Submitted to the Annals of Statistics

PSEUDO-LIKELIHOOD METHODS FOR COMMUNITY DETECTION IN LARGE SPARSE NETWORKS

By Arash A. Amini[*], Aiyou Chen[†], Peter J. Bickel[‡] and Elizaveta Levina[*]

[*] *University of Michigan,* [†] *Google, Inc., and* [‡] *University of California, Berkeley*

Many algorithms have been proposed for fitting network models with communities but most of them do not scale well to large networks, and often fail on sparse networks. Here we propose a new fast pseudo-likelihood method for fitting the stochastic block model for networks, as well as a variant that allows for an arbitrary degree distribution by conditioning on degrees. We show that the algorithms perform well under a range of settings, including on very sparse networks, and illustrate on the example of a network of political blogs. We also propose spectral clustering with perturbations, a method of independent interest, which works well on sparse networks where regular spectral clustering fails, and use it to provide an initial value for pseudo-likelihood. We prove that pseudo-likelihood provides consistent estimates of the communities under a mild condition on the starting value, for the case of a block model with two balanced communities.

1. Introduction. Analysis of network data is important in a range of disciplines and applications, appearing in such diverse areas as sociology, epidemiology, computer science, and national security, to name a few. Network data here refers to observed edges between nodes, possibly accompanied by additional information on the nodes and/or the edges, e.g., edge weights. One of the fundamental questions in analysis of such data is detecting and modeling community structure within the network. A lot of algorithmic approaches to community detection have been proposed, particularly in the physics literature (see [26], [14] for reviews). These include various greedy methods such as hierarchical clustering (see [24] for a review) and algorithms based on optimizing a global criterion over all possible partitions, such as normalized cuts [32] and modularity [27]. The statistics literature has been more focused on model-based methods, which postulate and fit a probabilistic model for a network with communities. These include the popular stochastic block model [20], its extensions to include varying degree distributions within communities [21] and overlapping communities [2, 3], and

Keywords and phrases: community detection, network, pseudo-likelihood

various latent variable models [16, 18].

The stochastic block model is perhaps the most commonly used and best studied model for community detection. For a network with n nodes defined by its $n \times n$ adjacency matrix A, this model postulates that the true node labels $c = (c_1, \cdots, c_n) \in \{1, \cdots, K\}^n$ are drawn independently from the multinomial distribution with parameter $\pi = (\pi_1, \cdots, \pi_K)$, where $\pi_i > 0$ for all i, and K is the number of communities, assumed known. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathbb{E}[A_{ij}|c] = P_{c_i c_j} \ , \tag{1}$$

where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix. The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops). The problem of community detection is then to infer the node labels c from A, which typically also involves estimating π and P.

There are many extensions of the block model, notably to mixed membership models [2], but we will only focus on one extension here that we use later in the paper. The block model implies the same expected degree for all nodes within a community, which excludes networks with "hub" nodes commonly encountered in practice. The degree-corrected block model [21] removes this constraint by replacing (1) with $\mathbb{E}[A_{ij}|c] = \theta_i \theta_j P_{c_i c_j}$, where θ_i's are node degree parameters which satisfy an identifiability constraint. If the degree parameters only take on a discrete number of values, one can think of the degree-corrected block model as a regular block model with a larger number of blocks, but that loses the original interpretation of communities. In [21] the Bernoulli distribution for A_{ij} was replaced by the Poisson, primarily for ease of technical derivations, and in fact this is a good approximation for a range of networks [30].

Fitting block models is non-trivial, especially for large networks, since in principle the problem of optimizing over all possible label assignments is NP-hard. In the Bayesian framework, Markov Chain Monte Carlo methods have been developed [33, 29] but they only work for networks with a few hundred nodes. Variational methods have also been developed and studied (see for example [2, 9, 22, 7]), and are generally substantially faster than the Gibbs sampling involved in MCMC, but still do not scale to the order of a million nodes. Another Bayesian approach based on a belief propagation algorithm was proposed recently by [13], and is comparable to ours in theoretical complexity but slower in practice (see more on this in Section 4).

In the non-Bayesian framework, a profile likelihood approach was proposed in [5]: since for a given label assignment parameters can be estimated

trivially by plug-in, they can be profiled out and the resulting criterion can be maximized over all label assignments by greedy search. The same method is used in [21] to fit the degree-corrected block model. The speed of the profile likelihood algorithms depends on exactly what search method is used and the number of iterations it is run for, but again these generally work well for thousands but not millions of nodes. A method of moments approach was proposed in [6], for a large class of network models that includes the block model as a special case. The generality of this method is an advantage, but it involves counting all occurrences of specific patterns in the graph, which is computationally challenging beyond simple special cases. Some faster approximations for block model fitting based on spectral representations are also available [25, 31], but the properties of these approximations are only partially known.

Profile likelihood methods have been proven to give consistent estimates of the labels when the degree of the graph grows with the number of nodes, under both the stochastic block models [5] and the degree-corrected version [37]. To obtain "strong consistency" of the labels, that is, the probability of the estimated label vector being equal to the truth converging to 1, the average graph degree λ_n has to grow faster than $\log n$, where n is the number of nodes. To obtain "weak consistency", i.e., the fraction of misclassified nodes converging to 0, one only needs $\lambda_n \to \infty$. Asymptotic behavior of variational methods has been studied by [9] and [7], and [13] analyzed their belief propagation method for both the sparse ($\lambda_n = O(1)$) and the dense ($\lambda_n \to \infty$) regimes, by non-rigorous cavity methods from physics, and established a phase transition threshold below which the labels cannot be recovered. In fact, it is easy to see that consistency is impossible to achieve unless $\lambda_n \to \infty$, since otherwise the expected fraction of isolated nodes does not go to 0. The results one can get for the sparse case, such as [13], can only claim that the estimated labels are correlated with the truth better than random guessing, but not that they are consistent. In this paper, for the purposes of theory we focus on consistency and thus necessarily assume that the degree grows with n. However, in practice we find that our methods are very well suited for sparse networks and work well on graphs with quite small degrees.

Our main contribution here is a new fast pseudo-likelihood algorithm for fitting the block model, as well as its variation conditional on node degrees that allows for fitting networks with highly variable node degrees within communities. The idea of pseudo-likelihood dates back to [4], and in general amounts to ignoring some of the dependency structure of the data in order to simplify the likelihood and make it more tractable. The main feature of

the adjacency matrix we ignore here is its symmetry; we also apply block compression, that is, divide the nodes into blocks and only look at the likelihood of the row sums within blocks. This leads to an accurate and fast approximation to the block model likelihood, which allows us to easily fit block models to networks with tens of millions of nodes. Another major contribution of the paper is the consistency proof of one step of the algorithm. The proof requires new and somewhat delicate arguments not previously used in consistency proofs for networks; in particular, we use the device of assuming an initial value that has a certain overlap with the truth, and then show the amount of overlap can be arbitrarily close to purely random. Finally, we propose spectral clustering with perturbations, a new clustering method of independent interest which we use to initialize pseudo-likelihood in practice. For sparse networks, regular spectral clustering often performs very poorly, likely due to the presence of many disconnected components. We perturb the network by adding additional weak edges to connect these components, resulting in regularized spectral clustering which performs well under a wide range of settings.

The rest of the paper is organized as follows. We present the algorithms in Section 2, and prove asymptotic consistency of pseudo-likelihood in Section 3. The numerical performance of the methods is demonstrated on a range of simulated networks in Section 4 and on a network of political blogs in Section 5. Section 7 concludes with discussion, and the Appendix contains some additional technical results.

2. Algorithms.

2.1. *Pseudo-likelihood.* The joint likelihood of A and c could in principle be maximized via the expectation-maximization (EM) algorithm, but the E-step involves optimizing over all possible label assignments, which is NP-hard. Instead, we introduce an initial labeling vector $e = (e_1, \ldots, e_n)$, $e_i \in \{1, \ldots, K\}$, which partitions the nodes into K groups. Note that for convenience we partition into the same number of groups as we assume to exist in the true model, but in principle the same idea can be applied with a different number of groups; in fact dividing the nodes into n groups with a single node in each group instead gives an algorithm equivalent to that of [28].

The main quantity we work with are the block sums along the columns,

$$b_{ik} = \sum_j A_{ij}\, 1(e_j = k) \ , \tag{2}$$

for $i = 1, \ldots, n$, $k = 1, \ldots, K$. Let $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iK})$. Further, let R be the $K \times K$ matrix with entries $\{R_{ka}\}$ given by

$$R_{ka} = \frac{1}{n} \sum_{i=1}^{n} 1(e_i = k, c_i = a). \tag{3}$$

Let $R_{k\bullet}$ be the kth row of R, and let $P_{\bullet l}$ be the lth column of P. Let $\lambda_{lk} = nR_{k\bullet}P_{\bullet l}$ and $\Lambda = \{\lambda_{lk}\}$.

Our approach is based on the following key observations: for each node i, conditional on labels $c = (c_1, \ldots, c_n)$ with $c_i = l$,

(A) $\{b_{i1}, \cdots, b_{iK}\}$ are mutually independent;
(B) b_{ik}, a sum of independent Bernoulli variables, is approximately Poisson with mean λ_{lk}.

With true labels $\{c_i\}$ unknown, each $\boldsymbol{b}_i$ can be viewed as a mixture of Poisson vectors, identifiable as long as Λ has no identical rows.

By ignoring the dependence among $\{\boldsymbol{b}_i, i = 1, \ldots, n\}$, using the Poisson assumption, treating $\{c_i\}$ as latent variables, and setting $\lambda_l = \sum_k \lambda_{lk}$, we can write the pseudo log-likelihood as follows (up to a constant):

$$\ell_{\mathrm{PL}}(\pi, \Lambda; \{\boldsymbol{b}_i\}) = \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l e^{-\lambda_l} \prod_{k=1}^{K} \lambda_{lk}^{b_{ik}} \right) \tag{4}$$

A pseudo-likelihood estimate of (π, Λ) can then be obtained by maximizing $\ell_{\mathrm{PL}}(\pi, \Lambda; \{b_i\})$. This can be done via the standard EM algorithm for mixture models, which alternates updating parameter values with updating probabilities of node labels. Once the EM converges, we update the initial block partition vector e to the most likely label for each node as indicated by EM, and repeat this process for a fixed number of iterations T.

