SUBGRAPH APPROXIMATIONS FOR LARGE DIRECTED GRAPHICAL MODELS

Constantin T. Yiannoutsos
Alan E. Gelfand

TECHNICAL REPORT No. 473
SEPTEMBER 27, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

(12)

SUBGRAPH APPROXIMATIONS FOR LARGE DIRECTED GRAPHICAL MODELS

Constantin T. Yiannoutsos
Alan E. Gelfand

TECHNICAL REPORT No. 473
SEPTEMBER 27, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

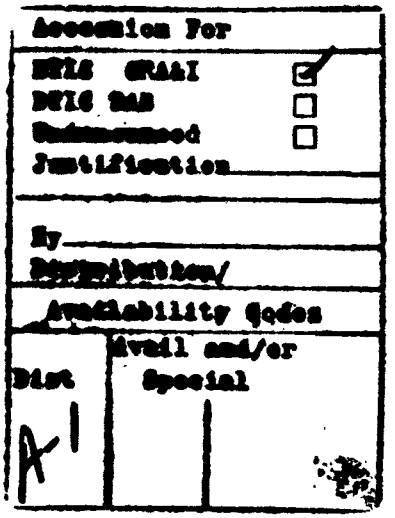
Approved for public release; distribution unlimited

Subgraph Approximations for Large Directed Graphical Models

Constantin T. Yiannoutsos

and

Alan E. Gelfand*



ABSTRACT

Graphical Models provide a powerful tool for the formulation of general statistical models. In a previous paper, (Yiannoutsos & Gelfand, 1991), the authors argued that sampling-based techniques provide a unified approach for the analysis of graphical models under general distributional specifications. These techniques include both noniterative and iterative Monte Carlo.

Our concern here is with very large graphical models whose size and complexity may prohibit analysis within a reasonable time frame. Typically in large systems however, interest focuses on the behavior of only a few critical nodes. Our proposal is to develop, for a particular node, an approximating subgraph which contains virtually as much information about the variable as the full network, but by virtue of its reduced size, enables rapid computational investigation. We provide an illustration using a 40-node graph. Though this is not as large as we would envision in practice, it is convenient in permitting full model calculations to enable assessment of our approximations.

KEY WORDS

Conditional independence, Gibbs sampler, Kullback-Leibler distance, L^1 distance, likelihood weighting, Monte Carlo, Propagation of Information.

*Constantin T. Yiannoutsos has just completed his Ph.D. in the Dep't of Statistics, University of Connecticut, Storrs, CT 06269. Alan E. Gelfand is professor of Statistics in the same department. Gelfand's research was supported in part by NSF grant DMS 8918563.

ending node). Cycles lead to logical implausibilities (Pearl, 1986). Probabilistically, the joint distribution of the variables is not uniquely defined.

One of the fundamental relations expressible by a DAG is that of precedence. Parental nodes precede their children, in a causal or temporal sense, inducing an ordering of the nodes in the graph. We enumerate the nodes in a top–down fashion, starting with those without parents (source nodes) at the first level, proceeding to their children at the next level, and so on to the bottom of the graph. Nodes allocated to the same level can be ordered arbitrarily permitting many equivalent enumerations. We also define the parental set of i as pa(i), and the set of children of i as ch(i) and denote the variable at node i by X_i.

Recall that two (possibly vector–valued) random variables X_1 and X_2 are conditionally independent given a third random variable X_3, denoted by $X_1 \perp\!\!\!\perp X_2 | X_3$, iff $f(x_1,x_2|x_3) = f(x_1|x_3)\ f(x_2|x_3)$, or, equivalently, iff, $f(x_1|x_2,x_3) = f(x_1|x_3)$ and $f(x_2|x_1,x_3) = f(x_2|x_3)$. Implicit in this definition is that use of conditional independence refines and qualifies general dependence relations. Conditional independence arises naturally for a DAG in terms of the set of predecessors. That is, if nodes i and j are not connected and i≺j (i precedes j), then

$$j \perp\!\!\!\perp i | pr(j-i) \iff j \perp\!\!\!\perp i | pa(j)$$

where pr(j–i) is the set of predecessors of j excluding i. This is equivalent to saying that *j* is conditionally independent of i given its parents.

As a result, the following factorization of the joint density of the random variables at the nodes in V ensues (Whittaker, 1990, p. 73). If n is the number of nodes in V,

$$f(x_1,\ldots,x_n) = \prod_{v=1}^{n} f(x_v | pa(v)) \qquad (1)$$

2. Subgraph Approximations

Model reduction is hardly a novel statistical idea since parsimony usually facilitates interpretation. In our case, however, the goal is to produce, for a particular variable, an

approximating subset which contains essentially as much information about the variable as the full model, but by virtue of its reduced size, enables rapid computational investigation.

Proximity between two variables in a DAG plays an important role. We argue that in such hierarchical structures, for a wide class of measures, the information in X_j about X_i decreases monotonically as j moves farther away from i. For a one dimensional chain we show that, in studying X_i, one can truncate the chain a certain number of steps away from i and obtain close approximations to exact marginal and conditional distributions for X_i. In a (two–dimensional) graphical structure, of course, there will be many chains, emanating from a particular node X_i suggesting the concept of a radius. The radius around a variable X_i contains all those variables at a constant number of edges away from X_i (the length of the radius). For instance, a subset of zero radius is the node itself, a subset of radius one is ch(i)∪pa(i)∪{i}, and so on until the complete graph is reacquired. Figure 2 provides an illustration the concept of the radius around a node. In this 20–node network, node 15 is our target (zero radius). Next to every other node, the distance from 15 has been entered.

[Insert figure 2 about here]

In the next three subsections, to support the use of subgraph approximations, we present theoretical evidence asserting the degeneration of information transmitted from nodes increasingly distant from the node of interest.

2.1 Information and Divergence

We recall the *Kullback–Leibler information divergence* between two measurable functions f(x) and g(x) with respect to an absolutely continuous measure ν, defined as

$$I_x(f;g) = E_f \log \frac{f(X)}{g(X)} = \int \log \frac{f}{g} f d\nu$$

Information divergence has the following properties (Whittaker, 1990, p. 94–99).

(i) The divergence is *well defined*. That is, the integral in $I_x(f;g)$ exists, even though it might be ∞.

(ii) The divergence is *additive* over independent variables. Suppose that X, Y are independent random variables. Then

$$I_{xy}(f;g) = I_x(f_x;g_x) + I_y(f_y;g_y)$$

(iii) The information divergence is *positive definite*, that is $I_x(f;g) \geq 0$, with equality holding if and only if $g(x)=f(x)$, except possibly on a set of measure zero.

(iv) The information divergence is *not symmetric*. That is, $I_x(f;g) \neq I_x(g;f)$. A symmetric version is

$$J_x(f;g) = I_x(f;g) + I_x(g;f)$$

which we call the *mean information divergence* between functions f and g (Kullback & Leibler, 1951).

A particular information divergence is the *information proper*, i.e. the information divergence testing conditional independence defined by

$$\begin{aligned} \mathrm{Inf}(X_1 \perp\!\!\!\perp X_2) &= I_x(f_{12};f_1f_2) \\ &= I_x(f_{1|2};f_1) = \int log \frac{f(x_1|x_2)}{f(x_1)} f(x_1,x_2)dx_1dx_2 \end{aligned}$$

The information proper is symmetric in X_1 and X_2, and achieves its minimum when X_1 and X_2 are independent.

