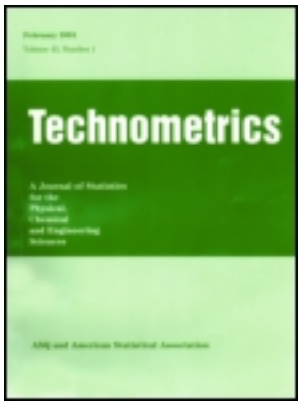



Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

# Bayesian Nonparametric Models for Community Detection

Jiqiang Guo [a], Alyson G. Wilson [b] & Daniel J. Nordman [c]
[a] Virginia Bioinformatics Institute, Virginia Tech University, Blacksburg, VA, 24061
[b] Department of Statistics, North Carolina State University, Raleigh, NC, 27695
[c] Department of Statistics, Iowa State University, Ames, IA, 50011
Accepted author version posted online: 28 May 2013.Published online: 22 Nov 2013.

To cite this article: Jiqiang Guo, Alyson G. Wilson & Daniel J. Nordman (2013) Bayesian Nonparametric Models for Community Detection, Technometrics, 55:4, 390-402, DOI: 10.1080/00401706.2013.804438

To link to this article: http://dx.doi.org/10.1080/00401706.2013.804438

# Bayesian Nonparametric Models for Community Detection

**Jiqiang GUO**

Virginia Bioinformatics Institute,
Virginia Tech University
Blacksburg, VA 24061
(*jqguo@vbi.vt.edu*)

**Daniel J. NORDMAN**

Department of Statistics
Iowa State University
Ames, IA 50011
(*dnordman@iastate.edu*)

**Alyson G. WILSON**

Department of Statistics
North Carolina State University
Raleigh, NC 27695
(*alyson_wilson@ncsu.edu*)

We propose a series of Bayesian nonparametric statistical models for community detection in graphs. We model the probability of the presence or absence of edges within the graph. Using these models, we naturally incorporate uncertainty and variability and take advantage of nonparametric techniques, such as the Chinese restaurant process and the Dirichlet process. Some of the contributions include: (a) the community structure is directly modeled without specifying the number of communities a priori; (b) the probabilities of edges within or between communities may be modeled as varying by community or pairs of communities; (c) some nodes can be classified as not belonging to any community; and (d) Bayesian model diagnostics are used to compare models and help with appropriate model selection. We start by fitting an initial model to a well-known network dataset, and we develop a series of increasingly complex models. We propose Markov chain Monte Carlo algorithms to carry out the estimation as well as an approach for community detection using the posterior distributions under a decision theoretical framework. Bayesian nonparametric techniques allow us to estimate the number and structure of communities from the data. To evaluate the proposed models for the example dataset, we discuss model comparison using the deviance information criterion and model checking using posterior predictive distributions. Supplementary materials are available online.

KEY WORDS: Chinese restaurant process; Dirichlet process: Markov chain Monte Carlo; Networks; Partition.



## 1. INTRODUCTION

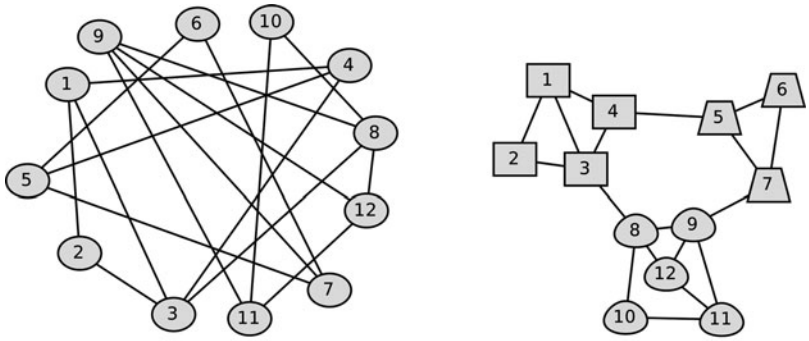
Many systems and organizations can be represented by *graphs* (*networks*) that are useful for studying social and scientific problems. In a network, we use *nodes* and *edges* to model entities and relationships, respectively. Here, we consider simple networks, where the nodes do not have attributes, the edges do not have directions or weights, and there are no self-loops or multiple edges.

Networks have proven to be useful in studying a wide variety of applications. Newman (2003), for example, classified networks into social networks (e.g., groups of people and their interactions), information networks (e.g., academic citations, World Wide Web), technological networks (e.g., the electrical grid), and biological networks (e.g., metabolic pathways or ecosystems). Because of their ubiquity, research on networks has been active from many different perspectives; see Newman (2003), Fortunato (2010), Karrer and Newman (2011), and Ball, Karrer, and Newman (2011) for reviews.

An interesting phenomenon exhibited by many networks is that the vertices can be partitioned into *communities*, where the nodes within a "cluster" often share more edges with each other than with nodes outside the cluster (Newman 2003, sec. 3.6). However, in most cases, the community structure is unknown. One important goal of studying networks is to develop approaches to identify communities using data on graph edges. Figure 1(a) presents a network example with three communities, which are difficult to see until the same graph is reorganized in Figure 1(b).

Community detection has been mainly studied in the computer science and physics literature (Girvan and Newman 2002; Hofman and Wiggins 2008; Lancichinetti, Fortunato, and Radicchi 2008; Leskovec et al. 2008; Karrer and Newman 2011), with relatively few articles taking a statistical point of view (although see Hofman and Wiggins 2008; Ball, Karrer, and Newman 2011; Karrer and Newman 2011). Some statistical approaches have used stochastic block models with a fixed (i.e., known) number of blocks (Nowicki and Snijders 2001; Ball, Karrer, and Newman 2011; Karrer and Newman 2011) and combined them with Bayesian analysis to perform community detection (cf. Kemp et al. 2006; Airoldi et al. 2008); we will return to discuss this literature shortly. However, many community detection algorithms lack "full" statistical developments in that they provide only point estimates of communities, while


TECHNOMETRICS, NOVEMBER 2013, VOL. 55, NO. 4
DOI: 10.1080/00401706.2013.804438




(a) A simple network example (b) Network of (a) reorganized

Figure 1. An illustration of community detection.



failing to address the uncertainty associated with the estimated community structures. These deficiencies suggest the potential for further statistical contributions.

We propose a series of Bayesian nonparametric statistical models to describe networks and perform community detection. These models employ Bayesian nonparametric models, such as the Chinese restaurant process (CRP; as in Kemp et al. 2006) and the Dirichlet process (DP) so that we do not need to specify the number of communities a priori. Bayesian inference for these models is possible via Markov chain Monte Carlo (MCMC), which allows the use of posterior samples for estimating community structures (and other model features), quantifying statistical significance, and assessing model fits.

We motivate our models and statistical methods using a well-known dataset as a benchmark: the network of National Collegiate Athletic Association (NCAA) American football games between Division I-A colleges during the regular season of Fall 2000, compiled by Girvan and Newman (2002). Most colleges in Division I-A participate in athletic conferences, which are groups of teams that play competitively against each other. Teams typically play at least half of their games within their conferences. The dataset is presented in Figure 2, and summaries of the number of edges per node and per community appear in Tables 1 and 2 of the supplementary materials. To apply community detection techniques, we suppose that we know only the 613 games (edges) between the 115 teams and use edge information to infer which teams form communities. We take their athletic conference memberships to be the "truth" of their community membership. This dataset does not include post-season games, so each team could play another at most one time.

To give some background, Hofman and Wiggins (2008) proposed and fit a statistical model for this football dataset; essentially the same model also appears in Ferry, Bumgarner, and Ahearn (2011) for community detection. However, we have found that their model is not able to capture several key characteristics of the data, resulting in some discrepancies between estimated communities and the truth. Our approach is to develop and motivate a series of statistical models, beginning from an initial data-generating model that resembles the model in Hofman and Wiggins (2008). However, we estimate our models in a fundamentally different manner than Hofman and Wiggins (2008) by conducting a fully Bayesian analysis (MCMC based). Comparing the estimated results from this initial model with the conference memberships indicates deficiencies with this model. To overcome these problems, we extend the initial model in two directions:

1. We generalize the notion of communities as a model parameter corresponding to a partition of the nodes, also allowing for some nodes to *fall outside communities altogether*.
2. We consider more flexibly parameterized models for the presence/absence of graph edges, which are built around, and may vary with, the community structures. In particular, some models allow for the probability of edges forming among nodes in a community to vary across communities, and also allow the probability of edges forming between nodes in different communities to vary across pairs of communities. Additionally, the within- and between-community edge probabilities may be modeled *differently*, which is often a realistic distinction to draw. In a given analysis one might, for example, anticipate stronger edge formations within communities than between communities.

