Detecting Overlapping Community Structures with PCA Technology and Member Index

Peiyan Yuan
School of Computer and Information Engineering
Henan Normal University
Xinxiang 453007, China
Tel.: +86-159-3734-8382
peiyan@htu.cn

Wei Wang
School of Computer and Information Engineering
Henan Normal University
Xinxiang 453007, China
wangwei3600@126.com

Mingyang Song
School of Computer and Information Engineering
Henan Normal University
Xinxiang 453007, China
a.aaooppi@163.com

ABSTRACT
The community structures reflect the basic property of social networks. How to effectively detect them is difficult and important. Traditional solutions such as nonnegative matrix factorization approaches have a high time and space complexity, resulting in a poor scalability. In this paper, we propose PCA-MI, a novel method to detect the overlapping community structures. Firstly, we use the principle component analysis (PCA) technology to extract key features of network information, and then employ membership index (MI) to classify nodes. Experimental results show that our approach can fast identify the overlapping community structures and achieve approximate Module-Q values as the traditional algorithms.

Keywords
Social networks; Overlapping community structure; Principle component analysis; Membership index.

1. INTRODUCTION
In social networks, many systems can be abstracted and described by network structures, and the latter forms different communities with different sizes, detecting community structures is one of the most significant topics in research community. Community structures reflect the social relationship of nodes. In general, a group of people in the real world can be considered as communities in the virtual space, i.e., nodes within a community have close connection and those in different community have relatively sparse connection [1-5]. Furthermore, people can join multiple groups and form overlapping community structures, which act as bridges to connect disjoint parts and help to diffuse information among nodes [6-7].

In the past several years, lots of overlapping community detection algorithms have been proposed, such as clique percolation method [8], nonnegative matrix factorization [9], the improved nonnegative matrix factorization algorithm based on Bayesian [10], etc. Note that most of them used the nonnegative matrix factorization (NMF) method to classify nodes, these algorithms require long time to calculate and more spaces to store matrixes, resulting in a poor scalability. Considering this question, in this paper, we first use principle component analysis (PCA) [11] technology to reduce the matrix dimension and extract key features of network information, we then employ membership index (MI) [12] to classify nodes. One node can join multiple communities if its MI value exceeds a threshold.

The rest of this paper is organized as follow. Section 2 introduces the related work. Section 3 presents the PCA-MI algorithm. We implement the simulation and evaluate the proposed method in Section 4. Finally, Section 5 concludes the paper and discusses the future research areas.

2. RELATED WORK
With the development of online social networks, the overlapping community has recently attracted more attention and becomes an important research topic. Many algorithms have been proposed in the past years [13]. In 2002, M. Girvan and M. E. J. Newman presented the GN algorithm [14], which repeatedly deleted edges with the biggest betweenness, so as to find the network community structure. In 2004, Newman proposed the Module-*Q* [2], a network partitioning index function, it is an important parameter to measure the accuracy of the community divided. Palla et al. put forward the pioneering overlapping community detection algorithm-clique percolation method (CPM) [8], which is mainly used to find the largest connected sub graph by k-cliques. Thus, nodes in network can be divided in several cliques and one node may belong to multi-cliques. In this way, overlapping communities are detected.

Zhang et al. [9] used nonnegative matrix factorization (NMF) to realize the community division, where each node only belongs to one community. Later, Psorakis et al. [10] extracted the overlapping community by using the Bayesian NMF. Li et al. [12] further extended the work of [10]. They used the characteristic matrix, rather than the adjacent matrix to effectively calculate the number of communities, and then classified nodes by employing the relationship between nodes and communities [15-17].

3. OVERLAPPING COMMUNITY DETECTION
In this section, we present PCA-MI algorithm. We first introduce the datasets in Section 3.1. In section 3.2, we discuss how to process the original data. Finally, we illustrate the PCA-MI algorithm in Section 3.3.

