The (un)supervised Detection of Overlapping Communities as Well as Hubs and Outliers via (Bayesian) NMF

Xiaochun Cao
Sch. of Comp. Sci. & Tech.,
Tianjin Univ.
State Key Lab of Information Security, Institute of Information Engineering, CAS
caoxiaochun@iie.ac.cn

Xiao Wang, Di Jin
Sch. of Comp. Sci. & Tech.,
Tianjin Univ.
{wangxiao_cv, jindi}@tju.edu.cn

Yixin Cao
Inst. for Com. Sci. & Cont,
Hungarian Academy of Sciences
yixin@sztaki.hu

Dongxiao He
Colg. of Comp. Sci. & Tech,
Jilin Univ.
hedongxiaojlu@gmail.com

ABSTRACT

The detection of communities in various networks has been considered by many researchers. Moreover, it is preferable for a community detection method to detect hubs and outliers as well. This becomes even more interesting and challenging when taking the unsupervised assumption, that is, we do not assume the prior knowledge of the number K of communities. In this poster, we define a novel model to identify overlapping communities as well as hubs and outliers. When K is given, we propose a normalized symmetric nonnegative matrix factorization algorithm to learn the parameters of the model. Otherwise, we introduce a Bayesian symmetric nonnegative matrix factorization to learn the parameters of the model, while determining K.Our experiment indicates its superior performance on various networks.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Application-Data Mining

Keywords: community, hubs, outliers, (Bayesian) NMF.

1. INTRODUCTION

The community structure is arguably the most fundamental property of most real-world networks. Considering the large scale of these networks, almost all studies on them inevitably starts from the detection of their community structures. It is now widely agreed that the communities usually overlap with each other and some members of the network do not belong to any communities, recognized as *outliers*. Also, some member of a community might be special in the sense that it is linked with almost all others, known as a *hub*. Needless to say, the community structure will greatly benefited from the simultaneous detection of hubs and outliers. This becomes more interesting and also more challenging when we do not assume the prior knowledge of the number K of communities. Pre-determining the number of communities artificially by guess, even by domain expert, will not always be plausible. Thus, we may have to learn the number of communities from the data.

Herein we propose a novel generative model consisting of a centrality matrix of $\mathbf{U}$ vertices and a degree matrix $\mathbf{H}$ of communities. When K is known, we apply a normalized symmetric nonnegative matrix factorization (NSNMF) algorithm based on KL divergence to learn the two sets of matrices under the condition. When K is unknown, we apply a Bayesian symmetric nonnegative matrix factorization (BSNMF) based method to learn the two matrices and determine the number of communities automatically. Once learned, matrices $\mathbf{U}$ and $\mathbf{H}$ enable us to identify overlapping communities, hubs, and outliers altogether.

2. METHODS

In this poster all networks are assumed to be undirected and unweighted. Our model contains two sets of parameters, w_z and u_{iz}. The *soft degree* w_z of community z is defined to be the sum of expected degrees of all vertices in community z. The *centrality* u_{iz} of vertex i in community z is defined to be the expected proportion of the degree of i in z. For each community z, it holds that $\sum_{i=1}^{N} u_{iz} = 1$. The expected number of edges that lie between vertices i and j can be written as $\hat{A}_{ij} = \sum_{z=1}^{K} u_{iz} w_z u_{jz}$. In matrix terminology, we have the *expected adjacency matrix* $\hat{\mathbf{A}} = \mathbf{UHU}^T$, where $\mathbf{U}$ is the $N \times K$ centrality matrix of vertices, and $\mathbf{H} = diag(\mathbf{w}^T)$ is a diagonal matrix obtained from $\mathbf{w}^T = (w_1, w_2, ..., w_z, ..., w_K)$.

2.1 Normalized symmetric NMF method

When K is known, we apply NSNMF method to learn the two matrices. Using KL divergence to measure the relaxation error, it is defined as

$$\min_{\mathbf{U},\mathbf{H}\geq 0} D(\mathbf{A}\|\hat{\mathbf{A}}) = \sum_{ij}(A_{ij}\log\frac{A_{ij}}{\hat{A}_{ij}} - A_{ij} + \hat{A}_{ij}). \quad (1)$$

WWW'14 Companion, April 7-11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577307.