All of the models that we consider are structured cases of stochastic block models that use Bernoulli trials to describe random edges between each pair of nodes in the graph. In a related work, Kemp et al. (2006) proposed a broad stochastic block model formulation, using a CRP to model partitions of the graph nodes. The most general model considered here is an extension of a model considered in Kemp et al. (2006), but with an important distinction that we allow probabilities for edges within a community to be described differently than edges between communities, while the model considered in Kemp et al. (2006) does not (see Section 4.2). In fact, even in the most simple model considered here from Hofman and Wiggins (2008) makes a distinction that intra- and intercommunity edges may behave differently, and this general notion often underlies the concept of community detection (cf. Fortunato 2010). Allowing such differences in formulating edge probabilities can improve model fits in community detection, as will be illustrated with the football dataset. Beyond this, the work of Kemp et al. (2006) does not allow nodes to exist outside of a community structure and the Bayesian inference approach differs substantially. Whereas we outline steps for full MCMC-based Bayesian inference in

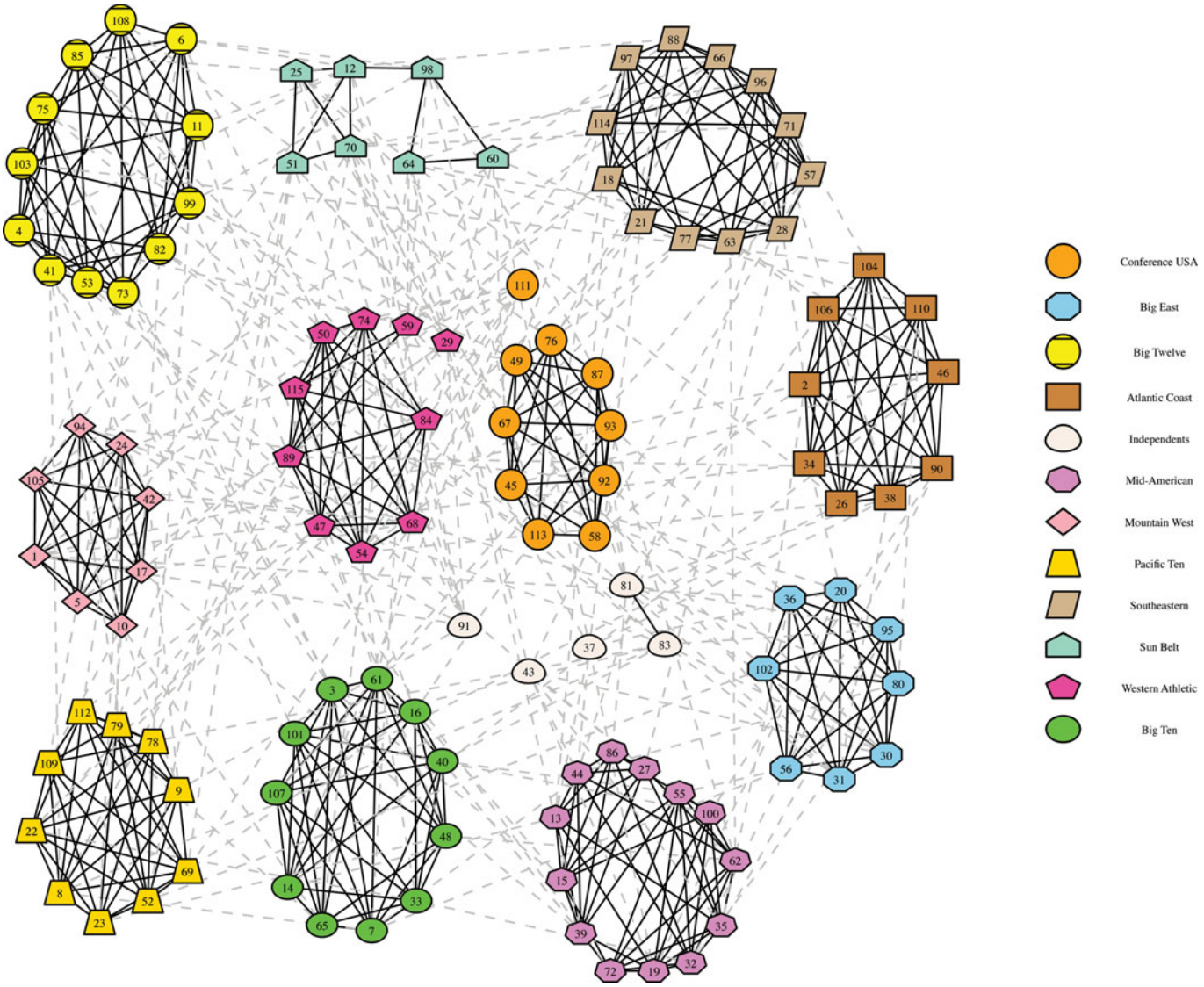



Figure 2. NCAA Division I-A football network data. (There are a total of 115 teams and 613 games in the regular season of Fall 2000.) The online version of this figure is in color.



our series of probability models and address point estimation of community structure from a decision theoretical framework, Kemp et al. (2006) determined a point estimator of community structure through a posterior maximization algorithm. Without posterior samples, their approach offers fewer numerical options in quantifying the estimation uncertainty in general and limits the capacity for model assessment. We examine issues of model fit and diagnostics, which are often overlooked in community detection.

The remainder of the article is organized as follows. In Section 2, we propose an initial model for community detection and point out where this model is lacking for approximating the football data. We describe a richer partition model in Section 3 and present a progression of more flexible models for describing the edge probabilities within and across communities in Section 4. In Section 5, we discuss model selection issues and goodness of fit. In Section 6, we discuss extensions and future research directions. In the supplementary materials, we give a short summary of several results related to the CRP and the DP.

## 2. AN INITIAL MODEL FOR COMMUNITY DETECTION

### 2.1 Partitions as Parameters

A common goal in community detection is to partition the node set $V = \{1, \ldots, n\}$ into nonoverlapping, nonempty collections of nodes. That is, we are looking for $K$ communities, denoted by sets $V_1, V_2, \ldots,$ and $V_K$, such that

$$V_i \neq \emptyset \text{ for } i = 1, \ldots, K; \quad V_i \cap V_j = \emptyset \text{ for } i \neq j; \quad \text{and } \cup_{j=1}^{K} V_j = V.$$

In our models, $K$ is unknown a priori, as is the membership of nodes in communities. For brevity, we denote a specific way to partition $V$ by $\boldsymbol{\pi}$. To avoid ambiguity, we refer to a *community* as one of the sets $V_k$ defining the *partition* $\boldsymbol{\pi}$. Alternatively, we can label each node in community $V_k$ with "$c_k$" ($c_k \neq c_{k'}$ if $k \neq k'$) and then use labels to represent a partition, that is, $\boldsymbol{\pi} \equiv$

$(c_1, c_2, \ldots, c_n)$. At times, we use both set and label notation to represent partitions.

In the following, a partition is considered to be a model parameter. Identifying communities, therefore, amounts to estimating $\boldsymbol{\pi}$. With this parameter and prior specification, an issue then arises in specifying a distribution over the support of $\boldsymbol{\pi}$ (all possible partitions). If the number of communities is fixed, the distribution on the resulting partitions (a restricted subset of the whole partition space) is induced by specifying distributions for the node labels over a fixed number $K$ of possibilities $\{1, \ldots, K\}$. Instead of fixing the number of communities in a partition a priori, we propose to use the CRP to specify a distribution for the partition $\boldsymbol{\pi}$ on the whole partition space; see also Kemp et al. (2006).

### 2.2 Chinese Restaurant Process

The CRP characterizes a random partition obtained from a sequence of "customers" sitting at tables in a restaurant (Pitman 2006, chap. 3.1). The first customer sits at a first table. After $n$ customers have been seated, the $(n+1)$th customer sits at an occupied table or a new table with probabilities $\frac{n_k}{\alpha+n}$ and $\frac{\alpha}{\alpha+n}$, respectively, where $n_k$ is the number of customers sitting at table $k$. (See the supplementary material for an illustrative example.) Customers at the same table form a community, and thus a partition is constructed.

For $n$ nodes, the CRP specifies a distribution for the partition, denoted by $\boldsymbol{\pi} \sim \text{CRP}_n(\alpha)$ ($n$ subsequently omitted), with probability mass function given by

$$\Pr(\boldsymbol{\pi} = (c_1, c_2, \ldots, c_n)) = \frac{\alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)!}{[1+\alpha]_{n-1}}, \quad \alpha > 0, \tag{1}$$

where $K$ is the number of tables occupied, $n_k$ is the number of customers at table $k$, and $[1+\alpha]_{n-1} = \prod_{i=1}^{n-1}(\alpha + i)$. Large values of $\alpha$ imply higher probabilities of occupying new tables in the CRP, and hence assign higher probabilities to partitions with more communities.

### 2.3 An Initial Statistical Model for Community Detection

To describe a model for the data-generating process (i.e., edges in a graph), suppose that a partition "informs" the presence/absence of graph edges. Let $\mathcal{M}$ denote the set of pairs of nodes, that is, $\mathcal{M} \equiv \{(i, j) : i < j;\ i, j \in \{1, 2, \ldots, n\}\}$. For any $(i, j) \in \mathcal{M}$, a binary variable $X_{ij} = 1/0$ indicates edge presence/absence between nodes $i, j$. Conditioning on a partition, the edge presences for all $(i, j) \in \mathcal{M}$ are assumed to be independent (IND) Bernoulli trials (i.e., $\Pr\{X_{ij} = 1\} = p_{ij}$). As a first intuitive model, if nodes $i$ and $j$ are in the same community, we let $p_{ij} = p_{\text{in}} \in (0, 1)$; otherwise, $p_{ij} = p_{\text{out}}$. That is, $p_{\text{in}}$ represents an intracommunity edge probability (considering nodes inside a community) and $p_{\text{out}}$ represents an intercommunity edge probability ("out" refers to considering a node outside of a given community). For illustration in this *initial model*, we impose a parameter constraint $p_{\text{out}} < p_{\text{in}}$, which is appropriate for reflecting stronger connections inside communities than between communities (cf. Ferry, Bumgarner, and Ahearn 2011, p. 1742). Alternatively, one can estimate $p_{\text{in}}$ and $p_{\text{out}}$ without constraints, as in Hofman and Wiggins (2008). (In all other models to follow, we allow intracommunity and intercommunity edge probabilities to simply differ, and leave the data to decide a stochastic ordering; the intracommunity edges probabilities typically emerge as larger for the football data.) Summarizing the initial model, we have

$$X_{ij} \mid p_{\text{in}}, p_{\text{out}}, c_1, \ldots, c_n \overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \quad 1 \le i < j \le n, \tag{2}$$

$$\boldsymbol{\pi} \equiv (c_1, c_2, \ldots, c_n) \sim \text{CRP}(\alpha), \quad \alpha > 0, \tag{3}$$

where $0 < p_{\text{out}} < p_{\text{in}} < 1$ and $p_{ij} = p_{\text{in}} \mathbb{1}_{[c_i = c_j]} + p_{\text{out}} \mathbb{1}_{[c_i \neq c_j]}$.

