Group-Corrected Stochastic Blockmodels for Community Detection on Large-scale Networks

Lijun Peng
Department of Mathematics & Statistics
Boston University
Boston, MA 02215
ljpeng@bu.edu

Luis E. Carvalho
Department of Mathematics & Statistics
Boston University
Boston, MA 02215
lecarval@math.bu.edu

Abstract

Community detection in networks is becoming increasingly important in many applications, especially in social sciences. Today's era of big data proposes a new challenge in this field—how to efficiently detect community structure on large-scale social networks. In this paper, we propose a stochastic blockmodel based on a logistic regression setup with group correction terms to better address this problem and conduct exact inference based on a *maximum a posteriori* (MAP) estimator. We demonstrate the novel proposed model and estimation on large real-world networks as well as simulated benchmark networks, and show that the proposed estimator is more computationally efficient and performs better when compared to the MAP estimator from classical degree-corrected stochastic blockmodels as well as other commonly used estimators suitable for large-scale networks.

1 Introduction

Despite the increased interest in characterizing the structure of social networks, community detection is arguably still an open problem in network analysis. For instance, while many solutions, such as degree-corrected stochastic blockmodels, are available for relatively small networks, these approaches become computationally intractable for large networks. Large networks typically require approaches such as spectral clustering, which may suffer from limited ability to detect small communities or to detect mixed membership, or membership in multiple communities.

Given the tremendous interest in big data in recent years, there has been extensive study in developing community detection approaches for large-scale networks. Commonly used methods include hierarchical clustering [1], spectral clustering [2, 3, 4] and variational approaches [5], which are heuristic and thus suitable for large networks. However, these methods do not address directly community detection but aim instead at partitioning the network according to edge densities between groups and fail to consider within-group connections.

Among many community detection approaches, we focus on parametric statistical approaches where inference on community structure is based on a *stochastic blockmodels* (SBM), in which the probability of an association between two nodes depends on the communities to which they belong [6, 7]. Here we adopt a hierarchical Bayesian SBM that regards probabilities of association as random and group membership as latent variables which allows for *degree-correction*, that is, models where the degree distribution of nodes within each group can be heterogeneous.

However, this traditional degree-corrected SBM is computationally prohibitive to fit on large networks. To address this hurdle, we consider *group* corrections based on "popularity" instead of degree corrections, and propose a simpler model that involves fewer parameters but that still captures community behavior. This model is introduced in Section 2. In Section 3 we discuss an efficient inferential procedure to estimate community assignments and regression coefficients. We compare our results to existing community detection models based on simulated data in Section 4. In addition, we test our model with four real world data sets with ground truth communities and demonstrate the improved performance of the proposed model compared to other community detection models and estimators. We conclude with a discussion and directions for future work in Section 5.

2 Group-Corrected Stochastic Blockmodel

Given a social network with N individuals and observed interactions $A_{ij} \in \{0, 1\}$ between individuals i and j, our goal is to identify K groups such that there are more interactions within groups and fewer connections between groups. This behavior is usually attributed to social assortativity—individuals with similar interests interact more intensively, "birds of a feather flock together"—and thus these groups are called "communities". We can then see these social associations as a graph with A as its adjacency matrix.

Following the approach of Peng and Carvalho [8], we can consider a Bayesian logistic model to infer network assortativity parameters γ, degree correction terms η, and community labels σ with

$$A_{ij} \,|\, \sigma, \eta, \gamma \sim \mathsf{Bern}\big[\text{logit}^{-1}(\gamma_{\sigma_i,\sigma_j} + \eta_i + \eta_j)\big]$$

as the likelihood. To guarantee identifiability, we set $\gamma_{ss} = 0$ for $s \in \{1, \ldots, K\}$, and to capture community behavior we set $\gamma_{rs} \leq 0$, $r \neq s$, w.p. 1 under the prior, that is, we expect that the probability of an interaction between communities is smaller than the probability of interaction within communities.

This model has $K(K-1)/2 + N$ parameters, and is computationally unfeasible for large K and/or N. To alleviate this problem, we propose to amend the above model in two ways:

1. We expect that, for large networks, individuals are approximately exchangeable modulo their "popularity". Thus, we pool individuals according to their popularity profiles, and assume that the degree distribution within each popularity class is homogeneous.

