Markov chain Monte Carlo for Structural Inference with Prior Information

Sach Mukherjee & Terence P. Speed
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
{sach,terry}@stat.berkeley.edu

April 5, 2007

Abstract

This paper addresses the question of making inferences regarding features of conditional independence graphs in settings characterized by the availability of rich prior information regarding such features. We focus on Bayesian networks, and use Markov chain Monte Carlo to draw samples from the relevant posterior over graphs. We introduce a class of "locally-informative priors" which are highly flexible and capable of taking account of specific information regarding graph features, and are, in addition, informative at a scale appropriate to local sampling moves. We present examples of such priors for beliefs regarding edges, groups and classes of edges, degree distributions and sparsity, applying our methods to challenging synthetic data as well as data obtained from a biological network in cancer.

1 Introduction

In recent decades, rich developments in computational methods have allowed statisticians to perform inference using increasingly realistic, complex data models. At the same time, in the broader research community there has been a growing interest in complex, multi-variable systems, a trend which has been greatly influenced by continuing advances in experimental methodologies capable of making measurements on such systems.

Bayesian statistics in particular has benefited greatly from these technical and scientific developments. Computational tools like Markov chain Monte Carlo have broadened the applicability of rich Bayesian models, and the increasingly close

integration between computational statistics and fields such as bioinformatics, finance and data mining has been accompanied by an increasing demand for statistical methods capable of taking account of relevant domain knowledge.

A specific trend which has begun to gather pace is an interest in studying systems characterized by multiple interacting components. For example, in the field of molecular biology, there has been a movement away from thinking about one gene or protein at a time to thinking about multiple genes and proteins acting in concert. Indeed, it is largely this type of thinking that characterizes so-called "systems" approaches to biology (see e.g. Kitano, 2002; Ideker and Lauffenburger, 2003). In statistical terms, this has led to much interest in multivariate methods, and in network-orientated models.

Graphical models (Pearl, 1988; Lauritzen and Spiegelhalter, 1988; Lauritzen, 1996; Jordan, 2004) are a class of statistical models which provide graph-based representations of conditional independence relationships between random variables. A graphical model consists of a graph G, describing a set of conditional independence statements, and parameters $\mathbf{\Theta}$ which specify conditional distributions implied by G. Often, the graph G is known, and inferential questions concern specific marginal and conditional distributions. Three decades of research have provided a rich array of theory and computer algorithms with which to address such questions. However, in many settings, questions of interest concern the conditional independence graph itself. For example, in molecular biology, we may be interested in saying something about which molecules or combinations of molecules influence one another; in the social sciences we may be interested in relationships between various economic and demographic variables. Such questions can often be cast, in a quite natural manner, in terms of features, such as edges, classes of edges, or paths, of conditional independence graphs.

The daunting nature of inference on graphical model structure is well known, and is largely due to the vast space of possible models in even moderately large domains. Yet, equally, in many settings, an understanding of the relevant domain may suggest that not every possible graph is equally plausible, and that certain features should be regarded as *a priori* more likely than others. Where available, such knowledge, even when uncertain, is surely a valuable resource, making the question of how to capture and exploit it an important one.

This paper seeks to address precisely this question, of making inferences regarding conditional independence graphs in the presence of prior knowledge regarding graph features. We focus on directed graphical models called Bayesian networks, and use Markov chain Monte Carlo (MCMC) for structural inference. Motivated by the kinds of questions alluded to above, the model averaging methods described in this paper are aimed not so much at recovering the correct graph, but more as a flexible device for addressing questions concerning features of graphs.

MCMC-based inference on conditional independence graphs is a topic which has attracted a great deal of interest in recent years in statistics as well as machine learning (Madigan et al., 1995; Dellaportas and Forster, 1999; Giudici and Green, 1999; Friedman and Koller, 2003; Giudici and Castelo, 2003; Tarantola, 2004; Dellaportas and Tarantola, 2005). Our work adds to the existing literature in two ways. Firstly, we place an emphasis on making use of rich prior information regarding graph structure. Much of the existing literature on structural inference has used flat priors on graphs (e.g. Madigan et al., 1995; Giudici and Castelo, 2003), or priors designed to promote sparse models by penalizing graphs with too many edges (e.g. Friedman and Koller, 2003; Jones et al., 2005). In contrast, we seek to take account of detailed information concerning features of graphs such as individual edges, classes of edges and degree distributions on vertices. In many domains, such beliefs follow, in a natural manner, from a consideration of the underlying science or semantics of the variables under study. We argue that such information can be profitably exploited in structural inference. We present priors which can be used in this fashion, and show examples of how these ideas can be put to use for practical problems. A second emphasis is on settings in which the number of observations is small relative to the complexity of the system under study. It is frequently observed that there is a "deluge" of data in modern science. Yet, in our experience, it is often the case that there is *not enough* data concerning rich, multivariate systems. In molecular biology, for example, obtaining large datasets can be costly and labour-intensive, especially if a high level of biochemical detail is desired. Equally, in the social sciences, there may be a natural limit on the number of units upon which measurements can be made, such as in the case of studying demographic or economic variables at the level of zip codes.

These elements - approximate inference on graphs, structural priors and sample size - are inter-related. At smaller sample sizes, posterior distributions over graph space tend to be diffuse, with significant probability mass in many regions of the space. This can serve as a motivation for the use of sampling methods, but at the same time can mean that such methods must explore ever larger, dispersed areas of graph space in order to account for a given fraction of probability mass. This in turn motivates the need to exploit prior information regarding competing graphs to guide inference and refine the questions being asked. Indeed, one of our key empirical findings is that in the analysis of small sample size data the use of structural priors leads to substantive gains in the accuracy of inferences regarding graph features. A Receiver Operating Characteristic or ROC analysis of decisions on individual edges, presented in Section 4, reveals large gains in sensitivity and specificity compared with both structural inference with a flat prior and simple pairwise associations.

A natural concern regarding informative priors on graphs is whether their use

amounts to "putting too much in" during inference. We hold the view that even strong priors on graphs can play a valuable role in sharpening questions being asked, in a manner analogous to a well thought out set of hypotheses, but with an added degree of flexibility and generality. Consider, as an example, the five variables illustrated in Figure 1. Suppose we knew, from outside knowledge, that the As tend to influence the Bs, and that the main question we were interested in addressing was which combination of As influence each of the Bs. One way of posing this question would be as a classical multiple decision problem. A second approach would be to perform structural inference on the variables, with a strong prior in favour of models in which As influence Bs. The structural prior would then play a role similar to the hypothesis formulation step in the first approach. Yet the structural analysis offers two key advantages. Firstly, it allows for the discovery of unexpected relationships, when such relationships are well-supported by the data. When the variables of immediate interest are embedded in a larger system such relationships could also include outside influences of one kind or another. Secondly, structural inference offers a mechanism by which to simultaneously address a range of possible questions concerning relationships between variables: once we have described a posterior distribution over models, we are free to evaluate probabilities or odds concerning essentially any structural features of interest.

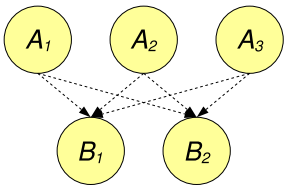


Figure 1: Priors for hypothesis formulation. A prior preferring models in which As influence Bs can play a role similar to a hypothesis formulation step in a multiple decision approach.

