Finding Fuzzy Communities in Directed Networks

Kun Zhao, Shao-Wu Zhang and Quan Pan

School of Automation, Northwestern Polytechnical University, Xi'an, China

Abstract. To comprehend the directed networks in a fuzzy view, we introduce a new matrix decomposition approach that reveals overlapping community structure in weighted and directed networks. This method decomposes a directed network into modules by optimally decomposing the asymmetric feature matrix of the directed network into two matrices separately representing the closeness degree from node to community and the closeness degree from community to node. Their combined result uncovers the community structures in a fuzzy sense in the directed networks. The illustrations on an artificial network and a word association network give reasonable results.

1 Introduction

A cogent module representation of a network will retain the important information about the network and highlight the underlying structures and the relationships in the network. Many researches have been devoted to the development of algorithmic tools for discovering communities [1]. Nearly all of these methods, however, are not intended for the analysis of directed network. Yet, directedness is an essential feature of many real networks. Ignoring direction may reduce considerably the information that one can extract from the network structure. In particular, neglecting link directedness when looking for communities may lead to partial, or even misleading, results. Very few algorithms [2], [3], [4] currently available are able to handle directed graphs, since the presence of directed links places a serious obstacle towards community detection problems.

Another subject that attracts much attention in network studies is the detection of overlapping communities, or fuzzy clustering. Specifically, many real world networks exhibit an overlapping community structure, which is hard to grasp with the classical graph clustering methods [5], [6], [7] where every node of the graph belongs to exactly one community. Up to now, only a small number of studies [8], [9], [10] have addressed the problem of overlapping community. Typically, there is an algorithm takes symmetrical non-negative matrix factorization (s-NFM) [10] into optimization framework and achieves explicit physical meaning for the clustering results, which are helpful for the network analysis after clustering. As for directed networks, however, symmetrical factorization could not treat the asymmetry which is resulted from the directedness in edges. To solve this problem, we constructed a new optimization framework based on the approximation to the directed feature matrix with matrices of two types of directed paths. In order to complete the framework, we also proposed a directfied and fuzzified variant of the modularity function first

Zhao K., Zhang S. and Pan Q. (2010).
Finding Fuzzy Communities in Directed Networks.
In *Proceedings of the 6th International Workshop on Artificial Neural Networks and Intelligent Information Processing*, pages 3-12

introduced by Newman [11]. New function provides a reasonable basis for the determination of the optimal number of communities. The clustering results contain abundant information and equally possess explicit physical meaning. We tested our method on a computer-generated graph and a real-world graph and gained significant and informative community divisions in both cases.

2 The Algorithm

2.1 Optimization Scheme for Directed Graphs

Consider a directed and weighted network $G(N,E)$, which can be described by the weighted adjacency matrix $A=[A_{ij}]_{n\times n}$ where n is the number of nodes, and $A_{ij}>0$ if and only if $(i,j)\in E$ and 0 otherwise. Let the feature matrix of G be $Y=[Y_{ij}]_{n\times n}$ where Y_{ij} denotes the similarity from node i to node j. Note that the relationship between a pair of nodes is easy to grasp in the sense of connecting path. As the path linking a pair of nodes increases, the relationship of the pair is enhanced. Then, in this paper, we make the path number as the central metric of various relationships in network.

Undirected graph is defined as a graph in which edges have no orientation. It is, therefore, no need to distinguish between the paths that start from a given node and the paths that arrive in it because they are essentially the same in undirected graphs. However, in directed graphs, these two types of paths are usually not equivalent since not all edges are bidirectional. Suppose that n nodes can be grouped into r overlapping communities. Here we introduce the concept of node-community similarity matrix $U^{\circ}=[U^{\circ}_{ik}]_{n\times r}$, which is non-negative, to represents the number of paths (or the similarity degree) from nodes to communities, and the concept of community-node similarity matrix $V^{\circ}=[V^{\circ}_{ik}]_{n\times r}$, which is non-negative, to represents the number of paths (or the similarity degree) from communities to nodes. Generally, U° concerns the outgoing edges of node, and V° concerns the incoming edges of node. Fig. 1 illustrates the difference between the two types of paths in directed network.

Fig. 1. Schematic illustrations of the two types of paths in directed graph (community → node and node → community).

