Article

Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 6, pp. 635–664
DOI: 10.3102/1076998615606110

Incorporating Covariates Into Stochastic Blockmodels

Tracy M. Sweet
University of Maryland

Social networks in education commonly involve some form of grouping, such as friendship cliques or teacher departments, and blockmodels are a type of statistical social network model that accommodate these grouping or blocks by assuming different within-group tie probabilities than between-group tie probabilities. We describe a class of models, covariate stochastic blockmodels (CSBMs), that incorporates covariates into blockmodels. These models not only estimate the effects of covariates in the presence of the block structure but also can determine differential covariate effects such as within blocks versus between blocks. For example, education researchers can now determine those factors that mitigate relationships both within schools and between schools. We introduce several CSBMs as examples and present a series of simulation studies to investigate both the feasibility and some operating characteristics as well as fit CSBMs to real network data.

Keywords: *social network analysis; blockmodels; Bayesian statistics*

1. Introduction

In educational settings, school professional networks offer insight into how social resources are shared and information is exchanged among teachers and other staff. For example, advice-seeking social networks can reveal which teachers are considered experts, whose advice is highly sought, which teachers are well-connected in their school's network, and which teachers are completely isolated. Often substantive researchers are interested in the factors associated with the presence of a particular relationship, since collegial interactions are associated with changes in teacher practice (Frank, Zhao, & Borman, 2004; Moolenaar, Daly, & Sleegers, 2010). For example, Spillane, Kim, and Frank (2012) and Hopkins, Lowenhaupt, and Sweet (2015) both find that being a formal leader increases the likelihood of providing advice and that teachers tend to seek and provide advice to other teachers who teach the same grade. Thus, it is not surprising that many standard statistical social network models, such as exponential random graph models (Wasserman & Pattison, 1996) and latent space models (Hoff, Raftery, & Handcock, 2002), accommodate covariates for inference about individual- or pair-level variables.

Current social network models however do not adequately accommodate ties between teachers in different schools. Analyzing ties across schools is relevant in

many contexts, since schools do not operate in isolation and may share teachers, coaches, or other personnel. Research on those ties that do exist across schools may be important for understanding the functionality of the district as a whole in addition to other factors such as which schools have more across-school ties, which teachers' attributes are related to these ties, or the general relationship between across-school ties and within-school ties. Consider a school district with k schools. The framework introduced by Sweet, Thomas, and Junker (2013) for modeling the networks in these k schools is only useful if we assume that these k schools are isolated. While important for many research questions, this framework does not currently allow for ties across schools.

Instead, we propose to treat the district as a single network and model the k school networks as subgroups or blocks within the larger district network. Social network models that accommodate subgroup structure are commonly known as blockmodels (White, Boorman, & Breiger, 1976) or stochastic blockmodels (Holland, Laskey, & Leinhardt, 1983), and these models generally assume that within-block ties have a different probability than across-block ties. The terms *a priori* stochastic blockmodel and *a posteriori* stochastic blockmodel refers to whether block membership is known or unknown, respectively, and are common in the blockmodeling literature (Anderson & Wasserman, 1992; Nowicki & Snijders, 2001; Wasserman & Anderson, 1987).

Because blockmodels historically have not included covariates, the purpose of this article is to define a class of models, covariate stochastic blockmodels (CSBMs), explore a variety of CSBMs, and illustrate how these models can be applied in educational research. A covariate extension of the stochastic blockmodel was originally introduced by Airoldi, Choi, and Wolfe (2011) and our work will explore similar models as well as some of the operating characteristics of these models, which has not been studied. There are several other models that incorporate covariates into extensions of stochastic blockmodels (Sweet, Thomas, & Junker, 2014; Tallberg, 2004; White & Murphy, 2014), but these extensions are relatively rare and these models have not been applied in education.

In Section 2, we formally introduce several CSBMs and describe our model-fitting algorithm. We then discuss possible applications for these models as well as consider potential issues in Section 2.2. Then, to illustrate the feasibility of these models as well as to assess potential model-fitting issues and parameter recovery, we present a series of simulation studies in Section 3, followed by an application using real advice-seeking network data in Section 4. We conclude by describing future work and additional applications in Section 5.

2. Model

The CSBM framework is a class of blockmodels that accommodates covariates in an additive way, so that tie probability is influenced by both block membership and covariates. We then assume covariates influence the probability of a

tie independently of the block membership, which is similar to the model proposed by Airoldi et al. (2011) and unlike other models that incorporate covariates into the block membership assignment mechanism (Tallberg, 2004; White & Murphy, 2014).

We define a CSBM using a hierarchical Bayesian model. Let Y denote the binary sociomatrix such that $Y_{ij} = 1$ implies a tie from individual i to individual j. If the tie is directed, we consider i the sender of the tie or relationship and j the receiver. For example, in an advice-seeking network, $Y_{ij} = 1$ indicates that i seeks advice from j. Let g_i be the block or group membership indicator vector for individual i such that $g_{ik} = 1$ if and only if i is in group k and 0 otherwise. Undirected relationships are also possible and CSBMs can accommodate these networks as well by assuming $Y_{ij} = Y_{ji}$.

CSBMs incorporate covariates into the model in an additive way through the log odds probability of a tie. Since the probability of a tie has support (0, 1), we transform the tie probability to the log odds or logit scale, so that the support is now $\mathbb{R}$. Then, we can easily incorporate a linear component βX_{ij}, where X_{ij} is any collection of node- or edge-level covariates and β is the vector of regression coefficients. For example, if teachers in a district are the nodes, then teacher attributes such as experience or position are node-level covariates. These covariates may be related to either the sender or the receiver such as the sender's experience or position. Edge-level or pairwise covariates are based on both teachers' attributes such as difference in the amount of professional development and the indicator that both teachers teach the same grade.

An example of such a CSBM is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ \beta &\sim MVN(\mu, \textstyle\sum^2), \\ g_i &\sim \text{Multinomial}(1, \theta), \\ B_{lm} &\sim \text{Beta}(a_{lm}, b_{lm}), \end{aligned} \tag{1}$$

where logit (B) is defined as logit (B_{lm}) for all entries of B. Note also that B is a $k \times k$ matrix, where k is the number of blocks or subgroups. Then, the entry denoted by $g_i^T B g_j$ is the probability of a tie from an individual in i's block to an individual in j's block. We generally assume that within-block ties are more likely than across-block ties, although the reverse is also possible. We enforce this assumption through the prior specification on B. Alternatively, we may also assume part or all of this matrix to be known a priori. Note also that B is symmetric if the network is undirected. We also include a prior for β as an example though the prior need not be normal.

Note that when block membership is not known a priori, it must be estimated, and in Equation 1, we use a multinomial prior for each g_i. One such example is a friendship network in which students cluster together in a finite number of subgroups, but the individual identities within each clique are not known. We define this model as the a posteriori CSBM. Note that the model proposed by Airoldi et al. (2011) differs from our a posteriori CSBM in that they also included random effects for each node.

There are many situations in education in which block membership is known. For example, given a network of schools, each school could be considered as a subgroup, and individual membership to each school is known and does not need to be part of the model. If the block memberships are known, then the model is called the a priori CSBM and is given below as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ \beta &\sim MVN(\mu, \textstyle\sum^2), \\ B_{lm} &\sim \text{Beta}(a_{lm}, b_{lm}). \end{aligned} \tag{2}$$

Note that the a priori CSBM can be thought of as a reparameterization of a logistic regression model, although we caution that this model should be interpreted carefully. Instead of using a B matrix or in our case a logit (B) to describe the additive effects of block membership, we can incorporate group–group probabilities through indicators and regression coefficients. Such a model is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ \text{logit}(p_{ij}) &= \textstyle\sum_l \sum_m \gamma_{lm} I_{[g_{il}=1]} I_{[g_{jm}=1]} + \beta X_{ij}, \\ \beta &\sim MVN(\mu, \boldsymbol{\Sigma}^2), \\ \gamma_{lm} &\sim \text{Beta}(a_{lm}, b_{lm}), \end{aligned} \tag{3}$$

where l and m index the sender's and receiver's blocks, respectively, and $\gamma_{lm} = \text{logit}(B_{lm})$ in Equation 2.