The remainder of this paper is organized as follows. We begin by reviewing basic ideas and notation for Bayesian networks and structural inference. We then turn our attention to priors on graphs. We introduce a class of priors which we call *locally-informative priors*. These priors are highly flexible and can take account of rich, specific information regarding graph features. In addition, they provide information at a scale appropriate to local sampling moves and, as a consequence, are particularly well-suited to our sampling-based approach. We present a number of examples of the use of our methods, including analyses of challenging synthetic data and of a biological network in breast cancer. We close the paper with a dis-

cussion of the key points and shortcomings of our work and some ideas for further research.

2 Background

2.1 Bayesian Networks

Bayesian networks (Pearl, 1988; Lauritzen, 1996) are a type of multivariate statistical model in which a directed acyclic graph describing conditional independence statements regarding a group of random variables is exploited to provide a compact description of their joint distribution. A Bayesian network consists of two elements: (i) a directed acyclic graph $G = (V(G), E(G))$, whose vertices V represent random variables $X_1 \dots X_p$ of interest, and whose edge-set E contains edges describing conditional independencies between those variables, and (ii) parameters $\mathbf{\Theta}$ which specify the conditional distributions implied by the graph. In particular, the graph G implies that each variable is conditionally independent of its non-descendants given its immediate parents. Importantly, this means that the joint distribution $P(X_1 \dots X_p)$ can be factorized into a product of local terms:

$$P(X_1 \dots X_p \mid G) \quad = \quad \prod_{i=1}^{p} P(X_i \mid \mathbf{Pa}_G(X_i)) \tag{1}$$

where, $\mathbf{Pa}_G(X_i)$ is the set of parents of X_i in graph G.

The goal of structural inference is make inferences regarding the graph G given observations of the variables $X_1 \dots X_p$. Let $\mathbf{X}$ represent a $p \times n$ data matrix, where n is the number of multivariate samples available. Using Bayes' theorem, the posterior probability of graph G can be written as follows:

$$P(G \mid \mathbf{X}) \quad = \quad \frac{p(\mathbf{X} \mid G) P(G)}{p(\mathbf{X})} \tag{2}$$

where, $p(\mathbf{X} \mid G)$ is the (marginal) likelihood and $P(G)$ is a prior distribution over directed acyclic graphs; we refer to the latter as a *structural prior*.

Now, the graph G does not in itself specify a full data model, since it describes only the conditional independence structure of the variables, but neither the form of the conditional distributions which relate child nodes to parents, nor the parameters of those distributions. If we assume that the form of the conditional distributions $P(X_i \mid \mathbf{Pa}(X_i))$ is known, we need only specify parameters to obtain a full model. Let $\mathbf{\Theta}$ represent a complete set of model parameters. Then, the marginal likelihood $p(\mathbf{X} \mid G)$ can be evaluated by integrating over parameters $\mathbf{\Theta}$:

$$p(\mathbf{X} \mid G) \quad = \quad \int p(\mathbf{X} \mid G, \mathbf{\Theta}) p(\mathbf{\Theta} \mid G) \, \mathrm{d}\mathbf{\Theta} \tag{3}$$

where, $p(\mathbf{\Theta})$ is a prior over parameters.

This paper is concerned mainly with inferences regarding the graph G itself, and the ideas presented below are applicable for any choice of conditional distributions and parameter priors under which the marginal likelihood $p(\mathbf{X} \mid G)$ can be evaluated. In our experiments, we follow previous authors (Cooper and Herskovits, 1992; Heckerman et al., 1995; Giudici and Castelo, 2003) in assuming parameter independence and using Multinomial conditionals and Dirichlet priors. This allows the marginal likelihood to be evaluated in closed form. Here, we reproduce the well-known result of Heckerman et al. (1995) and refer the interested reader to the reference for further details:

$$p(\mathbf{X} \mid G) \quad = \quad \prod_{i=1}^{p} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N_{ijk})} \tag{4}$$

where, N_{ijk} is the number of observations in which X_i takes the value k, given that $\mathbf{Pa}_G(X_i)$ has configuration j; q_i are the number of possible configurations of parents $\mathbf{Pa}_G(X_i)$; and r_i are the number of possible values of X_i. N'_{ijk} are Dirichlet hyperparameters. Finally, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$.

2.2 MCMC for structural inference

The posterior distribution $P(G \mid \mathbf{X})$ is a discrete distribution over the space $\mathcal{G}$ of all possible directed acyclic graphs with p vertices. We may rewrite (2) as follows, explicitly summing over graphs in the denominator:

$$P(G \mid \mathbf{X}) \quad = \quad \frac{p(\mathbf{X} \mid G)P(G)}{\sum_{G \in \mathcal{G}} p(\mathbf{X} \mid G)P(G)} \tag{5}$$

The number of possible graphs grows super-exponentially with the number of variables p. Indeed, Robinson (1973) has shown that the number $|\mathcal{G}_p|$ of possible directed acyclic graphs with p vertices is given by the following recurrence formula:

$$|\mathcal{G}_p| \quad = \quad \sum_{i=1}^{p} (-1)^{i+1} \binom{p}{i} 2^{i(p-i)} |\mathcal{G}_{(p-i)}|$$

where, $|\mathcal{G}_1| = 1$ and $|\cdot|$ indicates the cardinality of its argument.

This gives $|\mathcal{G}_2| = 3$, $|\mathcal{G}_3| = 25$, $|\mathcal{G}_{10}| \approx 4.2 \times 10^{18}$, $|\mathcal{G}_{14}| \approx 1.4 \times 10^{36}$ and so on. The number of possible graphs is therefore usually much too large to permit the distribution to be enumerated exhaustively. Thus, while we can evaluate the posterior probability of a graph upto a multiplicative constant, we cannot actually consider every possible graph in the course of inference.

The intractability of the sum in (5) motivates the use of stochastic simulation methods to approximate posterior distributions over graphs. *Markov Chain Monte Carlo* or *MCMC* represents a general class of such methods which are widely used in computational statistics. The basic idea of MCMC is to construct a Markov chain whose state space is the domain of the desired random quantity, and whose stationary distribution is its posterior. Then, simulating the Markov chain provides a means by which to make inferences based on the posterior distribution of interest.

In a *Metropolis-Hastings* sampler (Hastings, 1970), draws are made from a *proposal distribution* Q, which depends on current state, and then accepted or rejected in such a way as to guarantee that, asymptotically, they behave as draws from the desired target distribution. Here, following Madigan et al. (1995) and Giudici and Castelo (2003), we develop a MCMC sampler of the Metropolis-Hastings type for the purpose of simulating the posterior distribution $P(G \mid \mathbf{X})$ over conditional independence graphs.

Let $\eta(G)$ denote a *neighbourhood* around a directed acyclic graph G, consisting of every directed acyclic graph which can be obtained by adding, deleting or reversing a single edge in G. Define proposal distribution Q as follows:

$$Q(G';G) = \begin{cases} \frac{1}{|\eta(G)|} & \text{if } G' \in \eta(G) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Then, calculate the following *acceptance probability* α:

$$\alpha = \frac{P(G' \mid \mathbf{X})Q(G;G')}{P(G \mid \mathbf{X})Q(G';G)} \tag{7}$$

Since the proposal distribution is uniform over the relevant neighbourhood, the ratio $Q(G;G')/Q(G';G)$ may be written in terms of neighbourhood size:

$$\alpha = \frac{P(G' \mid \mathbf{X})|\eta(G)|}{P(G \mid \mathbf{X})|\eta(G')|} \tag{8}$$

A proposed graph G', drawn from Q, is then *accepted* with probability $\min(1, \alpha)$, and otherwise *rejected*. If accepted, G' is added to the sequence of samples drawn, and becomes the current graph. Else, G is added to the sequence of samples, and remains the current graph. As shown in Madigan et al. (1995) and Giudici and Castelo (2003), the proposal distribution Q gives rise to an irreducible Markov chain, since there is positive probability of reaching any part of the state space $\mathcal{G}$. Standard results (see, e.g., Tierney, 1994; Gilks et al., 1996) then guarantee that the Markov chain must converge to the desired posterior $P(G \mid \mathbf{X})$. The sampler described above is summarized in Algorithm 1.