Recall the classic NMF problem: $\min_{\mathbf{W},\ \mathbf{B}\geq 0} D(\mathbf{A}\|\mathbf{WB})$. We can use the traditional multiplicative updating rules to get $\mathbf{W}$ and $\mathbf{B}$. The columns of $\mathbf{W}$ and rows of $\mathbf{B}$ are divided by a normalized factor, which gives a normalized $N \times K$ matrix $\hat{\mathbf{W}}$ and a normalized $K \times N$ matrix $\hat{\mathbf{B}}$. The normalized factor of each column of $\mathbf{W}$ and each row of $\mathbf{B}$ can form the normalized factor vector $\mathbf{NW}$ and $\mathbf{NB}^T$, respectively. This transforms $D(\mathbf{A}\|\mathbf{WB})$ to $D(\mathbf{A}\|\hat{\mathbf{W}}\hat{\mathbf{H}}\hat{\mathbf{B}})$, where $\hat{\mathbf{H}} = diag(\mathbf{NW})diag(\mathbf{NB}^T)$. Further, since the adjacency matrix $\mathbf{A}$ is symmetric, the corresponding elements of $\mathbf{W}$ and $\mathbf{B}^T$ have a proportional relation. As a result, $\hat{\mathbf{W}} = \hat{\mathbf{B}}^T$. Since $\sum_{i=1}^{N} u_{iz} = 1$, we have $\min_{\hat{\mathbf{W}},\hat{\mathbf{B}}\geq 0} D(\mathbf{A}\|\hat{\mathbf{W}}\hat{\mathbf{H}}\hat{\mathbf{B}}) = \min_{\hat{\mathbf{W}},\hat{\mathbf{H}}\geq 0} D(\mathbf{A}\|\hat{\mathbf{W}}\hat{\mathbf{H}}\hat{\mathbf{W}}^T)$. The parameters are given by $\mathbf{U} = \hat{\mathbf{W}}$ and $\mathbf{H} = \hat{\mathbf{H}}$.

2.2 Bayesian symmetric NMF method

In this case that K is unknown, we turn to a Bayesian symmetric nonnegative matrix factorization (BSNMF) method, which determines K when learning the parameters.

The expected adjacency matrix $\hat{\mathbf{A}}$ can be rewritten as $\hat{\mathbf{A}} = (\mathbf{U}\mathbf{H}^{\frac{1}{2}})(\mathbf{U}\mathbf{H}^{\frac{1}{2}})^T = \mathbf{W}\mathbf{W}^T$. Thus, we need to figure out $\mathbf{W}$ and $\mathbf{B}$ such that $\hat{\mathbf{A}} = \mathbf{WB}$ and $\mathbf{B} = \mathbf{W}^T$. In order to solve the problem of model selection, we introduced the priors $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$ on both the columns of $\mathbf{W}$ and the rows of $\mathbf{B}$ [4]. These priors are the qualities that control the irrelevant columns of $\mathbf{W}$ and rows of $\mathbf{B}$ that do not contribute to the construction of $\mathbf{A}$. Given the adjacency matrix $\mathbf{A}$, we can obtain the posterior as:

$$\begin{aligned} -\log p(\mathbf{W},\mathbf{B},\boldsymbol{\beta}|\mathbf{A}) \propto & -\log p(\mathbf{A}|\mathbf{W},\mathbf{B}) - \log p(\mathbf{W}|\boldsymbol{\beta}) \\ & - \log p(\mathbf{B}|\boldsymbol{\beta}) - \log p(\boldsymbol{\beta}). \end{aligned} \quad (2)$$

Here we adopt the multiplicative update rules in [4], which are based on the fast fixed-pointed algorithm for $\mathbf{W}$ and $\mathbf{B}$. We remark that some columns of $\mathbf{W}$ and rows of $\mathbf{B}$ are possibly zero vectors, which indicates these columns and rows do not contribute to the construction of $\mathbf{A}$. As a result, the number of the non-zero columns of $\mathbf{W}$, i.e., non-zero rows of $\mathbf{B}$, gives the number of meaningful communities, i.e., K. Finally, we can obtain the degree matrix $\mathbf{H} = (\mathbf{1}_N^T\mathbf{W})^2$, and the centrality matrix $\mathbf{U} = \mathbf{W}(\mathbf{H}^{\frac{1}{2}})^{-1}$.