Since U° and V° respectively denote the number of paths from node to community and the number of paths from community to node, $U^{\circ}V^{\circ T}$ could further be an approximation of similarity between nodes. That is, we can use U° and V° to reconstruct Y:

$$U^{\circ}V^{\circ T} \to Y \tag{1}$$

For convenience, we hope to have the following approximation form:

$$\overline{U}\overline{S}\overline{V}^{T} \to Y \tag{2}$$

where $\overline{U}$ and $\overline{V}$ are non-negative matrix which are separately the column Frobenius normalization form of U° and V°, and the $r \times r$ non-negative diagonal matrix $\overline{S}$ stores the weights of the columns of $\overline{U}$ and $\overline{V}$. Note that Equation (1) is essentially equal to Equation (2), then

$$\overline{U}\overline{S}\overline{V}^{T} = U^{\circ}V^{\circ T} \tag{3}$$

Equation (2) leads us to the following Frobenius norm (Euclidean distance equation), which measures the fitness of the given matrices $\overline{U}$, $\overline{V}$ and $\overline{S}$ of graph *G(V,E)* by quantifying how precisely they approximate the network structural information *Y*:

$$\min_{W \geq 0} F_G(Y, \overline{U}, \overline{S}, \overline{V}) = \left\| Y - \overline{U}\overline{S}\overline{V}^{T} \right\|_{Fro}^{2} = \frac{1}{2}\sum_{ij}[(Y - \overline{U}\overline{S}\overline{V}^{T}) \circ (Y - \overline{U}\overline{S}\overline{V}^{T})]_{ij} \tag{4}$$

where $A \circ B$ means the Hadamard product (or element-by-element product) of matrices *A* and *B*.

Now the community detection problem is reduced to the optimization of F_G. In other words, we must find the optimal $\overline{U}$, $\overline{V}$ and $\overline{S}$ to minimize F_G. To solve this optimization problem, we will develope a modified Non-negative Singular Value Decomposition.

2.2 Method of Compressed Non-negative Singular Value Decomposition

Matrix factorization plays an important role in scientific computation. The commonly used one is singular value decomposition (SVD) [12]. It approximates one matrix with three lower rank matrices (including one rectangular matrix and two square matrices) with orthogonality constraints, in which the left and right singular vectors correspond to the column and row spaces of the original matrix. SVD has been successfully applied in both science and engineer areas [13]. However the results by SVD on real data always lose physical meaning because they usually contain negative values, and this can not be interpretable easily from intuitive insight. To make the results more interpretable, Liu [14] took the non-negativity constraints into SVD and developed a non-negative SVD (NNSVD).

NNSVD is very help for our theme. We only need to make appropriate modification on it. Firstly, both of $\overline{U}$ and $\overline{V}$ are not square matrices and they do not have orthogonality constraints. Secondly, $\overline{U}$ and $\overline{V}$ must be achieved by the normalization after matrix factorization. We take the above conditions in NNSVD and propose the iteratively update rules of a compressed Non-negative Singular Value Decomposition (c-NNSVD):

$$\begin{cases} U_{k+1} = U_k \circ \dfrac{[YV_kS_k]}{[U_kS_kV_k^TV_kS_k]} \\ S_{k+1} = S_k \circ \dfrac{[U_kYV_k]}{[U_k^TU_kS_kV_k^TV_k]} \\ V_{k+1} = V_k \circ \dfrac{[Y^TU_kS_k]}{[V_kS_kU_k^TU_kS_k]} \end{cases} \tag{5}$$

where U_k and V_k are $n \times r$ non-negative matrix and S_k is $r \times r$ non-negative diagonal matrix. The iteration starts from random matrices which are chosen from a normal distribution with mean 0, variance 1. $\overline{U}$ and $\overline{V}$ are obtained by the column normalization of *U* and *V* which are the optimal solution of iteration rules in Equation (4); and the non-negative diagonal matrix $\overline{S}$ stores the weights of the columns of $\overline{U}$ and $\overline{V}$. The *i*'th diagonal element of $\overline{S}$ corresponds to the *i*'th community, or the *i*'th column of matrix $\overline{U}$ and $\overline{V}$.

