Soft Partitioning in Networks via Bayesian Non-negative Matrix Factorization

Ioannis Psorakis*, **Stephen Roberts**
Pattern Analysis and Machine Learning Research Group
Department of Engineering Science
University of Oxford
Parks Road, OX1 3PJ Oxford, United Kingdom
{yannis,sjrob}@robots.ox.ac.uk

Ben Sheldon
Edward Grey Institute
Department of Zoology
University of Oxford
South Parks Road, OX1 3PS Oxford, United Kingdom
ben.sheldon@zoo.ox.ac.uk

Abstract

Identifying overlapping communities in networks is a challenging task. In this work we present a novel approach to community detection that utilizes the Bayesian non-negative matrix factorization (NMF) model to extract overlapping modules from a network. The scheme has the advantage of computational efficiency, soft community membership and an intuitive foundation. We present the performance of the method against a variety of benchmark problems and compare and contrast it to several other algorithms for community detection.

1 Introduction

Community structure, or modular organization, is a significant property of real-world networks as it is often considered to account for the functional characteristics of the system under study [16]. Although the notion of "community" appears intuitive [19, 22] (for example people form cliques in social networks and web pages of similar content have links to one another) there is no disciplined, context-independent definition of what communities are [22]; we adopt here the loose definition that these modules are subgraphs with more links connecting the nodes inside than outside them [19,22]. The task of identifying such subgraphs in a given network can be challenging [16], both in terms of recognition and computational feasibility.

In this work we propose a novel approach to community detection based on computationally efficient Bayesian Non-negative Matrix Factorization (NMF) [23]. The advantages of this methodology are i) soft partitioning solutions; each node is associated with a membership distribution over communities, describing its degree of participation to each module ii) excellent computational performance and scalability iii) the method does not suffer from the drawbacks of modularity optimization methods, such as the resolution limit iv) excellent module identification capabilities.

In the following section we present the theoretical foundations of our approach along with an illustrative example to provide intuition behind the method. Following the model formulation section,

*corresponding author

we test our algorithm on a variety of artificial and real-world benchmark problems and present our experimental results.

2 Theoretical Background

Consider the generative model of Fig. 1. The observed variable v_{ij} denotes the non-negative count of interactions that took place between two individuals i, j in a weighted undirected network with adjacency matrix $\mathbf{V} \in \mathbb{R}_+^{N \times N}$.

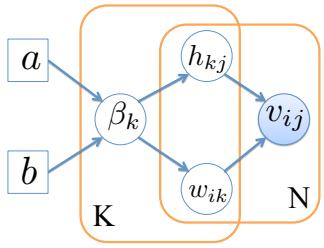


Figure 1: Graphical model showing the generation of count processes $\mathbf{V}$ from the latent structure $\mathbf{W}$ and $\mathbf{H}$, the components of which have scale hyperparameters β_k. The hyper-hyperparameters a, b are fixed.

We assume that the pair-wise interactions described in $\mathbf{V}$ are influenced by two unobserved non-negative matrices $\mathbf{W} \in \mathbb{R}_+^{N \times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ so that $\hat{\mathbf{V}} = \mathbf{WH}$. We hence model each interaction v_{ij} as drawn from a Poisson distribution with rate $\hat{v}_{ij} = \sum_{k=1}^{K} w_{ik} h_{kj}$. The inner rank K denotes the unknown number of communities and each element $k \in \{1, ..., K\}$ in row i of $\mathbf{W}$ and column j of $\mathbf{H}$ represents the contribution of a single latent community to $\hat{v}_{ij}$. In other words, the expected number of times $\hat{v}_{ij}$ two individuals i, j interact is a result of their *mutual participation* into the same communities.

In the typical community detection setting, the value of K is initially unknown. We address this issue by placing *automatic relevance determination* [13] priors β_k on the latent variables w_{ik}, h_{kj}, as presented in [23]. The effect of these priors is to "switch off" irrelevant columns of $\mathbf{W}$ and rows of $\mathbf{H}$ that do not contribute to explaining the observed interactions $\mathbf{V}$.

