Hybrid Mixed-Membership Blockmodel for Inference on Realistic Network Interactions

Edward K. Kao, *Member, IEEE*, Steven Thomas Smith, *Fellow, IEEE*, and Edoardo M. Airoldi

Abstract—This work proposes a novel hybrid mixed-membership blockmodel (HMMB) that integrates three canonical network models to capture the characteristics of real-world interactions: community structure with mixed-membership, power-law-distributed node degrees, and sparsity. This hybrid model provides the capacity needed for realism, enabling control and inference on individual attributes of interest such as mixed-membership and popularity. A rigorous inference procedure is developed for estimating the parameters of this model through iterative Bayesian updates, with targeted initialization to improve identifiability. For the estimation of mixed-membership parameters, the Cramér-Rao bound is derived by quantifying the information content in terms of the Fisher information matrix. The effectiveness of the proposed inference is demonstrated in simulations where the estimates achieve covariances close to the Cramér-Rao bound while maintaining good truth coverage. We illustrate the utility of the proposed model and inference procedure in the application of detecting a community from a few cue nodes, where success depends on accurately estimating the mixed-memberships. Performance evaluations on both simulated and real-world data show that inference with HMMB is able to recover mixed-memberships in the presence of challenging community overlap, leading to significantly improved detection performance over algorithms based on network modularity and simpler models.

Index Terms—Social networks, network models, network inference, overlapping communities, mixture models, bayesian methods, MCMC, cramér-rao bound, fisher information, identifiability, targeted initialization, seeded community detection, graph algorithms

1 Introduction

Network modeling and inference provide structural and formation insights as well as information on the key attributes of the population and individuals. Such information, often not directly observable, is valuable in many real-world applications. One example of such information is the individuals' community membership. Much work has been done both in network modeling and inference and in community detection. This section begins with a brief survey of previous models closely related to ours and representative works in community detection as an application area for network modeling and inference. It ends with a description of the overall structure and contributions of this paper, which proposes a novel hybrid model to capture the key characteristics of real-world networks, especially in the presence of overlapping communities.

1.1 Background

The three canonical network models that compose the novel hybrid model in this paper are the Erdős-Rényi model, the Chung-Lu model, and the mixed-membership stochastic blockmodel (MMB). Erdős-Rényi model the network with a sparsity parameter that indicates the probability of any edge to be present [1]. This simple model led to elegant mathematical properties such as the graph percolation theories, but does not express real-world network characteristics such as node degree variation and social structures. Chung, Lu, and Aiello model the node degrees variation using a power-law distribution, commonly known as the Chung-Lu model [2]. This model captures the realism of the small-world property with a few nodes having much higher degrees and the existence of a short path between any two nodes. Wang and Wong explain the social structure through group membership, using stochastic blockmodels where the blocks represent the communities through which the nodes interact based on their membership [3]. Airoldi et al. extend this model to allow mixed-membership for the nodes, known as the mixed-membership stochastic blockmodels [4]. The mixed-membership begins to capture the realism of overlapping communities which has received more attention in recent years.

Some recent network models have also attempted to capture the real-world network characteristics addressed in this paper. For example, both the works by Soufiani and Airoldi [5] and Yang and Leskovec [6] explicitly model the community overlap by allowing each node to belong to multiple communities. The model in Soufiani and Airoldi has the elegant interpretation of a multiple-scale decomposition referred to as the graphlet decomposition. Karrer and Newman [7] incorporate the degree correction terms to the blockmodel to capture both the community structure and the varying node degrees. Peixoto [8] builds upon Karrer and Newman's model and extends it to include multiple community memberships for each node. Peixoto's model is

- *E.K. Kao and S.T. Smith are with MIT Lincoln Laboratory, Lexington, MA 02421-6426. E-mail: {edward.kao, stsmith}@ll.mit.edu.*
- *E.M. Airoldi is with Harvard Statistics Department, Cambridge, MA 02138-2901. E-mail: airoldi@fas.harvard.edu.*

Manuscript received 6 Sept. 2017; revised 18 Jan. 2018; accepted 2 Mar. 2018. Date of publication 4 Apr. 2018; date of current version 11 Sept. 2019. (Corresponding author: Edward K. Kao.)
Recommended for acceptance by H. H. C. Iu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TNSE.2018.2823324

2327-4697 © 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

perhaps the closest to ours, so it will be included as a baseline performance comparison for community detection in Section 4. None of these existing models capture all three of the real-world network characteristics in the proposed hybrid model: community-based interactions with mixed-membership, varying node degree, and sparsity. Lastly, this proposed hybrid model is related to broader graph model classes with high dimensional parameter spaces [9], [10].

Community detection, with much existing work [11], is a natural application for network modeling and inference. A well-known algorithm is the spectral method by Newman [12] where detection is done through the eigenspectrum of the modularity matrix. Most existing algorithms maximize modularity by partitioning the network into communities (i.e. modules) that have much stronger interactions within than between them [13]. Smith et al. exploit network modularity through random walks on graph, for cued network detection where a few nodes in the community of interest is known a priori [14]. Some recent works detect communities through membership estimates in a generative model. Huang et al. propose a fast tensor method to the mixed-membership stochastic blockmodel [15], and Ball and Newman use the degree corrected stochastic blockmodel (DCSB) [16]. Peixoto's extension on the DCSB with multi-community memberships [8] can be used for community detection in the presence of overlapping communities. Section 4 selects from these existing approaches as baseline comparisons against our proposed model in performing community detection.

1.2 Contributions

This paper proposes in Section 2 a novel hybrid mixed-membership blockmodel (HMMB) that integrates three canonical network models to capture the characteristics of real-world networks: community-based interactions with mixed-membership, varying node degree, and sparsity. The hybrid model provides the richness for realism and enables inference on useful latent attributes such as mixed-membership and popularity. A rigorous inference procedure is developed in Section 3 for estimating the model parameters through iterative Bayesian updates with targeted initialization to improve identifiability. For the mixed-membership estimate, the Cramér-Rao bound is derived from quantifying the information content of the network data through Fisher information analysis. The effectiveness of the proposed inference is demonstrated in simulations where the estimated posterior intervals achieve covariances close to the Cramér-Rao bound while maintaining good truth coverage. As an application demonstration, the mixed-membership estimate is used for community detection in Section 4. With community overlap in both simulated and real-world graphs, the HMMB-based approach is shown to significantly outperform existing approaches based on network modularity and simpler models.

2 Hybrid Mixed-Membership Blockmodel (HMMB)

For a growing body of real-world network interactions data, a desirable statistical model should be rich and realistic enough to capture its fundamental characteristics. Examples of network interactions span a wide range of data types: in emails [17], social encounters [18], trips between locations [19], social network messages [20], tweets [21], relationship between accounts [22], wireless communications [23], co-authorship [24], co-sponsorship of legislation [25], and interactions between molecules and proteins [26]. The statistical model should be interpretable and intuitive so that estimating the model parameter will provide useful information on the individuals and the nature of their interactions. These are the key motivations for the design of the hybrid mixed-membership blockmodel, which simultaneously captures the characteristics of these real-world interaction networks: community-based interactions with mixed-membership, varying node degree, and sparsity. Section 4.3 demonstrates on DBLP co-authorship network the advantage of the HMMB over previous models which do not fully account for these characteristics.

2.1 Model Description

Let $G = (\mathscr{V}, \mathscr{E})$ denote a weighted and directed network consisting of $N = |\mathscr{V}|$ nodes. The edge set $\mathscr{E}$ is represented as an adjacency matrix $\boldsymbol{A}$ where each element (i.e. edge weight) a_{ij} is the number of interactions from node i to node j. Modeling the interactions as Poisson point processes over time, similar to Karrer and Newman's model [7], each weighted edge is drawn from a Poisson distribution with rate λ_{ij},

$$a_{ij} \sim \text{Poisson}(\lambda_{ij} T), \tag{1}$$

where the duration T is a time constant in units appropriate for the given application, and '$\sim$' denotes "is distributed as." For example, when modeling email interactions, T may be measured in hours. The Poisson interactions has the elegant interpretation as the accumulative count of Bernoulli events in infinitesimally small time intervals. This interpretation provides a connection back to the classical random graph models with Bernoulli edges. The HMMB models properties of the real-world networks, by combining three canonical network models:

- Power-law degree distribution and small world network: Chung-Lu model [2], [27]
- Overlapping communities: mixed-membership stochastic blockmodel [4], [28]
- Sparsity: Erdős-Rényi model [1]

This is accomplished by modeling the rate of interaction from each node i to j as

$$\lambda_{ij} = (\lambda_i \lambda_j) \times (\boldsymbol{\pi}_i^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_j) \times I_{ij}, \tag{2}$$

where the Chung-Lu term $\lambda_i \lambda_j$ captures variation in node degrees (i.e. activity level), the mixed-membership stochastic blockmodel term $\boldsymbol{\pi}_i^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_j$ captures community structure through nodal mixed-membership $\boldsymbol{\pi}$ and the community interaction block matrix $\boldsymbol{B}$, and the Erdős-Rényi indicator I_{ij} turns edges "on" and "off." The entire N-by-N frequency matrix $\boldsymbol{\Lambda}$ can be expressed in matrix form:

$$\boldsymbol{\Lambda} = (\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}) \circ (\boldsymbol{\Pi}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\Pi}) \circ \boldsymbol{I}, \tag{3}$$

where '$\circ$' denotes the Schur-Hadamard product (element-wise product). The Chung-Lu term $\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}$ allows some nodes to be much more active than the others. The frequency of

interaction between i and j is proportional to the activity level of both nodes. This well-known model is parameterized by $\boldsymbol{\lambda}$, a vector of N components and a hyperparameter α for how quickly the prior power-law distribution for each λ_i drops off: $p_0(\lambda_i) \propto \lambda_i^{\alpha}$. Empirical evidence from real-world networks show that α is typically between -3 and -2 [29]. For networks with minimum and maximum node degrees, a truncated power-law may be used. The mixed-membership stochastic blockmodel term, $\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Pi}$, captures the overlapping community structure. $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_N)$ is a K-by-N matrix representing each node's mixed-membership to the K communities. Each $\boldsymbol{\pi}_i$ sums to one (i.e. $\mathbb{1}^{\mathrm{T}}\boldsymbol{\Pi} = \mathbb{1}^{\mathrm{T}}$). The model parameter $\boldsymbol{B}$ is a K-by-K block matrix where each element b_{mn} is the strength of interaction between community m to n. Consequently, $\boldsymbol{\pi}_i^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\pi}_j$, captures the total strength of interaction from node i to node j based on their mixed-memberships. Lastly, the sparsity indicator matrix $\boldsymbol{I}$ acts as "on" and "off" switches for each edge drawn from a Bernoulli distribution with sparsity parameter s,

$$I_{ij} \overset{\text{iid}}{\sim} \text{Bern}(s). \tag{4}$$

In some applications, sparsity varies with community memberships. The HMMB can be generalized to this case by expanding sparsity to a K-by-K matrix, $\boldsymbol{S}$, so that the probability of edge switch I_{ij} to be "on" is $\boldsymbol{\pi}_i^{\mathrm{T}}\boldsymbol{S}\boldsymbol{\pi}_j$.

The mixed-memberships of each node are drawn from prior distributions that represent the lifestyles of the individuals. L lifestyles are defined, each with K pseudocounts indicating the lifestyle's tendency to participate in each community membership, represented by hyperparameter $\boldsymbol{X}$, a K-by-L mixed-membership pseudocount matrix. The lifestyle indicator vector for each node i, $\boldsymbol{l}_i$, is drawn from the $\text{Multinom}(1, \boldsymbol{\phi})$ prior distribution where hyperparameter $\boldsymbol{\phi}$ is a simplex vector of L components indicating the probability of belonging to each lifestyle. Finally, the mixed-membership for each node i, $\boldsymbol{\pi}_i$ is drawn from the $\text{Dirichlet}(\boldsymbol{X}\boldsymbol{l}_i)$ prior distribution where $\boldsymbol{X}\boldsymbol{l}_i$ is the membership pseudocounts of the lifestyle i belongs to. This generative procedure for the mixed-membership is useful for simulation, but in real-world data sets, information on the lifestyles is typically unavailable. Fig. 1 presents a plate diagram [30] on the generation process under HMMB.

