - W_{old}\| < 1e-4$;//community structure is stable.

12: **end for**

13: **return** W_*,H_*

Results and Discussion

In this section, we used both synthetic (computer-generated) and real-world networks to show IBNMF's effectiveness. The synthetic datasets enable us to test the algorithm's performance and stability, and the real datasets allow us to observe the method's accuracy under practical, real-world conditions.

Synthetic Networks

Our first synthetic network examples employ Newman's large set of artificial, computer-generated benchmark networks (GN benchmarks) [1]. Each graph was constructed with 128 vertices, and each vertex was connected to exactly $z = 16$ others. These vertices were divided into four separate communities such that some number z_{in} of each vertex's 16 connections were made to

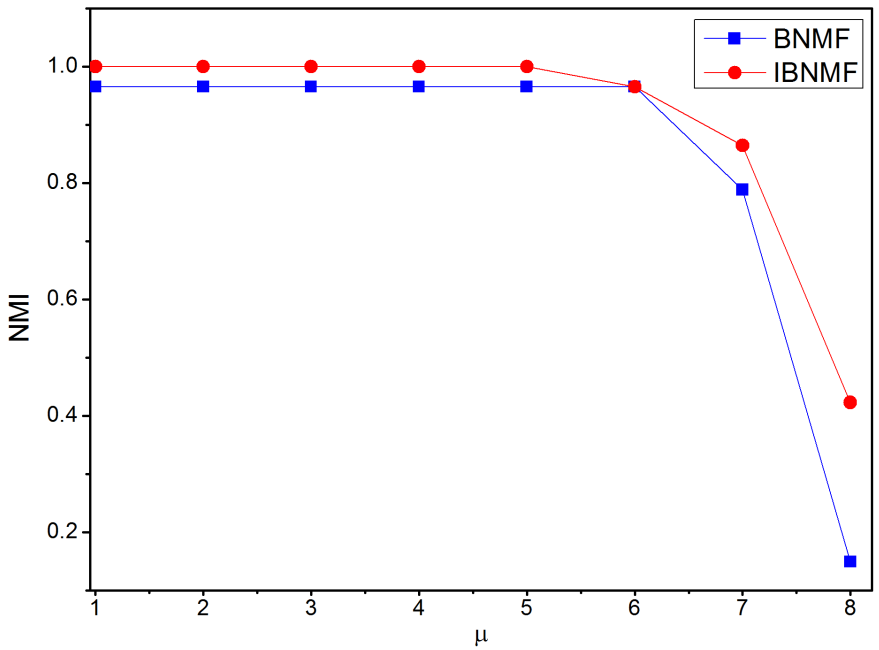

Figure 11. Average normalized mutual information for GN benchmarks.
doi:10.1371/journal.pone.0107884.g011

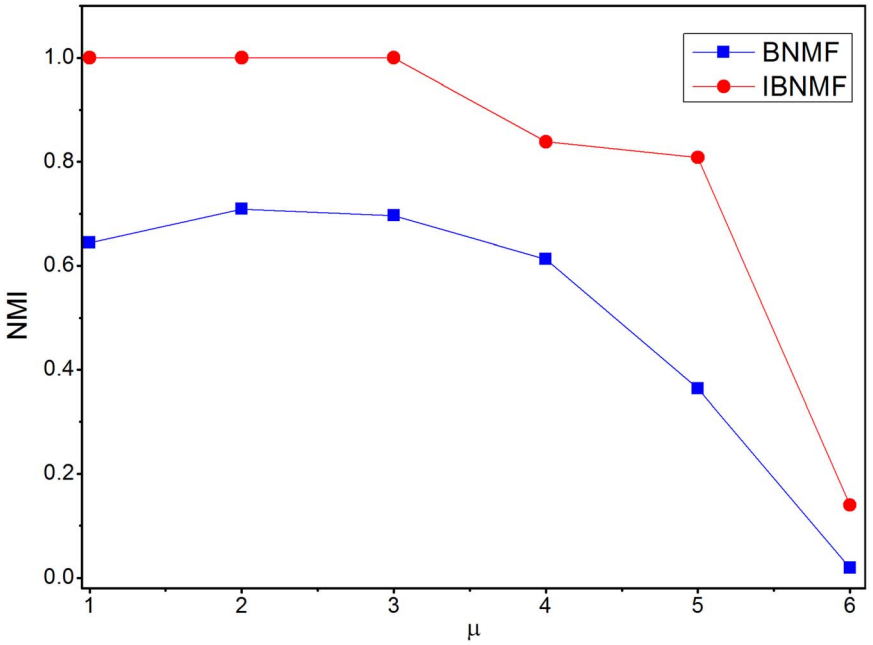

Figure 12. Average normalized mutual information for LFR benchmarks.
doi:10.1371/journal.pone.0107884.g012