Based on the graphical model of Fig. 1, we express the joint distribution over all variables as:

$$p(\mathbf{V}, \mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = p(\mathbf{V}|\mathbf{W}, \mathbf{H}) p(\mathbf{W}|\boldsymbol{\beta}) p(\mathbf{H}|\boldsymbol{\beta}) p(\boldsymbol{\beta}) \tag{1}$$

and hence the model posterior over model parameters given the observations is:

$$p(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}|\mathbf{V}) = \frac{p(\mathbf{V}|\mathbf{W}, \mathbf{H}) p(\mathbf{W}|\boldsymbol{\beta}) p(\mathbf{H}|\boldsymbol{\beta}) p(\boldsymbol{\beta})}{p(\mathbf{V})} \tag{2}$$

Noting that $p(\mathbf{V})$ is a constant w.r.t. the inference over the model's free parameters, we aim to maximize the model posterior given the observations. This is equivalent to minimizing the negative log posterior, which we may regard as an energy (or error) function $\mathcal{U}$. We hence define

$$\mathcal{U} = -\log p(\mathbf{V}|\mathbf{W}, \mathbf{H}) - \log p(\mathbf{W}|\boldsymbol{\beta}) - \log p(\mathbf{H}|\boldsymbol{\beta}) - \log p(\boldsymbol{\beta}) \tag{3}$$

where the first term is the log-likelihood of our data, derived from the probability $p(\mathbf{V}|\mathbf{W}, \mathbf{H}) = p(\mathbf{V}|\hat{\mathbf{V}})$ of observing every interaction v_{ij} given a Poisson rate $\hat{v}_{ij}$. The negative log likelihood is hence:

$$
\begin{aligned}
-\log p(\mathbf{V}|\hat{\mathbf{V}}) &= -\sum_{i=1}^{N}\sum_{j=1}^{N}\log p(v_{ij}|\hat{v}_{ij}) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(v_{ij}\log\frac{v_{ij}}{\hat{v}_{ij}} + \hat{v}_{ij} - v_{ij} + \frac{1}{2}\log(2\pi v_{ij})\right) + \text{const.} \qquad (4)
\end{aligned}
$$

Following [23] and similar models for probabilistic PCA [24] and ICA [4, 20, 21] we place independent half-normal priors over the columns of $\mathbf{W}$ and rows of $\mathbf{H}$ with zero mean and precision (inverse variance) parameters $\boldsymbol{\beta} \in \mathbb{R}^K = [\beta_1, ..., \beta_K]$. The negative log priors are then given by:

$$
-\log p(\mathbf{W}|\boldsymbol{\beta}) = -\sum_{i=1}^{N}\sum_{k=1}^{K}\log\mathcal{HN}(0,\beta_k^{-1}) = \sum_{i=1}^{N}\sum_{k=1}^{K}\left(\frac{1}{2}\beta_k w_{ik}^2\right) - \frac{N}{2}\log\beta_k + \text{const.} \qquad (5)
$$

$$
-\log p(\mathbf{H}|\boldsymbol{\beta}) = -\sum_{k=1}^{K}\sum_{j=1}^{N}\log\mathcal{HN}(0,\beta_k^{-1}) = \sum_{k=1}^{K}\sum_{j=1}^{N}\left(\frac{1}{2}\beta_k h_{kj}^2\right) - \frac{N}{2}\log\beta_k + \text{const.} \qquad (6)
$$

Each β_k controls the importance of community k in explaining the observed interactions; large values of β_k denote that the elements of column k of $\mathbf{W}$ and row k of $\mathbf{H}$ lie close to zero and therefore represent irrelevant communities. By assuming the β_k are independent [1] we place a standard Gamma distribution over them [1]. The negative log hyper-priors are thus:

$$
-\log p(\boldsymbol{\beta}) = -\sum_{k=1}^{K}\log\mathcal{G}(\beta_k|a,b) = \sum_{k=1}^{K}(\beta_k b - (a-1)\log\beta_k) + \text{const.} \qquad (7)
$$

The objective function $\mathcal{U}$ of Eq. (3) can be expressed as the sum of Eq. (4), (5), (6) and (7). To optimize for $\mathbf{W}, \mathbf{V}$ and $\boldsymbol{\beta}$ we follow [2, 11, 12, 23] by adopting the fast fixed-point algorithm presented in [23] with algorithmic complexity $\mathcal{O}(NK)$, which involves consecutive updates of $\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}$ until a convergence measure has been satisfied[2]. The pseudocode is presented in Algorithm 1.