Interactions between nodes in real-world networks often increase or decrease due to changes both at the individual and the network level. The HMMB does not explicitly model such dynamics. However, existing works have included dynamics for the components of the HMMB (e.g. dynamic mixed-membership blockmodel [31]), which point to ways that HMMB may be extended with dynamics in future work.

An efficient and scalable Python implementation of the HMMB network generator is provided in Table 1, with parameters chosen to generate a network of three overlapping communities. A single realization is illustrated using the graph-tool package [32] in Fig. 2.

2.2 Parameter Identifiability

The HMMB is a composite of three identifiable network models; inference will be performed on one identifiable model at a time conditional on the parameters from the other two. This is detailed in Section 3.1. The Erdős-Rényi

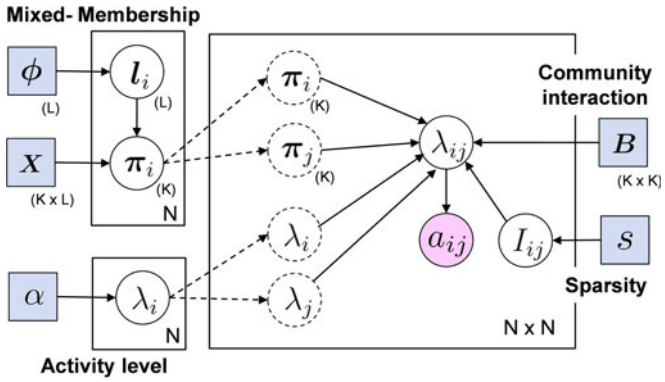

Fig. 1. Plate diagram of the full generative process of the HMMB, with K communities, N nodes (individuals), and L lifestyles. The blue squares indicate the model parameter and prior hyperparameters that are pre-specified in a simulation. White circles are the latent model parameters drawn from the prior distributions. The red circle is the resulting network drawn from the Poisson distribution according to the edge rates. The dotted circles indicate the nodal parameters used to draw each edge. The solid boxes specify the number of copies drawn for each parameter.

model has only the sparsity parameter s, and each edge switch I_{ij} is inferred independently; therefore, it is identifiable [33]. The Chung-Lu node degree parameter $\boldsymbol{\lambda}$ is shown to be identifiable by Perry and Wolfe [34]. The mixed-membership blockmodel is known to be identifiable, but belong to the general class of mixture models known to have symmetric multi-modal likelihood [35]. This multi-modality may present challenges in finding the true mode; however, this issue is mitigated by inference strategies such as targeted initialization and multiple starting points [28], [36], [37].

Joint model identifiability of the HMMB is not guaranteed by identifiability of its individual component models. Considering the joint model of Equation (3), identifiability issue arises when there exists a set of invariant parameter mappings across the three component model terms that result in the same $\boldsymbol{\Lambda}$, and therefore the same likelihood. Specifically, potential invariant mappings across the Chung-Lu and the blockmodel terms, $(\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}) \circ (\boldsymbol{\Pi}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{\Pi})$, must be examined. The Erdős-Rényi term $\boldsymbol{I}$ is a binary matrix of edge switches; therefore, invariant mappings on this term do not exist.

First, consider a positive scalar mapping on the Chung-Lu parameter, $\boldsymbol{\lambda} \mapsto w\boldsymbol{\lambda}$. The positive scalar multiplier w cannot be absorbed by the mixed-membership matrix $\boldsymbol{\Pi}$ because of its simplex constraint: $\mathbb{1}^{\mathrm{T}}\boldsymbol{\Pi} = \mathbb{1}^{\mathrm{T}}$, where $\mathbb{1}$ is a vector of ones. However, it can be easily absorbed by the block matrix with mapping, $\boldsymbol{B} \mapsto w^{-2}\boldsymbol{B}$. For special cases where there exists a subset of communities $\mathscr{K}$ that do not overlap in nodal memberships with the rest of the communities, equivalent invariant mappings exist on the subspace: $\boldsymbol{\lambda}_{\mathscr{V}_{\mathscr{K}}} \mapsto w\boldsymbol{\lambda}_{\mathscr{V}_{\mathscr{K}}}$ and $\boldsymbol{B}_{\mathscr{K}\mathscr{K}} \mapsto w^{-2}\boldsymbol{B}_{\mathscr{K}\mathscr{K}}$, where $\mathscr{V}_{\mathscr{K}}$ is the set of nodes with memberships inside $\mathscr{K}$.

More generally, consider any linear transform mapping on the Chung-Lu parameter, $\boldsymbol{\lambda} \mapsto \boldsymbol{W}\boldsymbol{\lambda}$, where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)^{\mathrm{T}}$ is an N-by-N invertible matrix. Each ijth term in the Chung-Lu term $(\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}})$ becomes $(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\lambda})(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\lambda})$, a quadratic function of the original $\boldsymbol{\lambda}$ with all cross terms $w_{im}w_{jn}\lambda_m\lambda_n$. Invariance requires

$$\left(\sum_{m=1}^{N} w_{im}\lambda_m\right)\left(\sum_{n=1}^{N} w_{jn}\lambda_n\right) = \lambda_i\lambda_j \qquad \text{for all } i, j, \lambda_i, \lambda_j. \tag{5}$$

TABLE 1
HMMB Generator Example Python Code

```
# 3 community network example with parameters
import numpy as np, numpy.random as npr, \
       random as rand, scipy.sparse as sps
N = 256
alpha = -2.5
powerlaw_min = 1.0 # min of the truncated power-law
phi = np.array([1, 1, 1, 0.1]); phi /= phi.sum()
X = np.array([[10, 1, 0, 1],   # last column is a
              [1, 10, 1, 1],   # lifestyle with highly
              [0, 1, 10, 1]])  # mixed membership
B = np.array([[10, 1, 0],
             [2, 10, 1],
             [0, 1, 10]])
s = 1.5*np.log(N)/N
T = 10
def dirichlet_withzeros(a):
    x = np.zeros_like(a).astype(np.float)
    x[a.nonzero()] = npr.dirichlet(a[a.nonzero()])
    return x
def create_sparse(fn): # sparse matrix from fn(i,j)
    return sps.csr_matrix(([fn(i,j)
        for i,j in zip(rows,cols)],(rows,cols)))

# Generate the network
L = npr.multinomial(1,phi,N).transpose()
Pi = np.apply_along_axis(dirichlet_withzeros,0,X.dot(L))
lmbda = (npr.pareto(-alpha-1,N) + 1) * powerlaw_min
num_edges = npr.binomial(N*N,s)
rows,cols = np.unravel_index(rand.sample(range(N*N),
    num_edges),(N,N)) # Erdos-Renyi
A = create_sparse(lambda i,j: npr.poisson(T*
    (lmbda[i]*lmbda[j]) # Chung-Lu * blockmodel
    *(Pi[:,i].transpose().dot(B).dot(Pi[:,j]))))

# Visualize the network
import graph_tool.all as gt
g = gt.Graph(directed=False)
eweight = g.new_edge_property("int")
g.add_edge_list([(i,j,A[i,j]) for i,j in np.transpose(
    (A>np.median(np.array(A[A.nonzero()]))).nonzero())],
    eprops=[eweight])
g.set_vertex_filter(gt.label_largest_component(g))
vertex_fill_color = g.new_vertex_property("string")
vertex_size = g.new_vertex_property("int")
for v in g.vertices():
    vertex_fill_color[v] = ("#"+"{:02x}"*3).format(
        *np.round(Pi[:,g.vertex_index[v]]*255).astype(int))
    vertex_size[v] = np.log1p(v.out_degree())*7
pos = gt.sfdp_layout(g,eweight=eweight)
gt.graph_draw(g,pos=pos,
    vertex_fill_color=vertex_fill_color,
    vertex_size=vertex_size)
```

The only linear transformation that meets this requirement is a scalar product of the identity matrix, $W = wI$, and nonlinear transformations are not possible because the right-hand side is homogeneous of degree 2, returning back to the scalar case addressed above. Having considered all potential invariant mappings across the HMMB component model terms, the HMMB identifiability can be summarized in the following theorem:

Theorem 1 (Identifiability of Hybrid Mixed-Membership Blockmodel). *The hybrid mixed-membership blockmodel is identifiable except for the arbitrary scalar multiplier w between the Chung-Lu parameters $\boldsymbol{\lambda}$ and the block matrix $\boldsymbol{B}$ in the invariant mappings*

$$\begin{aligned} \boldsymbol{\lambda} &\mapsto w\boldsymbol{\lambda}, \\ \boldsymbol{B} &\mapsto \frac{1}{w^2}\boldsymbol{B} \end{aligned} \tag{6}$$

that result in the same network interaction rates $\boldsymbol{\Lambda}$ and likelihood on the network data. For special cases where there exists a subset of communities $\mathscr{K}$ that do not overlap in nodal memberships with the rest of the communities,

$$\begin{aligned} \boldsymbol{\lambda}_{\mathscr{V}_{\mathscr{K}}} &\mapsto w\boldsymbol{\lambda}_{\mathscr{V}_{\mathscr{K}}}, \\ \boldsymbol{B}_{\mathscr{K}\mathscr{K}} &\mapsto \frac{1}{w^2}\boldsymbol{B}_{\mathscr{K}\mathscr{K}}, \end{aligned} \tag{7}$$

where $\mathscr{V}_{\mathscr{K}}$ is the set of nodes with memberships inside $\mathscr{K}$.

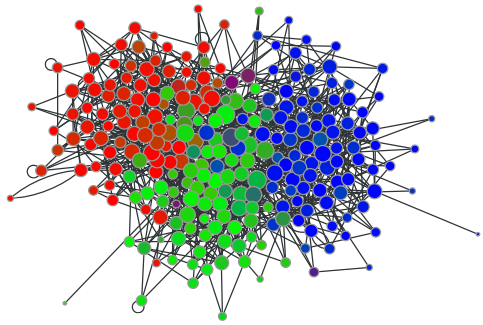

Fig. 2. HMMB example network with three overlapping communities. Node color is specified by assigning its fractional community memberships determined by the columns of the matrix Π to red-green-blue values. Node size is proportional to node degree. Weak edges and nodes with no edges are dropped for visual clarity.

The identifiability issues above are mitigated in the proposed inference procedure in Section 3.1, with specific strategies including the profile likelihoods, targeted parameter initializations, and multiple independent inference chains. Using the proposed inference procedure, identifiability is demonstrated empirically in Section 3.3 on networks of convenient sizes (hundreds of nodes). As the network increases in size, the number of parameters effectively grows at a rate of $O(N \log N)$ as the number of communities each node is in grows at a sub-linear rate of $O(\log N)$ and is eventually capped at a constant. The number of observed interactions grows typically at a rate of $O(N \log N)$ for realistic networks, to catch up with the number of parameters. Therefore, identifiability scales well with increasing network size.

3 Inference and Performance Bounds

Inferring the parameters under HMMB on an observed network provides valuable information that are not typically observable. For example, the mixed-memberships offer insights on the community structure and reveal members of each community. As an example, Section 4.1 demonstrates the application of community detection using the mixed-membership estimates. Individual activity levels, $\boldsymbol{\lambda}$, and the strength of interactions between and within communities, $\boldsymbol{B}$, also provide useful information. In applications involving social influence, node activity levels and community interaction are central to the analysis of the amount and structure of the influence [38]. Lastly, inferring the model parameters enables prediction of future interaction and generation of network data with similar characteristics.

