KYBERNETIKA — VOLUME *50* (2014), NUMBER 3, PAGES 363–377

VARIATIONS ON UNDIRECTED GRAPHICAL MODELS AND THEIR RELATIONSHIPS

David Heckerman, Christopher Meek and Thomas S. Richardson

We compare alternative definitions of undirected graphical models for discrete, finite variables. Lauritzen [7] provides several definitions of such models and describes their relationships. He shows that the definitions agree only when joint distributions represented by the models are limited to strictly positive distributions. Heckerman et al. [6], in their paper on dependency networks, describe another definition of undirected graphical models for strictly positive distributions. They show that this definition agrees with those of Lauritzen [7] again when distributions are strictly positive. In this paper, we extend the definition of Heckerman et al. [6] to arbitrary distributions and show how this definition relates to those of Lauritzen [7] in the general case.

Keywords: graphical model, undirected graph, Markov properties, Gibbs sampler, conditionally specified distributions, dependency network

Classification: 60E05, 62H99, 68T30

1. INTRODUCTION

Lauritzen [7, Ch.3] provides alternative definitions of undirected models for discrete variables with finite state spaces. In particular, given an undirected graph $\mathcal{G}$, he defines the following families of distributions:

- $\mathcal{M}_P(\mathcal{G})$: The family of distributions satisfying the pairwise Markov property relative to $\mathcal{G}$,
- $\mathcal{M}_L(\mathcal{G})$: The family of distributions satisfying the local Markov property relative to $\mathcal{G}$,
- $\mathcal{M}_G(\mathcal{G})$: The family of distributions satisfying the global Markov property relative to $\mathcal{G}$,
- $\mathcal{M}_F(\mathcal{G})$: The family of distributions that can be written as a product of potentials over the maximal cliques in the graph,
- $\mathcal{M}_E(\mathcal{G})$: The family of distributions that can be written as a limit of strictly positive distributions in $\mathcal{M}_F(\mathcal{G})$.

DOI: 10.14736/kyb-2014-3-0363

In addition, he defines families of distributions limited to strictly positive distributions among the first four families just listed, denoted $\mathcal{M}_P^+(\mathcal{G})$, $\mathcal{M}_L^+(\mathcal{G})$, $\mathcal{M}_G^+(\mathcal{G})$, and $\mathcal{M}_F^+(\mathcal{G})$, respectively. He shows that, whereas the strictly positive families are equal (and denoted $\mathcal{M}^+(\mathcal{G})$), the general families are not:

$$\mathcal{M}^+(\mathcal{G}) \subset \mathcal{M}_F(\mathcal{G}) \subset \mathcal{M}_E(\mathcal{G}) \subset \mathcal{M}_G(\mathcal{G}) \subset \mathcal{M}_L(\mathcal{G}) \subset \mathcal{M}_P(\mathcal{G}).$$

Lauritzen shows that each inclusion is strict by way of examples. A summary of this work is shown in the Venn diagram in Figure 1a.

In this paper, we contrast these model definitions with the definition via *dependency networks* introduced by Heckerman et al. [6]. Dependency networks are a natural extension of initial efforts to formalize undirected graphical models (a.k.a. Markov networks) and spatial statistical systems. As is discussed in Besag [4], several researchers including Lévy [8], Bartlett [3, Section 2.2], and Brooks [5] considered lattice systems where each variable X depends only on its nearest neighbors ne_X, and quantified the dependencies within these systems using the conditional probability distributions $p(x|\,\mathrm{ne}_X)$. A dependency network uses such distributions in conjunction with a single-site Gibbs sampler to define a joint distribution. Arnold et al. [2] give essentially the same definition under the name *conditionally specified distributions*. Yang et al. [12] contains another more recent application of the same idea.

Heckerman et al. [6] show that their definition, limited to strictly positive distributions, coincides with that of $\mathcal{M}^+(\mathcal{G})$. We extend their definition to include arbitrary distributions. We call the resulting set of distributions the *conditionally specified undirected graphical model* associated with $\mathcal{G}$ and denote it by $\mathcal{M}_C(\mathcal{G})$. We provide examples that demonstrate the relationships among $\mathcal{M}_C(\mathcal{G})$, $\mathcal{M}_P(\mathcal{G})$, $\mathcal{M}_L(\mathcal{G})$, $\mathcal{M}_G(\mathcal{G})$, $\mathcal{M}_F(\mathcal{G})$, and $\mathcal{M}_E(\mathcal{G})$. A summary of our work is shown in Figure 1b.

2. NOTATION AND DEFINITIONS

In this section, we review some basic graph-theoretic definitions and notation and define conditionally specified graphical models.

We use $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ to denote an undirected graph where $\mathbf{V} = \{A, B, C, \ldots\}$ denotes the set of vertices and $\mathbf{E}$ is the set of edges. We will denote an edge between two vertices A and B by $A - B$ and denote the set of neighbors of a vertex A by $\mathrm{ne}_{\mathcal{G}}(A)$. In order to associate a distribution with a graph we associate a random variable X_A with each vertex A. We will let $\mathcal{X}_A$ denote the state-space for X_A. We assume that each variable takes a finite set of possible values. We use x_V to denote one such value. We denote the set of variables corresponding to a set of vertices $\mathbf{C}$ by $\mathbf{X_C}$ with state-space $\mathcal{X}_\mathbf{C}$ and denote the set of all variables by $\mathbf{X} = \mathbf{X_V}$.

As mentioned, our definition of a conditionally specified undirected graphical model is based on the definition of dependency network given by Heckerman et al. [6]. Roughly, a dependency network for $\mathbf{X_V}$ consists of a graph $\mathcal{G}$ and a set of strictly positive conditional distributions, one for each vertex V, conditioned on the random variables corresponding to its neighbors in the graph.