randomly chosen members of its own community while the remaining $z_{out}=z-z_{in}$ connections were made to random members of other communities. This process produces graphs that have a known community structure, but are essentially random in other respects. As shown in Figure 4, when $z_{in}=15$, the vertexes have more intra-community connections than inter-community ones; when $z_{in}=11$, the vertexes also have more intra-community connections than inter-community ones; finally, when $z_{in}=8$, the vertexes have as many intra-community connections as inter-community ones. Note that in the third graph, the community structure is not clear and the vertices cannot be accurately divided into four parts as in the first and second graphs.

To evaluate the performance and stability of IBNMF with respect to determining the community structure, we choose the widely used measure called modularity Q [20,21], which can be given by:

$$Q=\sum_i (e_{ii}-a_i^2) \tag{8}$$

The modularity is the sum of the sub-modularities in different communities [20], which measures the density of intra-community connections and inter-community connections.

Using the synthetic benchmark networks, we tested the modularity and stability of our algorithm in comparison with the random initialization method (BNMF) as the ratio of intra-community connections to inter-community connections varied. After running our method and the random initialization method 100 times, we obtained the 600 sets of results shown in Figures 5, 6 and 7.

In these figures, we give the results of the two algorithms in terms of their stability and average performance as measured by the modularity. Generally, the experimental performance of IBNMF is better than that of the random initialization algorithm in terms of modularity. When $z_{in}=15$ and $z_{in}=11$, our method has a higher initial modularity and converges more rapidly to a better final result, and the final stable modularity is also higher than that of the random initialization method. Furthermore, when $z_{in}=8$, we also obtain a higher initial modularity and an average solution even though the network cannot be appropriately divided. Furthermore, the most important fact is that our method gives a

Table 3. Summary of statistics for modularity based on the use of different priors.

	Simple NNDSVD		Random	
	Mean	**Variance**	**Mean**	**Variance**
GN($z_{out}=1$)	0.6716	6.1007e–031	0.6482	0.0023
GN($z_{out}=5$)	0.4418	1.5252e–031	0.4267	8.7437e–004
GN($z_{out}=8$)	0.2347	2.5212e–031	0.2306	2.7471e–004
LFR($\mu=0.1$)	0.8617	0	0.8326	3.9899e–005
LFR($\mu=0.2$)	0.7616	0	0.7314	2.7275e–005
LFR($\mu=0.3$)	0.6636	1.3696e–032	0.6416	1.9708e–005

doi:10.1371/journal.pone.0107884.t003

Table 4. Ten real-world datasets used in this work.

Dataset	*n*	*M*	*K*	Description
Karate	34	78	2	Karate club [32]
Dolphins	62	159	2	Dolphin network [33]
Friendship6	68	220	6	High school friendship [26]
Friendship7	68	220	7	High school friendship [26]
Polbooks	105	441	3	US politics books [34]
Word	112	425	2	Word network [35]
Polblogs	1490	16718	-	Blogs about politics [36]
Football	115	613	-	American college football [21]
Net Science	1589	2742	-	Scientific collaboration networks [37]
Email	1133	5451	-	Email network [38]

doi:10.1371/journal.pone.0107884.t004

unique solution for 100 experiments, as indicated by the red lines in Figures 5, 6 and 7.

In short, when the community structure is clear, as shown in Figure 4, IBNMF obtains a stable solution that does not change as the number of iterations increases, and this solution is obtained in fewer steps than with BNMF. In addition, when the community structure is not clear, our method produces a unique solution, as represented by the red line, which is better than the BNMF results in terms of the average modularity.