   To this end, we assign a popularity class Z_i to the i-th individual. There are many ways to assign Z, including the use of other covariates, but here we settle with defining a *maximum* number L of popularity classes, ranking the observed degrees, and splitting them according to their quantiles. More precisely, if $d_i = \sum_{j \neq i} A_{ij}$ is the degree of i, we set $Z_i = j$ if $(j-1)/L < \sum_v I(d_v < d_i)/N \leq j/L$. Some degree strata might not have any nodes, and so the maximum popularity label might not be L.
2. To reduce the number of γ parameters, we set $\gamma_{rs} = 0$ for $r \neq s$, and require $\gamma_{ss} \geq 0$ w.p. 1 under the prior to keep community behavior.

With these two changes, the model has now $K + L$ parameters and is more computational amenable. The new likelihood is then

$$A_{ij} \,|\, \sigma, \eta, \gamma \sim \mathsf{Bern}\left[\text{logit}^{-1}\left(\sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j}\right)\right]. \tag{1}$$

The prior distributions are

$$\begin{aligned}(\gamma, \eta) &\overset{\text{ind}}{\sim} \prod_k I(\gamma_k \geq 0) N(0, \tau^2 I_{L+K}) \\ \sigma_i &\overset{\text{iid}}{\sim} \mathsf{MN}(1; \pi).\end{aligned} \tag{2}$$

Hyper-parameter τ^2 can be taken to be large for a weakly informative prior. Similarly, π informs about the expected size of the communities; for a flat prior we take $\pi = (1/K, \ldots, 1/K)$.

3 Model Inference

To estimate γ, η, and σ we explore a cyclic gradient descent method on the log posterior defined by (1) and (2) with two conditional group steps:

$$[\gamma, \eta \,|\, \sigma, A] \qquad \text{and} \qquad [\sigma_i \,|\, \sigma_{[-i]}, \gamma, \eta, A],$$

where $\sigma_{[-i]}$ denotes all labels but the i-th label. The update step on σ can get stuck in local maxima, and so we run this procedure from multiple starting points and select the fit with highest joint posterior probability. The next subsections detail these steps.

3.1 Updating γ and η

Conditional on community labels σ, we update γ and η using a ridge-regularized version of iteratively reweighted least squares (IRLS [9].) IRLS is an efficient and common method when fitting generalized linear models, and is well adapted to logistic regression. In this case, we have a design matrix X such that $A_{ij} \mid \sigma, \gamma, \eta \sim \mathsf{Bern}(\mu_{ij})$ with $\mu_{ij} = \text{logit}^{-1}(x_{ij}(\sigma)^\top \beta)$ and $\beta = (\gamma, \eta)$ according to (1). That is,

$$x_{ij}(\sigma)^\top \beta = \sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j}.$$

Then, defining $W \doteq \text{Diag}[\mu_{ij}(1-\mu_{ij})]$, the update is $\beta^{(t+1)} = V^{-1}X^\top W z^{(t)}$, with $V = X^\top W X + 1/\tau^2 I_{K+L}$ as the concentration and $z^{(t)} = X\beta^{(t)} + W^{-1}(A-\mu)$ as the "working response". In addition, to guarantee that $\gamma_{ss} \geq 0$ for every community s, we use an active-set method [10] when updating β.

3.2 Updating σ

Now, given the updated values of γ and η, we seek to update σ. A group update as in the previous step is however not possible, so we update each label σ_i in turn, conditional on the remaining labels $\sigma_{[-i]}$ and model parameters γ and η. From (1) and (2), we have that

$$\mathbb{P}(\sigma_i = k \mid \sigma_{[-i]}, \beta, A) \propto \pi_k \prod_{j \neq i} \frac{\exp\{A_{ij} x_{ij}^\top \beta\}}{1 + \exp\{x_{ij}^\top \beta\}}.$$

In practice, we do not compute the product above at each iteration but instead keep track of sufficient statistics when tentatively assigning $\sigma_i = k$ for $k = 1, \ldots, K$. We then pick $\sigma_i^{(t+1)}$ as the argument maximizer of $\mathbb{P}(\sigma_i \mid \sigma_{[-i]}^{(t)}, \beta^{(t)}, A)$ and update the sufficient statistics accordingly.

