SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
INFORMATION TECHNOLOGY
COMPUTER SCIENCE

Received
9 July 2013

Accepted
27 September 2013

Published
21 October 2013

Correspondence and requests for materials should be addressed to D.J. (jindi@tju.edu.cn)

Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization

Xiaochun Cao[1,2], Xiao Wang[1], Di Jin[1], Yixin Cao[3] & Dongxiao He[4]

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300072, China, [2]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, [3]Institute for Computer Science and Control, Hungarian Academy of Sciences, Budapest 1111, Hungary, [4]College of Computer Science and Technology, Jilin University, Changchun 130012, China.

Community detection is important for understanding networks. Previous studies observed that communities are not necessarily disjoint and might overlap. It is also agreed that some outlier vertices participate in no community, and some hubs in a community might take more important roles than others. Each of these facts has been independently addressed in previous work. But there is no algorithm, to our knowledge, that can identify these three structures altogether. To overcome this limitation, we propose a novel model where vertices are measured by their centrality in communities, and define the identification of overlapping communities, hubs, and outliers as an optimization problem, calculated by nonnegative matrix factorization. We test this method on various real networks, and compare it with several competing algorithms. The experimental results not only demonstrate its ability of identifying overlapping communities, hubs, and outliers, but also validate its superior performance in terms of clustering quality.

Many complex systems take the form of networks, sets of nodes or vertices joined together in pairs by links or edges, such as social networks, technological networks, biological networks, and information networks. The community structure is arguably the most fundamental property of most real-world networks, i.e., it is commonly observed that a group of vertices are densely connected; such a graph is called a community (or module) in literature. However, it would be oversimplifying if one assumes the communities make a partition of the whole network under concern. It errs in two senses. On one hand, it is very nature for a vertex to participate in more than one communities; i.e., communities are often overlapped[1]. On the other hand, some vertices might not participate in any community; i.e., we might have outliers. An outlier is not necessarily solitary, and it might have some negligible connection with some communities. Finally, not all vertices are born equal in a community, and some vertices of a community might be special in the sense that it is linked with almost all others; in literature, such a vertex is known as *hub*, *leader*, or *center*. Since real-world networks are inevitably huge, its analysis usually starts from the identification of its communities with overlapping under consideration. Needless to say, the community structure will greatly benefited from the simultaneous detection of hubs and outliers.

Albeit all of communities identification as well as hubs and outliers detection are well studied, we are not aware of any algorithm can furnish them together. One might be tempted to run specific algorithms and combine their results. This nevertheless does not always work, as different approaches might end with conflicting information. Herein we propose a novel method, namely CDNMF, which means "Community Detection with Nonnegative Matrix Factorization." CDNMF first describes these three types of roles using two sets of quantities, the centrality matrix of vertices and degree matrix of communities. Here a vertex's centrality represents its importance in a community, and hence the centrality matrix of vertices represents the vertices' importance in each community. An element of the degree matrix of communities, which is diagonal, indicates the degree of the community, and is equivalent to the summation of the expected degree of all vertices of this community. It then learns the two quantities by the multiplicative updating rule of NMF style. These matrices enable us to rank each vertex's centrality in each community, and use the community degree as cutting off criterion; as a result, the three types of vertices can be inferred together. Since the communities are retrieved independently, when we are working on a new community, we do not need to care whether a vertex of it belongs to a previously identified community or not. The overlapping communities are thus handled naturally. The importance of a hub in a community will ensure it to be ranked at the top of the community. After all communities have been decided, those vertices that

have not been included in any of them are outliers. In summary, CDNMF is capable of identifying overlapping communities as well as detecting hubs and outliers simultaneously.

The nature of the problem incurs a plethora of work in literature, and we only mention some new results that closely relate to us. For example, Xu *et al.*[2] introduced a method called SCAN which detects communities with overlapping vertices (which they called hubs1) and outliers in the network. Berton *et al.*[3] proposed a distance measure using random walk, and then introduced the dissimilarity index between pairs of vertices based on it. By ranking the dissimilarity index, outliers could be detected. But this method just focuses on finding the outliers, while it does not consider the detection of communities. However, it considers hubs as outliers, which seems to be unreasonable. Zhao *et al.*[4] considered that many networks contains vertices that do not fit in with any of the communities, and thus forcing all vertices into communities could distort the results. They extracted the cores of the networks and allowed for arbitrary structure in the remainder of the network, which could include weakly connected vertices, as the "background." But they defined most of the vertex in a network as outliers, which was too sweeping to be compared with the traditional definition of outliers. Chen and Saad[5] held the opinion that not every participating vertex in the network needed to belong to a community as before, and they proposed a method to extract meaningful dense subgraphs from given networks. However, they extracted dense subgraphs regardless of the rest vertices, and still their method does not have the capability to detect hubs.

Nonnegative Matrix Factorization (NMF)[6] is a feature extraction and dimensionality reduction technique in machine learning, which has been adapted to community detection recently. For example, Zarei *et al.*[7] presented a NMF-based algorithm for identifying fuzzy communities, where the new feature matrix, called the vertex-vertex correlation matrix was introduced. Psorakis *et al.*[8] presented an approach to community detection that utilizes a Bayesian nonnegative matrix factorization model to extract soft modules from networks. Wang *et al.*[9] proposed a symmetric NMF technique to detect overlapping communities in networks. Zhang and Yeung[10] proposed a community detection method called BNMTF, which is based on the bounded nonnegative matrix factorization. Using three factors in the factorization, they could explicitly model and learn the community membership of each vertex. However, the current NMF-based methods only focuses on the detection of communities, but none of them take into account the identification of vertex roles, such as hubs and outliers.