We now investigate the behavior of the information proper in hierarchical structures. Consider the hierarchical chain $X_n \ldots \rightarrow X_j \rightarrow \ldots \rightarrow X_i \rightarrow \ldots \rightarrow X_1$. The information contained in a variable about others on the chain is not constant. Intuitively, it should be larger for variables closer to it than it is for variables farther away. We now prove that this is so.

Lemma 1. Consider the hierarchical sequence $X_3 \to X_2 \to X_1$. For convex functions r on $\mathbb{R}^+$ we have

$$\int r\left(\frac{f(x_1|x_3)}{f(x_1)}\right) f(x_1)dx_1 \leq \int r\left(\frac{f(x_2|x_3)}{f(x_2)}\right) f(x_2)dx_2$$

Proof (DeGroot and Goel, 1986). Clearly,

$$\frac{f(x_1|x_2)}{f(x_1)} = \frac{f(x_2|x_1)}{f(x_2)}$$

Therefore,

$$\frac{f(x_1|x_3)}{f(x_1)} = \int \frac{f(x_1|x_2)f(x_2|x_3)}{f(x_1)}\,dx_2 = \int \frac{f(x_2|x_1)f(x_2|x_3)}{f(x_2)}\,dx_2$$

Since r is defined on $\mathbb{R}^+$ we have

$$\int r\left(\frac{f(x_1|x_3)}{f(x_1)}\right) f(x_1)dx_1 = \int r\left(\int \frac{f(x_2|x_3)}{f(x_2)} f(x_2|x_1)dx_2\right) f(x_1)dx_1$$

$$\leq \int \left(\int r\left(\frac{f(x_2|x_3)}{f(x_2)}\right) f(x_2|x_1)dx_2\right) f(x_1)dx_1$$

by Jensen's inequality. Switching the order of integration between X_2 and X_1 proves the lemma.

Corollary 1. In a hierarchical sequence, $X_n \to \ldots \to X_j \to \ldots \to X_i \to \ldots X_1$, we have, for a convex function r, defined on $\mathbb{R}^+$

$$\int r\left(\frac{f(x_1|x_j)}{f(x_1)}\right) f(x_1)dx_1 \leq \int r\left(\frac{f(x_i|x_j)}{f(x_i)}\right) f(x_i)dx_i$$

Proof. Realising that $f(x_1|x_i) = \int f(x_1|x_2)\ldots f(x_{i-1}|x_i)dx_2\ldots dx_{i-1}$ and that $X_2,\ldots,X_{i-1}$ do not enter in the relation above, the corollary follows immediately.

Hence we have

Theorem 1. In a hierarchical chain with $X_n \to X_{n-1} \to \ldots \to X_j \to \ldots \to X_i \to \ldots \to X_1$

$$\mathrm{Inf}(X_1 \| X_j) \leq \mathrm{Inf}(X_i \| X_j)$$

Proof. Since $\frac{f(.|x_i)}{f(.)} \log\left[\frac{f(.|x_i)}{f(.)}\right]$ is of the form $t \log t$, a convex function, from corollary 1 we obtain,

$$\int f(x_1|x_j) \log\left[\frac{f(x_1|x_i)}{f(x_1)}\right] dx_1 \leq \int f(x_i|x_j) \log\left[\frac{f(x_i|x_i)}{f(x_i)}\right] dx_i$$

Integrating both sides with respect to $f(x_j)$ produces the desired result.

We interpret theorem 1 as follows: X_j transmits less information to X_1 than to X_i, i.e. less to a further node than to a closer one. Whittaker (1990, p. 109) captures this in a slightly different way.

Corollary 2. In a hierarchical chain we have,

$$\mathrm{Inf}(X_1 \perp\!\!\!\perp X_2) = \mathrm{Inf}(X_1 \perp\!\!\!\perp X_2|X_3) + \mathrm{Inf}(X_1 \perp\!\!\!\perp X_3)$$

Proof. From conditional independence, it is clear, that $f_{123} = f_{1|2}f_{2|3}f_3$. From the definition of information proper, the desired relation is equivalent to,

$$\mathrm{Inf}(X_1 \perp\!\!\!\perp X_2|X_3) + \mathrm{Inf}(X_1 \perp\!\!\!\perp X_3) = E_f \log\left(\frac{f_{123}}{f_{1|3}f_{2|3}f_3}\right) + E_f \log\left(\frac{f_{13}}{f_1 f_3}\right)$$

$$= E_f \log\left(\frac{f_{1|2}f_{2|3}f_3}{f_{1|3}f_{2|3}f_3}\frac{f_{13}}{f_1 f_3}\right) = E_f \log\left(\frac{f_{1|2}}{f_1}\right) = \mathrm{Inf}(X_1 \perp\!\!\!\perp X_2)$$

The interpretation is that $\mathrm{Inf}(X_1 \perp\!\!\!\perp X_3) \leq \mathrm{Inf}(X_1 \perp\!\!\!\perp X_2)$, by the quantity $\mathrm{Inf}(X_3 \perp\!\!\!\perp X_1|X_2)$, which is, in a sense, the "degeneration" of information from X_3 to X_1 through its parent X_2. The implication is encouragement for the subgraph approximation.

2.2 L^1 Distance Between Distributions

Suppose, as an alternative to the information divergence between two densities, we consider the L^1 distance,

$$\|f-g\| = \int |f-g| \, d\mu \qquad (3.2)$$

Consider again the Markov sequence $X_n \rightarrow X_{n-1} \rightarrow, .. \rightarrow X_1$. The claim is that, in calculating marginal and conditional distributions for a particular X_i, it is unnecessary to

retreat back to X_n in order to recover most of the pertinent information about X_i. Increasingly distant predecessors and descendants should have less influence than closer ones. So suppose that the chain is truncated a number of steps away from X_i. For the resulting source variables, we replace their exact marginal density by an approximation. The following theorem shows that in doing so, we necessarily better approximate exact distributions in the L^1 sense as we proceed deeper into the subnetwork.

Theorem 2. Suppose that X_i and X_{i+1} are neighboring nodes in a Markov sequence, and that all marginal and conditional distributions are defined in L^1. Then

$$\|f(X_i)-f'(X_i)\| \le \|f(X_{i+1})-f'(X_{i+1})\|$$

where

$$f(X_i) = \int f(X_i|x_{i+1})f(x_{i+1})dx_{i+1},$$

and

$$f'(X_i)=\int f(X_i|x_{i+1}])f'(x_{i+1})dx_{i+1}$$

Proof.

$$\begin{aligned}
\|f(X_i)-f'(X_i)\| &= \int \left| f(x_i) - f''(x_i) \right| dx_i \\
&= \int \left| \int f(x_i|x_{i+1})f(x_{i+1})dx_{i+1} - \int f(x_i|x_{i+1})f'(x_{i+1})dx_{i+1} \right| dx_i \\
&= \int \left| \int f(x_i|x_{i+1}) \left(f(x_{i+1}) - f'(x_{i+1}) \right) dx_{i+1} \right| dx_i \\
&= \int \left| \int \left\{ f(x_i|x_{i+1}) \left(f(x_{i+1}) - f'(x_{i+1}) \right) \right\}^+ dx_{i+1} \right. \\
&\qquad \left. - \int \left\{ f(x_i|x_{i+1}) \left(f(x_{i+1}) - f'(x_{i+1}) \right) \right\}^- dx_{i+1} \right| dx_i \\
&\le \int \left| \int \left\{ f(x_i|x_{i+1}) \left(f(x_{i+1}) - f'(x_{i+1}) \right) \right\}^+ dx_{i+1} \right| dx_i \\
&\qquad + \int \left| \int \left\{ f(x_i|x_{i+1}) \left(f(x_{i+1}) - f'(x_{i+1}) \right) \right\}^- dx_{i+1} \right| dx_i
\end{aligned}$$

by the definition of the positive and negative part functions. Since all terms are nonnegative, the absolute values can be removed, and, upon reversing the order of integration,

$$\|f(X_i)-f'(X_i)\| \le \int \left(f(x_{i+1})-f'(x_{i+1})\right)^{+} dx_{i+1} + \int \left(f(x_{i+1})-f'(x_{i+1})\right)^{-} dx_{i+1}$$
$$= \int \left|f(x_{i+1}) - f'(x_{i+1})\right| dx_{i+1} = \|f(X_{i+1}) - f'(X_{i+1})\|$$

This theorem shows that whatever density approximations are used for the outer edge of a subnetwork, the approximations improve as we move further from the edge, i.e. as we do more processing incorporating the correct conditional densities. We can again interpret this as supporting subgraph approximations.