4 Experimental Results

In this section, we evaluate the performance of our proposed MAP estimator on both simulated benchmark networks and large-scale real-world networks. We compare it to KN estimator [7], Fast-Greedy (FG) estimator [11], Multi-Level (ML) estimator [12], Walktrap (WT) estimator [13] and Label Propagation (LP) estimator [14] through an empirical study in terms of the *normalized mutual information* (NMI) defined in [15]. The NMI measures the similarity between two community labels σ and $\widetilde{\sigma}$:

$$\text{NMI}(\sigma, \widetilde{\sigma}) = \frac{2\text{MI}(\sigma, \widetilde{\sigma})}{H(\sigma) + H(\widetilde{\sigma})},$$

where $\text{MI}(\sigma, \widetilde{\sigma})$ is the mutual information and $H(\sigma)$ is the entropy of σ. The NMI is bounded below by 0, when two labels are independent, and above by 1, when two labels are identical.

4.1 Empirical Study

In our empirical study we adopt a popular benchmark suite that models community behavior and heterogeneities in node degrees using a power law distribution [16]. The model has the following parameters: the exponents of the power law distribution on network degrees and community sizes, a and b, respectively; network average degree $\langle k \rangle >$; mixing parameter μ that captures the proportion of between-community edges; the ratio ρ between the size n of the network and its maximum degree, as a way to control community sizes relative to n, i.e., higher values ρ yield smaller communities. In our simulations we have set $a = 2$ and $b = 1$, and varied designs with $n \in \{500, 1000\}$, $\langle k \rangle \in \{10, 15, 20\}$, $\mu \in \{0.1, 0.2, \ldots, 0.6\}$, and $\rho \in \{2, 10\}$.

We generate 100 networks for each combination of the parameters mentioned above and assume that the number of popularity classes is at most $\lfloor n/10 \rfloor$. The NMIs of other estimators (KN, FG, ML, WT, and LP) and our proposed MAP estimator are summarized in Figure 1. We can conclude that the MAP estimator outperforms the other estimators on average in terms of the NMI, especially when the network is formed by relatively large communities (smaller ρ.) Not surprisingly, all estimators perform worse as the mixing parameter μ increases (so that the communities are defined in a weaker sense) or the average degree $\langle k \rangle$ decreases. We also measure running times to compare the practical computational complexities across methods, as seen in Figure 2. We see that our community detection procedure outperforms Karrer and Newman's (KN) significantly in computational time while showing better or comparable results. In addition, our MAP estimator is as computationally efficient as the FG, ML, and WT estimators. While the LP estimator outperforms all other estimators with respect to runtime, it often achieves the lowest NMI on average.

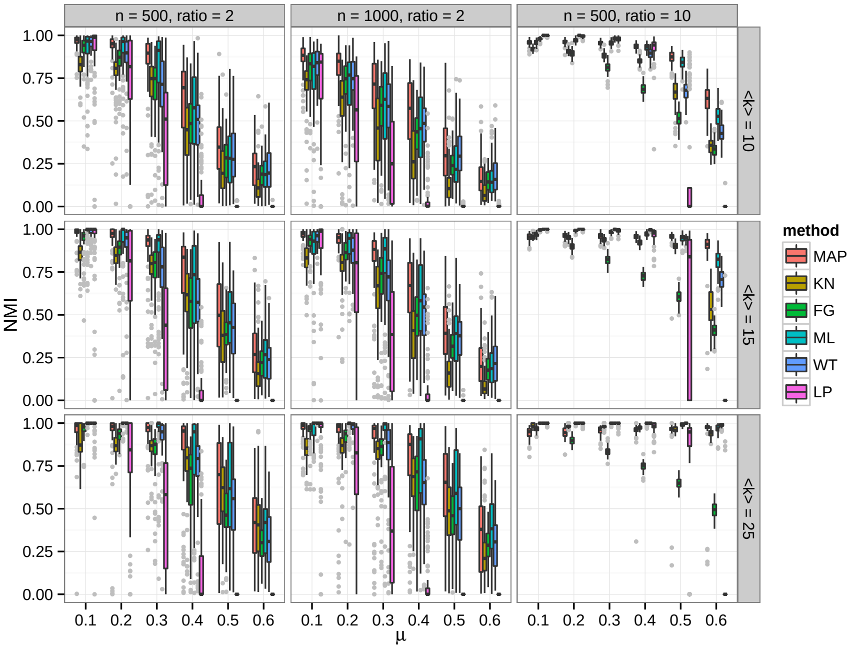

Figure 1: Benchmark networks of $n = 500$ and 1000 nodes, with different combinations of the average degree $\langle k \rangle$ and the ratio controlling the relative size of communities. Each boxplot corresponds to the NMI of the estimator over 100 graph realizations.

