The Bayesian Cut

Taborsky, Petr; Vermue, Laurent; Korzepa, Maciej; Morup, Morten

Published in:
IEEE Transactions on Pattern Analysis and Machine Intelligence

Link to article, DOI:
10.1109/TPAMI.2020.2994396

Publication date:
2021

Document Version
Peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

The Bayesian Cut

Petr Taborsky, Laurent Vermue, Maciej Korzepa, and Morten Mørup

Abstract—An important task in the analysis of graphs is separating nodes into densely connected groups with little interaction between each other. Prominent methods here include flow based graph cutting procedures as well as statistical network modeling approaches. However, adequately accounting for the holistic community structure in complex networks remains a major challenge. We present a novel generic Bayesian probabilistic model for graph cutting in which we derive an analytical solution to the marginalization of nuisance parameters under constraints enforcing community structure. As a part of the solution a large scale approximation for integrals involving multiple incomplete gamma functions is derived. Our multiple cluster solution presents a generic tool for Bayesian inference on Poisson weighted graphs across different domains. Applied on three real world social networks as well as three image segmentation problems our approach shows on par or better performance to existing spectral graph cutting and community detection methods, while learning the underlying parameter space. The developed procedure provides a principled statistical framework for graph cutting and the Bayesian Cut source code provided enables easy adoption of the procedure as an alternative to existing graph cutting methods.

Index Terms—normalized cut, ratio cut, graph cut, modularity, degree-corrected stochastic block modeling, Bayesian inference, incomplete gamma function, image segmentation.

1 Introduction

In the analysis of graphs, partitioning nodes into groups that are highly intra-connected with few inter-group connections has become important in disparate scientific fields - from network science for the identification of communities [1], [2], computer vision for image segmentation [3], [4] and the extraction of superpixel representations [5], scene reconstruction from large community photo collections [6], video decomposition [7], to physics for the splitting of materials [8]. In fact, many problems can be rephrased as a graph partitioning problem. This includes clustering problems based on pair-wise similarity in which graph partitioning approaches have found to have merits over traditional k-means and agglomerative hierarchical clustering procedures [9], and semi-supervised learning problems in which a popular solution procedure is to use graph cuts constrained according to the labelled observations [10], [11].

A variety of computational tools have been developed for graph partitioning. As such, methods based on minimizing flow between the separated entities have been devised based on various quality measures of cutting graphs. Two prominent procedures are the ratio cut [12] and normalized cut [3], for a review see also [4], [9]. On the other end, flexible in objective function, are methods minimizing certain classes of submodular energies in pairwise Markov Random Fields with applications in computer vision [13] and extended to certain nonsubmodular functions in [14]. Recently, inference in sparse graphs recovering true partitions using side information was introduced in [15]. While providing general optimisation frameworks these methods face scaling issues. Within network science a prominent procedure to identify communities is based on optimizing the modularity measure proposed in [1], which contrasts intra-group connectivity structure relative to the connectivity structure as would be expected according to the nodes' degree distribution. Within the social sciences identifying subgroups in graphs has been addressed using stochastic block-models (SBM) [16], [17] that identify homogeneous groups with similar connectivity profiles. This framework has been advanced to community detection by constraining parameters specifying intra-connectivity to be higher than inter-connectivity based on an information theoretic compression imposing intra and inter link constraints [18] or through Bayesian modeling constraining the parameters specifying intra and inter group link densities [19]. When partitioning networks a limitation of the SBM is that it is driven by grouping nodes according to their degree distribution. This issue has been alleviated by the degree-corrected stochastic block model (dc-SBM) proposed in [20] and its non-parametric Bayesian counterpart defined in [21]. Recently, it has been proven that modularity is a special case of maximum-likelihood estimation in the dc-SBM [22] assuming a planted l-partition model [23] in which link densities within l groups are specified only by two parameters; a within community η_{in} and between community strength η_{out} and further assuming the network is community structured, i.e. $\eta_{in} > \eta_{out}$. This then corresponds to the generalized modularity quality function proposed in [24] in which modularity is perfectly recovered when $\frac{\eta_{in}-\eta_{out}}{\log(\eta_{in})-\log(\eta_{out})} = 1$ [22].

In this paper, we propose a novel computational framework for cutting graphs into communities or groups that accounts for parameter uncertainty through Bayesian modeling. Our starting point is the dc-SBM in which we explicitly impose community structure requiring the parameters specifying intra-connectivity to be strictly larger than the corresponding inter-connectivity. Although less flexible than the dc-SBM, our model is more realistic than the planted l-partition model as we endow each community separate link-densities. We derive a Bayesian inference procedure and provide an analytical solution to the corresponding constrained integral representation. On three social networks

• *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark.*
E-mail: {ptab, lauve, mjko, mmor}@dtu.dk

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Summary of the notation used.

Notation	Meaning	Definition
A	Adjacency Matrix	
A_{ij}	Link strength between node i and j	
b_c	Hyperparameter link density gap between the inter clusters expected link density and expected link density in community c	
C	Number of communities/clusters	
G	Undirected Graph	
d_i	Degree of node i	$A_{ii}/2 + \sum_{j \neq i} A_{ij}$
D_c	Sum of node degrees in cluster c.	$\sum_{i:z_i=c} d_i$
n	Total number of nodes in graph G	
n_c	Number of nodes in cluster c	$\sum_{i:z_i=c} 1$
n_{out}	Number of nodes between the clusters	$\sqrt{n^2 - \sum_{c=1}^{C} n_c^2}$
N	Total number of links in the graph G	$\sum_i A_{ii}/2 + \sum_{i<j} A_{ij}$
N_c	Number of links in cluster c	$\sum_{i:z_i=c} A_{ii}/2$ $+\sum_{i:z_i=c, j<i:z_j=c} A_{ij}$
N_{out}	Number of links between the clusters	$N - \sum_{c=1}^{C} N_c$
z_i	Cluster assignment of node i	
$\boldsymbol{z}$	Set of node assignments z_i for all nodes n	$\{z_1, z_2, \ldots, z_n\}$
Z_G	Normalizing constant of the graph G	$\prod_{i<j} A_{ij}! \prod_i \frac{A_{ii}}{2}! 2^{\frac{A_{ii}}{2}}$.
Z_{BC}	Normalizing constant of the constrained distribution	see (5)
α_c	A priori assumed link counts within community c, $\alpha_c \in \mathbb{R}^+$	
α_{out}	A priori assumed link counts between communities, $\alpha_{out} \in \mathbb{R}^+$	
β_c	A priori assumed number of network entries within community, $\beta_c \in \mathbb{R}^+$	
β_{out}	A priori assumed number of network entries between communities, $\beta_{out} \in \mathbb{R}^+$	
$\boldsymbol{\eta}$	Set of all η parameters	$\{\eta_1, \ldots, \eta_C, \eta_{out}\}$
γ	Degree correction hyperparameter	
η_c	Parameter controlling expected density of links within cluster c	
η_{out}	Parameter controlling expected density of links between clusters	
ϕ_i	Weight of node i	
$\boldsymbol{\phi}$	Set of node weights ϕ_i for all nodes n	$\{\phi_1, \phi_2, \ldots, \phi_n\}$
θ_i	Node degree control weight for node i	$n_{z_i} \phi_i$
$B(\boldsymbol{x})$	Multivariate Beta function	$\frac{\prod_k \Gamma(x_k)}{\Gamma(\sum_k x_k)}$

we demonstrate the importance of correctly accounting for community-structure when clustering nodes in graphs and that our Bayesian approach to cutting graphs have merits in contrast to the prominent graph cutting procedures outlined above. This includes better recovery of the true underlying partitioning structure of nodes into groups and more reliable inference. We further highlight the utility of the procedure for image segmentation considering both the Fast Marching Method (FMM) of [25] and the mean color regional adjacency graph (RAG) of [26] where normalized cut is typically applied. Notably, our results are for illustrative purposes demonstrated in the context of social network modeling in which the true partitioning structure is known, and image segmentation in which results can easily be visually inspected. However, we note that the computational framework developed has application beyond social network modeling and computer vision to the many domains in which graph cuts are currently used.

2 METHOD

Let G be an undirected graph with adjacency matrix A (i.e., $A_{ij} = A_{ji}$) whose elements A_{ij} are equal to the number of links between nodes i and j for $i \neq j$ and for computational reasons [20] twice that number for $i = j$. Let further n define the total number of nodes in the graph.

Following the dc-SBM [20] we assume that G is partitioned into a fixed number of C communities and the number of links between nodes i and j follow a Poisson distribution:

$$A_{ij} = \begin{cases} Poisson(\theta_i \theta_j \eta_{z_i z_j}) \text{ for } i \neq j \\ Poisson(\frac{1}{2}\theta_i^2 \eta_{z_i z_i}) \text{ for } i = j \end{cases}, \quad (1)$$

in which the parameter η_{ce} controls the probability of links between communities c and e, θ_i regulates the probability of links connected to the node i based on the degree of that node, and z_i defines the community assignment of node i. The factor of $\frac{1}{2}$ for $i = j$ results from the factor of two in the definition of diagonal elements of the adjacency matrix. In particular in all presented application in this paper self-links A_{ii} are constant. For the social networks presented they are zeros given by data, while in image applications with well defined similarities (following a common sense that node/pixel is similar to itself) they obtain maximal similarity.

As noted in [20] typically in large scale applications (i.e. images) self-links do not play a role as their effect diminish with scale ($\sim 1/n$). If necessary they can be marginalized as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

suggested in [21]. Although it may be undesired to account for self-links they add to generality of the model that makes computations and (approximate) optimisation easier, i.e. [27].

