Infinite-degree-corrected stochastic block model

Tue Herlau,[*] Mikkel N. Schmidt,[†] and Morten Mørup[‡]
Section for Cognitive Systems, DTU Compute
Technical University of Denmark,

In Stochastic blockmodels, which are among the most prominent statistical models for cluster analysis of complex networks, clusters are defined as groups of nodes with statistically similar link probabilities within and between groups. A recent extension by Karrer and Newman incorporates a node degree correction to model degree heterogeneity within each group. Although this demonstrably leads to better performance on several networks it is not obvious whether modelling node degree is always appropriate or necessary. We formulate the degree corrected stochastic blockmodel as a non-parametric Bayesian model, incorporating a parameter to control the amount of degree correction which can then be inferred from data. Additionally, our formulation yields principled ways of inferring the number of groups as well as predicting missing links in the network which can be used to quantify the model's predictive performance. On synthetic data we demonstrate that including the degree correction yields better performance both on recovering the true group structure and predicting missing links when degree heterogeneity is present, whereas performance is on par for data with no degree heterogeneity within clusters. On seven real networks (with no ground truth group structure available) we show that predictive performance is about equal whether or not degree correction is included; however, for some networks significantly fewer clusters are discovered when correcting for degree indicating that the data can be more compactly explained by clusters of heterogenous degree nodes.

I. INTRODUCTION

The stochastic blockmodel (SBM) [1–3] has become a prominent tool for modeling group structure in complex networks [4]. However, as pointed out by Karrer and Newman [5], the stochastic blockmodel has a tendency to group nodes according to their degree such that high degree nodes group together even though their patterns of interactions with the remaining network may differ. This grouping thus reflects aspects of node degree rather than overall statistical patterns in the network. To alleviate this issue, Karrer and Newman introduced the degree corrected stochastic blockmodel (DCSBM) [5]. In their model, additional parameters modeling node degree heterogeneity are introduced allowing nodes of varying degree to be clustered together, and they demonstrate that including this degree correction reduces the tendency to group nodes according to their degree distribution [5] . The parameters in the DCSBM model are inferred using maximum likelihood (ML) estimation and since closed form expressions for the ML estimates of the additional degree correction parameters are available, the computational complexity of the inference procedure is similar to inference in the SBM.

Although Karrer and Newman demonstrate on several network datasets that degree correction leads to better performance [5], it is not obvious whether including a degree correction is always appropriate on real network data. Furthermore, the number of groups used in the analysis is likely to influence the results since groups of heterogenous node degree can be reasonably modelled by a number of homogenous subgroups. Not handling this issue in a principled manner could potentially confound the results. Finally, an important subject of network modelling is validation. Although many real networks are hypothesized to possess group structure, no ground truth clustering is available which makes it difficult to assess the goodness of the obtained clustering. A popular alternative is to measure the predictive performance on held out links in the network. In order to do this in a principled manner the methods must be able to handle missing entries in the network data as well as define a predictive distribution over the missing entries.

In this paper we address these three important challenges when modeling network data by the DCSBM:

- Can we infer the extent in which degree correction is necessary?
- How can we determine the number of components?
- How can we predict links in the DCSBM?

In particular, we formulate a non-parametric Bayesian generative model for the DCSBM. The number of components are inferred using the Chinese Restaurant Process which has previously been used to determine the number of components in stochastic blockmodels [6, 7]. Our generative model is characterized by admitting a simple inference procedure in which both the degree parameter and group interactions can be analytically marginalized out such that inference reduces to estimating the assignments of nodes to clusters as for the DCSBM. We address the link-prediction problem using Markov chain Monte Carlo (MCMC) imputation. By infering the hyper-parameter in the prior distribution of the parameters that account for heterogenous node degree our model is able to learn the extent to which a degree correction is necessary, possibly reducing to an uncorrected stochastic blockmodel. On synthetic as well as seven real networks we demonstrate the utility of our proposed model for determining the number of components, link-prediction, and inferring the magnitude of the parameter controlling degree correction.

[*] tuhe@dtu.dk
[†] mnsc@dtu.dk
[‡] mmor@dtu.dk

Past work on the SBM and DCSBM has not treated the problem of inferring components, presence of degree heterogeneity and link prediction under one unified framework. Although Bayesian approaches to inferring components and link prediction has a long history for the SBM [4, 6, 7], most work on the DCSBM has been focused on other inference methods. As noted, Karrer and Newman [5] treated the problem of inference in the DCSBM from a ML perspective. A related approach was taken by Peixoto [8] who considered degree-correction as constraints on a blockmodel ensemble and derived an entropy-based cost function. For the SBM, a method relying on a minimum description length based approach to learning has been proposed giving rise to an efficient maximization procedure [9]. The MDL approach by Rosvall et al. [10] allows degree correction but is otherwise analytically different from the DCSBM. For the DCSBM minimum-description length based procedures was considered by Peixoto [11] to give an efficient MCMC-based inference procedure, see also [12] for additional discussion of this approach and an application to the problem of estimating the number of components. The belief propagation method of Decelle et al. [13, 14] may also be applied to the DCSBM. More related to our approach is that of Yan et al. [15] who consider the problem of inferring the number of groups in the DCSBM from a model-selection perspective.