It is important to realize that without informative priors on the entires of the B matrix, the model is not identified. Without these priors, we make no assumption about the probability of within-block ties as compared to across-block ties. Without this or an analogous assumption, block structure no longer is meaningful and block membership does not contribute to the variability in network structure. Thus, if one chooses to fit the a priori CSBM using logistic regression, it should be done in a Bayesian framework taking advantage of informative priors, or if done in a frequentist setting, it should employ some kind of constraint to accommodate the block structure of the network.

Turning our attention to the covariate part of the CSBM, there are situations in which a covariate is believed to have a differential effect depending on whether the tie occurs within a block or between blocks. For example, elementary school

teachers tend to seek advice from teachers in their school, who teach the same grade (Spillane, Kim, & Frank, 2012), but this factor may not be a strong predictor of across-school advice ties. Alternatively, we might find that principals tend to seek advice outside their home schools because they are communicating with other principals, making it unlikely they will ask advice within their own schools. In these situations, we should model the covariates for ties within blocks and across blocks separately. In other situations, we may choose to model covariates as having constant effects across all individuals regardless of block membership.

To accommodate covariates that vary depending on the nature of the tie (whether it is within a block or across blocks), we introduce the a priori random covariate stochastic blockmodel (RCSBM) along with the a posteriori RCSBM. The a posteriori RCSBM is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_i + \beta_1 X_{ij} I_{g_i=g_j} + \beta_2 X_{ij} I_{g_i \neq g_i}\}}{1 + \exp\{g_i^T \text{logit}(B) g_i + \beta_1 X_{ij} I_{g_i=g_j} + \beta_2 X_{ij} I_{g_i \neq g_j}\}}, \\ g_i &\sim \text{Multinomial}(1, \theta), \\ \beta_1 &\sim MVN(\mu_1, \textstyle\sum_1^2), \\ \beta_2 &\sim MVN(\mu_2, \textstyle\sum_2^2), \\ \beta_{lm} &\sim \text{Beta}(a_{lm}, b_{lm}). \end{aligned} \tag{4}$$

Within each of these models, additional specifications are possible depending on the situation. For example, the covariate effect may vary depending on whether the tie occurs within blocks or between blocks, and we could also allow within-block covariate effects to vary by block. Then, $\beta_1 X_{ij} I_{g_i=g_j}$ becomes $\sum_k \beta_{1k} X_{ij} I_{g_i=g_j=k}$. Of course we could further extend this notion of random effects to allow random covariate effects depending on which pair of blocks are involved in the tie for between-block covariate effects, essentially producing a matrix of covariate effects. The possibility of block-level covariates also exists, but block covariates introduce an identifiability issue with respect to B and should be used only if B is fixed. If we fix B, we essentially force the within- and between-block probabilities to be known, which can greatly decrease the utility of the model. But if the interest is really on estimating between- and within-block covariate effects, then the elements of the B matrix merely operate to constrain the parameter space.

Additional parameters can also be added to the model depending on prior specification. For example, we specify $B_{lm} \sim \text{Beta}(a_{lm}, b_{lm})$ and we may want to estimate the hyperparameters a_{lm} and b_{lm}. Furthermore, we might also include additional dependence assumptions among parameters with how we specify prior distributions. For example, we might have reason to model within- and between-block covariate effects as coming from the same distribution. Finally, we can extend these models by adding an additional level of hierarchy. Rather than generating each g_i from the same multinomial distribution (θ), we could

instead generate g_i and g_j for each tie from their own multinomial distribution (θ_i and θ_j, respectively). This extension of the blockmodel was first introduced by Airoldi, Blei, Fienberg, and Xing (2008) and the covariate version by Sweet, Thomas, and Junker (2014).

2.1. Model Estimation

To fit CSBMs, we use a Markov chain Monte Carlo algorithm (MCMC; Gelman, Carlin, Stern, & Rubin, 2004), where the joint likelihood of the model can be written as:

$$P(Y|g,B,\beta)P(B)P(g)P(\beta) = \prod_{i\neq j} P(Y_{ij}|g_i,g_j,B)\prod_{l,m} P(B_{l,m})\prod_i g_i \mathrm{p}(\beta). \quad (5)$$

If block memberships are estimated, they are updated using Gibbs updates and the complete conditional distribution is given as:

$$\begin{aligned} &P(g_i|\ \ldots) \propto \text{Multinominal}(p),\\ &\log p_k = \textstyle\sum_{i\neq j} Y_{ij}[\text{logit}(B) + \beta X_{ij}] - \log\left[1 + \exp\left(\text{logit}(B) + \beta X_{ij}\right)\right], \end{aligned} \quad (6)$$

where p_k is the probability of belonging to group k.

All other parameters in the model are updated using Metropolis–Hastings updates. The regression coefficients β can be updated using a random walk proposal distribution, and depending on the specific model, it may be possible to update hyperparameters with Gibbs updates. To estimate B, we reparameterize B and update instead the log odds of B again using a normal random walk proposal distribution. Without covariates, closed form complete conditionals do exist and B can be updated using a Gibbs step (Sweet et al., 2014).

The MCMC algorithm is coded in *R* (R Development Core Team, 2010) and this code is available as supplemental material. Note that our code is for binary network data only. These models can also be estimated using the *R* package CID-networks (Dabbs, Junker, Sweet, & Thomas, 2014), which uses a probit link function and accommodates any type of tie data.

2.2. Considerations

2.2.1. Applications and models. CSBMs are a useful class of models for researchers interested in identifying and measuring factors associated with network ties in the presence of subgroups. Subgroup structure is common in many social organizations, either through formal organization (e.g., departments in a high school) or as an outcome of social interaction (e.g., proximity or homophily). Given a friendship network of students, students may self-group by grade but choose their friends based on a number of other characteristics such as race, gender, or common interests. We could also estimate attributes associated with friendships

across classrooms. Are these friendships based on other shared activities such as sports teams or neighborhood playgroups or are there certain attributes that these students possess such as age or having a leadership role?

Specifically, these models are best-suited for covariates that are unrelated to the subgroup. For example, consider an advice-seeking network of teachers in a high school and suppose the blocks represent departments. That is, teachers tend to seek out other teachers within their own department for advice. Then, we could use CSBMs to investigate the effects of teacher attributes such as race, gender, experience, and instructional practices on advice-seeking. Because college major tends to be strongly associated with teaching department, including that as an node-level or even as a dyad-level covariate is problematic since the block membership is based on department. This is analogous to the practice of removing collinear covariates in multiple regression (Weisberg, 2005) and does not detract from the utility of the model.

CSBMs are also a flexible class of models. CSBMs can accommodate a variety of block-dependent covariate effects, such as differential effects for within-group ties and between-group ties. In addition, CSBMs can accommodate other random effects for covariates. For example, we might need a model in which covariate effects vary by block. Consider elementary school friendship networks in which students are blocked by grade. We might hypothesize that the effect of gender on friendship ties becomes increasingly positive as students get older, that is, students select students of the same gender for friendship ties, so the effect of the tie-level covariate of "being the same gender" varies across the grade-level blocks. Such a block-dependent covariate could also be accomplished by incorporating block-specific covariates directly into the B matrix.

2.2.2. Estimation and identifiability. One issue common to all a posteriori blockmodel-fitting algorithms is that of label switching. When block membership is unknown, the CSBM algorithm estimates the group identity for all individuals, creating several identical solutions. Such an identifiability issue is handled in a number of ways (Celeux, 1998; Jasra, Holmes, & Stephens, 2005; Stephens, 2000). We choose to postprocess the block memberships and relabel in order of average node ID, which is a simple way to obtain a unique solution. Labeling of blocks becomes an issue when calculating classification rates, especially when block membership is not recovered with near-perfect accuracy.

Another potential identifiability issue that is more particular to CSBMs involves the B matrix and the number of ties across groups, since the inclusion of covariates may increase the number of between-block ties. Separating the effects of the covariate with the estimates of the B matrix may prove challenging in some circumstances, and for our particular applications, we assume between-block tie probabilities to be low (i.e., the off-diagonal elements of the B matrix are small) and suggest using informative priors or even fixing the off-diagonal values of the B matrix to ensure that covariate effects are not affected.

We also consider the challenge of estimating group membership. Local modality is certainly an issue with group membership estimation and we recommend using a clustering algorithm to generate starting values for group membership in the MCMC algorithm. If starting values for group membership are selected randomly or fixed so that all individuals begin in the same group, MCMC convergence in a local mode is more likely and this issue is incensed as subgroup density decreases and the number of blocks increases.