In order to keep analytic tractability of the constrained model that will be introduced later we assume all links between different communities are generated using the same value, i.e. $\eta_{ce} = \eta_{out}$ for $c \neq e$. We will also refer to η_{cc} simply as η_c and $\boldsymbol{\eta}$ as the set of all $\{\eta_1, \ldots, \eta_C, \eta_{out}\}$ parameters. Accordingly, the probability of graph G can be written as:

$$\begin{aligned} P(G|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{z}) &= \prod_{i<j} \frac{(\theta_i \theta_j \eta_{z_i z_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \theta_j \eta_{z_i z_j}) \\ &\times \prod_i \frac{(\frac{1}{2}\theta_i^2 \eta_{z_i z_i})^{A_{ii}/2}}{(A_{ii}/2)!} \exp(-\frac{1}{2}\theta_i^2 \eta_{z_i z_i}) \\ &= \frac{1}{Z_G} \eta_{out}^{N_{out}} \exp(-\frac{n_{out}^2}{2}\eta_{out}) \\ &\times \left[\prod_c \eta_c^{N_c} \exp(-\frac{n_c^2}{2}\eta_c)\right] \left[\prod_i \theta_i^{d_i}\right]. \end{aligned} \tag{2}$$

We have here used that $d_i = A_{ii}/2 + \sum_{j \neq i} A_{ij}$ is the degree of node i; $n_c = \sum_{i:z_i=c} 1$, $N_c = \sum_{i:z_i=c} A_{ii}/2 + \sum_{i:z_i=c, j<i:z_j=c} A_{ij}$ are respectively the number of nodes and links in community c; $n_{out}^2 = n^2 - \sum_{c=1}^{C} n_c^2$ and $N_{out} = N - \sum_{c=1}^{C} N_c$ with $N = \sum_i A_{ii}/2 + \sum_{i<j} A_{ij}$, whereas $Z_G = \prod_{i<j} A_{ij}! \prod_i \frac{A_{ii}}{2}! 2^{\frac{A_{ii}}{2}}$. Following [21], given partition $\boldsymbol{z}$, we define a constraint $\sum_{i:z_i=c} \theta_i = n_c$ and parametrize $\theta_i = n_{z_i}\phi_i$ such that parameters $(\phi_i)_{z_i=c}$ for each community c lie on a simplex. We endow all parameters with priors thereby accounting for uncertainty using Bayesian modeling. Thus, for given partition $\boldsymbol{z}$, we assign Dirichlet priors for the $(\phi_i)_{z_i=c}$ parameters of each community c. Further we impose Gamma priors for the elements of $\boldsymbol{\eta}$ and we obtain:

$$\begin{aligned} p(\boldsymbol{\phi}|\boldsymbol{z}) &= \prod_c \frac{1}{B(\gamma \mathbf{1}_{n_c})} \prod_{i:z_i=c} \phi_i^{\gamma-1}, \\ p(\boldsymbol{\eta}) &= \frac{\beta_{out}^{\alpha_{out}}}{\Gamma(\alpha_{out})} \eta_{out}^{\alpha_{out}-1} \exp(-\beta_{out}\eta_{out}) \\ &\times \prod_c \frac{\beta_c^{\alpha_{in}}}{\Gamma(\alpha_c)} \eta_c^{\alpha_c - 1} \exp(-\beta_c \eta_c), \end{aligned} \tag{3}$$

where $B(\boldsymbol{x}) = \frac{\prod_k \Gamma(x_k)}{\Gamma(\sum_k x_k)}$ denotes the multivariate Beta function, and γ is a hyperparameter that allows to infer the optimal strength of degree correction for a given graph such that if $\gamma \to \infty$, then $\phi_i \to \frac{1}{n_c}$ and $\theta_i \to 1$ and the model reduces to the corresponding SBM [21]. On the other hand, if $\gamma \to 0$, then $\phi_{i^*} \to 1$ and $\theta_{i^*} \to n_c$ for some node i^* in each community c and thus a network generated according to this prior becomes dominated by a few greedy nodes. α_c and α_{out} denotes the a priori assumed number of links within community c and between communities (i.e., the prior shape parameter of the Gamma distribution) whereas β_c and β_{out} denotes the corresponding a priori imposed number of network entries (i.e., the prior rate parameter of the Gamma distribution) within community c and between communities. Assuming further an uniform prior on $\boldsymbol{z}$, $P(\boldsymbol{z}) = C^{-n}$, we obtain:

$$\begin{aligned} P(G, \boldsymbol{z}) &= \int P(G|\boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{z}) p(\boldsymbol{\phi}) p(\boldsymbol{\eta}) P(\boldsymbol{z}) d\boldsymbol{\eta} d\boldsymbol{\phi} \\ &= \frac{C^{-n}}{Z_G} \frac{\Gamma(N_{out}+\alpha_{out})\beta_{out}^{\alpha_{out}}}{\left(\frac{n_{out}^2}{2}+\beta_{out}\right)^{N_{out}+\alpha_{out}} \Gamma(\alpha_{out})} \\ &\times \prod_c \frac{\Gamma(N_c+\alpha_c)\beta_c^{\alpha_c}}{\left(\frac{n_c^2}{2}+\beta_c\right)^{N_c+\alpha_c} \Gamma(\alpha_c)} \frac{B(\gamma \mathbf{1}_{n_c} + (d_i)_{i:z_i=c})}{B(\gamma \mathbf{1}_{n_c})} n_c^{D_c}, \end{aligned} \tag{4}$$

where $D_c = \sum_{i:z_i=c} d_i$ is the sum of node degrees in community c. The marginalized parameters $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_C, \eta_{out}\}$ can be interpreted as the densities of links within each community and between the communities respectively.

To ensure community structure in the graph, we presently restrict the model such that the within-community densities are larger than the between-community density. This has previously been considered in the context of the SBM [18], [19] but not in the context of the dc-SBM and without fully analytical tractable solutions to the constraints as presently derived. We constrain $\boldsymbol{\eta}$ parameters such that $\eta_c b_c \geq \eta_{out}$ for each community c where each b_c is a hyperparameter within range $[0, 1]$ specifying a density gap between the inter and intra community densities as considered in the context of the standard SBM in [19]. We introduce this constraint by defining the following constrained prior on the $\boldsymbol{\eta}$ parameters

$$\begin{aligned} p_{BC}(\boldsymbol{\eta}) &= \frac{1}{Z_{BC}} \eta_{out}^{\alpha_{out}-1} \exp(-\beta_{out}\eta_{out}) \\ &\times \left(\prod_{c=1}^{C} \eta_c^{\alpha_c - 1} \exp(-\beta_c \eta_c)\right) I(\boldsymbol{\eta}), \end{aligned} \tag{5}$$

where $I(\boldsymbol{\eta}) = \prod_c \chi_{[0;\infty[}(\eta_c - b_c \eta_{out})$ is an indicator function evaluating to 1 if the constraints are satisfied and zero otherwise ($\chi_{[a;b]}(x)$ is the standard step function evaluating to one if $x \in [a; b]$ and 0 otherwise). Z_{BC} is the normalizing constant of this constrained distribution. For a summary of the notation used see table 1.

Combining priors with the likelihood function and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

marginalizing the $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ parameters gives:

$$\begin{aligned}
p(G,\boldsymbol{z}) &= \int p(G|\boldsymbol{\phi},\boldsymbol{\eta},\boldsymbol{z})p(\boldsymbol{\phi}|\boldsymbol{z})p_{BC}(\boldsymbol{\eta})p(\boldsymbol{z})d\boldsymbol{\eta}d\boldsymbol{\phi} \\
&= \int \eta_{out}^{N_{out}+\alpha_{out}-1}\exp\left(-\eta_{out}(\frac{n_{out}^2}{2}+\beta_{out})\right) \\
&\times \left[\prod_c \eta_c^{N_c+\alpha_c-1}\exp\left(-\eta_c(\frac{n_c^2}{2}+\beta_c)\right)I(\boldsymbol{\eta})\right. \\
&\times \left.\frac{B\left(\gamma\mathbf{1}_{n_c}+(d_i)_{i:z_i=c}\right)}{B(\gamma\mathbf{1}_{n_c})}n_c^{D_c}\right]d(\boldsymbol{\eta}) \times \frac{C^{-n}}{Z_G Z_{BC}} \\
&= \int_0^\infty e^{-\eta_{out}(\frac{n_{out}^2}{2}+\beta_{out})}\eta_{out}^{N_{out}+\alpha_{out}-1} \\
&\times \prod_{c=1}^{C}\Gamma\left(N_c+\alpha_c,\eta_{out}\times\left(\frac{\frac{n_c^2}{2}+\beta_c}{b_c}\right)\right)d\eta_{out} \\
&\times \left[\prod_{c=1}^{C}\left(\frac{n_c^2}{2}+\beta_c\right)^{-(N_c+\alpha_c)}\right. \\
&\times \left.\frac{B\left(\gamma\mathbf{1}_{n_c}+(d_i)_{i:z_i=c}\right)}{B(\gamma\mathbf{1}_{n_c})}n_c^{D_c}\right] \times \frac{C^{-n}}{Z_G Z_{BC}},
\end{aligned} \tag{6}$$

where in the second step we used change of variables $s = \eta_c(\frac{n_c^2}{2}+\beta_c)$ to obtain each of c integrals in the form of an upper incomplete gamma function (in the following simply referred to as incomplete gamma function) given by [28]:

$$\Gamma(\alpha,x) = \int_x^\infty s^{\alpha-1}e^{-s}ds \tag{7}$$

A major challenge that remains and we presently solve is to analytically marginalize η_{out} in the above expression thereby solving analytically for the constraints specified by $I(\boldsymbol{\eta})$.

2.1 Marginalization of constrained η parameters

According to eq. (6) marginalizing under the constraint imposed by $I(\boldsymbol{\eta})$ requires the solution to an integral of the following form:

$$\int_0^\infty e^{-B_0 x}x^{\mu_0-1}\left(\prod_{c=1}^{C}\Gamma(\mu_c,B_c x)\right)dx, \tag{8}$$

(Marginalizing integral)

Where we have used the following substitutions, $x = \eta_{out}$, $\mu_c := N_c + \alpha_c$, $\mu_0 := N_{out} + \alpha_{out}$, $B_c := \frac{n_c^2}{2} + \beta_c$, $B_0 := \frac{n_{out}^2}{2} + \beta_{out}$, and ignored all terms independent on η_{out}. As a result, the μ_c and B_c elements in (8) relate respectively to scale and rate parameters of the involved incomplete gamma functions.

We outline what is to the best of our knowledge a novel approach solving integrals of the form presented in Eq.(8). We exploit the following known recurrence property of incomplete gamma functions (see Theorem 1 in [29]): $\Gamma(a+1,x) = a\Gamma(a,x) + x^a e^{-x}$ for $a \in \mathbb{R}, a > 0$. This can be considered a generalization of $\Gamma(n+1) = n\Gamma(n)$ to the incomplete Gamma function. By a simple recursion of this property we obtain

$$\Gamma(a,x) = \frac{\Gamma(a+K,x)}{(a)^{\dot{K}}} - x^a e^{-x}\sum_{i=0}^{K-1}\frac{x^i}{(a)^{\dot{i+1}}}$$

(K-recurrence of $\Gamma's$)

where $(a)^{\dot{n}}$ is the Pochhammer symbol (a.k.a. "rising factorial") defined as $(a)^{\dot{n}} = \Gamma(a+n)/\Gamma(a)$. This recursion formalizes idea of "shifting" of shape parameters of gamma distribution as shown in figure 1.