Table 1: FVCC comparison between NSNMF and other methods on real-networks with given K

FVCC (%)	Louvain[1]	CPM[3]	BNMF[4]	NSNMF
Karate	97.06	75.00	82.35	**100.00**
Dolphins	98.39	**100.00**	83.23	96.15
Friendship6	92.75	82.35	86.39	**94.23**
Friendship7	91.30	82.35	85.22	**94.34**
Polbooks	84.76	**88.51**	81.52	*87.10*
Word	58.93	62.16	55.36	**93.85**
Football	93.04	29.20	86.09	**94.55**
Polblogs	96.17	–	93.15	**98.22**

2.3 The detection of communities

We order each column of $\mathbf{U}$ in nonincreasing order. Let $\hat{\mathbf{U}}_z$ denote the ordered column vector of zth column of $\mathbf{U}$, then we can get $\mathbf{S}_z$ as the corresponding column vector for vertex indices of $\hat{\mathbf{U}}_z$. The degree matrix $\mathbf{H}$ provides the expected degree w_z of the zth community, and this quality is a criterion to cut the rank sequence. We add the vertex in $\mathbf{S}_z$ one by one from top to bottom to this community, until the sum of degrees of these vertices is no less than w_z for the first time. So the members of the zth community C_z are the vertices in $\mathbf{S}_z$. Now we get all the communities, all outliers expose themselves immediately: they are the remaining vertices. The hub in community z is the top vertex in $\mathbf{S}_z$.

Table 2: AC comparison between BSNMF and other methods on real-networks without given K

AC	Louvain[1]	CPM[3]	BNMF[4]	BSNMF
Lesmis	0.3343	0.3612	0.3736	**0.2764**
jazz	**0.3344**	0.6140	0.5578	*0.5331*
neural	**0.4816**	0.7486	0.7430	*0.4864*
metabolic	0.5244	0.6248	0.6336	**0.3717**
email	0.4298	0.5066	0.5429	**0.3263**
netscience	0.1035	0.2272	0.0416	**0.0040**

3. EXPERIMENTS

We make some quantitative comparisons with three community detection algorithms. The first one, called Louvain [1], is regarded as one of the best for vertex partition. The second one, called CPM [3], is the most prominent algorithm for overlapping community detection. The third one is called BNMF [4]. We evaluate NSNMF on eight real-world networks[1] that the number K of communities is known, and test BSNMF on six networks[1] that we do not know K. FVCC is an accuracy metric for networks with known communities and K, measuring the fraction of vertices classified correctly, which is suitable to test NSNMF, and the average conductance (AC) metric of communities [2] can be used to evaluate BSNMF on networks that we do not know the communities and K. The smaller the AC, the better the result.

The comparison results of NSNMF and BSNMF are summarized in Table 1 and Table 2, respectively. Our best results are marked by **boldface** and our second best results are marked by *italic*. To sum up, both of NSNMF and BSNMF dramatically outperforms the other methods in gerenal.

4. ACKNOWLEDGMENTS

Supported by NSFC (61332012, 61303110), R&D Program (2012 BAH07B01), 973 Program (2013CB329305), and 100 Talents Programme of CAS, RFDP (20130032120043), and Open Project Program of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (93K172013K02). Correspondence should be addressed to D.J. (jindi@tju.edu.cn)

5. REFERENCES

[1] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J Stat Mech*, 2008(10):P10008, 2008.

[2] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of WWW'10*, pages 631–640. ACM, 2010.

[3] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[4] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping community detection using Bayesian non-negative matrix factorization. *Phy Rev E*, 83(6):066114, 2011.

[1] http://www-personal.umich.edu/ mejn/netdata/