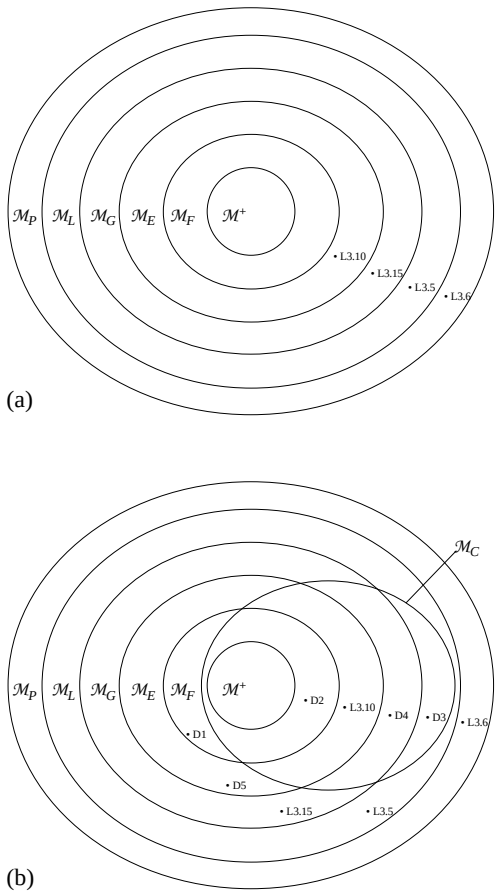

Fig. 1. (a) Relationships among definitions of undirected graphical models as described by Lauritzen [7]. (b) Additional relationships among the definitions in (a) and $\mathcal{M}_C(\mathcal{G})$. Labeled points correspond to example distributions demonstrating the non-emptiness of cells in the Venn diagram. Example distributions in Lauritzen [7] and this note are prefixed with "L" and "D", respectively.

More formally a dependency network is defined as follows:
A *Markov kernel* $K_V(x_V \mid \mathbf{x}_{\mathrm{ne}(V)})$ with *conditional domain* $\mathcal{N}_V$ is a function:

$$K_V : \mathcal{X}_V \times \mathcal{N}_V \;\rightarrow\; [0,1]$$

with $\mathcal{N}_V \subseteq \mathcal{X}_{\mathrm{ne}(V)}$ that satisfies:

$$1 = \sum_{x_V \in \mathcal{X}_V} K_V(x_V \mid \mathbf{x}^*_{\mathrm{ne}(V)}) \quad \text{for all } \mathbf{x}^*_{\mathrm{ne}(V)} \in \mathcal{N}_V .$$

A *dependency network specification* associated with graph $\mathcal{G}$ is a set of Markov kernels and conditional domain pairs, one for each vertex:

$$\{(K_V(x_V \mid \mathbf{x}_{\mathrm{ne}(V)}), \mathcal{N}_V) \mid V \in \mathbf{V}\}.$$

We also define:

$$\mathcal{N} \equiv \bigcap_{V \in \mathbf{V}} \{\mathbf{x}_\mathbf{V} \mid \mathbf{x}_\mathbf{V} = (\mathbf{x}_{\mathrm{ne}(V)}, \mathbf{x}_{\mathbf{V} \setminus \mathrm{ne}(V)}) \in \mathcal{N}_V \times \mathcal{X}_{\mathbf{V} \setminus \mathrm{ne}(V)}\}.$$

Thus $\mathcal{N} \subseteq \mathcal{X}_\mathbf{V}$ is the subset of vectors $\mathbf{x}_\mathbf{V} \in \mathcal{X}_\mathbf{V}$ such that the subvectors $\mathbf{x}_{\mathrm{ne}(V)} \in \mathcal{N}_V$ for every V.

A dependency network will be said to be *consistent* if there *exists* a distribution $P(\mathbf{x}_\mathbf{V})$ that has support only on a (possibly strict) subset of $\mathcal{N}$, so:

$$\text{for every } \mathbf{x}_\mathbf{V} \in \mathcal{X}_\mathbf{V}, \quad P(\mathbf{x}_\mathbf{V}) > 0 \;\Rightarrow\; \mathbf{x}_\mathbf{V} \in \mathcal{N}, \tag{1}$$

and, in addition,

$$\text{for all } V \in \mathbf{V} \text{ and } \mathbf{x}_\mathbf{V} \in \mathcal{N},\ P(\mathbf{x}_{\mathbf{V} \setminus \{V\}}) > 0 \;\Rightarrow\; P(x_V \mid \mathbf{x}_{\mathbf{V} \setminus \{V\}}) = K_V(x_V \mid \mathbf{x}_{\mathrm{ne}(V)}). \tag{2}$$

Note that it follows from (2) that any such distribution $P(\mathbf{x}_\mathbf{V})$ obeys the local Markov property for $\mathcal{G}$ and hence $P(\mathbf{x}_\mathbf{V}) \in \mathcal{M}_L(\mathcal{G})$. Further note that the requirement (1) implies that for all $\mathbf{x}_\mathbf{V} \in \mathcal{N}$, $V \in \mathbf{V}$, and $x_V^* \in \mathcal{X}_V$, if $P(\mathbf{x}_{\mathbf{V} \setminus \{V\}}) > 0$ and $P(x_V^* \mid \mathbf{x}_{\mathbf{V} \setminus \{V\}}) > 0$ then $(x_V^*, \mathbf{x}_{\mathbf{V} \setminus \{V\}}) \in \mathcal{N}$.

Given a distribution $P(\mathbf{x}_\mathbf{V})$ obeying the local Markov property for $\mathcal{G}$, there is a natural dependency network specification resulting from $P(\mathbf{x}_\mathbf{V})$ given by defining for each $V \in \mathbf{V}$:

$$\mathcal{N}_V \equiv \{\mathbf{x}_{\mathrm{ne}(V)} \mid P(\mathbf{x}_{\mathrm{ne}(V)}) > 0\} \subseteq \mathcal{X}_{\mathrm{ne}(V)}, \tag{3}$$

$$K_V(x_V \mid \mathbf{x}_{\mathrm{ne}(V)}) \equiv P(x_V \mid \mathbf{x}_{\mathrm{ne}(V)}).$$

It follows from the definitions that the support of $P(\mathbf{x}_\mathbf{V})$ is contained in the resulting set $\mathcal{N}$. That is, condition (1) is satisfied. The following is an immediate consequence:

Proposition 2.1. Given a distribution $P(\mathbf{x}_\mathbf{V})$ that obeys the local Markov property, the dependency network specification resulting from $P(\mathbf{x}_\mathbf{V})$ is consistent.