The following theorem presents application of the "shifting" method described above to solve the multidimensional incomplete gamma integral in equation (8) up to an arbitrary precision.

Theorem 2.1. *For every $C \in \mathbb{N}^+$, $\mu_i, B_i \in \mathbb{R}, \mu_i > 0, B_i > 0$ for $i \in \{1,...,C\}$ and $K \in \mathbb{N}^+$ following equality holds:*

$$\int_0^\infty e^{-xB_0}x^{\mu_0-1}\prod_{c=1}^{C}\Gamma(\mu_c,B_c x)dx \tag{9}$$

$$\begin{aligned}
&= \sum_{m_1=1}^{C}\sum_{\substack{m_2=1,\\ m_2\neq m_1}}^{C}\cdots\sum_{\substack{m_C=1,\\ m_C\neq m_1,\ldots,m_{C-1}}}^{C}\sum_{i_1=0}^{K-1}\cdots\sum_{i_C=0}^{K-1} \\
&\prod_{w=1}^{C}\frac{B_{m_w}^{\mu_{m_w}}(B_0+\sum_{j=1}^{w-1}B_{m_j})^{i_w}}{\left(\mu_0+\sum_{j=1}^{w-1}(\mu_{m_j}+i_j)\right)^{i_w\dot{+}1}} \\
&\times \frac{\Gamma\left(\mu_0+\sum_{j=1}^{C}(\mu_{m_j}+i_j)\right)}{(B_0+\sum_{j=1}^{C}B_{m_j})^{(\mu_0+\sum_{j=1}^{C}(\mu_{m_j}+i_j))}} \qquad (10) \\
&+ E(K), \qquad (11)
\end{aligned}$$

where the error term $E(K)$ satisfies $\lim_{K\to\infty}E(K) = 0$.

Proof. Detailed proof altogether with additional two proven lemmas is to be found in appendix. (6.3) □

To evaluate the joint distribution $p(G,\boldsymbol{z})$ the integral (8) is to be evaluated twice. First to compute prior normalization factor of hyperparameters (α's being gamma priors), denoted Z_{BC}, and second to evaluate the integral (6) with shape parameters μ's that are result of α's added together with link counts from the respective clusters.

While the former can be efficiently solved by theorem 2.1 as the prior values are typically small requiring a small value of K, the latter imposes substantial computational challenges especially for large and dense graphs where the use of theorem 2.1 becomes computationally heavy as the required K has to be in orders of magnitudes of the number of links in the largest cluster.

Rather than resorting to analytical integration one could opt for the use of point estimates in the large setting where the posterior distribution can be expected to be peaked and thereby point estimates to provide reasonable accuracy or apply simple normal approximations through the Laplace

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

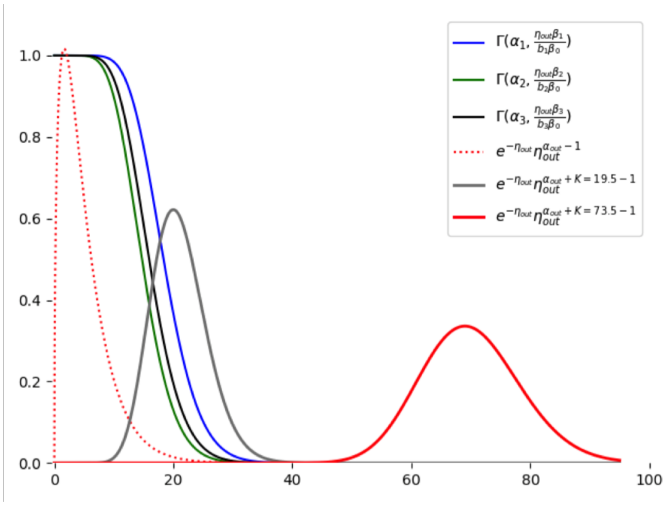

Fig. 1: Decomposition of integrand of Part A in lemma 6.1 into elements and shift of original gamma pdf (dotted red) using (K-recurrence of $\Gamma's$) to the inadequately shifted (gray curve K=19.5) and adequately shifted (red curve K=73.5) with close to zero-mass area of all the considered incomplete gamma functions (blue, green, and black curves) whereby the product in Part A becomes close to zero. As a result, the size of the shift controls the closeness to zero of Part A.

procedure also potentially accounting for the constraints using the result of the work of Hartman at al. [30] from 2017. Notably, a simple point estimate would be the maximum a posteriori of $\boldsymbol{\eta}$ under the required constraint and as the posterior is convex with convex constraints on $\boldsymbol{\eta}$ the MAP estimation of the constrained $\boldsymbol{\eta}$ is convex. Alternatively, η_{out} could be sampled and conditioned on the sampled value of η_{out}, η_c could be analytically marginalized using the incomplete Gamma function. While these approaches are scalable they are approximate and for the large scale setting we therefore opt for the following analytic procedure accounting explicitly for the uncertainty of $\boldsymbol{\eta}$ while keeping complexity at $O(C)$ for evaluating (8) which is the same as can be achieved by use of point estimates.

2.2 Large Scale Settings

Up until now there were no limitations set on values of η and in particular of hyperparameters μ and B, besides being real and positive. In large scale applications however, we are often facing large values of $\mu_{c,c\in\{1,\ldots,C\}}$. In such case, it is convenient to consider evaluation of the integral for integer values of the μ's. As we present in the following theorem, for integer μ's the integral is proportional to the CDF of the Negative Multinomial distribution with easy to evaluate limiting distribution. Notably, it is shown in section 3.1 that resorting to the integer setting imposes no significant constraints for most large scale applications.

Next we present main result of this section: exact evaluation of the integral (8) in case of integer shape parameters of involved gamma densities:

Theorem 2.2. *For $C \in \mathbb{N}^+$, $\mu_i \in \mathbb{N}^+$ and $B_i \in \mathbb{R}^+$, $i \in \{0, \ldots, C\}$ integral* (8) *is proportional to the cumulative distribution function of the Negative Multinomial (NMn) distribution and the following equality holds:*

$$\int_0^\infty e^{-xB_0} x^{\mu_0-1} \prod_{i\in\{1,\ldots,C\}} \Gamma(\mu_i, B_i x)dx$$

$$= \frac{\prod_{c=0}^{C}\Gamma(\mu_c)}{B_0^{\mu_0}} \times \tag{12}$$

$$\times \sum_{i_1=0}^{\mu_1-1} \cdots \sum_{i_c=0}^{\mu_c-1} \frac{\Gamma(\mu_0+i_1+\ldots+i_c)}{\Gamma(\mu_0)i_1!\ldots,i_c!} \left(\frac{B_0}{B}\right)^{\mu_0} \prod_{c=1}^{C}\left(\frac{B_c}{B}\right)^{i_c}, \tag{13}$$

where $B := \sum_{i=0}^{C} B_i$

Proof. To be found in Appendix (6.4). For Negative Multinomial distribution definition and properties refer to [31]. □

The connection to Negative Multinomial distribution shown in theorem 2.2 also allows for an interpretation of the marginalized posterior (8) probability. If we consider sequence of independent multinomial trials in each of which event E_i occurs with probability $p_{i,i\in\{0,\ldots,C\}}$, $\sum_{i=0}^{C} p_i = 1$ and let X_i be the frequency of $E_{i,i\in\{1,\ldots,C\}}$ "successes" before predefined number μ_0 of X_0 "failures" appears, then $(X_0, X_1, \ldots, X_C)$ follows the Negative Multinomial distribution NMn [31].

Hence integral (8) is proportional to the likelihood of observing μ_i "successes" (links within clusters) before number of "failures" (links between clusters) reaches at most μ_0, given that number of links in graph follows multinomial distribution with probability of links appearing in cluster c being $\frac{B_c}{B}$, which is positively related to the relative size of a cluster (proportion of nodes in cluster) c.

In the following section we make use of favourable asymptotics of the Negative Multinomial distribution to derive a fast evaluation of the integral for the large scale setting.

3 Inference

We presently show how to efficiently evaluate Theorem 2.1 for $C = 2$ clusters. In this case, the formula (omitting the error term) can be written as:

$$B_1^{\mu_1} B_2^{\mu_2} \Gamma(\mu_0) \sum_{m=1}^{2} \sum_{i_1=0}^{K-1} \sum_{i_2=0}^{K-1} \frac{(1+B_m)^{i_2}}{(1+B_1+B_2)^{\mu_T+i_1+i_2}} \times \frac{(\mu_0+i_1+1)^{(\mu_m-1)}\Gamma(\mu_T+i_1+i_2)}{\Gamma(\mu_0+\mu_m+i_1+i_2+1)},$$

where $\mu_T = \mu_0+\mu_1+\mu_2$. If we apply substitution $v = i_1+i_2$, we can rewrite the above expression as:

$$B_1^{\mu_1} B_2^{\mu_2} \Gamma(\mu_0) \sum_{m=1}^{2} \sum_{v=0}^{2(K-1)} \frac{(1+B_m)^{v}\Gamma(\mu_T+v)}{(1+B_1+B_2)^{\mu_T+v}} \times \sum_{i_1=0}^{min(v,K-1)} \frac{(\mu_0+i_1+1)^{(\mu_m-1)}}{(1+B_m)^{i_1}}.$$

We notice that the sums dependent on v or i_1 can be evaluated independently in $\mathcal{O}(K)$ time which allows for efficient evaluation compared to the original $\mathcal{O}(K^2)$ time.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

With regards to control of the approximation error Theorem 2.1 gives for arbitrary error thresholds ϵ the existence of K that evaluates this integral up to ϵ precision. However, the Theorem is not explicit about the choice of a sufficient value of K. One simple approach for finding K to control approximation error we used to produce the results presented in section 2.1 is to set K such that the mode of inter cluster link density $\frac{\mu_0+K-1}{B_0}$ is equal or greater than the q-quantile of all gamma distributions controlling intra clusters link densities. An accuracy is then controlled by setting values of q. Results of this application on karate network are shown in figure 2. There are many alternative choices for K, however, we found this approach to be easy and efficient in practice. For the purpose of error evaluation we compared results of Theorem 2.1 with results of the scipy.integrate.quad function from the scipy 1.2.0 python package. From the figure we can observe how increasing K, corresponding to increasing the q-quantile according to the method described above, controls the absolute error on the evaluation of the integral. For the results obtained in the following we used $q = 0.9999$, given that this guarantees an absolute error close to 10^{-5}, but in most cases will range around 10^{-9}.