It is also worth pointing out that NMI is less useful in comparing communities as K increases, as shown in Figure 1 when "ratio = 10": NMI is unusually high (above 0.75). To understand why this is the case, consider a simple illustration where:

$$\begin{array}{rcccccccccc} \sigma: & 1 & 1 & 2 & 2 & 3 & 3 & \cdots & \cdots & K & K \\ \widetilde{\sigma}: & 1 & 2 & 3 & \cdots & K & 1 & 2 & 3 & \cdots & K. \end{array}$$

In this case $H(\sigma) = H(\widetilde{\sigma}) = \log K$, and since $\mathrm{MI}(\sigma, \widetilde{\sigma}) = 2\log K - \log 2K$, $\mathrm{NMI}(\sigma, \widetilde{\sigma}) = 1 - \log 2 / \log K$ approaches 1 as K increases while the actual labels might be quite different. Similar loss of resolution happens to other measures, such as Binder's loss and adjusted Rand Index, so we focus on NMI since it is more standard.

4.2 Case Study

Next, we evaluate our estimator for community detection on a collection of large-scale real-world network datasets with ground-truth communities [17]. We consider three networks: an online social network *Youtube*, where nodes represent the users of Youtube, edges indicate the friendship formed by the users and ground-truth communities are defined by the user-defined interest groups;

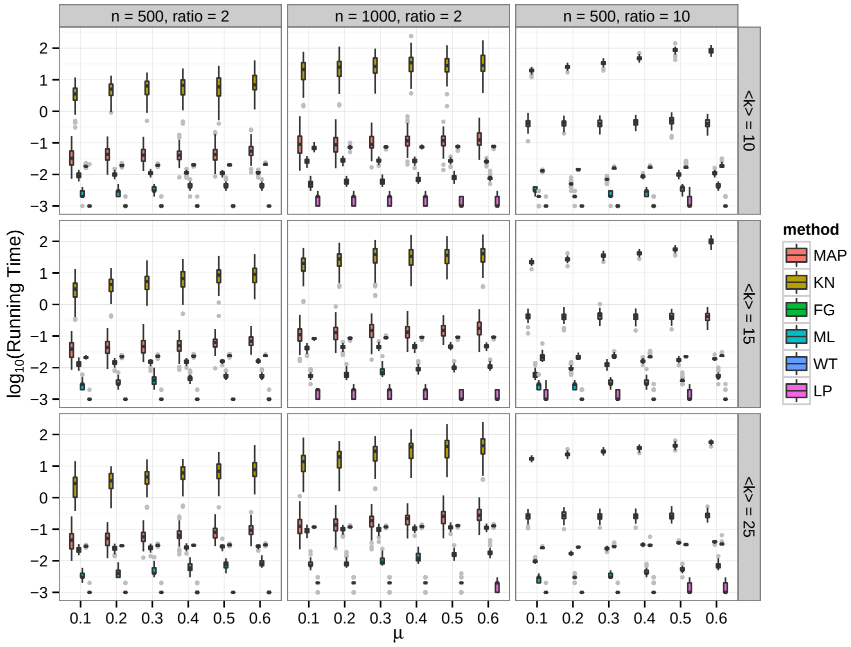

Figure 2: Benchmark networks of $n = 500$ and 1000 nodes, with different combinations of the average degree $\langle k \rangle$ and the ratio controlling the relative size of communities. Each boxplot corresponds to the $\log_{10}(\text{runtime})$ of the estimator over 100 graph realizations.

a co-authorship network *DBLP* where nodes represent authors published in a comprehensive list of research papers in computer science, edges indicate co-authorship in at least one paper and ground-truth communities are defined by Publication venues; a product co-purchasing network *Amazon*, where nodes represent products sold on Amazon website, edges indicate frequently co-purchase and ground-truth communities are defined by product categories provided by Amazon. All datasets are publicly available at the SNAP dataset, `http://snap.stanford.edu/data`.