This section describes the Bayesian joint inference procedure through iterative updates. Targeted initializations of the parameters are designed to improve identifiability. A fully Bayesian procedure is proposed, with an alternative maximum posterior update for some of the parameters for faster convergence at a cost on accuracy. With a special

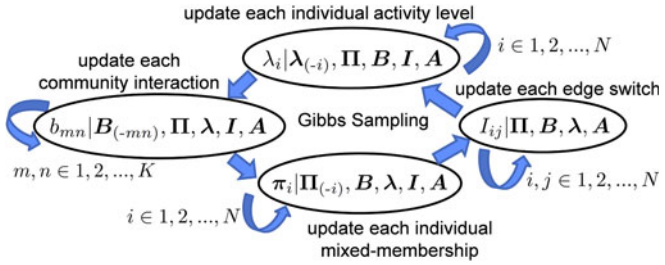

Fig. 3. Gibbs sampling procedure for iterative updates on the parameters one at a time, conditional on the other parameters. There are K communities and N nodes in the network. Specifics of each step are in Sections 3.1.2 and 3.1.3.

interest in the community membership, this section characterizes the information content of network data and the theoretical performance bound on the mixed-membership estimates. Lastly, the proposed estimation procedures are evaluated and compared across a range of reasonable model parameter settings, demonstrating empirically the identifiability of parameters in the HMMB.

3.1 Bayesian Parameter Estimation Procedure

Parameter estimation is performed through Bayesian methods [39] using iterative Monte Carlo updates. Given the observed network data, $\boldsymbol{A}$, the goal is to obtain the joint posterior distribution on the parameters of interest, namely the mixed-memberships, $\boldsymbol{\Pi}$, the activity level (node degree), $\boldsymbol{\lambda}$, of each node, and the community interaction structure, $\boldsymbol{B}$. The joint posterior captures both the likely values and the uncertainty of the estimates, which is a key advantage of Bayesian inference. In applications where decisions are made based on the inferred parameters, quantifying the uncertainty of the estimates is especially desirable. The edge switches $\boldsymbol{I}$ are latent variables in the posterior distribution. The joint posterior distribution is proportional to the likelihood times the prior,

$$p(\boldsymbol{\Pi}, \boldsymbol{\lambda}, \boldsymbol{B}, \boldsymbol{I} \mid \boldsymbol{A}) \propto \mathcal{L}(\boldsymbol{\Pi}, \boldsymbol{\lambda}, \boldsymbol{B}, \boldsymbol{I} \mid \boldsymbol{A}) p_0(\boldsymbol{\Pi}, \boldsymbol{\lambda}, \boldsymbol{B}, \boldsymbol{I}). \quad (8)$$

Note that the notation for the likelihood function throughout the paper is written with the model parameters as likelihood function argument, conditioned on the data. The joint likelihood function is a consequence of Equations (1), (2), and (4):

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Pi}, \boldsymbol{\lambda}, B, \boldsymbol{I}|\boldsymbol{A}) &= \prod_{i,j \in \mathbb{I}_1} \left(f(a_{ij}\,;\, \lambda_{ij}T)\, s\right) \prod_{i,j \in \mathbb{I}_0} \left(\delta_{a_{ij}0}\,(1-s)\right) \\ &\propto \exp\left(\sum_{i,j \in \mathbb{I}_1} \left(a_{ij} \log\left(\lambda_{ij}\right) - T\lambda_{ij}\right)\right) \prod i, j \in \mathbb{I}_0 \prod_{i,j \in \mathbb{I}_0} \delta_{a_{ij}0} \\ &\quad \times s^{|\mathbb{I}_1|}(1-s)^{|\mathbb{I}_0|},\end{aligned} \quad (9)$$

where $f(\,\cdot\,; \lambda_{ij}T)$ is the Poisson probability mass function with rate $\lambda_{ij}T$, $\mathbb{I}_1$ the set of "on" edges, $\mathbb{I}_0$ the set of "off" edges, and the index pair i, j refers to an edge. The $\delta_{a_{ij}0}$ term makes sure that only edges with no observed interaction can be "off." The joint posterior distribution is difficult and inefficient to sample directly. Therefore, the parameters will be iteratively updated one at a time, conditional on the other parameters. This is commonly known as Gibbs sampling for multivariate distributions [39]. Fig. 3 shows each step of the Gibbs update for HMMB. Fig. 4 summarizes the overall inference procedure. Two flavors of the Gibbs updates are developed. The first is purely based upon Markov Chain Monte Carlo (MCMC) by sampling from the conditional posterior distribution in each step. The second is a hybrid of sampling and maximizing the conditional posterior distribution, commonly known as the Monte Carlo Expectation Maximization (MCEM) [40], [41]. The MCMC approach has the theoretical property of converging to the joint posterior distribution, while the MCEM approach sacrifices this property for faster convergence with the greedy maximization steps. For both flavors, each parameter update step and the governing likelihood function, prior probability, and posterior probability are described in detail in Sections 3.1.2 and 3.1.3. A performance comparison between the MCMC and MCEM procedures is given later in Section 3.3 with discussion of results in Table 2.

Require: Network data in adjacency matrix $\boldsymbol{A}$
  for each independent inference chain **do**
    initialize $\boldsymbol{\Pi}^0$ using labels from spectral clustering
    initialize $\boldsymbol{\lambda}^0$ proportional to the observed node degrees
    initialize $\boldsymbol{B}^0$ to a roughly diagonal matrix
    initialize $\boldsymbol{I}^0$ to be "on" only for edges with interactions
    repeat {Gibbs sampling updates}
      for $i = 1$ to N **do**
        update λ_i^{t+1} conditional on $\boldsymbol{\lambda}_{(-i)}^{t+}, \boldsymbol{\Pi}^t, \boldsymbol{B}^t, \boldsymbol{I}^t$
      end for
      for $m, n = 1$ to K **do**
        update b_{mn}^{t+1} conditional on $\boldsymbol{B}_{(-mn)}^{t+}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{\Pi}^t, \boldsymbol{I}^t$
      end for
      for $i = 1$ to N **do**
        update $\boldsymbol{\pi}_i^{t+1}$ conditional on $\boldsymbol{\Pi}_{(-i)}^{t+}, \boldsymbol{B}^{t+1}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{I}^t$
      end for
      for $i, j = 1$ to N **do**
        update I_{ij}^{t+1} conditional on $\boldsymbol{\Pi}^{t+1}, \boldsymbol{B}^{t+1}, \boldsymbol{\lambda}^{t+1}$
      end for
      fix diagonal of $\boldsymbol{B}$ after some updates
      store $\boldsymbol{\Pi}^{t+1}, \boldsymbol{B}^{t+1}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{I}^{t+1}$ if $t + 1 >$ burn-in
    until convergence or maximum iteration reached
  end for
  return Posterior samples on $\boldsymbol{\Pi}, \boldsymbol{\lambda}, \boldsymbol{B}$, from the chain with the highest posterior probability

Fig. 4. HMMB Bayesian inference procedure using Gibbs sampling

The number of communities (K) remains fixed under each Gibbs run. As the true number of communities is typically unknown with real-world networks, a generous number is chosen to ensure the model has enough capacity to capture the community structure. Inference may be performed using a number of reasonable values for K. Allowing the number of communities to change during the inference will require proposal steps to jump back and forth between models with different parameter spaces. While allowing inference to find the optimal number of blocks is an attractive feature, designing effective proposals for it is a non-trivial task, but an interesting avenue of research.

Due to the potential identifiability challenges with multimodal likelihood of mixture models discussed in Section 2.2, several independent Monte Carlo chains are run to maximize the chance for finding the true mode. At the end of the run, the chain that finishes with the highest joint posterior probability is selected. This "best" chain has found the highest

mode on the complex posterior probability surface, and therefore returns the best representation of the posterior distribution [39].

The computational complexity of the proposed inference procedure is $O(N \log N)$. The bottleneck of the proposed procedure lies in the evaluation of the likelihood functions in Equations (11), (15), and (18) (to appear in Section 3.1.2). Evaluating these likelihood functions with Gibbs iterations, every edge with the switch I_{ij} turned "on" requires a matrix product in the mixed-membership blockmodel term, $\boldsymbol{\pi}_i^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_j$. Although the number of communities grow with the size of the network, the number of communities each node takes membership in should not. In other words, in real-world networks, the number of communities each node participates in is limited by that individual's capacity so should not grow unbounded with network size. As the size of the network grows, the mixed-membership vectors $\boldsymbol{\pi}_i$ becomes increasingly sparse, making the cost of evaluating $\boldsymbol{\pi}_i^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_j$ capped at some constant value. Therefore, the overall computational complexity scales with the number of "on" edges, which is $N \log N$ in real-world networks. This theoretical complexity has been verified with experimental runs on networks of orders hundreds to thousands of nodes. Finally, inference with the mixed-membership model has been proven to be scalable and performed on large networks by Gopalan and Blei [42] through variational inference. Scaling up the HMMB inference to large networks as future work will likely involve similar approximate inference, and efficient proposal steps similar to Peixoto's work with the degree-corrected stochastic blockmodel [43].

3.1.1 Parameter Initialization

The potential challenge in identifiability discussed in Section 2.2 is mitigated with targeted initialization [Fig. 4]. Each independent inference chain is initialized to values likely to be closer to the true mode to increase its chance for converging to it. Good initial values of the mixed-membership estimate, $\boldsymbol{\Pi}^0$, can be obtained by a fast segmentation algorithm on the network (e.g. spectral clustering by [12]). After the segmentation, nodes belonging to the same cluster are initialized to be members of the same community. This way, nodes with strong interactions will likely be initialized to the same communities. The edge switches are simply initialized to "on" for each edge with observed interaction, and "off" for all other edges. This initialization, $\boldsymbol{I}^0$, is also the maximum posterior estimate. A good initialization of the node degree parameters $\boldsymbol{\lambda}^0$ sets them to values proportional to the average weight of the observed in- and out-edges of each node,

$$\lambda_i^0 \propto \frac{\sum_j a_{ij} + \sum_j a_{ji}}{\sum_j I_{ij}^0 + \sum_j I_{ji}^0}. \tag{10}$$

This initialization is also the unbiased estimator for the Chung-Lu model with desirable asymptotic properties [44]. The proportionality expression here accommodates an arbitrary relative scaling between $\boldsymbol{\lambda}$ and $\boldsymbol{B}$ which will be explained shortly. Because within-community interactions are typically stronger than the between-community interactions, the block matrix $\boldsymbol{B}^0$ is initialized to roughly diagonal with small values in the off-diagonal entries.

As shown in Theorem 1, there exists one parameter redundancy in the shared relative scale between the block matrix, $\boldsymbol{B}$ and the node degree parameters $\boldsymbol{\lambda}$. This issue can be mitigated by fixing the diagonal entries of the block matrix $\boldsymbol{B}$ to a precomputed value during the Gibbs sampling. It is equivalent to taking a profile slice of the likelihood functions at these fixed values. A reasonable precomputed value is the maximum posterior estimate from the gradient ascent described by Equation (24) in Section 3.1.3 with uniform $\boldsymbol{\lambda}$, after a preliminary run of the Gibbs sampling with a diagonal block matrix. This precomputed value represents the model's best fit on the network using the within-community interactions (i.e. the diagonal of $\boldsymbol{B}$).

If prior knowledge is available (e.g., the memberships on a few nodes is known or observed), one can introduce such knowledge as informative priors on the relevant parameters.