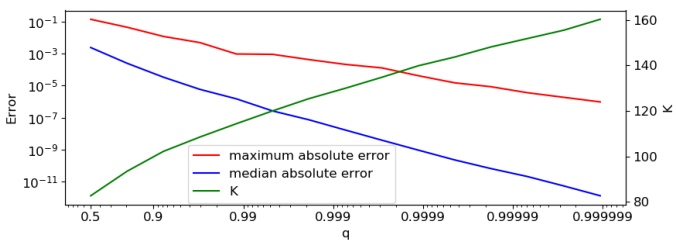

Fig. 2: Maximum and median absolute approximation error and corresponding number of added observations T for karate network based on 100 chains with 100 samples each.

Typically, a graph cut is obtained by optimizing a given cost function. In case of Bayesian Cut, the cost function is defined by the posterior distribution $p(\boldsymbol{z}|G)$ which specifies probability of every possible partition of graph G. While the full posterior would provide lots of insight into different ways of cutting the graph, due to its high complexity, it is not possible to determine it fully. Instead, the most reasonable approach is to search for the maximum of the posterior (MAP) $\boldsymbol{z}_{MAP} = \text{argmax}_{\boldsymbol{z}} p(\boldsymbol{z}|G)$. While one could opt for optimization of the posterior distribution of $\boldsymbol{z}$ possibly making use of wide arsenal of approximation methods i.e. [14], [13], [32] or other discrete optimisation methods [33] to this NP hard problem, we advocate using MCMC sampling (for reference see [34], Chapter 11) before performing optimization for a few reasons. First of all, optimization might get stuck in local maxima while sampling given enough time will find the global maximum. In practice, within the sampling budget, the sampler will likely focus on some high density region of the posterior, but it will still explore multiple modes within that region. A comparison of only using optimization compared to using the sampler can be found in the appendix, see section 6.3. Secondly, by using sampling we are able to infer values of specific hyperparameters to create a more plausible model that explains the observed data better and thus learn about the underlying structure of the problem. Finally, sampling produces not only a point estimate but an approximation of the true posterior (more or less accurate depending on its complexity and sampling budget) that can be used to answer more complex questions than what the most probable cut is. To perform MCMC sampling, we use Gibbs sampling and sample each element z_i of $\boldsymbol{z}$ independently:

$$p(z_i|G) = \frac{p(G, \boldsymbol{z})}{p(G, \boldsymbol{z}_{-i})} \propto p(G, \boldsymbol{z}).$$

We treat the hyperparameter γ as a random variable while fixing other parameters to a constant value. We use the non-informative prior $p(\gamma) = \gamma^{-1}$ and after each Gibbs sweep over all nodes in the graph, we perform 20 Metropolis-Hastings (MH) updates using the proposal distribution $\gamma^* = \gamma \exp(\epsilon), \epsilon \sim N(0, \sigma = 0.1)$. Alternatively, if one is not interested in inferring γ, it can be set to 1 which assumes any configuration of node-specific parameters $(\phi_i)_{z_i=c}$ of community c is equally probable (i.e., corresponding to the uniform distribution over the $(n_c$–$1)$-simplex). Furthermore, we fix all α and β parameters to a non-informative value 0.01 and set b to 1 unless specified otherwise. After running out of the sampling budget, we apply deterministic optimization by switching node assignments only when it leads to higher likelihood and we stop when in a full sweep over all nodes we do not observe any further improvement.

3.1 Inference for large graphs

Posterior distribution of η (expected density of links) in BC model has the same form as prior (due to the conjugacy between Poisson and Gamma) where 'shape' μ_c and 'rate' B_c, in general positive real parameters of the involved gamma densities, are by definition priors $\alpha_c \in \mathbb{R}^+$ and $\beta_c \in \mathbb{R}^+$ updated by the added number of links and nodes in cluster c respectively. This often results in large, in general real, values when dealing with large graphs. In order to find a fast evaluation algorithm first let us note that for the cutting of large graphs limiting ourselves to integer shape hyperparameters of both prior and posterior gamma densities while updating real 'rate' impose any relevant constraints in most applications as the prior is overwhelmed by the observed data. Technicaly speaking transformation from $\mu' = \lceil \mu \rceil, B' = \frac{\lceil \mu \rceil}{\mu} B$ keeps mean of posterior gamma distribution unchanged ($\frac{\mu'}{B'}$) while increases its variance (or uncertainty) ($\frac{\mu'}{B'^2}$) by factor diminishing with scale. Therefore and especially with uninformative priors resorting to integer 'shape' should have an insignificant and asymptotically zero effect on posterior for large values of μ and if not the general Theorem (2.1) should be applied.

Secondly, as shown in [31], a limiting distribution of the negative multinomial decomposes into a product of Poisson distributions as $\mu_0 \to \infty$. Making use of this limiting distribution we obtain a large scale (asymptotic) solution of our integral. In the following we make use of the fact that the cummulative density function (cdf) of a Poisson distributed variable $F_{Pois(\lambda)}(\mu_i - 1)$ can be written as $\Gamma(\mu_i - 1; \lambda)$. Let m denote the threshold beyond which the asymptotic is applied. To determine m we analyze in Figure 3 how well the asymptotic approximation of the marginalized integral (8) behaves. The figure shows that absolute error of log integral

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

is close to zero but for the bipartite setting, corresponding to a graph in which all links/similarities are between clusters while there is zero density of link/similarity within clusters. In such setting it is still possible to evaluate the integral exactly using Theorem (2.2) with complexity $O(N)$. However, if the observed graph G has bipartite structure (can be detected prior to application of the method) then the proposed asymptotic becomes expensive. This does not impose any issues for most applications, in particular, for image segmentation where bipartite structures are unlikely. Formally, proposed method to evaluate the marginalized integral (8) in large scale settings depends on sum of weights (in our case number of links) between clusters, μ_0:

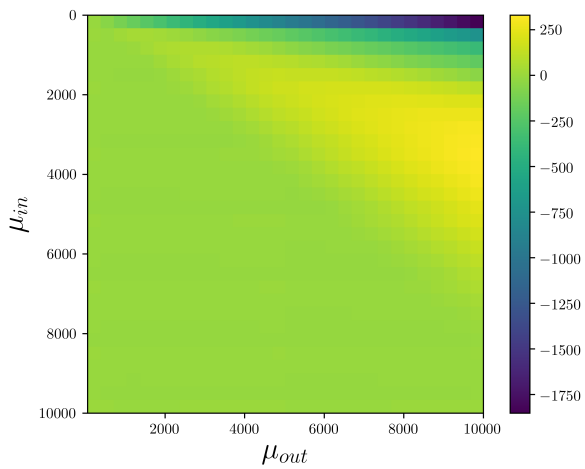

Fig. 3: Error of logarithm of integral (8) evaluated by "shifted" method of theorem (2.1) with bounded error of 10^{-5} and logarithm of same integral evaluated by asymptotic method of section (3.1). For experiments we fixed $B_0 = 60$ and $B_1 = B_2 = 70$ while ranging $\mu_{out} \in (51, 10^3)$ and $\mu_1 = \mu_2 = \mu_{in} \in (0, 10^3)$.

Large $\mu_0 > m$**:** If μ_0 is sufficiently large, then (41) resolves asymptotically into:

$$\prod_{i=1}^{C} \Gamma(\mu_i - 1; \mu_0 B_i / B_0) \tag{14}$$

Small $\mu_0 \leq m$**:** In this case there are 2 options:

- In case the smallest of μ_c's, $_{c\in\{1,\ldots,C\}}$, is sufficiently large $\min_c(\mu_c) > m$ we apply 'per-partes' on (41) to rotate elements of integral and asymptotic decomposition on each of C summands resulting in:

$$B_0^{-\mu_0} \prod_{i=0}^{C} \Gamma(\mu_i) - \sum_{j=1}^{C} \prod_{\substack{i=0,\\ i\neq j}}^{C} \Gamma(\mu_i - 1; \alpha_j B_i / B_0) \tag{15}$$

- Else, when one or more μ_c's, $_{c\in\{1,\ldots,C\}}$, are small $(\min_c(\mu_c) < m)$, asymptotic properties of NMn are of no use. This corresponds to a degenerated case when nodes within one or more clusters are dissimilar or respective clusters contain few nodes. In either case this does not correspond to a preferable cut. Let's note that it is unlikely that the Metropolis - Hastings/ Gibbs MCMC sampler appears to be sampling from an assignment corresponding to this case unless observed graph has aforementioned bipartite structure. In degenerate case sampler would need to accept low probability proposals against the imposed constraints on the link densities. So unless initial assignments of sampler are degenerate or number of clusters C is extremely large compared to nodes in the considered graph, it is unlikely to end up in such case during sampling. Anyway, in such case we evaluate the integral of theorem 2.2 directly at cost of higher complexity $O(N)$ instead of $O(C)$.

In the procedure above m represents a threshold above which asymptotic apply. In our image experiments (non bipartite structure) we applied $m = 50$ given results of fig. 3, striking balance between accuracy and runtimes. More elaborate and/or conservative choices may be better suited, depending on use.

3.2 Reference methods

We contrast the proposed BC to the corresponding dc-SBM without community constraints given by (4) as well as to modularity optimization (Mod), ratio-cut (RC) and normalized cut (NC). The solutions obtained by RC and NC were derived using the spectral clustering procedure described in [9] whereas the modularity objective was optimized using the spectral approach described in [1]. We note that the spectral optimization procedure may be suboptimal to other inference approaches, however, we presently use these solutions for illustrative purposes to characterize the methods and contrast favourable configurations by these approaches to the favorable configurations using the proposed BC procedure.