Algorithm 1 Community Detection using NMF

Require: adjacency matrix $\mathbf{V} \in \mathbb{R}_+^{N\times N}$, initial K_0, fixed Gamma hyperparameters a, b.

1: **for** $i = 1$ to n_{iter} **do**

2: $\quad \mathbf{H} \leftarrow \left(\frac{\mathbf{H}}{\mathbf{W}^\mathsf{T}\mathbf{1}+\mathbf{BH}}\right)\mathbf{W}^\mathsf{T}\left(\frac{\mathbf{V}}{\mathbf{WH}}\right)$

3: $\quad \mathbf{W} \leftarrow \left(\frac{\mathbf{W}}{\mathbf{1}\mathbf{H}^\mathsf{T}+\mathbf{WB}}\right)\left(\frac{\mathbf{V}}{\mathbf{WH}}\right)\mathbf{H}^\mathsf{T}$

4: $\quad \beta_k \leftarrow \frac{N+a-1}{\frac{1}{2}\left(\sum_i w_{ik}^2+\sum_j h_{kj}^2\right)+b}$

5: **end for**

6: $K_\star \leftarrow$ # of non-zero columns of $\mathbf{W}$ or rows of $\mathbf{H}$

7: **return** $\mathbf{W}_\star \in \mathbb{R}_+^{N\times K_*}, \mathbf{H}_\star \in \mathbb{R}_+^{K_*\times N}$

In the case of undirected graphs, $\mathbf{W}_\star = \mathbf{H}_\star^\mathsf{T}$ (as $\mathbf{V}$ is symmetric) and represents the $N \times K_\star$ incidence matrix of a bipartite graph of N nodes and $K_\star$ communities. Each element $w_{ik}^\star$ (or $h_{ki}^\star$) denotes the *degree of participation* of individual i into community k while each normalized row of $\mathbf{W}_\star$ (or column of $\mathbf{H}_\star$) expresses a *soft-membership* distribution over communities given a certain node. Therefore this bipartite graph describes the *overlapping* mesoscopic structure of our network, where nodes are allocated to multiple groups with varying participation score.

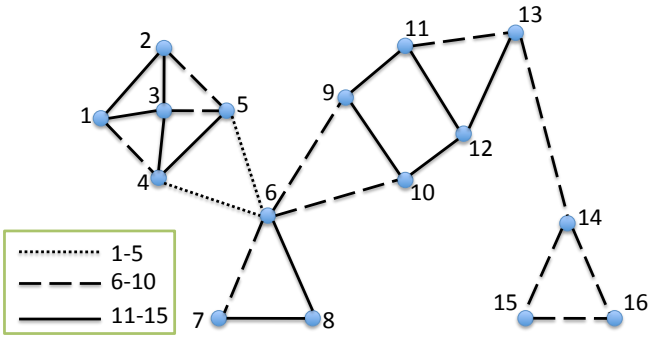
[1] This corresponds to the belief that the existence of one community is not dependent upon others. Clearly, there will be situations in which this can be extended to allow for a full inter-dependency between communities. We do not consider this here, however. Allowing dependency is similar to the notion of *structure priors* discussed in [18].

[2] A maximum number of iterations, or a tolerance on the cost function.