2.3. An Importance Sampling Argument

Retaining the same notation as in the previous subsections we now argue that

$$\text{var}\,\frac{f(X_i)}{f'(X_i)} \le \text{var}\,\frac{f(X_{i+1})}{f'(X_{i+1})} \qquad (2)$$

where, in both cases, expectations are taken with respect to f', i.e., $X_i \sim f'(X_i)$, on the left–hand side, $X_{i+i} \sim f'(X_{i+1})$ on the right–hand side. We interpret (2) from a Monte Carlo perspective. On average, $f'(X_i)$ is a better importance sampling density for (better matches) $f(X_i)$, than $f'(X_{i+1})$ is for $f(X_{i+1})$. Approximation improves as we move further from the edge of the subnetwork again encouraging subgraph approximations.

Lemma 2. If X_i and X_{i+1} are neighboring nodes in a Markov sequence, then

$$\frac{f(X_i)}{f'(X_i)} = E'\left[\frac{f(X_{i+1})}{f'(X_{i+1})}\,|X_i\right]$$

where the expectation E' is taken with respect to the conditional distribution $f'(X_{i+1}|X_i)$.

Proof. By definition, $f'(X_{i+1}|X_i) = \dfrac{f(X_i|X_{i+1})f'(X_{i+1})}{f'(X_i)}$. But then,

$$E'\left[\frac{f(X_{i+1})}{f'(X_{i+1})}\,|X_i\right] = \int \frac{f(x_{i+1})}{f'(x_{i+1})} f'(x_{i+1}|X_i)dx_{i+1}$$
$$= \int \frac{f(X_i|x_{i+1})f(x_{i+1})}{f'(X_i)} dx_{i+1} = \frac{f(X_i)}{f'(X_i)}.$$

Theorem 3. If X_i and X_{i+1} are neighboring nodes in a Markov sequence, then (2) holds.

Proof. The result is immediate from the Rao–Blackwell theorem.

3. Simulation approaches

Calculation of desired marginal and conditional distributions in a general mixed directed graphical model requires high dimensional integration/summation. Such calculations usually can not be carried out analytically, either exactly or approximately. Simulation methods offer a viable alternative. In particular, suppose the conditional information fixes a subset of the variables (possibly an empty set if marginalization is sought), X_0, of size n_0 to specified levels. The nodes in X_0 are called *evidence* nodes in Shachter and Peot (1989). Let X_u denote the complement of X_0, i.e. the *unobserved* nodes. We seek the conditional distribution of X_u given $X_0=x_0$ as well as that of X' given $X_0=x_0$, where X' denotes a generic component of X_u. Exact calculation of $f(X_u|X_0 = x_0)$ requires an $n - n_0$ dimensional integration/summation and calculation of $f(X'|X_0 = x_0)$ requires an additional integration or summation of dimension n_0-1. Envisioning n and possibly n_0 to be large, such computation will not be feasible by exact or approximate analytic methods. Hence, we turn to simulation strategies. Such approaches, implemented in conjunction with subgraph approximations, enable rapid calculation of $f(x'|X_0=x_0)$. In subsections 3.1 and 3.2 we briefly describe two simulation approaches. They are discussed for a general DAG which we envision as a subgraph. The Gibbs sampling approach is implemented for the example in section 4.

As observed in Smith and Gelfand (1992), there is an essential duality between a sample and the density (distribution) from which it is generated. Clearly the density generates the sample. But conversely, given a sample we can approximately recreate the density and its features. Thus our objective is to draw samples from $f(X_u|X_0 = x_0)$. Drawing observations from the joint distribution of X is straightforward. It may be done

in a "top down" fashion using the components of the factorization (1). That is, we sample all source nodes, then sample all their children, etc. The directed graphical model reveals which sampling orders are thus *feasible* and we shall in fact assume that the labeling of the X's from X_1 to X_n constitutes a feasible sampling order.

In attempting to sample $f(X_u | X_o = x_o)$ one might take draws from f(X) and retain those meeting the evidence $X_o = x_o$. Such rejection sampling is called logic sampling (see, e.g., Henrion 1988), and is very inefficient when the nodes in X_o are discrete but the event $X_o = x_o$ is rare; it is impossible if any of the nodes in X_o are continuous.

Note that, though we can not obtain $f(X_u | X_o = x_o)$ explicitly, we do know its form modulo a normalizing constant, i.e.,

$$f(X_u | X_o = x_o) \propto f(X_u, x_o) \tag{3}$$

where the right hand side of (3) is given in (1). In fact we can write

$$f(X_u | X_o = x_o) \propto \prod_{X_i \in X_o} f(X_i | pa(X_i)) \prod_{X_i \in X_u} f(X_i | pa(X_i)) \Big|_{(X_u, x_o)} \tag{4}$$

Suppose, as is traditional, that we refer to what is observed as "data" and what is unobserved as "parameters". Then the first product on the right side of (4) could be considered as a *likelihood* since here all the X_i are observed while the second product could be considered as a *prior* since here none of the X_i are observed. Then (4) is of the form Likelihood x Prior as in customary Bayesian modeling and we may bring to bear on our problem all of the recently discussed sampling–based technology for Bayesian calculations. This work includes noniterative Monte Carlo using importance sampling as summarized in Geweke (1989) as well as iterative or Markov chain Monte Carlo as described for Bayesian calculations in Gelfand and Smith (1990).

3.1 Independent or Noniterative Monte Carlo

Since we cannot sample from $f(X_u | X_o = x_o)$ directly, we employ a suitable importance sampling density (ISD). More precisely from a density say $g(X_u)$ we draw a

large sample denoted by X_{ut}, t=1,...,m, where $g(X_u)$ has the same support as $f(X_u|X_o=x_o)$. Expectations under $f(X_u|X_o=x_o)$, e.g. $E[h(X_u)|X_o=x_o]$, are approximated by the weighted sum

$$\hat{h}_m = \frac{\sum_{t=1}^{m} h(X_{ut})\cdot w_t}{\sum_{t=1}^{m} w_t} \qquad (5)$$

where $w_t = f(X_{ut}|X_o = x_o)/g(X_{ut})$. Moreover, resampling the X_{ut} with probabilities $g_t = w_t/\Sigma w_t$ produces samples approximately distributed according to $f(X_u|X_o=x_o)$ (see Rubin, 1988, Smith and Gelfand, 1992).