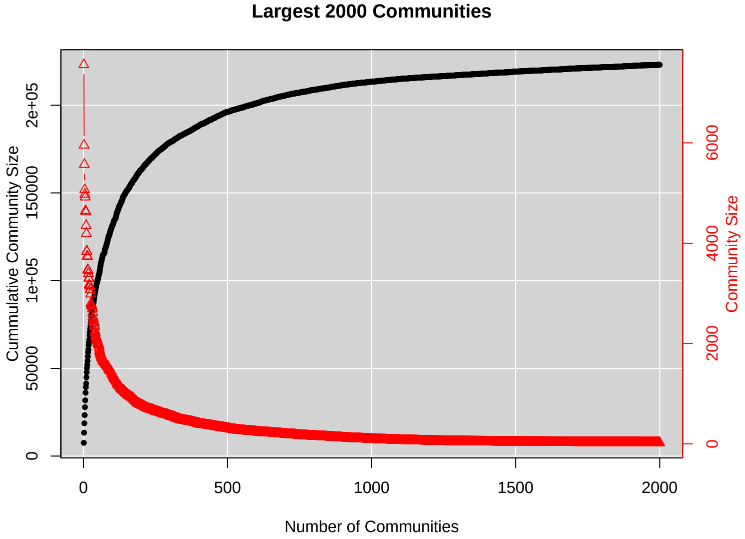

Figure 3: Co-authorship network *DBLP* with $k = 2{,}000$. Red: the decreasing sequence $\{|C_{(i)}|, i = 1, \ldots, k\}$; black: the cumulative size of the largest k communities.

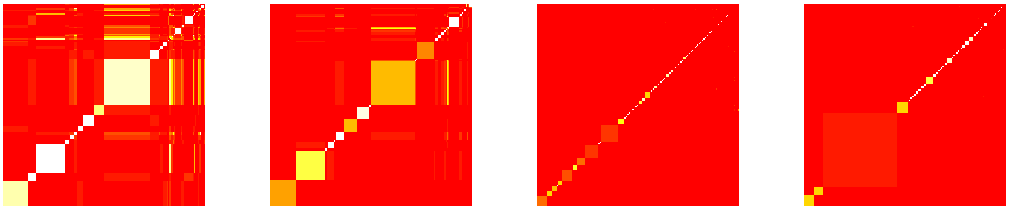

Figure 4: Community-by-community heatmaps showing the edge densities within and between communities based on our MAP estimator. The left two plots correspond to the Youtube network under the smallest and largest L. The right two plots correspond to the DBLP and Amazon networks under the largest L. Red indicates lower edge densities while white indicates higher edge densities.

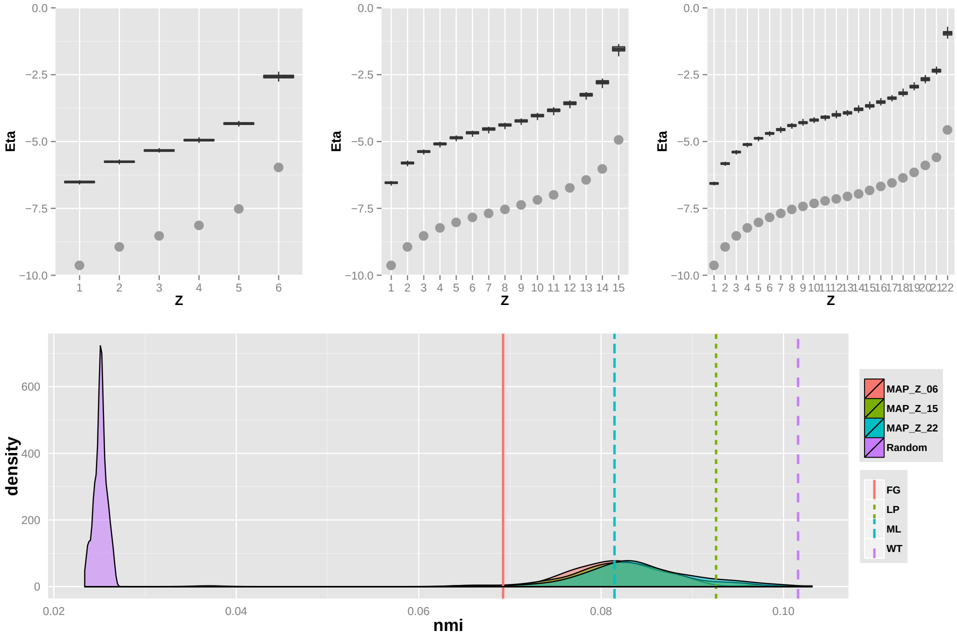

Figure 5: Online social network *Youtube*. Top: η_i (boxplots) and logit$(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class i. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML, and WT estimates relative to the ground truth.