For any labeling e, let $n_k(e) = \sum_i 1(e_i = k)$, $n_{kl}(e) = n_k(e)n_l(e)$ if $k \neq l$, $n_{kk}(e) = n_k(e)(n_k(e) - 1)$, and $O_{kl}(e) = \sum_{i,j} A_{ij}\, 1(e_i = k, e_j = l)$. We suppress the dependence on e whenever there is no ambiguity. The details of the algorithmic steps can be summarized as follows.

The pseudo-likelihood algorithm. Initialize labels e, and let $\hat{\pi}_l = n_l/n$, $\hat{R} = \mathrm{diag}(\hat{\pi}_1, \ldots, \hat{\pi}_K)$, $\hat{P}_{lk} = O_{lk}/n_{lk}$, $\hat{\lambda}_{lk} = n\hat{R}_{k\bullet}\hat{P}_{\bullet l}$, $\hat{P} = \{\hat{P}_{lk}\}$ and $\hat{\Lambda} = \{\hat{\lambda}_{lk}\}$. Then repeat T times:

1. Compute the block sums $\{b_{il}\}$ according to (2).
2. Using current parameter estimates $\hat{\pi}$ and $\hat{\Lambda}$, estimate probabilities for node labels by

$$\hat{\pi}_{il} = \mathbb{P}_{\mathrm{PL}}(c_i = l \mid \boldsymbol{b}_i) = \frac{\hat{\pi}_l \prod_{m=1}^K \exp(b_{im} \log \hat{\lambda}_{lm} - \hat{\lambda}_{lm})}{\sum_{k=1}^K \hat{\pi}_k \prod_{m=1}^K \exp(b_{im} \log \hat{\lambda}_{km} - \hat{\lambda}_{km})}.$$

3. Given label probabilities, update parameter values as follows:

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_{il}, \quad \hat{\lambda}_{lk} = \frac{\sum_i \hat{\pi}_{il} b_{ik}}{\sum_i \hat{\pi}_{il}}.$$

4. Return to step 2 unless the parameter estimates have converged.
5. Update labels by $e_i = \arg\max_l \hat{\pi}_{il}$ and return to step 1.
6. Update $\hat{P}$ as follows: $\hat{P}_{lk} = \big(\sum_{i,j} A_{ij} \hat{\pi}_{il} \hat{\pi}_{jk}\big)/n_{lk}(e)$.

In practice, in step 6 we only include the terms corresponding to $\hat{\pi}_{il}$ greater than some small threshold. The EM method is fitting a valid mixture model as long as the identifiability condition holds, and is thus guaranteed to converge to a stationary point of the objective function [35]. Another option is to update labels after every parameter update (that is, skip step 4.) We have found empirically that the algorithm above is more stable, and converges faster. In general, we only need a few label updates until convergence, and even using $T = 1$ (one-step label update) gives reasonable results with a good initial value. The choice of the initial value of e, on the other hand, can be important; see more on this in Section 2.3.

2.2. *Pseudo-likelihood conditional on node degrees.* For networks with hub nodes or those with substantial degree variability within communities, the block model can provide a poor fit, essentially dividing the nodes into low-degree and high-degree groups. This has been both observed empirically [21] and supported by theory [37]. The extension of the block model designed to cope with this situation, the degree-corrected block model [21], has an extra degree parameter to be estimated for every node, and writing out a pseudo-likelihood that lends itself to an EM-type optimization is more complicated. However, there is a simple alternative: consider the pseudo-likelihood conditional on the observed node degrees. Whether these degrees are similar or not will not then matter, and the fitted parameters will reflect the underlying block structure rather than the similarities in degrees.

The conditional pseudo-likelihood is again based on a simple observation:

(C) If random variables X_k are independent Poisson with means μ_k, their distribution conditional on $\sum_k X_k$ is multinomial.

Applying this observation to the variables $(b_{i1}, \cdots, b_{iK})$, we have that their distribution conditional on labels c with $c_i = l$ and the node degree $d_i =$

$\sum_k b_{ik}$ is multinomial with parameters $\theta_{lk} = \frac{\lambda_{lk}}{\lambda_l}$. The conditional log pseudo-likelihood (up to a constant) is then given by,

$$\ell_{\mathrm{CPL}}(\pi, \Theta; \{\boldsymbol{b}_i\}) = \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l \prod_{k=1}^{K} \theta_{lk}^{b_{ik}} \right) , \tag{5}$$

and the parameters can be obtained by maximizing this function via the EM algorithm for mixture models, as before. We again repeat the EM for a fixed number of iterations updating the initial partition vector after the EM has converged. The algorithm is then the same as that for unconditional pseudo-likelihood, with steps 2 and 3 replaced by:

2′. Based on current estimates $\hat{\pi}$ and $\{\hat{\theta}_{lk}\}$, let

$$\hat{\pi}_{il} = \mathbb{P}_{\mathrm{CPL}}(c_i = l \,|\, \boldsymbol{b}_i) = \frac{\hat{\pi}_l \prod_{m=1}^{K} \hat{\theta}_{lm}^{b_{im}}}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{m=1}^{K} \hat{\theta}_{km}^{b_{im}}}.$$

3′. Given label probabilities, update parameter values as follows:

$$\hat{\pi}_l = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{il}, \quad \hat{\theta}_{lk} = \frac{\sum_i \hat{\pi}_{il} b_{ik}}{\sum_i \hat{\pi}_{il} d_i}.$$

2.3. *Initializing the partition vector.* We now turn to the question of how to initialize the partition vector e. Note that the full likelihood, pseudo-likelihoods ℓ_{PL} and ℓ_{CPL}, and other standard objective functions used for community detection such as modularity [27] can all be multi-modal. The numerical results in Section 4 suggest that the initial value cannot be entirely arbitrary, but the results are not too sensitive to it. We will quantify this further in Section 4; here we describe the two options we use as initial values, both of which are of independent interest as clustering algorithms for networks.

2.3.1. *Clustering based on 1- and 2-degrees.* One of the simplest possible ways to group nodes in a network is to separate them by degree, say by one-dimensional K-means clustering applied to the degrees as in [10]. This only works for certain types of block models identifiable from their degree distributions, and in general K-means does not deal well with data with many ties, which is the case with degrees. Instead, we consider two-dimensional K-means clustering on the pairs $(d_i, d_i^{(2)})$, where $d_i^{(2)}$ is the number of paths of length 2 from node i, which can be obtained by summing the rows of A^2.

2.3.2. *Spectral clustering with perturbations.* A more sophisticated clustering scheme is based on spectral properties of the adjacency matrix $A = \{A_{ij}\}$ or its graph Laplacian. Let $D = \mathrm{diag}(d_1, \ldots, d_n)$ be diagonal matrix collecting node degrees. A common approach is to look at the eigenvectors of the normalized graph Laplacian $L = D^{-1/2}AD^{-1/2}$, choosing a small number, say $r = K-1$, corresponding to r largest (in absolute value) eigenvalues, with the largest eigenvalue omitted (see, e.g., [32]). These vectors provide an r-dimensional representation for nodes of the graph, on which we can apply K-means to find clusters; this is one of the versions of spectral clustering, which was analyzed in the context of the block model by [31].

We found that this version of spectral clustering tends to do poorly at community detection when applied to sparse graphs, say, with expected degree $\lambda < 5$. The r-dimensional representation seems to collapse to a few points, likely due to the presence of many disconnected components. We have found, however, that a simple modification performs surprisingly well even for values of λ close to 1. The idea is to connect all disconnected components which belong to the same community by adding artificial "weak" links. To be precise, we "regularize" the adjacency matrix A by adding $\alpha/p \times \lambda/n$ multiplied by the adjacency matrix of an Erdos-Renyi graph on n nodes with edge probability p, where α is a constant. We found that, empirically, $\alpha/p = 0.25$ works well for the range of n considered in our simulations, and that the results are essentially the same for all $p > 0.1$ Thus we make the simplest and computationally cheapest choice of $p = 1$, adding a constant matrix of small values, namely, $0.25(\lambda/n)1_n 1_n^T$ where 1_n is the all-ones n-vector, to the original adjacency matrix. The rest of the steps, i.e., forming the Laplacian, obtaining the spectral representation and applying K-means, are performed on this regularized version of A. We note that to obtain the spectral representation, one only needs to know how the matrix acts on a given vector; since $(A + 0.25(\lambda/n)1_n 1_n^T)x = Ax + 0.25(\lambda/n)(\sum_i x_i)1_n$, the addition of the constant perturbation does not increase computational complexity. We will refer to this algorithm as spectral clustering with perturbations (SCP), since we perturb the network by adding new low-weight "edges".

3. Consistency results. By consistency we mean consistency of node labels (to be defined precisely below) under a block model as the size of the graph n grows. For the theoretical analysis, we only consider the case of $K = 2$ communities. We condition on the community labels $\{c_i\}$, that is, we treat them as deterministic unknown parameters. For simplicity, we only consider the case of balanced communities, each having $m = n/2$ nodes,

which naturally leads us to use the class prior estimates $\hat{\pi}_1 = \hat{\pi}_2 = 1/2$ in (10). We call this assumption **(E)** (for equal class sizes):

(E) Assume each class contains $m = n/2$ nodes, and set $\hat{\pi}_1 = \hat{\pi}_2 = 1/2$.

Without loss of generality, we can take $c_i = 1$ for $i \in \{1, 2, \ldots, m\}$.

As an intermediate step in proving consistency for the block model introduced in Section 1, we first prove the result for a *directed* block model. Recall that for the (undirected) block model introduced earlier, one has

$$\text{(undirected)} \quad A_{ij} \sim \mathrm{Ber}(P_{c_i c_j}), \text{ and } A_{ji} = A_{ij}, \text{ for } i \leq j. \tag{6}$$

In the directed case, we assume that all the entries in the adjacency matrix are drawn independently, that is,

$$\text{(directed)} \quad \widetilde{A}_{ij} \sim \mathrm{Ber}(\widetilde{P}_{c_i c_j}), \text{ for all } i, j. \tag{7}$$

We will use different symbols for the adjacency and edge-probability matrices in the two cases. This is to avoid confusion when we need to introduce a coupling between the two models. In both cases, we have assumed that diagonal entries of the adjacency matrices are also drawn randomly (i.e., we allow for self-loops as valid within-community edges.) This is convenient in the analysis with minor effect on the results.