Nonnegative Matrix Factorization (NMF)[6] is a feature extraction and dimensionality reduction technique in machine learning, which has been adapted to community detection recently. For example, Zarei *et al.*[7] presented a NMF-based algorithm for identifying fuzzy communities, where the new feature matrix, called the vertex-vertex correlation matrix was introduced. Psorakis *et al.*[8] presented an approach to community detection that utilizes a Bayesian nonnegative matrix factorization model to extract soft modules from networks. Wang *et al.*[9] proposed a symmetric NMF technique to detect overlapping communities in networks. Zhang and Yeung[10] proposed a community detection method called BNMTF, which is based on the bounded nonnegative matrix factorization. Using three factors in the factorization, they could explicitly model and learn the community membership of each vertex. However, the current NMF-based methods only focuses on the detection of communities, but none of them take into account the identification of vertex roles, such as hubs and outliers.

Also of note is several related works[11–14] which adopts similar models. But rather than using loss function, they adopts the likelihood probability as the goal, and take a different algorithmic approach such as expectation-maximization algorithm to learn the model. Still, as with the above NMF-based methods, they only considers the detection of communities, and does not refer to the hubs or outliers.

Results

In this section, we demonstrate the effectiveness of our method CDNMF at exploring the three kinds of vertex roles by applying it on some real-world datasets. The experimental results verify that CDNMF can reveal rich information on these networks.

Real world networks examples. The School Friendship Network was compiled from the National Longitudinal Study of Adolescent Health[15]. It is based on self-reporting from students, which are from different grades, from grade 7 to grade 12. But in grade 9, there are two subgraphs, which correspond to the groups of white and black students, respectively.

By setting the group number $K = 6$, we fit our model to the school friendship network data. Figure 1 shows our community result, which roughly matches the ground-truths of this network. The hubs

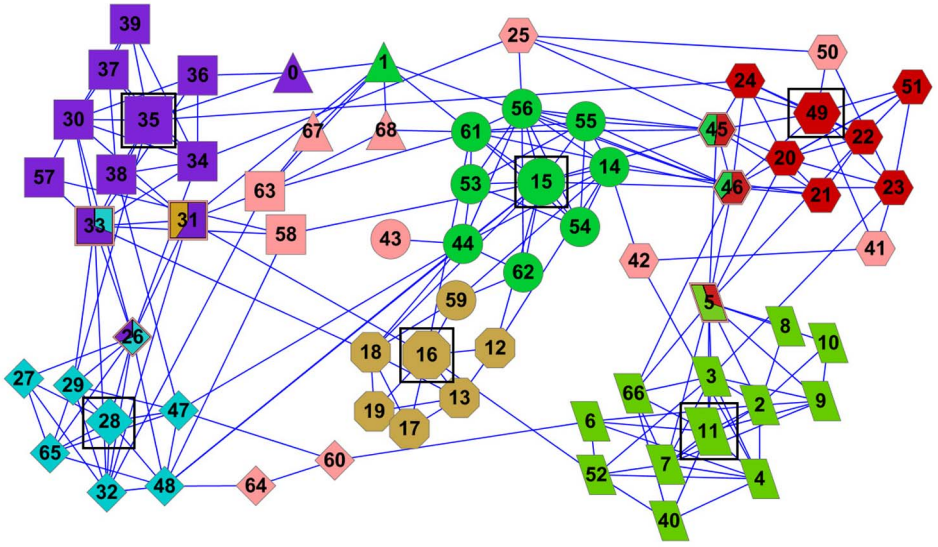

Figure 1 | Our community result. Here different shapes represent the real community membership, and different colors represent different communities detected (except pink vertices). The hubs are marked by black boxes, such as vertices 15 and 35; the overlapping vertices are shown as pie vertices, such as vertices 45 and 46; and the rest vertices, colored by pink, are outliers, such as vertices 25 and 42.

are drawn larger (in black boxes), and they all have strong links to other students in their communities. The overlapping vertices are those students who have many relations in different communities, which imply that they communicate with these communities frequently, although they still mainly belong to the grades where they are in the ground-truths. What's more, the distribution of overlapping students often lies between the adjacent grades. This phenomenon is very sensible in the reality where students in the adjacent grades often communicate more frequently than that in non-adjacent grades. Besides, Xie *et al.*[16] posed a problem that some vertices (such as vertex 42) serves as a bridge between groups but do not have particular coherence to any group, and it is still not clear whether these vertices are really meaningful or necessary to be considered as "overlapping." Here we give the answer. In our result, we consider vertices 42, 58, and 60 as outliers, and find that there are key differences between these vertices and the overlapping ones. They all have weak links to different communities, implying their relations with students either in their grades or in other grades are weak. This is a different behavior from the overlapping students, which implies these outliers are not as "gregarious" as the overlapping ones. Hence maybe they should be given more care about. Therefore, we believe it has more sense to assign such "bridge" vertices as outliers. Obviously, the identification of these three types of roles reveals more important and interesting information, and gives us a better understanding of this network.

The Dolphins Social Network was reported by Lusseau[17]. In this network, dolphins represented as vertices have a link with each other if they are observed together more often than expected by chance over a period of seven years from 1994 to 2001. It is mainly divided into the male dolphins and female dolphins, which are marked by the cycle vertices and square vertices, respectively (see Figure 2).

By setting the group number $K = 2$, we fit our model to the dolphins social network data. Our community result has been shown in Figure 2 with different colors. As we can see, "sn100" is an overlapping vertex lying between the two communities and has some links to both of them; it is thus not proper to assign it to only one community. Besides, this vertex has the highest value of betweenness[18], leading to the fission of the dolphin community of doubtful sound into subgroups, which is clearly playing an important role holding the network together. Notice that the betweenness is a measure of the influence of individuals in a network over the flow of information between others, which makes sense to consider it as an overlapping vertex of the two communities. Moreover, our method successfully finds the outliers, such as "zig," "smn5," "pl," most of which possess a same behaviour that just have one or two links with each community. Especially, "pl" belongs to the male community in the reality but has more links with the female community, thus some other community detection methods often misclassified it to be a female dolphin[19]. Differently, our method neither misclassifies it to the female community nor assigns it to the male community, but considers it as an outlier. This assignment provides a new insight and involves deeper understanding for this network.