We had to pre-process the ground-truth communities in these datasets for two reasons: first, they allowed for mixed membership which can result in "redundant" communities due to overlap; second, interestingly, community sizes vary considerably, ranging from two to hundreds of thousands, and so some communities are negligible and can introduce noise in the reference. To address these issues, we filter the communities: we first order them by their sizes $|C_{(i)}|, i = 1, \ldots, K$, then take the cumulative size of the largest k communities $| \cup_{i \leq k} C_{(i)}|$ (increasing trend) as well as the sequence $\{|C_{(i)}|, i = 1, \ldots, k\}$ (decreasing trend) for some $k \leq K$, and finally pick the top $k^\star$ communities that comprise at least 95% of the network. Figure 3 shows an example of this procedure on the DBLP co-authorship network where $k^\star = 1,000$ is a good choice. Next, we consider the induced sub-graph generated by the nodes in the largest $k^\star$ communities and further shrink the number of communities by merging communities that are closely connected by performing hierarchical clustering. We regard each connected component in a group as a separate ground-truth community and provide an analysis on the largest connected component.

For each of these networks, we vary the number of popularity classes and carry out our proposed MAP estimation procedure. Figure 4 shows edge densities within communities and between communities based on the MAP estimator. It is apparent that edge densities are higher within commu-

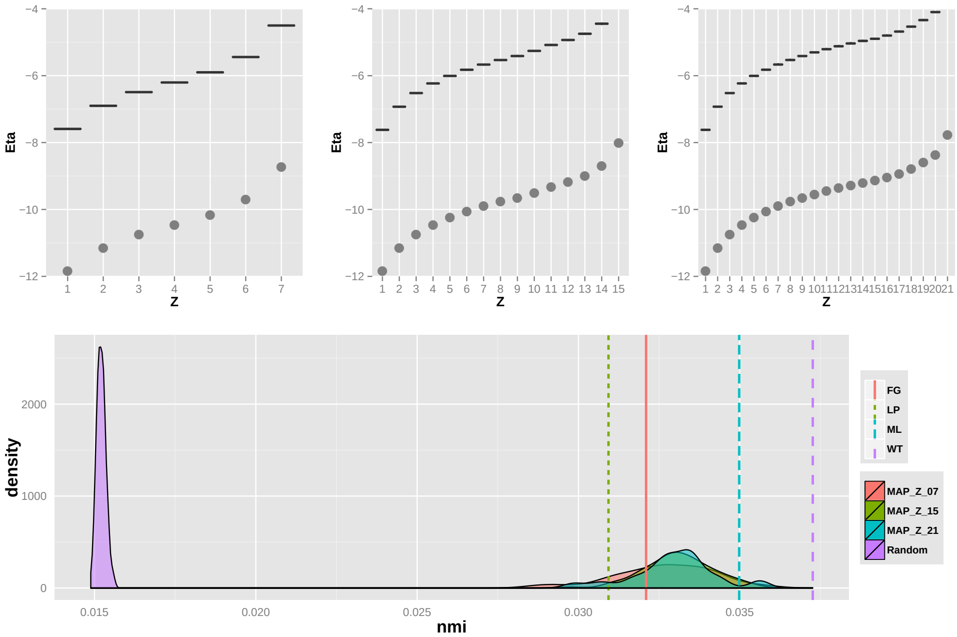

Figure 6: Co-authorship network *DBLP*. Top: η_i (boxplots) and logit$(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class i. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML, and WT estimates relative to the ground truth.