To evaluate the modularity score of a given partition we use the modularity objective function described in [1], given by

$$Q(\boldsymbol{z}) = \frac{1}{4m} \sum_{c=1}^{C} \left[\sum_{i:z_i=c} \left(\sum_{j:z_j=c} (A_{ij} - \frac{k_i k_j}{2m}) - \sum_{j:z_j\neq c} (A_{ij} - \frac{k_i k_j}{2m}) \right) \right], \qquad m = \frac{1}{2} \sum_{i=1}^{n} k_i \tag{16}$$

To evaluate solutions in the domain of NC and RC we use their respective cost functions as defined in [9]

$$RC(\boldsymbol{z}) = \frac{1}{2} \sum_{c=1}^{C} \frac{1}{n_c} \sum_{i:z_i=c} \sum_{j:z_j\neq c} A_{ij}, \tag{17}$$

$$NC(\boldsymbol{z}) = \frac{1}{2} \sum_{c=1}^{C} \frac{1}{K_c} \sum_{i:z_i=c} \sum_{j:z_j\neq c} A_{ij}. \tag{18}$$

3.3 Visualization technique for solution landscape

To show the solutions supported by each procedure we plot the solution landscapes similar to the method proposed in [35]. These landscapes were created by obtaining a set of V $\boldsymbol{z}$ vectors for the models under scrutiny. This set of vectors is expanded by 50% to cover the in between solution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

space through pseudo-random vectors, i.e. a new vector is generated by randomly taking two distinct $\boldsymbol{z}$ vectors and combining half of the elements of each vector. The score or likelihood for each vector is subsequently obtained by running the specified model with each unique $\boldsymbol{z}$ vector. As measure of distance between partition vectors we use the Variation of Information [36] between all V vectors. The resulting $V \times V$ dimensional distance matrix is reduced to two dimensions using Multidimensional Scaling [37]. Discrete Sibson Interpolation [38] is subsequently used to obtain a meshgrid of the remaining two dimensions.

4 Results and Discussion

In the following we analyze the properties of the proposed Bayesian Cut (BC) model for community detection in social networks and image segmentation for computer vision.

We first present results on a set of simple synthetic networks (Section 4.1) followed by results for community detection in social networks (Section 4.2) that have an advantage of available "ground truth" as well as unified definition of an adjacency matrix across methods we compare with. Hence presented comparison provides insights on graph cutting performance more clearly than in the subsequent image applications where cuts are used in connection with disparate similarity matrices. In Section 4.3 two often used similarity matrices are presented to demonstrate utility of BC model as a generic tool for graph cuts. We further apply multiple cluster solutions on the images.

The Bayesian Cut source code used for these experiments is provided through a public source code repository, hosted on Github (https://github.com /DTUComputeCognitiveSystems /bayesian_cut), and through the Python Package Index (https://pypi.org/project/bayesian-cut/) to allow a straightforward installation of the package. To ensure accessibility and reproducibility of the results, the repository includes image of "bears" used in experiments (original downloaded from: https://images.app.goo.gl /Mvdra73AwjfRfp629) and the software, that is accompanied by instructions on how to use the package and Jupyter Notebooks that show how the results were obtained.

4.1 Synthetic networks

We test the proposed algorithm on synthetic networks to demonstrate the effect of imposed connectivity constraint on the inference. In this experiment, we fix the total number of nodes to $n = 100$ and links to $N = 1000$. We assume networks are partitioned into two communities having equal number of nodes ($n_1 = n_2 = \frac{n}{2}$) and links ($N_1 = N_2 = N_{in}$). For different values of intra- to inter-community link density ratio ($\eta_{in}/\eta_{out} = \frac{2N_{in}}{N_{out}}$), we generate network to match these predefined properties. We fix γ to 10^6 (to remove effects of degree correction), α_{out} to 10^{-6} (to remove the difference coming from marginalizing the constrained vs. unconstrained prior) while keeping values of the other hyperparameters as specified in Section 3. In Figure 4, we show the posterior densities (up to a constant) of the partition for a wide range of η_{in}/η_{out} density ratios for a constrained (Bayesian Cut) and corresponding unconstrained (dc-SBM) model to demonstrate the effect of the constraint. Condition $\eta_{in}/\eta_{out} < 1$ represents an extent of constraint violation - the closer it is to 0, the stronger the violation. At extreme of 0, the partition represents a bipartite network which is a structure exactly opposite to a community structure. As it can be seen in the figure, unconstrained dc-SBM assigns very high probability to partitions where there is very distinct difference between intra- and inter-community densities, even if inter-community density is higher. On the other hand, the constrained model penalizes partitions that violate the constraint and assigns them even lower probability than to partitions with the density of links uniformly distributed over the whole graph.

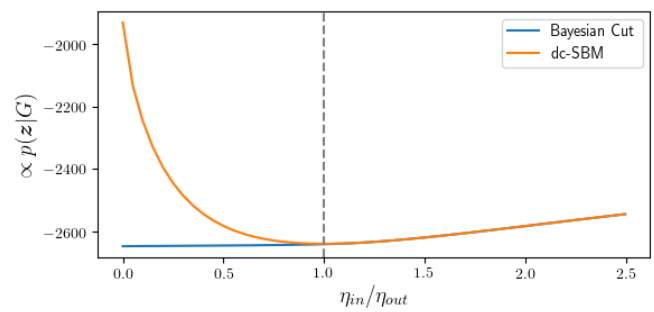

Fig. 4: Experiment on synthetic networks confirms that constrained BC model "Bayesian cut" strongly penalizes partitions that violate graph connectivity constraint, $\eta_{in} \geq \eta_{out}$, compared to unconstrained "dc-SBM" model that assigns very high probability to partitions where there is very distinct difference between intra- and inter-community densities, even if inter-community density is higher.

4.2 Community detection in social networks

For community detection the properties of the proposed Bayesian Cut (BC) model are analyzed based on three real world social networks and contrasted to ratio-cut, normalised cut, modularity and the unconstrained dc-SBM. The networks considered are:

Karate: A social undirected network studied by Zachary [39] of ties in a Karate club that turned out to split in two. The network consists of 34 nodes and 78 edges and was partitioned using modularity in [1].

Polblogs: The political blogosphere (Polblogs) network on US politics assembled by [40]. We consider the largest connected component of the network in the undirected form used in the dc-SBM analysis of [20] which contains 1222 nodes and 16714 edges.

HIV-1: Sexual partnership network extracted from the first study (Colorado Springs Project 90) in HIV Transmission Network Metastudy Project [41]. We consider the largest connected component of the network consisting of 1888 nodes and 2096 edges.

In all analyses we used $C = 2$ corresponding to the ground-truth structure of the split in Karate club and political blogs along party line. Notably, when $C = 2$ there is only one η_{out} parameter in the dc-SBM and our analyses correspond to the dc-SBM parametrization with and with-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

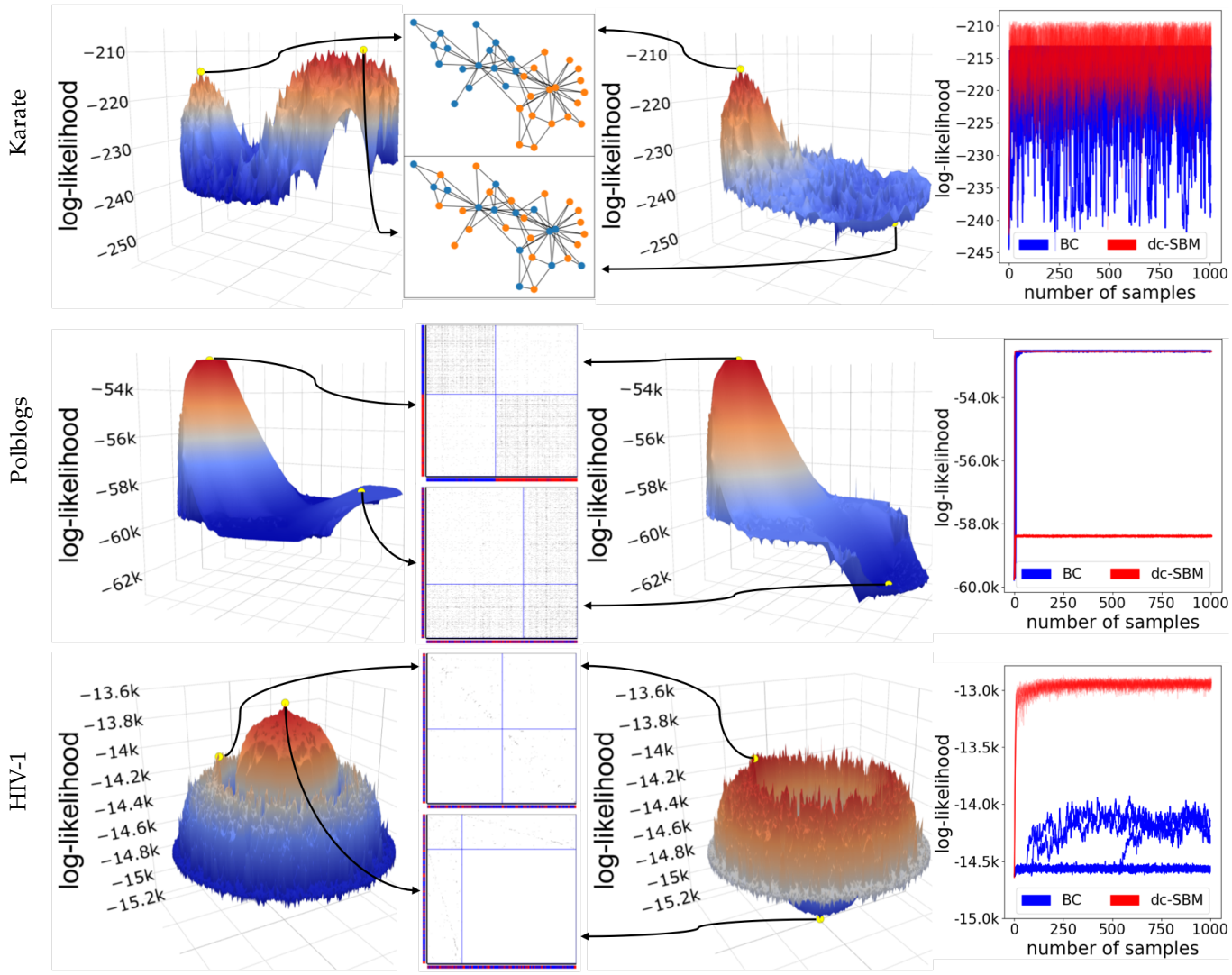

Fig. 5: Comparison of the dc-SBM (left column) and BC (right column) solution landscapes based on $p(\boldsymbol{z}|G)$ as well as the resulting cuts performed on the three networks. The outer right column shows the trace plots obtained running each model with 15 chains and 1000 samples. The dc-SBM model exhibits for all three networks modes and thus resulting cuts that violate the constraint $\eta_{in} \geq \eta_{out}$. In the corresponding adjacency matrices it can be seen that whenever the constraint is violated (lower adjacency matrix/network for each example), the off-diagonal blocks have a higher density than at least one of the diagonal blocks. In contrast, the proposed BC model gives those regions of the solution landscape that violate the constraint lower likelihoods, which leads only to modes and thus resulting cuts that do not violate the constraint.

out the community constraint. For model inference in the dc-SBM and BC we use Gibbs sampling to infer $\boldsymbol{z}$.