Algorithm 1 A Metropolis-Hastings sampler for structural inference.

(1) Initialize graph $G^{(1)}$, set $t = 1$, $G \leftarrow G^{(1)}$

(2) **Propose** $G' \sim Q(G'; G)$

(3) **Accept** G' with probability $\min(1, \alpha)$, $\alpha = \frac{P(G'|\mathbf{X})Q(G;G')}{P(G|\mathbf{X})Q(G';G)}$.

(4) **Update** If G' is accepted, $G^{(t+1)} \leftarrow G'$, $G \leftarrow G^{(t+1)}$ else $G^{(t+1)} \leftarrow G$. Set $t \leftarrow t+1$

(5) While $t < T$, repeat (2)-(4).

During sampling, we only need the posterior distribution in order to compute acceptance ratio α. This means that the unnormalized quantities $p(\mathbf{X} \mid G')P(G')$ and $p(\mathbf{X} \mid G)P(G)$ are sufficient for our purposes. We have discussed the marginal likelihood $p(\mathbf{X} \mid G)$ above; we turn our attention to the prior $P(G)$ below. We note also that efficient local computations, as described in Heckerman et al. (1995) and Giudici and Castelo (2003), suffice to compute the Bayes factor $p(\mathbf{X} \mid G')/p(\mathbf{X} \mid G)$ at each iteration.

As shown in Algorithm 1, iterating "propose", "accept" and "update" steps gives rise to samples $G^{(1)} \ldots G^{(T)}$. An important property of these samples is that, provided the Markov chain has converged to its stationary distribution, they provide a means by which to compute the expectation of essentially any function on graphs. Specifically, if $\mathbb{E}[\phi(G)]_{P(G|\mathbf{X})}$ is the expectation, under the posterior, of a function $\phi(G)$, then

$$\hat{\mathbb{E}}[\phi(G)] \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\phi(G^{(t)}) \tag{9}$$

is, by standard results, an asymptotically valid estimator of $\mathbb{E}[\phi(G)]_{P(G|\mathbf{X})}$.

An important special case of (9), which we shall make use of below, concerns the posterior probability of an individual edge e, or $P(e \mid \mathbf{X})$. We may write $P(e \mid \mathbf{X})$ as a posterior expectation as follows:

$$\begin{aligned} P(e \mid \mathbf{X}) &= \sum_{G \in \mathcal{G}} P(e \mid G, \mathbf{X}) P(G \mid \mathbf{X}) \\ &= \sum_{G \in \mathcal{G}} I_{E(G)}(e) P(G \mid \mathbf{X}) \\ &= \mathbb{E}[I_{E(G)}(e)]_{P(G|\mathbf{X})} \end{aligned}$$

where, I_A is the indicator function for set A.

Then, applying (9), we may use samples $G^{(1)} \dots G^{(T)}$ to obtain an asymptotically valid estimate of $\mathbb{E}[I_{E(G)}(e)]$:

$$\hat{\mathbb{E}}[I_{E(G)}(e)] \quad = \quad \frac{1}{T}\sum_{t=1}^{T} I_{E(G^{(t)})}(e) \tag{10}$$

where, $G^{(t)} = (V(G^{(t)}), E(G^{(t)}))$.

3 Priors on graphs

In this Section, we discuss the use of prior information concerning graph features. We begin with a motivating example which highlights some of the different kinds of prior beliefs which are encountered in practice and which we might like to take account of during inference. We then introduce a class of priors on graphs which we call *locally-informative priors*. These priors are quite general in nature, and are well-suited to structural inference using the sampler described above. We provide examples of locally-informative priors for information regarding individual edges, classes of edges, degree distributions and sparsity. We close the Section by looking at the use of proposal distributions based on structural priors.

3.1 A motivating example

We begin with a motivating example taken from cancer biology, which is paradigmatic of the broad class of structural inference problems with which this paper is concerned. Our choice of example is motivated by our own applied interests, but questions of this general type arise in many areas of biology, the social sciences, and data mining, so we invite the reader to substitute in its place a motivating example of her own choice.

Table 1 shows 14 proteins which are components of a biological network called the *Epidermal Growth Factor Receptor* or *EGFR* system. Here, each protein is either a *ligand*, *receptor* or *cytosolic protein*; for our present purposes, these may be regarded as well-defined classes of variable.

Our general goal is to infer structural features of the biological network in which these components participate. We model the relevant biochemical connectivity in terms of conditional independence. Then, questions regarding relationships between molecular components can be expressed, in a natural fashion, as questions regarding features of conditional independence graphs. (Naturally, this conceptualization raises important semantic issues, but in light of the largely methodological goals of the present paper, we do not discuss these here.)

Table 1: Some components of the Epidermal Growth Factor Receptor system.

Protein	**Type**	**Protein**	**Type**
EGF	*Ligand*	GAP	*Cytosolic protein*
AMPH	*Ligand*	SHC	*Cytosolic protein*
NRG1	*Ligand*	RAS	*Cytosolic protein*
NRG2	*Ligand*	Raf	*Cytosolic protein*
EGFR	*Receptor*	MEK	*Cytosolic protein*
ERBB2	*Receptor*	ERK	*Cytosolic protein*
ERBB3	*Receptor*		
ERBB4	*Receptor*		

The biochemistry of the system provides us with some prior knowledge regarding graph features, which we would like to take account of during inference. Some illustrative examples of the kind of knowledge which might be available include:

(S1) Ligands influence cytosolic proteins via ligand-receptor interactions. As a consequence, we do not expect them to directly influence cytosolic proteins. Equally, we do not expect either receptors or cytosolic proteins to directly influence ligands.

(S2) Certain ligand-receptor binding events occur with particularly high affinity; these include EGF and AMPH with EGFR, NRG1 with ERBB3, and NRG1 and NRG2 with ERBB4. Equally, the receptors EGFR, ERBB3 and ERBB4 are all capable of influencing the state of ERBB2 (via heterodimer formation and transphosphorylation). Also, there is much evidence indicating that Raf can influence MEK, which in turn can influence ERK.

(S3) Since we observe ligand-mediated activity at the level of cytosolic proteins, we expect to see a path from ligands to receptors, and from receptors to cytosolic proteins.

Without going into a great deal of biological detail it is clear that these beliefs correspond to information regarding graph structure: (S1) contains information concerning classes of edges; (S2) contains information regarding specific edges and (S3) contains information regarding edges between classes of vertices.

3.2 Locally informative priors

Since the scale of Metropolis-Hastings moves is controlled by the proposal distribution Q, it makes sense to use a prior which is informative at the same scale

as the proposal. In this Section we introduce a class of priors on graphs which are designed to match the local scale of our Metropolis-Hastings sampler: we call these *locally-informative priors*. We first describe locally-informative priors at a very general level, and go on to provide specific examples of such priors for beliefs regarding individual edges, classes of edges, degree distributions and graph sparsity. We highlight a number of connections to existing ideas concerning priors on graphs, and finally offer a few comments on the contrast between global and local information in the context of sampling.

3.2.1 Concordance functions

Let $f(G)$ be a real-valued function on graphs which:

(i) is increasing in the *degree* to which graph G accords with prior beliefs, and

(ii) typically takes on more than one distinct value in a neighbourhood $\eta(G)$.