An Erdös–Rényi (ER) model is often used to model "completely random" graphs (see Fortunato 2010). In this model, the edges are formed for all pairs of nodes as independent and identically distributed (IID) Bernoulli trials with probability $p \in (0, 1)$. Because the edge presence probability between any two nodes is equal, with no preference in forming edges among subgroups of nodes, an ER graph has no community structure. From this perspective, we can test whether there is community structure by testing hypothesis $H_0$: $p_{\text{in}} = p_{\text{out}}$ under this initial model.

For this initial model, the supplementary material describes an MCMC algorithm for obtaining posterior samples based on the following priors. Let $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ (mean $a_\alpha / b_\alpha$ and $a_\alpha, b_\alpha > 0$). For $(p_{\text{in}}, p_{\text{out}})$, we first specify $p_{\text{in}} \sim \text{Beta}(a_{\text{in}}, b_{\text{in}})$, $p_{\text{out}} \sim \text{Beta}(a_{\text{out}}, b_{\text{out}})$ $(a_{\text{in}}, b_{\text{in}}, a_{\text{out}}, b_{\text{out}} > 0)$, and restrict the joint distribution of $(p_{\text{in}}, p_{\text{out}})$ to $0 < p_{\text{out}} < p_{\text{in}} < 1$. The posterior distribution for $p_{\text{in}}$, $p_{\text{out}}$, $\alpha$, and $\boldsymbol{\pi} \equiv (c_1, c_2, \ldots, c_n)$ is therefore proportional to

$$\begin{aligned} f(\boldsymbol{\pi}, p_{\text{in}}, p_{\text{out}}, \alpha \mid X) \propto & \left( \prod_{(i,j) \in \mathcal{M}} f(X_{ij} \mid p_{\text{in}}, p_{\text{out}}, c_i, c_j) \right) \\ & \cdot f(\boldsymbol{\pi} \mid \alpha) \cdot f(p_{\text{in}}, p_{\text{out}} \mid a_{\text{in}}, b_{\text{in}}, a_{\text{out}}, b_{\text{out}}) \\ & \cdot f(\alpha \mid a_\alpha, b_\alpha), \end{aligned} \tag{4}$$

where $f$ generally denotes a density function.

### 2.4 Point Estimator of Partition Based on Posterior Risk

MCMC draws provide posterior samples of the partition parameter $\boldsymbol{\pi}$. However, it is difficult to find a point estimate of the partition (community structure) in terms of a posterior mean or median. In the Bayesian framework, a point estimator can generally be defined by minimizing a posterior expected loss, so that obtaining a point estimate of the partition then becomes a question of how to suitably define a loss function. We consider a loss function from Lau and Green (2007), which is based on pairwise coincidences (whether pairs of nodes are in a community or not) and defined as follows. For a partition $\boldsymbol{\pi} = (c_1, \ldots, c_n)$ and an estimator $\hat{\boldsymbol{\pi}} = (\hat{c_1}, \ldots, \hat{c_n})$, let

$$L(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) = \textstyle\sum_{(i,j) \in \mathcal{M}} \left( \ell_1 \mathbb{1}_{[c_i = c_j]} \cdot \mathbb{1}_{[\hat{c_i} \neq \hat{c_j}]} + \ell_2 \mathbb{1}_{[c_i \neq c_j]} \cdot \mathbb{1}_{[\hat{c_i} = \hat{c_j}]} \right)$$

denote the cost for declaring that the true partition is $\hat{\boldsymbol{\pi}}$ when it is really $\boldsymbol{\pi}$, where $\ell_1, \ell_2 > 0$ are two tuning parameters and $\mathbb{1}_{[\cdot]}$ denotes the indicator function. The loss is increased by $\ell_1$ (or $\ell_2$) whenever two nodes belonging to the same community in $\boldsymbol{\pi}$ (or $\hat{\boldsymbol{\pi}}$) are in different communities within $\hat{\boldsymbol{\pi}}$ (or $\boldsymbol{\pi}$). Then the

posterior expected loss is

$$\mathrm{E}(L(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) \mid X) = \sum_{(i,j)\in\mathcal{M}} \big(\ell_1 \Pr\{c_i = c_j \mid X\} \cdot \mathbb{1}_{[\hat{c}_i \neq \hat{c}_j]} + \ell_2 \Pr\{c_i \neq c_j \mid X\} \cdot \mathbb{1}_{[\hat{c}_i = \hat{c}_j]}\big), \quad (5)$$

and, as indicated in Lau and Green (2007), minimizing (5) is equivalent to maximizing

$$\sum_{(i,j)\in\mathcal{M}} \mathbb{1}_{[\hat{c}_i = \hat{c}_j]} \left(\Pr\{c_i = c_j \mid X\} - \frac{\ell_2}{\ell_1 + \ell_2}\right). \quad (6)$$

In our implementation, we let $\ell_1 = \ell_2$. After posterior samples have been drawn, we use empirical frequencies to approximate $\Pr\{c_i = c_j \mid X\}$ for all node pairs $(i, j) \in \mathcal{M}$. We then find the posterior sample iteration that maximizes (6), which serves as an estimate of the partition. As a result, an estimated partition $\hat{\boldsymbol{\pi}}$ is an approximate maximizer of (6).

### 2.5 Fitting the Football Data

Here, we focus on community detection for the football data, and we fit the initial model with relatively noninformative priors ($a_{\text{in}} = 2, b_{\text{in}} = 1, a_{\text{out}} = 1, b_{\text{out}} = 2, a_\alpha = 1$, and $b_\alpha = 1$). Using the MCMC algorithm presented in the supplementary material, we obtained an estimated community structure $\hat{\boldsymbol{\pi}}$ for the football data. The results are presented in Figure 3, with additional summaries in Table 3 of the supplementary materials.

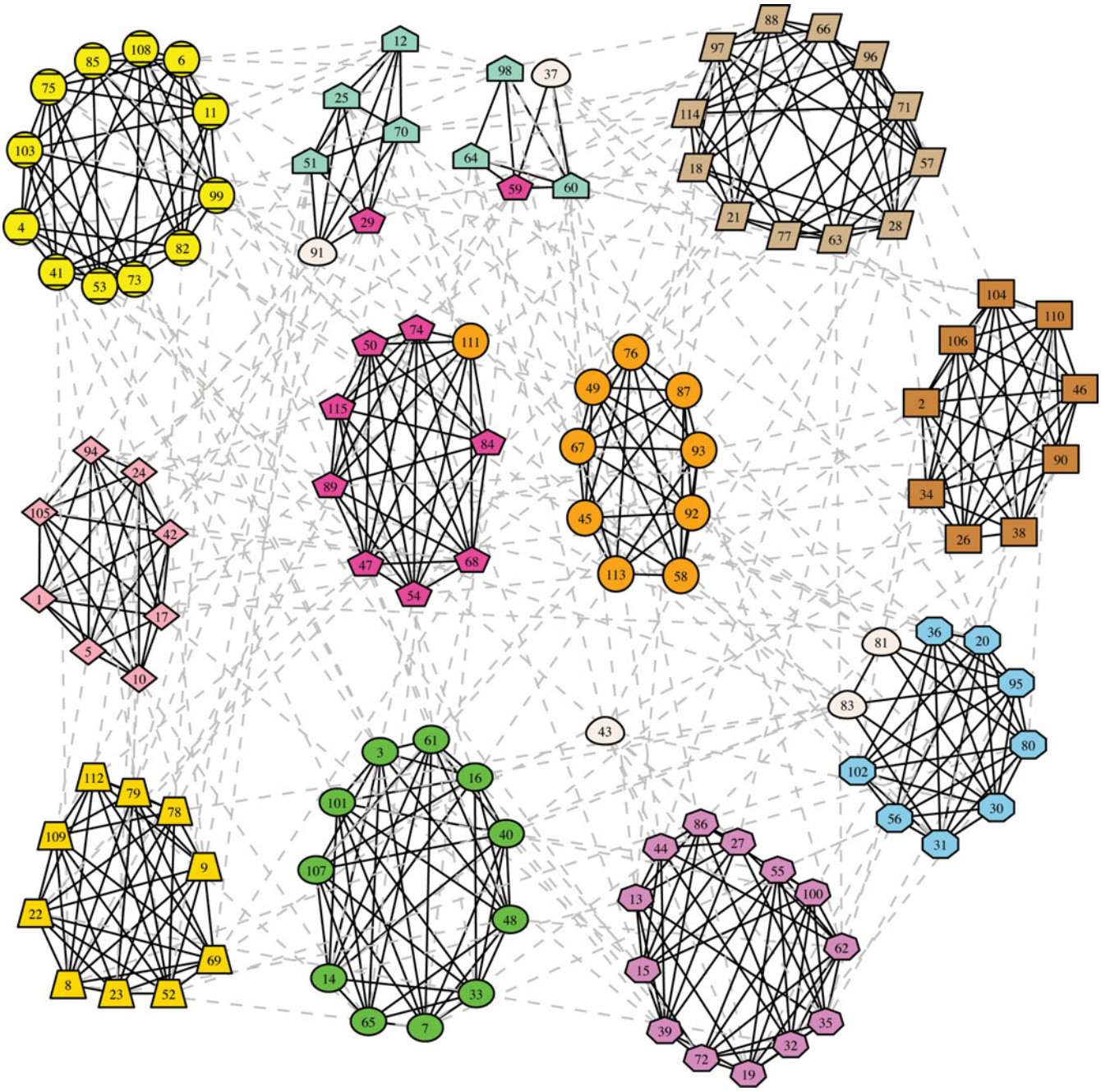


Figure 3. An estimate of communities in the football network using the initial model. The original conferences are indicated as in Figure 2. The online version of this figure is in color.