The Political Books Network was compiled by Valdis Krebs[20]. This network represents books about US politics sold by Amazon. com. Edges represent frequent co-purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon. The political viewpoints of these books are given by "liberal," "neutral" and "conservative," respectively, which are taken as the ground-truth in our experiment.

By setting the group number $K = 3$, we fit our model to the political books network data. Our community result is shown as Figure 3. Because the topological structure of the "neutral" community is not clear, it's a common challenge for most community detection algorithms. Here, our method successfully finds the domains of "conservative," "neutral" and "liberal," respectively. For instance, vertex "Why America Slept" is an overlapping vertex between "conservative" and "neutral," which means it is often co-purchased by the same buyer. Although it is marked by "neutral," we infer that it may contain both of the two viewpoints but mainly belongs to "neutral" part. More interestingly, the overlapping vertices are all between "conservative" and "neutral," or between "liberal" and "neutral," but not between "conservative" and "neutral." It makes sense that the same buyer seldom buys two books with the clearly opposite political views, but has some probability to choose two books with similar or relative soft views. In addition, we find the hubs "A National Party No More" and "Bushwhacked" in the "conservative" and "liberal" communities, respectively. We guess these two books may be very popular in the two communities, which is correctly the situation in the reality. Considering the detected outliers, most of them locate at the borderline in the network. It implies they have weak links to other vertices, and probably are not as popular as other books in each community. In summary, our method can not only detect the community structure, but also provide some more useful information for this network.

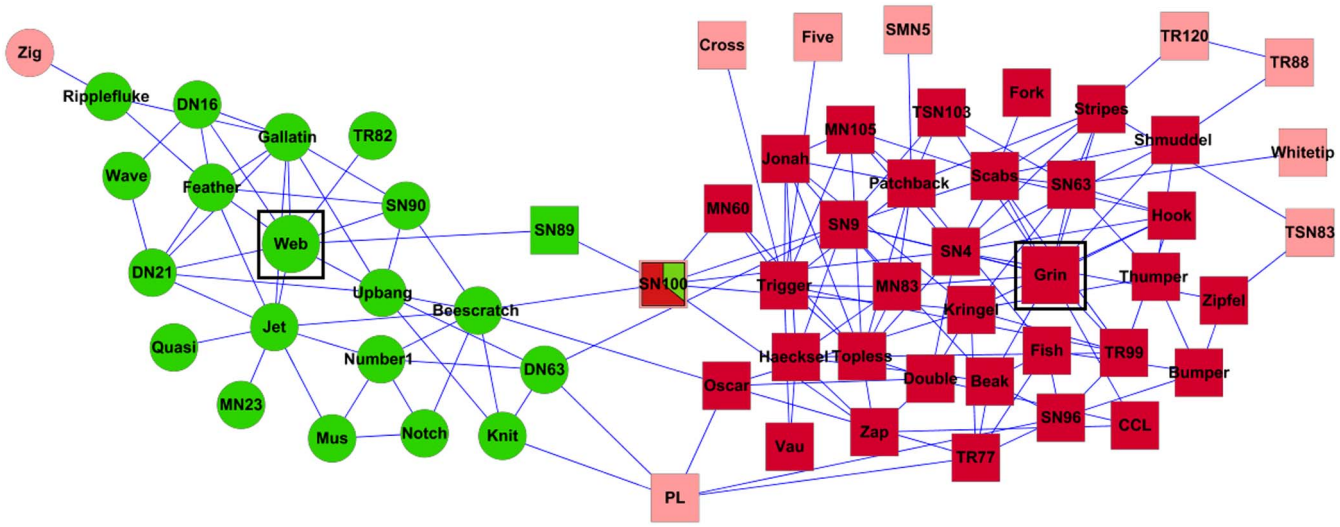

Figure 2 | Our community result for the dolphin social network. Here different shapes represent the real community membership, and different colors represent different communities detected (except pink vertices). The hubs (vertices "Web" and "Grin") are shown in the black boxes; the overlapping vertex "SN100" is shown by pie vertex; and the rest, colored by pink, such as vertices "Zig," are outliers.