3.1.2 Bayesian Sampling Updates (MCMC)

After the parameters are initialized properly, the main computation through iterative updates begins, as shown in Figs. 3 and 4. This section presents the fully Bayesian approach using the MCMC sampling. Because most of the posterior densities in the updates do not have closed-form distributions, the Metropolis-Hasting algorithm is used to accept or reject the proposal based on the posterior ratio between the current sample and the proposal. When the proposal distribution is not symmetric, a Hasting correction ratio is added to maintain the detailed balanced condition. As mentioned previously in Section 3.1.1, a few parameters may be fixed. All other parameters will be updated iteratively as described below:

Update λ_i^{t+1} conditional on $\boldsymbol{\lambda}_{\bar{i}}^{t^+}, \boldsymbol{\Pi}^t, \boldsymbol{B}^t, \boldsymbol{I}^t$. The subscript notation $\bar{i}$ on a vector denotes the vector without element i. For the node degree update on each node i conditional on the other parameters, the governing conditional likelihood function is derived from Equations (1) and (2), resulting in

$$\begin{aligned}\mathcal{L}(\lambda_i \mid \boldsymbol{\lambda}_{\bar{i}}, \boldsymbol{\Pi}, \boldsymbol{B}, \boldsymbol{I}, \boldsymbol{A}) = \\ \exp\Bigg(\sum_{j \in I_{ij}=1} \Big(a_{ij} \log(\lambda_i \lambda_j \; T \; \boldsymbol{\pi}_i^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_j \Big) + \\ \sum_{j \in I_{ji}=1} \Big(a_{ji} \log(\lambda_i \lambda_j \; T \; \boldsymbol{\pi}_j^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\pi}_i \Big) \Bigg),\end{aligned} \tag{11}$$

where the index j refers to all source/target nodes to/from node i. The posterior distribution is

$$p(\lambda_i \mid \boldsymbol{\lambda}_{\bar{i}}, \boldsymbol{\Pi}, \boldsymbol{B}, \boldsymbol{I}, \boldsymbol{A}) \propto \mathcal{L}(\lambda_i \mid \boldsymbol{\lambda}_{\bar{i}}, \boldsymbol{\Pi}, \boldsymbol{B}, \boldsymbol{I}, \boldsymbol{A})\, p_0(\lambda_i). \tag{12}$$

Real-world networks often follow power-law degree distribution [2], which gives a natural prior for the Chung-Lu parameters λ_i,

$$p_0(\lambda_i) \propto \lambda_i^{\alpha}. \tag{13}$$

Typically, the exponent α is between -3 and -2 in real-world networks [29], so one may simply fix α to a reasonable value or update it as another step in the MCMC with a symmetric normal proposal distribution, $\alpha^{t^+} \sim \mathrm{Normal}(\alpha^t, \sigma_\alpha^2)$, and posterior, $p(\alpha \mid \boldsymbol{\lambda}) \propto \prod_{1 \le i \le N} \lambda_i^{-\alpha}$.

For updates on λ_i, a symmetric normal proposal distribution is used, $\lambda_i^{t^+} \sim \text{Normal}(\lambda_i^t, \sigma_\lambda^2)$. The variance σ_λ^2 of the normal proposal controls the step size of the proposal from the current sample. The acceptance probability, a_λ, of the proposed update is

$$a_\lambda = \begin{cases} \min\left(\dfrac{p(\lambda_i^{t^+} \mid \boldsymbol{\lambda}_{\bar{i}}, \boldsymbol{\Pi}, B, \boldsymbol{I}, \boldsymbol{A})}{p(\lambda_i^t \mid \boldsymbol{\lambda}_{\bar{i}}, \boldsymbol{\Pi}, B, \boldsymbol{I}, \boldsymbol{A})}, 1\right) & \text{if } \lambda_i^{t^+} > \epsilon_\lambda; \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

No Hasting correction ratio is needed because the proposal distribution is symmetric. Note that the new proposal is automatically rejected if it is below the minimal value ϵ_λ typically set close to zero because nodes cannot have negative degrees. So with probability α_λ, the proposed sample will be accepted (i.e. $\lambda_i^{t+1} = \lambda_i^{t^+}$), otherwise, the current sample will be kept (i.e. $\lambda_i^{t+1} = \lambda_i^t$).

Update b_{mn}^{t+1} conditional on $B_{\overline{mn}}^{t^+}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{\Pi}^t, \boldsymbol{I}^t$. The subscript notation $\overline{mn}$ on a matrix denotes the matrix without element mn. For updating the block matrix one element, b_{mn}, at a time conditional on the other parameters, the governing conditional likelihood function is

$$\mathcal{L}(b_{mn} \mid B_{\overline{mn}}, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{I}, \boldsymbol{A}) = \exp\left(\sum_{i,j \in \mathbb{I}_1} \left(a_{ij} \log\left(\boldsymbol{\pi}_i^{\mathrm{T}} B^* \boldsymbol{\pi}_j\right) - \lambda_i \lambda_j T \, \boldsymbol{\pi}_i^{\mathrm{T}} B^* \boldsymbol{\pi}_j\right)\right), \tag{15}$$

where B^* consists of the element being updated, b_{mn} and the other elements $B_{\overline{mn}}$. The posterior distribution is

$$p(b_{mn} \mid B_{\overline{mn}}, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{I}, \boldsymbol{A}) \propto \mathcal{L}(b_{mn} \mid B_{\overline{mn}}, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{I}, \boldsymbol{A})\, p_0(b_{mn}). \tag{16}$$

Typically, a flat prior is used for b_{mn}. However, given prior knowledge on the structure of the block matrix, one may add an informative prior accordingly (e.g., prior with mass at zero on off-diagonal elements for sparse B). Similar to the node degrees parameter update, a symmetric normal proposal distribution is used, $b_{mn}^{t^+} \sim \text{Normal}(b_{mn}^t, \sigma_B)$. With a non-negative constraint, the acceptance probability a_b of the proposed update is

$$a_b = \begin{cases} \min\left(\dfrac{p(b_{mn}^{t^+} \mid B_{\overline{mn}}, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{I}, \boldsymbol{A})}{p(b_{mn}^t \mid B_{\overline{mn}}, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{I}, \boldsymbol{A})}, 1\right) & \text{if } b_{mn}^{t^+} \geq 0; \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Update $\boldsymbol{\pi}_i^{t+1}$ conditional on $\boldsymbol{\Pi}_{\bar{i}}^{t^+}, B^{t+1}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{I}^t$. The subscript notation $\bar{i}$ on a matrix denotes the matrix without column i. The conditional likelihood function for mixed-membership updates on each node i is

$$\mathcal{L}(\boldsymbol{\pi}_i \mid \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, \boldsymbol{I}, \boldsymbol{A}) = \exp\left(\sum_{j \in I_{ij}=1} \left(a_{ij} \log\left(\boldsymbol{\pi}_i^{\mathrm{T}} B \boldsymbol{\pi}_j\right) - \lambda_i \lambda_j \; T \; \boldsymbol{\pi}_i^{\mathrm{T}} B \boldsymbol{\pi}_j\right) + \sum_{j \in I_{ji}=1} \left(a_{ji} \log\left(\boldsymbol{\pi}_j^{\mathrm{T}} B \boldsymbol{\pi}_i\right) - \lambda_i \lambda_j \; T \; \boldsymbol{\pi}_j^{\mathrm{T}} B \boldsymbol{\pi}_i\right)\right), \tag{18}$$

which is similar to the likelihood for the block matrix. The posterior distribution is

$$p(\boldsymbol{\pi}_i \mid \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, \boldsymbol{I}, \boldsymbol{A}) \propto \mathcal{L}(\boldsymbol{\pi}_i \mid \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, \boldsymbol{I}, \boldsymbol{A})\, p_0(\boldsymbol{\pi}_i). \tag{19}$$

Without additional information on the different lifestyles of the population, as is typically the case, one can simply use a flat prior $p_0(\boldsymbol{\pi}_i) \propto 1$. Because the space for $\boldsymbol{\pi}_i$ is a simplex (i.e. $\mathbb{1}^{\mathrm{T}} \boldsymbol{\Pi} = \mathbb{1}^{\mathrm{T}}$), a logistic-normal proposal function is used, inspired by the work from [45]. Let $\boldsymbol{\pi}_i^t$ denote the current sample on the mixed-membership of i, the proposed update, denoted $\boldsymbol{\pi}_i^{t^+}$, is governed by

$$\boldsymbol{\pi}_i^{t^+} \sim \text{logistic}\Big(\text{Normal}\big(\text{logit}(\boldsymbol{\pi}_i^t), \boldsymbol{\Sigma}_{\boldsymbol{\pi}}\big)\Big), \tag{20}$$

where the magnitude of a typically diagonal $\boldsymbol{\Sigma}_{\boldsymbol{\pi}}$ controls the step size of the proposal from the current sample. The acceptance probability $a_{\boldsymbol{\pi}}$ of the proposed update is

$$a_{\boldsymbol{\pi}} = \min\left(\frac{p(\boldsymbol{\pi}_i^{t^+} \mid \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, \boldsymbol{I}, \boldsymbol{A}) Q(\boldsymbol{\pi}_i^t; \boldsymbol{\pi}_i^{t^+})}{p(\boldsymbol{\pi}_i^t \mid \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, \boldsymbol{I}, \boldsymbol{A}) Q(\boldsymbol{\pi}_i^{t^+}; \boldsymbol{\pi}_i^t)}, 1\right), \tag{21}$$

where $Q(\boldsymbol{\pi}_i^t; \boldsymbol{\pi}_i^{t^+})$ is the logistic-normal density function with mean $\boldsymbol{\pi}_i^{t^+}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\pi}}$ evaluated at $\boldsymbol{\pi}_i^t$. The Hasting correction ratio $Q(\boldsymbol{\pi}_i^t; \boldsymbol{\pi}_i^{t^+})/Q(\boldsymbol{\pi}_i^{t^+}; \boldsymbol{\pi}_i^t)$ is needed because the logistic-normal proposal is not symmetric.

Update I_{ij}^{t+1} conditional on $\boldsymbol{\Pi}^{t+1}, B^{t+1}, \boldsymbol{\lambda}^{t+1}$. Lastly, updating the edge switches, I_{ij}, is fairly straight forward because each switch is conditionally independent of each other. For edges with observed interactions (i.e. $a_{ij} > 0$), I_{ij} must be 1. For edges with zero observed interactions (i.e. $a_{ij} = 0$), the governing posterior distribution is

$$\begin{aligned} &p(I_{ij} = 1 \mid B, \boldsymbol{\lambda}, \boldsymbol{\Pi}, \boldsymbol{A}) \\ &\quad = \frac{p(I_{ij} = 1 \mid a_{ij} = 0)\; p_0(I_{ij} = 1)}{p(a_{ij} = 0, I_{ij} = 1) + p(a_{ij} = 0, I_{ij} = 0)} \Big| \, B, \boldsymbol{\lambda}, \boldsymbol{\Pi} \\ &\quad = \frac{\exp(-\lambda_{ij} T) s}{\exp(-\lambda_{ij} T) s + (1 - s)}, \end{aligned} \tag{22}$$

where λ_{ij} is calculated using Equation (2) on the current parameter samples. Updating I_{ij}^{t+1} can simply be done by a Bernoulli draw with probability equal to the expression in Equation (22). The sparsity parameter s can be set to a reasonable value, or updated as another step in the MCMC with a closed-form posterior, $s^{t+1} \sim \text{Beta}(\sum_{1 \leq i,j \leq N} I_{ij} + 1, N^2 - \sum_{1 \leq i,j \leq N} I_{ij} + 1)$. When the network is large, it may become infeasible to update I_{ij} for every edge such that $a_{ij} = 0$. A practical implementation is to simply turn all such I_{ij} off which is equivalent to fixing $\boldsymbol{I}$ at the maximum posterior estimate. Alternatively, one may selectively update such I_{ij} only on edges likely with low frequency λ_{ij}.

3.1.3 Maximum Posterior Updates (MCEM)

An alternative to the MCMC procedure is to maximize the posterior probability when updating some of the parameters. This is commonly known as the Monte Carlo Expectation Maximization [40], [41]. The MCEM may converge faster due to its greedy nature, but does not capture the full posterior distribution and is more prone to being trapped in local modes. Because this paper takes a special interest in estimating the posterior distribution of the mixed-memberships, the MCEM developed here maximizes the posteriors on the node degree and block matrix parameters. Maximization is

done through gradient ascent. A performance comparison between the MCMC and MCEM procedures is given later in Section 3.3 with discussion of results in Table 2.

Alternative update for λ_i^{t+1} conditional on $\boldsymbol{\lambda}_{\bar{i}}^{t^+}, \boldsymbol{\Pi}^t, B^t, \boldsymbol{I}^t$. Maximizing the posterior on the node degree parameters [Equation (12)] is performed by gradient ascent:

$$\lambda_i^{t+1} = \lambda_i^t + \gamma_\lambda^t \left(\sum_{j \in I_{ij}=1} \left(\frac{a_{ij}}{\lambda_i^t} - \lambda_j^t \, T \, \boldsymbol{\pi}_i^{\mathrm{T}} B \boldsymbol{\pi}_j \right) + \sum_{j \in I_{ji}=1} \left(\frac{a_{ji}}{\lambda_i^t} - \lambda_i^t \, T \, \boldsymbol{\pi}_j^{\mathrm{T}} B \boldsymbol{\pi}_i \right) - \frac{\alpha}{\lambda_i^t} \right), \tag{23}$$

where γ_λ^t is the step size, and the gradient is the derivative of the log posterior evaluated at λ_i^t. A sensible update strategy is the batch mode where all of the λ_i take one gradient ascent step together at a time. This is not only more efficient to compute, but also makes sense for a more balanced climb in the space spanned by $\boldsymbol{\lambda}$.