However, though the dependency network specification given by a distribution $P(\mathbf{X}_\mathbf{V})$ (obeying the local Markov property for $\mathcal{G}$) will be consistent by definition, there may exist another distribution $P^*(\mathbf{x}_\mathbf{V}) \neq P(\mathbf{x}_\mathbf{V})$ satisfying (1) and (2). This leads to the following:

A distribution $P(\mathbf{x}_\mathbf{V})$ obeying the local Markov property for $\mathcal{G}$ will be said to be in the *conditionally specified model* $\mathcal{M}_C(\mathcal{G})$ if $P(\mathbf{x}_\mathbf{V})$ is the unique consistent distribution that has the natural dependency network specification resulting from $P(\mathbf{x}_\mathbf{V})$. Equivalently, $P(\mathbf{x}_\mathbf{V}) \in \mathcal{M}_C(\mathcal{G})$ if $P(\mathbf{x}_\mathbf{V})$ obeys the local Markov property for $\mathcal{G}$ and there is no other distribution $P^*(\mathbf{x}_\mathbf{V})$ such that for all V, $P^*(x_V \mid x_{\mathbf{V} \setminus \{V\}}) = P(x_V \mid x_{\mathbf{V} \setminus \{V\}})$.

We will also associate with a dependency network specification a set of *single-variable Gibbs transition kernels* defined on $\mathcal{N}$, one for each vertex V, as follows:

$$M_V(\mathbf{x}^*_{\mathbf{V}}, \mathbf{x}_{\mathbf{V}}) = P(\mathbf{x}_{\mathbf{V}} \mapsto \mathbf{x}^*_{\mathbf{V}}) \equiv I(\mathbf{x}_{\mathbf{V}\setminus\{V\}} = \mathbf{x}^*_{\mathbf{V}\setminus\{V\}}) K_V(x^*_V \mid \mathbf{x}_{\mathrm{ne}(V)}),$$

where $I(\cdot)$ is the indicator function, $\mathbf{x}_{\mathbf{V}}, \mathbf{x}^*_{\mathbf{V}} \in \mathcal{N}$ and $M_V(\mathbf{x}_{\mathbf{V}}, \mathbf{x}^*_{\mathbf{V}})$ is a $|\mathcal{N}| \times |\mathcal{N}|$ transition matrix.

Proposition 2.2. Given a consistent dependency network specification, a distribution P satisfying (1) and (2), and states $\mathbf{x}_{\mathbf{V}}$, $\mathbf{x}^*_{\mathbf{V}}$ such that $P(\mathbf{x}_{\mathbf{V}}) > 0$ and $M_V(\mathbf{x}^*_{\mathbf{V}}, \mathbf{x}_{\mathbf{V}}) > 0$ then $P(\mathbf{x}^*_{\mathbf{V}}) > 0$, and thus $\mathbf{x}^*_{\mathbf{V}} \in \mathcal{N}$.

Thus given an initial state $\mathbf{x}_{\mathbf{V}}$ in the support of P (and hence in $\mathcal{N}$), with probability one the transition kernel M_V takes us to a state $\mathbf{x}^*_{\mathbf{V}}$ that is also in $\mathcal{N}$.

P r o o f . Suppose otherwise, so that $P(x^*_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}}) > 0$, but $P(\mathbf{x}^*_{\mathbf{V}}) = 0$. Note that by definition of M_V, $\mathbf{x}^*_{\mathbf{V}\setminus\{V\}} = \mathbf{x}_{\mathbf{V}\setminus\{V\}}$. Since $P(\mathbf{x}^*_{\mathbf{V}}) = P(x^*_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}}) P(\mathbf{x}_{\mathbf{V}\setminus\{V\}})$, this implies that $P(\mathbf{x}_{\mathbf{V}\setminus\{V\}}) = 0$. However, this is a contradiction since $0 < P(\mathbf{x}_{\mathbf{V}}) \leq P(\mathbf{x}_{\mathbf{V}\setminus\{V\}})$. □

We will consider a two-stage Gibbs sampling scheme whereby a single vertex $V \in \mathbf{V}$ is picked randomly and then a new state $\mathbf{x}^*_{\mathbf{V}}$ is obtained from the Markov kernel M_V. We will refer to this as *the Gibbs sampler associated with the dependency network specification*. It is not difficult to see that the (multiple-variable) Gibbs transition kernel for this overall scheme is then:

$$M \equiv \frac{1}{|\mathbf{V}|} \sum_{V \in \mathbf{V}} M_V.$$

Proposition 2.2 shows that given a starting point $\mathbf{x}_V$ within the support of P (hence in $\mathcal{N}$), the Gibbs sampler will only transition to other points within the support of P (hence in $\mathcal{N}$).

A distribution μ on $\mathcal{N}$ will be said to be *a stationary distribution* of the Gibbs sampler if $\mu M = \mu$. Recall that given a Markov chain on a set $\mathcal{N}$, a state n_2 is said to be *accessible* from n_1, written $n_1 \to n_2$, if after sufficiently many transitions, there is a non-zero probability of transitioning from n_1 to n_2. A state n_1 is said to be *essential* if for every n_2 such that $n_1 \to n_2$, it also holds that $n_2 \to n_1$. Lastly a chain is *irreducible* if, for any two states $n_1, n_2 \in \mathcal{N}$, it holds that $n_1 \to n_2$ and $n_2 \to n_1$.

Lemma 2.3. A consistent dependency network specification resulting from a distribution $P(\mathbf{x}_{\mathbf{V}}) \in \mathcal{M}_L(\mathcal{G})$ will give rise to a unique stationary distribution if and only if the Markov chain resulting from the associated Gibbs sampling scheme is irreducible.

P r o o f . The Ergodic Theorem (see e. g. Norris [11, p.53]) states that if a Markov chain is irreducible then there exists a unique stationary distribution $\mu(\mathbf{x}_{\mathbf{V}}) \equiv P(\mathbf{x}_{\mathbf{V}})$ for the Markov chain given by M.

We now show the converse. First note that by Proposition 2.2, given an initial starting value in $\mathcal{N}$ the Markov chain will remain in $\mathcal{N}$. Since, by construction, every

state of the chain is essential, if the chain is reducible then $\mathcal{N}$ may be decomposed into disjoint components $\mathcal{N}^1, \ldots, \mathcal{N}^p$ such that the Gibbs sampling scheme is irreducible on each subset $\mathcal{N}^i$. By the Ergodic Theorem, there is a unique stationary distribution with support $\mathcal{N}^i$, call this distribution μ^i. Note that $\mu^i M = \mu^i$. However, since μ^i and μ^j have disjoint support (for $i \neq j$) it follows that, if the Gibbs sampling scheme is not irreducible, then there will not be a unique stationary distribution. □

In addition we have the following result:

Lemma 2.4. Given a consistent dependency network specification that results in an irreducible Gibbs sampler with unique stationary distribution μ, it follows that for any state $\mathbf{x}^*_{\mathbf{V}}$ such that $\mu(\mathbf{x}^*_{\mathbf{V}}) > 0$, and any vertex V, $\mu(x_V \mid \mathbf{x}^*_{\mathrm{ne}(V)}) = K_V(x_V \mid \mathbf{x}^*_{\mathrm{ne}(V)})$.