While these approaches represent important contributions to the problem of jointly modelling degree heterogeneity and block structure, none of the current proposals are based on a Baysian generative model and allow joint inference of degree-correction, number of components and missing links using a MCMC-based approach.

II. METHODS

Let $\boldsymbol{A}$ be the adjacency matrix of an undirected observed network of n nodes such that A_{ij} is the number of links between node i and j. We allow a positive number of self-links A_{ii} in our model definition (note that in the original formulation of DCSBM [5] A_{ii} is defined as twice the number of self-links). The DSCBM model [5] for an undirected graph assumes that the links between nodes i and j follow a Poisson distribution

$$\text{for } i \neq j\text{: } A_{ij} \sim \mathcal{P}\big(\theta_i \eta_{z_i z_j} \theta_j\big). \tag{1}$$

The parameter $\eta_{\ell m}$ controls the probability of links between nodes in group ℓ and m, $z_i = \ell$ indicate node i is assigned to group ℓ and θ_i is a node specific parameter that regulates this link probability and thus accounts for heterogenous node degrees. The model is subject to the constraint that $\sum_i \delta_{z_i \ell} \theta_i = 1$ for all groups ℓ, i.e. the sum of the θ_i within each group is one.

We presently propose a non-parametric Bayesian generative model that extends the DCSBM dubbed the Infinite Degree Corrected Stochastic Blockmodel (IDCSBM). Like the DCSBM we also maintain node weights θ_i to control the degree, however, to arrive at a Bayesian formulation we assume the weights within each group are drawn from a Dirichlet distribution. More precisely, for each group ℓ containing n_ℓ nodes, we introduce a n_ℓ-dimensional vector of weights $(\phi_i)_{z_i=\ell}$ drawn from a Dirichlet distribution and define $\theta_i = n_\ell \phi_i$ in eq. (1).

The scaling by n_ℓ makes the average degree of any given node independent on the size of the group the node belongs to. The full model now consists of (i) generating a random partition, (ii) generating the interaction between each group of the partition $\eta_{\ell m}$ from a gamma distribution, (iii) for each group, generate $(\phi_i)_{z_i=\ell}$ from a Dirichlet distribution and rescale with n_ℓ, and finally (iv) use eq. (1) to generate the number of links A_{ij} between node $i \neq j$.

The full model is given generatively below. The symbols $\mathcal{D}$ denote the Dirichlet distribution and $\mathcal{G}$ the gamma distribution. For analytical convenience the model assumes a particular parametrization of the self-links A_{ii}, a point we will return to later.

$$\begin{array}{lrll}
 & \boldsymbol{z} \sim \mathrm{CRP}(\alpha), & \textit{clusters,} & (2)\\
\text{for } \ell \geq 0 & (\phi_i)_{z_i=\ell} \sim \mathcal{D}(\gamma \mathbf{1}_{(n_\ell)}) & & \\
 & \theta_i = n_{z_i} \phi_i, & \textit{relative degree,} & (3)\\
\text{for } \ell \leq m & \eta_{\ell m} \sim \mathcal{G}(\kappa, \lambda), & \textit{link rate,} & (4)\\
\text{for } i < j & A_{ij} \sim \mathcal{P}(\theta_i \eta_{z_i z_j} \theta_j), & \textit{link weight,} & (5)\\
\text{for } i = j & A_{ii} \sim \mathcal{P}\left(\dfrac{1}{2}\theta_i^2 \eta_{z_i z_i}\right). & &
\end{array}$$

In the above $\mathbf{1}_{(n_\ell)}$ is a vector of ones with length n_ℓ, $N = \sum_{\ell=1}^{L} n_\ell$ is the total number of nodes and L is the number of groups. As a prior over the node partitions $\boldsymbol{z}$ we use the Chinese Restaurant Process (CRP) parameterized by a single parameter α controlling the distribution of group size [16]. A potential advantage of the CRP over for instance a uniform prior over partitions is that the CRP is consistent under projections whereas the uniform prior is not. The simplest example is the case where $\boldsymbol{z}$ is a partition of two nodes assigned to the same group (i.e. $z_1 = z_2 = 1$) and we consider a partition obtained by including a third node. In this case for the CRP it holds: $p(z_1 = z_2 = 1|\alpha) = p(z_1 = z_2 = 1, z_3 = 1|\alpha) + p(z_1 = z_2 = 1, z_3 = 2|\alpha)$, however for the uniform prior the left-hand side is $\frac{1}{2}$ and the right-hand side $\frac{2}{5}$.