The directed model is a natural extension of the block model when one considers the pseudo-likelihood approach; in particular, it is the model for which the pseudo-likelihood assumption of independence holds. It is also a useful model of independent interest in many practical situations, in which there is a natural direction to the link between nodes, for example, in email, web, routing, and some social networks. The model can be traced back to the work of [19] and [34] in which it has been implicitly studied in the context of more general exponential families of distributions for directed random graphs.

Our approach is to prove a consistency result for the directed model, with an edge-probability matrix of the form

$$\widetilde{P} = \frac{1}{m} \begin{pmatrix} a & b \\ b & a \end{pmatrix}. \tag{8}$$

Note that the only additional restriction we are imposing is that $\widetilde{P}$ has the same diagonal entries. Both a and b depend on n and can in principle change with n at different rates. This is a slightly different parametrization from the more conventional $P_n = \rho_n S$ [5], where S (and π) do not depend on n, and $\lambda_n = \rho_n \pi^T S \pi$. We use this particular parametrization here because we only

consider the case $K = 2$, and it makes our results more directly comparable to those obtained in the physics literature, e.g., [13].

A coupling between the directed and the undirected model that we will introduce allows us to carry the consistency result over to the undirected model, with the edge-probability matrix

$$P = \frac{2}{m}\begin{pmatrix} a & b \\ b & a \end{pmatrix} - \frac{1}{m^2}\begin{pmatrix} a^2 & b^2 \\ b^2 & a^2 \end{pmatrix}. \tag{9}$$

Asymptotically, the two edge-probability matrices have comparable (to first order) expected degree and out-in-ratio (as defined by [13]), under mild assumptions. The average degrees for $\widetilde{P}$ and P are $a+b$ and $2(a+b)-\frac{1}{m}(a^2+b^2)$, respectively. The latter is $\sim 2(a+b)$ as long as $\frac{1}{2m}\frac{a^2+b^2}{a+b} \leq \frac{a+b}{n} \to 0$. The condition is satisfied as soon as the average degree of the directed model has sublinear growth: $a+b = o(n)$. The same holds for out-in-ratios.

For our analysis, we consider an E-step of the CPL algorithm. It starts from some initial estimates $\hat{a}$, $\hat{b}$ and $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$ of parameters a, b and π, together with an initial labelling e, and outputs the label estimates

$$\widehat{c}_i(e) = \arg\max_{k\in\{1,2\}} \Big\{ \log \hat{\pi}_k + \sum_{\ell=1}^{2} b_{i\ell}(e) \log \hat{\theta}_{k\ell}(e) \Big\}, \quad i \in [n], \tag{10}$$

where $\hat{\theta}_{k\ell}$ are the elements of the matrix obtained by row normalization of $\hat{\Lambda} = [nR(e)\hat{P}]^T$. Here $R = R(e)$ is the confusion matrix as defined in (3) and $\hat{P}$ is given by either (8) or (9), depending on the model, with a and b replaced with their estimates $\hat{a}$ and $\hat{b}$.

The key assumption of our analysis is that the initial labeling has a certain overlap with the truth (we will show later that the amount of overlap is not important). One situation where this might naturally arise is survey data, when some small fraction of nodes has been surveyed about their community membership. Another possibility is to run some other crude algorithm first to obtain a preliminary result. More formally, we consider an initial labeling $e = (e_i) \in \{1,2\}^n$, which is balanced (that is, assigns equal number of nodes to each label) and *matches exactly* γm *labels in community* 1, for some $\gamma \in (0,1)$. We do not assume that we know which labels are matched, or the value of γ. It is easy to see that this is equivalent to e matching exactly γm labels in each of the two communities. Assuming γm to be an integer, let $\mathcal{E}^\gamma = \mathcal{E}_n^\gamma$ denote the collection of such labelings,

$$\mathcal{E}^\gamma = \mathcal{E}_n^\gamma = \Big\{ e \in \{1,2\}^n : \sum_{i=1}^{m} 1\{e_i = 1\} = \gamma m = \sum_{i=m+1}^{n} 1\{e_i = 2\} \Big\}. \tag{11}$$

Our goal is to obtain a uniform result guaranteeing the consistency of CPL iteration (10) for any initial labelling in $\mathcal{E}^\gamma$. In particular, this guarantees consistency for any initial labelling of strength at least γ, even if it is obtained by an algorithm operating on the same adjacency matrix used by CPL. As will become clear in the course of the proof of Theorem 1, although $\{\hat{\theta}_{k\ell}\}$ depend on $R(e)$ (which in turn depends on γ) and $\hat{P}$, under the stated (idealized) assumptions, we do not need to know their exact values in order to implement rule (10). In particular, we do not need to know γ. We can plug in any number in $(0,1) \setminus \{\frac{1}{2}\}$ for γ and get the same estimates. Note that the value of $\gamma = 1/2$ corresponds to "no correlation" between the true and the initial labeling, whereas $\gamma = 0$ and $\gamma = 1$ both correspond to perfect correlation (the labels are either all true or all flipped).

Let us consider the directed case first. We take the following (directed-case) mismatch ratio

$$\widetilde{M}_n(e) = \frac{1}{n}\sum_{i=1}^{n} 1\{\widehat{c}_i(e) \neq c_i\},$$

as our measure of performance (i.e., the loss function), where $\widehat{c}_i(e)$ are computed based on the directed adjacency matrix $\widetilde{A}$. The counterpart for the undirected case is denoted by $M_n(e)$. Note that the notion of consistency based on convergence of this quantity matches the "weak" consistency discussed in [37], rather than the "strong" consistency used by [5]. Define

$$\tau_n^2 = \frac{(a-b)^2}{a+b}, \tag{12}$$

and let $h(p) = -p\log p - (1-p)\log(1-p)$, $p \in [0,1]$ be the binary entropy function. Let us also consider the collection of estimates $(\hat{a}, \hat{b})$ which have the same ordering as true parameters (a, b),

$$\mathcal{P}_{a,b} = \big\{(\hat{a}, \hat{b}) : \ (\hat{a} - \hat{b})(a - b) > 0\big\}.$$

Then, we have the following result.

Theorem 1 (Directed case). *Assume (E), and let $\gamma \in (0,1) \setminus \{\frac{1}{2}\}$. Let the adjacency matrix $\widetilde{A}$ be generated according to the directed model (7) with edge-probability matrix (8), and assume $a \neq b$. Then, there exists a sequence $\{u_n\} \subset \mathbb{R}_+$ such that*

$$\log u_n + \log\log u_n \geq \log\left(\tfrac{4}{e}h(\gamma)\right) + \frac{1}{4}(1-2\gamma)^2\tau_n^2 \tag{13}$$

and

$$\mathbb{P}\Big[\sup_{(\hat{a},\hat{b})\in\mathcal{P}_{a,b}}\ \sup_{e\in\mathcal{E}_n^\gamma}\widetilde{M}_n(e)\geq\frac{4h(\gamma)}{\log u_n}\Big]\leq\exp\big(-n[h(\gamma)-\kappa_n]\big),\tag{14}$$

where $\kappa_n=\frac{1}{n}\{-\log[\frac{\gamma(1-\gamma)}{\pi}n]+\frac{1}{6n}\}=o(1)$.

In particular, if $\tau_n^2\to\infty$, *we have* $u_n\to\infty$ *and the CPL estimate is uniformly consistent.*

Remark 1. We think of γ as fixed, but it is possible to let $\gamma=\gamma_n\to\frac{1}{2}$, making the problem harder as n grows. We still get consistency as long as $(1-2\gamma_n)^2\tau_n^2\to\infty$.

Remark 2. While the labels are of primary interest in community detection, one may also be interested in consistency of the estimated parameters. Under strong consistency in the sense of [5], consistency of the natural plug-in estimates of the block model parameters follows easily, but here we only show weak consistency of the labels. However, in the directed model the pseudo-likelihood function we defined is in fact exactly the likelihood of $\boldsymbol{b}_i$s. Parameter estimates (say $\hat{a}$ and $\hat{b}$) obtained by the EM algorithm converge to a local maximum of this function. As a consequence of Theorem 1, these estimates are also consistent (for a and b). Since the likelihood is smooth with bounded derivatives, one may be able to use standard arguments to show that the estimated parameters are a unique local maximum in a neighborhood of the truth, and even derive their asymptotic normality along (see, e.g., Theorem 6.2.1, p. 384 of [8]). We do not pursue this direction here.

We now turn to the undirected case. Let

$$a_\gamma=\gamma a+(1-\gamma)b.\tag{15}$$

Theorem 2 (Undirected case). *Assume (E), and let* $\gamma\in(0,1)\setminus\{\frac{1}{2}\}$. *Let the adjacency matrix* A *be generated according to the undirected model* (6) *with edge-probability matrix* (9), *and assume* $a\neq b$. *In addition, assume*

$$2(1+\varepsilon)a_\gamma\leq\varepsilon(1-2\gamma)(a-b)\tag{16}$$

for some $\varepsilon\in(0,1)$. *Then, there exist sequences* $\{u_n\},\{v_n\}\subset\mathbb{R}_+$ *such that* $\{u_n\}$ *satisfies* (13), *and* $\{v_n\}$ *satisfies*

$$\log v_n+\log\log v_n\ \geq\ \log\big(\tfrac{4}{e}h(\gamma)\big)+\frac{\varepsilon^2}{1+\varepsilon/3}a_\gamma,$$

and

(17)
$$\mathbb{P}\Big[\sup_{(\hat{a},\hat{b})\in\mathcal{P}_{a,b}}\ \sup_{e\in\mathcal{E}_n^{\gamma}} M_n(e) \geq 4h(\gamma)\Big(\frac{1}{\log u_n}+\frac{1}{\log v_n}\Big)\Big] \ \leq\ 3\exp\big(-n[h(\gamma)-\kappa_n]\big)$$

where $\kappa_n = o(1)$ is as defined in Theorem 1.

In particular, if $\tau_n^2, a_\gamma \to \infty$, we have $u_n, v_n \to \infty$ and the CPL estimate is uniformly consistent.

The proofs of both theorems can be found in Section 6.