Alternative update for b_{mn}^{t+1} conditional on $B_{\overline{mn}}^{t^+}, \boldsymbol{\lambda}^{t+1}, \boldsymbol{\Pi}^t, \boldsymbol{I}^t$. Similarly, the posterior on the block matrix parameters, shown in Equation (16), is maximized via

$$b_{mn}^{t+1} = b_{mn}^t + \gamma_B^t \sum_{i,j \in \mathbb{I}_1} \left(\frac{a_{ij}}{\boldsymbol{\pi}_i^{\mathrm{T}} B^t \boldsymbol{\pi}_j} - \lambda_i \lambda_j \, T \right) \pi_{im} \pi_{jn}, \tag{24}$$

where γ_B^t is the step size, and the gradient is the derivative of the log posterior evaluated at b_{mn}^t. The batch update is again a sensible strategy where all of the b_{mn} are updated together one step at a time.

3.2 Fisher Information and Performance Bounds

With a focus on estimating the mixed-memberships for its application in community detection (see Section 4), this section derives a theoretical membership estimation performance bound through Fisher information analysis. This result reveals what makes certain network data more informative than the others. It also provides the best case estimation performance. Such knowledge can help guide researchers in collecting enough and the most informative data, to achieve the desired accuracy and precision on the membership estimate.

The results presented here consist of the Fisher information matrix, Cramér-Rao bound, and the asymptotic posterior distribution of the mixed-membership estimate $\boldsymbol{\pi}_i$ for each node i, conditional on the other parameters. This is a best case scenario because in practice the estimates on the other parameters have uncertainty and possibly bias as well. Fisher information analysis of the data on the mixed-membership begins with the log of the likelihood function in Equation (18). Because $\boldsymbol{\pi}_i$ stays inside a $(K-1)$ simplex (i.e., $\mathbb{1}^{\mathrm{T}} \boldsymbol{\Pi} = \mathbb{1}^{\mathrm{T}}$), the redundant dimension is removed through the reparameterization

$$\boldsymbol{\pi}_i = \left(\pi_{i1} \; \pi_{i2} \; \ldots \; \pi_{i,K-1} \; \; 1 - \sum_{k=1}^{K} \pi_{ik} \right), \tag{25}$$

where the K-th component is entirely determined by the others. The Fisher information on the reparameterized parameters $\boldsymbol{\pi}_i$ is a $(K-1)$-by-$(K-1)$ matrix determined by the negative expected curvature of the log-likelihood function:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\pi}_i)_{mn} &= -E_{\boldsymbol{A}} \left[\frac{\partial^2 \ell(\boldsymbol{\pi}_i \,|\, \boldsymbol{\Pi}_{\bar{i}}, B, \boldsymbol{\lambda}, I, A)}{\partial \pi_{im} \partial \pi_{in}} \right] \\ &= T \left(\sum_{j \in I_{ij}=1} \frac{\lambda_i \lambda_j \big((B_{m\cdot} - B_{K\cdot}) \boldsymbol{\pi}_j \big) \big((B_{n\cdot} - B_{K\cdot}) \boldsymbol{\pi}_j \big)}{\boldsymbol{\pi}_i^{\mathrm{T}} B \boldsymbol{\pi}_j} + \sum_{j \in I_{ji}=1} \frac{\lambda_i \lambda_j \big(\boldsymbol{\pi}_j^{\mathrm{T}} (B_{\cdot m} - B_{\cdot K}) \big) \big(\boldsymbol{\pi}_j^{\mathrm{T}} (B_{\cdot n} - B_{\cdot K}) \big)}{\boldsymbol{\pi}_j^{\mathrm{T}} B \boldsymbol{\pi}_i} \right), \end{aligned} \tag{26}$$

where $m, n \in 1, 2, \ldots, K-1$ and $B_{m\cdot}$ denotes the mth row of matrix B. While the precise information content on each membership estimate is captured by the entire Fisher information matrix, each of the m-th diagonal element of the Fisher information matrix, $\mathcal{I}(\boldsymbol{\pi}_i)_{mm}$, gives an interpretable first order indication on the information content of the data on node i's membership to community m:

$$\mathcal{I}(\boldsymbol{\pi}_i)_{mm} = T \left(\sum_{j \in I_{ij}=1} \frac{\lambda_i \lambda_j \big((B_{m\cdot} - B_{K\cdot}) \boldsymbol{\pi}_j \big)^2}{\boldsymbol{\pi}_i^{\mathrm{T}} B \boldsymbol{\pi}_j} + \sum_{j \in I_{ji}=1} \frac{\lambda_i \lambda_j \big(\boldsymbol{\pi}_j^{\mathrm{T}} (B_{\cdot m} - B_{\cdot K}) \big)^2}{\boldsymbol{\pi}_j^{\mathrm{T}} B \boldsymbol{\pi}_i} \right). \tag{27}$$

Agreeing with intuition, this quantity shows that information content is proportional to the amount of time, T, during which data is collected. It also accumulates over all the incoming and outgoing edges to i. The information content of each such edge is proportional to the node degrees product $\lambda_i \lambda_j$ (the Chung-Lu term), and is maximized when either most of the interactions between i and j can be explained through community m or none of such interactions can be explained through community m (i.e. making the squared difference term large), normalized by the total interaction strength based on community memberships. This result reveals the kind of edges that contributes the most to the information content towards estimating each community membership of a given node.

The Fisher information matrix leads to the asymptotic posterior distribution of the mixed-membership $\boldsymbol{\pi}_i$ of each node i, through the Bernstein-von Mises Theorem [46] under mild regularity conditions:

Theorem 2 (Asymptotic Posterior Distribution of Mixed-Membership). *Let $\boldsymbol{A}$ be a network generated by the hybrid mixed-membership blockmodel. The conditional posterior distribution of node i's mixed-membership $\boldsymbol{\pi}_i$ converges asymptotically, with increasing network data, in distribution to normal with mean at the true value and covariance equal to the inverse of the Fisher information matrix derived in Equation (26):*

$$\boldsymbol{\pi_i} \,|\, \boldsymbol{\Pi}_{\bar{i}}, \boldsymbol{\lambda}, I, B, \boldsymbol{A} \xrightarrow{D} \mathrm{Normal}\big(\boldsymbol{\pi}_i, \mathcal{I}(\boldsymbol{\pi}_i)^{-1} \big). \tag{28}$$

The inverse of the Fisher information matrix, $\mathcal{I}(\boldsymbol{\pi}_i)^{-1}$, is also commonly known as the Cramér-Rao lower bound, which gives the minimal covariance of the maximum likelihood estimate. The Bernstein-von Mises Theorem provides a Bayesian interpretation of the Cramér-Rao lower bound in Theorem 2. As a best case analysis by conditioning on the other parameters, this bound becomes tight asymptotically

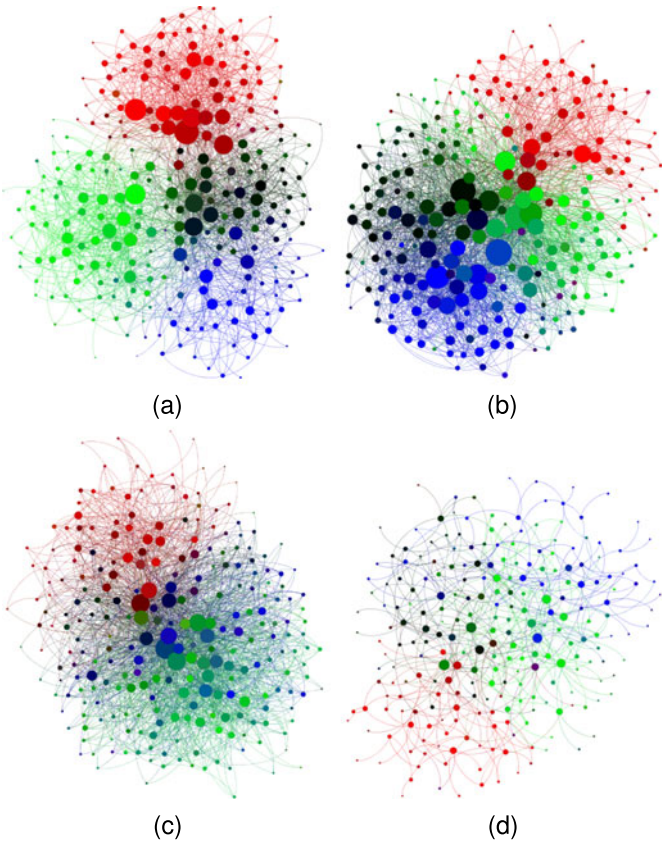

Fig. 5. Networks layouts with (a) baseline settings, (b) strong between-community interactions, (c) highly mixed memberships, and (d) sparse edges. A node's color is determined by its memberships to the first three communities which fill the Red-Green-Blue channels. Membership to the fourth community makes a node darker as it does not contribute to the three color channels. Larger nodes are those with higher degrees.

as the parameter estimates converge toward their true values with an increasing amount of data. This result is useful for characterizing the best case performance in inferring the mixed-membership parameters given the amount of network data available. Practically, this informs researchers on how much data to collect in order to meet the desired estimation performance.

3.3 Parameter Estimation Performance

This section characterizes the performance of the estimation procedure described above, on networks simulated over a range of reasonable and realistic values on the key model parameters. Overall, the proposed procedure with targeted parameter initialization performs well by covering the true values with posterior intervals of width close to the theoretical Cramér-Rao bound, across a reasonable range of parameter settings. This demonstrates empirically the parameter identifiability of the HMMB. In challenging settings where the community structure is no longer distinct due to strong between-community interactions or highly mixed memberships, the procedure's performance drops off gracefully and remains adequate, highlighting the strength of inference with HMMB for networks with overlapping communities.

The estimation procedure is evaluated across the four main network features as they are varied by sweeping through the corresponding model parameters, as described in Section 2 and Fig. 1: (1) strength of between-community interactions governed by the block matrix B, (2) level of mixed-memberships by the mixed-membership pseudocount matrix X, (3) number of high-degree nodes by the power-law exponent α, and (4) network density by the sparsity parameter s. Fig. 5 provides visualization of 256-node network samples over a time span of 100 units with four communities across these features. A node's color is determined by its memberships to the first three communities which fill the Red-Green-Blue channels. Membership to the fourth community makes a node darker as it does not contribute to the three color channels. Larger nodes are those with higher degrees. Fig. 5a shows the baseline setting with a small fraction of high-degree nodes (power-law exponent $\alpha = -2.9$), medium network density ($s = 20\%$), and moderate levels of between-community interaction and mixed-memberships:

$$X = \begin{bmatrix} 5 & 0.1 & 0.1 & 1 \\ 0.1 & 5 & 1 & 0.1 \\ 0.1 & 0.1 & 2 & 0.5 \\ 0.1 & 1 & 0.3 & 3 \end{bmatrix}, \tag{29}$$

$$B = \begin{bmatrix} 2.3 & 0.07 & 0 & 0 \\ 0.3 & 2 & 0 & 0 \\ 0 & 0 & 2.5 & 0.4 \\ 0 & 0.3 & 0 & 3 \end{bmatrix}. \tag{30}$$

Each node adopts one of the four lifestyles (the rows in X) with equal probability, $\boldsymbol{\phi} = (0.25\ 0.25\ 0.25\ 0.25)$. The baseline case has overlapping communities but each of the four communities still has its distinct cluster. Fig. 5b shows a network sample with strong between-community interactions, generated with the baseline setting but multiplying the off-diagonal entries of the block matrix B by 3. Distinct clusters are no longer visible for each community. Fig. 5c shows a network sample with a high level of mixed-memberships seen in the mixed-colors of most nodes, by keeping the baseline setting but multiplying the off-diagonal entries of the pseudocount matrix X by 3. Again, distinct clusters are not visible. Fig. 5d shows a network sample with the baseline setting but lowering the network density s to 5 percent. For brevity, a visualization with more high-degree nodes is skipped here, because it looks just like the baseline case with more high-degree nodes.