The selection of g becomes the primary concern. The more closely g resembles $f(X_u,x_o)$ the more efficient g is in terms of sample size m. Hence a good ISD is characterized as having fairly constant weights w_t. Such stability might be measured through the variance of the w_t (Geweke, 1989) or their entropy (Ferguson, 1983). A natural choice for g would be the prior $\prod_{X_i \in X_u} f(X_i|pa(X_i))\Big|_{(X_u,x_o)}$ which, provided that it is proper, it can be readily sampled using a feasible sampling order. This choice of g results in weights $w_t = \prod_{X_i \in X_o} f(X_i|pa(X_i))\Big|_{(X_{ut},x_o)}$ which would naturally be called likelihood weights (see Shachter and Peot, 1989), i.e., bigger weights are attached to "more likely" X_{ut}'s. Such w_t are not likely to be very stable since we would not expect the prior to "match" the Likelihood x Prior form.

A more refined selection of g can be obtained as follows. Partition X_u into a set of discrete and a set of continuous variables denoted by X_u^d and X_u^c respectively. We consider an ISD which samples equally likely over the domain of X_u^d and then, given $X_u^d = x_u^d$, draws $X_u^c \sim g(X_u^c|X_u^d = x_u^d)$. The joint density of X_u under this ISD is proportional to the conditional density $g(X_u^c|X_u^d)$. Thus to match $f(X_u, x_o)$, for each $X_u^d = x_u^d$ we would choose $g(X_u^c|x_u^d)$ to match $f(X_u^c, x_u^d, x_o)$. Methods for developing an efficient ISD for a nonstandardized continuous density have received much attention lately (see e.g. Geweke,

1989, Oh & Berger, 1991, West, 1991). In most of this work the resultant ISD is a mixture of normal or t distributions.

3.2 Dependent or Iterative Monte Carlo

Markov chain Monte Carlo in the form of the Gibbs sampler was applied by Pearl (1987), to causal models involving binary variables. Yiannoutsos and Gelfand (1991) generalize this approach to arbitrary directed graphical models. We briefly summarize their discussion.

Implementation of the Gibbs sampler to draw observations from $f(\mathbf{X}_u|\mathbf{X}_0=\mathbf{x}_0)$ proceeds by making draws from the univariate complete conditional distributions $f(X'|\mathbf{X}'_u,\mathbf{x}_0)$. Given a starting state vector for $\mathbf{X}_u$, the components of $\mathbf{X}_u$ are typically sampled in the natural order, sometimes referred to as a raster scan, with the conditional levels of $\mathbf{X}'_u$ updated after each sampling while $\mathbf{X}_0$ remains "clamped" at $\mathbf{x}_0$. One pass through all of the components of $\mathbf{X}_u$ is called an iteration. After l such iterations a sampled vector $\mathbf{X}_u^{(l)}$ will result. Under mild conditions as $l \to \infty$, $\mathbf{X}_u^{(l)}$ converges in distribution to an observation from $f(\mathbf{X}_u|\mathbf{X}_0=\mathbf{x}_0)$. Such conditions mandate that we can not permit any purely deterministic nodes in our graphical model. Technically we merely remove any such nodes, adjusting parents and children accordingly.

Let us be more specific about the form of the complete conditional distribution for X'. It is clear that $f(X'|\mathbf{X}'_u,\mathbf{X}_0=\mathbf{x}_0)$ is proportional to (1). Moreover, with regard to factorization, only terms involving X' as a child or as a parent need be considered so that

$$f(X'|\mathbf{X}'_u,\mathbf{X}_0=\mathbf{x}_0) \propto \left(f[X'|pa(X')] \prod_{V\in ch(X)} f[V|pa(V)]\right)\Big|_{(\mathbf{X}_u,\mathbf{x}_0)} \qquad (6)$$

Thus, the only variables involved in the complete conditional distribution of X', are its parents, its children, and the parents of these children. This set has been called the Markov blanket by Pearl (1986). Since typically dependence in the model is sparse, only a few of the terms in (1) appear in (6).

Turning to the sampling itself we recall that by virtue of the Markovian updating the conditional levels in (6) will change with each draw of X'. In addition, the form of (6) will almost never be a standard distribution even if the individual terms in the product are. For discrete variables sampling is routine upon standardization/summing of (6) although for ordinal variables rejection methods are available (Devroye, 1986). For continuous variables we might also consider a rejection method (see Devroye, 1986 or Ripley 1987). For instance an envelope function for (6) is $Mf[X'|pa(X')]$ where $M=\sup_{X'} \prod_{V\in ch(Y)} f(V|pa(V))\Big|_{(X_u,x_o)}$. Typically, $f[X'|pa(X')]$ is readily sampled and M is not difficult to compute though it must be recomputed with changing levels of X'_u. In practice such envelopes tend to be very inefficient producing very small acceptance rates. This is not surprising since the concentration of mass for $f[X'|pa(X')]$ may be quite different from that of (6).

Alternative "black box" Gibbs sampler for graphical models handle continuous variables by approximate inversion of the probability integral transform as in, e.g., Tanner (1991), by the use of a modified ratio of uniforms method as described in Wakefield, Gelfand and Smith (1991), by adaptive normal kernel density approximations to (6) (see Silverman, 1986, §5.3).

4. Example; Large Network Processing by Subnetwork Approximations

In this section, we illustrate subgraph approximations to desired marginal and conditional distributions and demonstrate that with increasingly broader subnetworks, approximations improve. To determine which variables are included in a subnetwork of radius r around a particular node, consider the following simple procedure:

(1). Determine the matrix of distances of every variable in the network from all others. This is accomplished by determining the shortest path from one variable to the

other (least number of conditioning steps).

(2). Choose a variable i (anyone will do since this process exhausts the complete network in one pass). Enter a value of 1 in all squares i,j in the grid where $j \in pa(i) \cup ch(i)$. Enter a value of 2 for all nodes k in the parent or children set of each j if a lower value has not been already entered (the lower value is always kept). Repeat until no nodes are left.

(3). Include in the subnetwork of the particular variable all those variables that are within the given distance from it, by consulting its line in the above distance matrix.

(4). When finished, check for cases where parental lists are incomplete and add missing parents, accordingly, to the subgraph.

To illustrate why step 4 is necessary, consider the case r=1. Suppose a child of the node of interest has a parent not included in the subgraph. When we attempt to analyze the subgraph as a DAG, we will be unable to use this node's conditional distribution as defined in the full graph.

After step 4 the reduced network will have a set of source nodes. These nodes require an approximate marginal distribution since the exact marginal is unknown and the conditional distribution cannot be used. Rough approximations can be based upon a few replications of the complete network. In particular for discrete source variables taking on, say, k levels, we smooth observed proportions by adding 1/k to each count. For continuous source variables we use a mixture of the conditional distributions at that node based upon the replications of the complete model. Arguments in section 2 suggest that with increasing radius the quality of these approximations becomes less consequential.

In our illustrative 40–node network, nodes 5, 13, 15, 20, 25, 33, 35 and 40 are continuous with the remainder binary. Linear logit models were used for the discrete variables, that is,

$$p(X_i=1) = \frac{\exp\left(\beta_0 + \sum_{j \in pa(X_i)} X_j \beta_j \right)}{1 + \exp\left(\beta_0 + \sum_{j \in pa(X_i)} X_j \beta_j \right)} \qquad (7)$$

in the discrete case, with $p(X_i=0) = 1-p(X_i=1)$. In the continuous case normality on the log–scale was assumed with

$$\mu_{pa(X_i)} = \exp\left(\beta_0 + \sum_{j \in pa(X_i)} X_j \beta_j\right) \qquad (8)$$

Variances were assumed constant though, if appropriate, forms analogous to (8) could be used.

[Insert figure 3 about here]

A general strategy for implementing subgraph approximation is as follows. Start with a subnetwork of small radius and continue by increasing the radius in a step wise manner until successive density estimates are stable. Typically, satisfactory estimates can be obtained employing subnetworks of radius far smaller than the maximum, though this is by no means guaranteed. When dependence is very strong, as information is filtered through successive hierarchical steps, it may be that only minimal amounts are lost.