It is unsurprising that increasing the number of parameters in any CSBM increases the number of possible identification issues and we conclude our discussion with the a posteriori RCSBM in which both group membership is unknown and covariate effects differ for across-block and within-block ties. This particular model is likely to be the most difficult to fit of the CSBMs presented in this article. Estimates for covariates are necessarily dependent on block membership estimates and errors in one create errors in the other. These models are likely best-suited for networks with dense subgroups and most likely small networks, for example, a class friendship network or small organization network.

Despite these issues, which are not difficult to circumvent, these models are quite useful. In Section 3, we explore these issues in more detail though simulation studies to illustrate under which conditions these models can and should be used. Then, we present empirical examples of fitting these models in Section 4.

3. Simulations

We conducted three simulation studies to illustrate the utility of these models but also to investigate some underlying operating characteristics that may affect parameter estimation. We first explore whether estimating the B matrix influences covariate effect parameter recovery in a CSBM. In the next simulation study, we explore the effect of covariates on recovering group membership, and in the third study, we explore the effects of subgroup density on recovering group membership. These simulations are not meant to be exhaustive; merely they are an exploration into these models and preliminary results can be used to inform future work in this area.

3.1. Simulation 1: Investigating Possible Identifiability Between-Group Tie Probability Matrix and Covariate Parameters

There may be an identifiability issue when simultaneously estimating the within- and between-group tie probability matrix (B) and regression coefficients. When all elements of B can vary and β is unconstrained, changes in within-block probability of a tie may result from a change in β or a change in entries of B. For example, a positive covariate may increase the number of both within- and between-group ties and this effect may instead be captured in the estimates of the entries of B as opposed to the covariate effect. Thus, to investigate this possibility, we explore covariate parameter recovery when B is fixed or estimated.

We simulated 20-node networks of three subgroups and varied both the effects and the numbers of covariates. Each covariate was then included in our data generating model as both a *sender* covariate and a *receiver* covariate, that is, $X_{ij} = X_i$ for all j for sender covariates and $X_{ij} = X_j$ for all i for receiver covariates. We considered four combinations of one or two pairs of covariate effects: $\beta = (2, -1), (3, 1), (2, 1, -1\ -3), (-2, 1, 1\ -3)$, where sender effects are listed first, and three different covariate variances, $\sigma^2 = 0.1, 1, 25$. Note that these parameter choices include sender and receiver covariates with similar effects as well as opposite effects to reflect a range of effects on between- and within-block tie probabilities. The generative model is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ X_{ij} &\sim N(0, \sigma^2), \\ B &= \begin{pmatrix} 0.25 & 0.01 & 0.01 \\ 0.01 & 0.25 & 0.01 \\ 0.01 & 0.01 & 0.25 \end{pmatrix}, \\ g_i &\sim \text{Multinomial}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \end{aligned} \tag{7}$$

where $\sigma^2 = 0.1, 1, 25$.

We simulated 100 data sets for each simulation for each of the four covariate combinations. We then fit the simulated data set using three different models that differ only with respect to the B matrix estimation. Model 1 assumes the B matrix is known, Model 2 assumes the off-diagonal elements of the B matrix are known, but the diagonal elements must be estimated, and Model 3 assumes the entire B matrix is unknown. Thus, Model 3 is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{\text{logit}\ g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ B_{mm} &\sim \text{Beta}(3, 1), \\ B_{lm} &\sim \text{Beta}(1, 30), \end{aligned} \tag{8}$$

where g_i is not estimated. To assess model fit, we examine parameter recovery. The results from our MCMC algorithm are essentially samples from the posterior distributions for each parameter. Using equal-tailed 95% credible intervals (CIs), we can determine whether a parameter is accurately recovered based on whether it is contained in the 95% CI. For each model fit, we report the coverage probability, that is, the proportion of the 100 simulations that each parameter is recovered.

Table 1 displays the parameter recovery rates and we find that parameters are recovered approximately 88–98% of the time. Coverage probabilities do not

appear to be influenced by the model, which suggests that estimating B does not affect covariate parameter estimation. Coverage may decrease slightly as the number of covariates increases, but it is not obvious from these results. We also note that coverage does not seem to vary based on the variance of the X.

3.2. Simulation 2: Effects of Covariates on Group Membership Recovery

Another potential estimation issue may occur when jointly estimating group membership and covariates. We suspect that increasing numbers of covariates would decrease the accuracy with which we can estimate group membership. Thus, in this simulation, we explore the effects of additional covariates on the group membership estimation.

We generated networks of size 30 and 75 with three and five subgroups, respectively, and included either one, two, or four pairs of sender and receiver covariates. For each combination of group number and covariate effects, we simulated 100 data sets from the following CSBM:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ X_i, X_j &\sim N(0, \sigma^2), \\ B_{mm} &= .2, \\ B_{lm} &= .01, \\ g_i &\sim \text{Multinomial}(\vec{1}/K), \end{aligned} \tag{9}$$

where K is the number of subgroups and $\sigma^2 = .01, .1, 1$.

We then fit the simulated network data and covariates using an a posteriori CSBM. To optimize convergence, we used a community detection random walk algorithm through the `igraph` *R* package (Csardi & Nepusz, 2006) to determine group membership starting values. This particular algorithm, originally introduced by Pons and Latapy (2005), uses a random walk along the network as part of a distance measure between two vertices and then employs a hierarchical clustering algorithm to determine how nodes are clustered into communities, and we constrain the number of clusters to be equal to the number of blocks. Note that there are a variety of other community detection algorithms, for example, `igraph` has eight algorithms including methods by Girvan and Newman (2002) and Raghavan, Albert, and Kumara (2007). For methods that scale to very large networks, see Amini, Chen, Bickel, and Levina (2013).

Our MCMC algorithm generates draws for group membership for each node and we use the posterior mode as the estimated block membership. We then compare the estimated group membership with the true group membership, using an algorithm that permutes the rows to optimize the trace of the classification matrix. This then generates the labels that best align with the true groups. To

TABLE 1.
(Simulation 1) Coverage Probabilities for Covariate Parameters

		True β											
Model	Var(X)	(2, 1)		(3, 1)		(2, 1, −1, −3)				(−2, 1, 1, −3)			
1	0.1	.91	.97	.95	.93	.97	.95	.93	.91	.95	.97	0.92	.92
	1	.92	.95	.95	.94	.93	.98	.91	.98	.95	.91	.93	.98
	25	.91	.91	.93	.95	.95	.94	.93	.93	.91	.92	.93	.95
2	0.1	.92	.97	.93	.95	.96	.95	.96	.91	.95	.95	.90	.87
	1	.89	.94	.92	.93	.95	.96	.91	.96	.94	.90	.93	.96
	25	.90	.92	.94	.95	.90	.93	.93	.94	.90	.96	.91	.94
3	0.1	.92	.95	.92	.94	.94	.95	.97	.91	.93	.96	.88	.90
	5	.92	.97	.94	.97	.92	.95	.90	.97	.95	.90	.89	.95
	25	.92	.95	.95	.95	.94	.95	.96	.94	.93	.94	.93	.96

Note. We fit each network with three models: B is fixed (Model 1), the diagonal elements of B are estimated (Model 2), and all of B is estimated (Model 3). CSBM = covariate stochastic blockmodel.

assess our group membership estimates, we use the classification rate defined as the proportion of nodes that were correctly classified.

Classification rates for node membership are summarized in Table 2. The mean classification rate is determined by taking the average of the 100 classification rates computed for each model fit. Along with the mean, we include the standard deviation (*SD*) and two tie density measures: the within-group and between-group tie density. We define the within-group density as the number of observed ties that occur within the same block divided by the total number of ties within the same block and between-group density as the number of observed ties that occur between different blocks divided by the total number of possible ties between different blocks. This is not based on the model since we use the true block memberships to calculate these values.

Block membership recovery does not vary greatly across conditions, but there are some detectable patterns. When the variance of the covariates increases to 1, block classification rates are slightly lower. Similarly, we also notice that in general, the eight covariate models have the best classification rates, but there isn't a noticeable difference between two and four covariates. The models when $\beta = (2, -1)$ do have slightly lower classification rates than the other two- and four-covariate models, but it's not clear that this is due to the values of the covariate effects, since we do not observe a similar pattern for $\beta = (2, 1, -1, -3)$.