A typical output of the Bayesian estimation procedure is an interval that covers p percent of the posterior distribution. A desirable posterior interval captures the true value of the parameter p percent of the time in the narrowest interval possible. Therefore, performance can be summarized by the frequency of truth coverage and the interval width [39]. For each network setting, the results are averaged over inference on 25 simulated networks. The parameter estimates of interest are the mixed-memberships $\boldsymbol{\Pi}$, the node degrees $\boldsymbol{\lambda}$, and the block matrix B. For brevity, results are reported as the average over all the nodes and communities. As discussed in Section 3.1.1, there exists a arbitrary scaling between the block matrix and the node degree parameters, which is dealt with by rescaling them to match the truth before evaluation. Because the block matrix diagonal is scaled according to the truth, it is not included in the evaluation.

Table 2 shows a comparison between variants of the proposed estimation procedure, on networks with the baseline setting (see Fig. 5b) and the more challenging setting

TABLE 2
Frequentist Coverage* of the 90 Percent Posterior Interval and the Interval Width in Parentheses

Network setting	Parameter of interest	Estimation strategy			
		MCMC with full init strategy	MCMC with parameter init without fixing block diag	MCMC with no init strategy	MCEM with full init strategy
Baseline network	Π	83.0% (0.089)	69.4% (0.098)	57.7% (0.171)	82.3% (0.082)
	λ	85.9% (0.046)	92.4% (0.095)	51.5% (0.148)	–
	B	70.7% (0.049)	67.0% (0.026)	62.3% (0.007)	–
Strong between-community interactions	Π	78.4% (0.102)	71.0% (0.112)	61.1% (0.148)	74.8% (0.095)
	λ	84.0% (0.043)	91.5% (0.095)	42.7% (0.118)	–
	B	74.0% (0.066)	66.7% (0.056)	64.7% (0.013)	–

*Frequentist coverage is measured as the percentage of the instances where the posterior interval covers the true value.

with strong between-community interactions (see Fig. 5b). The variants include the MCMC procedure described in Section 3.1.2 with no initialization strategy, partial initialization strategy (without fixing the block matrix diagonal), and full initialization strategy as described in Section 3.1.1. The MCEM procedure described in Section 3.1.3 with full initialization is also included as a comparison.

The first three columns of results in Table 2 show that the full initialization strategy described in Section 3.1.1 improves the estimation performance, evident in the better truth coverage and generally narrower interval. The targeted parameter initialization started the MCMC in a more feasible region of the large parameter space, increasing the likelihood of converging upon the true posterior mode. Fixing the block matrix diagonal to a precomputed value effectively addressed the identifiability issue from the arbitrary scale between the block matrix diagonal and the node degrees. The coverage for λ under partial initialization strategy is a few percentage points better than that under full initialization, but at the big cost of doubling the interval width. Without fixing the block matrix diagonal, the MCMC visits a wide range of values for λ due to the arbitrary scale, resulting in a wide posterior interval that more likely covers the truth. This does not indicate better performance for λ under partial initialization.

Comparing the first and the last columns in the result shows that the MCEM procedure performs roughly on par with the MCMC procedure, with slightly lower coverage and narrower intervals. This is not surprising due to the more greedy nature of the MCEM in maximizing the posterior when updating λ and B. As a result, MCEM does not render posterior distributions on λ and B, so the result on those parameters is left blank. Under both the baseline and the strong between-community interactions settings, the MCEM procedure found the posterior mode in slightly fewer iterations than MCMC. In this simulation, there is not enough gain in convergence speed for MCEM to justify sacrificing the the nice property of capturing the full posterior distribution.

As the MCMC procedure with full initialization strategy provides the best performance, it will be the focus of evaluation for the rest of this section. The first column of Table 2 shows that the 90 percent posterior interval slightly undercovers the true Π and λ by a few percentage points in the baseline setting. This is caused by fixing the block matrix diagonal to pre-computed values with some errors. This is a small price to pay for mitigating the arbitrary scale between the block matrix and λ. Coverage on B is worse but reasonable, again, due to the error on the fixed diagonal. In the more challenging setting under strong between-community interactions, the coverages on Π and λ drop slightly but are reasonable. This highlights the strength of modeling and inference under HMMB for highly overlapping communities with no distinct clusters (see Fig. 5b). Moreover, the mixed-membership estimates nearly achieve the precision of the theoretical Cramér-Rao lower bound derived in Section 3.2, indicating the estimation procedure has good statistical efficiency. In practice, this bound may not be reached by any unbiased estimator, because it is derived conditional on all other parameters (see Theorem 2) as a best-case analysis.

Having established the good baseline performance of the MCMC procedure with full initialization, we now evaluate its performance as the four main features of the HMMB are varied from the baseline setting. Fig. 6 shows the estimation performance as the scale of the block matrix off-diagonal entries varies from 0.25 to 5. This scale corresponds to the strength of between-community interactions, with a scale of 1 being the baseline setting in Table 2 and Fig. 5c, and a scale of 3 being the challenging setting in Table 2 and Fig. 5b. As the scale increases, the communities overlap more and more and eventually lose any structure of distinct clusters, making it difficult to estimate the community memberships as the posterior distribution becomes flatter and increasingly multi-modal. The estimation performance drops off gracefully as seen in Fig. 6a, with reasonable truth coverage even with very strong between-community interactions. This highlights the strength of modeling and inference under

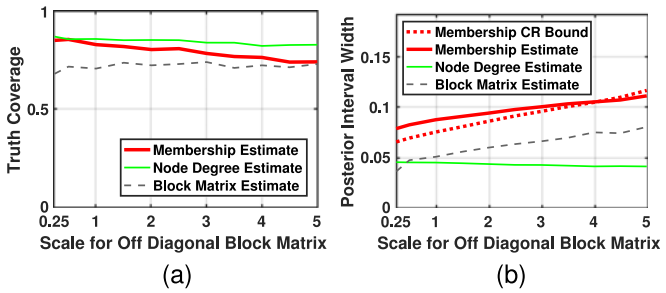

Fig. 6. Estimation performance as the strength of between-community interactions varies, in terms of (a) truth coverage and (b) width of the 90 percent posterior interval.

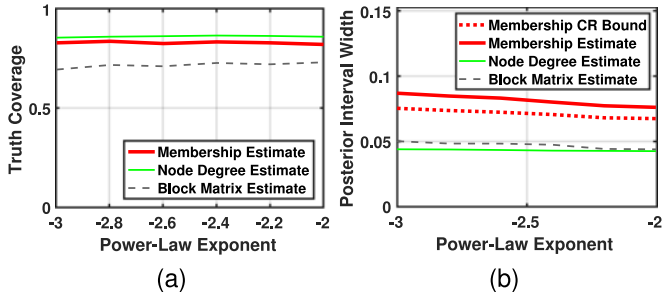

Fig. 7. Estimation performance as the number of high-degree nodes varies, in terms of (a) truth coverage and (b) width of the 90 percent posterior interval.

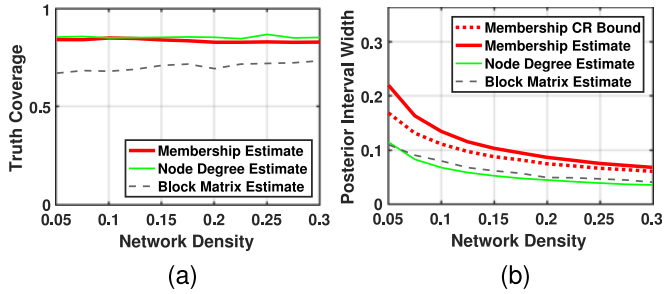

Fig. 8. Estimation performance as the network density varies, in terms of (a) truth coverage and (b) width of the 90 percent posterior interval.

HMMB for highly overlapping communities. The posterior interval widths for $\mathbf{\Pi}$ and $\boldsymbol{B}$ also increase gracefully as shown in Fig. 6b. The node degree interval becomes narrower due to more total interactions overall. The mixed-membership interval width stays slightly above the Cramér-Rao lower bound as expected but falls below the bound when the scale is at 4. This is likely due to biased mixed-membership estimates and the bias-variance trade-off [47]. The Cramér-Rao lower bound rises with stronger between-community interactions, pointing to a decrease in the Fisher information content of the interactions because they can be explained through multiple communities. This matches the intuitions from the theoretical result in Equation (26).

The estimation performance over varying levels of mixed-memberships by scaling the off-diagonal entries of the pseudocount matrix (X) is very similar to those in Fig. 6, as the community overlap increases. The figures and discussion are omitted for brevity.

Fig. 7 shows the estimation performance over varying numbers of high-degree nodes as the negative power-law exponent (α) goes from -3 to -2, the range of values observed in real-world networks [29]. Less negative exponent leads to more high-degree nodes. Truth coverage remains steady and the posterior interval widths decrease gradually. Again, the mixed-membership interval width tracks the Cramér-Rao lower bound nicely, indicating statistical efficiency. The decreasing Cramér-Rao lower bound matches the theoretical result in Equation (26) which indicates that an increase of high-degree nodes leads to higher Fisher information content.

Fig. 8 shows the estimation performance as the network density (the sparsity parameter s) varies from 0.05 to 0.3. A visualization of a sparse network with $s = 0.05$ is shown in Fig. 5d. Truth coverage is steady and the posterior interval widths decrease with increasing network density. Higher density increases the total number of interacting edges and Fisher information content as shown in Equation (26). Accordingly, variance decreases for all parameter estimates and the mixed-membership interval width gets closer to the decreasing Cramér-Rao lower bound.

Overall, the proposed MCMC procedure with full initialization strategy is robust across the ranges of realistic network settings. For gathering statistics on performance in a timely fashion, the networks evaluated in this paper are kept at hundreds of nodes. At the baseline parameter setting, we have verified that the inference procedure performs just as well on larger networks with thousands of nodes and six communities. However, each run takes hours on a regular desktop in the current unoptimized Matlab implementation. Scaling up to larger networks is a direction for future work.

4 Community Detection Results

A natural application for the HMMB is community detection, using the mixed-membership estimate from Section 3.1. Community detection has garnered much interest especially in the last ten years, both in terms of algorithm development and real-world applications. The discovery of communities and their members gives insight to network formation and identifies individuals of interest for targeted campaigns (e.g., public health interventions, promotional programs, advertisement, etc.). For a brief survey of the existing work, please see Section 1.1.

This section demonstrates HMMB's utility in community detection and compares its results against the closely related degree-corrected stochastic blockmodel with overlapping groups [8] and other approaches that exploits network modularity, including the well-known spectral clustering [12], modularity maximization [13], [48], and random walks on graph [14], [49], [50]. Specifically, this demonstration is on the problem of cued community detection. Results on simulated and real data over a range of realistic settings show that HMMB significantly outperforms the other methods in the presence of overlapping communities.

4.1 Cued Community Detection Approaches

The objective of cued community detection is to identify all the members (i.e. nodes) belonging to a community of interest, given prior knowledge on a number of its members (i.e. cue nodes). Typically, only a very small number of cue nodes are given. The nodes belonging to the community of interest are the foreground nodes and the rest, background nodes. Here, in a mixed-membership setting, the foreground nodes are defined as those with membership to the community of interest exceeding a given threshold (i.e. 50 percent).

HMMB-based community detection uses the mixed-membership estimate in Section 3.1. The targeted communities are determined as those which the cue nodes are estimated to have a high level of membership in. Detection is declared on the other nodes if a significant percentage (i.e. above detection threshold) of its estimated membership posterior on any of the targeted communities exceeds the level that defines the foreground community.