In figure 3 the complete network is presented, with four subnetworks of node 20 included in figures 4a through 4d. The approximations of the marginal distribution of node 20 based on each subnetwork are plotted in figure 5.

[Insert figures 4a–4d about here]

They are compared to the correct marginal distribution (in solid). We see a gradual improvement of the approximations to the true density. In Table 1, the point estimates of the means, produced by the successive subnetworks around node 20 are given.

[Insert figure 5 about here]

Radius	Estimate
1	11.6471
2	10.6072
3	9.6503
4	10.4683
5	10.3841
Network	10.0830

Table 1. Point estimates of the means of subgraph approximations for node 20.

Finally, the subnetwork approximations of the *conditional* distribution around node 20 are plotted. Nodes 1, 11, 13, 14, 27, 31, 33, 34 were set to fixed levels. Each subnetwork approximation includes some fixed nodes while excluding others. As seen in figure 6 the approximations based on the subnetworks of radius 3 and 4 are sufficient to obtain an adequate estimate of the conditional distribution of node 20.

[Insert figure 6 about here]

References

Darroch, J. N., Lauritsen, S. L. and Speed, T. P. (1980). Markov Fields and Log–Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, 8, 522–539.

Degroot, M. H. and Goel, P. K. (1986) Information and Bayesian Hierarchical Models. *Unpublished Manuscript*.

Devroye, L. (1986) *Non–Uniform Random Number Generation*. Springer–Verlag, New York.

Ferguson, T. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions. *Recent Advances in Statistics*, M. Haseeb Rizvi et al. eds., Academic Press: New York, 287–302.

Gelfand, A. E. and Smith A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398–409

Geman, S. and Geman, D (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Simulation. *Econometrica*, 57, 1317–1339.

Henrion, M. (1988). Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling. In *Uncertainty in Artificial Intelligence*, Vol 2, J. Lemmer, and L. N. Kanal (Eds.), North–Holland: Amsterdam, 149–164.

Lauritsen, S. L. (1990a). Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models. *Research Report* 90–18, Aalborg University.

Lauritsen, S. L. (1990b). Mixed Graphical Association Models, *Scandinavian Journal of Statistics*, 16, 273–306.

Lauritsen, S. L. & Spiegelhalter, D. J. (1988). Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems. *Journal of the Royal Statistical Society*, series B , 50, 57–224.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087–1091.

Oh, M–S, and Berger, J. O. (1991). Adaptive Imputation Sampling in Monte Carlo Integration. *Journal of Statistical Computing and Simulation* (to appear).

Pearl, J. (1986). Propagation and Structuring in Belief Networks. *Artificial Intelligence*, 29, 241–288.

Pearl, J. (1987). Evidential Reasoning Using Stochastic Simulation of Causal Models. *Artificial Intelligence*, 32, 247–257.

Ripley, B. (1989). *Stochastic Simulation*. John Wiley and Sons, New York.

Rubin, D. (1988). Using the SIR Algorithm to Simulate Posterior Distributions. In *Bayesian Statistics 3*. Eds. J. Bernardo et al., Oxford University Press, 395–402.

Shachter, R. D.& Peot, M. A. (1989). Evidential Reasoning Using Likelihood Weighting. *Artificial Intelligence*, (submitted).

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Smith, A. F. M., & Gelfand, A. E. (1991). Bayesian Statistics Without Tears: A Sampling–Resampling Perspective. *The Americal Statistician* (to appear).

Tanner, M. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*. Lecture Notes In Statistics, Springer–Verlag, New York.

Wakefield, J., Gelfand, A. E., & Smith, A. F. M. (1991). Efficient Computation of Random Variates via the Ratio–of–Uniforms Method. Statistics and Computing, 1, 129–134.

West, M. (1991). Approximating Posterior Distributions by Mixtures. In *Bayesian Statistics 4*, eds. J. Bernardo et al, Oxford University Press (to appear).

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley and sons, New York.

Yiannoutsos, C. T., & Gelfand, A. E. (1991). Simulation Approaches for Calculations in Directed Graphical Models. *Technical Report*, 91–23, Department of Statistics, University of Connecticut.

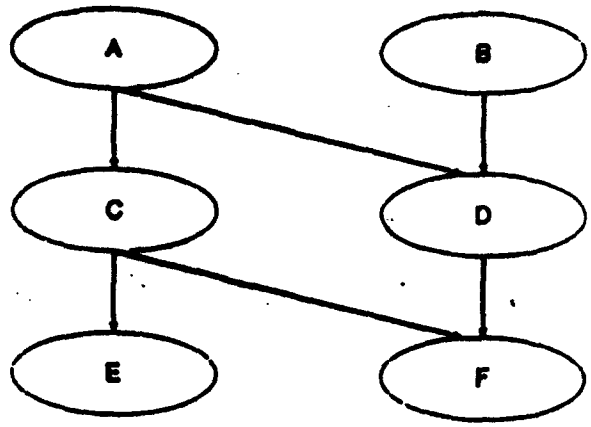


Figure 1: An Illustrative six node Directed Graphical Model

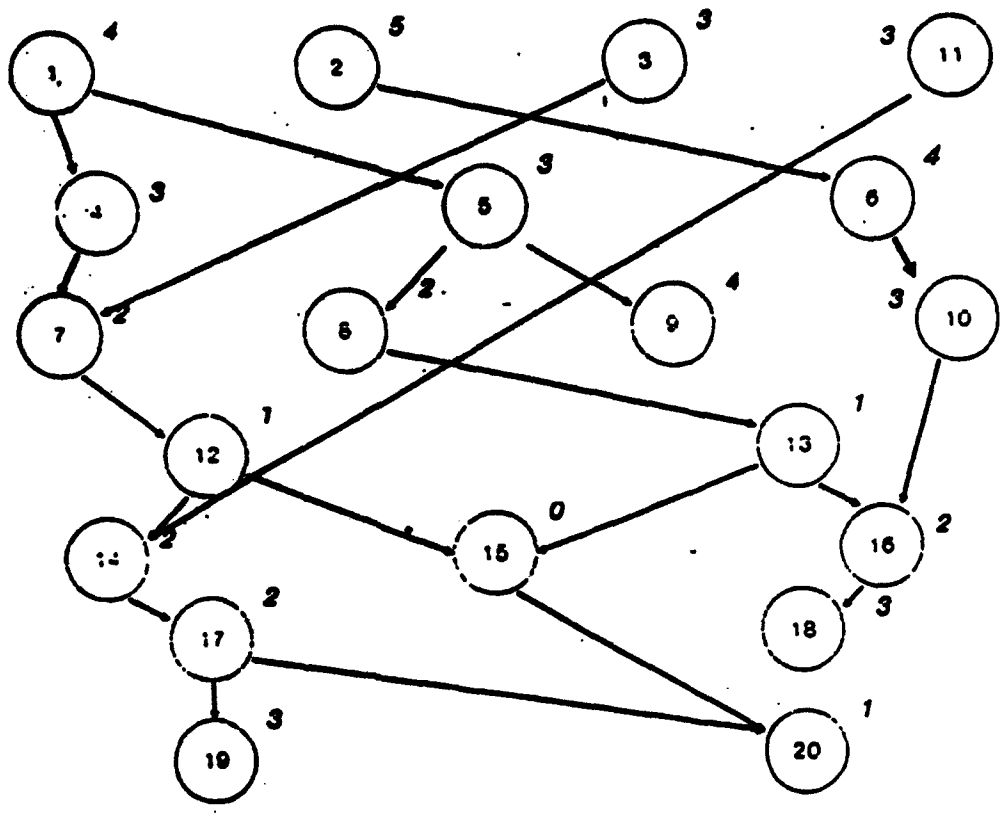


Figure 2: An Illustrative twenty node graph with distances from the target node 15 indicated