4.2.1 Comparison of dc-SBM and BC

Figure 5 shows the results of the unconstrained dc-SBM and our Bayesian Cut (BC) procedure for $b_1 = b_2 = 1$, i.e. imposing the constraint $\eta_1 \geq \eta_{out}$ and $\eta_2 \geq \eta_{out}$. Furthermore, a non-informative prior is used, i.e. $\alpha_{in} = \alpha_{out} = \beta_{in} = \beta_{out} = 0.01$. For the Karate network (top panel) we observe that the conventional Bayesian dc-SBM (given by the likelihood in (4)) creates a substantially different solution from our proposed BC. While our BC peaks around the true split of the Karate network, we observe that the samples of the conventional dc-SBM concentrate around two modes of the distribution in which the other mode represents a configuration that does not comply with the notion of community structure, but has a significantly higher likelihood.

For the larger Polblogs network we again observe that the dc-SBM exhibits one mode that does not comply with the community structure and creates a split leading to one community with high link density and one community with a bipartite structure, while the mode shared with our proposed model corresponds well to a separation along political orientation (i.e., democrat vs. republican). In the bottom panel for the HIV-1 network we observe a substantial difference between the dc-SBM and our proposed BC procedure with no shared modes. Here the unconstrained model identifies a bipartite structure in which one community has very low link density as compared to the inter community link density, whereas the constrained model by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

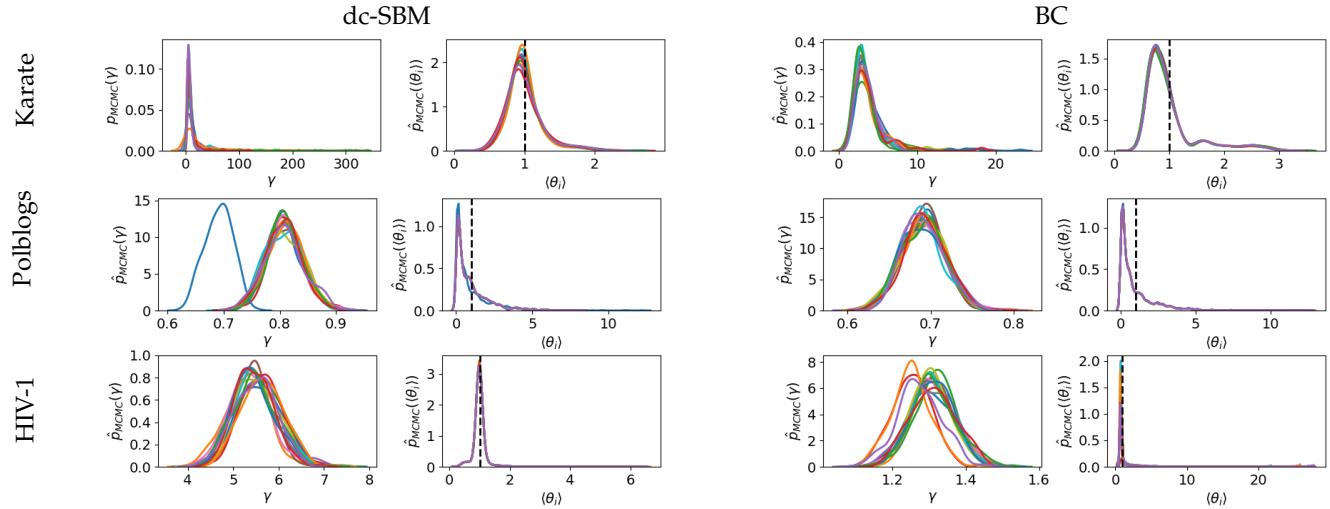

Fig. 6: Gamma inference and resulting node degree correction (theta) of the dc-SBM and BC for all three networks. The dc-SBM and BC models show substantial differences, since the parameter inference resulting from the BC model is more reliable, because it contrary to the dc-SBM does not get stuck in local optima that violate the constraint.

only giving community structure support strives to separate the network according to identifying separate communities.

Overall it can be seen that the BC with its constraints is more in line with the natural splits in the Karate and Polblogs networks and suggests a more sensible split for the HIV-1 network. The sub-optimal congruity of the unconstrained model can be attributed to the local modes of the posterior observed in Figure 5 that are unsupported by the BC procedure.

On the outer right side in Figure 5 the convergence of the dc-SBM and BC is illustrated for 15 chains and 1000 samples. Notably, we observe that for the Karate and political blogosphere networks the unconstrained model explores the mode not complying with community structure. For the political blogosphere the inference for most of the chains is stuck in the local sub-optimal mode of the posterior distribution, incapable of escaping this mode by the Gibbs sampler and recovering the underlying correct structure leading to the lower cut shown in the middle panel of figure 5. In contrast all chains of the BC model converge to the underlying partitioning structure for both networks. When considering the HIV-1 network it can be observed that the solution space supporting community structure is consisting of a vast number of local optima, contrary to the non-community supporting structure, which has a strong global mode. This is causing the community structure inferred to be less reliable and the chains to end in local modes of the community constrained posterior.

The influence of BC and dc-SBM on inferring the parameter controlling for degree (γ) is shown in figure 6. Here the gamma inference as well as the node degree correction distribution of each chain of both the dc-SBM and BC model is shown for the three networks. Focusing on the left column, a substantial difference within the inference of the γ parameter, i.e. controlling the degree correction, is observable. Subsequently, the derived θ parameters differ based on the modes preferred by the models. As previously shown, the dc-SBM model often gets stuck in local modes or exhibits globally preferred modes that do not support the community structure. Accordingly, the parameter inference is biased by the modes in the non-community structure region in those cases. In the above analysis we used non-informative priors on $\boldsymbol{\eta}$, however, we could also impose an informed prior favoring community structure in the dc-SBM. This and role of constraint parameter b is further addressed exemplary on the karate network in the appendix, section 6.4.

TABLE 2: Comparison of Cuts running 100 chains with 1000 samples without and with (in parenthesis) deterministic optimization in terms of their modularity value (Mod.) and correspondence to ground truth partition structure (avaible for Karate and Polblogs) as quantified using normalized mutual information (NMI). $\langle\cdot\rangle$ denotes average value and $\lceil\cdot\rceil$ maximum value.

	Score	RC	NC	MOD	dc-SBM	BC
Karate	$\langle$NMI$\rangle$	0.415	0.732	-	0 (0)	**0.837 (0.837)**
	$\lceil$NMI$\rceil$	0.578	0.732	**0.837**	0 (0)	**0.837 (0.837)**
	$\langle$Mod.$\rangle$	0.236	0.356	-	-0.267 (-0.258)	**0.371 (0.371)**
	$\lceil$Mod.$\rceil$	0.313	0.356	**0.371**	-0.267 (-0.258)	**0.371 (0.371)**
Polblogs	$\langle$NMI$\rangle$	0.017	0.017	-	0.143 (0.146)	**0.717 (0.718)**
	$\lceil$NMI$\rceil$	0.017	0.017	0.693	0.727 (0.737)	**0.739 (0.739)**
	$\langle$Mod.$\rangle$	0.001	0.001	-	-0.057 (-0.062)	**0.426 (0.426)**
	$\lceil$Mod.$\rceil$	0.001	0.001	0.424	**0.426 (0.426)**	**0.426 (0.426)**
HIV	$\langle$Mod.$\rangle$	0.045	0.045	-	-0.363 (-0.365)	**0.185 (0.411)**
	$\lceil$Mod.$\rceil$	0.045	0.045	0.190	-0.357 (-0.359)	**0.385 (0.463)**

4.2.2 Comparison of dc-SBM and BC to Modularity, NC and RC

In Table 2 we quantify the correspondence as measured by normalized mutual information (NMI) between the inferred partitions and the partition defined by the underlying split with highest support for each of the considered methods in the Karate network and separation according to party line in Polblogs. Furthermore, we measure the adherence to community structures of each model by calculating the modularity for the inferred partitions using the formula defined in eq. 16. For each calculated metric and network we point out the average and maximum score achieved by that particular method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