3.1 Datasets
In this paper, we use the following four real datasets traces, called Dolphins, College Football and two email datasets from Shujutang [18].

Dolphin social network records the activities of New Zealand wide Kiss Dolphin, including 62 dolphins. These dolphins construct a social network based on their daily contact behaviors. If two Dolphins have a contact, there exists an edge between them. The College Football dataset reflects the representation of the schedule of the American College Football Game in the year of

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China

DOI 10.4108/eai.18-6-2016.2264206

2000 regular season. Nodes represent teams and edges denote whether two teams meet in the game. In the two Email datasets, one edge appears when two users send emails to each other. Table 1 summarizes the basic properties of the four datasets.

Table 1. Real datasets

Dataset	Number of nodes	Number of edges
Dolphins[10]	62	159
College Football [14]	115	613
Email1	928	16624
Email2	1133	5452

3.2 Processing original data

The raw data need to be processed to meet the demand of matrix operation. Firstly, the original data object can be abstracted as a node in the network, the weight of edges between two objects (*i, j*) can be represented by 0 and 1, respectively. Wherein, 1 denotes that there exists a link between nodes and 0 otherwise. We here use the Dolphins dataset as an example and demonstrate the 0-1 relations in Fig.1.

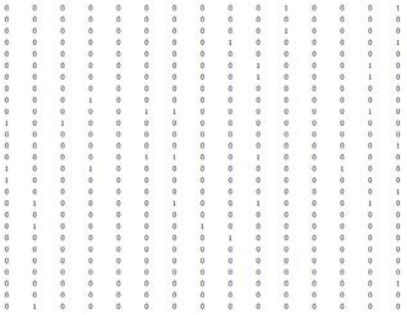


Figure 1. Dolphins adjacency matrix.

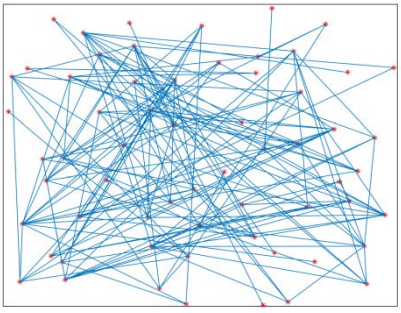


Figure 2. Dolphins network topology.

To show the relationship among nodes more intuitively, we also plot the network topology of the Dolphins dataset as shown in Fig. 2, where we connect two nodes with a straight-line if the weight of their edge is 1. We can see that that the network topology is very complex and it is difficult to partition communities precisely. Furthermore, if the number of nodes increases, the adjacency matrix will be huger and the network topology will become denser, which results in a long time to detect the community structures, a lightweight community detection algorithm is hence needed. We next discuss how to reduce the dimension of datasets with PCA technology.

3.3 Principle Component Analysis

The principle component analysis technology is generally used to extract the main features of datasets and neglect the relatively minor parts. The main purpose of PCA is to reduce dimension of the data matrix and extract the main information. We here introduce how to use PCA to detect the overlapping community structure by the following steps.

(1) Construct the adjacency matrix $W_{n\times n}$, where the value of x_{ij} equals to 1 if two nodes have an edge and 0 otherwise.

$$W=\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nn} \end{bmatrix} \quad (1)$$

(2) Get the standardized matrix by data standardization process.

$$y_{ij}=\frac{x_{ij}-\overline{x_j}}{S_j} \quad (i,j=1,2,.....n) \quad (2)$$

Where,

$$\overline{x_j}=\frac{\sum_{i=1}^{n}x_{ij}}{n},$$

$$s_j^2=\frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)^2}{n-1}(j=1,2,3,\cdots n)$$

(3) Compute the covariance matrix C_{ij}.

$$c_{ij}=\begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} \quad (3)$$

Where,

$$c_{ij}=\mathrm{cov}(y_i,y_j)=\frac{\sum_{k=1}^{n}(y_{ki}-\mu_{y_i})(y_{kj}-\mu_{y_j})}{n-1},$$

$$\mu_{y_i}=\frac{\sum_{k=1}^{n}y_{k_i}}{n},\mu_{y_j}=\frac{\sum_{k=1}^{n}y_{k_j}}{n}$$