Similarly, inference using the degree-corrected stochastic blockmodel with overlapping groups [8] produces mixed-membership estimate and is used in the same way to

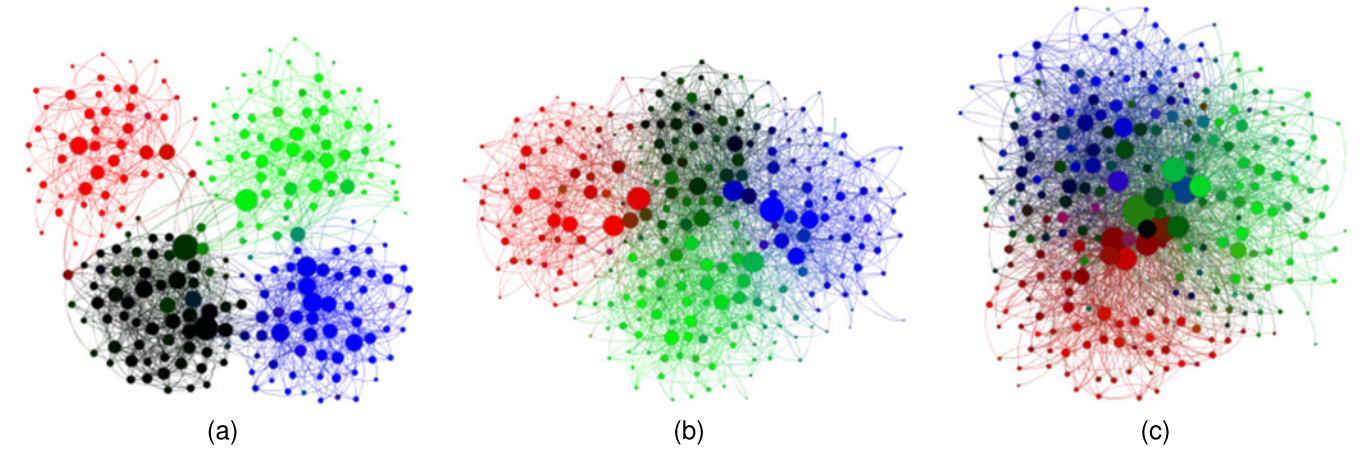

Fig. 9. Networks with increasing levels of community overlap at (a) 0.25, (b) 1, and (c) 2, from higher between-community interactions and mixed-memberships. A nodes color is determined by its memberships to the first three communities which fill the Red-Green-Blue channels. Membership to the fourth community makes a node darker as it does not contribute to the three color channels. Larger nodes are those with higher degrees. The foreground community to be detected is the fourth community.

produce cued community detection results in this section. This approach is the closest to the HMMB in spirit and in terms of model complexity, accounting for node degree variation and community overlap. However, it differs from the HMMB by allowing for multiple community memberships instead of mixed-memberships, and it does not account for sparsity explicitly.

Another intuitive method to accomplish this task is to simply propagate the "threat" from the cue nodes along the edges of the network and declare detection on nodes based on the "threat" propagated to them [14]. This works very much like a heat diffusion process on the network through the edges, where the cue nodes are the "sources" of the heat. This type of method can be formally defined through random walks on graph.

The popular modularity maximization [13], [48] and spectral clustering [12] methods for community detection produce hard partition of the network, instead of soft membership scores like the methods above. So for each detection run, they result in a single probability of detection (PD) and false alarm (PFA) point, instead of the whole receiver operating characteristic (ROC) curve. Detections are simply declared on all the nodes in the same partition as the cue nodes.

Performance of the HMMB will be compared against these four approaches that cover a wide spectrum of existing approaches in community detection. Compared to the HMMB and the DCSB with overlapping groups, threat propagation, modularity maximization, and spectral clustering have much lower computational complexity.

4.2 Detection Results on Simulated Network Data

The section begins with evaluation on simulated networks, with increasing levels of community overlap, a known challenge to the existing community detection algorithms [6] on real-world networks. As the level of community overlap increases, the community structure loses its distinct clusters and the individual communities are harder to tease apart, making community detection more difficult. Under HMMB, the level of community overlap can be adjusted through the strength of the between-community interactions and the level of mixed-membership, as shown in Section 3.3. Evaluation is done on networks with low, moderate, and high level of community overlap, with examples shown in Fig. 9. The moderate community overlap setting here corresponds to the baseline in Section 3.3. Performance is evaluated with 1, 3, and 5 cue nodes, for a range of realistic settings. The mixed-membership of the nodes (i.e., the node colors in Fig. 9) is not known.

Fig. 10 is a 3-by-3 matrix of receiver operating characteristic curves and points that summarize detection results at each community overlap level and number of cue nodes. For each setting, 25 independent draws of the network with 256 nodes and 30 independent random draws of the cue nodes for each network provide 750 instances of detection performance. For clarity and parsimony, the median performance of all the runs is plotted in the figure. For approaches that report soft membership scores, a ROC is computed by varying the detection threshold on the nodes. The objective is to detect nodes (i.e. foreground community) that have at least 50 percent of its true mixed-membership under the community of interest, colored black in Fig. 9. Cue nodes are not included in the evaluation. In real-world applications, the true number of communities is not typically known. To make sure that the HMMB and spectral clustering have enough capacity to express the community structure, it is practical to have a generous number of communities to fit the data. The detection runs here use a higher number of communities ($K = 6$) than the truth ($K = 4$). Peixoto's inference using the DCSB with overlapping groups determines the optimal number of communities by minimizing entropy under that model. Modularity maximization using the Louvain method also determines the optimal number of communities. Threat propagation performs random walks on graphs from the cue nodes and does not model the number of communities.

The first column of the ROC matrix shows that under a low level of community overlap, all five approaches perform almost perfectly. However, with a moderate level of community overlap (second column), modularity maximization, spectral clustering, and threat propagation perform much worse. This is not surprising because these approaches depend on the strength of the within-community interactions over the between-community interactions (i.e., modularity), which diminishes as the community overlap increases. Both

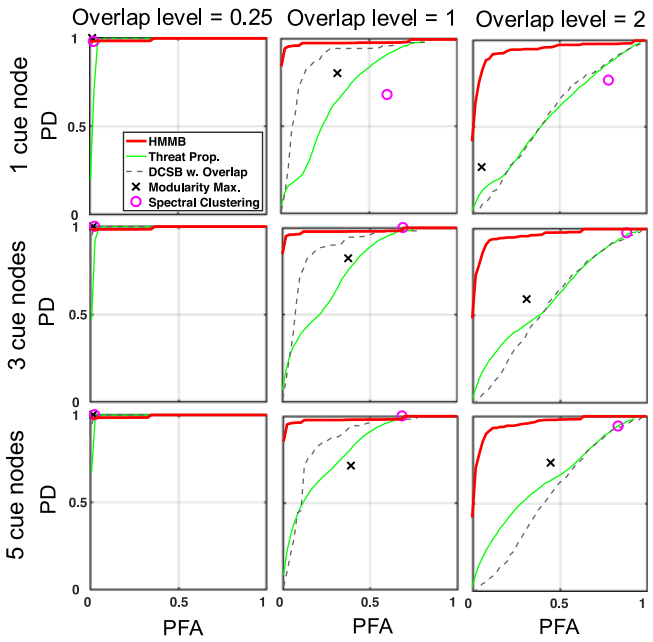

Fig. 10. Detection results on the simulated networks for the HMMB and the four baseline approaches: threat propagation [14], DCSB with overlapping groups [8], modularity maximization [48], and spectral clustering [12]. For clarity and parsimony, the median performance of all the runs is reported.

the DCSB with overlapping groups and the HMMB approach handles community overlap better by accounting for it during estimation, and is able to identify the foreground nodes through the mixed-membership estimate. The HMMB outperforms the DCSB with overlapping groups due to its explicit modeling for network sparsity and unevenly-mixed memberships. The third column of the ROC matrix shows that the HMMB performs reasonably even under a high level of community overlap. Having more cue nodes improves the performance of threat propagation as random walks from more cue nodes reach more target nodes before the threats are spilled over to the background nodes. Overall, Fig. 10 highlights the strength of HMMB for community detection, especially in the presence of overlapping communities.

4.3 Detection Results on Co-Authorship Network Data

This section evaluates the community detection performance on a real-world network that represents the co-authorship between researchers in the machine learning, computer vision, and signal processing field, shown in Fig. 11. The nodes represent authors and the edges co-authorship between them.

This network is constructed from the DBLP bibliography database [51] using papers published from 1983 to 2014 in a signal processing conference (ICASSP), computer vision conferences (ICCV and CVPR), and machine learning conferences (ICML, NIPS, and KDD). The mixed-membership is each author's fraction of publications in the three types of conferences. For data richness and network connectivity, the authors included are the top 314 most prolific authors in these conferences. Fig. 11 shows a significant amount of community overlap. The mixed-memberships of the nodes are not known a priori so the detection algorithm is only

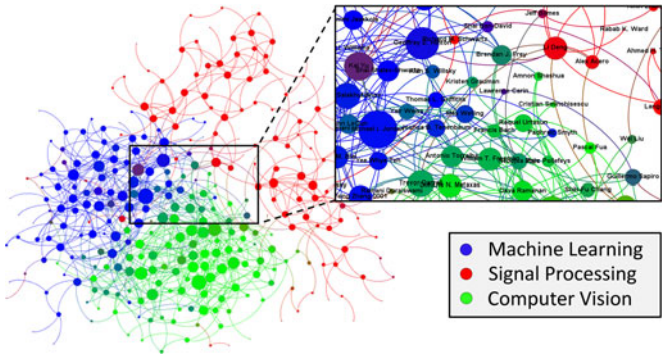

Fig. 11. A DBLP co-authorship network where the communities are the type of conference the publications took place. Nodes are colored by their community mixed-membership in each of the three conference types, filling the Red-Green-Blue channels. Node size indicates the number of co-authorships.

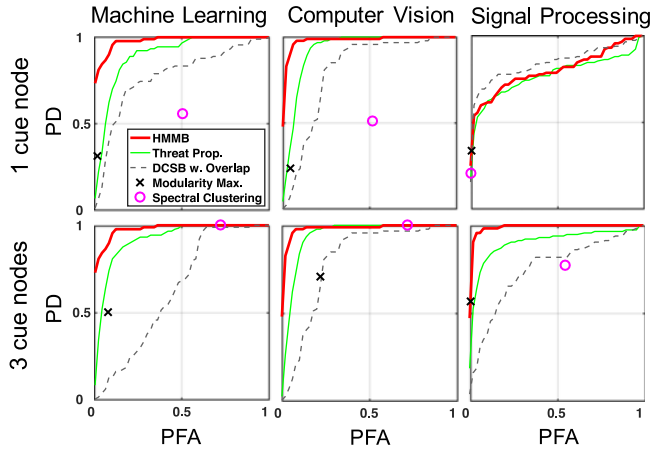

Fig. 12. Detection results on the DBLP co-authorship network in Fig. 11, for the HMMB and the four baseline approaches: threat propagation [14], DCSB with overlapping groups [8], modularity maximization [48], and spectral clustering [12]. For clarity and parsimony, the median performance of all the runs is reported.

given the network in Fig. 11 without the node color. It is reasonable to fit this network to the HMMB and spectral clustering with four communities ($K = 4$) because there are 4 visible clusters in the uncolored layout: two sparse clusters at the top and to the right, and two dense clusters on the bottom left. The DCSB with overlapping groups and the modularity maximization approaches infer the optimal number of communities, and threat propagation does not model the number of communities.

Fig. 12 summarizes the detection result using a 2-by-3 ROC matrix for having 1 and 3 cue nodes on each of the 3 conference fields being the community of interest. For each of the 6 setups, 1000 detection runs are produced from independent draws of the cue nodes. For clarity and parsimony, the median performance of all the runs is plotted in the figure.