Proof. By consistency there exists a distribution μ satisfying (2). By definition, any such distribution μ satisfies for all V, $\mu M_V = \mu$, and thus $\mu M = \mu$. The uniqueness of the stationary distribution then implies the conclusion. □

This then leads to the following characterization of $\mathcal{M}_C(\mathcal{G})$:

Theorem 2.5. A distribution $P(\mathbf{x}_{\mathbf{V}}) \in \mathcal{M}_L(\mathcal{G})$ is also in $\mathcal{M}_C(\mathcal{G})$ if and only if the associated Gibbs sampler is irreducible.

Proof. ($\Leftarrow$) This follows immediately from Lemmas 2.3 and 2.4.
($\Rightarrow$) Suppose the Gibbs sampler is reducible, with disjoint components $\mathcal{N}^1, \ldots, \mathcal{N}^p$, $p > 1$. For a given i, consider the distribution $P^i(\mathbf{x}_{\mathbf{V}}) \equiv P(\mathbf{x}_{\mathbf{V}} \mid \mathcal{N}^i)$.

Consider a point $\tilde{\mathbf{x}}_{\mathbf{V}} \in \mathcal{N}^i$. It follows immediately that $P(\tilde{\mathbf{x}}_{\mathbf{V}}, \mathcal{N}^i) = P(\tilde{\mathbf{x}}_{\mathbf{V}})$. Further, by definition of $\mathcal{N}^i$ for any $x'_V \in \mathcal{X}_V$, if $P(x'_V, \tilde{\mathbf{x}}_{\mathbf{V}\setminus\{V\}}) > 0$ then $(x'_V, \tilde{\mathbf{x}}_{\mathbf{V}\setminus\{V\}}) \in \mathcal{N}^i$. Consequently, $P(\tilde{\mathbf{x}}_{\mathbf{V}\setminus\{V\}}, \mathcal{N}^i) = P(\tilde{\mathbf{x}}_{\mathbf{V}\setminus\{V\}})$. It then follows that if $(x_V, \mathbf{x}_{\mathbf{V}\setminus\{V\}}) \in \mathcal{N}^i$ then $P^i(x_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}}) = P(x_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}})$. By construction of $\mathcal{N}^i$, $P^i(\mathbf{x}_{\mathbf{V}})$ implies a dependency network specification leading to an irreducible Markov chain. Hence it follows from ($\Leftarrow$) that $P^i(\mathbf{x}_{\mathbf{V}}) \in \mathcal{M}_C(\mathcal{G})$. Thus for all V, and all $\mathbf{x}_{\mathbf{V}} \in \mathcal{X}_{\mathbf{V}}$,

$$P^i(\mathbf{x}_{\mathbf{V}\setminus\{V\}}) > 0 \quad \Rightarrow \quad P^i(x_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}}) = P(x_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}}) = P(x_V \mid \mathbf{x}_{\mathrm{ne}(V)}). \tag{4}$$

However, since $P^i(\mathbf{x}_{\mathbf{V}})$ has support on a subset of $\mathcal{N}^i$ and $\mathcal{N}^i \subset \mathcal{N}$, (4) establishes that $P^i(\mathbf{x}_{\mathbf{V}})$ satisfies the original dependency network specification given by P. The same argument may be carried through for $j \neq i$, leading to a distribution $P^j(\mathbf{x}_{\mathbf{V}})$ also obeying the original specification. However, $P^i(\mathbf{x}_{\mathbf{V}}) \neq P^j(\mathbf{x}_{\mathbf{V}})$ since the distributions have disjoint support. Hence $P(\mathbf{x}_{\mathbf{V}}) \notin \mathcal{M}_C(\mathcal{G})$. □

It follows from the proof of Theorem 2.5 that a distribution $P(\mathbf{x}_{\mathbf{V}}) \in \mathcal{M}_L(\mathcal{G}) \setminus \mathcal{M}_C(\mathcal{G})$ is a mixture of two or more distributions in $\mathcal{M}_C(\mathcal{G})$ with disjoint supports. These distributions in $\mathcal{M}_C(\mathcal{G})$ are uniquely determined by the full conditionals $P(x_V \mid \mathbf{x}_{\mathbf{V}\setminus\{V\}})$; it is solely the mixing distribution that is not identified. More formally we have shown the following:

Proposition 2.6. If $P(\mathbf{x_V}) \in \mathcal{M}_L(\mathcal{G}) \setminus \mathcal{M}_C(\mathcal{G})$ then for some $k \geq 2$, for every $V \in \mathbf{V}$ there exists a partition

$$\mathcal{X}_V = \mathcal{X}_V^{(1)} \dot{\cup} \cdots \dot{\cup} \mathcal{X}_V^{(k)},$$

and there exist distributions $P^{(1)}(\mathbf{x_V}), \ldots, P^{(k)}(\mathbf{x_V}) \in \mathcal{M}_C(\mathcal{G})$, whereby $P^{(i)}(\mathbf{x_V})$ has support on the Cartesian product $\mathcal{X}_{\mathbf{V}}^{(i)} \equiv \prod_{V \in \mathbf{V}} \mathcal{X}_V^{(i)}$, and there is a distribution $\tilde{P}(i)$ with support $\{1, \ldots, k\}$ such that

$$P(\mathbf{x_V}) = \sum_{i=1}^{k} \tilde{P}(i) \cdot P^{(i)}(\mathbf{x_V}). \tag{5}$$

Hence $P(\mathbf{x_V})$ has support on $\bigcup_{i=1}^{k} \mathcal{X}_{\mathbf{V}}^{(i)} \subsetneq \mathcal{X}_{\mathbf{V}}$.

Given a distribution $P(\mathbf{x_V})$ we define the *Hamming graph for* $P(\mathbf{x_V})$ to be a graph $\mathcal{H}$ defined as follows:

(1) The vertex set of $\mathcal{H}$ corresponds to the set of *values* $\mathbf{x_V}$ in the support of $P(\mathbf{x_V})$;

(2) There is an edge between the vertices corresponding to two distinct support points $\mathbf{x_V}$, $\mathbf{x_V^*}$, if there is a unique $V \in \mathbf{V}$ such that $\mathbf{x}_{\mathbf{V}\setminus\{V\}} = \mathbf{x}^*_{\mathbf{V}\setminus\{V\}}$; thus two support points are joined by an edge only if they differ in the value they assign to a unique vertex V.