Remark 3. Condition (16) can be met for a fixed $\varepsilon \in (0,1)$ by choosing γ sufficiently small and an upper bound on b/a in terms of γ. For example, for $\varepsilon = \frac{1}{2}$ and $\gamma < \frac{1}{8}$, we have (16) if

$$\frac{b}{a} \leq \frac{1-8\gamma}{7-8\gamma}.$$

Remark 4. The parameter τ_n^2 controlling consistency is the same as the one reported in [13] and [23]. There the concern is with recovering a labelling which is positively correlated with the truth, and the threshold of success is observed to be $\tau_n^2 \geq 2$. A similar lower bound was given by [12] for spectral clustering. Here, we are concerned with moving from a positively correlated labelling to one with an asymptotically vanishing mismatch ratio (i.e., $\widetilde{M}_n(e) = o_p(1)$), which is why we need $\tau_n^2 \to \infty$.

4. Numerical results. Here we investigate the performance of both the unconditional and conditional pseudo-likelihood algorithms on simulated networks, as well as that of spectral clustering with perturbations. We simulate two scenarios, one from the regular stochastic block model and one from the degree-corrected block model, to assess the performance in the presence of hub nodes. Throughout this section, we fix $K = 3$ and $\pi = (1/3, 1/3, 1/3)$. Conditional on the labels, the edges are generated as independent Bernoulli variables with probabilities proportional to $\theta_i\theta_j P_{ij}$. The parameters θ_j are drawn independently from the distribution of Θ with $\mathbb{P}(\Theta = 0.2) = \rho$, $\mathbb{P}(\Theta = 1) = 1-\rho$. We do not enforce the identifiability scaling constraint on θ at this point as it is absorbed into the scaling of the matrix P in (18) below. We consider two values of ρ: $\rho = 0$, which corresponds to the regular block model, and $\rho = 0.9$, which corresponds to a network where 10% of the nodes can be viewed as hubs.

The matrix P is constructed as follows. It is controlled by two parameters: the "out-in-ratio" β [13], which we will vary from 0 to 0.2, and the weight vector w, which determines the relative degrees within communities. We consider two values of w: $w = (1, 1, 1)$ (no information about communities is contained in node degrees) and $w = (1, 5, 10)$ (degrees themselves provide relevant information for clustering). If $\beta = 0$, we set $P^{(0)} = \mathrm{diag}(w)$, a diagonal matrix. Otherwise, we set the diagonal of $P^{(0)}$ to $\beta^{-1}w$ and set all off-diagonal elements to 1. We then fix the overall expected network degree λ, which is the natural parameter to control [5] and which we will vary from 1 to 15. Then we rescale $P^{(0)}$ to obtain this expected degree, giving the final P

$$P = \frac{\lambda}{(n-1)(\pi^T P^{(0)} \pi)(\mathbb{E}\Theta)^2} P^{(0)}. \tag{18}$$

To compare our results to the true labels, we will use normalized mutual information (NMI). One can think of the confusion matrix R as a bivariate probability distribution, and of its row and column sums R_{i+} and R_{+j} as the corresponding marginals. Then the NMI is defined by [36] as $\mathrm{NMI}(c, e) = -\sum_{i,j} R_{ij} \log \frac{R_{ij}}{R_{i+}R_{+j}} (\sum_{i,j} R_{ij} \log R_{ij})^{-1}$, and is always a number between 0 and 1 (perfect match). It is useful to have a few benchmark values of NMI for reference: for example, for large n, matching 50%, 70%, and 90% of the labels correspond to values of NMI of approximately 0.12, 0.26, and 0.58, respectively.

All figures show the performance of the following methods: K-means clustering on 1- and 2-degrees (DC), spectral clustering (SC), spectral clustering with perturbations (SCP), unconditional pseudo-likelihood (UPL) initialized with either DC or SCP, and conditional pseudo-likelihood (CPL), with the same two initial values for labelings. The number of outer iterations for UPL and CPL is set to $T = 20$; n, λ, ρ and the number of replications N are specified in the figures.

Figures 1 and 2 show results on estimating the node labels with varying β and λ, respectively. Generally, smaller β and larger λ make the problem easier, as we expect. In principle, degree-based clustering gives no information about the labels with uniform weights w, and only a moderate amount of information with non-uniform weights, so it serves as an example of a poor starting value for pseudo-likelihood. Regular spectral clustering performs well with uniform weights, but very poorly with non-uniform weights; we conjecture that this is due to a limitation of K-means. Spectral clustering with perturbation, on the other hand, performs very well in all scenarios. Apart from being a useful general method on its own, it also serves as an example of a good starting value for pseudo-likelihood.

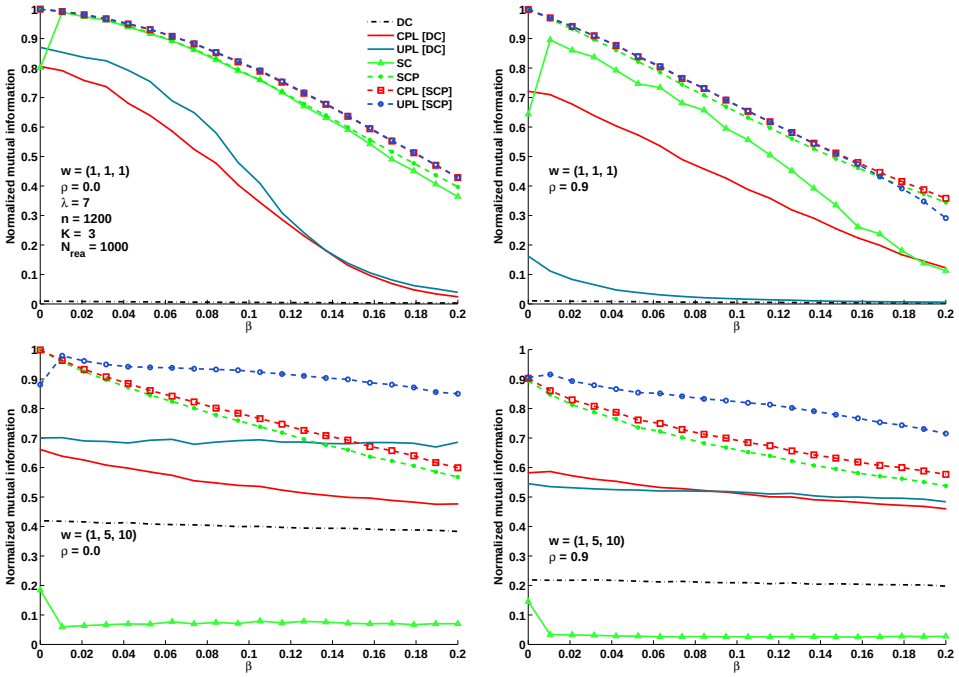


FIG 1. *The NMI between true and estimated labels as a function of "out-in-ratio" β.*

Figures 1 and 2 show that pseudo-likelihood achieves large gains over a poor starting value, giving surprisingly good results even when starting from the uninformative degree clustering in the case of $w = (1, 1, 1)$. One exception is unconditional pseudo-likelihood with $\rho = 0.9$ and $w = (1, 1, 1)$, which shows that conditioning is necessary to accomodate variation in degrees when the starting value is not very good. When spectral clustering with perturbation is used as a starting value, which is already very good, UPL and CPL do not have much room to do better, although UPL still provides a noticeable improvement, being overall the best method when initialized with SCP. It appears that a good starting value overcomes the limitations of the regular block model for networks with hubs, effectively ruling out the competing solution which divides nodes by degree.

Finally, Figure 3 shows run times for all the methods for the case of the regular block model ($\rho = 0$) with different community weights ($w = (1, 1, 1)$ and $w = (1, 5, 10)$). The times shown for UPL and CPL do not include the time to compute the initial value, which is shown separately. For the case $w = (1, 1, 1)$, all methods take roughly the same amount of time. For the case $w = (1, 5, 10)$, spectral clustering (SC) takes considerably more time

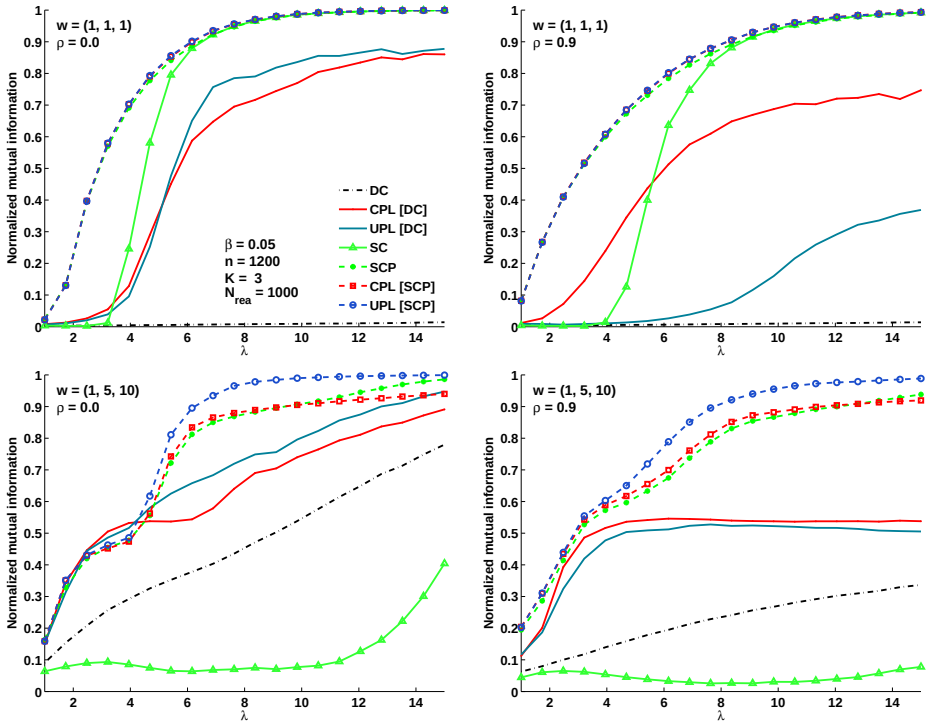


FIG 2. *The NMI between true and estimated labels as a function of average expected degree* λ.