In comparing the results from our initial model with those in Hofman and Wiggins (2008), there is only one difference. In Hofman and Wiggins (2008), node 43 is estimated to be in the Mid-American conference, that is, in the same conference with nodes 13, 15, etc. In our estimate, node 43 does not belong to any community at all, forming a single-node community. When compared with the truth, both our initial model and that of Hofman and Wiggins (2008) have problems with identifying the "independent teams" that do not belong to any conference in NCAA football. If we consider these independent teams as one community, the initial model would explain the observed data poorly, as only one game was played among independent teams. As a result, it does not seem reasonable to model these independent teams as a true community (at least not as in (2)), as the pattern of edges among the independent teams is not different from the pattern of edges between independent teams and teams in other conferences. The phenomenon of independent teams in the football data (nodes not falling into communities) is not an uncommon behavior in networks, and some research has suggested that single-node communities can be quite reasonable (see Donetti and Muñoz 2004).

As another way to view the inadequacy of our initial model (and that of Hofman and Wiggins 2008), we consider computing the posterior density for the football data at our point estimate of communities and at the true partition (the true conferences). The ratio of the posterior evaluated at the estimated partition to the true partition is about exp(130.903), implying that the truth is not a plausible possibility at all under the initial model. This again suggests that the initial model has problems in describing the football data.

In particular, we notice that teams in small conference play almost every other team in their conference plus a few out-of-conference games. However, for teams in big conferences, such as the Big Twelve and the Mid American, the teams are not able to play every other team in their conferences, since the total number of games played in a season is fixed. Consequently, it is reasonable to consider more flexible models, allowing the probabilities of games (edges) between two teams in the same conference to change across different conferences/communities.

In Section 3, we propose a model to address the issue of the independent teams. In Section 4, we then extend the data-generating model to allow more flexible "edge presence probabilities" across different conferences (communities).

## 3. AN ERDÖS–RÉNYI GROUP OF NODES

Because the initial model fails to identify the independent teams (teams that do not belong to any conference) in the football data, we introduce an ER group to model these teams (see Section 2.3 for an ER graph definition).

### 3.1 Extending the Partition Model

We extend the initial model to incorporate the possibility that some nodes do not belong to the community structure. These "unstructured" nodes are modeled as from an ER graph, so that the probability of an edge between a node in this "ER group" and any other node is the same across the whole network. The remaining "structured" nodes are assumed to be partitioned into a community structure.

We replace the initial data model (2) with the following sampling model,

$$X_{ij} \mid p_{\text{in}}, p_{\text{out}}, p_{\text{ER}}, c_1, \ldots, c_n \overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \quad 1 \le i < j \le n,$$

where

$$p_{ij} = \begin{cases} p_{\text{in}} & \text{if } c_i = c_j \neq c_{\text{ER}}, \\ p_{\text{out}} & \text{if } c_i \neq c_j, c_i \neq c_{\text{ER}}, c_j \neq c_{\text{ER}}, \\ p_{\text{ER}} & \text{if } c_i = c_{\text{ER}} \text{ and/or } c_j = c_{\text{ER}}. \end{cases}$$

In the above specification, label notation for partitions is used, in which $c_{\text{ER}}$ denotes the label for the ER group. We let $p_{\text{in}}, p_{\text{out}}, p_{\text{ER}} \in (0, 1)$, where $p_{\text{ER}}$ denotes the edge presence probability between a node in the ER group with any other node in the graph.

For Bayesian inference, we extend the prior model (3) for the partition $\boldsymbol{\pi} \equiv (c_1, \ldots, c_n)$, in the presence of an ER group. Using a restaurant analogy, we first reserve a table as the ER table, and we refer to other tables as "CRP tables." The first customer sits at the ER table with probability $q$ or the first of the CRP tables with probability $1 - q$. Then, given that $n$ customers have been seated among a total of $K$ CRP tables and the ER table, the next customer would choose a table according to

$$\begin{cases} \Pr(\text{ER table} \mid \text{previous customers}) = q, \\ \Pr(\text{table } k \mid \text{previous customers}) = (1-q)\dfrac{n_k}{n^* + \alpha}, \\ \quad k = 1, 2, \ldots, K, \\ \Pr(\text{new table} \mid \text{previous customers}) = (1-q)\dfrac{\alpha}{n^* + \alpha}, \end{cases} \tag{7}$$

where $n_k$ $(k = 1, \ldots, K)$ is the number of customers at CRP table $k$, and $n^* = \sum_{k=1}^{K} n_k$ is the number of customers sitting at CRP tables prior to the next customer. Let $\boldsymbol{\pi} \sim \text{CRP-ER}(\alpha, q)$ denote the partition distribution defined by the above probability rules. We illustrate the partition model with an ER group in Figure 4.

The probability for a specific partition $\boldsymbol{\pi}$ defined by CRP-ER$(\alpha, q)$ is given by

$$\begin{aligned} \Pr(\boldsymbol{\pi} &= (c_1, c_2, \ldots, c_n)) \\ &= q^{n_{\text{ER}}}(1-q)^{n-n_{\text{ER}}} \frac{\alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)!}{[1+\alpha]_{n^*-1}}, \end{aligned} \tag{8}$$

where $n_{\text{ER}}$ is the number of nodes in the ER group, $n_k$ is the number of nodes in table $k$ of the CRP group, $n^* = \sum_{k=1}^{K} n_k = n - n_{\text{ER}}$. Because the probability in (8) is a function of only community sizes, the order of seating customers is exchangeable (Pitman 1996), as in the case of the CRP. This property plays an important role in developing the MCMC algorithms for models based on this partition model.

### 3.2 Fitting the Football Data with an Erdös–Rényi Group

The MCMC algorithm proposed for the initial model can be easily adapted for the model with an ER group. For priors, we have the same specification for $p_{\text{in}}$, $p_{\text{out}}$, and $\alpha$ as in the initial

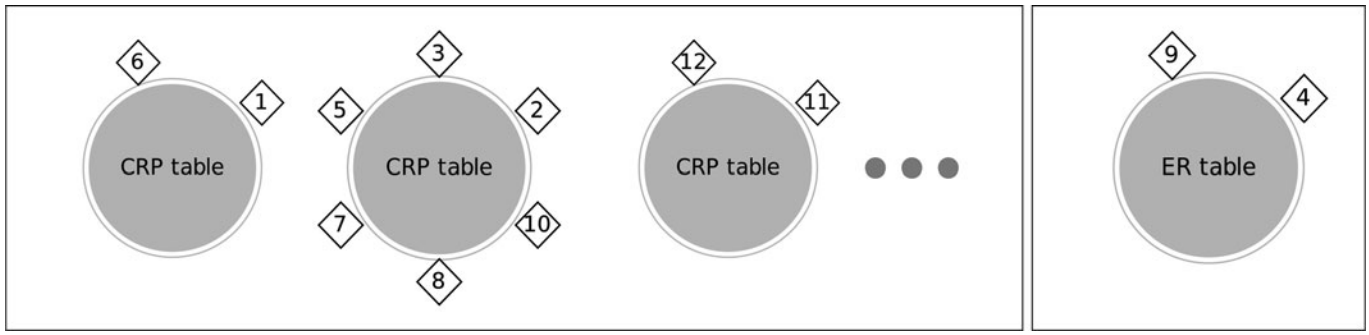



Figure 4. An example partition drawn from CRP-ER($\alpha = 1, q = 0.15$) for 12 nodes.



model. For $q$ and $p_{\text{ER}}$, we use Beta distribution and, in particular, Beta(1, 1) when fitting the football data.

When we fit the model with an ER group to the football data, our estimated partition did not have any nodes in the ER group, although there were posterior draws that did. Consequently, we can evaluate the posterior probability that a particular node is in the ER group, that is, $\Pr\{c_i = c_{\text{ER}} \mid X\}$, and Table 1 presents these probabilities for independent teams (i.e., $i = 37, 43, 81, 83, 91$) as well as for the nonindependent teams with nonzero $\Pr\{c_i = c_{\text{ER}} \mid X\}$. Note that the posterior was multimodal, and MCMC chains with different starting values converged to different modes of the posterior distribution. After running many chains, we found that all but one of the local modes had negligible probability.

Note that if we consider nodes 91, 12, 25, and 70 in Figure 3, which were identified as a community under the initial model, there is a game between each pair of these teams. This explains why node 91 is not identified as an independent team (belonging to the ER group) by the posterior probability in Table 1. This points to an issue with the "truth" itself (the conference labels in these data compiled by Girvan and Newman 2002). While the games were played in 2000, the conference structure is apparently from 2002. Team 91 (Utah State) was independent in 2002, but was in the Big West Conference in 2000. The Big West Conference stopped sponsoring football after the 2000 season, and we observe that Utah State played teams that eventually formed the Sun Belt Conference in 2001. This supports the estimated community for Team 91 in Figure 3.

## 4. EXTENSIONS: COMMUNITY-VARYING EDGE PROBABILITIES

After introducing an unstructured or ER group, the resulting model still exhibits some problems as a stochastic representation of the football data. Since we know the football conference structure, we use this information to gain insight for developing additional model extensions. We first examine the proportion of games played within individual conferences. Consider the Big Ten Conference (Figure 2) composed of 11 teams with 44 edges (games played) within as an example. A simple method to estimate a $p_{\text{in}}$ for this conference is the ratio of number of games played to the total number of pairs of teams, that is, $\widehat{p_{\text{in}}} = 44/55 = 0.8$. Similarly, for other conferences such as Mountain West Conference, Big Twelve Conference, and Mid-American Conference, we have $\widehat{p_{\text{in}}} = 1.0, 0.727$, and $0.641$, respectively. Clearly, these estimates of $p_{\text{in}}$ differ by conference, which suggests that a model representation with a single $p_{\text{in}}$ for the whole network is not flexible enough (even after including an ER group). As a result, we consider model extensions to allow community-varying $p_{\text{in}}$ parameters.

In the following subsections, we consider models where edge probabilities can vary by community (or pairs of communities), and the probabilities for inter- and intracommunity edges are permitted to differ.