We call f a *concordance function*, since it indicates concordance with prior beliefs. Then, a *locally-informative prior* is a prior of the following form:

$$P(G) \quad \propto \quad g(f(G)) \tag{11}$$

where, g is a monotone increasing function.
In all our experiments, we use:

$$g(f(G)) \quad = \quad \lambda^{f(G)}, \;\; \lambda \geq 1 \tag{12}$$

with the parameter λ used to control the strength of the prior.

3.2.2 Examples

The crucial element in the formulation (11) is the concordance function f. If f typically takes on multiple distinct values in local neighbourhoods, the resulting prior will be informative in such local neighbourhoods.

It turns out to be relatively straightforward to specify concordance functions corresponding to prior beliefs of various kinds. We provide now a few examples of concordance functions.

Individual edges. Suppose we believe that certain edges are *a priori* more plausible than others. Such beliefs may be *positive* or *negative*, depending on whether we belief the edges are likely to be present or absent in the data-generating graph. Let E_+ denote a set of edges concerning which we have positive belief ("positive edge set") and E_- be a set of edges concerning which we have negative belief ("negative

edge set"). We assume that these two sets are disjoint, such that $E_+ \cap E_- = \emptyset$. Then, we suggest the following concordance function:

$$f(G) \quad = \quad |E(G) \cap E_+| - |E(G) \cap E_-| \tag{13}$$

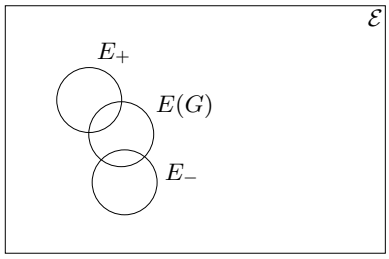


Figure 2: Positive and negative edge sets; $\mathcal{E}$ denotes the set of all possible edges.

This is essentially a counting function on individual edges, which attains its maximum value $|E_+|$ if and only if G contains all the positive edges and no negative edges. Importantly, the function can take on a range of possible values, and is sensitive to local changes in graph structure involving edges which are members of either the positive or negative edge sets.

In the motivating example presented above, (S1) contains negative prior information, while (S2) contains positive prior information regarding individual edges. Such information can be captured in quite natural way using (13). We note also that the notion of specifying a particular prior graph $G_0 = (V_0, E_0)$, and penalizing graphs on the basis of the number of edges by which they differ from G_0 (Heckerman et al., 1995) is a special case of (13), with $E_+ = E_0$ and $E_- = E_0^c$.

In the remainder of the paper we will use (13) to capture prior beliefs concerning individual edges. We note however that the following, more general concordance function allows beliefs regarding individual edges to be weighted in accordance with their strength:

$$f(G) \quad = \quad \sum_i w_i I_{E(G)}(e_P^i) \tag{14}$$

where, $\{e_P^i\}$ denotes a set of edges concerning which we have prior beliefs, either positive or negative, and w_i are edge-specific weights.

The simpler function in (13) is then a special case of this general counting function with $w_i = 1$ if the corresponding edge e_P^i represents a positive prior belief and $w_i = -1$ if edge e_P^i represents a negative prior belief.

Classes of edges. Concordance function (13) may also be used to capture beliefs regarding classes of edges. Let $\{\mathcal{C}_k\}$ be a set of classes into which vertices $v \in V$ can be categorised. Suppose we wish to penalize graphs displaying edges between vertices of type i and j. This can be accomplished simply by using the concordance function(13) with a negative edge set E_- containing all such edges:

$$E_- \quad = \quad \{e = (v_k, v_l) \cdot \mathcal{C}(v_k) = \mathcal{C}_i, \mathcal{C}(v_l) = \mathcal{C}_j\} \tag{15}$$

Positive priors on classes of edges can be defined in a similar fashion.

Higher-level graph features. In many cases, we may wish to capture prior knowledge concerning higher-level graph features which cannot be described by reference to individual edges. To take but one example, we may believe that there ought to be at least one edge between certain classes of vertices, as in (S3) above. Examples of knowledge of this kind are abundant in molecular biology, where the classes may represent distinct types of molecule thought to influence one other in specific ways.

As above, let $\{\mathcal{C}_k\}$ be vertex classes. Also, let $\mathcal{C}(v)$ denote the class to which a vertex v belongs. Let $E_{\mathcal{C}}$ be a set of ordered pairs of classes such that $(\mathcal{C}_i, \mathcal{C}_j) \in E_{\mathcal{C}}$ means that we have a belief that there ought to be at least one edge from class $\mathcal{C}_i$ to class $\mathcal{C}_j$. Then, the following concordance function captures the belief that there ought to be at least one edge between specified pairs of classes:

$$f(G) \quad = \quad \sum_{(\mathcal{C}_i, \mathcal{C}_j) \in E_{\mathcal{C}}} I_{\mathbb{Z}^+} \left[\sum_{(v_1, v_2) \in E(G)} \delta \left((\mathcal{C}(v_1), \mathcal{C}(v_2)), (\mathcal{C}_i, \mathcal{C}_j) \right) \right] \tag{16}$$

where, $\mathbb{Z}^+$ is the set of positive integers.

Thus, (16) counts the number of pairs in $E_{\mathcal{C}}$ which are represented by at least one edge in a graph. We provide a practical example and application of a concordance function of this kind in an analysis of a protein network presented below.

Degree distributions. Suppose we have reason to believe that the degree distribution of the graph is likely to be scale-free. This is a property which has been investigated widely in recent years, in contexts ranging from systems biology to the structure of the internet (refs). The *degree* $\deg(v)$ of a vertex v is the total number of edges in which vertex v participates. The degree distribution of a graph G is a function $\pi_G(\delta)$ describing the total number of vertices having degree δ:

$$\pi_G(\delta) \quad = \quad \sum_{v \in V(G) \cdot \deg(v) = \delta} 1 \tag{17}$$

A graph is said to have a *scale-free* degree distribution if π_G follows a power-law with $\pi_G(\delta) \propto \delta^{-\gamma}, \gamma > 0$ such that $\log(\pi_G(\delta))$ should be approximately

linear in $\log(\delta)$. Accordingly, we suggest using the negative correlation coefficient between $\log(\pi_G(\delta))$ and $\log(\delta)$ as a concordance function to capture the extent to which the degree distribution of a graph G can be regarded as scale-free:

$$f(G) \quad = \quad -r(\log(\pi_G(\delta)), \log(\delta)) \tag{18}$$

where, $r(\cdot, \cdot)$ denotes the correlation coefficient of its arguments.

Again, since the concordance function (18) is sensitive to local changes in graph structure, it will typically take on multiple distinct values in a given neighbourhood.

Sparsity. In many settings, it can be advantageous to promote parsimonious models by using priors which promote sparse graphs. Such priors differ from those discussed already in that they aim to promote a general, statistically desirable feature rather than capture specific domain knowledge. Here, we describe two ways of promoting sparsity: by penalizing large *in-degrees* and by penalizing the *total number of edges*.

Since Bayesian networks factorize joint distributions into local terms conditioned on parent configurations, model complexity typically grows - often very rapidly - with the number of parents. Controlling the in-degree of graphs can therefore be an effective means of controlling model complexity. The *in-degree* $\mathrm{indeg}(v)$ of a vertex $v \in V$ is the number of edges in edge-set E leading into v, that is $\mathrm{indeg}(v) = |\{(v_i, v_j) \cdot (v_i, v_j) \in E, v_j = v\}|$. Let $\Delta(G) = \max_{v \in V(G)} \mathrm{indeg}(v)$ be the *maximum in-degree* of graph G. Then, the following concordance function expresses a preference for graphs having in-degree not exceeding λ_{indeg}:

$$f(G) \quad = \quad \min(0, \lambda_{\mathrm{indeg}} - \Delta(G)) \tag{19}$$

An alternative way to promote sparsity is by penalizing the total number of edges in a graph, for example using a Binomial distribution over the total number of edges, with parameters set to ensure an expected number of edges equal to the number of variables p, and an appropriate maximum number of possible edges (Buntine, 1991; Jones et al., 2005).