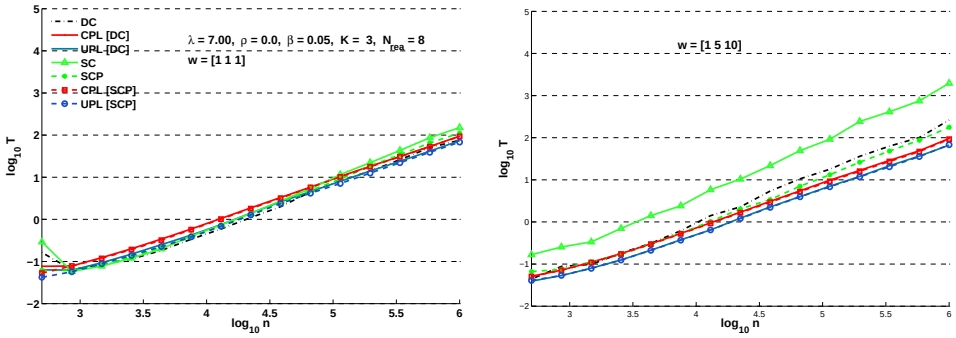


FIG 3. *The runtime in seconds as a function of the number of nodes (log-log scale).*

than the rest. On the other hand, SCP takes nearly the same time as it takes for $w = (1, 1, 1)$, and it slightly outperforms DC for larger values of n. This might be explained, in part, by the sparse matrix multiplication required for DC, which is both time and memory-consuming for large n. Generally, SCP

provides an excellent starting value, with low computational complexity in a variety of situations.

We have also done some brief comparisons with the belief propagation (BP) method of [13]. Direct fair comparison is difficult because of the different platform for the belief propagation code and the different way in which it handles initial values; generally, we found that while the computing time of belief propagation scales with n at the same rate as ours, BP is slower by a constant factor of about 10. In terms of accuracy of community detection, in the examples we tried BP was either similar to or a little worse than pseudo-likelihood.

5. Example: a political blogs network. This dataset on political blogs was compiled by [1] soon after the 2004 U.S. presidential election. The nodes are blogs focused on US politics and the edges are hyperlinks between these blogs. Each blog was manually labeled as liberal or conservative by [1], and we treat these as true community labels. Following [21], we ignore directions of the hyperlinks and analyze the largest connected component of this network, which has 1222 nodes and the average degree of 27. The distribution of degrees is highly skewed to the right (the median degree is 13, and the maximum is 351).

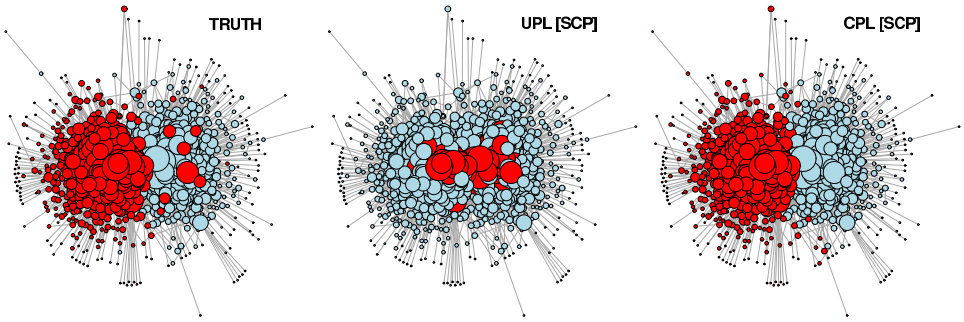


FIG 4. *Political blogs data: true labels and unconditional and conditional pseudo-likelihoods (UPL and CPL) initialized with spectral clustering with perturbations (SCP). Node size is proportional to log degree.*

The results in Figure 4 show that the conditional pseudo-likelihood produces a result closest to the truth, as one would expect in view of highly variable degrees. Its result is also very close to those obtained by profile maximum likelihood for the degree-corrected block model and by two different modularities [21, 37]. Unconditional pseudo-likelihood, on the other hand, puts high-degree nodes in one group and low-degree nodes in the other. This is very close to the block model solution [21]. This example confirms that

the unconditional and conditional pseudo-likelihood methods are correctly fitting the block model and the degree-corrected block model, respectively.

6. Proofs of consistency results. Due to symmetry, we can assume without loss of generality that $\gamma \in (0, \frac{1}{2})$. Similarly, we can assume $a > b$. Then, for any $(\hat{a}, \hat{b}) \in \mathcal{P}_{a,b}$ we have $\hat{a} > \hat{b}$. These will be our standing assumptions throughout the proofs.

6.1. *Proof of Theorem 1 (directed case).* Let us introduce the following notation

$$\begin{aligned} \mathcal{C}_\ell &= \{i : \ c_i = \ell\}, \\ \mathcal{S}_k &= \mathcal{S}_k(e) = \{i : \ e_i = k\}, \\ \mathcal{S}_{k\ell} &= \mathcal{S}_{k\ell}(e) = \mathcal{S}_k \cap \mathcal{C}_\ell, \end{aligned}$$

for $k, \ell = 1, 2$. As long as $e \in \mathcal{E}^\gamma$, we have $|\mathcal{C}_\ell| = |\mathcal{S}_k| = m$ for all $k, \ell = 1, 2$ and

$$|\mathcal{S}_{11}| = |\mathcal{S}_{22}| = \gamma m, \qquad |\mathcal{S}_{12}| = |\mathcal{S}_{21}| = (1-\gamma)m. \tag{19}$$

Under the equal priors assumption (E), the CPL estimate (10) simplifies to

$$\widehat{c}_i(e) = \arg\max_{k \in \{1,2\}} \Big\{ \sum_{m=1}^{2} \widetilde{b}_{im}(e) \log \hat{\theta}_{km}(e) \Big\},$$

where $\{\widetilde{b}_{im}\}$ are obtained by block compression of the directed adjacency matrix $\widetilde{A}$.

Let us focus on $i \in \mathcal{C}_1$ from now on. Then, $\widehat{c}_i(e) = 1$ if

$$\widetilde{b}_{i1}(e) \log \frac{\hat{\theta}_{11}(e)}{\hat{\theta}_{21}(e)} + \widetilde{b}_{i2}(e) \log \frac{\hat{\theta}_{12}(e)}{\hat{\theta}_{22}(e)} > 0. \tag{20}$$

For $e \in \mathcal{E}^\gamma$, we have $r_{k\ell}(e) = n^{-1}|\mathcal{S}_{k\ell}|$, implying that $R(e) = \frac{1}{2}\left(\begin{smallmatrix} \gamma & 1-\gamma \\ 1-\gamma & \gamma \end{smallmatrix}\right)$, where $R(e)$ is defined in (3). It is then not hard to see that after row normalization of $\hat{\Lambda} = [nR(e)\hat{P}]^T$, we obtain $\hat{\theta}_{11}(e) = \hat{\theta}_{22}(e) = \gamma \frac{\hat{a}}{\hat{a}+\hat{b}} + (1-\gamma)\frac{\hat{b}}{\hat{a}+\hat{b}}$, and $\hat{\theta}_{12}(e) = \hat{\theta}_{21}(e) = \gamma \frac{\hat{b}}{\hat{a}+\hat{b}} + (1-\gamma)\frac{\hat{a}}{\hat{a}+\hat{b}}$.

Since by assumption $\hat{a} > \hat{b}$ and $\gamma \in (0, \frac{1}{2})$, it follows that $\hat{\theta}_{11} < \hat{\theta}_{21}$. Then, (20) is equivalent to $\widetilde{b}_{i1}(e) - \widetilde{b}_{i2}(e) < 0$. Recalling that $\widetilde{b}_{ik}(e) =$

$\sum_{j=1}^m \widetilde{A}_{ij} 1\{e_i = k\} = \sum_{j \in \mathcal{S}_k} \widetilde{A}_{ij}$, we can write the condition as

$$\widetilde{\xi}_i(\sigma(e)) = \sum_{j=1}^n \widetilde{A}_{ij}\sigma_j(e) < 0, \quad \text{where} \quad \sigma_j(e) = \begin{cases} 1 & e_j = 1 \\ -1 & e_j = 2 \end{cases},$$

and $\sigma(e) = (\sigma_1(e), \ldots, \sigma_n(e))$. Let $\Sigma^\gamma = \Sigma_n^\gamma$ be the set of all $\sigma(e)$ with $e \in \mathcal{E}^\gamma$, that is,

$$\Sigma^\gamma = \Sigma_n^\gamma = \Big\{\sigma \in \{-1,1\}^n : \sum_{j=1}^m 1\{\sigma_j = 1\} = \gamma m\Big\}.$$

For $\ell = 1, 2$, let $\widetilde{M}_{n,\ell}(e) = \frac{1}{m}\sum_{i \in \mathcal{C}_\ell} 1\{\widehat{c}_i(e) \neq c_i\}$ be the fraction of mismatches over community ℓ. Note that the overall mismatch is

$$\widetilde{M}_n(e) = \frac{1}{2}[\widetilde{M}_{n,1}(e) + \widetilde{M}_{n,2}(e)]. \tag{21}$$

Since we are focusing on $i \in \mathcal{C}_1$, we are concerned with $\widetilde{M}_{n,1}(e)$.

Let us define, for $\sigma \in \{-1, +1\}^n$ and $r \geq 0$,

$$\widetilde{N}_{n,1}(\sigma; r) = \sum_{i=1}^m 1\{\widetilde{\xi}_i(\sigma) \geq -r\}.$$

Then, we have

$$\sup_{e \in \mathcal{E}^\gamma} \widetilde{M}_{n,1}(e) \leq \sup_{\sigma \in \Sigma^\gamma} \frac{\widetilde{N}_{n,1}(\sigma; 0)}{m}$$

where the inequality is due to treating the ambiguous case $\widetilde{\xi}_i(\sigma) = 0$ as error. We now set out to bound this in probability. Let us start with a tail bound on $\widetilde{\xi}_i(\sigma)$ for fixed σ and i.