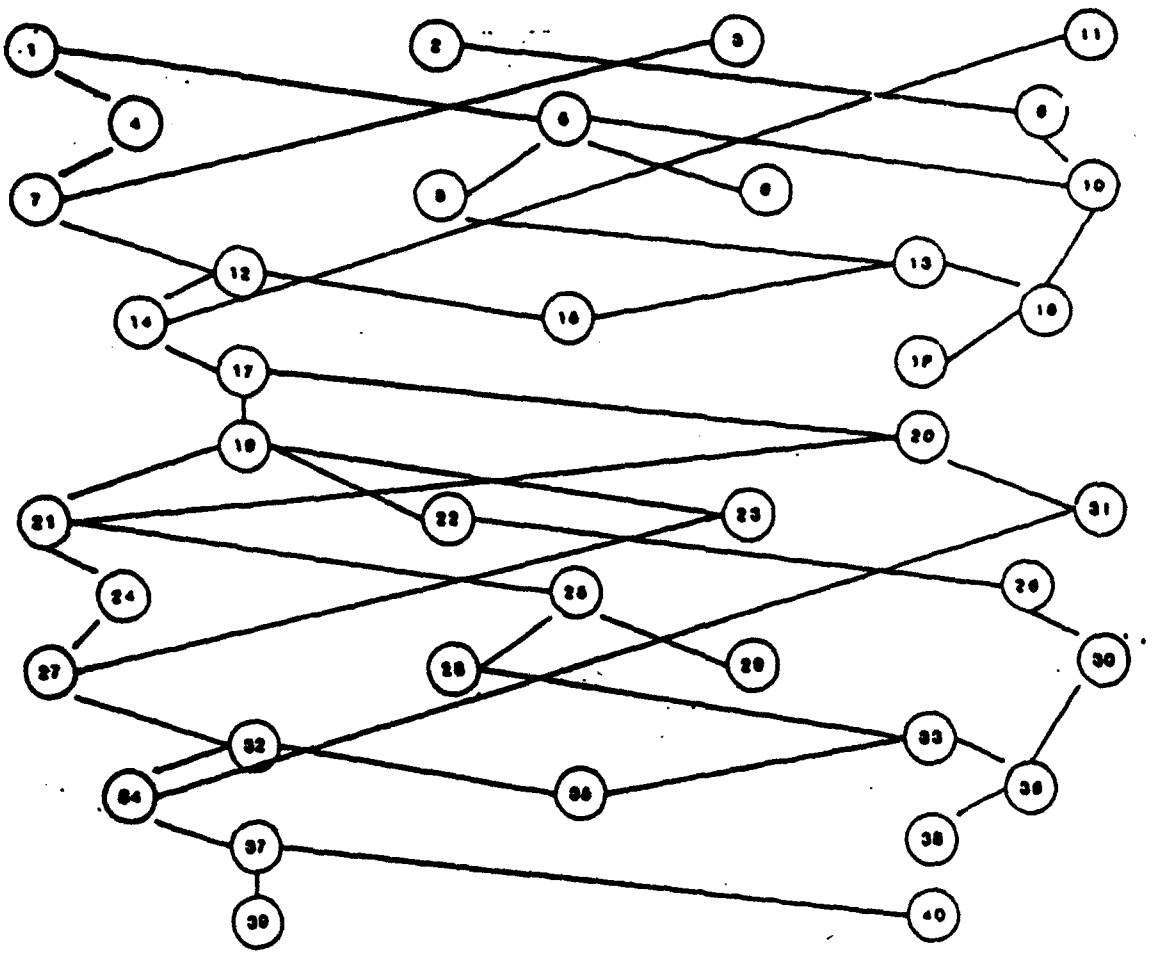
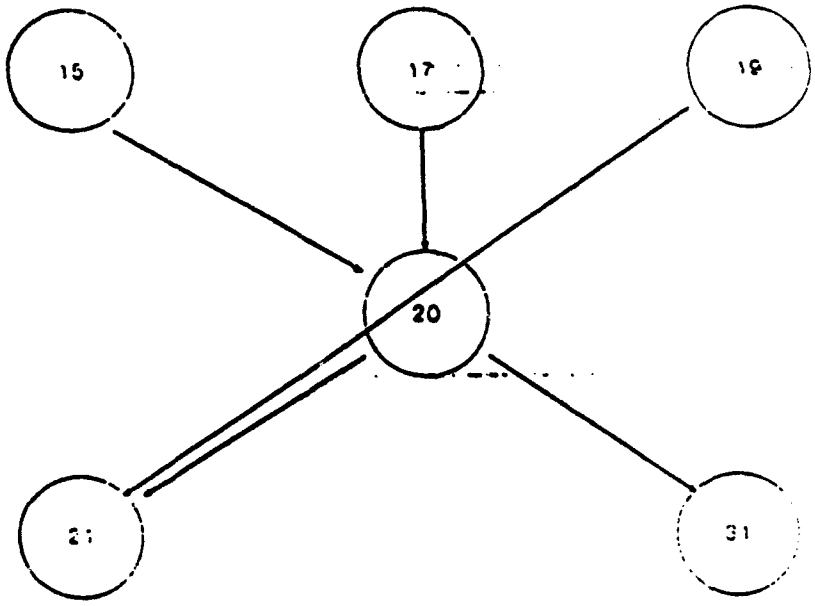


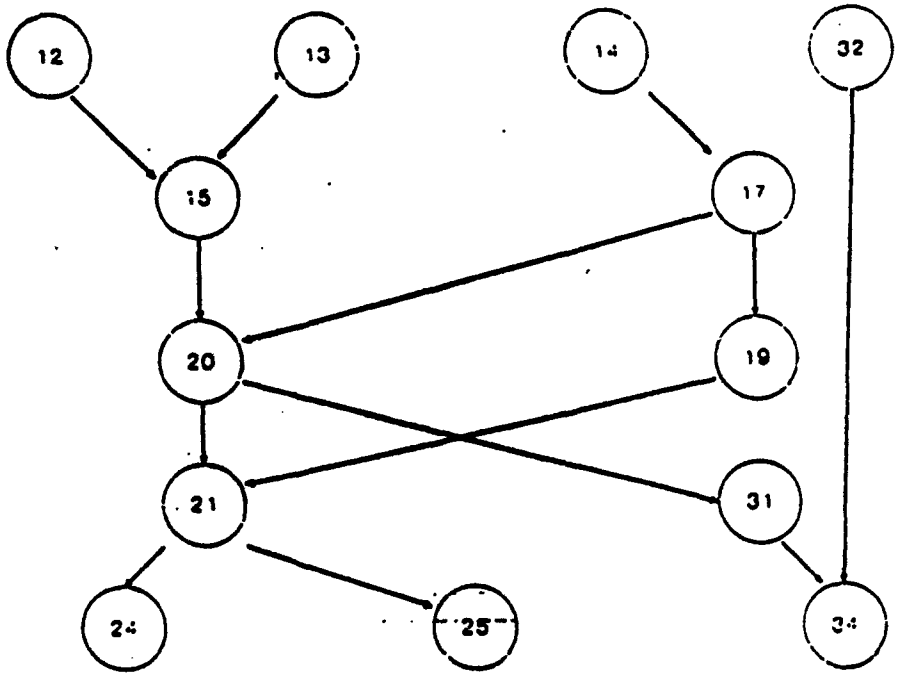
Figure 3: A forty node example

Figure 4: Subgraphs for node 20 of radius 1,2,3, and 4 using the graph of figure 3.