The Hamming graph $\mathcal{H}$ is said to be *connected* if there is a path from any vertex $\mathbf{x_V}$ to any other vertex $\mathbf{x_V^*}$.

Theorem 2.7. A distribution $P(\mathbf{x_V})$ obeying the local Markov property for $\mathcal{G}$ is in $\mathcal{M}_C(\mathcal{G})$ if and only if the Hamming graph $\mathcal{H}$ for $P(\mathbf{x_V})$ is connected.

P r o o f . By construction of the Hamming graph, the Gibbs sampling scheme defined by M is irreducible if and only if the Hamming graph is connected. The result then follows from Theorem 2.5. □

Heckerman et al. [6] consider dependency network specifications including those that are not necessarily consistent. Their motivation for this definition of (general) dependency networks is that it is both straightforward and computationally efficient to learn conditional distributions separately and then combine them via Gibbs sampling to obtain a joint distribution. They note that asymptotically (in sample size), the conditional distributions will converge to the true conditional distributions and hence be consistent.

The choice of a single-site Gibbs sampler in our definitions is not arbitrary. The local Markov property is itself a "single-site" property. As noted above, this property will be satisfied by any distribution satisfying (2). It also leads to more computationally efficient inference and to a graphical representation. We can extend the definition of a consistent dependency network to include k-site Gibbs sampling, but we then require a k-site version of the local Markov property to provide similar benefits. Namely, we

say that a joint distribution satisfies the *k-order local Markov property relative to* $\mathcal{G}$ if, for any pair $(\mathbf{A}, \mathbf{S})$ of disjoint subsets of $\mathbf{V}$ such that the size of $\mathbf{A}$ is at most k and $\mathbf{S}$ separates $\mathbf{A}$ from the remaining vertices in $\mathcal{G}$, then $\mathbf{X_A}$ and $\mathbf{X_{V\setminus(A\cup S)}}$ are independent given $\mathbf{X_S}$. Note that, when k represents a sufficiently large number of the nodes in graph $\mathcal{G}$, the k-order local Markov implies the global Markov property. In the remainder of this paper, we concentrate on the case where $k = 1$.

3. RELATING THE MODELS

We now examine the relationships between the various families for a given graph $\mathcal{G}$.

3.1. Strictly positive distributions

As mentioned, Lauritzen [7] shows that $\mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_G^+(\mathcal{G}) = \mathcal{M}_L^+(\mathcal{G}) = \mathcal{M}_P^+(\mathcal{G}) \equiv \mathcal{M}^+(\mathcal{G})$, and Heckerman et al. [6] show that $\mathcal{M}_C^+(\mathcal{G}) = \mathcal{M}^+(\mathcal{G})$. Here, we show that $\mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_E^+(\mathcal{G}) = \mathcal{M}_C^+(\mathcal{G})$, where $\mathcal{M}_E^+(\mathcal{G})$ and $\mathcal{M}_C^+(\mathcal{G})$ are the families $\mathcal{M}_E(\mathcal{G})$ and $\mathcal{M}_C(\mathcal{G})$, respectively, limited to strictly positive distributions.

Lemma 3.1. $\mathcal{M}_E^+(\mathcal{G}) = \mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_C^+(\mathcal{G})$.

Proof. From the definition of $\mathcal{M}_F$, we know that $\mathcal{M}_F^+(\mathcal{G}) \subseteq \mathcal{M}_E^+(\mathcal{G})$. Furthermore, we know that $\mathcal{M}_E(\mathcal{G})$ is a subset of the set of *global Markov* distributions $\mathcal{M}_G(\mathcal{G})$ [7, p. 42] and $\mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_G^+(\mathcal{G})$ [7, p. 34]. From these facts we have $\mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_E^+(\mathcal{G})$.

From the definition of $\mathcal{M}_C$, we know that $\mathcal{M}_C^+(\mathcal{G}) \subseteq \mathcal{M}_L^+(\mathcal{G})$. Furthermore, the set of strictly positive *pairwise Markov* distributions $\mathcal{M}_P^+(\mathcal{G})$ and $\mathcal{M}_G^+(\mathcal{G})$ are equal to $\mathcal{M}_L^+(\mathcal{G})$ (see, e. g. Lauritzen [7, p. 34]). From the Hammersley–Clifford Theorem, $\mathcal{M}_F^+(\mathcal{G}) = \mathcal{M}_P^+(\mathcal{G})$ (see, e. g. Lauritzen [7, p. 36]). Therefore, we have established $\mathcal{M}_C^+(\mathcal{G}) \subseteq \mathcal{M}_F^+(\mathcal{G})$. It is thus sufficient to prove that $\mathcal{M}_F^+(\mathcal{G}) \subseteq \mathcal{M}_C^+(\mathcal{G})$. Let P be in $\mathcal{M}_F^+(\mathcal{G})$. The factorization of the distribution guarantees that P is in $\mathcal{M}_L(\mathcal{G})$. Furthermore, by positivity, the Hamming graph $\mathcal{H}$ has vertex set $\mathcal{X}_\mathbf{V}$ and thus is connected. Hence $P \in \mathcal{M}_C(\mathcal{G})$ by Theorem 2.7. □

3.2. $\exists\, \mathcal{G} : \mathcal{M}_L(\mathcal{G}) \setminus (\mathcal{M}_G(\mathcal{G}) \cup \mathcal{M}_C(\mathcal{G})) \neq \emptyset$

In the remainder of Section 3, we show that certain cells in the Venn diagram of Figure 1b are not empty. We do so with specific examples. Example distributions given by Lauritzen [7] are prefixed with the letter "L." New examples are prefixed with the letter "D." We first consider examples from Lauritzen [7], as some of the examples we introduce are based on them.

An important tool in our constructions will be the characterization of $\mathcal{M}_C(\mathcal{G}_C)$ given in Theorem 2.7. Several of our examples are constructed by adapting previous known examples so that their Hamming graph is connected. The basic idea is that given a graph $\mathcal{G}^*$ over a set of variables $\mathbf{V}^*$ and a distribution $P^*(\mathbf{V}^*)$ that is in one of the sets shown in Figure 1a, we may construct an example that is also in $\mathcal{M}_C(\mathcal{G}_C)$ by considering an extended graph $\mathcal{G}$ with vertices $\mathbf{V}^* \cup \{T\}$, such that the induced subgraph on $\mathbf{V}^*$ is still $\mathcal{G}^*$ and T is adjacent to every vertex in $\mathbf{V}^*$. Finally we let T be a binary variable

and define $P(\mathbf{V}) = P(\mathbf{V}^* \mid T)P(T)$ where $P(T = t) = 0.5$ for $t = 0, 1$, $P(\mathbf{V}^* \mid T{=}0)$ is the uniform distribution on $\mathcal{X}_{\mathbf{V}^*}$, while $P(\mathbf{V}^* \mid T{=}1) = P^*(\mathbf{V}^*)$.