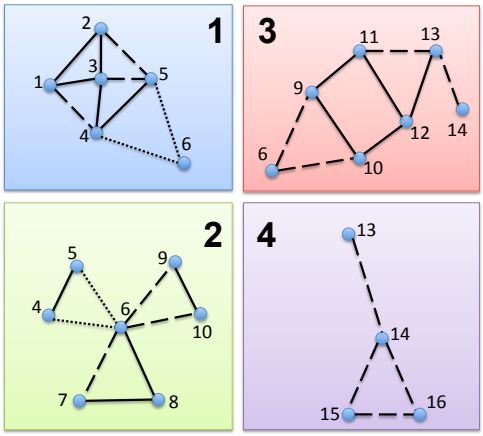
In the next section, we present an illustrative example of this community extraction scheme, followed by experimental results from various artificial and real-world networks.

3 Applications

We start by presenting an illustration of our community detection scheme and form of results, using the simple toy graph of Fig. 2(a) that has $N = 16$ nodes and $M = 25$ links of various weights. We applied NMF based community detection and extracted $K_\star = 4$ overlapping groups as shown in Fig. 2(b). We can see that these communities share the boundary nodes, which lie on high flow paths in the network and are usually difficult to classify in one module or another [22] by traditional hard-partitioning methods.



(a) A simple undirected network of $N = 16$ nodes and $M = 25$ links.

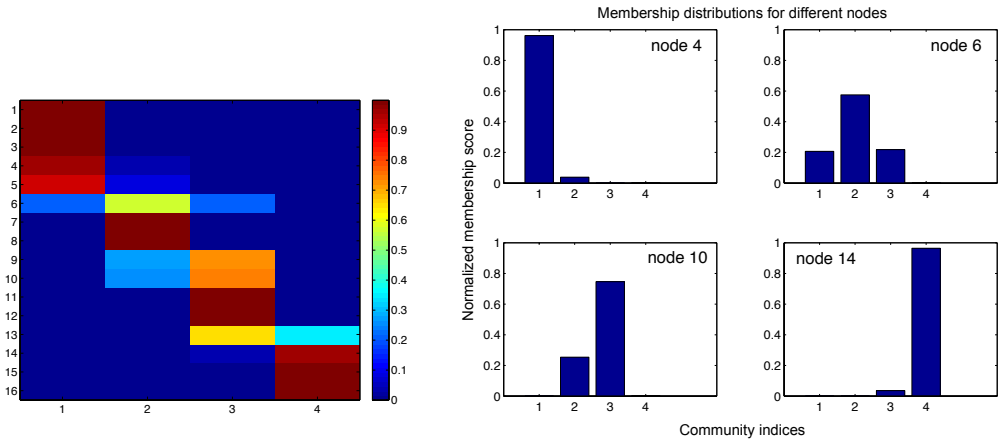


(b) Decomposed into $K_\star = 4$ overlapping communities.

Figure 2: We use the weighted undirected network of 2(a) to illustrate our community detection scheme. Our algorithm extracted four overlapping modules, as shown in 2(b).

Allocating nodes to multiple modules as in Fig. 2(b), is only one part of the solution. We also capture the *degree of participation* of individuals to each community by using the incidence matrix $\mathbf{W}_\star$ described in the previous section. Fig. 3(a) shows $W_\star \in \mathbb{R}_+^{16\times 4}$ where different colors indicate various levels of participation of nodes to communities. We can see that the matrix is not in a clear block diagonal form, as an individual can have some form of membership into multiple groups.

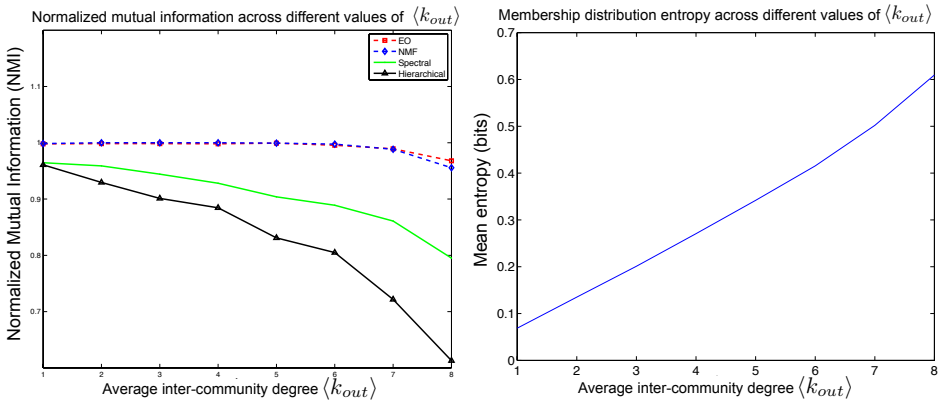
In our framework, community allocation is not a Boolean decision but a *belief*; each node is assigned a membership distributed over communities, as seen in Fig. 3(b). We can see that mediator nodes of high "betweenness", such as $i = 6$, have a more entropic distribution (similar to the concept of "bridgeness" [14]) while for nodes such as $i = 4$ or $i = 14$ we have much more confident allocations.