Table 1. The posterior probability that a team is in the ER group. All independent teams, as well as nonindependent teams with nonzero probabilities, are listed

| | Independent teams | | | | | Other teams | |
|---|---|---|---|---|---|---|---|
| Node | 37 | 43 | 81 | 83 | 91 | 64 | 98 |
| $\Pr\{c_i = c_{\text{ER}} \mid X\}$ | 0.113 | 0.254 | 0.381 | 0.381 | 0 | 0.0011 | 0.00082 |

### 4.1 Varying Intracommunity and Single Intercommunity Edge Probabilities

We first extend the initial model (2) to allow different $p_{\text{in}}$'s (within-community edge probabilities) for each community. We do not initially consider an ER group. Given a partition $\boldsymbol{\pi}$ for which we have $K$ communities, we let $p_{\text{in},1}, \ldots, p_{\text{in},K}$ be the $p_{\text{in}}$'s for communities $1, \ldots, K$, respectively. As part of a prior specification (or as viewed as a hierarchical model), we suppose that $p_{\text{in},1}, \ldots, p_{\text{in},K}$ are obtained as IID draws from a continuous distribution $G$ on $(0, 1)$, implying zero probability for $p_{\text{in},k} = p_{\text{in},k'}$ $(k \neq k')$. In particular, here we take $G$ to be a Beta distribution, so that

$$p_{\text{in},1}, \ldots, p_{\text{in},K} \mid \boldsymbol{\pi} \overset{\text{IID}}{\sim} G, \quad G \equiv \text{Beta}(tm, t(1-m)), \quad t > 0,\ 0 < m < 1,$$

where $t$ and $m$ denote hyperparameters. Given a partition $\boldsymbol{\pi}$ dividing the nodes set $V = \{1, \ldots, n\}$ into communities $V_1, \ldots, V_K$, we again let $c_i = k$ if node $i$ is in community $k$ and note that community $k$ $(k = 1, \ldots, K)$ has a distinct intracommunity edge probability, $p_{\text{in},k}$. The probability of edge presence for two nodes in different communities is $p_{\text{out}}$ for the whole network (single intercommunity edge probability). In summary,

the proposed model is

$$\begin{aligned} X_{ij} \mid \boldsymbol{\pi}, p_{\text{in},1}, \ldots, p_{\text{in},K}, p_{\text{out}} &\overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \\ p_{\text{in},1}, \ldots, p_{\text{in},K} \mid \boldsymbol{\pi} &\overset{\text{IID}}{\sim} G, \\ \boldsymbol{\pi} \equiv (c_1, \ldots, c_n) &\sim \text{CRP}(\alpha), \end{aligned} \tag{9}$$

where $p_{ij} = p_{\text{in},k}$ if $c_i = c_j = k$; or $p_{ij} = p_{\text{out}}$ if $c_i \neq c_j$.

Equivalently, the above model (9) can be represented using a DP. We let node $i$ ($i = 1, 2, \ldots, n$) be associated with a $p_{\text{in}}$ parameter, denoted by $p_{\text{in}}^i$ (where $p_{\text{in}}^i = p_{\text{in},k}$ if node $i$ is in community $k$). Then model (9) can be written as

$$\begin{aligned} X_{ij} \mid p_{\text{in}}^1, \ldots, p_{\text{in}}^n, p_{\text{out}} &\overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \\ p_{\text{in}}^1, \ldots, p_{\text{in}}^n &\overset{\text{IID}}{\sim} P, \\ P &\sim \text{DP}(\alpha, G), \end{aligned} \tag{10}$$

where $p_{ij}$ is $p_{\text{in}}^i$ if $p_{\text{in}}^i = p_{\text{in}}^j$ or $p_{\text{out}}$ otherwise, and $G$ is $\text{Beta}(tm, t(1-m))$. Hence, nodes with the same $p_{\text{in}}^i$ values are treated as belonging to the same community, so $p_{\text{in}}^1, \ldots, p_{\text{in}}^n$ define a particular community structure. The model represented by the DP allows us to employ characteristics of the DP for developing MCMC algorithms for this model; see Guo (2011) for details.

We can also introduce an ER group in (9). For the CRP tables, we have different $p_{\text{in}}$'s for each community. The probability that a node in the ER group shares an edge with any other node (including those in the ER group) is $p_{\text{ER}}$. Using labels $(c_1, \ldots, c_n)$ to denote the partition, we again let $c_i = k$ if node $i$ belongs to community $k = 1, \ldots, K$, and let $c_i = c_{\text{ER}} \notin \{1, 2, \ldots, K\}$ if node $i$ belongs to the ER group. The model that allows multiple $p_{\text{in}}$'s and an ER group is summarized as

$$\begin{aligned} X_{ij} \mid \boldsymbol{\pi}, p_{\text{in},1}, \ldots, p_{\text{in},K}, p_{\text{out}}, p_{\text{ER}} &\overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \\ p_{\text{in},1}, \ldots, p_{\text{in},K} \mid \boldsymbol{\pi} &\overset{\text{IID}}{\sim} \text{Beta}(tm, t(1-m)), \\ \boldsymbol{\pi} &\sim \text{CRP-ER}(\alpha, q), \end{aligned}$$

where $p_{ij} = p_{\text{in},k}$ if $c_i = c_j = k$; $p_{ij} = p_{\text{out}}$ if $c_i \neq c_j \neq c_{\text{ER}}$; or $p_{ij} = p_{\text{ER}}$ if $c_i = c_{\text{ER}}$ and/or $c_j = c_{\text{ER}}$.

For fitting this model to the football data, we specify $p_{\text{ER}} \sim \text{Beta}(1, 1)$; $q \sim \text{Beta}(1, 1)$; $t \sim \text{Gamma}(1, 0.1)$ with mean 10 and variance 100; $m \sim \text{Beta}(1, 1)$; $p_{\text{out}} \sim \text{Beta}(1, 2)$; and $\alpha \sim \text{Gamma}(1, 1)$. Using this more flexible model to estimate the community structure (with results shown in Figure 1 and Table 4 in the supplementary material), nodes 81 and 83 are successfully identified as in the ER group (i.e., as independent teams), and we have the following posterior probabilities: $\Pr\{c_{37} = c_{\text{ER}} \mid X\} = 0.45$, $\Pr\{c_{81} = c_{\text{ER}} \mid X\} = 0.92$, and $\Pr\{c_{83} = c_{\text{ER}} \mid X\} = 0.95$. Hence, this model provides stronger evidence that nodes 37, 81, and 83 are independent teams than the previous model in Section 3.2 (one $p_{\text{in}}$ and $p_{\text{out}}$ parameter for the network, and $\boldsymbol{\pi} \sim \text{CRP-ER}(\alpha, q)$).

### 4.2 Varying Intra- and Intercommunity Edge Probabilities

We next consider a further natural model extension, allowing a different $p_{\text{out}}$ parameter for any pair of communities. Given a network that has $K$ communities among the "structured" nodes, we let $p_{\text{in},k}$ be the intracommunity edge probability $p_{\text{in}}$ for community $k$ ($k = 1, 2, \ldots, K$), and $p_{\text{out},(k,l)}$ be the intercommunity edge probability $p_{\text{out}}$ between communities $k$ and $l$ ($1 \leq k < l \leq K$). Further, we let

$$p_{\text{in},k} \mid \boldsymbol{\pi} \overset{\text{IID}}{\sim} \text{Beta}(tm, t(1-m)), \quad t > 0,\ 0 < m < 1, \tag{11}$$

$$p_{\text{out},(k,l)} \mid \boldsymbol{\pi} \overset{\text{IID}}{\sim} \text{Beta}(hw, h(1-w)), \quad h > 0,\ 0 < w < 1, \tag{12}$$

where we model $\boldsymbol{\pi}$ by CRP-ER$(\alpha, q)$ from Section 3.1, and $t$, $m$, $h$, $w$ are hyperparameters. As mentioned in Section 1, a hierarchical model similar to the above appears in Kemp et al. (2006) but without the distinction of (11)–(12). Continuing the progression, the model for edge presence is

$$\begin{aligned} &X_{ij} \mid \boldsymbol{\pi}, p_{\text{in},1}, \ldots, p_{\text{in},K}, p_{\text{out},(1,2)}, \ldots, p_{\text{out},(K-1,K)}, \\ &p_{\text{ER}} \overset{\text{IND}}{\sim} \text{Bernoulli}(p_{ij}), \end{aligned}$$

where

$$p_{ij} = \begin{cases} p_{\text{in},k} & \text{if } c_i = c_j = k, \\ p_{\text{out},(k,l)} & \text{if } c_i = k \neq c_{\text{ER}},\ c_j = l \neq c_{\text{ER}},\ c_i \neq c_j, \\ p_{\text{ER}} & \text{if } c_i = c_{\text{ER}} \text{ and/or } c_j = c_{\text{ER}}. \end{cases}$$