We also note that the five-block model fits have slightly better classification rates than the three-block model fits. Although we believe that increasing the numbers of blocks generally increases the difficulty in estimating block

TABLE 2.
(Simulation 2) A posteriori CSBM Block Membership and Classification Rates and Within- and Between-block Densities

		True β				
Three Blocks Var(X)		(2, 1)	(3,1)	(2, 1, −1, −3)	(−2, 1, −1, −3)	(3, −2, 1, 0, 3, −2, 1, 0)
.01	Mean (*SD*)	.71 (.12)	.72 (.12)	.71 (.11)	.72 (.12)	.74 (.13)
	Within	.20	.20	.20	.21	.22
	Between	.01	.01	.01	.01	.01
.1	Mean (*SD*)	.68 (.12)	.73 (.12)	.71 (.12)	.72 (.12)	.83 (.11)
	Within	.22	.24	.25	.25	.29
	Between	.01	.01	.02	.02	.03
1	Mean (*SD*)	.66 (.10)	.66 (.11)	.67 (.13)	.67 (.12)	.73 (.13)
	Within	.31	.35	.38	.37	.40
	Between	.05	.11	.15	.15	.19
Five Blocks						
.01	Mean (*SD*)	.78 (.11)	.78 (.12)	.77 (.12×)	.76 (.11)	.83 (.10)
	Within	.20	.21	.21	.21	.21
	Between	.01	.01	.01	.01	.01
.1	Mean (*SD*)	.70 (.11)	.81 (.09)	.72 (.12)	.72 (.11)	.89 (.07)
	Within	.22	.24	.25	.25	.27
	Between	.01	.02	.02	.02	.03
1	Mean (*SD*)	.53 (.08)	.69 (.10)	.64 (.11)	.63 (.11)	.77 (.11)
	Within	.31	.35	.37	.37	.41
	Between	.05	.10	.14	.14	.21

membership, the five-block networks had 75 nodes and the three-block networks had 30 nodes. Thus, on average, the five-block networks had 15 nodes per block and the three-block networks has 10 nodes per block. While an increase from 30 nodes to 75 nodes may not seem like a huge leap, this corresponds to an increase in the number of ties from 870 to 5,550. In addition, there appears to be little relationship between tie densities and classification rates. In our experiences fitting noncovariate versions of SBMs, we find classification rates are generally best when there are large differences in within-block and between-block tie densities and the between-block tie density is very low, so it may at first seem surprising that this pattern did not hold for the CSBMs. This is likely due to the fact that covariate effects affect both within-block and between-block tie densities, so the observed within-block and between-block tie densities are not informative in the presence of covariates. The converse is also likely true: When we condition on the covariates, classification rates improve.

We also estimated parameter recovery with the a posteriori CSBM to explore the effects of estimating group membership on covariate parameter recovery. Table 3 reports coverage rates which noticeably decrease as covariate variance increases, and this result aligns with our finding of reduced classification rates with larger variances. A decrease in parameter recovery as covariate variance increases was not observed in Section 3.1, but covariate effect parameter recovery is likely correlated with block membership recovery.

We also notice that $\beta = (2, -1)$ was recovered less often than other covariate effects, which indicates a possible issue with estimating opposing sender and receiver effects. Covariate effects of $(2, -1)$ suggest individuals with larger values of that covariate x are more likely to send ties, but that individuals with larger x are less likely to receive ties. We found that block membership recovery was also smaller, which suggests poor β recovery is due to poor block classification. Such an identification issue does make sense; these covariate effects results in a systematic increase in ties between some individuals and decrease in ties between others, a structure that is modeled through the blockmodel and thus negated when estimating covariate effects. However, in most of the cells, covariate effect parameter recovery was noticeably better than block membership recovery, which suggests the likelihood of an additional or alternative explanation. Finally, we were quite surprised that parameter recovery was better for the three-block model than the five-block model, given the large increase in network size and the higher classification rates for the five-block model. One explanation might be that incorrectly assigning a node to a block has greater negative impacts on covariate parameter estimation, as the number of possible wrong blocks increases.

3.3. Simulation 3: Effects of Subgroup Density and Number of Subgroups on Group Membership Recovery

Networks are generally sparse with subgroups being quite dense. There are situations in which the subgroups are also somewhat sparse, so we investigate what ranges of subgroup density can be used to accurately recover group membership. In this simulation, we consider networks with a variety of numbers of nodes, within-block tie probabilities, and numbers of blocks; we consider networks generated from 3-, 5-, and 10-block CSBMs; and we consider network sizes that are 10 times, 15 times, or 20 times the number of blocks For example, we generate networks with 30, 45, and 60 nodes for a three-block CSBM. The data generating model is given as:

TABLE 3.
(Simulation 2) Covariate Effect Parameter Recovery in the A Posteriori CSBM

	Three-Block			Five-Block		
	Var(X)			Var(X)		
β	0.01	0.1	1	0.01	0.1	1
2	.95	.85	.65	.98	.57	.09
−1	.90	.91	.83	.92	.83	.49
3	.92	.94	.83	.90	.83	.70
1	.96	.97	.93	.91	.93	.91
2	.95	.88	.84	.90	.79	.80
1	.99	.89	.94	.93	.79	.83
−1	.97	.90	.91	.90	.87	.86
−3	.94	.88	.87	.91	.78	.70
−2	.91	.92	.87	.90	.78	.76
1	.97	.93	.89	.93	.77	.82
1	.97	.92	.94	.91	.84	.83
−3	.96	.85	.83	.89	.82	.65
3	.94	.93	.89	.94	.96	.91
−2	.95	.95	.90	.97	.92	.91
1	.91	.96	.91	.95	.97	.89
0	.94	.98	.92	.86	.91	.94
3	.94	.95	.90	.95	.94	.91
−2	.94	.96	.97	.94	.99	.94
1	.96	.95	.92	.97	.92	.95
0	.97	.94	.96	.98	.93	.94

$$
\begin{aligned}
Y_{ij} &\sim \text{Bernoulli}(p_{ij}),\\
p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}},\\
X_{ij} &\sim N(0, .1),\\
g_i &\sim \text{Multinomial}(\vec{1}/G),
\end{aligned}
\tag{10}
$$

where $\beta = (-2, 1, -1, 3)$ for all networks, and we used three pairs of within-group and between-group tie probabilities, $(B_{mm}, B_{em}) = (0.35, 0.01), (0.2, 0.01)$, and $(0.1, 0.001)$. Note that we specifically chose to also vary the between-block tie probabilities to explore the relationship between within-block and between-block tie probabilities and classification rates.

For each combination of network size, number of blocks, and B, we simulated 100 networks. Table 4 shows the mean classification rate and *SD* for block membership, and in general, increasing the number of nodes per block (i.e., increasing network size and keeping the number of blocks fixed) improves block membership estimation. We also notice that block classification rates dramatically improve, as the within-block tie probability increases from .2 to .35, and classification rates do not generally increase when B_{ii} increases from .1 to .2, but this is unsurprising because the off-diagonals of the B matrix are different. One might assume that accurate block membership estimation depends on absolute differences between within-block and between-block tie probabilities, but theoretical results with a two-block SBM suggest that the values of such a threshold depend on network size as well as the within- and between-block tie probabilities (Abbe, Bandeira, & Hall, 2014) and our results appear to align with these findings. When $B_{ii} = .2$, we find within-block densities of .25 and between-block densities of .02 in the generated data given the true blocks. When $B_{ii} = .1$, those densities were .15 and .002, and these pairs of densities resulted in similar block membership estimation rates. One explanation is that as within-block densities decrease, the between-block densities must decrease at a faster rate to recover block membership, and Abbe et al. (2014) report an exact threshold that is similar for the two-block model.