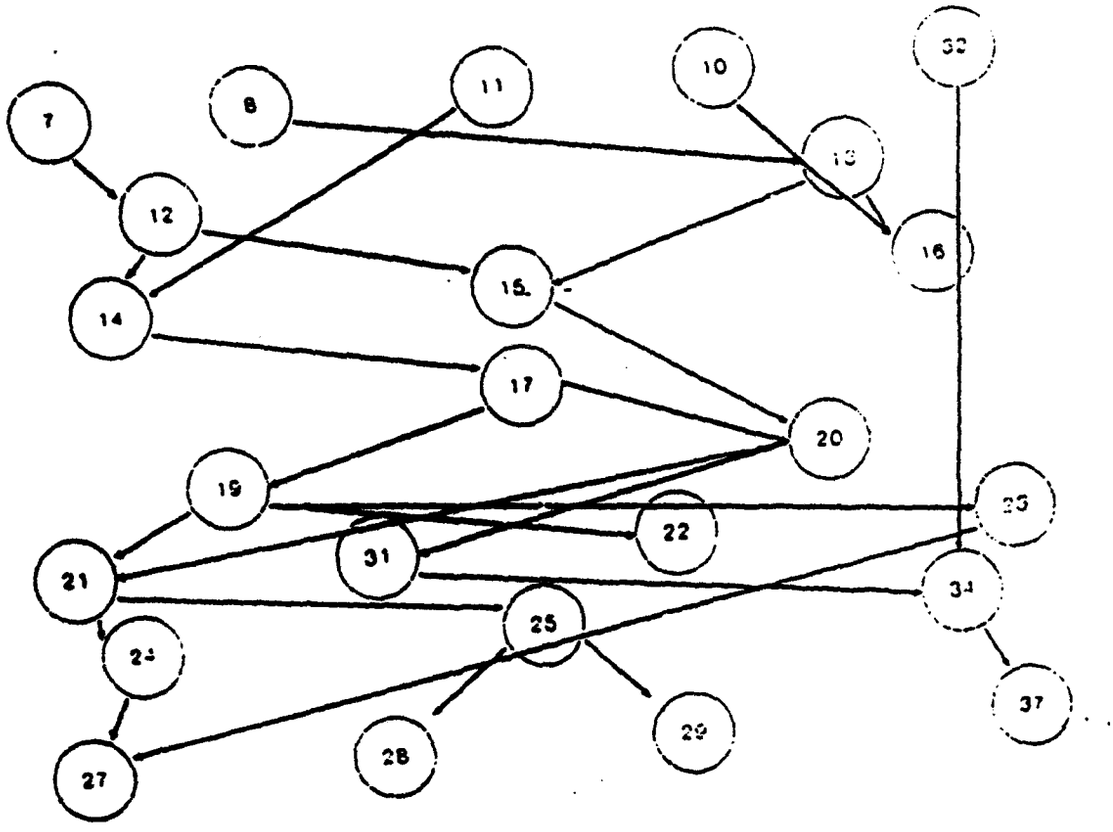
4a) r=1



4b) r=2



+c) r=3


7
8
11
10
32
13
12
16
15
14
17
20
19
33
22
31
21
34
25
24
37
27
28
29

+d) r=4

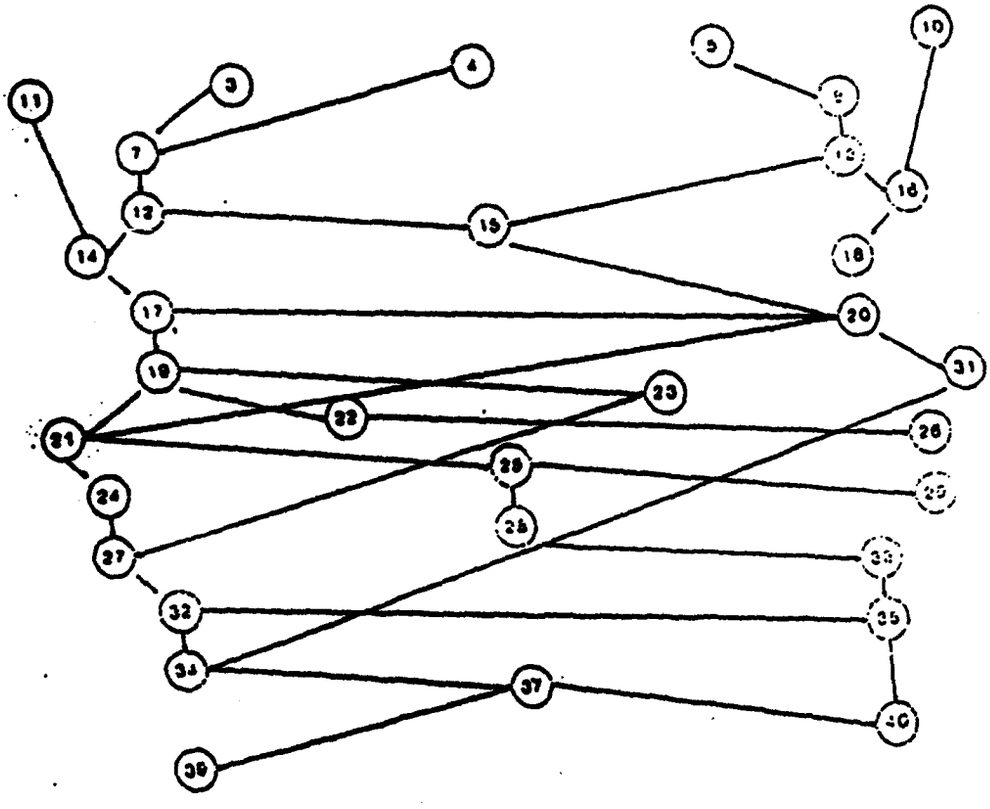

11
3
4
9
10
7
12
15
16
14
18
17
20
31
19
23
22
21
26
28
24
25
27
33
32
35
34
37
40
39