Our second synthetic network examples are based on a Lancichinetti-Fortunato-Radicchi (LFR) benchmark network [22], which more accurately reflects the properties of real-world networks. In LFR benchmark networks, distributions of node degrees and community sizes follow power laws with exponents γ and β. The network cohesion is controlled by two mixing parameters $1-\mu$ and μ, which denote the fraction of a node's neighbors in its own community and the fraction of neighbors that are in the other communities, respectively. In this paper, the parameters of the LFR benchmark were set as follows: the number of nodes equals 1000, the average degree is 15, the maximum degree is equal to 50, and the mixing parameter μ ranges from 0.1 to 0.3. The number of runs is set to 10. Moreover, we evaluate the performance and stability of IBNMF using modularity; the results presented in Figures 8, 9 and 10 demonstrate that our IBNMF method has a higher initial modularity and rapidly converges to a better final result.

Sensitivity Analysis

Furthermore, we use normalized mutual information (NMI) [23] to evaluate the sensitivity of our method on synthetic networks (GN and LFR). The free parameters used here include z_{out} and μ. We vary parameter z_{out} from 1 to 8 and parameter μ from 0.1 to 0.6. The number of runs is set to 10, and the average NMI results are shown in Figures 11 and 12. From these two figures, one can observe the following: (i) the results of both the BNMF and IBNMF models decrease as z_{out} or μ increases; and (ii) IBNMF consistently outperforms BNMF on both benchmarks. From the above results, we can also see that IBNMF outperforms BNMF with respect to the iteration times. The detailed iteration times of IBNMF and BNMF that are required to obtain a steady solution are shown in Tables 1 and 2.

To analyze the sensitivity of the modularity for different priors, we perform a statistical analysis of the mean and variance by using simple NNDSVD and the random initialization, as shown in Table 3. From the experimental results, one can observe the following: (i) IBNMF obtains a higher mean modularity value than random initialization BNMF; and (ii) the simple NNDSVD initialization model is more stable than the random initialization model. The higher mean value and lower variance indicate that IBNMF has better and more stable performance for the GN and LFR benchmarks.

We have also tested our method on numerous real-world networks. In the next section, we provide detailed accuracy results of our method for the community detection of specific examples.

Table 5. A comparison of IBNMF with the Louvain, BNMTF, BNMF, RCBNMF, CBNMF, SSNMF and RSNMF methods for six real networks with FVCC.

FVCC	IBNMF	Louvain	BNMTF	BNMF	RCBNMF	CBNMF	SSNMF	SNMF
Karate	100	97.10	93.62	79.41	100	95.88	100	91.76
Dolphins	96.77	96.67	82.97	83.39	87.29	73.39	91.94	85.79
Friendship6	84.06	92.70	76.35	88.99	91.16	88.15	84.06	84.19
Friendship7	92.75	91.30	87.58	91.45	90.87	89.30	92.75	86.70
Polbooks	82.86	84.80	72.91	79.60	81.11	78.63	81.90	74.85
Word	66.38	58.95	59.58	57.34	63.71	61.20	54.46	56.10

The abbreviations of different initialized nonnegative matrix factorizations: RCBNMF is BNMF with *"random Acol"* initialization; CBNMF is BNMF with clustering initialization; SSNMF is SNMF with our initialization.

doi:10.1371/journal.pone.0107884.t005

Table 6. A comparison of IBNMF with MMSB, RN, and Infomap for six real networks with FVCC.

FVCC	IBNMF	MMSB	RN	Infomap
Karate	100	94.12	64.71	82.35
Dolphins	96.77	62.90	98.39	58.06
Friendship6	84.06	84.06	18.84	84.06
Friendship7	92.75	75.36	18.84	92.75
Polbooks	82.86	76.19	80.95	78.10
Word	66.38	55.36	49.11	51.79

doi:10.1371/journal.pone.0107884.t006

Real Networks

While synthetic networks provide a reproducible and well-controlled testing platform for our community structure algorithm, it is desirable to test the algorithm on real-world networks as well. To this end, we selected ten datasets representing real-world communities and compared the results of IBNMF with those of several state-of-the-art methods. In Table 4, our real-world network datasets are described by the vertex number n, edge number m and actual community number k. "Friendship6" network and "Friendship7" network are the same high school friendship network based on two different ground-truths [24]. All of the networks that we used here were obtained from Newman's website [25], except for "Friendship", which was obtained from Add Health in [26]. The methods that we used for comparison include the Louvain method [27], which is one of the best approaches for vertex partition [24]; Newman's fast algorithm [28], which is one of the most widely used methods for community detection; the mixed-membership stochastic block model (MMSB) [29], which is based on a Bayesian model of networks that allows nodes to participate in multiple communities; RN [30], which is based on a minimization of the Hamiltonian of a Potts-like spin model; Infomap [31], which is based on optimally compressing the information in the structure of the graph; BNMTF and SNMF methods, which are NMF based community detection ones; and other initialization methods.