According to Equation (3), to gain the number of paths from nodes to communities (U°) and the number of paths from communities to nodes (V°), the weights in $\overline{S}$ should be properly assigned to U° and V°. Intuitively, the weight of the input paths of node should be quantitively equated with the weight of the output paths of node. Therefore, the weights in $\overline{S}$ are equally distributed by:

$$\begin{cases} U^\circ = \overline{U}\overline{S}^{\frac{1}{2}} \\ V^\circ = \overline{V}\overline{S}^{\frac{1}{2}} \end{cases} \tag{6}$$

By Equation (6), we gain the number of paths from nodes to communities and the number of paths from communities to nodes. Note that their sum can produce an integrated closeness degree between nodes and communities, which is necessary for the directed network analysis:

$$W^\circ = U^\circ V^\circ = (\overline{U} + \overline{V})\overline{S}^{\frac{1}{2}} \tag{7}$$

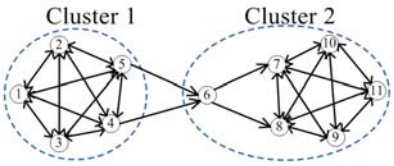
If one do not want to separately consider U° and V°, the integrated quantity, W°, would give a consolidated result which combines the two directions. To specifically illustrate the difference among U°, V° and W°, we apply our method on a 11-nodes network studied in [3], as follows:

Let *r* have a value of 2; the output of Equations (5),(6) and (7) is:

$$U^{\circ T} = \begin{bmatrix} 0.533 & 0.533 & 0.533 & 0.550 & 0.550 & 0.109 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.004 & 0.004 & 0.163 & 0.543 & 0.543 & 0.533 & 0.533 & 0.533 \end{bmatrix}$$

$$V^{\circ T} = \begin{bmatrix} 0.533 & 0.533 & 0.533 & 0.543 & 0.543 & 0.163 & 0.004 & 0.004 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.109 & 0.550 & 0.550 & 0.533 & 0.533 & 0.533 \end{bmatrix}$$

$$W^{\circ T} = \begin{bmatrix} 1.066 & 1.066 & 1.066 & 1.093 & 1.093 & 0.272 & 0.004 & 0.004 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.004 & 0.004 & 0.272 & 1.093 & 1.093 & 1.066 & 1.066 & 1.066 \end{bmatrix}$$



$$L_{ij} = \begin{cases} A_{ij}, & i \neq j \\ -d_i, & i = j \end{cases} \tag{8}$$

where d_i is the degree of node i. Diffusion kernel , the exponential of matrix L, is defined as:

$$K = \exp(\beta L) = \lim_{n\to\infty}\left(1+\frac{\beta L}{n}\right)^n = 1+\beta L+\frac{\beta^2}{2}L^2+\frac{\beta^3}{3!}L^3+\cdots \tag{9}$$

where β is a positive constant to control the degree of diffusion. In undirected networks, the resulting matrix K is symmetric and positive definite. It is a valid kernel. A similarity matrix Y can be obtained by normalizing the kernel matrix K in such a way:

$$Y_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}} \tag{10}$$

Note that, the diffusion kernel of undirected network starts with a symmetric adjacency matrix. However, in directed network, the adjacency matrix A is not symmetric. Therefore, directed networks should have an alternative form of kernel which could be traced back to a different Laplacian. The Laplacian of undirected and weighted network is the following matrix:

$$L^d_{ij} = \begin{cases} A_{ij}, & i \neq j \\ -d_i^{out}, & i = j \end{cases} \tag{11}$$

where d_i^{out} is the out degree (weighted) of node i. It is naturally an asymmetric matrix. So, its kernel matrix K^d=exp(βL^d) and the resulted feature matrix are also asymmetric. Frankly speaking, K^d do reflect the number of directed paths from one node to another in an asymmetric manner. In this paper, we choose β= 0.1 in the feature matrices in our study.

2.4 Directfied and Fuzzified Variant of the Modularity Function

If a priori knowledge of the community number is absent, the optimal number of communities should be determined by some computational methods in a self-consistent way without human intervention. Recently, a concept of modularity function Q introduced by Newman and Girvan [11] has been broadly used as a valid measure for community structure. It comes from the notion that: only if the number of edges within communities is significantly higher than would be expected purely by chance can we justifiably claim to have found significant community structure. The original modularity of a network is then defined as:

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij}-\frac{d_i d_j}{2m}\right]\cdot\delta_{C_iC_j} \tag{12}$$

where A_{ij} is an element of the adjacency matrix, δ_{ij} is the Kronecker delta symbol, and c_i is the label of the community to which vertex i is assigned.

Then one maximizes Q over possible divisions of the network into communities, the maximum being taken as the best estimate of the true communities in the network. So, the optimal number of fuzzy communities can be determined by the modularity function Q which gains its maximum on a certain value of r.