HMMB outperforms the other approaches in all 6 setups except when detecting the signal processing community with only one cue node. Because the signal processing community appears to consist of two sub-clusters, the HMMB splits the signal processing community between two different community labels when fitting with four communities ($K = 4$). When only one cue node is available, only the nodes that are in the same sub-cluster as the cue node is detected before many false alarms occur. Having more cue nodes mitigates

this problem because of the higher chance that at least one cue node falls in each of the sub-clusters, as seen in the ROC for 3 cue nodes on the signal processing community. If the true number of communities is known a priori ($K = 3$), the HMMB ROC improve to the 3-cue-node curves seen in Fig. 12, regardless of the number of cue nodes. The DCSB with overlapping group has the best performance on the signal processing community with one cue node. This is due to its ability to correctly identify the optimal number of communities to be 3, the true number of communities, avoiding splitting up the signal processing community. However, its performance overall suffers from not being able to reliably estimate the mixed-memberships under this level of community overlap. Neither spectral clustering nor modularity maximization perform well at this level of community overlap. However, modularity maximization is able to detect some of the target nodes with few false alarms, by correctly identifying some tightly connected clusters within each of the three communities. Threat propagation offers a reasonable compromise with much lower computational requirement, for not explicitly modeling the community memberships, and its performances improves with more cue nodes.

The results here demonstrate HMMB's advantage over simpler models and methods that depend on modularity, in real-world networks with overlapping communities. The advantage of the HMMB approach here is not as dominant as the result on simulated networks, likely due to model misspecification. In practice, the success of HMMB for community detection on real-world networks will vary depending on how well the model describes the underlying network data.

5 Conclusions

This paper presents the hybrid mixed-membership blockmodel, that incorporates key characteristics of the real-world networks: community-based interactions with mixed-membership, node degree variation, and sparsity. Bayesian inference procedures are developed and shown to estimate the model parameters well on simulated networks over a range of realistic settings. These parameter estimates give insight to the formation of the network and reveal information on each individual that is useful in real-world applications. One such application is community detection. In the presence of community overlap, the HMMB approach is shown in realistic settings to significantly outperform existing models and common community detection approaches which depend on much stronger within-community than between-community interactions (i.e. modularity), due to its ability to model and account for the mixed-memberships and between-community interactions. Future work includes applying and adapting the HMMB model and inference procedure to additional real-world problems, proposing a collection and sampling strategy to maximize the information content on the key parameters of interest, adding dynamics to some of the model parameters, deriving a criterion and inference procedure to converge with an optimal number of communities, and optimizing and adapting the inference procedure for large networks.

Acknowledgments

The authors would like to thank MIT Lincoln Laboratory for funding this research through Lincoln Scholarship and the CANDiD Line, the Office of Naval Research (ONR), and the National Science Foundation (NSF). This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. The Harvard portion of this work is also supported by the National Science Foundation (NSF) under Awards No. IIS-1149662 and IIS-1409177, and by the Office of Naval Research (ONR) under Awards No. N00014-14-1-0485 and N00014-17-1-2131. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the ASD(R&E), NSF, and ONR.

References

[1] P. Erdős and A. Rényi, "On the evolution of random graphs," *Magyar Tud. Akad. Mat. Kutató Int. Közl*, vol. 5, pp. 17–61, 1960.
[2] F. Chung and L. Lu, "The average distances in random graphs with given expected degrees," *Proc. Nat. Acad. Sci.*, vol. 99, no. 25, pp. 15 879–15 882, 2002.
[3] Y. J. Wang and G. Y. Wong, "Stochastic blockmodels for directed graphs," *J. Am. Statistical Assoc.*, vol. 82, no. 397, pp. 8–19, 1987.
[4] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, no. 1981–2014, 2008, Art. no. 3.
[5] H. A. Soufiani and E. Airoldi, "Graphlet decomposition of a weighted network," in *Proc. 15th Int. Conf. Artif. Intell. Statist.*, Apr. 21–23, 2012, pp. 54–63.
[6] J. Yang and J. Leskovec, "Community-affiliation graph model for overlapping network community detection," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 1170–1175.
[7] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E*, vol. 83, no. 1, 2011, Art. no. 016107.
[8] T. P. Peixoto, "Model selection and hypothesis testing for large-scale network models with overlapping groups," *Phys. Rev. X*, vol. 5, Mar. 2015, Art. no. 011033.
[9] A. Athreya, D. E. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe, "Statistical inference on random dot product graphs: A survey," *arXiv: 1709.05454*, 2017.
[10] L. Lovász and B. Szegedy, "Limits of dense graph sequences," *J. Combinatorial Theory, Series B*, vol. 96, no. 6, pp. 933–957, 2006.
[11] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3, pp. 75–174, 2010.
[12] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, no. 3, 2006, Art. no. 036104.
[13] M. E. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.
[14] S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, and S. Philips, "Bayesian discovery of threat networks," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5324–5338, Oct. 2014.
[15] F. Huang, U. Niranjan, M. U. Hakeem, and A. Anandkumar, "Fast detection of overlapping communities via online tensor methods," *arXiv:1309.0787*, 2013.
[16] B. Ball, B. Karrer, and M. E. Newman, "An efficient and principled method for detecting communities in networks," *Phys. Rev. E*, vol. 84, 2011, Art. no. 036103.
[17] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, 2006.
[18] J. Diesner and K. M. Carley, "Exploration of communication networks from the enron email corpus," presented at the *SIAM Int. Conf. Data Mining: Workshop Link Anal. Counterterrorism Secur.*, Newport Beach, CA, USA, 2005.
[19] S. T. Smith, K. D. Senne, S. Phiips, E. K. Kao, and G. Bernstein, "Covert network detection," *Lincoln Laboratory J.*, vol. 20, pp. 47–61, 2013.
[20] S. A. Golder, D. M. Wilkinson, and B. A. Huberman, "Rhythms of social interaction: Messaging within a massive online network," in *Communities and Technologies*. New York, NY, USA: Springer, 2007, pp. 41–66.

[21] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 65–74.
[22] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 10–17.
[23] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, "Context-aware small cell networks: How social metrics improve wireless resource allocation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5927–5940, Nov. 2015.
[24] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Inf. Process. Manage.*, vol. 41, no. 6, pp. 1462–1480, 2005.
[25] J. H. Fowler, "Connecting the congress: A study of cosponsorship networks," *Political Anal.*, vol. 14, no. 4, pp. 456–487, 2006.
[26] A.-L. Barabasi and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature Rev. Genetics*, vol. 5, no. 2, pp. 101–113, 2004.
[27] W. Aiello, F. Chung, and L. Lu, "A random graph model for power law graphs," *Exp. Math.*, vol. 10, no. 1, pp. 53–66, 2001.
[28] E. M. Airoldi, X. Wang, and X. Lin, "Multi-way blockmodels for analyzing coordinated high-dimensional responses," *Ann. Appl. Statist.*, vol. 7, no. 4, 2013, Art. no. 2431.
[29] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.
[30] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer, 2006.
[31] W. Fu, L. Song, and E. P. Xing, "Dynamic mixed membership blockmodel for evolving networks," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 329–336.
[32] T. P. Peixoto, "The graph-tool python library," 2014. [Online]. Available: http://graph-tool.skewed.de
[33] B. P. Olding and P. J. Wolfe, "Inference for graphs and networks: Extending classical tools to modern data," *arXiv:0906.4980*, 2009.
[34] P. O. Perry and P. J. Wolfe, "Null models for network data," *arXiv:1201.5871*, 2012.
[35] M.-L. G. Buot, S. Hoşten, and D. St. P. Richards, "Counting and locating the solutions of polynomial systems of maximum likelihood equations, II: The Behrens-Fisher problem," *Statistica Sinica*, vol. 17, pp. 1343–1354, 2007.
[36] I. Mukherjee and D. M. Blei, "Relative performance guarantees for approximate inference in latent dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1129–1136.
[37] X. Nguyen, "Posterior contraction of the population polytope in finite admixture models," *Bernoulli*, vol. 21, no. 1, pp. 618–646, 2015.
[38] E. K. Kao, E. M. Airoldi, and D. B. Rubin, "Causal inference under network interference: A network potential outcome framework with Bayesian imputation," in preparation.
[39] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, vol. 2. New York, NY, USA: Taylor & Francis, 2014.
[40] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, no. 1/2, pp. 5–43, 2003.
[41] G. C. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Am. Statistical Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.
[42] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proc. Nat. Acad. Sci.*, vol. 110, no. 36, pp. 14 534–14 539, 2013.
[43] T. P. Peixoto, "Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models," *Phys. Rev. E*, vol. 89, Jan. 2014, Art. no. 012804.
[44] N. Arcolano, K. Ni, B. A. Miller, N. T. Bliss, and P. J. Wolfe, "Moments of parameter estimates for chung-lu random graph models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 3961–3964.
[45] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nature Methods*, vol. 7, no. 12, pp. 1009–1015, 2010.
[46] A. W. Van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[47] J. K. Blitzstein and J. Hwang, *Introduction to Probability*. Boca Raton, FL, USA: CRC Press, 2014.
[48] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statistical Mech.: Theory Exp.*, vol. 2008, no. 10, 2008, pp. P10008.
[49] L. Lovász, "Random walks on graphs," *Combinatorics, Paul Erdos Is Eighty*, vol. 2, pp. 1–46, 1993.
[50] F. Chung and W. Zhao, "Pagerank and random walks on graphs," in *Fete of Combinatorics and Computer Science*. New York, NY, USA: Springer, 2010, pp. 43–62.
[51] M. Ley, "DBLP computer science bibliography," 2005.

Edward K. Kao (M'03) received the BSc degree in electrical engineering and computer science from Carnegie Mellon University, in 2003, the MSc degree in electrical engineering from Johns Hopkins University, in 2006, and the PhD degree in statistics from Harvard University, in 2017. He is a staff member with the Intelligence and Decision Technologies Group, MIT Lincoln Laboratory. For the last 10 years, he has been working on research topics in the exploitation of graph and network data, where actionable intelligence is derived from interactions and relationships between entities. These topics include threat network detection and analysis, causal experiments on social networks, statistical models for networks, high performance computing for large networks, and information content, optimal sampling, and experimental design for network inference. He is the recipient of the MIT Lincoln Scholarship and the National Defense Science & Engineering Graduate Fellowship, from 2011 to 2016. He also received in 2013 the Tom R. Ten Have Research Award for his work in causal inference on networks, and in 2015 the Google Best Poster Award at the New England Statistical Symposium. He is a member of the IEEE.

Steven Thomas Smith (M'86-SM'04-F'17) received the BASc degree in electrical engineering and honours mathematics from the University of British Columbia, Vancouver, BC, in 1986 and the PhD degree in applied mathematics from Harvard University, Cambridge, Massachusetts, in 1993. He is an expert in radar, sonar, and signal processing who has made pioneering and wide ranging contributions through his research and technical leadership in estimation theory, resolution limits, and signal processing and optimization on manifolds. He has more than 20 years experience as an innovative technology leader with statistical data analytics, both theory and practice, and broad leadership experience ranging from first-of-a-kind algorithm development for groundbreaking sensor systems to graph-based intelligence architectures. His contributions span diverse applications from optimum network detection, geometric optimization, geometric acoustics, superresolution, and nonlinear parameter estimation. He is a recipient of the SIAM Outstanding Paper Award, in 2001, the IEEE Signal Processing Society Best Paper Award, in 2010, and has been an invited speaker as an original inventor of some of the big advances in signal processing over the last decade. He was an associate editor of the *IEEE Transactions on Signal Processing* from 2000 to 2002, and serves on the IEEE Sensor Array and Multichannel and Big Data committees. He has taught signal processing courses at Harvard, MIT, and for the IEEE. He is a senior staff member with MIT Lincoln Laboratory. He is a fellow of the IEEE.

Edoardo M. Airoldi received the BSc degree in mathematical statistics and economics from Bocconi University, the PhD degree in computer science, and the MSc degree in statistics from Carnegie Mellon University. He is an associate professor of statistics with Harvard University, where he founded and directed the Harvard Laboratory for Applied Statistical Methodology & Data Science from 2009 to 2017. His current research interests include statistical theory and methods for designing and analyzing experiments on large networks, and, more generally, on modeling and inferential issues that arise in analyses that leverage network data. His work has appeared in journals across statistics, computer science and engineering, including the *Annals of Statistics*, the *Journal of the American Statistical Association*, the *Journal of Machine Learning Research*, the *Proceedings of the National Academy of Sciences*, and the *Nature*. He has received several research awards including an Alfred Sloan Research Fellowship and a Shutzer Fellowship from the Radcliffe Institute of Advanced Studies. He delivered an IMS Medallion Lecture at the Joint Statistical Meetings, in 2017.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.