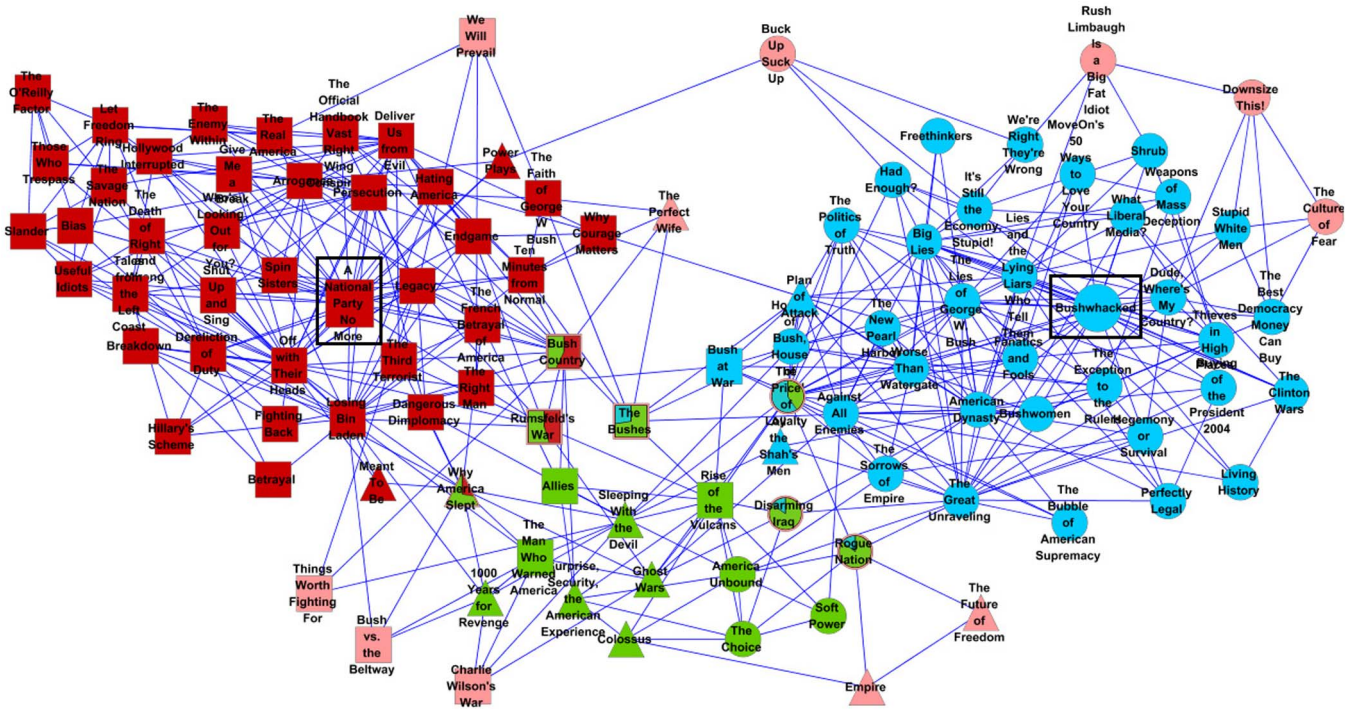

Figure 3 | Our community result for the political books network. Here different shapes represent the real community membership, and different colors represent the communities got by CDNMF (except pink vertices). Especially, the hubs are shown in the black box; the overlapping vertices are shown by pie vertices; and the outliers are marked by pink color.

The Karate Club Network[21] has become a *de facto* testbed for community detection algorithms. A disagreement developed between the administrator (vertex 33) of the club and the club's instructor (vertex 1), which ultimately resulted in the instructor's leave and starting a new club. These two groups are used as the ground-truth in our study.

By setting the group number $K = 2$, we fit our model to the karate club network data. Our community result is shown in Figure 4, which roughly corresponds to the actual communities of this network. Especially, vertices 1, 33, and 34 are hubs found by CDNMF, vertices 3, 9, 31, and 32 are overlapping vertices, and vertex 17 is an outlier. In fact, vertex 17 locates at the borderline position of the left community, and it only connects the other two unimportant vertices 6 and 7, which causes it has only weak association with this community, and thus it is considered as an outlier vertex. Differently, vertex 12 has only one link with this community, but it connects with the club's instructor (vertex 1) directly, which means it may be as well as an important vertex. For this reason, it should not be found as an outlier vertex but a community vertex, which is correctly the result of our method. Our method successfully finds the overlapping communities, the hubs, and outlier simultaneously. Therefore, it can be regarded as a helpful supplement to vertex divisions by introducing some more information from the identification of vertex roles.

Result comparisons. Here we use CDNMF on ten widely used real-world networks, and compare it with several well-known community detection methods. The networks used are shown in table 1, where n and m denotes the numbers of vertices and edges, respectively, and K denotes the actual number of communities in the network. Note that, "Friendship6" and "Friendship7" denote the same network, but they used different ground-truths; the last two networks "Jazz" and "Neural" do not have known communities. The methods compared include: Louvain method[22] which is regarded as one of the best for vertex partition, CPM (Clique Percolation Method)[1] which is the most prominent algorithm for overlapping community detection, and BNMF[8]

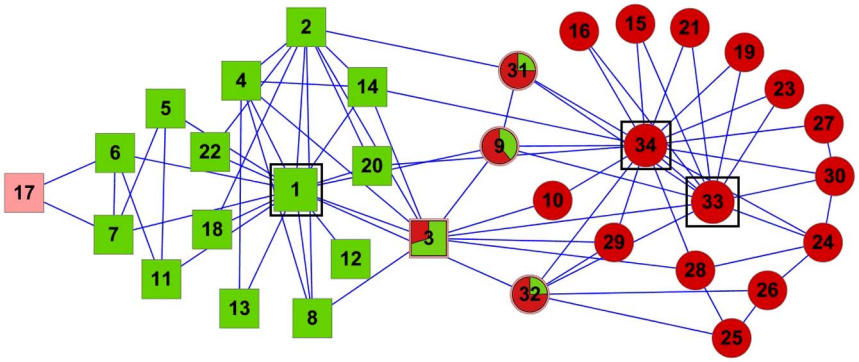

Figure 4 | Our community result for the karate club network. Here different shapes represent the real community membership, and different colors represent different detected communities (except pink vertices). The hubs (vertex 1, 33 and 34) are shown in the black boxes. The overlapping vertices, such as vertex 3, are shown by pie vertices, and the rest, colored by pink, are outliers.

Table 1 | Real-world networks used here

Datasets	*n*	*m*	*K*	Descriptions
Karate	34	78	2	Zachary's karate club[21]
Dolphins	62	159	2	Dolphin social network[17]
Friendship6	68	220	6	High school friendship network[15]
Friendship7	68	220	7	High school friendship network[15]
Polbooks	105	441	3	Books about US politics[20]
Word	112	425	2	Word network[23]
Polblogs	1,490	16,718	2	Blogs about politics[24]
Jazz	198	2,742	-	Jazz musicians[25]
Neural	297	2,148	-	Neural network of C. elegans[26]

and BNMTF[10] which both are community detection methods based on NMF. In order to sufficiently evaluate the performance of different algorithms, we adopt two sets of comparisons in terms of accuracy and community quality, respectively.