(a) Color map of the incidence matrix $W_\star \in \mathbb{R}_+^{16\times 4}$ (b) Probability of membership distributions for various nodes in our toy network.

Figure 3: Fig. 3(a) shows the node allocations proposed by our algorithm. Colors close to red indicate strong participation of node i (vertical axis) to community k (horizonal axis). Fig. 3(b) shows example (normalized) rows of $\mathbf{W}_\star$ that correspond to the membership distribution of different nodes.

Having such soft-membership distributions not only allows us to describe our confidence in assigning node i to community k, but also to quantify the degree of "fuzziness" in the network. In Fig. 3(b) nodes such as $i = 6$ that lie on community boundaries have a membership distribution that is closer to uniform. We hence expect our method to indicate networks with less degrees of modular organization. We apply the NMF method to realizations of the very popular Newman-Girvan (NG) random graph [8]. This benchmark tests the module identification capabilities of a method against an artificial graph of $N = 128$ nodes, observed solution of $C = 4$ communities (with $n = 32$ nodes each), average degree of $\langle k \rangle = 16$ and an ever-increasing *inter-community* degree $\langle k_{out} \rangle$ that controls the module cohesiveness of the network.

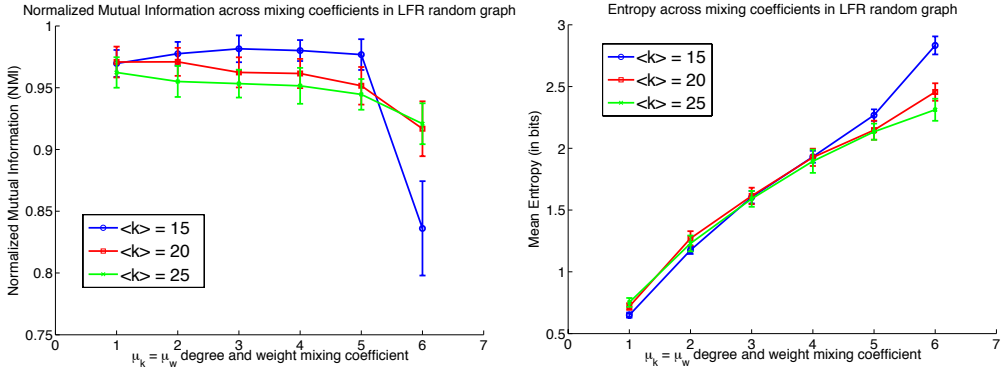


(a) Normalized Mutual Information, value range 0-1 (b) Mean entropy of probability membership distributions (bits).

Figure 4: Fig. 4(a) compares the NMF approach against Extremal Optimization (EO), Spectral Partitioning (Spectral) and Hierarchical Clustering (Hierarchical) in partitioning Newman-Girvan artificial graphs. Each point is the mean of 100 graph realizations. Fig. 4(b) shows the increase in uncertainty in assigning nodes to communities, as we increase the fuzziness of modular organization in NG graphs. Each point is the mean of 100 graph realizations.