Example 3.5 in Lauritzen [7] illustrates that, relative to some graph, there are distributions that are locally Markov but neither globally Markov nor able to be conditionally specified.

(L3.5) A distribution P for five variables X_U, X_W, X_X, X_Y, and X_Z where X_U and X_Z are independent, $P(X_U = 1) = P(X_Z = 1) = P(X_U = 0) = P(X_Z = 0) = 1/2$, and $X_W = X_U$, $X_Y = X_Z$, and $X_X = X_W X_Y$.

For the graph $\mathcal{G}_C$, U—W—X—Y—Z, Lauritzen [7] shows that this distribution is in $\mathcal{M}_L(\mathcal{G}_C)$ but not $\mathcal{M}_G(\mathcal{G}_C)$. The distribution is not in $\mathcal{M}_C(\mathcal{G}_C)$ because its Hamming graph is not connected.

3.3. $\exists\, \mathcal{G} : \mathcal{M}_G(\mathcal{G}) \setminus (\mathcal{M}_E(\mathcal{G}) \cup \mathcal{M}_C(\mathcal{G})) \neq \emptyset$

The following distribution from Lauritzen [7, Example 3.15] and Matúš and Studený [10] illustrates that, relative to some graph, there are distributions that are globally Markov but neither extended Markov nor able to be conditionally specified.

(L3.15) A distribution P for variables X_A, X_B, X_C, and X_D, where all variables have three possible values a, b, and c, and each of the following nine states have probability equal to 1/9:

$$\begin{array}{lll} (a,a,a,a), & (b,a,b,c), & (c,a,c,b), \\ (a,b,b,b), & (b,b,c,a), & (c,b,a,c), \\ (a,c,c,c), & (b,c,a,b), & (c,c,b,a). \end{array}$$

For the four-cycle graph $\mathcal{G}_4$ that contains the edges A — B, B — C, C — D and A — D, Lauritzen [7] and Matúš and Studený [10] show that this distribution is in $\mathcal{M}_G(\mathcal{G}_4)$ but not $\mathcal{M}_E(\mathcal{G}_4)$. The distribution is not in $\mathcal{M}_C(\mathcal{G}_4)$ because any two points of support differ in three variables, which implies that its Hamming graph is not connected.

3.4. $\exists\, \mathcal{G} : (\mathcal{M}_E(\mathcal{G}) \cap \mathcal{M}_C(\mathcal{G})) \setminus \mathcal{M}_F(\mathcal{G}) \neq \emptyset$

The following distribution from Lauritzen [7, Example 3.10] and Moussouris [9] illustrates that, relative to some graph, there are distributions that cannot be factored but are in the set of extended Markov distributions and can be conditionally specified.

(L3.10) A distribution P for four binary variables X_A, X_B, X_C, X_D with support only on the points

$$\begin{array}{llll} (0,0,0,0), & (1,0,0,0), & (1,1,0,0), & (1,1,1,0), \\ (0,0,0,1), & (0,0,1,1), & (0,1,1,1), & (1,1,1,1), \end{array}$$

and equal probability mass on each point. That is, for instance, $P(X_A = 0, X_B = 0, X_C = 0, X_D = 0) = 1/8$.

Consider the four-cycle graph $\mathcal{G}_4$, as used in the previous example. This distribution is not in $\mathcal{M}_F(\mathcal{G}_4)$ [7, p.37] but is in $\mathcal{M}_E(\mathcal{G}_4)$ [7, p.40]. It is straightforward to verify that the univariate conditionals of the distribution define a univariate Gibbs sampler with the correct stationary distribution. In particular, note that each point with support is Hamming distance one from two other points with support and that every point with support is reachable from every other point. Therefore, the distribution is in $\mathcal{M}_C(\mathcal{G}_4)$.

3.5. $\exists \mathcal{G} : \mathcal{M}_F(\mathcal{G}) \setminus \mathcal{M}_C(\mathcal{G}) \neq \emptyset$

We now come to new examples.

The following distribution illustrates that, relative to some graph, there are distributions that cannot be conditionally specified but factor.

(D1) A uniform distribution for two binary random variables X_A, X_B with support only on the points $(0, 0)$, $(1, 1)$.

Such a distribution is in $\mathcal{M}_F(A \text{ — } B)$ because the graph is complete and all distributions for two variables can be represented by the trivial factorization. The distribution is not in $\mathcal{M}_C(A \text{ — } B)$ because the two points of support have a Hamming distance of two, which means that there is no way for a single-site Gibbs sampler to visit both of the points in the support of the distribution (infinitely often) without visiting points not in the support.

3.6. $\exists \mathcal{G} : (\mathcal{M}_F(\mathcal{G}) \cap \mathcal{M}_C(\mathcal{G})) \setminus \mathcal{M}_C^+(\mathcal{G}) \neq \emptyset$

The following distribution illustrates that, relative to some graph, there are (non-strictly positive) distributions that factor and can be conditionally specified.

(D2) A distribution for two ternary random variables X_A, X_B each taking values in $\{0, 1, 2\}$. We will define the distribution to be uniform on the following seven combinations in the set:

$$(\{0, 1, 2\} \times \{0, 1, 2\}) \setminus \{(0, 2), (2, 0)\}$$

with support only on these combinations.

As argued for distribution (D1) above, such a distribution is in $\mathcal{M}_F(A \text{ — } B)$ and $\mathcal{M}_E(A \text{ — } B)$. In addition, the distribution is in $\mathcal{M}_C(A \text{ — } B)$ because the Hamming graph for this distribution is clearly connected, since it contains the following eight edges:

$$(i + 1, i) \text{ — } (i, i) \text{ — } (i, i + 1) \text{ for } i = 1, 2,$$
$$(i - 1, i) \text{ — } (i, i) \text{ — } (i, i - 1) \text{ for } i = 2, 3.$$

However, clearly the distribution is not positive.