We again consider regression coefficient parameter recovery, and the patterns for covariate effect recovery given in Table 5 are less consistent with the patterns observed for block membership classification rates. For example, we notice that block estimation improves as as the number of nodes per block increases, but this pattern is not consistent for covariate effects. We also saw block classification rates decline as the number of blocks increases while keeping the expected number of nodes per block constant, for example, the first row of Table 5 compares networks with an average of 10 nodes per block and we see a slight decrease in parameter recovery. But for covariate effects, this decline is not steep nor is it constant for every covariate parameter. In fact, we notice such a monotonic decline most with $\beta = 3$, which may be indicative of difficulty of recovering a parameter closer to the boundary of the parameter space as opposed to an issue with parameter recovery. The best evidence that parameter recovery rates align with block membership estimation is when we compare within-block tie densities. For example, comparing covariate effect recovery rates for $B_{ii} = .2$ versus $B_{ii} = .35$, parameter recovery generally appears to be better in the latter set of simulations for most but not all cells.

However, Table 5 indicates that covariate parameter recovery is generally quite good even in situations in which block membership estimation was less than ideal. For example, consider rows 1, 2, and 4 of Table 5. For all cells but one, block membership classification rates are less than .80, but parameter recovery rates suggest decent parameter recovery. These patterns suggest that covariate estimates may indeed be affected by extremely poor block membership, but

TABLE 4.
(Simulation 3) The Average Block Membership Classification Rates and Standard Deviations

	3-Block		5-Block		10-Block	
$B_{ii} = .1$	$n = 30$	.71 (.13)	$n = 50$	.62 (.11)	$n = 100$	.54 (.08)
	$n = 45$	.83 (.12)	$n = 75$	.76 (.11)	$n = 150$	.69 (.09)
	$n = 60$	.92 (.06)	$n = 100$	.89 (.06)	$n = 200$	.83 (.07)
$B_{ii} = .2$	$n = 30$	.72 (.12)	$n = 50$	.70 (.10)	$n = 100$	.54 (.08)
	$n = 45$	.94 (.07)	$n = 75$	.72 (.11)	$n = 150$	.77 (.08)
	$n = 60$	.97 (.02)	$n = 100$	.96 (.03)	$n = 200$	.93 (.03)
$B_{ii} = .35$	$n = 30$	.98 (.03)	$n = 50$	.94 (.06)	$n = 100$	.88 (.06)
	$n = 45$	.99 (.02)	$n = 75$	.97 (.01)	$n = 150$	.98 (.01)
	$n = 60$	1.00 (.01)	$n = 100$	.99 (.01)	$n = 200$	.99 (.01)

TABLE 5.
(Simulation 3) Covariate Parameter Recovery Rates

	3-Block					5-Block					10-Block				
	β	−2	1	−1	3	β	−2	1	−1	3	β	−2	1	−1	3
$B_{ii} = .1$	$n = 30$	.97	.92	.99	.92	$n = 50$	.88	.98	.94	.89	$n = 100$	.83	.95	0.92	.67
	$n = 45$	.83	.94	.93	.93	$n = 75$	.93	.98	.91	.83	$n = 150$	.81	.94	0.91	.63
	$n = 60$	.90	.94	.95	.95	$n = 100$	.93	.93	.97	.87	$n = 200$	.89	.92	1.00	.79
$B_{ii} = .2$	$n = 30$	.92	.93	.92	.85	$n = 50$	.93	.96	.93	.89	$n = 100$	.88	.91	.99	.85
	$n = 45$	.90	.95	.97	.95	$n = 75$	.78	.77	.84	.82	$n = 150$	.87	.97	0.93	.86
	$n = 60$	.95	.94	.93	.96	$n = 100$	.95	.93	.98	.97	$n = 200$	.97	.97	0.97	.92
$B_{ii} = .35$	$n = 30$	.95	.96	.93	.92	$n = 50$	.92	.97	.96	.95	$n = 100$	.98	.96	0.93	.98
	$n = 45$	.91	.98	.97	.94	$n = 75$	.96	.92	.94	.94	$n = 150$	.96	.96	0.96	.94
	$n = 60$	.96	.95	.96	.94	$n = 100$	.92	.95	.92	.95	$n = 200$	.92	.96	0.91	.94

that block membership recovery is not necessary for decent covariate estimation. Finally, we note that we did not observe the poor parameter recovery seen in Simulation 2 with the five-block network and $\beta = (2, -1)$. In this simulation, we used generative values of $\beta = (-2, 1, -1, 3)$, which do not have opposing sender/receiver effects.

4. Fitting CSBMs to Education Data

To illustrate the a priori CSBM and a posteriori CSBM, we use advice-seeking network data (Hopkins, Spillane, Jakopovic, & Heaton, 2013; Spillane &

Hopkins, 2013) taken from one suburban district in the Midwestern United States in 2013, and we use the pseudonym Auburn Park for the name of the district. The data include school staff surveys as well as social network data from 14 elementary schools. Staff members were asked to name the individuals to whom they seek instructional advice and information, and they were able to nominate any other staff member in the district.

The advice-seeking network is given in Figure 1, and this particular representation is simply an abstraction of the adjacency matrix. A black box in row s and column r corresponds to $Y_{sr} = 1$ and indicates the presence of a tie from the individual s to individual r, and the lack of a tie is indicated by the absence of a black box. Note that teachers are ordered by school, so that block structure by school is visible. Due to the large size of the network ($n = 389$), the block structure is more apparent in this representation than in a typical network plot of vertices and arrows.

Note also that the network is not symmetric since advice seeking is not always reciprocated. Thus, in our context, our senders are the individuals seeking advice or information and the receivers are the individuals providing advice or information. The tie then indicates an advice/information relationship from the seeker to the provider.

The network is overall quite sparse, which is unsurprising given the size of the network and the fact that most individuals have few opportunities to interact with one another because they work in different schools. Table 6 shows the number of staff in each school and several measures of density at both the network and school levels. The overall density and within-school and between-school tie densities for the full network of 389 professionals are shown along with within-school tie densities for each school. Note that we define density as the proportion of observed ties out of all possible ties which in this situation is 389×388, and we define within-school tie density as the total number of ties between teachers in the same school divided by the sum of the number of possible ties within each school, that is, $\frac{\sum_{ij}^{n} Y_{ij} I_{g_i = g_j}}{\sum_k n_k (n_k - 1)}$, where g_i and g_j are the school IDs for teacher i and j, and n_k is the number of teachers in school k, so that $\sum_k n_k = N$. The between-school tie density is defined as the number of ties observed between teachers in different schools divided by the total number of possible ties between schools, $\frac{\sum_{ij}^{n} Y_{ij} I_{g_i \neq g_j}}{N(N-1) - \sum_k n_k (n_k - 1)}$.

The overall tie density is very low (.01) with a noticeable difference between within-school (.11) and between-school (.002) densities. Within-school tie density also varies slightly, ranging from .09 to .20, and the higher densities are to some degree associated with smaller schools ($r = -.72$). Despite these low densities, there is still a substantial difference between within-school and between-school tie densities, which suggests that a CSBM is an appropriate model.

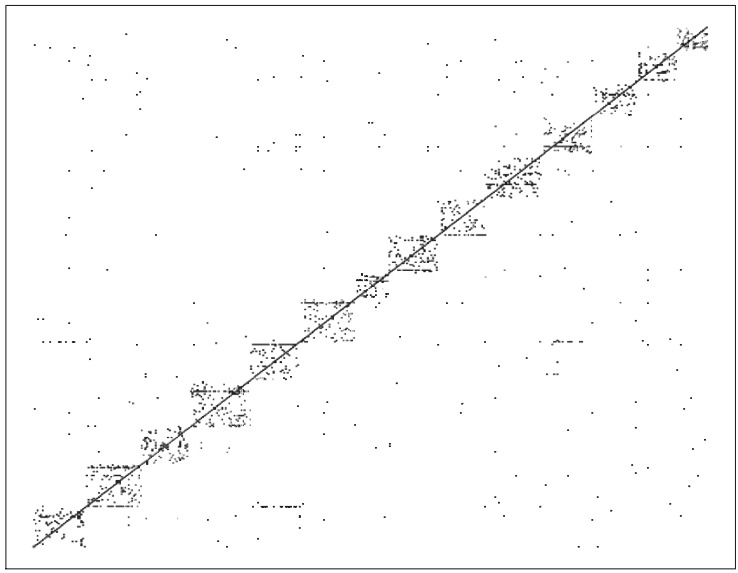

FIGURE 1. *The Auburn Park advice-seeking adjacency matrix represented visually. A black box represents a tie and the presence of a white box represents the absence of a tie. Individuals are ordered by school. Note that within-school ties are more likely than between-school ties although within-school ties are quite sparse even in some of the smaller schools.*

In addition, we explore two measures of node centrality: in-degree and out-degree. Out-degree is the number of ties each node sends and in-degree is the number of ties received by a node. Mean in-degree and out-degree across all 389 nodes are both 3.77, but the distributions of in-degree and out-degree are quite different. Figure 2 show histograms for in-degree (left) and out-degree (right). In general, staff members tend to seek advice from two to six people, whereas the majority of staff are not solicited for advice at all. The staff members who do provide advice tend to provide advice to a large number of staff members.