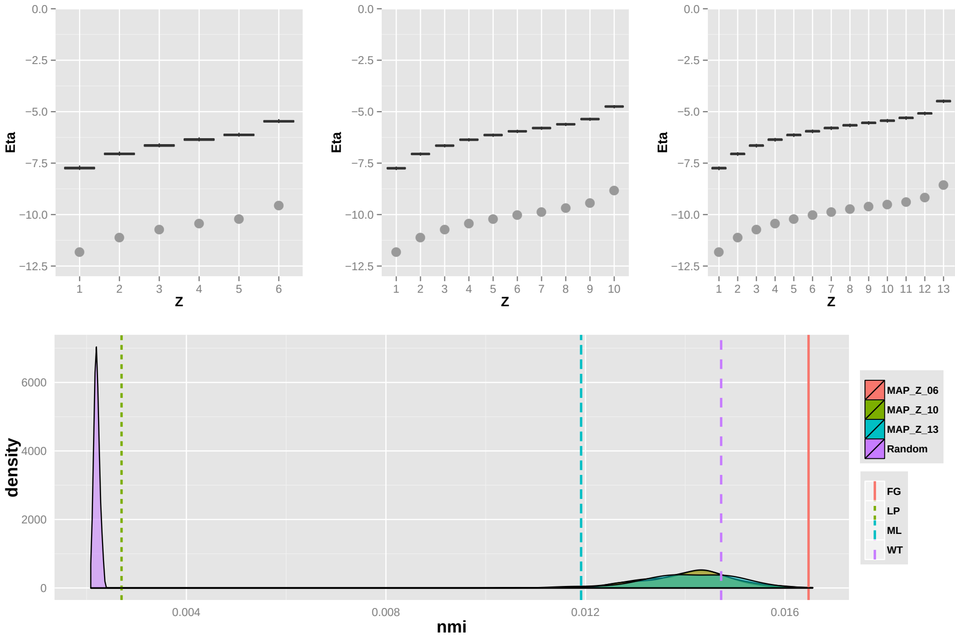

Figure 7: Product co-purchasing network *Amazon*. Top: η_i (boxplots) and logit$(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class i. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML, and WT estimates relative to the ground truth.

nities than between communities. Since the KN estimator used for comparison in Section 4.1 is practically unfeasible on large networks due to long running times, we generate random community assignments for comparison. The top panels in Figures 5, 6, and 7 plot the estimated η_i (boxplots) and the average normalized degrees in logit scale (points) within each popularity class; it is clear that η is closely related to the degree of nodes. The remaining panels in Figures 5, 6, and 7 show a comparison among FG, LP, ML, WT estimators, our MAP estimator and randomly generated labels in terms of NMI. We conclude that our MAP estimator performs as well as FG, LP, ML, and WT estimators on real-world large-scale networks while outperforms the random labels since it yields higher NMI on average.

5 Discussion

In this paper we have introduced a Bayesian model based on group-corrected stochastic blockmodels that is tailored for community detection in large networks. Our model aims to capture gregarious community behavior by requiring that the probability of within-community associations to be no smaller than that of between-community associations. What's more, we take degree heterogeneity into consideration by proposing a model that is more parsimonious and that makes group corrections based on popularity. The method has connections to some existing methods, but is expected to be more efficient, interpretable, and suitable for large-scale networks. We have demonstrated, based on on simulated benchmark networks as well as real-world networks with ground-truth references, that our MAP estimator performs better than or comparably to other estimators, at similar computational costs.

For future work, we plan to extend the proposed model to incorporate parameters for group attributes and to expand the formulation to account for count, categorical, and ordinal data using a broader generalized linear model formulation. Other directions for future work, albeit not related to community detection, include proposing more powerful measures for comparing communities, especially when the number of communities is large relative to the network. The number of communities K in out current model is assumed be known given the ground truth communities; we intend to investigate a procedure or criterion that efficiently selects K, such as Bayesian Information Criterion (BIC) [18, 19].

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Springer, 2001.

[2] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[3] Donghui Yan, Ling Huang, and Michael I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 907–916, New York, NY, USA, 2009. ACM.

[4] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 08 2011.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] Vladimir Batagelj, Andrej Mrvar, and Patrick Doreian. Generalized blockmodeling with pajek, 2005.

[7] B. Karrer and M.E.J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[8] L. Peng and L. E. Carvalho. Bayesian degree-corrected stochastic block models for community detection. *arXiv:1309.4796v1*, 2013.

[9] P. McCullagh and J. A. Nelder. *Generalized linear models (Second edition)*. London: Chapman & Hall, 1989.

[10] Jorge Nocedal and Steve J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, Berlin, 2006. NEOS guide http://www-fp.mcs.anl.gov/otc/Guide/.

[11] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, August 2004.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[13] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293, 2004.

[14] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), 2007.

[15] Leon Danon, Albert Daz-guilera, and Jordi Duch. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, page 09008, 2005.

[16] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(1):046110, 2008.

[17] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 3:1–3:8, New York, NY, USA, 2012. ACM.

[18] C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[19] Xin Gao and Peter X.-K. Song. Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540, 2010.