3.7. $\exists \mathcal{G} : \mathcal{M}_C(\mathcal{G}) \setminus \mathcal{M}_G(\mathcal{G}) \neq \emptyset$

We now present a distribution that is in $\mathcal{M}_C(\mathcal{G})$ but not $\mathcal{M}_G(\mathcal{G})$ for some $\mathcal{G}$. The construction is based on Example 3.5 in [7] (see § 3.2 above).

The support of the distribution in this example does not permit it to be sampled from via a univariate Gibbs sampler, because there are only four support points:

$$(0,0,0,0,0), \quad (1,1,0,0,0), \quad (0,0,0,1,1), \quad (1,1,1,1,1),$$

and they are all separated by Hamming distance at least two.

To define our distribution, we add an additional variable T to the graph $U - W - X - Y - Z$, which is a neighbor of $\{U, W, X, Y, Z\}$.

(D3) The distribution over 6 binary variables $\{X_T, X_U, X_W, X_X, X_Y, X_Z\}$ where $P(X_T = 1) = 0.5$, $P(X_U, X_W, X_X, X_Y, X_Z | X_T = 0) = 2^{-5}$ (i. e., uniform over the states) and $P(X_U, X_W, X_X, X_Y, X_Z | X_T = 1) = P(X_U | X_T = 1) P(X_Z | X_T = 1) P(X_W | X_U, X_T = 1) P(X_Y | X_Z, X_T = 1) P(X_X | X_W, X_Y, X_T = 1)$ with

$$\begin{aligned} P(X_U = 1 \,|\, X_T = 1) &= 0.5, \\ P(X_Z = 1 \,|\, X_T = 1) &= 0.5, \\ P(X_W = a \,|\, X_U = a, X_T = 1) &= 1.0, \\ P(X_Y = a \,|\, X_Z = a, X_T = 1) &= 1.0, \\ P(X_X = ab \,|\, X_W = a, X_Y = b, X_T = 1) &= 1.0, \end{aligned}$$

where a, b are the possible values (0 or 1) of the variables.

The resulting distribution $P(X_T, X_U, X_W, X_X, X_Y, X_Z)$ has support on the 36 point space:

$$\{0\} \times \{0,1\}^5 \cup \{1\} \times \{(0,0,0,0,0), (1,1,0,0,0), (0,0,0,1,1), (1,1,1,1,1)\}.$$

It is simple to see that there is now a path between any two points in the support space such that each pair of adjacent points has Hamming distance one. The Markov chain resulting from the Gibbs sampler will be irreducible and ergodic, and hence will have a unique limiting distribution. This limiting distribution will be as described above.

It now only remains to observe that the distribution obeys all of the conditional independence relations required by the Local Markov property applied to the graph:

$$\begin{array}{lcll} X_U & \perp\!\!\!\perp & X_X, X_Y, X_Z & | \quad X_W, X_T, \\ X_W & \perp\!\!\!\perp & X_Y, X_Z & | \quad X_U, X_X, X_T, \\ X_X & \perp\!\!\!\perp & X_U, X_Z & | \quad X_W, X_Y, X_T, \\ X_Y & \perp\!\!\!\perp & X_U, X_W & | \quad X_X, X_Z, X_T, \\ X_Z & \perp\!\!\!\perp & X_U, X_W, X_X & | \quad X_Y, X_T, \end{array}$$

but does not obey the global property because

$$X_W \not\perp\!\!\!\perp X_Y \mid X_X, X_T.$$

3.8. $\exists \mathcal{G} : (\mathcal{M}_C(\mathcal{G}) \cap \mathcal{M}_G(\mathcal{G})) \setminus \mathcal{M}_E(\mathcal{G}) \neq \emptyset$

We apply a similar construction to that used in Section 3.7 to Example 3.15 in Lauritzen [7] (see § 3.3).

(D4) The distribution over ternary variables X_A, X_B, X_C, X_D and binary variable X_T, where $P(X_T = 0) = P(X_T = 1) = 0.5$, $P(X_A, X_B, X_C, X_D | X_T = 0) = 3^{-4}$, and $P(X_A, X_B, X_C, X_D | X_T = 1)$ is the distribution specified in **(L3.15)**.

The resulting distribution is globally Markov with respect to the graph $\mathcal{G}_5$ that contains only the edges $A - B$, $B - C$, $C - D$, $A - D$, $T - A$, $T - B$, $T - C$ and $T - D$. This follows because, given $X_T = 0$, the variables X_A, X_B, X_C, X_D are marginally independent. Also, we have

$$X_A \perp\!\!\!\perp X_C \mid X_B, X_D, X_T = 1$$

because the conditional distribution over (X_A, X_C) is degenerate. Likewise, we have

$$X_B \perp\!\!\!\perp X_D \mid X_A, X_C, X_T = 1.$$

To demonstrate that the distribution is not in $M_E(\mathcal{G}_5)$, we follow the argument given in Lauritzen [7, p.42]. First note that each of the pairs (X_A, X_B), (X_B, X_C), (X_C, X_D), (X_A, X_D) are uniformly distributed under P given $T = 0$ and given $T = 1$. Hence, each of the triples

$$(X_A, X_B, X_T), (X_B, X_C, X_T),$$
$$(X_C, X_D, X_T), (X_A, X_D, X_T)$$

are uniformly distributed on the 18 element space: $\{a, b, c\} \times \{a, b, c\} \times \{0, 1\}$. Hence P has the same clique marginals as the uniform distribution Q defined by taking $Q(X_A, X_B, X_C, X_D, X_T) = 2^{-1}3^{-4}$. Q is clearly in $M_E(\mathcal{G}_5)$. It then follows by Lemma 3.14 in Lauritzen [7] that $P \notin M_E(\mathcal{G}_5)$, because distributions in $M_E(\mathcal{G}_5)$ are uniquely identified via their clique marginals.

Finally, we observe that $P \in M_C(\mathcal{G}_5)$ because every point with support is reachable from every other point via a sequence of points which are Hamming distance one apart.

3.9. $\exists \mathcal{G} : \mathcal{M}_E(\mathcal{G}) \setminus (\mathcal{M}_F(\mathcal{G}) \cup \mathcal{M}_C(\mathcal{G})) \neq \emptyset$

The following distribution illustrates that, relative to some graph, there are distributions in the set of extended Markov distributions that do not factor and cannot be conditionally specified.