For this model specification, it becomes challenging to perform Bayesian inference directly because the number of parameters are changing for different $\boldsymbol{\pi}$. Note that the previous model (10) involved drawing a parameter $p_{\text{in}}^i$ for each node $i = 1, \ldots, n$ in the DP representation, where nodes with the same $p_{\text{in}}$ values belonged to the same community; in this way, the $n$ such $p_{\text{in}}$ parameters automatically determined community assignments and, within the posterior MCMC-sampler, the $p_{\text{in}}$ values could be updated for each node in a conditionally stepwise manner. That is, the last model had a tractable and consistent form for "book-keeping" the parameters and community structure. In contrast, for the current model with multiple $p_{\text{out}}$ parameters, drawing MCMC-based posterior samples becomes problematic as the type of $p_{\text{out}}$ parameters inherently vary with a changing community structure and consequently it becomes intractable and even nonsensical to similarly stepwise update and summarize such parameters. To overcome this estimation difficulty, we integrate out the $p_{\text{in}}$'s and $p_{\text{out}}$'s. If we use node sets $(V_1, \ldots, V_K, V_{\text{ER}})$ and node labels $(c_1, c_2, \ldots, c_n)$ to denote the partition $\boldsymbol{\pi}$, in which $V_{\text{ER}}$ is the node set for the ER group, then the posterior distribution can be written as follows:

$$\begin{aligned} &f\big(\boldsymbol{\pi}, t, m, h, w, p_{\text{ER}}, q, p_{\text{in},1}, \ldots, p_{\text{in},K}, \\ &\quad p_{\text{out},(1,2)}, \ldots, p_{\text{out},(K-1,K)} \mid X\big) \\ &\propto \prod_{k=1}^{K} \Bigg( \prod_{i \in V_k,\ j \in V_k,\ i<j} f(X_{ij} \mid p_{\text{in},k}) \Bigg) \\ &\quad \cdot \prod_{k=1}^{K} f(p_{\text{in},k} \mid tm, t(1-m)) \\ &\quad \prod_{1 \leq k < l \leq K} \Bigg( \prod_{i \in V_k,\ j \in V_l} f\big(X_{ij} \mid p_{\text{out},(k,l)}\big) \Bigg) \\ &\quad \cdot \prod_{1 \leq k < l \leq K} f\big(p_{\text{out},(k,l)} \mid hw, h(1-w)\big) \end{aligned} \tag{13}$$

$$\prod_{i\in V_{\mathrm{ER}}}\prod_{j\neq V_{\mathrm{ER}}} f(X_{ij}\mid p_{\mathrm{ER}})\cdot \prod_{i\in V_{\mathrm{ER}},\ j\in V_{\mathrm{ER}},\ i<j} f(X_{ij}\mid p_{\mathrm{ER}})$$
$$f(c_1,c_2,\ldots,c_n\mid\alpha,q)f(p_{\mathrm{ER}}\mid\cdot)f(\alpha\mid\cdot)f(q\mid\cdot)$$
$$\times\ f(t\mid\cdot)f(m\mid\cdot)f(h\mid\cdot)f(w\mid\cdot).$$

In (13), $f(X_{ij}\mid p_{\mathrm{in},k})$ is a Bernoulli$(p_{\mathrm{in},k})$ probability mass function conditional on $p_{\mathrm{in},k}$ having Beta$(tm, t(1-m))$ distribution. So by conjugacy, we can integrate out $p_{\mathrm{in},k}$ for $k = 1,\ldots,K$ as

$$\begin{aligned}&\int p_{\mathrm{in},k}^{I_k}(1-p_{\mathrm{in},k})^{N_k-I_k} f(p_{\mathrm{in},k}\mid tm, t(1-m))\,d\,p_{\mathrm{in},k}\\ &= \int p_{\mathrm{in},k}^{I_k}(1-p_{\mathrm{in},k})^{N_k-I_k}\frac{1}{B(tm,t(1-m))}p_{\mathrm{in},k}^{tm-1}\\ &\quad\times(1-p_{\mathrm{in},k})^{t(1-m)-1}\,d\,p_{\mathrm{in},k}\\ &= \frac{B(tm+I_k, t(1-m)+N_k-I_k)}{B(tm,t(1-m))},\end{aligned}$$

where $I_k$ is the total number of edges in community $k$, $N_k$ is the total number of pairs of nodes in community $k$ (if community $k$ has $n_k$ nodes, then $N_k = n_k(n_k-1)/2$), and $B(\cdot,\cdot)$ denotes the Beta function. Similarly, we can integrate out the $p_{\mathrm{out},(k,l)}$'s in (13). This reduces the parameters in the posterior distribution to $\boldsymbol{\pi}$, $\alpha$, $q$, $p_{\mathrm{ER}}$, $t$, $m$, $h$, $w$. The approach of updating $\boldsymbol{\pi}$ in the algorithm for the initial model then can be applied for this model. For all parameters except $\alpha$, we either use a Gibbs sampling or a Metropolis–Hastings algorithm. Specifying priors for $h$ and $w$ is similar to specifying $t$ and $m$ in the previous models; see Guo (2011) for more details.

For the football data, we specify noninformative prior as $h \sim$ Gamma$(1, 0.1)$ and $w \sim$ Beta$(1, 1)$ with the prior of other parameters as in Section 3.2. We present counts of intra- and intercommunity edges (Table 5) and summaries for the parameter estimation (Table 6) in the supplementary material. The estimate of the community structure is presented in Figure 5: we are able to identify two of the independent teams (nodes 81 and 83) with this model. Nodes 37 and 43 (also independent teams) are estimated to be in single-node communities. It appears that the probabilities of nodes 37 and 43 being connected with other nodes are a bit different from $p_{\mathrm{ER}}$ (associated with nodes 81 and 83) and are perhaps more flexibly described as arising from the parent distribution of $p_{\mathrm{out},(k,l)}$. As a result, theses nodes are not estimated to be in the ER group, though nodes in single-node communities are qualitatively similar to nodes in the ER group, in that they do not belong to any community.

Comparing Figure 5 with the true conferences in Figure 2, we find, perhaps surprisingly, that estimation from this model splits the Mid-American, Southeastern, and Big Twelve Conferences. A close inspection of these three conferences (Figure 5) indicates that these three conferences can be literally split into two "smaller" communities, where there is a game between each pair of teams inside these smaller communities. In this sense, the estimation results in Figure 5, while initially unexpected, are quite reasonable and intuitive.

### 4.3 Relating Intracommunity Edge Probability to Community Size

We now consider a much simpler data generation model to more directly model the possibility that some communities share the same intracommunity edge probabilities ($p_{\mathrm{in}}$'s). In the football data, every team plays nearly the same number of games in a season. As a result, when a conference has a relatively large number of teams, the chance that any two teams play a conference game becomes smaller. To express this idea, we propose the following simplification of (11) for the $p_{\mathrm{in}}$'s in the structured CRP group, that is, $p_{\mathrm{in},k} = p_{\mathrm{in}_a}\mathbb{1}_{[n_k\le 9]} + p_{\mathrm{in}_b}\mathbb{1}_{[n_k>9]}$, where $n_k$ is the community size. We chose nine teams due to the structure of the season; teams often play eight conference games. We let $p_{\mathrm{in}_a}$ and $p_{\mathrm{in}_b}$ have the same prior distribution, Beta$(a_{in}, b_{in})$, where $a_{\mathrm{in}} > 0$, $b_{\mathrm{in}} > 0$ are the prior parameters. For the model estimation, we integrate out $p_{\mathrm{in}_a}$, $p_{\mathrm{in}_b}$, and $p_{\mathrm{out},(k,l)}$'s, analogously to (13).

For fitting the football data, we use the same prior specification for parameters that appeared in Section 4.2 (the model with varying intra- and intercommunity edge probabilities), and we let the prior for $p_{\mathrm{in}_a}$ and $p_{\mathrm{in}_b}$ be Beta$(1, 1)$. The community structure estimated using Bayesian analysis of this model with two $p_{\mathrm{in}}$'s and multiple $p_{\mathrm{out}}$'s is almost the same as that from the most flexible model with multiple $p_{\mathrm{in}}$'s, multiple $p_{\mathrm{out}}$'s, and an ER group (see Figure 5 and Section 4.2). This suggests that, for the football data, a simplified but slightly more structured model (based on community size) would also provide an adequate description of the data.

## 5. MODEL SELECTION AND ASSESSMENT

In Table 2, we list all models considered for the football data in Sections 2, 3, and 4. We next explore model diagnostics for the football data.

### 5.1 Model Selection

Model selection and assessment are important considerations, which we next investigate for the series of models proposed for the football data. We first use the deviance information criterion (DIC) to compare these models. Like many model selection criteria, the DIC (Spiegelhalter et al. 2002) is the sum of a

Table 2. Summary of statistical models for community detection

| Model | Model description |
|---|---|
| PIN-POUT | One $p_{\mathrm{in}}$ and one $p_{\mathrm{out}}$ |
| PIN-POUT-ER | One $p_{\mathrm{in}}$ and one $p_{\mathrm{out}}$ with ER group |
| MPIN-POUT-ER | Varying intracommunity edge probability and single intercommunity edge probability (multiple $p_{\mathrm{in}}$'s and one $p_{\mathrm{out}}$) with ER group |
| MPIN-MPOUT-ER | Varying intra- and intercommunity edge probability (multiple $p_{\mathrm{in}}$'s and multiple $p_{\mathrm{out}}$'s) with ER group |
| TPIN-MPOUT-ER | Relating intracommunity edge probability to community size (two $p_{\mathrm{in}}$'s and multiple $p_{\mathrm{out}}$'s) with ER group |