Accuracy comparisons. There are various standard measures that can be used to compare the known community structure and the one delivered by the algorithm. CDNMF does not force every vertex into a community, and some of them are detected as outliers. This situation appears for CPM algorithm in like manner. Thus, for fair comparison, we choose the widely-used FVCC, which measures the fraction of vertices classified correctly[18], as the accuracy metric here. Table 2 shows the results of different algorithms in terms of FVCC index. Notice that CPM cannot get the community result on political blogs network within 24 hours. As we can see, our method CDNMF has the best performance on four of the seven networks, and it is also competitive with the other method on the left three networks.

Quality comparisons. The second evaluation criterion is the average conductance (*AC*) of communities with weights, which extends the conductance used by Leskovec *et al.*[27], mapping the weighted value of conductance for all the communities in a cover. The conductance can be simply thought of as the ratio between the number of edges inside the community and those leaving it. More formally, the conductance is defined as follows:

$$\phi(S) = c_s / \min(Vol(S),\ \ Vol(V \backslash S)), \tag{1}$$

where $c_s = |\{(u, v):\ u \in S,\ v \notin S\}|$, $Vol(S) = \sum_{u \in S} d_u$, and d_u is the degree of vertex u. Thus, more community-like sets of vertices have lower conductance. Consequently, the *AC* can be defined as

$$AC = \frac{1}{\sum_{i=1}^{K} N(C_i)} \sum_{i=1}^{K} N(C_i)\phi(C_i), \tag{2}$$

where K denotes the number of communities, C_i denotes the ith community, and $N(C_i)$ denotes the number of vertices in C_i.

Table 3 shows the results of different algorithms in terms of average conductance. Generally, the performance of CDNMF is still better than the other four algorithms in terms of the *AC* quality. To sum up, our algorithm is very effective on real-world networks in terms of both accuracy and quality. Therefore, as we can see, CDNMF can not only detect three types of vertices roles providing richer information of networks, but also find community results with highly accuracy and quality.

Table 2 | Comparing CDNMF with Louvain, CPM, BNMF, and BNMTF on seven real networks with known community structures in terms of FVCC

FVCC (%)	Louvain	CPM	BNMF	BNMTF	CDNMF
Karate	97.06	75.00	82.35	52.94	**100**
Dolphins	96.77	**100**	83.23	67.74	98.11
School6	92.75	82.35	86.39	26.09	**96.55**
School7	91.30	82.35	85.22	36.23	**94.74**
Polbooks	84.76	**88.57**	81.52	79.05	84.54
Word	58.93	62.16	55.36	**72.32**	59.02
Polblogs	96.17	----	93.15	88.72	**97.50**

Table 3 | Comparing CDNMF with Louvain, CPM, BNMF, BNMTF, and CDNMF on nine real-world networks in terms of *AC* quality

AC	Louvain	CPM	BNMF	BNMTF	CDNMF
Karate	0.2739	0.5795	0.2282	0.8333	**0.1851**
Dolphins	0.2437	0.3748	0.1281	0.8824	**0.0537**
School6	0.2418	0.3054	0.2477	0.7579	**0.2117**
School7	0.2418	0.3054	0.2577	0.5433	**0.2375**
Polbooks	**0.0703**	0.2137	0.1546	0.4703	0.1120
Word	0.5292	0.7267	0.3645	0.8852	**0.3622**
Polblogs	0.0762	----	**0.0642**	0.0791	0.1634
Jazz	0.2587	0.614	0.5488	0.2953	**0.2281**
Neural	0.3661	0.7486	0.7373	0.4687	**0.3197**

Applications. We use our method CDNMF on two applications in biology science and cognitive psychology respectively, which are the molecular-biological network of protein-protein interactions and the network of word associations, to show its superior performance over the existing methods in solving real-world problems. Different from the networks used before, these two ones considered here possess rich metadata which describe the structural and functional roles of each vertex. Therefore, we can evaluate the performance of different methods by measuring how well the discovered community structures reflect the metadata, which seems to be more convincing than using quality metrics designed only based on network topology.

In the following, we will offer two types of comparisons for each network. The first one is with CPM[1], which is the most prominent algorithm for overlapping community detection. CPM takes some vertices of the network as background and does not classify them into any community. For fairness, when comparing with it, we take the subgraph processed by CPM as the targeted network. But the drawback of this comparison is that, it is on a subgraph rather than on a whole network. For this reason, we offer the second type of comparison with Louvain[22], which is regarded as one of the best algorithm for vertex partition by[28]. In these two comparisons, we use the number K of communities got by CPM and Louvain respectively as the given community number of our model.

Protein-protein interaction network. In the first application, we considered a protein-protein interaction (PPI) network from *Saccharomyces cerevisiae*[29]. It contains 2,640 vertices and 6,600 links, where vertices represent proteins and links denote pairwise physical interactions in the yeast.

For this network, we use the Gene Ontology (GO) terms[30], which are the most elaborate protein annotations available, as its metadata. It provides controlled vocabulary (GO terms) which describes certain aspects of protein characteristics (function, location, etc). Here we measured the quality of detected community structure by utilizing GO term enrichment analysis, which finds common biological meaning (i.e., significant shared GO terms) for the proteins within each community. Enrichment is computed using hyper-geometric test[31], and each shared GO term was assigned a p-value to quantify the significance of gene-term enrichment. For the quantitative evaluation of community structure quality, we used the average of numbers of significantly enriched GO terms (i.e., GO terms with p-value less than a threshold) for all communities as quality metric. Different thresholds for significance of gene-term enrichment may lead different results. For fairness, we set 10 different thresholds for the significance test.