In the fuzzy clustering method of Nepusz [8], a fuzzified variant of the modularity Q is presented as:

$$Q_f = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{d_i d_j}{2m}\right]\cdot s_{ij} \tag{13}$$

where $s_{ij} = \sum_{k=1}^{r} H_{ki} H_{kj}$ and H_{ki} is the fuzzy membership degree of node i to the community k. The probability of the event that vertex i belongs to the same community as vertex j becomes the dot product of their membership vectors, resulting in the similarity measure s_{ij}, which can be used in place of $\delta_{C_i C_j}$ to obtain a fuzzified variant of the modularity.

As for directed network, Newman [2] presented a new modularity function Q_d, which is generally applicable for directed networks:

$$Q_d = \frac{1}{M}\sum_{i,j}\left[A_{ij} - \frac{d_i^{out} d_j^{in}}{M}\right]\cdot \delta_{C_i C_j} \tag{14}$$

where A_{ij} is defined in the conventional manner to be 1 if there is an edge from j to i and zero otherwise, and d_i^{out} is the out-degree of node i and d_j^{in} is the in-degree of node j and M is the total number of directed edges in the network. Indeed edge i-j make larger contributions to this expression if d_j^{in} and/or d_i^{out} is small.

Each of the above two modularity, Q_f and Q_d, has its advantages which the other one does not has. To combine their advantages, we propose another variant of the modularity Q as:

$$Q_{df} = \frac{1}{M}\sum_{i,j}\left[A_{ij} - \frac{d_i^{out} d_j^{in}}{M}\right]\cdot s_{ij} \tag{15}$$

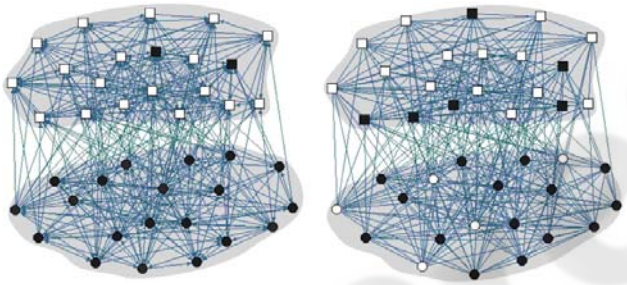
which can be applied to fuzzy clustering in directed networks.

The modularity can be either positive or negative, with positive values indicating the possible presence of community structure. One can search for community structure precisely by looking for the divisions of a network that have positive, and preferably large, values of the modularity. In order to determine the optimal number of fuzzy communities in directed networks, we iteratively increase r and choose the one which results in the highest modularity Q_{df}.

3 Test of the Method

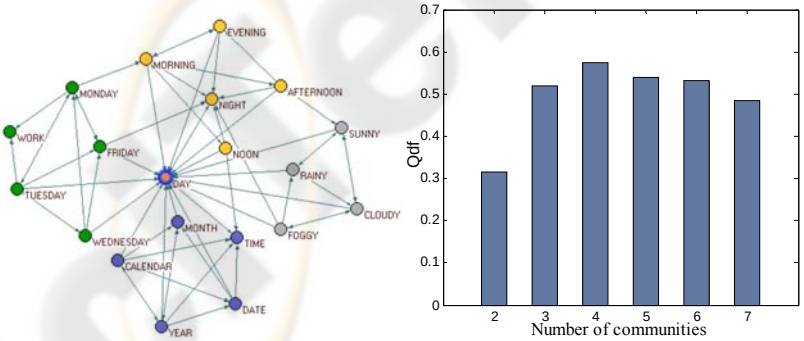
3.1 Random Graph

For illustrative purposes, we consider an artificial computer-generated network, designed specifically to test the performance of the algorithm. As Fig. 3a shows, this network is generated with N = 40 nodes, split into two communities containing 20 nodes each. We put 120 directed edges in each community at random and 120 directed edges between the two communities at random. The edges that fall within groups are biasedly assigned directions so that they are more likely to point from one group to another. As we apply c-NNSVD on this network, the two communities are detected almost perfectly: just two nodes out of 40 are misclassified. This is confirmed in Fig. 3(a), which shows the results of the application of our method. If we ignore the directions, however, using the algorithm presented in [10], there is nearly no community structure to be found in this network, as Fig. 3(b) shows.



(a) (b)

Fig. 3. Community assignments for the two-community random network described in the text from (a) the algorithm of this paper and (b) an undirected clustering algorithm in [10]. The true community assignments are denoted by vertex shape or shaded region. The different colors represent different communities obtained by the algorithms.



(a) (b)

Fig. 4. The communities of the word *DAY* in the South Florida Free Association coupled with the determination of the optimal number of communities. (a) By the method presented in this paper, the word DAY is discovered to be the overlapping node which has the largest membership degree to the yellow group and the second-largest membership degree to the blue group. (b) Histogram of Q_{df} for different choices on number of communities.