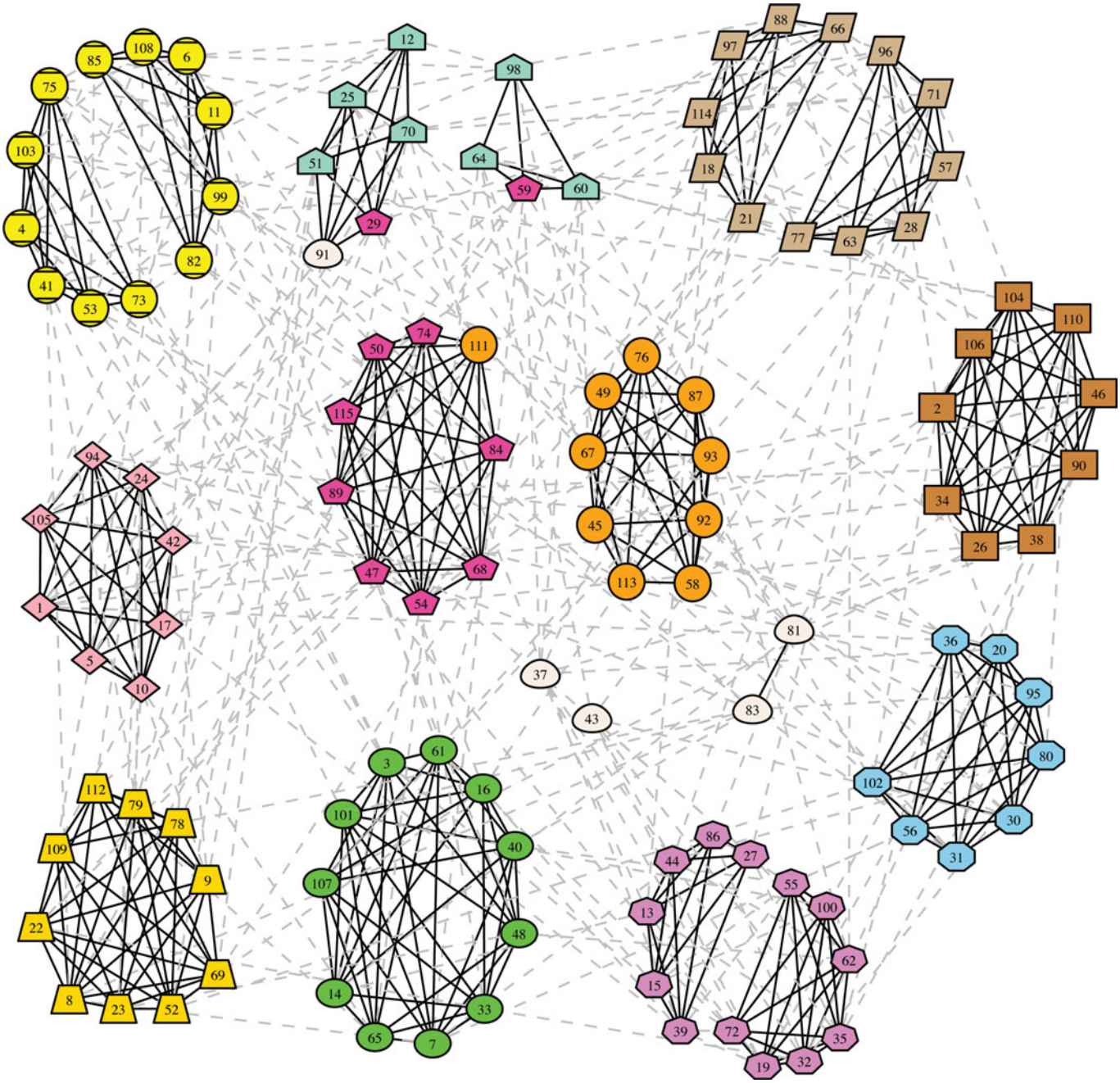



Figure 5. An estimate of communities in the football network using the model with varying intra- and intercommunity edge probabilities, and an ER group. Nodes 81 and 83 are in the ER group from the estimate, while nodes 37 and 43 are in single-node communities. The original conferences are indicated as in Figure 2. The online version of this figure is in color.

model goodness-of-fit term and a penalty for model complexity, so that smaller DIC values are interpreted as indicating better models.

Define *deviance* as $D(\theta) = -2 \log p(y \mid \theta)$, where $p(y \mid \theta)$ is the likelihood. Define the *posterior mean deviance* as $\overline{D} = E_{\theta \mid y}(D(\theta) \mid y)$. Let $D(\tilde{\theta})$ be the deviance evaluated at some estimator $\tilde{\theta}$ for $\theta$ (typically the posterior mean, though the posterior median is also justifiable, see Celeux et al. 2006). The expression for the DIC is $\text{DIC} = D(\tilde{\theta}) + 2p_D = \overline{D} + p_D$, where $p_D = \overline{D} - D(\tilde{\theta})$ is interpreted as an approximate dimension for the model parameters. For the football data, the likelihood function $p(y \mid \theta)$ is formed from the probabilities for edge presence/absence as before, where edge probabilities are integrated out in the case of multiple $p_{\text{in}}$'s and/or $p_{\text{out}}$'s. The DIC is then evaluated from MCMC-based posterior samples.

In our models, it is problematic to evaluate the typical DIC, as we do not have the posterior mean for the partition. As in Celeux et al. (2006), we first use the maximum a posteriori probability (MAP) estimator instead of the mean. We refer to this DIC using MAP estimator as "$\text{DIC}_1$." We also use the estimator of the partition that minimizes the posterior expected loss combined with the posterior mean of other quantitative parameters to serve as $\tilde{\theta}$. We refer to this DIC as "$\text{DIC}_2$." In Table 3, we list the DICs for the models we fit for the football data. From

Table 3. DICs and dimensionalities ($p_D$) of different models for fitting the football data

| Model | $DIC_1$ ($p_D$) | $DIC_2$ ($p_D$) |
|---|---|---|
| PIN-POUT | 2222.51 (4.03) | 2222.34 (3.86) |
| PIN-POUT-ER | 2206.7 (0.13) | 2111.95 (–94.45) |
| MPIN-POUT-ER | 2199.02 (6.82) | 2192.91 (0.71) |
| MPIN-MPOUT-ER | 2108.81 (14.15) | 2104.61 (9.95) |
| TPIN-MPOUT-ER | 2108.39 (11.66) | 2100.2 (3.47) |

this table of DICs, we can see that the most flexible model (MPIN-MPOUT-ER) is the best among the first four models in the table, and the best model overall is the one that incorporates the information on community sizes (TPIN-MPOUT-ER).

In framing model comparisons in terms of a numerical fit-complexity measure, some caution is needed in interpreting the DIC, as with similar model selection criteria. This criterion itself has been subject to debate (see the discussion following Spiegelhalter et al. 2002 and Celeux et al. 2006). Additionally, some deviance-based model dimensionalities ($p_D$) in Table 3 do not make sense in terms of model complexity (number of model parameters), which potentially indicates that some models do not fit the football data well. For example, when fitting the simplest model PIN-POUT-ER, we have difficulty identifying the independent teams, resulting in poor estimation of the parameter $p_{ER}$ and a negative $p_D$ value. Of course, in the process of developing models, a model selection criterion is one guide but not necessarily the most definitive. Other forms of model assessment are also helpful to consider.

### 5.2 Model Checking

In conjunction with the DIC, we consider another approach to assess model fits. Note first that comparing models in terms of how their estimated communities match the true community structure (Figure 2) is problematic, because some community labels in Figure 2 are a bit misleading. Independent teams are not a community in the usual sense, and the conference structure is seemingly from 2002, not 2000 when the games were played. That is, we used the data, including conference labels, from Girvan and Newman (2002) and discovered the issues with these labels (the nominally "true" conferences) when examining anomalies in our analyses. In particular, the Sun Belt Conference did not sponsor football until 2001; Team 29 (Boise State) was in the Big West in 2000 and the Western Athletic Conference in 2002; Team 111 (Texas Christian) was in the Western Athletic Conference in 2000 but in Conference USA in 2002; Team 59 (Louisiana Tech) was independent in 2000 but in the Western Athletic Conference in 2002. Consequently, we focus on examining how consistent the models are with data (presence of edges), and in particular we use the posterior predictive checking approach (Gelman et al. 2004).

In posterior predictive checking, we use every iteration of the posterior samples to replicate a new dataset of edge presence, $X^{rep} = [X_{ij}^{rep}]_{1\le i<j\le n}$. From all iterations of posterior samples, we obtain a distribution of $X^{rep}$, which is not easy to directly summarize given the incidence matrix form of $X^{rep}$. As a result, we choose a test quantity of interest, $T(X)$, to be a scalar summary of the edge data $X = [X_{ij}]_{1\le i<j\le n}$, which can be easily compared with the distribution of $T(X^{rep})$. For the football data, for example, we chose the number of edges (games) between the whole network, the Sun Belt Conference, the independent teams, and the Mid-American Conference, denoted by $T_1(X)$, $T_2(X)$, $T_3(X)$, and $T_4(X)$, respectively.

To simulate $X^{rep}$ for the initial model (2), we simply use the posterior samples of $p_{in}$ and $p_{out}$ and $\boldsymbol{\pi} \equiv (c_1, \ldots, c_n)$ to determine the edge presence between any pair of nodes using the Bernoulli distribution with probability either $p_{in}$ or $p_{out}$ depending on their labels. Similarly, we simulate replicated data for the model with one $p_{in}$, one $p_{out}$, and an ER group. For the model with multiple $p_{in}$'s and one $p_{out}$, since $p_{in}$'s are also drawn in the posterior samples, we apply the same approach as for the initial model. For the model with multiple $p_{in}$'s (and two $p_{in}$'s) and multiple $p_{out}$'s, since we have integrated out $p_{in}$'s and $p_{out}$'s, for every iteration, we need to first draw samples for the $p_{in}$'s and the $p_{out}$'s given the partition $\boldsymbol{\pi}$ and other parameters. Once we have samples of both the edge presence probabilities ($p_{in}$'s and $p_{out}$'s) and the partition, $\boldsymbol{\pi}$, we can easily generate a replication of $X^{rep}$.

In Figure 6, we present the comparison of $T(X)$ with the distribution of $T(X^{rep})$ for the quantities defined above $T_1$, $T_2$, $T_3$, and $T_4$ over four different models. The vertical lines indicate the test quantities observed for the football data: $T_1(X) = 613$, $T_2(X) = 1$, $T_3(X) = 10$, and $T_4(X) = 50$. We can see that all models look plausible for $T_1(X)$ and $T_2(X)$ but not all for $T_3(X)$ and $T_4(X)$. In particular, the predictive distributions for $T_3(X)$ and $T_4(X)$ based on model PIN-POUT-ER are not consistent with the data, which supports the negative model dimensionality value $p_D$ in Table 3, as mentioned in Section 5.1. Models with multiple $p_{in}$'s (and with one $p_{out}$ or multiple $p_{out}$'s) seem to fit the data well in terms of these test quantities. In the simplified model with two $p_{in}$'s (related to community sizes) and multiple $p_{out}$'s, we find results (not presented) comparable to the most flexible model (multiple $p_{in}$'s and $p_{out}$'s varying across communities).