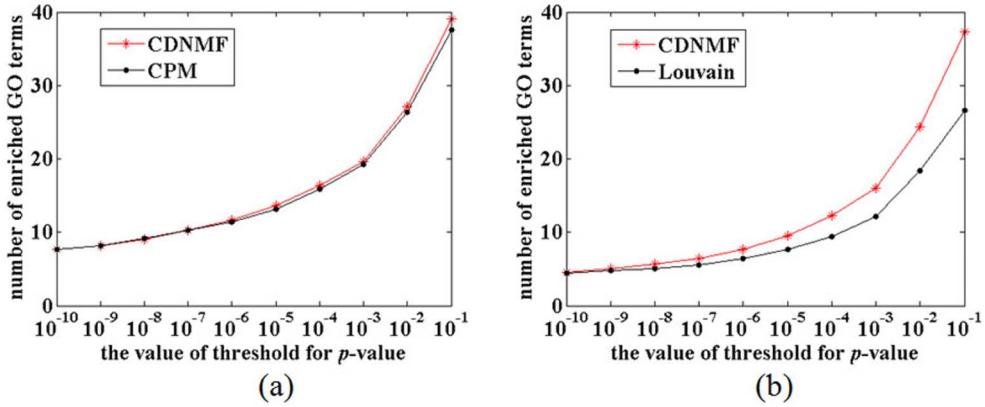

Figure 5 | Two types of comparisons in the PPI network, measured by the average of numbers of significantly enriched GO terms of communities. The 10 different thresholds were set for the significance test. (a) Comparison of our method and CPM. (b) Comparison of our method and Louvain.

The comparison of our method and CPM is shown in Figure 5(a). For seven in ten thresholds, communities attained by CDNMF always get much more enriched GO terms than that of CPM, which means our communities can better reflect the metadata. Of note, our method is run on the subgraph sifted by CPM, and the filtered network is more suitable for CPM than the original one, which makes this comparison partial to CPM. In this sense, our method still gets more significant communities. Thus, it can better show the superiority of our method over CPM.

Furthermore, the comparison of our method and Louvain is shown in Figure 5(b). It shows that our result is always better than that of Louvain as the threshold for the significance test is varied, which indicates it has better community results from the perspective of genes function.

Word association network. The network analyzed here comes from the word associations, which is constructed from the University of South Florida Free Association Norms data set[32] in the manner of[1]. This network contains 5,017 vertices and 29,148 links, where each vertex represents a word, and each link between two words indicates that people always associated one endpoint of the link with its other one.

For this network, we use the WordNet database, which is built for semantic analysis, as the metadata[33]. This database assigned a set of meanings/definitions or senses to each word (known as Synsets). We define a pair of words to be similar when they belong to a same Synset. To assess the quality of detected community structure, we compute the enrichment of vertex pair similarity[34]. Particularly, enrichment is the average metadata similarity between all pairs of vertices that share a community, divided by the average metadata similarity between all pairs of vertices. The larger the enrichment, the better the community structure is.

First, we compare our method with CPM. The enrichment of our result is 51.28, which is much larger than that of CPM (30.75). It indicates that, even in the case of using CPM's subgraph, the community result got by our method is still more reasonable than that of CPM in terms of real semantic. Thereafter, we compare our method with Louvain on the whole network. The enrichment of our result is 17.77, which is larger than that of Louvain (15.97). This result shows that, when compared with Louvain, our method can always get the better community result from the perspective of semantic analysis.

Discussion

In this work, we have proposed a novel method CDNMF based on NMF. Compared to previous work on roles identification of vertices in networks, CDNMF uses the centrality representation of vertices in each community, which enables us to identify not only all communities, but the different roles of vertices in the same run, including hubs and outliers. In contrast to other NMF-based methods, CDNMF avoids an artificial threshold in the detection of overlapping communities, which makes it much easier to implement. The experiments on various real-world networks, clearly demonstrated the superior performance of our method. We would like to draw attention to some information that are detected by CDNMF and possibly useful but missed by other algorithms that were applied on the same data sets.

Let us also point out some possible improvements to our method. In the current method, the number K of communities has to be predefined. This is not unique to our method, but commonly observed in all similar model-based methods. To surmount this obstacle, several methods have been proposed in literature, e.g., the minimum description length principle[11,14] and multi-objective optimization[35,36], neither of which, however, is compatible with our framework or can be adapted in a natural way. We leave it open for the future work.

Methods

This section describes our model and the optimization problem, presents the a simple algorithm to solve this problem, and then show how it reveals all the sought-after information.

Generative model. Let $N = (V, E)$ be an undirected and unweighted network, where V is a set of n vertices and E is a set of m edges each of which connects a pair of vertices in V. Let $\mathbf{A}$ denote the adjacency matrix of N. It is an $N \times N$ binary matrix where for $1 \le i, j \le N$, the entry A_{ij} is 1 if and only if there is an edge between vertices i and j; by definition, $A_{ii} = 0$ for any $1 \le i \le N$. Assume there can be at most K communities, and K is known a priori.

Our model will have two sets of parameters, w_z and u_{iz}. For a community z, the *soft degree* w_z of z is defined to be the sum of expected degrees of all vertices in z. For any pair of vertex i and community z, the centrality u_{iz} of i in z is defined as the expected proportion of vertex degree of i in z; by definition, $\sum_i u_{iz} = 1$. Note that we assume a soft membership of communities.