Notice the role played by γ in the Dirichlet distribution in eq. (3). If $\gamma \to \infty$, we will have $\phi_i \to \frac{1}{n_\ell}$ for $z_i = \ell$ or simply $\theta_i \to 1$ for all i (the limits are understood in distribution) and the model is thus independent of degree in eq. (1). On the other hand, for $\gamma \to 0$, within each group ℓ a single node, i^*, will have mass $\theta_{i^*} = n_\ell$ and the network become very nearly entirely dominated by a few greedy nodes. We return to the properties of the model in section II B. The advantage of a Bayesian formulation is that we can not only infer θ_i, but also a distribution of the degree-correction variable γ representing the appropriateness of modelling degree heterogeneity for the network.

Collecting variables of the same type the joint density factorizes as:

$$\begin{aligned}
&p(\boldsymbol{A}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{z}|\alpha, \gamma, \kappa, \lambda)\\
&\quad = p(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{z}) p(\boldsymbol{\eta}|\kappa, \lambda) p(\boldsymbol{\phi}|\boldsymbol{z}, \gamma) p(\boldsymbol{z}|\alpha).
\end{aligned} \tag{6}$$

The model thus depend on parameters $(\alpha, \gamma, \kappa, \lambda)$. While one could fix these at a particular value, a more principled approach we have taken is to introduce vague non-informative priors and sample these as well [17]. Either choice has no effect on the following derivation below. In our notation the relevant densities are

$$p(\boldsymbol{z}|\alpha) = \frac{\alpha^L \Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{\ell=1}^{L} \Gamma(n_\ell) \quad \textit{(Chinese retaurant process)} \tag{7}$$

$$\mathcal{D}(\boldsymbol{x}|\boldsymbol{\gamma}) = \frac{1}{B(\boldsymbol{\gamma})} \prod_i x_i^{\gamma_i - 1},\; B(\boldsymbol{\gamma}) = \frac{\prod_i \Gamma(\gamma_i)}{\Gamma(\sum_i \gamma_i)}, \tag{8}$$

$$\mathcal{G}(x|\kappa,\lambda) = \frac{1}{G(\kappa,\lambda)} x^{\kappa-1} e^{-\lambda x},\; G(\kappa,\lambda) = \lambda^{-\kappa}\Gamma(\kappa). \tag{9}$$

The advantage of the present formulation is the use of the Dirichlet distribution within each group, and the particular parametrization of A_{ii}, that allow the node weights as well as group interactions to be integrated out analytically. To see this we introduce the short-hand notation for between and within-group link counts

$$N^+_{\ell m} = \begin{cases} \sum_{\substack{i:z_i=\ell,\\ j:z_j=m}} A_{ij} & \ell \neq m \\ \sum_{\substack{i \le j:\\ z_i=z_j=\ell}} A_{ij} & \ell = m \end{cases},\; N_{\ell m} = \begin{cases} n_\ell n_m & \ell \neq m \\ \frac{n_\ell n_\ell}{2} & \ell = m \end{cases}. \tag{10}$$

as well as node degrees $k_i = \sum_j A_{ij}$ and $\hat{k}_i = k_i + A_{ii}$. It now follows by some algebra

$$\begin{aligned} p(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{z}) &= \prod_{i<j} \frac{(\theta_i \eta_{z_i z_j} \theta_j)^{A_{ij}}}{A_{ij}! e^{\theta_i \eta_{z_i z_j} \theta_j}} \prod_i \frac{\left(\frac{\theta_i^2 \eta_{z_i z_i}}{2}\right)^{A_{ii}}}{A_{ii}! e^{\frac{1}{2}\theta_i^2 \eta_{z_i z_i}}} \\ &= \frac{\prod_i 2^{-A_{ii}}}{\prod_{i \le j} A_{ij}!} \prod_{\ell \le m} \eta_{\ell m}^{N^+_{\ell m}} e^{-\eta_{\ell m} N_{\ell m}} \prod_i \theta_i^{k_i + A_{ii}} \\ &= \frac{\prod_i 2^{-A_{ii}}}{\prod_{i \le j} A_{ij}!} \left[\prod_{\ell \le m} \eta_{\ell m}^{N^+_{\ell m}} e^{-\eta_{\ell m} N_{\ell m}}\right] \prod_\ell n_\ell^{\hat{k}_\ell} \prod_{i:z_i=\ell} \phi_i^{\hat{k}_i}. \end{aligned} \tag{11}$$

$$p(\boldsymbol{\eta}|\kappa,\lambda) = \prod_{\ell \le m} \frac{1}{G(\kappa,\lambda)} \eta_{\ell m}^{\kappa-1} e^{-\eta_{\ell m}\lambda}, \tag{12}$$

$$p(\boldsymbol{\phi}|\boldsymbol{z},\gamma) = \prod_\ell \frac{1}{B(\gamma \mathbf{1}_{(n_\ell)})} \frac{\prod_{i:z_i=\ell} \Gamma(\gamma)(\frac{\theta_i}{n_\ell})^{\gamma-1}}{n_\ell \Gamma(n_\ell \gamma)}. \tag{13}$$