3.2 Word Association Graph

We examined a directed network obtained from the South Florida Free Association norms list [17] (containing 10617 nodes and 63788 links), where the weight of a directed link from one word to another indicates the frequency that the people in the survey associated the end point of the link with its start point. We picked a sub-network with 20 nodes from the list and chose 4 as the optimal number of clusters, see details in Fig. 4(b) which indicates that the peak for Q_{df} of 0.5745 is achieved at $r = 4$. For illustration in Fig. 4(a), we showed the (colour coded) modules of the word *DAY* obtained by c-NNSVD, with the overlap emphasized in nested color. According to its different meanings, this word participates in four, strongly internally connected modules. The green community can be associated with work days. The yellow community consists of day times, the gray community contains common adjectives of day related to weather, and the blue community can be associated with the calendar. Separately, the closeness degrees of the word *DAY* to the four communities is 0.018, 1.355, 0.019 and 0.059, which indicate that the yellow group is the dominant community of node *DAY* and the blue group follows.

4 Conclusions

In this paper we presented a new algorithm for identifying overlapping communities in directed networks based on two matrices of similarity between node and community. An integrated quantity was proposed to give a consolidated result and it was shown, through several examples that this leads to detection of the overlapping community structure of the directed network.

Acknowledgements

We thank Liu Weixiang for the useful discussion. The work was supported in part by the National Natural Science Foundation of China under Grant NO. 60775012 and NO. 60634030.

References

1. Danon, L., Duch, J., Diaz-Guilera, A. and Arenas, A. Comparing Community Structure Identification. Stat. Mech. (2005) P09008.
2. Leicht, E. A. and Newman, M. E. J. Community Structure in Directed Networks. Phys. Rev. Lett. (2008) 100, 118703.
3. Nepusz, T. and Bazsó, F. Likelihood-based Clustering of Directed Graphs. In: IEEE 3rd International Symposium on Computational Intelligence and Intelligent Informatics. (2007) Agadir, Marokkó, 28.
4. Palla, G., Farkas, I., Pollner, P., Derenyi, I. and Vicsek, T. Directed Network Modules. Phys. New. J. (2007) 186.

5. Girvan, M. and Newman, M. E. J. Community Structure in Social and Biological Networks. Proc. Natl. Acad. Sci. USA, (2002) 99, 7821−7826.
6. Ravasz, E., Somera, A. L. and Mongru. D. A. Hierarchical Organization of Modularity in Metabolic Networks. Science, (2002) 297, 1551−1555.
7. Capocci, A., Servedio, V. D. P., Caldarelli, G. and Colaiori, F. Detecting Communities in Large Networks. Physica A, (2005) 352, 669−676.
8. Nepusz, T., Petróczi, A., Négyessy, L. and Bazsó, F.. Fuzzy Communities and the Concept of Bridgeness in Complex Networks. Phys. Rev. E, (2008) 77, 1539-3755.
9. Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature, (2005) 435, 814–818.
10. Zhao,K., Zhang, S. W. and Pan, Q.. Fuzzy Analysis for Overlapping Community Structure of Complex Network. In: IEEE International Conference on Chinese Control and Decision Conference(CCDC), Submitted for publication (2010).
11. Newman, M. E. J. and Girvan, M. Finding and Evaluating Community Structure in Networks. Phys. Rev. E, (2004) 69, 026113.
12. Golub, G. H. and Van Loan, C. F. Matrix Computations (3nd ed.). Johns Hopkins University Press. (1996)
13. Jolliffe, I. *Principal Component Analysis*. Springer (2002).
14. Liu, W., Tang, A., Ye, D., and Ji, Z. Nonnegative Singular Value Decomposition for Microarray Data Analysis of Spermatogenesis. Technology and Applications in Biomedicine. (2008) 225-228.
15. Kondor, R. I. and Lafferty, J. Diffusion Kernels on Graphs and Other Discrete Structures. 19th International Conference on Machine Learning (ICML), (2002)315−322.
16. Fouss, F. and Yen, L. An experimental Investigation of Graph Kernels on Two Collaborative Recommendation Tasks. In: IEEE International Conference on Data Mining (ICDM), (2006) 18-22.
17. Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. The University of South Florida Word Association, Rhyme, and Word Fragment Norms. (1998) Retrieved from: http://www.usf.edu/FreeAssociation/.