To compare the performances of our method with the algorithms mentioned above, we adopt accuracy comparison and community modularity as measures for real-word datasets.

Accuracy comparisons. Various measures can be used to compare the given community structure with the one discovered by the algorithm. Here, we take fraction of vertices classified correctly (FVCC) [1], as a metric of accuracy comparison. The methods for comparison include the following: Louvain, RN, Infomap, BNMTF, and SNMF. Newman's fast algorithm is not included in this comparison, as it was not designed for FVCC. To test the influence of simple NNDSVD and a random initialization method, SNMF, SSNMF, IBNMF, and BNMF are also compared in our experiment. Furthermore, to test the influence of simple NNDSVD and other initialization methods, RCBNMF and CBNMF are also included. The abbreviations of the various initialized NMFs are introduced in Table 5.

Table 5 and 6 are the experimental results of different community detection algorithms based on FVCC index. As can be seen, IBNMF gives better results than other community detection methods and has the best performance in real-world networks. Compared with the random initialization method, simple NNDSVD initialization gives better results: both BNMF and SNMF have better performance on real-world networks. In addition, compared with other initialization methods such as *"random Acol"* and clustering, simple NNDSVD initialization also gives the best performance. In fact, IBNMF requires fewer iterations to obtain a unique result than the other initialization methods.

Modularity comparisons. As mentioned above, modularity is one of the most widely used indexes for community detection. Here, we select the modularity as our second evaluation criterion. In previous experiments, NNDSVD initialization has exhibited better performance than the other initialization methods. Thus, the methods for comparison include the Louvain method, MMSB, RN, Infomap, Newman's fast algorithm, SSNMF, and BNMTF.

Table 7. A comparison of IBNMF with the Louvain, MMSB, RN, Infomap, Newman's fast, SSNMF and BNMTF methods for nine real networks with modularity.

Dataset	IBNMF	Louvain	MMSB	RN	Infomap	Newman's fast	SSNMF	BNMTF
Karate	0.406	0.419	0.332	0.406	0.402	0.379	0.388	0.372
Dolphins	0.512	0.514	0.253	0.379	0.529	0.500	0.507	0.507
Friendship	0.586	0.590	0.500	0.400	0.595	0.585	0.583	0.524
Polbooks	0.519	0.520	0.451	0.527	0.527	0.486	0.506	0.492
Word	0.227	0.291	0.121	–0.0002	0.031	0.291	0.284	0.267
Polblogs	0.509	0.425	0.230	…	0.425	0.419	0.413	0.404
Football	0.594	0.604	0.261	0.601	0.601	0.572	0.592	0.570
Net Science	0.821	0.848	0.734	0.734	0.807	0.848	0.804	0.782
Email	0.540	0.548	0.190	0.008	0.538	0.477	0.532	0.511

doi:10.1371/journal.pone.0107884.t007

Table 7 gives the results of different algorithms in terms of the average modularity. As can be seen, our IBNMF has competitive performance even though it was not designed for the purpose of modularity maximization, unlike Louvain and Newman's fast method. Furthermore, our algorithm has the advantage of providing higher accuracy for community detection. In conclusion, our approach gives a better and more stable result than other initialization methods with a shorter running time.

Conclusions

In this paper, we present a novel method, IBNMF, for community detection, which adopts a simple NNDSVD initialization based on BNMF to achieve better and more stable results than other community detection methods. Experimental results show that IBNMF can determine the community membership in both synthetic and real-world networks. The proposed approach is more accurate and offers competitive performance to that of the RN, Infomap, Louvain and Newman's fast methods even though it is not designed for the purpose of modularity maximization. In contrast to other initialized NMF methods, our method is computationally efficient and obtains a better and more stable result with less running time.