Inserting into eq. (6), collecting terms and exploiting the conjugacy of the Dirichlet and Gamma distributions to the Poisson distribution we can analytically marginalize (i.e., collapse) $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ to obtain

$$\begin{aligned} p(\boldsymbol{A},\boldsymbol{z}|\alpha,\gamma,\kappa,\lambda) &= \int d\boldsymbol{\eta} d\boldsymbol{\phi}\; p(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{z}) p(\boldsymbol{\eta}|\kappa,\lambda) p(\boldsymbol{\phi}|\boldsymbol{z},\gamma) p(\boldsymbol{z}|\alpha) \\ &= \frac{1}{\prod_{i \le j} A_{ij}! \prod_i 2^{A_{ii}}} \prod_{\ell \le m} \frac{G\left(N^+_{\ell m} + \kappa, N_{\ell m} + \lambda\right)}{G(\kappa,\lambda)} \\ &\times \left[\prod_\ell \frac{B\left(\gamma \mathbf{1}_{(n_\ell)} + (\hat{k}_i)_{i:z_i=\ell}\right)}{B(\gamma \mathbf{1}_{(n_\ell)})} n_\ell^{\hat{k}_\ell}\right] \frac{\alpha^L \Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{\ell=1}^{L} \Gamma(n_\ell). \end{aligned} \tag{14}$$

In the above derivation we exploit that $\sum_{z_i=\ell} \theta_i = n_\ell$ and thus the derivation requires access to the entire network. As a result, the inference of our generative model is reduced to determining the posterior distribution of the assignment of nodes to groups, $\boldsymbol{z}$.

The assignment matrix $\boldsymbol{z}$ is inferred using standard Gibbs sampling [6], and using the Bayesian framework we can treat the hyperparameters γ, α, λ and κ as random variables. In particular, we will invoke the non-informative prior $p(x) \propto x^{-1}$ for all four parameters and infer them using random-walk Metropolis updates of the form $x^* = \exp(\log x + z)$, $z \sim N(0, \sigma = 0.1)$. For each Gibbs sweep over $\boldsymbol{z}$, we performed 20 Metropolis-Hastings updates of the hyperparameters. While Metropolis-Hastings with random proposals is not very computational efficient, we noticed throughout the experiments this step had a small computational cost compared to sampling $\boldsymbol{z}$.

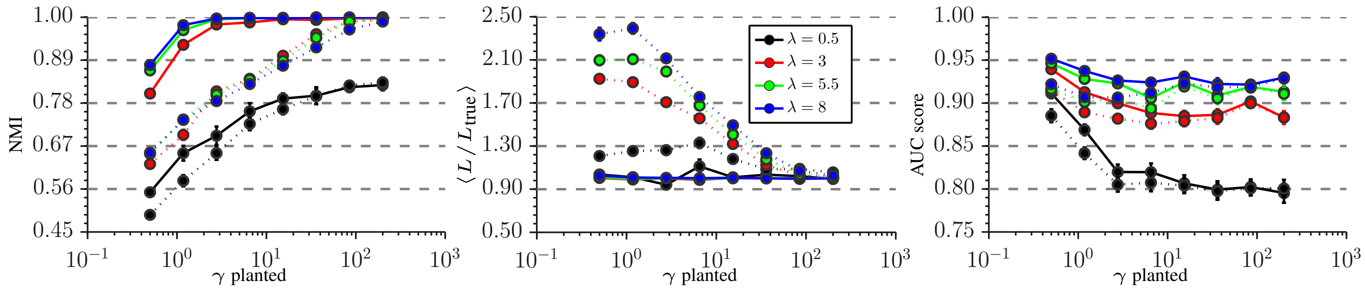
A. Imputation and link prediction

Missing (unobserved) links commonly occur in network and predicting missing links is an important goal of network modelling. Comparing the prediction of a model on unobserved data to the actual value is furthermore a popular way to validate a model. In addition the self-links A_{ii} are often unknown or, if the network cannot contain self-links such as the case of a friendship network, they should be treated as auxiliary variables that are integrated out.

For the IDCSBM the (marginalized) expression for $\boldsymbol{z}$ in eq. (14) requires access to all entries in the adjacency matrix and so it is not possible to marginalize over missing data simply by ignoring the corresponding terms in the likelihood function. To overcome this difficulty we marginalize over missing entries by formulating a Markov chain Monte Carlo algorithm jointly over the parameters and the missing links. This is done by sampling $\boldsymbol{z}$ and the hyperparameters using Gibbs sampling and random-walk Metropolis Hastings, and then conditionally on $\boldsymbol{A}$ and $\boldsymbol{z}$ drawing values of $\eta_{\ell m}$ and $(\phi_i)_i$ conditional on the full matrix $\boldsymbol{A}$ and assignments $\boldsymbol{z}$ and conditionally on these values draw the values of $\boldsymbol{A}$ corresponding to the missing links from the Poisson distribution eq. (5). This corresponds to imputing the missing values from their predictive distribution in each step of the MCMC algorithm and, assuming convergence of the Markov chain, is equivalent to marginalizing out the missing links. We use this framework both to handle self-links but also for link prediction in general. Another popular method to predict missing data is simply replacing missing entries of $\boldsymbol{A}$ with 0 [4, 5, 18], however as the diagonal of $\boldsymbol{A}$ is often fully missing, and the poisson rate for A_{ii} is proportional to θ_i^2, this approach would create an undesirable bias for θ_i.