In Fig. 4(a) we plot our module identification performance based on the Normalized Mutual Information (NMI) criterion [5] while in 4(b) we monitor our allocation confidence based on the mean entropy (in bits) of each node membership distribution. We can see that as we make the network fuzzier by increasing $\langle k_{out} \rangle$, our method "responds" by increasing the degree of node participation to multiple communities. An attractive aspect of this test is that the increase in entropy (see Fig. 4(b)) does not affect the module identification performance (we see from Fig. 4(a) that NMI remains close to unity) and is stable for the vast majority of $\langle k_{out} \rangle$ values. We also provide in Fig. 4(a) the NMI performance of some popular hard-partitioning methods (Extremal Optimization [6], Spectral Partitioning [15], Hierarchical Clustering [3] [22]) for comparison.

We extend the above test to the more complicated case of Lancichinetti-Fortunato random graphs (LFR) that have been proposed more recently [10]. In this setting, community cohesion is controlled by *mixing parameters* μ_k and μ_w, that denote the expected fraction of intercommunity degrees and weights per node. Other configuration parameters include, among others, the total number of nodes N, the average degree $\langle k \rangle$ or the exponents of the degree γ_1 and community size γ_2 distributions. We tested our method for a (decaying) range of values for μ_k, μ_w (where we set $\mu_k = \mu_w$), in weighted graphs of $N = 1000$ nodes and various values of $\langle k \rangle$, as seen in Fig. 5(a). In the same spirit with the NG graph case, in Fig. 5(b) we monitor the mean entropy of membership distributions per node (in bits) to quantify the confidence of our node allocations to communities. In Fig. 5(a) we can see that our model has an excellent module identification performance and starts to fail only when the mixing coefficients μ have values greater than 0.5, denoting no community organization in the graph. On the other hand, the increasing fuzziness of the network (based on μ) is captured in the mean entropy of the membership distributions; as the community structure is less cohesive, we are less confident into allocating each node to a group.



(a) Normalized Mutual Information, value range 0-1 (b) Mean entropy of probability membership distributions (bits).

Figure 5: Results of the NMF method on realizations of the LFR random graphs for $N = 1000$ and different values for the average degree $\langle k \rangle$ and community cohesion μ parameters. Each point represents the mean and standard deviation over 100 graph realizations.

We conclude this section by presenting the performance of our community detection method on a variety of popular benchmark datasets and compare it against the Extremal Optimization (EO) [6] and Louvain [3] methods. In contrast to the artificial graphs we used above, the absence of an "observed solution" for these problems prevents us from using the Normalized Mutual Information criterion for performance evaluation. Instead we use the extremely popular *modularity* Q [17], which is a measure of how "statistically surprising" the intra-community link density is of a proposed network partition. For the purposes of the experiment, we remove the overlapping aspect of the NMF solutions by assigning a node to a single community; the one for which is has the maximum degree of membership. Although this "greedy allocation" scheme omits the wealth of information provided by our model solutions, it is necessary in order to perform modularity comparisons against hard-partitioning methods such as EO and Louvain. For each dataset we ran all the methods 100 times and monitored the values of *modularity* Q along with the number $K_\star$ of extracted communities.

[3] using angular distance and complete-linkage clustering.

Louvain has very stable behaviour across different runs so we have omitted the standard deviation of modularity and community sizes for each dataset. For NMF initialization we used $K_0 = N$ with hyperparameters $a = 5$ and $b = 2$, giving a vague prior. We note that the results of the method are not sensitive, within reason, to changes in these values.

Table 1: Modularity results for NMF, EO and Louvain methods

Dataset	NMF	EO	Louvain
Dolphins	0.47 ± 0.03	0.51 ± 0.01	0.52
Books US Politics	$0.52 \pm \epsilon$	0.48 ± 0.01	0.50
Les Miserables	0.53 ± 0.02	0.53 ± 0.01	0.57
College Football	$0.60 \pm \epsilon$	0.58 ± 0.01	0.60
Jazz Musicians	0.43 ± 0.01	0.42 ± 0.01	0.44
C. elegans metabolic	0.36 ± 0.01	0.40 ± 0.09	0.43
Network Science	0.83 ± 0.01	0.86 ± 0.01	0.95
Facebook Caltech	0.38 ± 0.01	0.37 ± 0.01	0.37