Acknowledgments

We thank reviewers for their valuable suggestions and constructive comments. In addition, we thank Dr. Ioannis Psorakis for providing us with the Matlab code of their algorithm and Li Qiannan, Lu Min and Zuo Haichao for interesting discussions and useful advice.

Author Contributions

Conceived and designed the experiments: XT TX XF GY. Performed the experiments: XT TX. Analyzed the data: XT TX XF GY. Contributed reagents/materials/analysis tools: XT TX. Contributed to the writing of the manuscript: XT TX XF GY. Designed the software used in analysis: TX XT.

References

1. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99: 7821–7826.
2. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Physical review E 80: 056117.
3. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
4. Psorakis I, Roberts S, Ebden M, Sheldon B (2011) Overlapping community detection using bayesian non-negative matrix factorization. Physical Review E 83: 066114.
5. Wang F, Li T, Wang X, Zhu S, Ding C (2011) Community discovery using nonnegative matrix factorization. Data Mining and Knowledge Discovery 22: 493–521.
6. Zhang Y, Yeung D-Y (2012) Overlapping community detection via bounded nonnegative matrix tri-factorization. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. 606–614.
7. Boutsidis C, Gallopoulos E (2008) SVD based initialization: A head start for nonnegative matrix factorization. Pattern Recognition 41: 1350–1362.
8. Albright R, Cox J, Duling D, Langville AN, Meyer C (2006) Algorithms, initializations, and convergence for the nonnegative matrix factorization. Tech. rep. 919. NCSU Technical Report Math 81706.
9. Wild S (2003) Seeding non-negative matrix factorizations with the spherical k-means clustering. University of Colorado.
10. Jia Y-B (2013) Singular Value Decomposition.
11. Psorakis I, Roberts S, Sheldon B (2010) Soft partitioning in networks via bayesian non-negative matrix factorization. Adv Neural Inf Process Syst.
12. MacKay DJ (1995) Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6: 469–505.
13. Tan VY, Févotte C (2009) Automatic relevance determination in nonnegative matrix factorization. SPARS'09-Signal Processing with Adaptive Sparse Structured Representations.
14. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61: 611–622.
15. Roberts S, Choudrey R (2003) Data decomposition using independent component analysis with prior constraints. Pattern Recognition 36: 1813–1825.
16. Roberts S, Choudrey R (2005) Bayesian independent component analysis with prior constraints: An application in biosignal analysis. Deterministic and Statistical Methods in Machine Learning. Springer. 159–179.
17. Choudrey RA, Roberts SJ (2003) Variational mixture of Bayesian independent component analyzers. Neural Computation 15: 213–252.
18. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52: 155–173.
19. Seung D, Lee L (2001) Algorithms for non-negative matrix factorization. Advances in neural information processing systems 13: 556–562.
20. Pujol JM, Béjar J, Delgado J (2006) Clustering algorithm for determining community structure in large networks. Physical Review E 74: 016107.
21. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Physical review E 69: 026113.
22. Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E 80: 016118.
23. Lancichinetti A, Fortunato S, Kertész J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11: 033015.
24. Cao X, Wang X, Jin D, Cao Y, He D (2013) Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization. Scientific reports, 3.
25. Real-world networks we used. Mark Newman's website. Available: http://www-personal.umich.edu/~mejn/netdata/. Accessed 2014 Aug 28.
26. Weinberg BA (2007) Social interactions with endogenous associations. Technical report, National Bureau of Economic Research.
27. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008: P10008.
28. Newman ME (2004) Fast algorithm for detecting community structure in networks. Physical review E 69: 066133.
29. Gopalan PK, Blei DM (2013) Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences 110: 14534–14539.
30. Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. Physical Review E 80: 016109.
31. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105: 1118–1123.
32. Zachary W (1977) An Information Flow Modelfor Conflict and Fission in Small Groups1. Journal of anthropological research 33: 452–473.
33. Lusseau D (2003) The emergent properties of a dolphin social network. Proceedings of the Royal Society of London Series B: Biological Sciences 270: S186–S188.
34. Newman ME (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103: 8577–8582.
35. Newman ME (2006) Finding community structure in networks using the eigenvectors of matrices. Physical review E 74: 036104.
36. Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. Proceedings of the 3rd international workshop on Link discovery. ACM. 36–43.
37. Newman ME (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical review E 64: 016132.
38. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Physical review E 68: 065103.