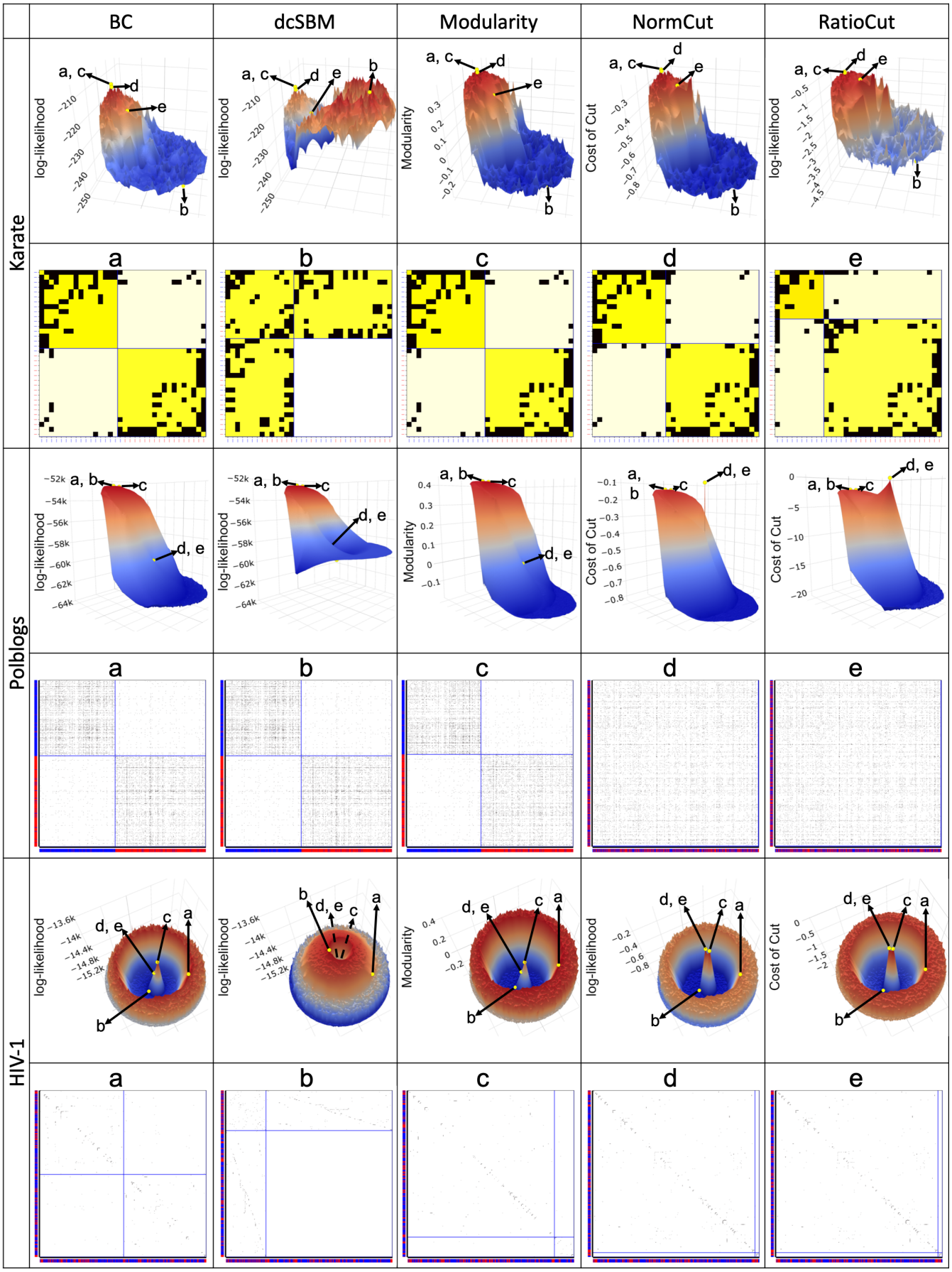

Fig. 7: Solution landscape comparison of BC, dc-SBM, Modularity, NormCut, RatioCut on the three networks. To explore the space, 100 samples from 15 chains were taken for the Bayesian methods, while for the spectral cuts 200 different solutions were generated for each method by randomly alternating 1% of the links within the networks. The costs of Normcut and Ratiocut are inverted to allow for direct landscape comparisons.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

We observe here that the BC achieves superior or on par performance on all three networks. In these results we again observe that the BC differs substantially from the dc-SBM, which is explained by the underlying supported configurations of the model likelihood $P(\boldsymbol{z}|G)$ shown in figure 5. In figure 7 we explore the solution space also of the ratio-cut (RC), normalized cut (NC) and modularity (Q) and how these solutions are supported by their corresponding objective functions.

We notice that the solutions supported (and thus the inference landscape) by the proposed BC is more in agreement with these existing community detection/graph partitioning procedures than the dc-SBM. However, we also observe notable differences of the proposed Bayesian Cut (BC) and these alternative partitioning procedures. In particular, neither RC nor NC provide as balanced solutions as the BC and they provide higher support for solutions further away from the underlying community structure. Here we pay particular attention to the cuts proposed for the Polblogs and HIV-1 network as these appear unsubstantiated due to the fact that they exclude a very small group of persons from the overall population.

For Polblogs both RC and NC exhibit very extreme and local optima in their solution landscape, which lead to a cut that excludes 4 persons from the other 1218 persons in both cases. In the case of RC, defined in eq 17, the dominance of the cost by the flow, i.e. links between two groups, is obvious. Since these two group are only connected by 1 inter-link the cost for performing this is cut is extremely low. In contrast, the true cut along party lines leads to one group with 662 nodes and one with 560, which share 1217 inter-links. To obtain lower costs, no-more than 76 inter-links would be allowed.

One way of alleviating this strong influence of the cut flow is to use NC, defined in eq. 18, which does not divide the cut flow by the number of nodes, but according to the degree of the cluster. However, even though this subtle difference changes the solution landscape in non-community supporting regions as shown in figure 7, the preference for cuts that separate unbalanced groups having very low flow remains. In this case the extreme cut leaves the small group with a degree of 5 and the bigger group with a degree of 16710, which results in very low costs. The above mentioned cut of our model results in a degree of 9464 and 8467 for the group with 662 nodes and 560 nodes respectively. In this case 894 inter-links would already give lower costs for our cut, which shows the improvement over RC, but still is not sufficient.

The highest congruence can be found between the BC and the Modularity method, confirming the community detection support of our proposed BC. Here we observe that both methods exhibit almost identical solution landscapes, which is reflected in the identical or very similar solution landscapes obtained by the methods. Interestingly, for the HIV-1 network the BC obtains a solution with a significantly higher modularity than the spectral modularity method itself identifies. In addition, this solution seems to be more balanced, since it achieves almost equally sized groups, while the proposed solution of the spectral modularity method partitions the network into a small and a large group. This highlights that BC strives for balanced modular structures.

4.3 Image Segmentation

In following we present results of image segmentation suitable for foreground-background or scene recognition. We compare BC model to NC and dc-SBM (with shared density of links out η_{out} in case of more than two segments $C > 2$).

NC implementations are often in practice combined with specific similarity matrices and we make use of the following two widely used procedures:

Mean color RAG: Mean color Regional Adjacency Graph is used to compute similarity matrices on super pixel graphs (RAG) [26] that serves as an input for NC in popular python package for image processing skimage https://scikit-image.org/docs/dev/api/skimage.future.graph.html. To compare with the BC method "cameraman" image and "coffee"available in the skimage package was used.

Fast Marching Method (FMM): This method (a.k.a. geodesical distance) is besides many used in the Graclus software presented in [25]. We used the MATLAB implementation of Jianbo Shi from https://www.cis.upenn.edu/~jshi/software/ to generate the FMM similarity matrix. Graclus software optimizes normcut objective in a hierarchical manner [25] with results presented at https://www.cis.upenn.edu/~jshi/software/demo2.html. For comparison purposes we use the public image of "baby" from the same site.

These methods produce similarity matrices $\mathbf{S}$ with elements in $[0; 1]$. To convert them into graphs with countable links required by the BC model we follow similar procedure as aforementioned Graclus software [25]. Graclus runs $\mathbf{A} = \lceil 100 * \mathbf{S} \rceil$ while BC implements $\mathbf{A} = \lfloor 100 * \mathbf{S} \rfloor$, both element wise.

Results of the BC model applied on images of "cameraman" and "bears" using RAG can be found in figure (8), "coffee" is presented in Appendix (13) and the results on "baby" using FMM can be found in figure (9). Notably BC model was applied on similarity matrices produced by respective implementations of RAG and FMM described above without further adjustments. In case of RAG and "bears" we adjust sigma for the Gaussian similarity kernel [1] in case of "cameraman" we leave it on default setting. "Cameraman" and "bears" experiments with Mean Color RAG have been ran with no degree correction (corresponds to hyper parameter γ set extremely large 10^7).

In all applications mentioned BC performs on par or superior to the compared methods (not necessarily state of the art though). In two partitions version considered for the "cameraman" the BC method separates objects from sky. In case of the four partition scenario used on "bears" BC recognizes foreground objects (cub and surrounding), background and adult bear while the other methods only partially succeed. For the "coffee cup" in appendix the BC

1. future.graph.rag_mean_color(img, labels1, mode='similarity', sigma=70**2, segmentation.slic(img, compactness=0.3, n_segments=100)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2994396, IEEE Transactions on Pattern Analysis and Machine Intelligence

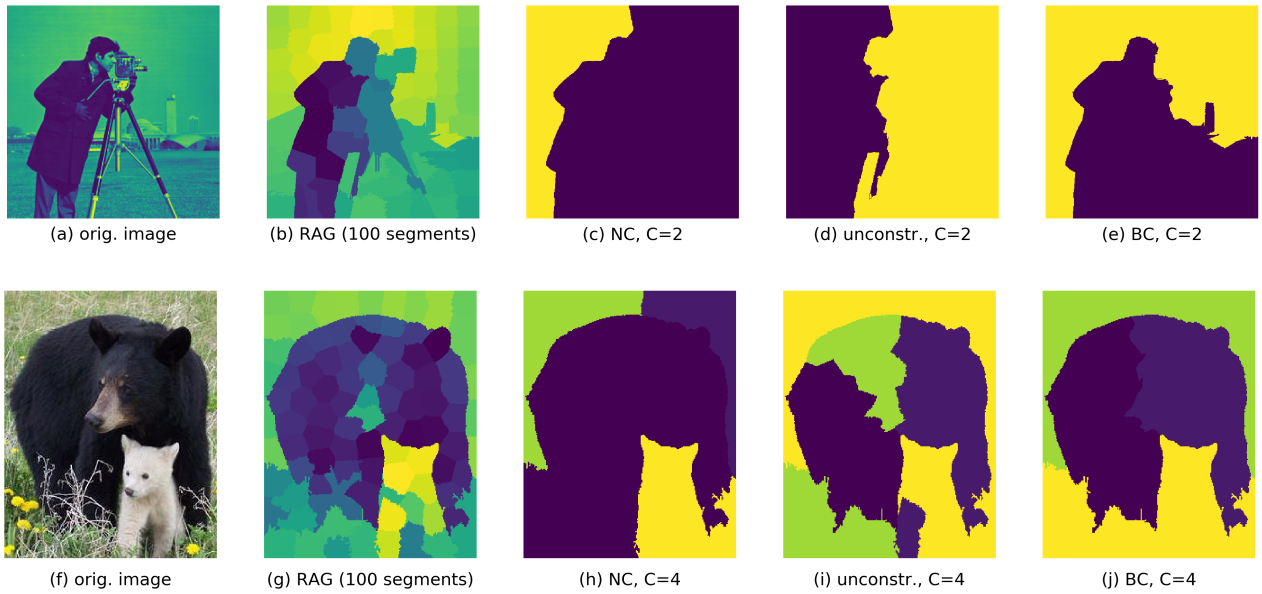

Fig. 8: **Top panel, Cameraman:** Resulting cut of BC model for C=2 (e) segments compared with unconstrained dc-SBM with shared η_{out} as well as spectral Norm cut. Fig (b) shows for reference RAG super pixel graph that is used to compute similarity matrix. Results demonstrate on par or better results of BC against referenced methods. Also it shows effect of constraint included in the model (we emphasized the effect of constraint by setting b=10 corresponding to 10x times higher within segment links density compared to links density among segments): (e) vs (d) a constraint model improves the results. Resulting cuts were obtained from 50 MCMC chains, 1000 samples each.
Bottom panel, Bears: Resulting cut of BC model for C=4 segments (j) compared with unconstrained dc-SBM with shared η_{out} (i) as well as spectral Norm cut (h) applied on mean color RAG similarity matrix. Similar to previous results figures BC demonstrates on par or better results against referenced methods. Resulting cuts were obtained from 50 MCMC chains, 1000 samples each with hyperparameters set on $b = 10^3$ and without degree correction

Fig. 9: Bayesian cut (BC) applied on similarity matrix obtained by fast marching method implemented by Jianbo Shi from https://www.cis.upenn.edu/j̃shi/software/. Resulting cut is sampled MAP obtained from 20 MCMC chains, 1000 samples each. (a) original, (b) $C = 2, b = 10$, with degree correction hyper parameter γ being inferred. Maximum of its posterior obtained at: $\gamma_{MAP} = 4.07$, (c) $C = 2, b = 10, \gamma = 0.0001$.

removes more of the background than the unconstrained dc-SBM and captures more of the coffee cup object than NC. In case of the "baby" image see figure 9 the effect of degree correction parameter γ controlling "greediness" of clusters is showed. In the more greedy settings, (c) as opposed to gamma being inferred in option (b), fixing it to "greedy" mode recognizes focal object's boundary more complete yet produces artifacts.