4.1. Fitting A Priori CSBM to Auburn Park Data

We first fit an a priori CSBM to the advice network and include four individual-level indicator covariates: two are sender-related and two are receiver-related. We have two leader indicator variables: an individual is a content leader and an individual is a noncontent leader, such as a principal. For each of these two variables, we have seeker covariates and receiver covariates. The covariates are then: sender is a content leader, receiver is a content leader, sender is a noncontent leader, and receiver is a noncontent leader. While we expect that

TABLE 6.
Density Estimates for the Auburn Park Advice-Seeking Network: Overall Density and Within-School and Between-School Densities Indicate Teachers Interact With Teachers Within Their Own School More Than Teachers Outside Their School

Density	Within Density	Between Density
.010	.113	.002

School Size (n_k)	Within Density
30	.090
32	.121
28	.114
35	.111
28	.103
32	.978
20	.153
28	.136
28	.094
33	.107
28	.123
26	.089
23	.134
18	.196

Note. Within-school densities for each school are also provided to show variability by school and smaller networks tend to be more dense.

content leaders are no more or less likely to seek advice, we do expect that they would be sought for advice more often than other teachers.

Furthermore, we might hypothesize that these two classes of formal leaders influence advice seeking differentially depending on whether individuals are in the same or different schools, and CSBMs allow us to examine ties across blocks. Thus, we included an indicator Z for whether two individuals are in the same block, that is, $Z_{ij} = 1$ if i and j are in the same block and 0 otherwise. The full CSBM is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ \beta X_{ij} &= (\beta_{1a} X_{1ij} + \cdots \beta_{4a} X_{4ij}) Z_{ij} + (\beta_{1b} X_{1ij} + \cdots \beta_{4b} X_{4ij})(1 - Z_{ij}), \\ B_{mm} &\sim \text{Beta}(2, 1), \\ \beta_i &\sim N(0, 10). \end{aligned} \tag{11}$$

We consider group membership fixed to the home school for each staff member. If teachers divide their time among multiple schools, we used the school

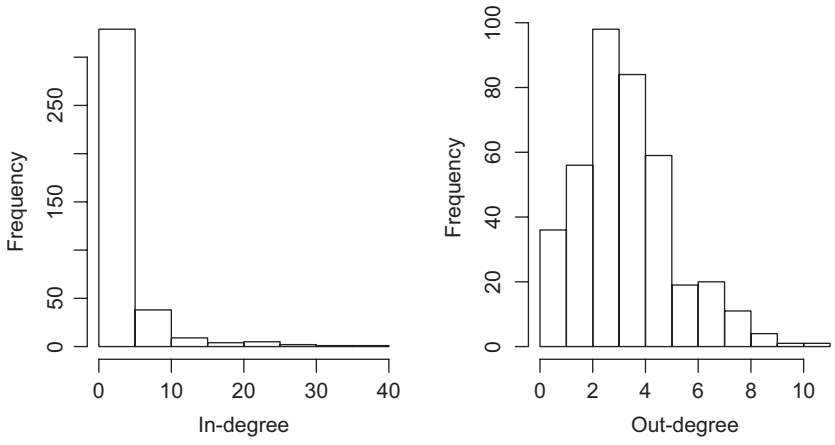

FIGURE 2. *Histograms of in-degree (left) and out-degree (right) suggest that staff members seek advice from several people, but that very few people receive requests for advice and those who do are likely to receive requests from a very large number of people.*

reported by the teacher as their designated school. Note that for this particular model, we only estimate the diagonal entries of the B matrix as parameters and constrain the off-diagonal elements to be fixed at .002. We also chose a fairly uninformative prior, a Beta distribution (2,1), for the diagonal elements. Similarly, we used a weak prior, $N(0,10)$, for the covariate effects. For an adequate posterior sample of 360, we excluded a burn-in sample of 2,000 iterations and retained every 50th sample of our 20,000 step chain.

To estimate regression coefficients in the CSBM, we use *expected a posteriori* (EAP) estimates along with 95% equal-tailed CIs. These estimates are given in Table 7 and can generally be interpreted in a similar way as logistic regression coefficients. As predicted, staff members sought leaders for advice or information throughout the Auburn Park district, but there is a difference between content leaders and formal leaders. Content leaders were much more likely than other staff to be sought for advice within their own school (2.10), but only slightly more likely to be sought for advice from someone at a different school (0.44), whereas noncontent formal leaders were only slightly more likely than other staff to be sought for advice within their own school (0.40) and more likely to be sought for advice from staff in other schools (0.80). Regarding seeking advice, we found that both content and noncontent formal leaders were more likely to seek advice outside of their home schools (1.16 and 2.12, respectively) but were somewhat less likely to seek advice within their home schools (−.28 and −.36, respectively). This suggests that resource and information gathering among leaders likely occurs at the district level between schools and then is disseminated throughout one's home school.

TABLE 7.
Posterior Means (EAP) and Equal-Tailed Credible Intervals From the A Priori CSBM Fit

		Within Block		Between Block	
Covariate		EAP	95% CI	EAP	95%CI
Content leader	Sender	−0.28	[−0.58, 0.01]	1.16	[0.85, 1.45]
	Receiver	2.10	[1.92, 2.28]	0.44	[0.18, 0.70]
Other leader	Sender	−0.36	[−0.53, −0.19]	2.12	[1.90, 2.35]
	Receiver	0.40	[0.24, 0.54]	0.80	[0.56, 1.04]

4.2. Fitting a Posteriori CSBM to Auburn Park Data

When block memberships are unknown, we can use an a posteriori CSBM to estimate them. To illustrate fitting this model, we again use the Auburn Park advice network data. To explore how group membership is recovered, we now treat covariate effects as being constant across all blocks, but respective covariates are the same as before, indicators that the sender is a content leader, receiver is a content leader, sender is a noncontent leader, and receiver is a noncontent leader.

The fitted a posteriori CSBM is given as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} &= \frac{\exp\{g_i{}^T \text{logit}(B) g_j + \beta X_{ij}\}}{1 + \exp\{g_i{}^T \text{logit}(B) g_j + \beta X_{ij}\}}, \\ \beta X_{ij} &= \beta_1 X_{1ij} + \ldots + \beta_4 X_{4ij}, \\ g_i &\sim \text{Multinomial}\left(1, \frac{1}{14}\right), \\ B_{mm} &\sim \text{Beta}(10, 80), \\ \beta_i &\sim N(0, 10), \end{aligned} \tag{12}$$

where the group membership prior distribution is the multinomial distribution with equal probability of belonging to one of the 14 groups. We use a stronger prior to aid with block estimation and use a Beta distribution centered at the average within-group tie density. For covariate effects, we again use a weak normal prior with a large variance.

We first examine group membership estimates. The posterior distributions for g for the first 60 individuals in the network are shown in Figure 3. The first 30 individuals all report belonging to the same school, but group membership is unclear for these individuals. The posterior distributions imply that many of these individuals appear to belong to 2–3 groups or even more groups for some extreme cases. Group membership is more clear for Individuals 31–60, almost all of whom are estimated to belong to Group 13 with high probability.

Our classification matrix showing the true block (school) membership for each individual versus the estimated block membership is given in Table 8.[1] The classification rate is 62.7%, which is similar to the classification rates found in Section 3.3.

Furthermore, Table 8 reveals that some of the schools are being grouped together as one school and others are being divided. For example, individuals in Schools 3 and 9 are estimated as belonging to the same school. The same is true for Schools 5 and 12 and Schools 7, 11, and 13. Such a lack of distinction between blocks is likely due to both the low within-school density and the large number of blocks. For example, both Schools 5 and 12 have slightly lower within-school density and School 5 has relatively high across-group density. In other circumstances, it is less clear. For example, there does not appear to be a particular reason that School 7 is grouped with Schools 11 and 13. As a result of these consolidations, the model estimates that the 389 school staff belong to 10 or 11 blocks.