Combining concordance functions. Finally we note that multiple concordance functions $\{f_i(G)\}$ may be combined using a function $g(f_1(G), f_2(G), \ldots)$ which is monotone increasing in each of its arguments. We suggest

$$P(G) \propto \prod_i \lambda_i^{f_i(G)} \tag{20}$$

where the λ_i's are strength parameters for the concordance functions.

3.2.3 Local and global information

The Metropolis-Hastings acceptance ratio α in (7) may be written as a product

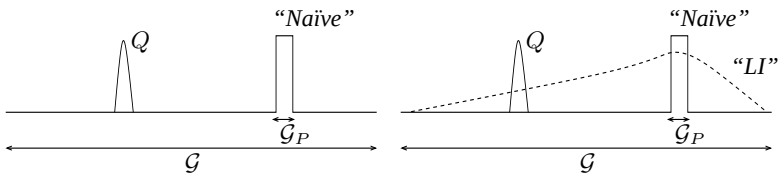
$$\alpha \;=\; \frac{P(\mathbf{X} \mid G')}{P(\mathbf{X} \mid G)} \times \frac{Q(G;G')}{Q(G';G)} \times \frac{P(G')}{P(G)}$$

of a likelihood ratio, Hastings ratio and prior odds. This means that the only way in which prior knowledge enters into the sampling process is via the prior odds $P(G')/P(G)$ in favour of a proposed graph over a current graph. As a consequence, the behavior of the ratio $P(G')/P(G)$ during sampling is central to the effectiveness of the prior. In particular, a prior which typically takes on a range of values over the local neighbourhoods in which the proposal distribution operates will tend to provide informative prior odds ratios during sampling.

To illustrate this point, consider a naïve prior of the following form:

$$P(G) \propto \begin{cases} \lambda & \text{if } G \in \mathcal{G}_P \\ 1 & \text{otherwise} \end{cases} \tag{21}$$

where, $\mathcal{G}_P$ is a set containing all graphs which fully accord with prior beliefs and $\lambda \geq 1$ is the prior odds in favour of such graphs. For our motivating example, such a prior would take on the value $k\lambda$ (for some constant k) only for graphs displaying all characteristics in (S1), (S2) and (S3) and k otherwise.



(a) Naïve prior (b) Locally-informative (LI) prior

Figure 3: Local versus global information. The relationship between the scales of the proposal and prior distributions play a key role in ensuring that the prior provides information during sampling. Here, (a) a naïve prior is flat across a typical local neighbourhood, while (b) a locally-informative prior takes on a range of values in such a neighbourhood.

This naïve prior expresses an clear preference for graphs with certain features and is therefore informative at a global scale. Yet, for relatively specific prior information and a moderate number of variables, typically $|\mathcal{G}_P| \ll |\mathcal{G}|$, such that the prior will very often take on the *same* value across a given neighbourhood and therefore provide relatively little information at a local scale. Such a prior may

therefore, in practice, operate as a *de facto* flat prior on graphs. It is in this sense that it is important to reconcile the scale of the prior and the proposal distribution, and in this sense that the priors we have put forward are locally-informative. Figure 3 shows schematically why it is that a naïve prior tends to be flat across local neighbourhoods, while a locally-informative prior tends to show variation in such neighbourhoods.

3.3 Prior-based proposals

The proposal distribution (6) is uniform over a neighbourhood $\eta(G)$. Yet the prior $P(G)$ provides potentially valuable information regarding which graphs are *a priori* most likely. A natural idea, then, is to use this information to guide the proposal mechanism. However, care must be taken to ensure that (i) irreducibility of the Markov chain is maintained, and (ii) that the Hastings factor $Q(G;G')/Q(G';G)$ and the prior odds $P(G')/P(G)$ do not simply cancel each other out or result in a lowering of the acceptance probability for *a priori* likely graphs. Due to the second of these concerns, we do not recommend the use of prior-based proposals as a matter of course, but rather only when necessitated by especially complex, multi-modal graph spaces and in the presence of strong prior information. In such settings, and for integer-valued concordance functions f, we suggest a proposal distribution of the following form:

$$Q_P(G';G) \propto \begin{cases} \lambda_Q & \text{if } P(G') > P(G) \\ 1 & \text{if } P(G') = P(G) \\ 1/\lambda_Q & \text{if } P(G') < P(G) \\ 0 & \text{if } G' \notin \eta(G) \end{cases} \tag{22}$$

where, $\lambda_Q \geq 1$ is a parameter controlling the strength with which the proposal mechanism prefers *a priori* likely graphs.

This ensures that (i) all graphs in $\eta(G)$ have a non-zero probability of being proposed, thereby preserving irreducibility, and (ii) the Hastings factor is at most on the order of λ_Q^2, such that for $P(G) \propto \lambda^{f(G)}$, setting $\lambda > \lambda_Q^2$ ensures that *a priori* likely graphs remain likely to obtain good acceptance ratios.

4 Experiments

4.1 Simulation

4.1.1 Data

We simulated data for the $p = 14$ variables described previously in Table 1, using the data-generating graph shown in Figure 4(a). One of our stated goals was to ad-

dress the question of structural inference at small sample sizes; accordingly we set the sample size for our simulation dataset to $n = 200$. Details of our data-generating model are as follows: the random variables are binary $\{0, 1\}$; all conditional distributions are Bernoulli, with success parameter p depending upon the configuration of the parents. In particular, root nodes are sampled with $p = 0.5$, while for each child node, $p = 0.8$ if at least one parent takes on the value 1, and $p = 0.2$ otherwise. This gives each child node a relationship to its parents which is similar to a logical OR.

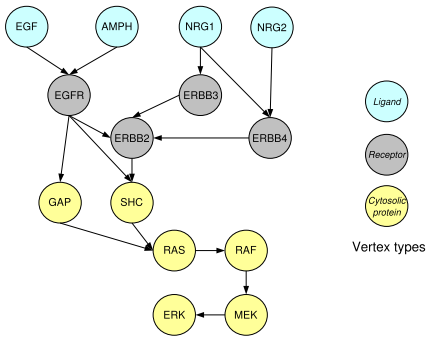
4.1.2 Priors

The graph shown in Figure 4(a) is based on the biochemistry of the Epidermal Growth Factor Receptor system alluded to the motivating example presented above. We constructed locally informative priors corresponding to the beliefs (S1) and (S2) described in Section 3.1 above. Using (15), we defined a a negative edge set E_- from (S1); (S2) defined a positive edge set E_+ in a natural manner. The concordance function (13) was then used with these edge sets. (S3) was not used in these experiments. To investigate the effects of weaker priors and of priors containing erroneous information, we also constructed a *partial prior* and a *mis-specified prior*. The partial prior uses (S2) but not (S1) and therefore contains information on some specific edges, but no information on classes of edges. The mis-specified prior includes in its negative edge set edges from Raf to MEK and from MEK to ERK, and in its positive edge set an edge from Ras to ERK. This prior allows us to consider a realistic setting in which the prior is largely reasonable but contains a number of egregiously false beliefs. Finally, as a baseline comparison, we also computed results using a "flat" prior with $P(G) = k$. For all locally-informative priors, we set $\lambda = e$.