B. Properties of the model

An important property of the model is that it can accurately learn the degree distribution of the data and the link-density



(a)

(b)

(c)

FIG. 1: (Color online) IDCSBM and ISBM results on simulated networks. The plots show the normalized mutual information (NMI), the ratio of estimated to true number of components L_{frac} as well as the area under curve (AUC) of the receiver operator characteristics as computed by running the proposed methods on networks produced from the generative model of the IDCSBM with different values of λ and γ. The fully drawn lines indicate results for the IDCSBM, dotted lines indicate results for ISBM.

between the groups. Suppose $\boldsymbol{A}_0$ is an observed network and let $\boldsymbol{z}$ be any fixed cluster. Conditional on $\boldsymbol{A}_0$ and $\boldsymbol{z}$ we may compute the posterior over $\boldsymbol{\eta}$, $\boldsymbol{\theta}$ and check if these distributions accurately reflect relevant properties of $\boldsymbol{A}_0$. First notice from eq. (11) the posterior distributions of $\boldsymbol{\eta}, \boldsymbol{\theta}$ are

$$p(\eta_{\ell m}|\boldsymbol{A}_0,\boldsymbol{z}) = \mathcal{G}(\eta_{\ell m} \mid N^+_{\ell m}+\kappa, N_{\ell m}+\lambda) \tag{15}$$

$$p\left(\left(\frac{\theta_i}{n_\ell}\right)_{z_i=\ell}|\boldsymbol{A}_0,\boldsymbol{z}\right) = \mathcal{D}\left(\left(\frac{\theta_i}{n_\ell}\right)_{z_i=\ell} \mid \gamma\mathbf{1}_{n_\ell}+(\hat{k}_i)_{z_i=\ell}\right) \tag{16}$$

Recall for two Poisson distributed random variables $X \sim \mathcal{P}(a)$, $Y \sim \mathcal{P}(b)$ their sum is Poisson with rate $a+b$: $X+Y \sim \mathcal{P}(a+b)$. This, along with the derivation eq. (11), allows us to compute various properties of the model.

First consider the total interaction strength between two groups ℓ and m. The interaction $\sum_{i\leq j}\delta_{z_i=\ell}\delta_{z_j=m}A_{ij}$, considered as a random variable, is then distributed as $\mathcal{P}(\eta_{\ell m}N_{\ell m})$. If $X \sim \mathcal{P}(\lambda)$ then $\mathbb{E}[X]=\lambda$ and so the average between-group interaction is (the expectation is with respect to $p(\cdot|\boldsymbol{A}_0,\boldsymbol{z})$)

$$\mathbb{E}\left[\sum_{i\leq j}\delta_{\substack{z_i=\ell,\\ z_j=m}}A_{ij}\right] = \mathbb{E}\left[N_{\ell m}\eta_{\ell m}\right] = \frac{N_{\ell m}(N^+_{\ell m}+\kappa)}{N_{\ell m}+\lambda}. \tag{17}$$

For analytical simplicity, we will consider the degree plus the diagonal element. To this end define the degree of node i as $d_i = \sum_j A_{ij} + A_{ii}$. Since each A_{ij} is Poisson distributed the degree too is a Poisson random variable. If $z_i = \ell$ then d_i's distribution is given by

$$d_i \sim \mathcal{P}\left(\sum_{j\neq i}\theta_i\eta_{\ell z_j}\theta_j + 2\frac{\theta_i^2\eta_{\ell\ell}}{2}\right) = \mathcal{P}\left(\theta_i\sum_m \eta_{\ell m}n_m\right). \tag{18}$$

We may now compute the average, again with respect to $\boldsymbol{A}_0$ and fixed $\boldsymbol{z}$:

$$\begin{aligned}
\mathbb{E}[d_i] &= \mathbb{E}\left[\theta_i\sum_m \eta_{\ell m}n_m\right] \\
&= n_\ell\frac{\hat{k}_i+\gamma}{\sum_{j:z_j=\ell}\hat{k}_j+\gamma n_\ell}\sum_m\frac{N^+_{\ell m}+\kappa}{N_{\ell m}+\lambda}n_m \\
&= (\hat{k}_i+\gamma)\sum_m\frac{N_{\ell m}2^{\delta_{\ell m}}}{N_{\ell m}+\lambda}\frac{N^+_{\ell m}+\kappa}{\sum_h N^+_{\ell h}2^{\delta_{\ell h}}+\gamma n_\ell}.
\end{aligned} \tag{19}$$

Assuming the groups are fairly large, and in the low limit of the prior γ, the sum will be 1 to first order. The derivations eq. (17) and (19) show in the limit of large systems the relative influence of the prior terms will vanish and the model will accurately capture the between-group link density as well as the node degree.