Lemma 1. *For any $\sigma \in \Sigma^\gamma$ and $t \in (0, 3(a+b)]$, we have*

$$\mathbb{P}\big[\widetilde{\xi}_i(\sigma) \geq -(1-2\gamma)(a-b) + t)\big] \leq \exp\Big(-\frac{t^2}{4(a+b)}\Big). \tag{22}$$

Proof of Lemma 1. We apply the classical Bernstein's inequality for sums of independent bounded random variables. Let $\alpha_{ij} = \mathbb{E}[\widetilde{A}_{ij}]$. Note that $|\widetilde{A}_{ij}\sigma_j - \mathbb{E}[\widetilde{A}_{ij}\sigma_j]| \leq \max(\alpha_{ij}, 1-\alpha_{ij}) \leq 1$. For $i \in \mathcal{C}_1$, we have

$$\begin{aligned}\mathbb{E}\,\widetilde{\xi}_i(\sigma) = \sum_{j=1}^n \alpha_{ij}\sigma_j &= \sum_{j \in \mathcal{S}_{11}} \frac{a}{m}(1) + \sum_{j \in \mathcal{S}_{22}} \frac{b}{m}(-1) + \sum_{j \in \mathcal{S}_{21}} \frac{a}{m}(-1) + \sum_{j \in \mathcal{S}_{12}} \frac{b}{m}(1) \\ &= (a-b)\gamma + (-a+b)(1-\gamma) = -(1-2\gamma)(a-b).\end{aligned}$$

where $\mathcal{S}_{k\ell}$ is defined based on labeling e which correspond to σ. In addition, since $\mathrm{var}(\widetilde{A}_{ij}) \le \alpha_{ij}$, we have

$$v = \sum_{j=1}^{n} \mathrm{var}(\widetilde{A}_{ij}\sigma_j) \le \sum_{j\in\mathcal{C}_1} \alpha_{ij} + \sum_{j\in\mathcal{C}_2} \alpha_{ij} = m\frac{a}{m} + m\frac{b}{m} = a+b.$$

Bernstein's inequality implies

$$\mathbb{P}\big[\widetilde{\xi}_i(\sigma) \ge \mathbb{E}\,\widetilde{\xi}_i(\sigma) + t\big] \le \exp\Big(-\frac{t^2}{2(v+t/3)}\Big).$$

Noting that for $t/3 \le (a+b)$, we have $2(v+t/3) \le 4(a+b)$ completes the proof. □

We also need a tail bound on $\widetilde{N}_{n,1}(\sigma;r)$. Let us define

$$p_i(r) = \mathbb{P}\big[\widetilde{\xi}_i(\sigma) \ge -r\big], \quad \bar{p}_1(r) = \frac{1}{m}\sum_{i=1}^{m} p_i(r). \tag{23}$$

Note that these probabilities do not depend on the particular value of $\sigma \in \Sigma^\gamma$, due to symmetry. In fact, they also do not depend on i due to symmetry, hence $p_i(r) = \bar{p}_1(r)$ for all i. We have the following lemma.

Lemma 2. *For $u > 1/e$,*

$$\mathbb{P}\Big[\frac{1}{m}\widetilde{N}_{n,1}(\sigma;r) \ge e\,u\bar{p}_1(r)\Big] \le \exp\big(-e\,m\bar{p}_1(r)u\log u\big) \tag{24}$$

Proof of Lemma 2. Follows from Lemma 5 in Appendix A, by noting that $\big\{\,1\{\widetilde{\xi}_i(\sigma) \ge -r\}\,\big\}_{i=1}^{m}$ are independent Bernoulli random variables. □

Now we apply Lemma 1 with $t = (1-2\gamma)(a-b) \le 3(a+b)$. Note that $\frac{a-b}{a+b} \le 1 \le \frac{3}{1-2\gamma}$, for $\gamma \in (0, \frac{1}{2})$. Noting that the RHS of (22) does not depend on i and using (23), we get

$$\bar{p}_1(0) \le \exp\Big\{-\frac{1}{4}(1-2\gamma)^2\frac{(a-b)^2}{a+b}\Big\}.$$

The cardinality of the set Σ^γ is $\binom{m}{\gamma m}^2 = (e^{m[h(\gamma)+\kappa_{2m}]})^2$ where $h(\cdot)$ is the binary entropy function and $\kappa_{2m} = \kappa_n$ is as defined in the statement of the theorem. Applying Lemma 2 with $u = u_n$ and the union bound, we obtain

$$\mathbb{P}\Big[\sup_{\sigma\in\Sigma^\gamma}\frac{1}{m}\widetilde{N}_{n,1}(\sigma;0) \ge e\,u_n\bar{p}_1(0)\Big] \;\le\; \exp\big\{m\big[2h(\gamma) - e\,\bar{p}_1(0)u_n\log u_n + 2\kappa_n\big]\big\}.$$

Pick u_n such that

$$u_n \log u_n = \frac{4h(\gamma)}{e\,\bar{p}_1(0)}.$$

It follows, using $m = n/2$, that

$$\mathbb{P}\Big[\sup_{\sigma\in\Sigma^\gamma} \frac{1}{m}\widetilde{N}_{n,1}(\sigma;0) \geq \frac{4h(\gamma)}{\log u_n}\Big] \leq \exp\{-[h(\gamma)-\kappa_n]n\}.$$

By symmetry the same bound holds for $\sup_\sigma \frac{1}{m}\widetilde{N}_{n,2}(\sigma;0)$. It follows from (21) that the same holds for $\sup_e M_n(e)$. This completes the proof of Theorem 1.

6.2. *Proof of Theorem 2 (undirected case).* Recall that A and $\widetilde{A}$ are the adjacency matrices of the undirected and directed cases, respectively. Let us define $\xi_i(\sigma)$, $M_{n,\ell}(e)$, $N_{n,\ell}(\sigma,r)$ as we did in the directed case, but based on A instead of $\widetilde{A}$. For example, $\xi_i(\sigma) = \sum_{j=1}^n A_{ij}\sigma_j$.

Our approach is to introduce a *deterministic coupling* between A and $\widetilde{A}$, which allows us to carry over the results of the directed case. Let

$$A = T(\widetilde{A}), \quad [T(\widetilde{A})]_{ij} = \begin{cases} 0, & \widetilde{A}_{ij} = \widetilde{A}_{ji} = 0, \\ 1, & \text{otherwise} \end{cases}. \tag{25}$$

In other words, the graph of A is obtained from that of $\widetilde{A}$ by removing directions. Note that

$$P_{kl} = \mathbb{P}(A_{ij}=1) = 1 - \mathbb{P}(\widetilde{A}_{ij}=0)\mathbb{P}(\widetilde{A}_{ji}=0) = 2\widetilde{P}_{kl} - \widetilde{P}_{kl}^2\ ,$$

which matches the relation between (8) and (9). From (25), we also note that

$$A_{ij} \geq \widetilde{A}_{ij}, \quad \text{for all } i,j. \tag{26}$$

Let us now upper-bound $\xi_i(\sigma)$ in terms of $\widetilde{\xi}_i(\sigma)$. Based on (26), only those σ_j that are equal to 1 contribute to the upper bound. More precisely, let $D_{ij} = A_{ij} - \widetilde{A}_{ij} \geq 0$, and take $i \in \mathcal{C}_1$ from now on. Then

$$\begin{aligned} \xi_i(\sigma) - \widetilde{\xi}_i(\sigma) &= \sum_{j\in\mathcal{S}_1} D_{ij}\sigma_j + \sum_{j\in\mathcal{S}_2} D_{ij}\sigma_j \\ &= \sum_{j\in\mathcal{S}_1} D_{ij} - \sum_{j\in\mathcal{S}_2} D_{ij} \\ &\leq \sum_{j\in\mathcal{S}_1} D_{ij}. \end{aligned} \tag{27}$$

We further notice that $D_{ij} \le \widetilde{A}_{ij} + \widetilde{A}_{ji}$. To simplify notation, let us define

$$\widetilde{A}_{i*}(\sigma) = \sum_{j \in \mathcal{S}_1} \widetilde{A}_{ij}, \quad \widetilde{A}_{*i}(\sigma) = \sum_{j \in \mathcal{S}_1} \widetilde{A}_{ji} \tag{28}$$

where the dependence on σ is due to $\mathcal{S}_1$ being derived from σ (recall that $\mathcal{S}_1 = \mathcal{S}_1(\sigma) = \{j : \sigma_j = 1\}$). Thus, we have shown

$$\xi_i(\sigma) \le \widetilde{\xi}_i(\sigma) + \widetilde{A}_{i*}(\sigma) + \widetilde{A}_{*i}(\sigma). \tag{29}$$

Recall from definition (15) that $a_\gamma = \gamma a + (1-\gamma) b$.

Lemma 3. *Fix $\varepsilon > 0$. For $i \in \mathcal{C}_1$, we have*

$$\mathbb{P}\big[\widetilde{A}_{i*}(\sigma) > (1+\varepsilon) a_\gamma\big] = \mathbb{P}\big[\widetilde{A}_{*i}(\sigma) > (1+\varepsilon) a_\gamma\big] \le \exp\Big\{-\frac{\varepsilon^2}{1+\varepsilon/3} a_\gamma\Big\}.$$

Proof of Lemma 3. The equality of the two probabilities follows by symmetry. Let us prove the bound for $\widetilde{A}_{i*}(\sigma)$. We apply Bernstein's inequality. Note that

$$\begin{aligned}\mu = \mathbb{E}\Big[\sum_{j \in \mathcal{S}_1} \widetilde{A}_{ij}\Big] &= \sum_{j \in \mathcal{S}_{11}} \mathbb{E}[\widetilde{A}_{ij}] + \sum_{j \in \mathcal{S}_{12}} \mathbb{E}[\widetilde{A}_{ij}] \\ &= \sum_{j \in \mathcal{S}_{11}} \frac{a}{m} + \sum_{j \in \mathcal{S}_{12}} \frac{b}{m} = a\gamma + b(1-\gamma) = a_\gamma.\end{aligned}$$

Since $\sum_{j \in \mathcal{S}_1} \mathrm{var}(\widetilde{A}_{ij}) \le \mu$, we obtain

$$\mathbb{P}\Big[\sum_{j \in \mathcal{S}_1} \widetilde{A}_{ij} \ge \mu + t\Big] \le \exp\Big(-\frac{t^2}{2(\mu + t/3)}\Big).$$