## 6. DISCUSSION

This work was sponsored by the Sandia National Laboratories Networks Discovery, Characterization and Prediction project (Kegelmeyer et al. 2010). The project's broad goal was to develop analysis capabilities to address adversarial networks engaged in "weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms smuggling, and other illicit activities." In addition, these networks can depend on secondary networks for "financial, supply chain, communication, recruiting, and fund-raising activities." These networks are typically complex and dynamic, which adds to the complexity in identifying and studying them. Of particular interest was the identification of structure within such networks (i.e., community partitions), which can frequently convey information about the roles of various nodes, and possibly contribute to the prediction of subsequent evolution. In addition, the project was focused on the computational scalability of the analysis methodologies and their potential for providing decision support for intelligence analysis.

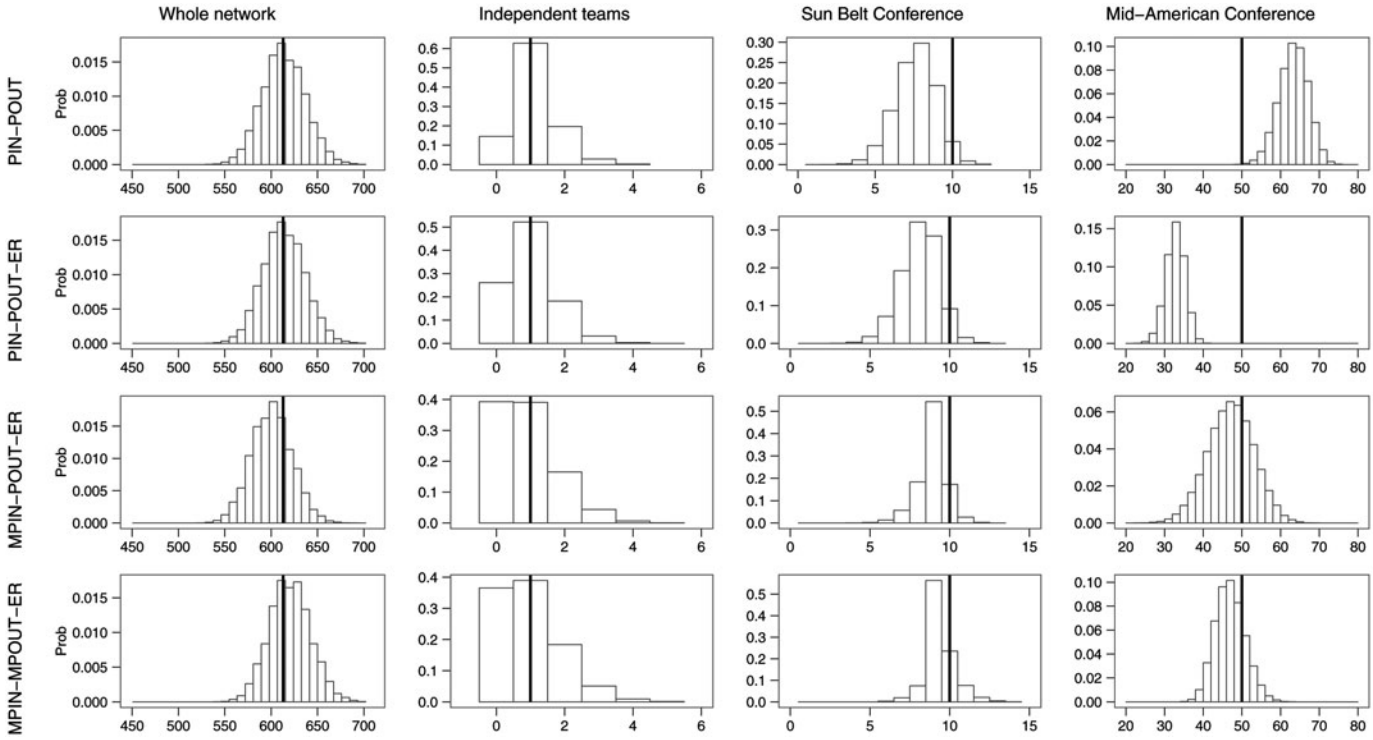



Figure 6. Posterior predictive checking using $T_1(X)$, $T_2(X)$, $T_3(X)$, and $T_4(X)$.



This work focuses on a small piece of the broad goal and develops a series of statistical models for community detection that can support uncertainty quantification. We considered an initial statistical model based on a Bayesian nonparametric partition model for the community structure in a network. Basic partition models in Bayesian nonparametric methods, such as CRP and related DP, allow inference without specifying the number of communities a priori for a variety of data models. The initial data model reflected an underlying notion of differing intra- and intercommunity edge probabilities. Motivated by problems from fitting the football data, we extended the initial model from two perspectives. First, by introducing an ER group within a model, nodes were allowed to fall outside community structure entirely. Second, we incorporated increasingly more general and flexible probabilities for edges between node pairs, which varied with the community structure and differed by the type of edge (within or between communities). While the resulting models are extensions of stochastic block models in the community detection literature (see Kemp et al. 2006; Karrer and Newman 2011), the inferential approach for community detection was a fully Bayesian one (estimation based on sampling the posterior distribution) rather than simply obtaining a point estimator. Our modeling and inference ideas for the football data—such as MCMC algorithms, a decision theoretical framework for obtaining communities, model selection, and assessments—can generalize other problems involving community detection.

Computation is often a concern in model fitting. While our focus is primarily on model development, it is worth noting that we provide workable algorithms for fitting the models. Extensions to the partition models could also be considered using, for example, a Pitman–Yor process (of which the CRP is a special case; see the supplementary materials). However, as the number of nodes increases, more efficient computational approaches will become necessary. Currently, our algorithms for implementing MCMC-based Bayesian inference become impractical for network data with over a few thousand nodes, due to community updating steps per node. One direction is to adapt the method of merging and splitting partitions as in Jain and Neal (2004). Other algorithms may be able to speed up the computation and thus to work with larger network data.

As the partition itself is unwieldy to work with, we are not able to directly study the convergence of partitions as a model parameter in terms of Markov chains. As a result, an approach to compare partitions quantitatively that can be used for the diagnostics in MCMC simulation might be helpful with these types of block models.

Finally, the models proposed in the article are for graphs with undirected edges and without weights or attributes. Future research might involve models to incorporate other covariate information on network-related data and applications. For example, Miller, Griffiths, and Jordan (2009) had introduced a Bayesian nonparametric model where the edge probability between node pairs depends on a weighting of binary-valued latent vectors associated with each node. It would be interesting to place such models in the context of community detection, which may require determining communities (e.g., node partitions) from a nontrivial topological space of the latent variables.

## SUPPLEMENTARY MATERIALS

**Code:** C code for implementing MCMC algorithms for the proposed models fitting the football network data (zip file).

**Supplementary text:** Additional details for fitting the football network data using the proposed models, an introduction to

the Dirichlet Process and Pitman–Yor process, and details of the MCMC algorithm for the initial model (pdf).

## ACKNOWLEDGMENTS

Alyson Wilson's work on this article took place while she was at IDA Science and Technology Policy Institute Washington, DC, and Iowa State University, Ames, IA.

*[Received April 2012. Revised February 2013.]*

## REFERENCES

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [390]

Ball, B., Karrer, B., and Newman, M. E. J. (2011), "Efficient and Principled Method for Detecting Communities in Networks," *Physical Review E*, 84, 036103. [390]

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006), "Deviance Information Criteria for Missing Data Models," *Bayesian Analysis*, 1, 651–674. [399,400]

Donetti, L., and Muñoz, M. A. (2004), "Detecting Network Communities: A New Systematic and Efficient Algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, 10, P10012. [395]

Ferry, J. P., Bumgarner, J. O., and Ahearn, S. T. (2011), "Probabilistic Community Detection in Networks," in *Proceedings of 14th International Conference on Information Fusion, IEEE*, pp. 1741–1748. [391,393]

Fortunato, S. (2010), "Community Detection in Graphs," *Physics Reports*, 486, 75–174. [390,391,393]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC. [400]

Girvan, M., and Newman, M. E. J. (2002), "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 7821–7826. [390,391,396,400]

Guo, J. (2011), "Bayesian Methods for System Reliability and Community Detection," Ph.D. dissertation, Iowa State University, Ames, IA. Available at *http://lib.dr.iastate.edu/etd/12240/* [397,398]

Hofman, J. M., and Wiggins, C. H. (2008), "Bayesian Approach to Network Modularity," *Physical Review Letters*, 100, 258701. [390,391,393,395]

Jain, S., and Neal, R. M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182. [401]

Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [390,401]

Kegelmeyer, P., Cook, B., Rogers, D., et al. (2010), "Network Discovery, Characterization, and Prediction: A Grand Challenge LDRD Final Report," Technical Report SAND2010-8715, Sandia National Laboratories. [400]

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006), "Learning Systems of Concepts With an Infinite Relational Model," in *Proceedings of the 21st National Conference on Artificial intelligence—Volume 1*, *AAAI Press*, pp. 381–388. [390,391,393,397,401]

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008), "Benchmark Graphs for Testing Community Detection Algorithms," *Physical Review E*, 78, 046110. [390]

Lau, J. W., and Green, P. J. (2007), "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [393,394]

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008), "Statistical Properties of Community Structure in Large Social and Information Networks," in *Proceeding of the 17th International Conference on World Wide Web*, New York: ACM, pp. 695–704. [390]

Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2009), "Nonparametric Latent Feature Models for Link Prediction," in *Advances in Neural Information Processing Systems 22, Neural Information Processing Systems Foundation*, pp. 1276–1284. [401]

Newman, M. E. J. (2003), "The Structure and Function of Complex Networks," *SIAM Review*, 45, 167–256. [390]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [390]

Pitman, J. (1996), "Some Developments of the Blackwell-MacQueen Urn Scheme," *IMS Lecture Notes–Monograph Series*, 30, 245–267. [395]

———, J. (2006), *Combinatorial Stochastic Processes*, Berlin: Springer-Verlag. [393]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society,* Series B, 64, 583–639. [398,400]