(4) Determine the characteristic value matrix (*A*) and the eigenvector matrix (*P*) according to the covariance matrix *C*. We can notice that the elements are ordered with a decreasing trend along the diagonal line in the eigenvalue matrix.

$$\lambda_1\geq\lambda_2\geq\cdots\lambda_n\geq 0 \quad (4)$$

Where,
$$A=\begin{bmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\ 0 & \lambda_2 & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & \lambda_n\end{bmatrix},$$

$$P=\begin{bmatrix}p_{11} & p_{12} & p_{13} & \cdots & p_{1n}\\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn}\end{bmatrix}$$

(5) The characteristic value reflects the contribution of element, and the individual proportion of the principal components is shown as following.

$$\lambda_i \Big/ \sum_{k=1}^{n}\lambda_k \ (i=1,2,\cdots\cdots n) \qquad (5)$$

(6) Extract the main features. We select the top $m(1\le m\le n)$ columns whose cumulative eigenvalue is greater than 0.85, and thus, the dimension can be reduced.

$$\sum_{k=1}^{m}\lambda_k \Big/ \sum_{k=1}^{n}\lambda_k \ge 0.85 \quad (m=1,2.....,n) \qquad (6)$$

(7) According to the value of *m*, we then extract the first *m* columns from the eigenvector matrix *P*, to constitute a new eigenvector matrix *R*, where *m* represents the number of network communities and the elements of each row indicate the degree of relations between nodes and the corresponding community.

(8) Classify nodes. In this paper, we use the Membership Index (MI) [12] to further demonstrate the degree of membership between node *i* and community *j*.

$$MI_{ij}=\frac{R_{ij}-\min R_{ij}}{\max R_{ij}-\min R_{ij}} \qquad (7)$$

In this formula, max R_{ij} and min R_{ij} denote the maximum and minimum of each row in matrix *R*. Node *i* belongs to the community *j* if $MI_{ij}\ge\sigma$. Using this method, node *i* can join multiple communities if the corresponding membership values are bigger than the threshold. Obviously, different threshold values have different community structures, how to select an appropriate threshold is a rather difficult issue. We here use the module *Q* to ascertain the value of σ, i.e., we select the threshold σ which results the maximum module *Q* value. The Module-Q is a common metric to evaluate the stability of communities.

$$Q=\frac{1}{2l}\sum_{ij}\left[W_{ij}-\frac{k_i k_j}{2l}\right]\delta\left(c_i,c_j\right) \qquad (8)$$

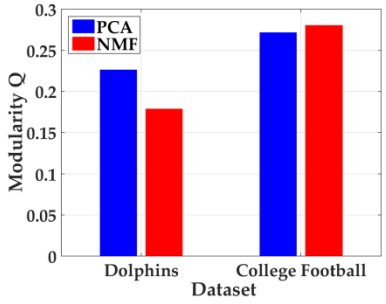
Where, C_i represents nodes, when c_i and c_j belong to the same community, the value of function $\delta\left(c_i,c_j\right)$ is 1, otherwise 0. In an undirected graph, the degree of node means the number of edges relevant to the node and the probability of an edge existing between nodes *i* and *j* is $k_i k_j/2l$. The parameter $l=\frac{1}{2}\sum_{ij}W_{ij}$ is used to indicate the total number of edges in the undirected graph.

4. EXPERIMENT AND ANALYSIS

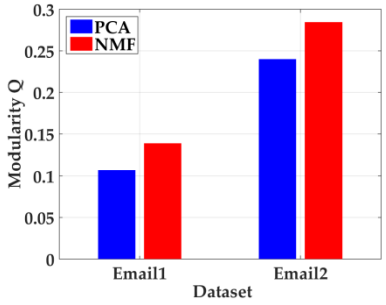
In this paper, we compare the PCA algorithm with classic non-negative matrix factorization algorithm. The evaluation metrics include the module *Q* value and time complexity. We use the term "PCA" to denote our proposed algorithm and the term "NMF" to denote the non-negative matrix factorization algorithm [10] in the following figures.