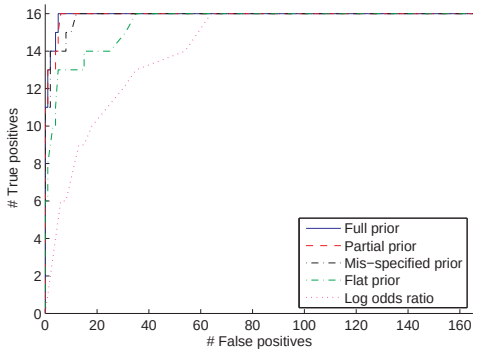
4.1.3 Results

Receiver Operating Characteristic or *ROC curves* are plots of true positive against false positive rates, and provide an 'at a glance' summary of error rates across a range of thresholds. The fact that we know the true data-generating graph allowed us to generate ROC curves from true and false positive calls on individual edges. Let $G^* = (V^*, E^*)$ denote the true data-generating graph. As before, let $P(e \mid \mathbf{X})$ denote the posterior probability of an edge $e = (v_i, v_j)$. Then, the set of edges called at threshold $\tau \in [0, 1]$ is

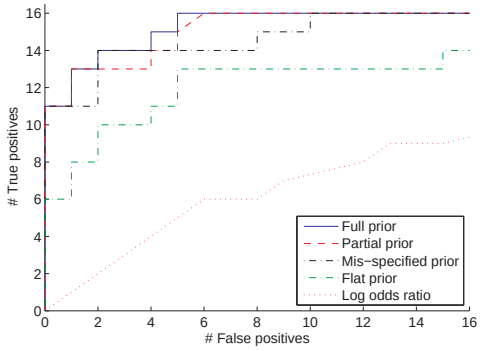
$$E_\tau \quad = \quad \{e = (v_i, v_j) \cdot v_i \in V, v_j \in V, P(e \mid \mathbf{X}) \geq \tau\},$$



(a) Data-generating graph



(b) Full



(c) Detail

Figure 4: ROC curves, synthetic data. (a) Data-generating graph; (b) Full ROC curves; (c) Detail of (b).

the number of true positives is $|E_\tau \cap E^*|$ and the number of false positives is $|E_\tau \setminus E^*|$.

ROC curves were then obtained by plotting, for each sampler, the number of true positives against the number of false positives parameterized by threshold τ. Thus, these curves are computed by comparison with the "gold-standard" edge-set E^*; they provide our key comparative result. Figure 4(b),(c) show ROC curves obtained using each of the full, partial, and mis-specified locally-informative priors, and, for comparison, the flat prior and absolute *log odds ratios* $|\psi_{ij}|$ computed for each pair (i, j) of variables; these provided a natural measure of association between pairs of binary variables.

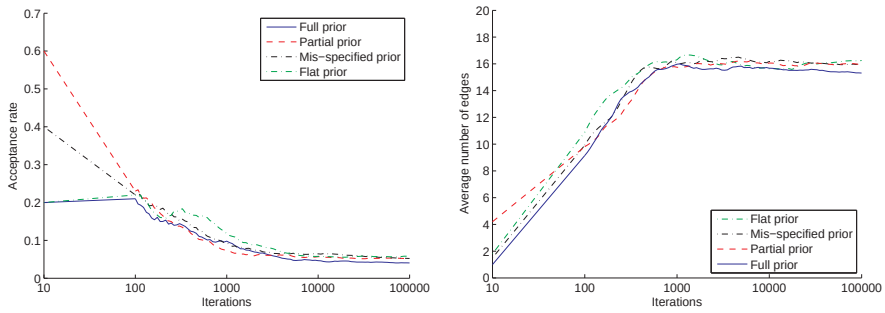
The locally-informative priors provide substantial gains in sensitivity and specificity: the full and mis-specified priors called 11 edges and the partial prior called 9 edges correctly before encountering a false positive. The full prior discovered all 16 edges in the data-generating graph at the cost of only 5 false positives; the partial prior required 8, the mis-specified prior 9 and the flat prior 30 false positives to recover all true edges. The log-odds ratio did substantially worse than any of the Bayesian network analyses, requiring 62 false positives to find all edges in the data-generating graph.

Figure 5(a) shows, for each sampler, the average probability of acceptance plotted against number of MCMC iterations. Figure 5(b) shows the average number of edges plotted against number of MCMC iterations. Interestingly, although none of the priors used in these experiments explicitly promoted sparsity, the samplers are all able to discover sparse graphs, with an average number of edges close to the true value of 16. We based our inferences on a single, long run of $T = 100000$ iterations for each sampler, with 5000 samples discarded as "burn-in" in each case. For diagnostic purposes, we also performed several short ($T = 20000$) runs using each sampler. Figure 6 shows profiles obtained from these diagnostic runs; in each case the monitored quantities converged within a few thousand iterations.

Taking advantage of our knowledge of the correct graph, we also computed, for each sampler, the average distance, across all samples drawn, from the data-generating graph. We used the *squared Frobenius norm* $\|\cdot\|_F^2$ to quantify distance from true graph G^*. This is given by:

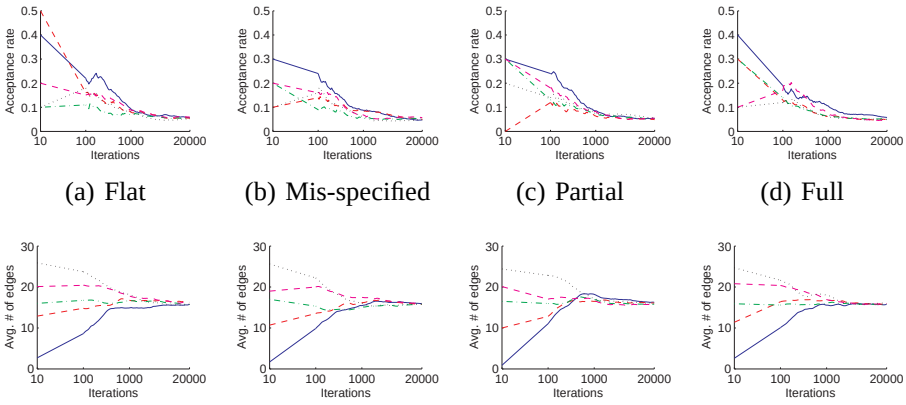
$$\|\mathbf{G}^{(t)} - \mathbf{G}^*\|_F^2 \quad = \quad \sum_{i=1}^{p} \sum_{j=1}^{p} |(\mathbf{G}^{(t)} - \mathbf{G}^*)_{ij}|^2,$$

where $\mathbf{G}^{(t)}$ and $\mathbf{G}^*$ are adjacency matrices corresponding to graphs $G^{(t)}$ and G^* respectively, and $(\cdot)_{ij}$ denotes the $(i, j)^{th}$ element of its (matrix) argument. For the flat, mis-specified, partial and full priors the average distances were 16, 9.23, 8.64 and 6.91 respectively. This accords with the ROC results presented above,



(a) Acceptance rates (b) Average number of edges

Figure 5: Acceptance rate and number of edges, synthetic data. (a) Average acceptance rate and (b) average number of edges for synthetic data, plotted against number of sampling iterations.



(a) Flat (b) Mis-specified (c) Partial (d) Full

(e) Flat (f) Mis-specified (g) Partial (h) Full

Figure 6: Diagnostic runs for synthetic data. Five short ($T = 20000$) runs were performed using each sampler: (a)-(d) show acceptance rates, (e)-(h) show number of edges plotted against number of sampling iterations for flat, mis-specified, partial and full priors.