Figure 5: Approximations to the marginal distribution of node 20

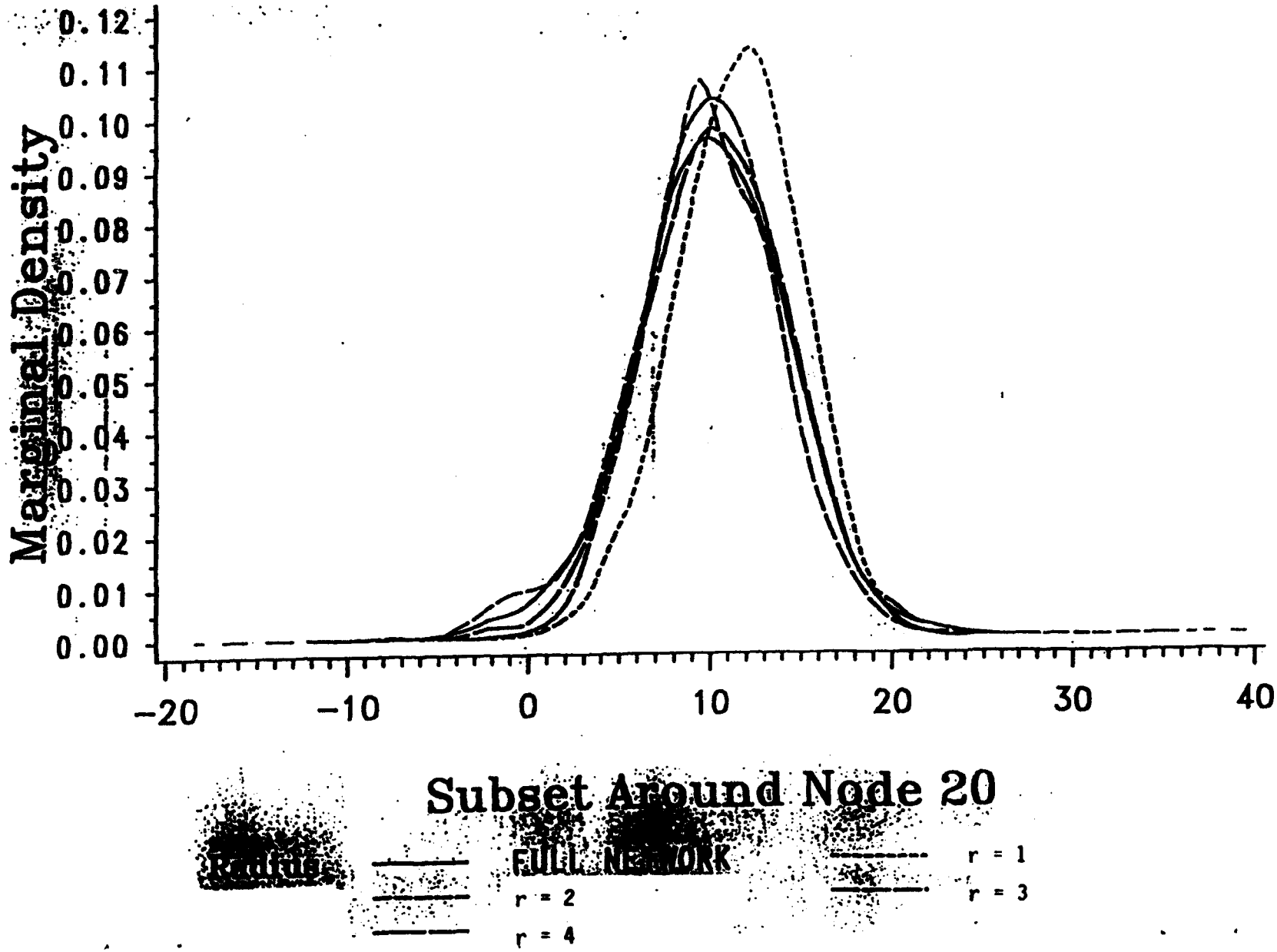
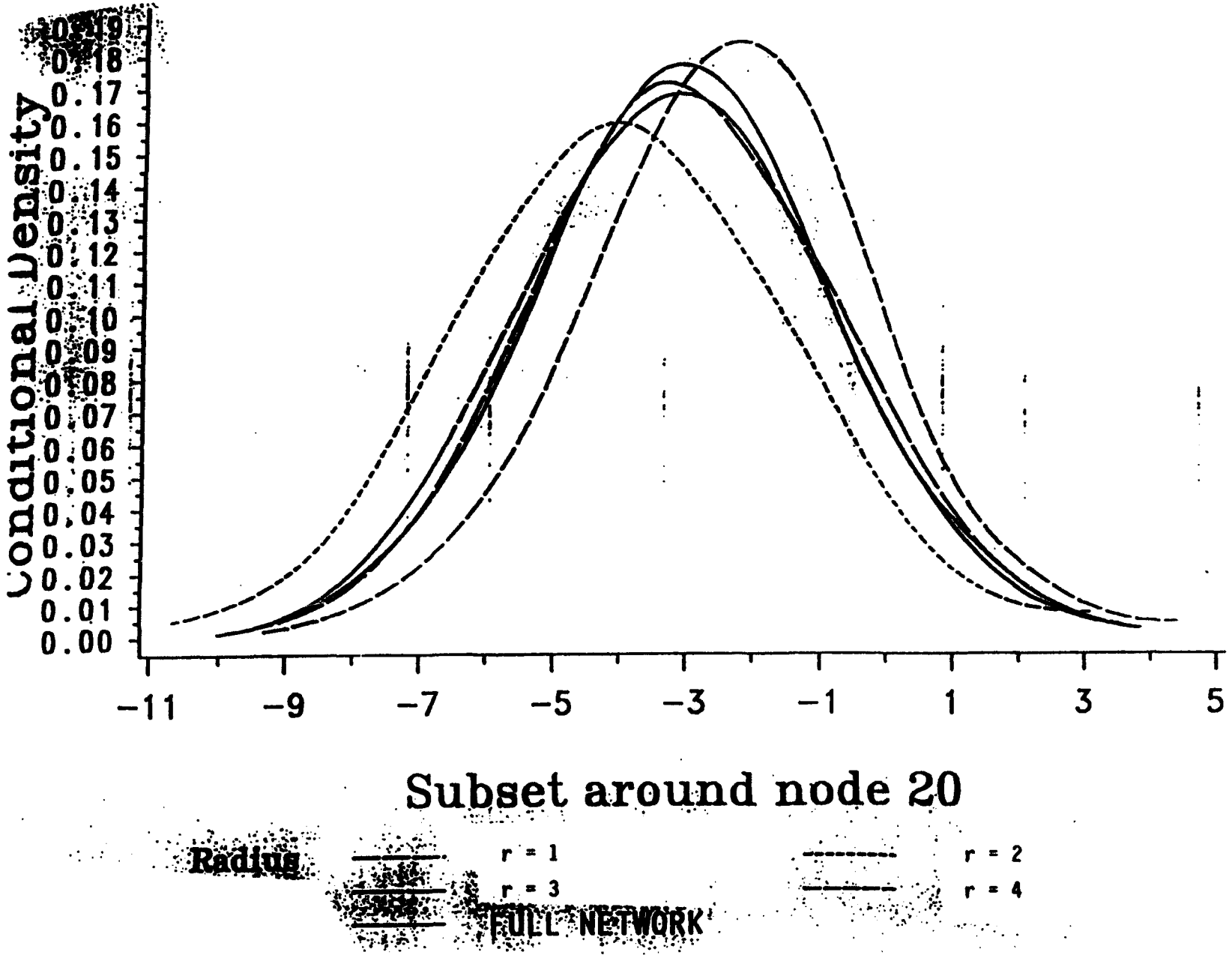


Figure 6: Approximations to the conditional distribution of node 20



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 473	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Subgraph Approximations for Large Directed Graphical Models		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Constantin T. Yiannoutsos and Alan E. Gelfand		8. CONTRACT OR GRANT NUMBER(s) N0025-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 111		12. REPORT DATE Sept. 27, 1993 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Conditional independence, Gibbs sampler, Kullback-Leibler distance, L^1 distance, likelihood weighting, Monte Carlo, Propagation of Information.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

DD FORM 1473, JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

ABSTRACT

Graphical Models provide a powerful tool for the formulation of general statistical models. In a previous paper, (Yiannoutsos & Gelfand, 1991), the authors argued that sampling-based techniques provide a unified approach for the analysis of graphical models under general distributional specifications. These techniques include both noniterative and iterative Monte Carlo.

Our concern here is with very large graphical models whose size and complexity may prohibit analysis within a reasonable time frame. Typically in large systems however, interest focuses on the behavior of only a few critical nodes. Our proposal is to develop, for a particular node, an approximating subgraph which contains virtually as much information about the variable as the full network, but by virtue of its reduced size, enables rapid computational investigation. We provide an illustration using a 40-node graph. Though this is not as large as we would envision in practice, it is convenient in permitting full model calculations to enable assessment of our approximations.