4.1 Module-Q

We plot the value of Module-Q in Fig. 3. It can be found that the two algorithms have a close performance in this metric. PCA has a high Module-Q at the Dolphins whereas NMF achieves a better performance at the Email datasets. Both of them reach a balance at the College Football dataset.



(a)



(b)

Figure 3. Modularity Q.

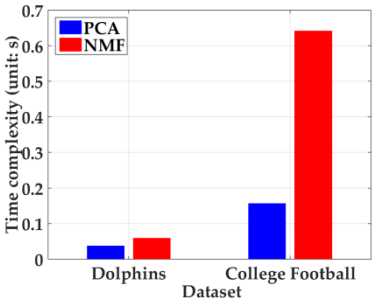
4.2 Time complexity

We first analyze the time complexity of PCA: The key step of PCA is computing eigenvectors and eigenvalues. In MATLAB, the time complexity of the spectral decomposition function is close to $O(n^3)$, this is the time-consuming step in the process of matrix dimensionality reduction.

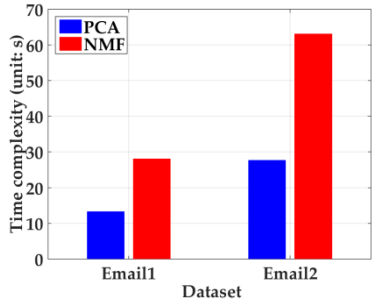
We now analyze the time complexity of NMF. NMF algorithm conducts a very centralized matrix multiplication during the

iterative process, resulting in a high time complexity. The core part of NMF algorithm is the matrix product of $A_{n\times k}$ and $B_{k\times n}$. It requires $k*n*k$ times multiplications and let r denote the number of iterations and we get the upper limit of the time complexity is $O\left(n^3 r\right)$.

We use the timer function TIC, TOC in the MATLAB to output the running time of the two algorithms. We observe that, PCA algorithm has the obvious advantages over the classical algorithm in the aspect of time complexity. In order to further analyze the differences between the two algorithms, we plot the running time in Fig. 4.



(a)



(b)

Figure 4. Time complexity.

The running time of PCA algorithm is less than non-negative matrix factorization algorithm. Simultaneously, when the dataset has fewer nodes, the difference of two algorithms is negligible. This is mainly because the processing time of datasets of Dolphins and College Football is less than 0.1 second. But with the increasing number of nodes, the running time also increases quickly and the difference is obvious. For example, at the Email2 dataset, the PCA algorithm has a running time 27 seconds and that of NMF is over 60 seconds. This phenomenon demonstrates that the PCA algorithm has a better scalability and help to fast detect the overlapping community structures in large social networks.

5. CONCLUSIONS

In this paper, we propose a novel method PCA-MI to detect the overlapping community structures. Experimental results demonstrate that our method can improve the accuracy of overlapping community structures detection and reduce the time complexity compared to the traditional algorithms. In the future, we will focus on the impact of the overlapping community structure on the message diffusion process.

6. REFERENCES

[1] Meo, P. D., Ferrara, E., Fiumara, G., Provetti, A. 2014. On Facebook, most ties are weak. *COMM. ACM*. 57, 11, 78-84. DOI= http://dl.acm.org/citation.cfm?id=2629438.

[2] Newman, M. E. J. 2004. Analysis of weighted networks. *Phys. Rev. E*. 70, 5 (November. 2004), 056131. DOI= http://journals.aps.org/pre/abstract/10.1103/PhysRevE.70.056131.