Table 2: Community sizes for NMF, EO and Louvain methods

Dataset	NMF	EO	Louvain
Dolphins	6.67 ± 0.83	4 ± 0	5
Books US Politics	6.23 ± 0.62	4.04 ± 0.4	3
Les Miserables	9.97 ± 0.78	4.96 ± 1.72	6
College Football	8.86 ± 0.79	8 ± 0	10
Jazz Musicians	8.57 ± 8.89	4 ± 0	4
C. elegans metabolic	15.69 ± 1.14	7.96 ± 1.06	10
Network Science	342.53 ± 5.28	58.24 ± 12.36	418
Facebook Caltech	24.28 ± 1.72	6.84 ± 1.82	10

From Table 1 we can see that our approach performs competitively despite not being an algorithm designed with the aim of maximizing modularity, unlike either EO or Louvain methods. Additionally, it has the advantage of providing *soft partitioning* solutions for community membership and has *low computational overhead*. Finally, although our method favors sparse solutions, it does not suffer from the resolution limit [7] of modularity optimization methods such as EO, where smaller groups are merged together [7, 19], leading a to smaller number of communities, as seen in Table 2.

4 Conclusion

This paper describes a novel approach to community detection that adopts a Bayesian non-negative matrix factorization model to achieve soft partitioning of a network, assigning each node a soft-membership distribution over all the extracted communities. This allows us not only to capture the fuzziness of the network (via the entropy of the membership distribution) but also to improve network cartography techniques [9] by identifying central and peripheral nodes in modules. Network visualization tools can also be improved in this manner, as the degree of membership over different communities can be utilized to position an individual in a cloud of nodes. The approach is computationally efficient and offers performance comparable to state-of-the-art methods, even when forced to make greedy allocations. Indeed the computational performance advantages over large data sets allow the NMF approach to be run many times compared to a single run of less efficient approaches. Clearly this allows for the selection of the best performing run, or a small ensemble of high-modularity solutions.

Acknowledgements

The authors would like to thank Mark Ebden, Nick Jones and Mason Porter for valuable comments. Ioannis Psorakis is funded from a grant via Microsoft Research, for which we are most grateful.

References

[1] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley, 1994.

[2] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. In *Computational Statistics and Data Analysis*, pages 155–173, 2006.

[3] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] R. A. Choudrey and S. J. Roberts. Variational mixture of Bayesian independent component analyzers. *Neural Computation*, 15(1):213–252, 2003.

[5] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, 2005(09):P09008, 2005.

[6] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72(2):027104, 2005.

[7] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[8] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[9] R. Guimera and L. A. N. Amaral. Cartography of complex networks: modules and universal roles. *J. Stat. Mech.*, 2005(02):P02001, 2005.

[10] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys. Rev. E*, 80(1):016118, 2009.

[11] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorisation. *Nature*, 401:788–791, October 1999.

[12] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2000.

[13] D. J. C. Mackay. Probable networks and plausible predictions a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[14] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E*, 77(1):016107, Jan 2008.

[15] M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.

[16] M. E. J. Newman. *Networks: an Introduction*. Oxford University Press, 2010.

[17] M. E. J Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[18] W. Penny and S. J. Roberts. Bayesian multivariate autoregressive models with structured priors. *IEE Proceedings on Vision, image and signal processing*, 149(1):33–41, 2002.

[19] M. A. Porter, J. P. Onnela, and P. J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097 and 1164–1166, 2009.

[20] S. J. Roberts and R. A. Choudrey. Data decomposition using independent component analysis with prior constraints. *Pattern Recognition*, 36(8):1813 – 1825, 2003.

[21] S. J. Roberts and R. A. Choudrey. Bayesian independent component analysis with prior constraints: An application in biosignal analysis. In Joab Winkler, Mahesan Niranjan, and Neil Lawrence, editors, *Deterministic and Statistical Methods in Machine Learning*, volume 3635 of *Lecture Notes in Computer Science*, pages 159–179. Springer Berlin / Heidelberg, 2005.

[22] Fortunato S. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010.

[23] V. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *SPARS09 - Signal Processing with Adaptive Sparse Structured Representations (2009)*, pages 1–19, 2009.

[24] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622, 1999.