Under this model, an expected edge $<i, j>$ can be generated in the following process. First, one selects a community z with degree w_z, then community z selects vertices i, j as a pair using probabilities u_{iz}, u_{jz} respectively, finally vertices i, j form the edge. Summing over communities z, the expected number of edges that lies between vertices i and j can be written as

$$\hat{A}_{ij} = \sum_z u_{iz} w_z u_{jz}, \tag{3}$$

Using the format of matrix, (3) can be evaluated as

$$\hat{\mathbf{A}} = \mathbf{U}\mathbf{H}\mathbf{U}^T, \tag{4}$$

where $\hat{\mathbf{A}}$ denotes the expected adjacency matrix of network N. Here, $\mathbf{U} = (u_{iz})_{n \times K}$ is the centrality matrix of vertices, where each element u_{iz} denotes the centrality of vertex i in community z, subject to $\sum_i u_{iz} = 1$. And $\mathbf{H} = diag(\mathbf{w}^T)$ is the degree matrix of communities, where $\mathbf{w}^T = (w_1, w_2, \ldots, w_z, \ldots, w_K)$. Obviously, $\mathbf{H}$ is a diagonal matrix, where each diagonal element w_z denotes the soft degree of community z.

Model learning. Our task here is to learn the model mentioned before. We first define it as an optimization problem, and then infer the parameters by best fitting the observed network and the model specified in (4).

We use squared loss to measure the relaxation error. The loss function can be then formulated as

$$\begin{aligned} \min_{U,H\geq 0} L(\mathbf{A},\mathbf{U},\mathrm{H}) &= \left\|\mathbf{A}-\hat{\mathbf{A}}\right\|_F^2 = \left\|\mathbf{A}-\mathbf{U}\mathrm{H}\mathbf{U}^T\right\|_F^2. \\ s.t. \quad & \mathbf{1}_n^T\mathbf{U}=\mathbf{1}_K^T, \\ & \mathbf{H}=diag(\mathbf{w}^T), and \\ & \mathbf{w}^T=(w_1,w_2,...,w_z,...,w_K). \end{aligned} \tag{5}$$

As $\mathbf{H}$ is a diagonal matrix, the expected adjacency matrix $\hat{\mathbf{A}}$ can be rewritten as

$$\hat{\mathbf{A}}=\mathbf{U}\mathrm{H}\mathbf{U}^T=\mathbf{U}\mathrm{H}^{1/2}\mathrm{H}^{1/2}\mathbf{U}^T=(\mathbf{U}\mathrm{H}^{1/2})(\mathbf{U}\mathrm{H}^{1/2})^T=\mathrm{X}\mathrm{X}^T. \tag{6}$$

Then, we can transform the optimization problem of (5) to be an equivalent problem of nonnegative matrix factorization:

$$\min_{X\geq 0}\left\|\mathbf{A}-\mathbf{X}\mathbf{X}^T\right\|_F^2. \tag{7}$$

According to[37], by using gradient descent method, we obtain the multiplicative updating rule of NMF style for the element X_{ij} in $\mathbf{X}$:

$$X_{ij}\leftarrow X_{ij}\left(\frac{1}{2}+\frac{(\mathbf{A}\mathbf{X})_{ij}}{(2\mathbf{X}\mathbf{X}^T\mathbf{X})_{ij}}\right). \tag{8}$$

Now, the optimization of (7) is to iteratively solve (8) by choosing a set of initial values. When it converges, we can infer the model parameters using $\mathbf{X}$. Using (6), we can obtain the degree matrix of communities by

$$\mathbf{H}=(1_n^T\mathbf{X})^2, \tag{9}$$

and we then get the centrality matrix $\mathbf{U}$ of vertices by

$$\mathbf{U}=\mathbf{X}(\mathbf{H}^{1/2})^{-1}. \tag{10}$$

Identifying overlapping communities, hubs, and outliers. When having the centrality matrix $\mathbf{U}$ of vertices and the degree matrix $\mathbf{H}$ of communities, here we introduce a method for detecting the overlapping communities, hub vertices, and outlier vertices.

As each column of $\mathbf{U}$ denotes the centralities of all vertices in this community, we rank all the vertices in each column according to their values in decreasing order. For any community z, the zth column of ordered $\mathbf{U}$ is denoted by $\hat{\mathbf{U}}_z$:

$$\hat{\mathbf{U}}_z^T=(\hat{U}_1^{(z)},\hat{U}_2^{(z)},...,\hat{U}_j^{(z)},...,\hat{U}_n^{(z)}), \tag{11}$$

where $\hat{U}_1^{(z)}\geq\hat{U}_2^{(z)}\geq...\geq\hat{U}_j^{(z)}\geq...\geq\hat{U}_n^{(z)}$ and $1\leq j\leq n$. Obviously, the upper the vertex, the more important it is in this community. The corresponding vertex vector of $\hat{\mathbf{U}}_z^T$ is denoted by $\mathbf{I}_z$:

$$\mathbf{I}_z^T=(I_1^{(z)},I_2^{(z)},...,I_j^{(z)},...,I_n^{(z)}), \tag{12}$$

where $1\leq j\leq n$, and $I_j^{(i)}$ represents the vertex index corresponding to the value $\hat{U}_j^{(i)}$.

From $\mathbf{H}$, we get the expected degree of the zth community w_z. Then we add the vertex in $\mathbf{I}_z$ one by one from top to bottom to this community, until the sum of degrees of these vertices is larger than w_z. The real degree of community z is then evaluated by

$$D_p^{(z)}=\sum_{j=1}^{p}\sum_{q=1}^{n}A_{I_j^{(z)}q}, \tag{13}$$

where p indicates that number of vertices having been added in the zth community. Then the members of the zth community C_z is:

$$C_z=\begin{cases}\left\{I_j^{(z)}|1\leq j\leq p\right\}, & if\ |D_P-w_z|\leq|D_{p-1}-w_z|,\\ \left\{I_j^{(z)}|1\leq j\leq p-1\right\}, & if\ |D_p-w_z|>|D_{p-1}-w_z|.\end{cases} \tag{14}$$

As we can see, these communities will overlap with each other when they are overlapping in nature. We then get the overlapping communities.