In summary, the presented image segmentation results by BC are on-par or superior to NC and the unconstrained version. However, we noted during experiments that the multiple MCMC runs produced slightly different cuts confirming that the inference is prone to sub optimal solutions and multiple restarts are therefore recommended.

5 Conclusion

We have proposed the Bayesian Cut (BC) advancing the degree-corrected stochastic block-model (dc-SBM) to explicitly account for community structure. In contrast to the dc-SBM only one parameter specified inter-group connectivity strength (η_{out}), however, in contrast to the generalized modularity as conforming to an l-partition model with shared link density across communities the proposed BC include more flexible community specific link-densities. We derived a fully Bayesian procedure and demonstrated that the imposed community constraints are analytically tractable even for large graphs by deriving a novel general solution to integrals involving multiple incomplete gamma functions. We expect the presented small and large scale solutions to the integral will have applications beyond community detection in social networks and image segmentation considered in this paper. For instance, for collapsed inference in the performance analysis of cognitive radio networks [42]. We observed that the constraints had significant impact on the inference providing more reliable results in compliance with ground truth for network exhibiting community structure

and it was also empirically confirmed that the constraint had merits for image segmentation in computer vision. We also observed that strictly enforcing community structure enabled to identify configurations where traditional block-modeling would identify bipartite structure. Notably, our Bayesian Cut provides favorable partitions when compared to traditional graph cutting procedures such as the ratio and normalized cut. In particular, we empirically observed that our BC procedure has meritorious properties balancing the partitions more favorable than these existing graph partitioning procedures. We have also derived fast large scale multiple cluster solution that presents generic tool for Bayesian inference.

We presently considered a uniform prior on the partition $P(\boldsymbol{z}) = C^{-n}$ to highlight the influence of the specification of the likelihood $p(G|\boldsymbol{z})$ in identifying partitions. However, we note that within the Bayesian modeling framework other (non-uniform) priors could be applied including the Pólya-urn (i.e., marginalized Dirichlet-Categorical) representation and its infinite limit given by the non-parametric Chinese restaurant process (CRP) also used in stochastic block-modeling [43].

Overall this work presents generic graph based clustering method that can be applied on wide range of similarity matrices. For illustrative purposes we presently applied our BC approach in the context of identifying communities in social networks and image segmentation, however, the approach extends to the many applications in which graph cuts are used. Flexibility with regards to similarity matrix allows for possible applications in areas such as scene reconstruction from large set of community photos [6], where the image set is partitioned into groups of related images, based on the visual structure represented in the image connectivity graph for the collection. Connectivity graph and corresponding similarity matrix is based on scale invariant feature transform, SIFT [44], that extracts image representative features that are used to find matches and define similarity between each image pair. Another possible area of application is Video summarization and scene detection [7], where similarity used for graph partitioning are based on color similarity and temporal frame distance.

In the outlook, although MCMC sampling are suitable for network structure inference, in order to find optimal cuts, future work should investigate alternatives while keeping the properties of the proposed framework. Further concerning image segmentation, this work made use of two popular similarities, Fast Marching Method and Mean Color, that rather relate pixels based on color intensities as opposed to spatial features. As suggested above we leave as future work to explore possibilities of BC applied on other existing or new similarities as well as extension of hereby presented bayesian generative hierarchical BC model to allow for contextual spatial or other features [34].

References

[1] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[2] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[4] B. Peng, L. Zhang, and D. Zhang, "A survey of graph theoretical approaches to image segmentation," *Pattern Recognition*, vol. 46, no. 3, pp. 1020–1038, 2013.

[5] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1356–1363.

[6] N. Snavely, I. Simon, M. Goesele, R. Szeliski, and S. M. Seitz, "Scene reconstruction and visualization from community photo collections," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1370–1390, 2010.

[7] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on circuits and systems for video technology*, vol. 15, no. 2, pp. 296–305, 2005.

[8] M. Witman, S. Ling, P. Boyd, S. Barthel, M. Haranczyk, B. Slater, and B. Smit, "Cutting materials in half: A graph theory approach for generating crystal surfaces and its prediction of 2d zeolites," *ACS central science*, vol. 4, no. 2, pp. 235–245, 2018.

[9] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[10] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 19–26.

[11] J. Wang, T. Jebara, and S.-F. Chang, "Semi-supervised learning using greedy max-cut," *Journal of Machine Learning Research*, vol. 14, no. Mar, pp. 771–800, 2013.

[12] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 11, no. 9, pp. 1074–1085, 1992.

[13] V. Kolmogorov and R. Zabih, "What energy functions can be minimizedvia graph cuts?" *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 2, pp. 147–159, 2004.

[14] V. Kolmogorov and C. Rother, "Minimizing nonsubmodular functions with graph cuts-a review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 7, pp. 1274–1279, 2007.

[15] D. J. Foster, D. Reichman, and K. Sridharan, "Inference in sparse graphs with pairwise measurements and side information," *arXiv preprint arXiv:1703.02728*, 2017.

[16] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[17] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American statistical association*, vol. 96, no. 455, pp. 1077–1087, 2001.

[18] M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.

[19] M. Mørup and M. N. Schmidt, "Bayesian community detection," *Neural computation*, vol. 24, no. 9, pp. 2434–2456, 2012.

[20] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016107, 2011.

[21] T. Herlau, M. N. Schmidt, and M. Mørup, "Infinite-degree-corrected stochastic block model," *Physical review E*, vol. 90, no. 3, p. 032819, 2014.

[22] M. E. Newman, "Equivalence between modularity optimization and maximum likelihood methods for community detection," *Physical Review E*, vol. 94, no. 5, p. 052315, 2016.

[23] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures & Algorithms*, vol. 18, no. 2, pp. 116–140, 2001.

[24] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[25] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 11, pp. 1944–1957, 2007.

[26] A. Trémeau and P. Colantoni, "Regions adjacency graph applied to color image segmentation," *IEEE Transactions on image processing*, vol. 9, no. 4, pp. 735–744, 2000.

[27] Y. Zhang, Z. Ghahramani, A. J. Storkey, and C. A. Sutton, "Continuous relaxations for discrete hamiltonian monte carlo," in *Advances in Neural Information Processing Systems*, 2012, pp. 3194–3202.

[28] R. AlAhmad, "Products of incomplete gamma functions," *Analysis*, vol. 36, no. 3, pp. 199–203, 2016.

[29] G. Jameson, "The incomplete gamma functions," *The Mathematical Gazette*, vol. 100, no. 548, pp. 298–306, 2016.

[30] M. Hartmann, "Extending owen's integral table and a new multivariate bernoulli distribution," *arXiv preprint arXiv:1704.04736*, 2017.

[31] M. Sibuya, I. Yoshimura, and R. Shimizu, "Negative multinomial distribution," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 409–426, Dec 1964. [Online]. Available: https://doi.org/10.1007/BF02868583

[32] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, "Network flows," 1988.

[33] R. G. Parker and R. L. Rardin, *Discrete optimization*. Elsevier, 2014.

[34] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[35] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," *Science advances*, vol. 3, no. 5, p. e1602548, 2017.

[36] M. Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*. Springer, 2003, pp. 173–187.

[37] I. Borg and P. Groenen, "Modern multidimensional scaling: theory and applications," *Journal of Educational Measurement*, vol. 40, no. 3, pp. 277–280, 2003.

[38] S. W. Park, L. Linsen, O. Kreylos, J. D. Owens, and B. H. Hamann, "Discrete sibson interpolation," 2006.

[39] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[40] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[41] M. Morris and R. Rothenberg, "Hiv transmission network metastudy project: An archive of data from eight network studies, 1988–2001," 2011.

[42] B. Van Nguyen, H. Jung, D. Har, and K. Kim, "Performance analysis of a cognitive radio network with an energy harvesting secondary transmitter under nakagami-m fading," *IEEE Access*, vol. 6, pp. 4135–4144, 2018.

[43] M. N. Schmidt and M. Morup, "Nonparametric bayesian modeling of complex networks: An introduction," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 110–128, 2013.

[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[45] G. Jameson, "A simple proof of stirling's formula for the gamma function," *The Mathematical Gazette*, vol. 99, no. 544, pp. 68–74, 2015.

Petr Taborsky received his M.Sc. degree in Mathematics, Mathematical Statistics and Probability at Charles University, Prague. Currently he's PhD understudy at the Section for Cognitive Systems at DTU Compute, Technical University of Denmark. He's also heading Digital Labs_ at Telenor Danmark and his research interests include federated machine learning, neural networks, and complex network modeling.

Laurent Vermue received his M.Sc. degree in Industrial Engineering and Management at the Technical University of Berlin and MMSc. degree in Management Science and Engineering at the Tongji University. Currently he is a Ph.D. student at the Section for Statistics and Data Analysis and the Section for Cognitive Systems at DTU Compute, Technical University of Denmark. His research interests include machine learning, complex network modeling and open research software.

Maciej Korzepa received his M.Sc. degree in Digital Media Engineering at the Technical University of Denmark. He is currently a Ph.D. student at the Section for Cognitive Systems at DTU Compute, Technical University of Denmark. His research interests include machine learning, intelligent interfaces, and complex network modeling.

Morten Mørup received his M.S. and Ph.D. degrees in applied mathematics at the Technical University of Denmark and he is currently Professor at the Section for Cognitive Systems at DTU Compute, Technical University of Denmark. He has been associate editor of IEEE Transactions on Signal Processing and his research interests include machine learning, neuroimaging, and complex network modeling.