Setting $t = \varepsilon\mu$ completes the proof. □

From (29), it follows that

$$\xi_i(\sigma) \ge 0 \implies (\widetilde{\xi}_i(\sigma) \ge -r) \vee (\widetilde{A}_{i*}(\sigma) \ge r/2) \vee (\widetilde{A}_{*i}(\sigma) \ge r/2)$$

which $\vee$ is the logical OR. This can be seen (as usual) by noting that if the RHS does not hold, then $\widetilde{\xi}_i(\sigma) + \widetilde{A}_{i*}(\sigma) + \widetilde{A}_{*i}(\sigma) < 0$, implying $\xi_i(\sigma) < 0$. Translating to indicator functions,

$$1\{\xi_i(\sigma) \ge 0\} \;\le\; 1\{\widetilde{\xi}_i(\sigma) \ge -r\} + 1\{\widetilde{A}_{i*}(\sigma) \ge r/2\} + 1\{\widetilde{A}_{*i}(\sigma) \ge r/2\}$$

Averaging over $i \in \mathcal{C}_1$ (that is, applying $m^{-1}\sum_{i=1}^m$), we get

$$\frac{1}{m}N_{n,1}(\sigma;0) \le \frac{1}{m}\widetilde{N}_{n,1}(\sigma;r) + \frac{1}{m}\widetilde{Q}_{n,1*}(\sigma;r/2) + \frac{1}{m}\widetilde{Q}_{n,*1}(\sigma;r/2) \tag{30}$$

where $\widetilde{Q}_{n,1*}(\sigma;t) = \sum_{i=1}^m 1\{\widetilde{A}_{i*}(\sigma) \ge t\}$, and similarly for $\widetilde{Q}_{n,*1}(\sigma;t)$. Note that $\widetilde{Q}_{n,1*}(\sigma;t)$ and $\widetilde{Q}_{n,*1}(\sigma;t)$, while not independent, have the same distribution by symmetry, so we can focus on bounding one of them. The key is that each one is a sum of iid terms, e.g., $\{\widetilde{A}_{i*}\}_{i=1}^m$.

We have a bound on $m^{-1}\widetilde{N}_{n,1}(\sigma;r)$ from Lemma 2. We can get similar bounds on the $\widetilde{Q}$-terms. To start, let

$$q_i(r) = \mathbb{P}\big[\widetilde{A}_{i*}(\sigma) \ge r/2\big], \quad \overline{q}_1(r) = \frac{1}{m}\sum_{i=1}^m q_i(r), \tag{31}$$

similar to (23), and note that these quantities too are independent of the particular choice of $\sigma \in \Sigma^\gamma$.

Lemma 4. *For $u > 1/e$,*

$$\mathbb{P}\Big[\frac{1}{m}\widetilde{Q}_{n,1*}(\sigma;r/2) \ge e\,u\overline{q}_1(r)\Big] \le \exp\big(-e\,m\overline{q}_1(r)u\log u\big) \tag{32}$$

Proof of Lemma 4. Follows from Lemma 5 in Appendix A, by noting that $\big\{\,1\{\widetilde{A}_{i*}(\sigma) \ge r/2\}\,\big\}_{i=1}^m$ is an independent sequence of Bernoulli variables. □

The same bound holds for $\frac{1}{m}\widetilde{Q}_{n,*1}(\sigma;r/2)$. Recall the definition of $\bar{p}_1(r)$ from (23). Using (30) and Lemmas 2 and 4, we get

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{\sigma\in\Sigma^\gamma}\frac{1}{m}N_{n,1}(\sigma;0) \ge e\,[u_n\bar{p}_1(r) + 2v_nq_1(r)]\Big] \\
&\quad\le\ \mathbb{P}\Big[\sup_{\sigma\in\Sigma^\gamma}\frac{1}{m}\widetilde{N}_{n,1}(\sigma;r) \ge e\,u_n\bar{p}_1(r)\Big] + 2\mathbb{P}\Big[\sup_{\sigma\in\Sigma^\gamma}\frac{1}{m}\widetilde{Q}_{n,1*}(\sigma;r/2) \ge e\,v_n\overline{q}_1(r)\Big] \\
&\quad\le \exp\big\{m\big[2h(\gamma) - e\,\bar{p}_1(r)u_n\log u_n + 2\kappa_n\big]\big\} + 2\exp\big\{m\big[2h(\gamma) - e\,\overline{q}_1(r)v_n\log v_n + 2\kappa_n\big]\big\}.
\end{aligned}$$

as long as $u_n, v_n > 1/e$. Now, take $r/2 = (1+\varepsilon)a_\gamma$, so that Lemma 3 implies

$$\overline{q}_1(r) \le \exp\Big\{-\frac{\varepsilon^2}{1+\varepsilon/3}a_\gamma\Big\}.$$

Now, in Lemma 1, take $t = (1-2\gamma)(a-b) - 2(1+\varepsilon)a_\gamma$. Note that the assumption

$$2(1+\varepsilon)a_\gamma \le \varepsilon(1-2\gamma)(a-b)$$

implies $t \geq (1-\varepsilon)(1-2\gamma)(a-b) > 0$. In addition $t \leq (1-2\gamma)(a-b) \leq 3(a+b)$ as before. Thus, the chosen t is valid for Lemma 1. Furthermore, $-(1-2\gamma)(a-b)+t = -r$. Hence, the lemma implies

$$\bar{p}_1(r) \leq \exp\Big\{-\frac{1}{4}[(1-\varepsilon)(1-2\gamma)]^2 \frac{(a-b)^2}{a+b}\Big\}.$$

Pick u_n and v_n such that

$$u_n \log u_n = \frac{4h(\gamma)}{e\,\bar{p}_1(r)}, \quad v_n \log v_n = \frac{4h(\gamma)}{e\,\overline{q}_1(r)}.$$

The rest of the argument follows as in the directed case. This completes the proof of Theorem 2.

7. Discussion. The proposed pseudo-likelihood algorithms provide fast and accurate community detection for a range of settings, including large and sparse networks, contributing to the long history of empirical success of pseudo-likelihood approximations in statistics. For the theoretical analysis, we did not focus on the convergence properties of the algorithms, since standard EM theory guarantees convergence to a local maximum as long as the underlying Poisson or multinomial mixture is identifiable. The consistency of a single iteration of the algorithm was established for an initial value that is better than purely arbitrary, as long as, roughly speaking, the graph degree grows and there are two balanced communities with equal expected degrees. The theory shows that this local maximum is consistent, and unique in a neighborhood of the truth, so in fact there is no need to assume that EM has converged to the global maximum, an assumption which is usually made in analyzing EM-based estimates.

We conjecture that additional results may be obtained under weaker assumptions if one focuses simply on estimating the parameters of the block model rather than consistency of the labels, just like one can obtain results for a labeling correlated with the truth (instead of consistent) under weaker assumptions discussed in Remark 4. For example, in a very recent paper [11], results are obtained under very weak assumptions for the mean squared error of estimating the block model parameter matrix P (which in itself does not guarantee consistency of the labels). While the primary interest in community detection is estimating the labels rather than the parameters, we plan to investigate this further to see if and how our conditions can be relaxed. We will also investigate label consistency for more general cases, such as unbalanced communities and the case $K > 2$.

While in theory any "reasonable" initial value guarantees convergence, in practice the choice of initial value is still important, and we have investigated a number of options empirically. Spectral clustering with perturbations, which we introduced primarily as a method to initialize pseudo-likelihood, deserves more study, both empirically (for example, investigating the optimal choice of the tuning parameter), and theoretically. This is a topic for future work.

Acknowledgements. We would like to thank Roman Vershynin (Mathematics, University of Michigan) for highly illuminating discussions. This research is supported by the NSF Focused Research Group grant DMS-1159005. E. Levina is also supported by NSF grant DMS-1106772.

REFERENCES

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election. In *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Machine Learning Research*, 9:1981–2014, 2008.

[3] B. Ball, B. Karrer, and M. E. J. Newman. An efficient and principled method for detecting communities in networks. *Physical Review E*, 34:036103, 2011.

[4] J. E. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc., Ser. B*, 36:192–236, 1974.

[5] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009.

[6] P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

[7] P. J. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. 2012. arXiv:1207.0865.

[8] P. J. Bickel and K. A. Doksum. *Mathematical statistics: Basic ideas and selected topics–2nd edition (updated printing)*, volume 1. Pearson Prentice Hall, 2007.

[9] A. Celisse, J.-J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

[10] A. Channarond, J.-J. Daudin, and S. Robin. Classification and estimation in the stochastic block model based on the empirical degrees. 2011. arxiv:1110.6517.

[11] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. 2012. arXiv:1212.1247.

[12] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR Workshop and Conference Proceedings*, 23:35.1 – 35.23.

[13] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84:066106, 2012.

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

[15] L. J. Gleser. On the distribution of the number of successes in independent trials. *Ann. Probab.*, 3(1):182–188, 1975.

[16] M. D. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *J. R. Statist. Soc. A*, 170:301–354, 2007.

[17] W. Hoeffding. On the distribution of the number of successes in independent trials. *Ann. Math. Statist.*, 27:713–721, 1956.

[18] P. D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, Cambridge, MA, 2007.

[19] P. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs:. *J. of Amer. Statist Assoc.*, 76(373):62–65, 1981.

[20] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983.

[21] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011.

[22] M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 2010.

[23] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. arXiv:1202.1499, 2012.

[24] M. E. J. Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330, 2004.

[25] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(3):036104, Sep 2006.

[26] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006.

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

[28] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci. USA*, 104:9564–9569, 2007.

[29] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[30] P. O. Perry and P. J. Wolfe. Null models for network data. 2012. arXiv:1201.5871v1.

[31] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *Annals of Statistics*, 39(4):1878–1915, 2011.

[32] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[33] T. Snijders and K. Nowicki. Estimation and prediction for stochastic block-structures for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[34] Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[35] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[36] Y. Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Springer, 2003.

[37] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Annals of Statistics*, 2012. arxiv.org/1110.3854.

APPENDIX A: POISSON-TYPE TAIL BOUND

Here is a lemma which we used quite often in proving consistency results in Section 6.