Table 2: Posterior odds in favour of the path ERBB2→SHC→RAS (correct) over ERBB2→GAP→RAS (incorrect).

Prior type	**Posterior odds**
Flat	32
Mis-specified	213
Partial	748
Full	366

Table 3: Posterior probabilities for path RAF→MEK→ERK.

Prior type	**Posterior probability**
Flat	0.79
Mis-specified	0.74
Partial	0.95
Full	0.95

and suggests that the priors are indeed capable of guiding sampling towards good regions of graph space.

The samples $G^{(1)} \dots G^{(T)}$ obtained using each sampler can be used to compute posterior probabilities or odds for more-or-less arbitrary graph features. Here, we present two examples of this type of analysis. The protein Ras is the "entry-point" to the ERK pathway; ERBB2 is an important up-stream influence on this pathway. Suppose we wished to ask whether ERBB2 activates Ras via SHC or GAP. A natural way to capture the evidence in favour of these two specific hypotheses is via the posterior odds in favour of the path ERBB2→SHC→RAS over the path ERBB2→GAP→RAS. Table 2 shows these odds ratios for each of the four samplers. What is interesting is that even though none of the prior information used specifically concerns the paths considered, the locally-informative priors give stronger odds in favour of the correct path than the flat prior. In fact, the strongest odds results from the partial prior which does not even contain information on classes of edges. This highlights the role priors can play in constraining inference to the benefit of questions concerning graph features generally, and not just those questions concerning features captured in the prior.

A second example is shown in Table 3 and concerns posterior probabilities for the path RAF→MEK→ERK. The full and partial priors give the highest probabilities for this correct feature: this is not surprising in light of the fact that they contain information, from (S2), in favour of this path. What is interesting is that the mis-specified prior finds the complete path in 74% of its samples, despite the fact that it contains information which explicitly *penalizes* graphs containing this

very path.

4.2 Proteomic data

The *Mitogen-Activated Protein Kinase* or *MAPK pathway* is a a biochemical pathway which plays a central role in cellular signaling and whose aberrant functioning is heavily implicated in a number of cancers. Despite many years of intensive research, cancer-specific features of the MAPK pathway remain poorly characterized, and relatively little is known regarding connectivity specific to certain important subtypes of proteins called *phospho-forms* and *isoforms*. In this Section we present some of the results obtained in an analysis of this system using the structural inference methods developed here. The aims of the present paper are primarily methodological; we therefore defer a full discussion of the biological details and experimental implications of our work to a forthcoming paper. We note that our investigation into this system started out as a simpler correlational analysis. The complex nature of relationships between components in cancer pathways motivated us to move towards a multivariate approach, while the need to take account of rich but uncertain biochemical knowledge and to sharpen questions concerning specific features of the pathway motivated the work we have described here.

4.2.1 Data

Proteomic data were obtained for the 14 protein phospho-forms and isoforms shown in Table 4; these included two isoforms of the protein *c-Raf*; four phosphoforms of *MAPK/ERK Kinase* or *MEK*; two isoforms of *Extracellular Regulated Kinase* or *ERK*; four isoforms of *Protein Kinase C* or *PKC*, and two phosphoforms of *Akt*. The data were obtained from an assay performed by Kinexus Inc. (Vancouver, Canada) on a panel of 18 breast cancer cell lines. The data were pre-processed by (i) setting all zeros to 1/100 of the smallest non-zero value, (ii) taking logs and (iii) discretizing around the median for each protein. This gave rise to binary data for each of the 14 proteins.

4.2.2 Priors

Our prior beliefs concerning the network can be summarized as follows:

(P1) The Rafs are expected to have edges going only to isoforms and phosphoforms of MEK; the MEKs are expected to have edges only to ERKs; the AKTs are expected to have edges only to Rafs.

(P2) There is expected to be at least one edge going from Rafs to MEKs and from MEKs to ERKs.

Table 4: Some protein isoforms and phosho-forms from the Mitogen-Activated Protein Kinase or MAPK pathway.

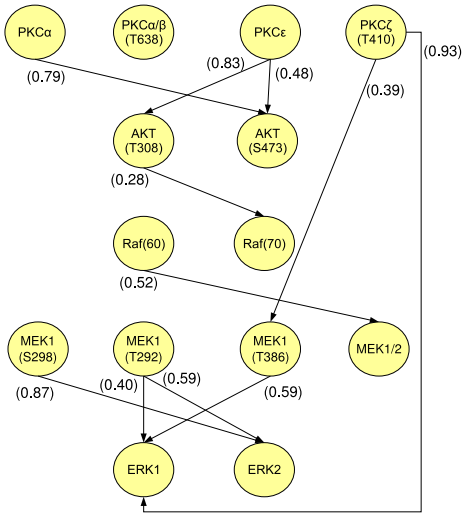
Protein	**Isoforms/phosho-forms**	**Notes**
Raf	Raf(60) Raf(70)	These are isoforms of c-Raf
MEK	MEK1(S298) MEK1(T292) MEK1(T386) MEK1/2	MEK stands for **M**APK/**E**RK **K**inase; also known as MAPK Kinase
ERK	ERK1 ERK2	ERK is **E**xtracellular **R**egulated **K**inase; also known as MAPK
AKT	AKT (T308) Akt (S473)	Also known as Protein Kinase B
PKC	PKCα PKCα/β(T638) PKCϵ PKCζ (T410)	PKC stands for **P**rotein **K**inase **C**

(P3) We do not expect the in-degree of any node to exceed 3.

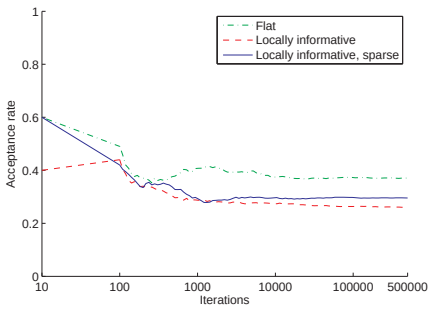
We constructed locally-informative priors for these beliefs: using (15), (P1) defined a negative edge set E_- for concordance function (13); (P2) and (P3) were captured using concordance functions (16) and (19) respectively. The three functions were combined multiplicatively, using (20), with all λ_i's set to 100. A prior-based proposal was used, using (22) with λ_Q set to 4. This is an example of using a relatively strong prior to refine a scientific question: in this case we are primarily interested in patterns of connectivity between subtypes of proteins Raf, MEK and ERK, but remain interested in other possibilities also.

4.2.3 Results

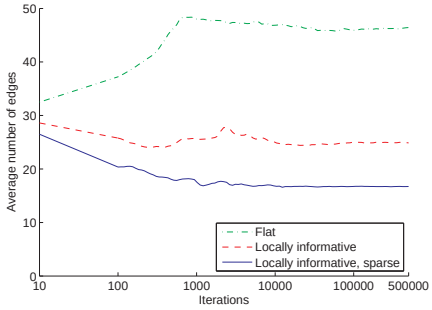
Figure 7(a) shows the single most probable graph encountered during sampling, using a locally-informative prior for (P1)-(P3). Each edge e is annotated with the corresponding posterior probability $P(e \mid \mathbf{X})$. Figure 7(b) shows average acceptance probabilities, while Figure 7(c) shows average number of edges plotted against number of iterations. Here, "locally-informative" refers to a locally-informative prior which does not explicitly promote sparsity, while "locally-informative, sparse" refers to a locally-informative prior which additionally promotes sparsity. That is, the first prior encodes (P1) and (P2) but not (P3), while the second encodes (P1),