To further expound on block estimation, we also calculated lower bounds on classification rates to perhaps illustrate that the classification rates are not terrible. In fact, if block membership were randomly assigned, we'd expect a classification rate near .155, a value empirically derived through Monte Carlo simulation. Thus, our model is detecting a good bit of subgroup structure. Another metric we could use is the proportion of ties that are correctly assigned as belonging to the same or different groups. We found that 93.7% of our pairs of nodes were correctly classified as belonging to either the same group or different groups, as compared to an expected value of 82.5% when group membership is assigned at random. Regardless of which measure we use, the a posteriori CSBM is able to correctly estimate some of the observed subgroup structure.

Despite underwhelming classification rates, the covariate effects appear to be well-recovered. To compare the effects of estimating block membership on covariate effects, we also fit the analogous a priori CSBM using the same four covariates. The results from each model are given in Table 9. In general, the covariate estimates from the a posteriori CSBM are very similar to the analogous estimates from the a priori CSBM, and the 95% CIs overlap, substantially in some cases. The a posteriori CSBM 95% CIs are slightly wider than the corresponding a priori CSBM intervals which is unsurprising given in the additional parameters in the a posteriori CSBM. In addition, we compare B_{mm} in the a priori CSBM and a posteriori CSBM model fits. The mean EAP estimates for B_{mm} are .09 and .06 for the a priori and a posteriori CSBM fits, respectively. Although entries of the B matrix adjusted for some of the block misclassification, we also notice that the a posteriori CSBM point estimates and corresponding intervals are slightly more positive than the a priori CSBM, suggesting that incorrect classification of individuals may bias covariate effects.

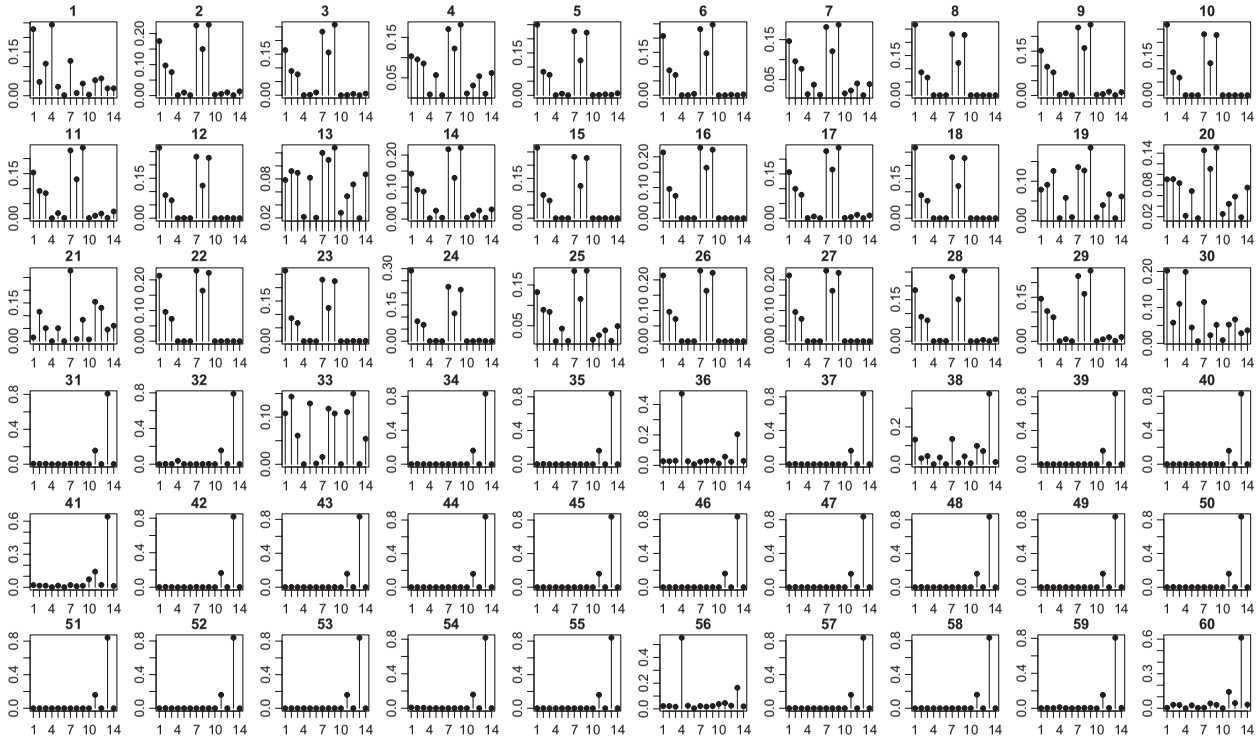

FIGURE 3. *Group membership posterior distributions for the first 60 individuals in the network. In many cases, group membership in a single group is clear, but for many other individuals, group membership appears to divided between several groups or in some extreme cases among all 14 groups. Note that Individuals 1–30 report belonging to the same school and Individuals 31–60 report belonging to the same school.*

TABLE 8.
Classification Matrix Showing the True Block (School) Membership for Each Individual Versus the Estimated Group Membership

		School ID													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Estimated School ID	1	27	0	1	0	4	1	1	0	0	0	5	0	0	0
	2	0	28	0	0	1	0	0	0	0	0	0	0	0	1
	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	4	2	2	0	31	0	1	1	0	0	1	0	1	0	0
	5	0	0	0	0	5	0	0	0	0	0	3	0	0	0
	6	0	0	0	0	0	27	6	0	2	0	0	0	2	0
	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	1	1	0	1	0	2	0	19	0	0	0	1	0	0
	9	0	1	24	0	0	1	0	9	23	0	3	2	1	0
	10	0	0	0	0	0	0	0	0	1	32	0	0	0	0
	11	0	0	1	1	2	0	1	0	1	0	1	1	1	0
	12	0	0	0	0	15	0	0	0	0	0	2	14	0	0
	13	0	0	0	1	1	0	11	0	1	0	14	0	19	0
	14	0	0	1	1	0	0	0	0	0	0	0	7	0	17
n_{school} WSD		30	32	28	35	28	32	20	28	28	33	28	26	23	18
		0.09	0.12	0.11	0.11	0.10	0.10	0.15	0.14	0.09	0.11	0.12	0.09	0.13	0.20

Note. The school size (n_{school}) and WSD are also shown. The classification rate is 62.7%, mainly due to several schools being estimated as a single school. WSD = within-school density.

TABLE 9.
Regression Coefficient Estimates for Indicator Variables

Covariate		A Priori CSBM		A Posteriori CSBM	
Content leader	Sender	0.27	[0.05, 0.50]	0.36	[0.12, 0.62]
	Receiver	2.24	[2.09, 2.38]	2.42	[2.22, 2.59]
Other leader	Sender	−0.11	[−0.25, 0.04]	0.06	[−0.13, 0.25]
	Receiver	0.57	[0.43, 0.71]	0.81	[0.64, 0.98]

5. Discussion

We presented a class of models, CSBMs, which allows researchers to model social network data in the presence of existing subgroups or blocks. Students and teachers in schools often form subgroups, either through formal organization, such as departments or schools, or through informal selection, such as working in the same area or teaching the same group of students. While models that relate individual traits to these subgroups are useful, researchers are also interested in the relationships that form or persist in the presence of this subgroup structure, and CSBMs are particularly suited for these types of research questions.

One particularly novel aspect of the CSBM framework is the ability to estimate covariate effects between subgroups, which has not been addressed in the education literature; in fact, a similar model was only recently introduced in the statistical data mining literature (Airoldi et al., 2011). Because CSBMs incorporate covariates independently of subgroup structure, they are unique in that covariates can vary both between and within blocks. In fact, understanding why individuals form ties outside of their subgroup (and with whom) is of particular interest to education researchers (Spillane, Hopkins, & Sweet, in press).