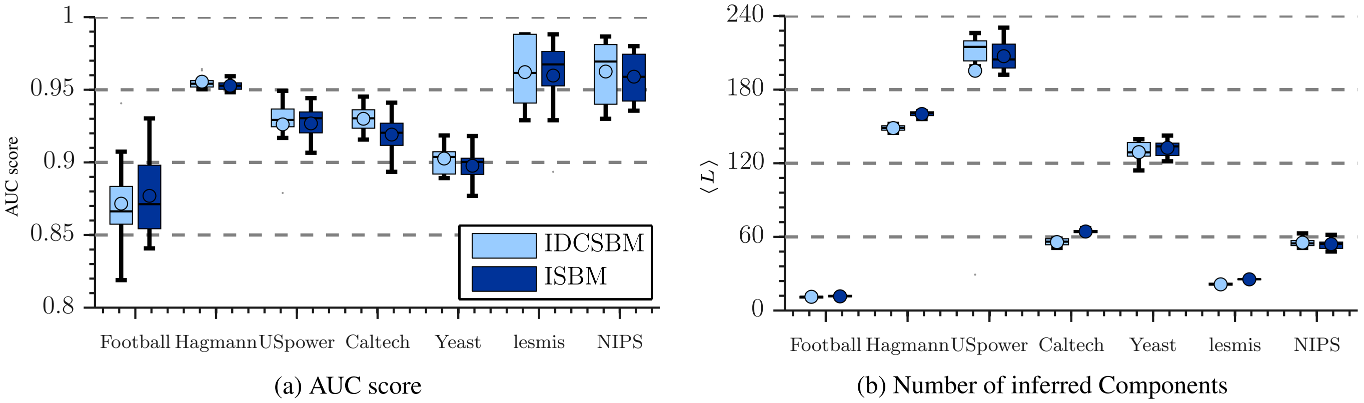
III. RESULTS AND DISCUSSIONS

We analyze synthetic datasets generated from our model as well as seven real networks from the literature.

A. Synthetic data

In our synthetic simulation studies we generated networks of $N = 80$ nodes from our generative model with the parameters κ and α fixed at $\kappa = 0.5$ and $\alpha = 4$ and under different values of λ and γ.

Each such network was analyzed using out Infinite Degree Corrected Stochastic Block Model (IDCSBM) as well as the corresponding infinite SBM (ISBM) without degree correction. In figure 1 the normalized mutual information (NMI), the ratio of true number of components to estimated number of components $L_{\text{frac}} = \langle\frac{L}{L_{\text{true}}}\rangle$ and the area under curve (AUC) of the receiver operator characteristic are given (error bars indicate standard deviation of the mean where the deviation is computed over 10 restarts of the sampler). In the analysis we ran the samplers for 1000 iterations and discarded the first half



(a) AUC score

(b) Number of inferred Components

FIG. 2: (Color online) IDCSBM and ISBM results on the seven real network. To the left is shown AUC scores on held-out links and to the right the number of inferred groups L, results are averaged over 10 random restarts. The degree corrected and non-corrected method perform roughly similar with a tendency for the degree-corrected model to find fewer groups.

as burnin. The AUC scores were computed by treating 5% of the links and a similar number of non-links as missing.

From the plot of the NMIs we see that the degree corrected model (IDCSBM) better recovers the true generated group structure than the uncorrected model (ISBM) and as expected the performance of the two methods converge as γ increases corresponding to networks which does not exhibit degree heterogeneity. Furthermore, the IDCSBM recovers the correct number of groups whereas the ISBM generates more than the true number of groups in order to account for the effect of a skewed degree distribution. The predictive performance as quantified by the AUC scores are more or less similar with a tendency of slightly better predictions for the IDCSBM. As expected this is most notable for small values of γ. We further observe that structure is better recovered when the contrast in the interactions are high as influenced by the values of λ. This too can be expected since very sparse networks presumably has little recoverable structure.