After getting all the communities, the outliers set O in the network can be then calculated as:

$$O=V-\bigcup_{z=1}^{K}C_z, \tag{15}$$

and the hubs set B is evaluated as:

$$B=\left\{I_j^{(z)}|j=1,1\leq z\leq K\right\}. \tag{16}$$

In this way, if a vertex is a hub, it is ranked in the top in the column, sequentially, it can be easily detected. If a vertex is an outlier, it is ranked below the cut position, i.e., it links to other communities via weak relations. If a vertex resides in the overlapping region of communities r and s, it will have high centrality in both these two communities and be added to them simultaneously. In addition, in some particular applications, we may consider the first two vertices as the hubs, such as the two leaders in the cycle community in the karate network (see Figure 4). Moreover, different from other models[8,11,14], which need the specified threshold to detect overlapping communities, our CDNMF can detect three types of vertex roles including overlapping communities without any threshold.

1. Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**(7043), 814–818 (2005).
2. Xu, X., Yuruk, N., Feng, Z. & Schweiger, T. SCAN: a structural clustering algorithm for networks. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining pages 824–833 ACM (2007).
3. Berton, L., Huertas, J., Araujo, B. & Zhao, L. Identifying abnormal nodes in complex networks by using random walk measure. In Evolutionary Computation (CEC), 2010 IEEE Congress on pages 1–6 IEEE (2010).
4. Zhao, Y., Levina, E., & Zhu, J. Community extraction for social networks. *Proceedings of the National Academy of Sciences* **108**(18), 7321–7326 (2011).
5. Chen, J. & Saad, Y. Dense subgraph extraction with application to community detection. *Knowledge and Data Engineering, IEEE Transactions on* **24**(7), 1216–1230 (2012).
6. Lee, D. D. & Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999).
7. Zarei, M., Izadi, D. & Samani, K. Detecting overlapping community structure of networks based on vertex–vertex correlations. Journal of Statistical Mechanics: Theory and Experiment **2009**(11), P11013 (2009).
8. Psorakis, I., Roberts, S., Ebden, M. & Sheldon, B. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E* **83**(6), 066114 (2011).
9. Wang, F., Li, T., Wang, X., Zhu, S. & Ding, C. Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery* **22**(3), 493–521 (2011).
10. Zhang,Y. & Yeung, D. 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).
11. Ren, W., Yan, G., Liao, X. & Xiao, L. Simple probabilistic algorithm for detecting community structure. *Physical Review E* **79**(3), 036111 (2009).
12. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Physical Review E* **83**(1), 016107 (2011).
13. Ball, B., Karrer, B. & Newman, M. E. J. An efficient and principled method for detecting communities in networks. *Physical Review E* **84**(3), 036103 (2011).
14. Shen, H., Cheng, X. & Guo, J. Exploring the structural regularities in networks. *Physical Review E* **84**(5), 056111 (2011).
15. This work uses data from Add Health, a program project designed by Udry, Richard J. Bearman, Peter S. & Harris, Kathleen Mullan, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516–2524 (addhealth@unc.edu).
16. Xie, J., Kelley, S. & Szymanski, B. K. Overlapping community detection in networks: the state of the art and comparative study. *ArXiv e-prints* Oct. 2011.
17. Lusseau, D. The emergent properties of dolphin social network. *Biol. Lett. Proc. R. Soc. Lond. B* vol. **270**, pp. S186–S188 Nov. 2003.
18. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**(12), 7821–7826 (2002).
19. White, S. & Smyth, P. A spectral clustering approach to finding communities in graphs. In Proceedings of the 5th SIAM international conference on data mining pages 76–84 (2005).
20. Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**(23), 8577–8582 (2006).
21. Zachary,W. W. An information flow model for conflict and fission in small groups. *Journal of anthropological research* pages 452–473 (1977).
22. Blondel, V., Guillaume, J., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008 (2008).
23. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical review E* **74**(3), 036104 (2006).

24. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* pages 36–43 ACM (2005).
25. Gleiser, P. M. & Danon, L. Community structure in jazz. *Advances in complex systems* **6**(04), 565–573 (2003).
26. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
27. Leskovec, J., Lang, K. J. & Mahoney, M. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web* pages 631–640 ACM (2010).
28. Fortunato, S. Community detection in graphs. *Physics Reports* **486**(3), 75–174 (2010).
29. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. & Eisenberg, D. Dip: the database of interacting proteins. *Nucleic acids research* **28**(1), 289–291 (2000).
30. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**(1), 25–29 (2000).
31. Altman, D. G. Practical statistics for medical research. London: Chapman and Hall, **1991**, 396–403.
32. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida, word association, rhyme, and word fragment norms. <http://w3.usf.edu/FreeAssociation/>.
33. Fellbaum, C. ed. WordNet: An electronic lexical database. Cambridge, MA: MIT Press. 1998.
34. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**(7307), 761–764 (2010).
35. Pizzuti,C. A multiobjective genetic algorithm to find communities in complex networks. *Evolutionary Computation, IEEE Transactions on* **16**(3), 418–430 (2012).
36. Gong, M., Ma, L., Zhang, Q. & Jiao, L. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. *Physica A: Statistical Mechanics and its Applications* (2012).
37. Wang, D., Li, T., Zhu, S. & Ding, C. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* pages 307–314 ACM (2008).

Acknowledgements

This work was supported by National Basic Research Program of China (2013CB329305), National Natural Science Foundation of China (61332012, 61303110), 100 Talents Programme of The Chinese Academy of Sciences, Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06030601), and Open Project Program of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (93K172013K02).

Author contributions

X.C., X.W., D.J. and Y.C. designed the research; X.W., D.J. and D.H. performed the research, analyzed the data and prepared the figures and tables; all authors reviewed the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Cao, X., Wang, X., Jin, D., Cao, Y. & He, D. Identifying overlapping communities as well as hubs and outliers via nonnegative matrix factorization. *Sci. Rep.* **3**, 2993; DOI:10.1038/srep02993 (2013).