In addition to illustrating model feasibility with real network data, we conducted several simulation studies not only to further demonstrate the utility of these models in practical settings but also to show how robust these models are, specifically with respect to covariate parameter recovery. For example, we hypothesized a possible identifiability issue with estimating the B matrix, but we found little effect of either estimating B or the variance of the covariate on covariate parameter recovery. We did find that parameter recovery generally decreases when estimating group membership, especially when group membership estimation is poor, but parameters are generally recovered with decent probability, even as the variance of the covariates increases. Finally, we found that very low within-group tie density negatively impacted block membership estimation. Our exploration with a real-world data set also corroborated these findings, since covariate effects were quite similar under the a priori CSBM and the posteriori CSBM, where group membership was not well recovered.

Regarding methodological research, there are many open areas for future work. Given the proposed models, we examined block membership recovery and covariate parameter recovery under certain conditions, but these simulations could be extended. In particular, we employed equal-sized blocks, and one open question is whether variability in relative block size affects block membership estimation. In the Auburn Park example, blocks sizes varied and we found that block size did not appear to be related to block recovery. In fact, the smallest school was recovered fairly well (1 node was misclassified and 7 other nodes were incorrectly estimated as belong to this school). In our own experience, we find within-block density to be the strongest predictor of block recovery even when working with blocks very different in size. Generally, if the block has at least a few nodes and the within-block density is above a certain threshold, it will be recovered even when among blocks several times larger in size.

Another area of research is on block-varying covariate effects. We fit models in which covariate effects varied within and between blocks, but certainly there are situations in which we would want covariates to vary across all combinations of interaction. Our simulation studies were largely not affected by identifiability issues, but identifiability is not a trivial issue and determining the circumstances in which parameter recovery is compromised is very important. Focus should also be paid when covariates conflate with block structure. Finally, we assumed independence between block membership and covariates when generating data, but we might consider how including a relationship between block membership and covariates affects not only parameter estimation but also model fit in general.

Other models are also possible within this framework. We may be interested in introducing a mixture component and investigate covariate recovery when using a mixed membership stochastic blockmodel rather than a stochastic blockmodel. Sweet et al. (2014) introduced such a model but did not explore any of the operating characteristics or consider possible estimation issues. Other models for clustered network data, such as the latent position cluster model (Handcock, Raftery, & Tantrum, 2007), could also be considered.

The CSBM is a conditionally independent network model in that we assume the ties are independent of one another given the parameters in the model (covariate effect and block membership). In fact, one of the unique benefits to fitting network data with conditionally independent network models is that covariates, random effects, and latent structural components such as blocks or latent spaces (Hoff et al., 2002) can be combined in additive ways. Depending on the network structure, assumptions about the network nodes and ties, and specific research question, we could combine these components to build a very specific social network model to address that research question. There is currently some work in this area (Dabbs, Adhikari, et al., 2014; Dabbs, Junker, et al., 2014) and we believe this to be an area of active research in the future.

Modeling social network ties as independent is a poor assumption, as these relationships rarely form or persist in isolation. In the case of the stochastic blockmodel, we believe that incorporating relevant covariates into the model will improve the model in two important ways. First, combining covariates and a stochastic blockmodel should improve the power of the model to detect either covariate effects and block membership, and this work is currently being explored in Dabbs, Adhikari, et al. (2014). Second and perhaps more importantly, adding relevant covariates makes the assumptions of the model more believable. Network ties, particularly in a large network such as Auburn Park (Section 4), are unlikely to be independent conditional on only block membership; it is more accurate to model ties as independent given both block membership and several important covariates. In fact, Spillane et al. (in press) use sender, receiver, and edge-level covariates in their CSBM.

Thus, CSBMs have great potential to inform future educational research involving relational or network data in addition to methodological research on social networks. Education researchers have a class of flexible models that can both cluster individuals and provide information about relations within and between these clusters. For example, Spillane, Hopkins, and Sweet (in press) fit CSBMs to illustrate differences in covariate effects within and across schools with regard to how advice and information is shared throughout the district. They found that subject-specific leaders tend to seek advice outside of their own schools but rarely within their own schools even though they are much more likely to provide advice than other staff. Understanding how social resources are shared within a school district then informs district policies and decisions. Since school leaders tend to obtain information outside of their schools, district leaders may encourage organizational routines to enable these interactions. Finally, these models are especially easy to understand because the covariate effects can be interpreted in a similar way as generalized linear model coefficients, since covariates are incorporated in an additive way.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was in part supported by the Institute for Education Sciences under Award #R305D12004.

Note

1. Note that these labels are not the same as those described in Figure 3.

Supplemental Material

The online data supplements are available at http://jeb.sagepub.com/supplemental

References

Abbe, E., Bandeira, A. S., & Hall, G. (2014). Exact recovery in the stochastic block model. *arXiv preprint arXiv*:1405.3267.

Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, *9*, 1981–2014.

Airoldi, E. M., Choi, D. S., & Wolfe, P. J. (2011). Confidence sets for network structure. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *4*, 461–469.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, *41*, 2097–2122.

Anderson, C., & Wasserman, S. (1992). Building stochastic blockmodels. *Social Networks*, *14*, 137–161.

Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In P. J. Green & R. Rayne (Eds.), *COMPSTAT* (pp. 227–232). Heidelberg, Germany: *Springer*.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*.

Dabbs, B., Adhikari, S., Thomas, A. C., Sweet, T., Sadinle, M., & Junker, B. W. (in press). *Conditionally independent Dyad network models*.

Dabbs, B., Junker, B. W., Sweet, T. M., & Thomas, A. C. (2014). *CIDnetworks: Generative models for complex networks with conditionally independent dyadic structure*, *R package version 3.0.0*. Pittsburgh, PA: Carnegie Mellon University.

Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education*, *77*, 148–171.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. London, England: CRC Press.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*, 7821–7826.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*, 301–354.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*, 1090–1098.

Holland, P., Laskey, K., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*, 109–137.

Hopkins, M., Lowenhaupt, R., & Sweet, T. M. (2015). Organizing instruction in new immigrant destinations: District infrastructure and subject-specific school practice. *American Educational Research Journal*, *52*, 408–439.

Hopkins, M., Spillane, J. P., Jakopovic, P., & Heaton, R. M. (2013). Infrastructure redesign and instructional reform in mathematics. *The Elementary School Journal*, *114*, 200–224.

Jasra, A., Holmes, C., & Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, *20*, 50–67.

Moolenaar, N., Daly, A., & Sleegers, P. (2010). Occupying the principal position: Examining relationships between transformational leadership, social network position, and schools' innovative climate. *Educational Administration Quarterly*, *46*, 623.

Nowicki, K., & Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*, 1077–1087.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In P. Yolum, T Gungor, F Gurgen, & C. Ozturan (Eds.), *Computer and information sciences-ISCIS 2005* (pp. 284–293). Berlin, Germany: Springer.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*, 036106.

Spillane, J. P., & Hopkins, M. (2013). Organizing for instruction in education systems and school organizations: How the subject matters. *Journal of Curriculum Studies*, *45*, 721–747.

Spillane, J. P., Hopkins, M., & Sweet, T. (in press). Intra- and inter-school instructional interactions: Exploring conditions for instructional knowledge production within and between schools. *American Journal of Education*.

Spillane, J. P., Kim, C. M., & Frank, K. A. (2012). Instructional advice and information providing and receiving behavior in elementary schools exploring tie formation as a building block in social capital development. *American Educational Research Journal*, *119*, 72–102.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*, 795–809.

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, *38*, 295–318.

Sweet, T. M., Thomas, A. C., & Junker, B. W. (2014). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In E. Airoldi, D. Blei, E. Erosheva, & S. Fienberg (Eds.), *Handbook on mixed membership models and their applications (Chap*. *22*, pp. 463–488). Boca Raton, FL: Chapman & Hall/CRC.

Tallberg, C. (2004). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, *29*, 1–23.

Wasserman, S., & Anderson, C. (1987). Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, *9*, 1–36.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika*, *61*, 401–425. doi: 10.1007/BF02294547

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). New York, NY: John Wiley.

White, A., & Murphy, T. B. (2014). Mixed-membership of experts stochastic blockmodel. *arXiv:1404.0221*[stat.CO].

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology*, *81*, 730–780.

Author

TRACY M. SWEET is an Assistant Professor in the Department of Human Development and Quantitative Methodology, University of Maryland, 1230A Benjamin Building, College Park, MD 20742, USA; email: tsweet@umd.edu. Her research focuses on developing statistical social network models for education applications.

Manuscript received October 21, 2014
Revision received July 10, 2015
Accepted August 6, 2015