B. Real data

We analyzed the following seven networks

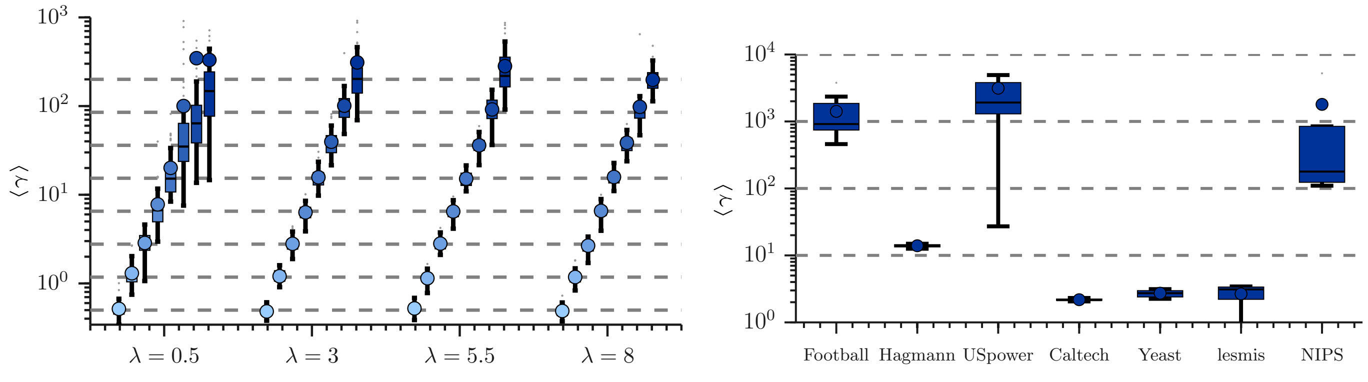
- *Football:* Undirected unweighted network of American football games between 115 Division IA colleges in the Fall 2000 [19].
- *Hagmann:* Undirected weighted network of the number of links between 998 brain regions as estimated by tractography from diffusion spectrum imaging across five subjects [20]. I.e., the graph of each subject has been symmetrized, thresholded at zero and the five subject graphs added together.
- *USPower:* Undirected unweighted network of 4941 nodes representing the topology of the Western States Power Grid of the United States compiled by [21].
- *Caltech:* The Caltech39 social network of 769 students from the Facebook100 dataset (available at `http://datahub.io/dataset/facebook100`).
- *Yeast:* The interaction network between 2361 proteins of yeast [22].
- *Lesmis:* Undirected and weighted graph of the coappearences of 77 characters in Les Miserables by Victor Hugo [23].
- *NIPS:* Undirected weighted network of the number of co-authorships between 234 authors of papers presented at the Neural Information Processing Systems 1-12 (available at `http://www.cs.nyu.edu/~roweis/data.html`).

In figure 2 is shown the results for the IDCSBM and the ISBM on the seven networks in terms of AUC score treating 5% of the links (and a similar number of non-links) as missing. Furthermore, the numbers of estimated components by the two models are given. The samplers were run for 1000 iterations (half discarded as burnin) and the results are averaged over 10 restarts.

From figure 2 it can be seen that in general the performance in predicting link as quantified by the AUC scores are on par for the IDCSBM and ISBM. However, as observed also in the synthetic study the IDCSBM model extracts less components than the ISBM for the Hagmann, Caltech, and Lesmis networks. Thus, the model allocates less groups when compared to the ISBM that allocates additional clusters in order to compensate for its lack of ability to explicitly account for degree.

Another way to examine this effect is to look at the degree distribution within each group. Since the groups have vastly different sizes it is hard to summarize this effect into a single number, however if we consider a fixed group structure $\boldsymbol{z}$ and a single group ℓ of size n_ℓ we may compute the empirical mean $\mathbb{E}[k_\ell] = \frac{1}{n_\ell}\sum_{i:z_i=\ell} k_i$ and standard deviation $\mathrm{std}[k_\ell] = \sqrt{\frac{1}{n_\ell}\sum_{i:z_i=\ell}(k_i - \mathbb{E}[k_\ell])^2}$ of the degree within this group.

In figure 4 we plotted the average of the empirical standard deviation of the degree distribution as a function of group size, that is, for each point (k, y) in figure 4, y is an estimate of



(a) Recovery of γ for artificial networks

(b) Inferred values of γ

FIG. 3: (Color online) Inferred values of $\langle\gamma\rangle$ for the artificial (left) and for the real (right) networks. The box plots show the inferred mean of γ for each of the 10 (or 50) MCMC chains (on artificial/real networks). For the artificial network (left), the networks are grouped according to the planted value of λ (controlling link density), and each of the eight boxes in a group corresponds to a planted value of γ, the planted values indicated by the horizontal lines. In the limit of good sampling the boxes should lie on the dotted lines. As shown, the sampler infer the correct value of degree-correction for the artificial networks except for very sparse networks ($\lambda = 0.5$). For the real networks the model infer very different degrees of node heterogeneity.

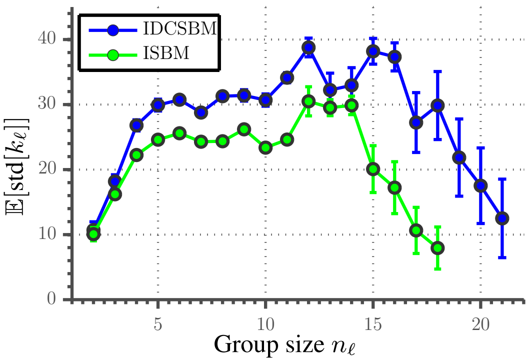


FIG. 4: (Color online) Variance of degree heterogeneity for the ISBM and IDCSBM for the Hagmann dataset. Each point (k, y) is an estimate of the standard deviation of the degree distribution for nodes in a group ℓ of size $n_\ell = k$, see main text for details.

$\mathbb{E}[\mathrm{std}[k_\ell]]$where the expectation is conditional on $n_\ell = k$. This quantity is easily estimated based on the last 500 states of a MCMC chain. The error bars are the standard deviation of the mean of each point based on 10 random restarts of the sampler.