[3] Zhao, J., Wu, J., Xu, K. 2010. Weak ties: subtle role of information diffusion in online social networks. *Phys. Rev. E*. 82 (July. 2010), 87-94. DOI= http://journals.aps.org/pre/abstract/10.1103/PhysRevE.82.016105.

[4] Shi, X., Adamic, L. A., Strauss, M.J. 2007. Networks of strong ties. *Physica A Statistical Mechanics*. 378, 1 (May. 2007), 33-47. DOI= http://xueshu.baidu.com/s?wd=paperuri%3A%28df07230950e6a1bf62f34e8b200a0f41%29&filter=sc_long_sign&sc_ks_para=q%3DNetworks%20of%20strong%20ties&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8.

[5] Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Barabási, A. L. 2007. Structure and tie strengths in mobile communication networks. *PNAS*. 104, 18, 7332-7336. DOI= http://www.pnas.org/content/104/18/7332.short.

[6] Qin, Y., Tian X., Wu, W., Wang, X. 2015. Mobility weakens the distinction between multicast and unicast. *IEEE. ACM. TN* 2015, 99, 1-14. DOI= http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7084699&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D7084699.

[7] Reid, F. and Hurley, N. 2011. Diffusion in networks with overlapping community structure. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops* (Vancouver, Canada, December 11-11, 2011). ICDMW '11. 969-978. DOI= http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6137486&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6137486.

[8] Palla, G., Derényi, I., Farkas, I., Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*. 435, (June. 2005), 814-818. DOI= http://www.nature.com/nature/journal/v435/n7043/abs/nature03607.html.

[9] Zhang, S., Wang, R. S., Zhang, X. S. 2007. Uncovering fuzzy community structure in complex networks. *Phys. Rev. E*. 76, (October. 2007), 046103. DOI= http://journals.aps.org/pre/abstract/10.1103/PhysRevE.76.046103.

[10] Psorakis, I., Roberts, S., Ebden, M., Sheldon, B. 2011. Overlapping community detection using bayesian non-negative matrix factorization. *Phys. Rev. E*. 83, (June. 2011), 066114. DOI= http://journals.aps.org/pre/abstract/10.1103/PhysRevE.83.066114.

[11] Yuan, P. and Tang, S. 2015. Community-based immunization in opportunistic social networks. Physica A Statistical Mechanics. 420, (February. 2015), 85-97. DOI= http://www.sciencedirect.com/science/article/pii/S037843711400942X.

[12] Li, Y., Li, B., GUO, Z. 2014. Research on overlapping community detection in networks using non-negative matrix factorization. *Journal of System Simulation.* 26, (June 2011), 643-649. DOI= http://journals.aps.org/pre/abstract/10.1103/PhysRevE.83.066114.

[13] Xie, J., Kelley, S., Szymanski, B. K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM. C. SURV*. 2013, 45,4 (August 2013), 115-123. DOI= http://dl.acm.org/citation.cfm?id=2501657.

[14] Girvan, M. and Newman, M.E.J. 2002. Community structure in social and biological networks. *PNAS*. 99, 12 (April. 2002), 7821-7826. DOI= http://www.pnas.org/content/99/12/7821.short.

[15] Zhang, Y. and Yeung, D.Y. 2012. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (Beijing, China, August 12-16, 2012). KDD '12. ACM, New York, NY, 606-614. DOI= http://dl.acm.org/citation.cfm?id=2339629.

[16] Yang, J.; Leskovec, J. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the 6th ACM international conference on Web search and data mining* (Rome, Italy, February 04 – 08, 2013). WSDM '13. ACM, New York, NY, 587-596. DOI= http://dl.acm.org/citation.cfm?id=2433471.

[17] Yang, J. and Leskovec, J. 2012. Community-Affiliation graph model for overlapping network community detection. In *Proceedings of the 12th International Conference on Data Mining*. IEEE '12, Brussels, Belgium, December 10-13, 1170-1175. DOI= http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6413734&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6413734.

[18] Shujutang. Available online: http://www.datatang.com/.