(a) Single most probable graph

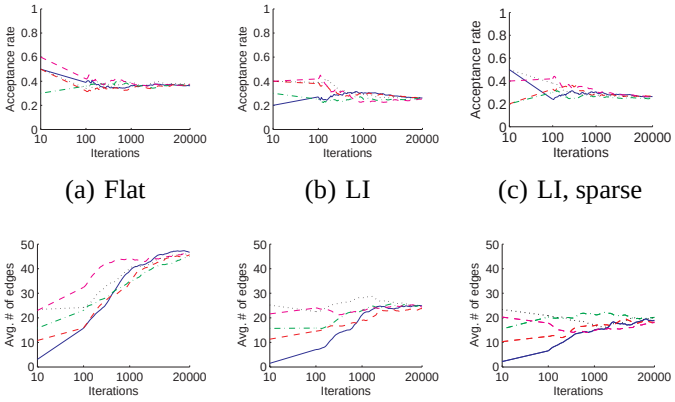


(b) Acceptance probability



(c) Number of edges

Figure 7: Results from protein data. (a) The single most probable graph sampled using a sparse, locally-informative prior, each edge is annotated with its posterior probability; (b) average acceptance rate and (c) average number of edges plotted against sampling iterations.



(a) Flat (b) LI (c) LI, sparse

(d) Flat (e) LI (f) LI, sparse

Figure 8: Diagnostic runs for protein data. Five short ($T = 20000$) runs were performed using each sampler: (a)-(c) show average acceptance rates, (d)-(f) show average number of edges plotted against number of sampling iterations, for flat, locally-informative (LI) and sparse, locally-informative (LI,sparse) priors.

(P2) and (P3). We note that in contrast to the simulation experiments, in this case the flat prior is unable to control model complexity, with the average number of edges converging to a relatively high value. Also, there is a difference in the level of sparsity obtained by the two locally-informative priors, with the sparse locally-informative prior sampling models which are noticeably more parsimonious. We used single, long runs of $T = 500000$ iterations in each case, with 5000 samples discarded as "burn-in". For diagnostic purposes, we also performed several short ($T = 20000$) runs using each sampler; these are shown in Figure 8.

5 Discussion

In this paper, we have discussed the use of rich structural priors for model averaging in Bayesian networks. In our view, such priors play two related roles. Firstly, they provide a means by which to capture valuable domain knowledge regarding graph features. Secondly, they allow us to refine or sharpen questions of interest, in effect playing a role analogous to formulating an initial set of hypotheses, but with much greater flexibility. This flexibility, combined with the well-known robustness of Bayesian model averaging, means that it is possible to obtain useful results even when priors are mis-specified, essentially by borrowing strength from a large space of models, many of which accord only partially with prior beliefs.

We saw also that the use of structural priors can lead to very substantive gains at small sample sizes. Much of the literature on MCMC-based structural inference has focused on moderate-to-large sample sizes: for example, in a recent paper, Giudici and Castelo (2003) analyzed a dataset with $p = 6$ and $n = 1846$. In contrast, our experiments focused on quite challenging settings in which there are both a greater number of variables and far fewer observations. Although, unsurprisingly, we found that the basic sampling approach does not do well in this setting, we discovered that reasonably well-specified priors do indeed permit effective inference under these conditions, yielding substantive gains even when the features of eventual interest were not described in the prior, or when the priors were partially mis-specified.

Bayesian formulations can, in general, be viewed as a form of penalized likelihood, and in that sense they quite naturally promote parsimonious models. An additional penalty on model complexity in the form of a sparsity prior may therefore be unnecessary in many cases. However, when a paucity of data, or a mis-specified model, exacerbate problems of over-fitting, explicit sparsity priors can play a useful role. Indeed, we saw that for the small-sample protein data, a sparsity-promoting prior had a noticeable effect on controlling model complexity. In such settings, we recommend examining the average number of edges in sampled graphs to decide whether or not such priors are called for.

We note that our structural priors do *not* satisfy so-called prior equivalence in that they allow us to express a prior preference for one graph over another even when both graphs imply the same conditional independence statements. Thus, we may express a prior preference for $A \rightarrow B$ over $B \rightarrow A$, despite the fact that both graphs describe the same likelihood model. This property allows us to express preferences derived from domain knowledge. For example, if we believe that A precedes B in time, or that A is capable of physically influencing B, we may express a preference for $A \rightarrow B$ over $B \rightarrow A$. Viewed in this way, the ability of non-equivalent priors to incorporate outside information into inference is a useful, even desirable, property.

In a recent paper, Friedman and Koller (2003) proposed an interesting approach to structural inference, in which samples are drawn from the space of *orders*, where an order $\prec$ is defined as a total order relation on vertices such that if $X_i \in \mathbf{Pa}_G(X_j)$ then $i \prec j$. The appeal of this approach lies in the fact that the space of orders is much smaller than the space of graphs. On the other hand, the use of order space means that structural priors must be translated into priors on orders, and inferences on graph features must be carried out via order space. This turns out to place restrictions on the kinds of structural priors which can be utilized, and moreover makes it difficult to compute the posterior probabilities of arbitrary graph features. Furthermore, the authors own experiments show that sampling in

order space offers no advantage at smaller sample sizes. In contrast, we find that remaining in graph space offers real advantages in terms of being able to specify rich priors in a natural and readily interpretable fashion and, just as important, in making inferences regarding essentially arbitrary features of graphs. Also, as we have seen, the use of such priors can lead, in turn, to much improved performance at small sample sizes.

There remains much to be done in extending the methods presented in this paper to higher-dimensional problems. One approach to making such problems tractable would be to place strong priors on some parts of the overall graph. This would, in effect, amount to using background knowledge to focus limited inferential power on the least well-understood, or scientifically most interesting, parts of the graph.

Our current applied efforts are directed towards questions in cancer biology. We have found the ability to specify rich, interpretable priors directly on graphs and make posterior inferences on features of graphs such as edges, groups of edges and paths to be valuable in casting biologically interesting questions within a statistical framework. We therefore hope that the methods presented here will prove useful in a number of settings where questions of this kind need to be addressed.

Acknowledgements: The authors would like to thank Rich Neve, Paul Spellman, Laura Heiser and other members of Joe Gray's laboratory at Lawrence Berkeley National Laboratory for a productive, ongoing collaboration and for providing the proteomic dataset used in this paper. SM was supported by a Fulbright-AstraZeneca postdoctoral fellowship.

References

W. Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60, 1991.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

P. Dellaportas and J. J. Forster. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615, 1999.

P. Dellaportas and C. Tarantola. Model determination for categorical data with factor level merging. *Journal of the Royal Statistical Society, Series B*, 67(2): 269–283, 2005.

N. Friedman and D. Koller. Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, 50 (1):95–125, 2003.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.

P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo Model Search for Data Mining. *Machine Learning*, 50(1):127–158, 2003.

P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97, 1970.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3): 197–243, 1995.

T. Ideker and D. Lauffenburger. Building with a scaffold: emerging strategies for high-to low-level cellular modeling. *Trends in Biotechnology*, 21(6):255–262, 2003.

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, 20(4):388–400, 2005.

M. I. Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

H. Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, 2002.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

D. Madigan, J. York, and D. Allard. Bayesian Graphical Models for Discrete Data. *International Statistical Review/Revue Internationale de Statistique*, 63(2):215–232, 1995.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

R. W. Robinson. Counting labeled acyclic digraphs. In F. Harary, editor, *New Directions in Graph Theory*. Academic Press, 1973.

C. Tarantola. MCMC model determination for discrete graphical models. *Statistical Modelling*, 4(1):39, 2004.

L. Tierney. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22(4):1701–1762, 1994.