LEMMA 5. *Consider $X_1, X_2, \ldots, X_m$ to be independent Bernoulli variables with $\mathbb{E}[X_i] = p_i$. Let $S_m = \sum_{i=1}^m X_i$, $\mu = \mathbb{E}[S_m] = \sum_{i=1}^m p_i$, and $\overline{\mu} = m^{-1}\mu$. Then, for any $u > 1/e$, we have*

$$\mathbb{P}\big(\frac{1}{m} S_m > e\, u\overline{\mu}\big) \leq \exp(-e\, m\overline{\mu}\, u \log u).$$

PROOF. We apply a direct Chernoff bound. Let $S_m^* \sim \mathrm{Bin}(m, \overline{\mu})$. Then, by a result of Hoeffding [17] (also see [15]), $\mathbb{E}g(S_m) \leq \mathbb{E}g(S_m^*)$ for any convex function $g : \mathbb{R} \to \mathbb{R}$. Letting $g(x) = e^{\beta x}$, we obtain for $\beta > 0$,

$$\begin{aligned}
\mathbb{P}(S_m > t) &\leq e^{-\beta t}\mathbb{E}(e^{\beta S_m^*}) \\
&= e^{-\beta t}(1 + \overline{\mu}(e^t - 1))^m \\
&\leq e^{-\beta t} \exp\big\{m\overline{\mu}(e^t - 1)\big\}
\end{aligned}$$

where we have used $(1 + x)^m \leq \exp(mx)$. The RHS is the Chernoff bound for a Poisson random variable with mean $\mu = \sum_i p_i$, and can be optimized to yield

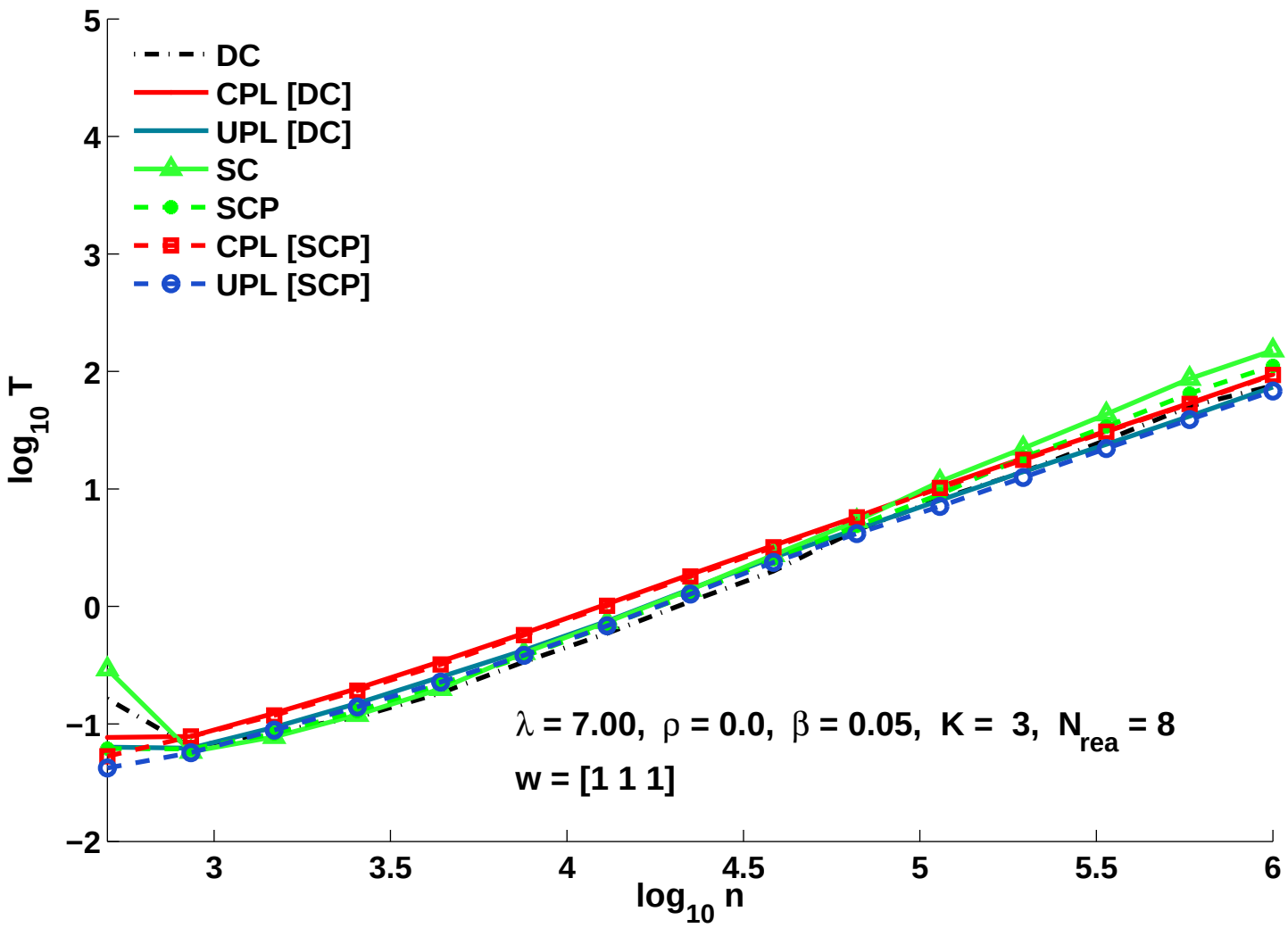
$$\mathbb{P}(S_m > t) \leq \frac{e^{-\mu}(e\mu)^t}{t^t}, \quad \text{for } t > \mu.$$

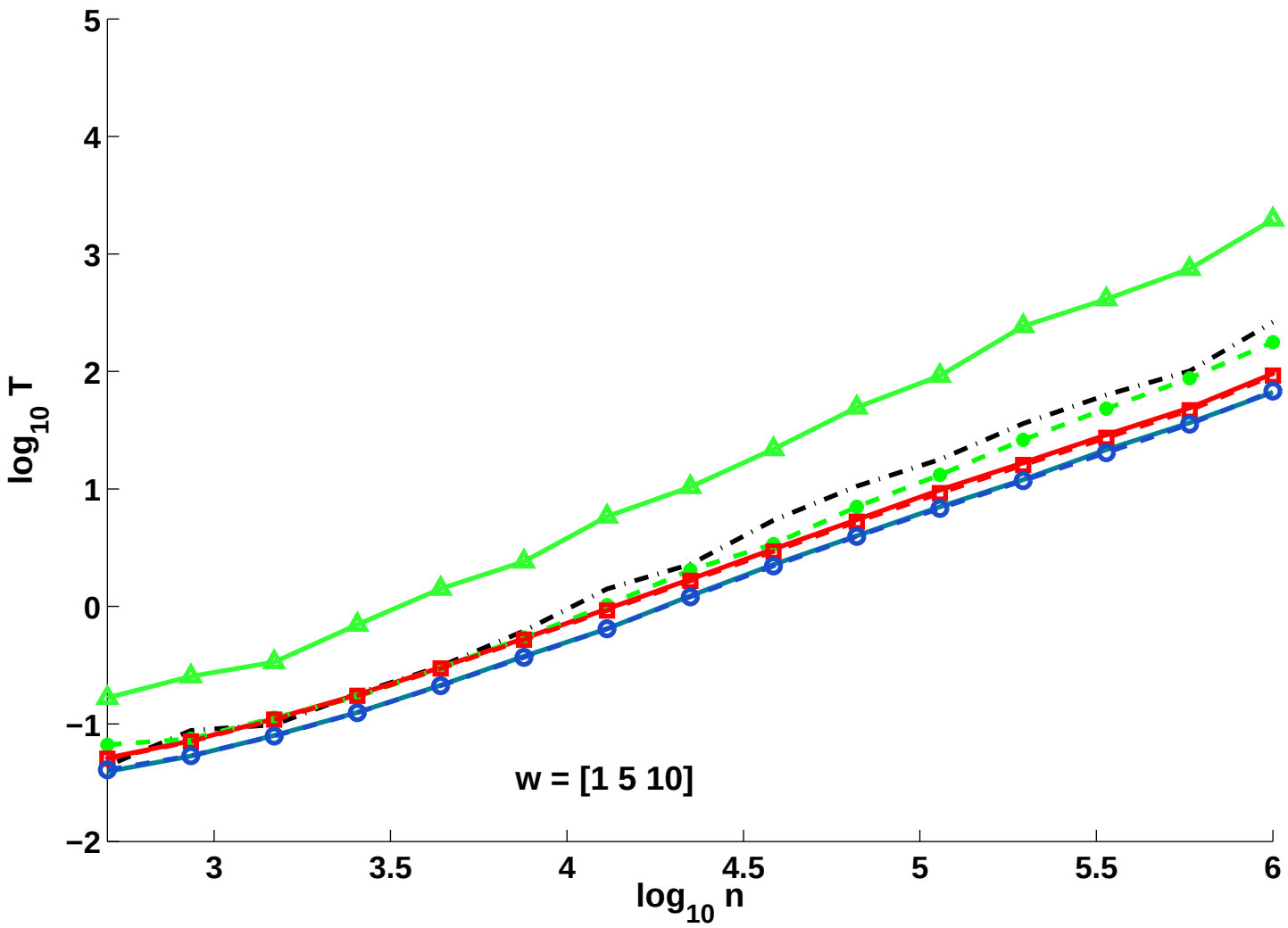
Letting $t = eu\mu$ for $u > 1/e$ and noting that $e^{-\mu} \leq 1$, we get $\mathbb{P}(S_m > e\, u\mu) \leq (1/u)^{e\, u\mu}$ which is the desired bound. □

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109-1107
E-MAIL: aaamini@umich.edu
E-MAIL: elevina@umich.edu

GOOGLE, INC
1600 AMPHITHEATRE PKWY
MOUNTAIN VIEW, CA 94043
E-MAIL: aiyouchen@google.com

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CA 94760
E-MAIL: bickel@stat.berkeley.edu


DC
CPL [DC]
UPL [DC]
SC
SCP
CPL [SCP]
UPL [SCP]
λ = 7.00, ρ = 0.0, β = 0.05, K = 3, N_rea = 8
w = [1 1 1]
log10 T
log10 n
5
4
3
2
1
0
−1
−2
3
3.5
4
4.5
5
5.5
6


5
4
3
2
1
0
−1
−2
log10 T
3
3.5
4
4.5
5
5.5
6
log10 n
w = [1 5 10]