As can be seen, the IDCSBM discover larger groups of nodes confirming our previous findings in figure 2 and, more importantly, the variance of the degree distribution within groups is larger than for the ISBM for all groups sizes. This show the compensation for degree heterogeneity not only affect a few large groups the IDCSBM lump together and the ISBM split apart, but groups of all sizes.

To better understand the role of γ, we examined the behaviour of the mean value of γ, $\langle\gamma\rangle$, across the random restarts of the chains both for the artificial and real datasets (see figure 3). For the artificial datasets (figure 3a) we grouped the networks according to the value of λ and γ used to generate the networks and plot the value of $\langle\gamma\rangle$ across the 50 restarts. Consistent with the other findings, the model has more difficulties recovering the true value of γ for very low link density ($\lambda = 0.5$) or when the planted value of γ is very high, here 200 as the highest value. The later finding may be related to this value not being favoured by the prior. However the sampler generally recovers the planted value of γ well across chains.

For the real networks (figure 3b), the recovered values of $\langle\gamma\rangle$ across chains show quite high variability for some of the larger networks indicating they may exhibit mixing times significantly longer than the 1000 iterations used here. Notice that since high values of γ is associated with a nearly vanishing effect of the degree, we see the model correctly identifies the skewed degree distribution of the social network Caltech and Yeast, while indicating the effect of degree for the (very strongly) community-structured network Football and the spatially embedded USPower network is vanishing.

IV. CONCLUSION

In this paper we extended the degree corrected stochastic block model (DCSBM) [5] to a non-parametric Bayesian generative model (the IDCSBM). The advantage of the proposed model being that the number of blocks, i.e. the distribution of the number of groups can be inferred, extending the model to an infinite representation similar to what has previously been done for the regular stochastic block model [6, 7]. By exploiting the model is formulated generatively we have derived a Markov chain Monte Carlo algorithm which handles missing links explicitly by marginalizing over missing entries. We have further shown we can learn the parameter γ in the process and thereby determine the extent to which networks can use the degree correction parameter $\boldsymbol{\theta}$ introduced in the degree corrected stochastic block model. We have shown analytically that under wide conditions the model will be able to accurately model between-group link density as well as node degree.

On synthetic and real networks we demonstrate that the IDCSBM can result in a more compact representation of network structure. The IDCSBM also tends to use fewer components than the ISBM while accounting equally well for the networks as quantified by the AUC link prediction scores. On synthetic data with degree-heterogeneity we have shown the proposed model, which corrects for degree skewness, is able to infer the parameters controlling degree heterogeneity correctly and obtain both a more compact and accurate representation. As expected, this also translates into improved link prediction. On real network data, we have shown that a model which captures degree skewness does not dominate a model which does not in terms of link prediction, however the IDCSBM is able to consistently learn vastly different values of γ and thereby the presence or absence of degree heterogeneity.

ACKNOWLEDGMENTS

This project was supported by the Lundbeck Foundation, grant no. R105-9813.

[1] H. C. White, S. A. Boorman, and R. L. Breiger, American journal of sociology , 730 (1976).

[2] P. W. HollandKathryn Blackmond and S. Leinhardt, Social networks **5**, 109 (1983).

[3] K. Nowicki and T. A. B. Snijders, Journal of the American Statistical Association **96**, 1077 (2001).

[4] R. Guimer and M. Sales-Pardo, Proceedings of the National Academy of Sciences **106**, 22073 (2009).

[5] B. Karrer and M. E. J. Newman, Phys. Rev. E **83**, 016107 (2011).

[6] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, in *AAAI*, Vol. 3 (2006) p. 5.

[7] Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel, Uncertainity in Artificial Intelligence (UAI2006) (2006).

[8] T. P. Peixoto, Physical Review E **85**, 056122 (2012).

[9] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **104**, 7327 (2007).

[10] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008).

[11] T. P. Peixoto, Physical Review Letters **110**, 148701 (2013).

[12] T. P. Peixoto, Physical Review E **89**, 012804 (2014).

[13] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborov, Physical Review Letters **107**, 065701 (2011).

[14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborov, Physical Review E **84**, 066106 (2011).

[15] X. Yan, C. R. Shalizi, J. E. Jensen, F. Krzakala, C. Moore, L. Zdeborova, P. Zhang, and Y. Zhu, arXiv preprint arXiv:1207.3994 (2012).

[16] D. Aldous, cole d't de Probabilits de Saint-Flour XIII1983 , 1 (1985).

[17] E. Jaynes, *Probability theory: the logic of science* (Cambridge University Press, 2003).

[18] A. Clauset, C. Moore, and M. E. Newman, Nature **453**, 98 (2008).

[19] M. Girvan and M. E. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).

[20] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, PLoS biology **6**, e159 (2008).

[21] D. J. Watts and S. H. Strogatz, nature **393**, 440 (1998).

[22] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, *et al.*, Nucleic acids research **31**, 2443 (2003).

[23] D. E. Knuth, *The Stanford GraphBase: a platform for combinatorial computing*, Vol. 4 (Addison-Wesley Reading, 1993).