(D5) A distribution P over 5 binary variables $\{X_A, X_B, X_C, X_D, X_T\}$ with support only on the 16 points:

$$(0,0,0,0,0), \quad (1,0,0,0,0), \quad (1,1,0,0,0), \quad (1,1,1,0,0),$$
$$(0,0,0,1,0), \quad (0,0,1,1,0), \quad (0,1,1,1,0), \quad (1,1,1,1,0),$$

$$(0,1,0,0,1), \quad (0,1,0,1,1), \quad (1,1,0,1,1), \quad (1,0,0,1,1),$$
$$(0,1,1,0,1), \quad (0,0,1,0,1), \quad (1,0,1,0,1), \quad (1,0,1,1,1),$$

with equal probability mass on each point. That is, for instance, $P(X_A = 0, X_B = 0, X_C = 0, X_D = 0, X_T = 0) = 1/16$.

We again consider graph $\mathcal{G}_5$. Every point in the second set of eight support points is Hamming distance greater than or equal to two from every other point in the first set of eight points in the support. Thus the distribution is not in $\mathcal{M}_C(\mathcal{G}_5)$. Next we show that the distribution is not in $\mathcal{M}_E(\mathcal{G}_5)$ using the same technique as Lauritzen does for his Example 3.10 [7, p.38]. Aiming at a contradiction, we assume that the distribution factors with clique potentials $\psi_{ABT}(\cdot), \psi_{BCT}(\cdot), \psi_{CDT}(\cdot), \psi_{ADT}(\cdot)$. Then

$$1/16 = P(0,0,0,0,0) = \psi_{ABT}(0,0,0)\psi_{BCT}(0,0,0)\psi_{CDT}(0,0,0)\psi_{ADT}(0,0,0).$$

But also

$$0 = P(0,0,1,0,0) = \psi_{ABT}(0,0,0)\psi_{BCT}(0,1,0)\psi_{CDT}(1,0,0)\psi_{ADT}(0,0,0).$$

From which it follows that

$$\psi_{BCT}(0,1,0)\psi_{CDT}(1,0,0) = 0.$$

Since

$$1/16 = P(0,0,1,1,0) = \psi_{ABT}(0,0,0)\psi_{BCT}(0,1,0)\psi_{CDT}(1,1,0)\psi_{ADT}(0,1,0)$$

it is the case that $\psi_{BCT}(0,1,0) \neq 0$. It follows that $\psi_{CDT}(1,0,0) = 0$. However, this contradicts the fact that

$$1/16 = P(1,1,1,0,0) = \psi_{ABT}(1,1,0)\psi_{BCT}(1,1,0)\psi_{CDT}(1,0,0)\psi_{ADT}(1,0,0).$$

It remains to show that $P \in \mathcal{M}_E(\mathcal{G}_5)$. We use the same proof technique as Lauritzen [7, p. 40] and describe a sequence of distributions $P_n \in \mathcal{M}_F(\mathcal{G}_5)$ whose limit as $n \to \infty$ is distribution **(D5)**:

$$P_n(a,b,c,d,t) = \frac{n^{t(1-(ab+bc+cd-ad-b-c+1))+(1-t)(ab+bc+cd-ad-b-c+1)}}{16+16n}.$$

The expression $f(a,b,c,d) \equiv ab+bc+cd-ad-b-c+1$, as in Example 3.13 of Lauritzen [7, p. 40], is equal to one for points that agree in the first four coordinates with a point in the first eight support points and is zero otherwise. Thus, $(1-t) \times f(\cdot)$ is one for the first set of eight support points and zero otherwise. Similarly, $t \times (1 - f(\cdot))$ is one for the second set of eight support points and zero otherwise. Therefore, the expression in the exponent of the numerator is one for each point in the support and zero otherwise which justifies the normalizing constant. Each distribution P_n factors according to $\mathcal{G}_5$ because each product of variables that occurs in the numerator is a subset of the cliques of the graph.

4. SUMMARY AND DISCUSSION

We have shown that all definitions agree for strictly positive distributions, but disagree as described in Figure 1b for general distributions. In closing, we make several observations and raise questions for future work.

$\mathcal{M}_C(\mathcal{G})$ is not always convex. As an example, consider the domain $\{X_A, X_B, X_C, X_D, X_E\}$, the graph $\mathcal{G}_5$, and the two distributions:

(D5a): The distribution with equal support on only the points

$$(0,0,0,0,0),\quad (1,0,0,0,0),\quad (1,1,0,0,0),\quad (1,1,1,0,0),$$
$$(0,0,0,1,0),\quad (0,0,1,1,0),\quad (0,1,1,1,0),\quad (1,1,1,1,0);$$

(D5b): The distribution with equal support on only the points

$$(0,1,0,0,1),\quad (0,1,0,1,1),\quad (1,1,0,1,1),\quad (1,0,0,1,1),$$
$$(0,1,1,0,1),\quad (0,0,1,0,1),\quad (1,0,1,0,1),\quad (1,0,1,1,1).$$

Whereas both **(D5a)** and **(D5b)** are in $\mathcal{M}_C(\mathcal{G}_5)$, their equal mixture, distribution **(D5)**, is not (see Section 3.9).

The pointwise limit of a set of distributions in $\mathcal{M}_C(\mathcal{G})$ is not necessarily in $\mathcal{M}_C(\mathcal{G})$. For example, consider domain $\{X_A, X_B\}$, the complete graph $\mathcal{G}$, and the set of distributions parameterized by $0 < \epsilon < 1$ such that $P(0,0) = P(1,1) = (1-\epsilon)/2$ and $P(0,1) = P(1,0) = \epsilon/2$. All such distributions are in $\mathcal{M}_C(\mathcal{G})$, but their limit, as ϵ goes to zero, is not. It follows that the maximum likelihood estimate for the limit distribution lies on the boundary of $\mathcal{M}_C(\mathcal{G})$ — that is, the maximum likelihood estimate for this distribution does not exist under $\mathcal{M}_C(\mathcal{G})$.

We noted earlier that the class $\mathcal{M}_C(\mathcal{G})$ can be extended to include k-site Gibbs sampling. We should note that such an extension (regardless of k) does not yield a class equal to $\mathcal{M}_L(\mathcal{G})$. For example, consider the distribution with two support points $(1,1,\ldots,1)$ and $(0,0,\ldots,0)$ and any non-complete graph without isolated vertices (i. e. every vertex has a neighbor). The distribution satisfies the local Markov condition for the graph but is not in $\mathcal{M}_C(\mathcal{G})$, because all variables would need to change at the same time.

Finally, we note that for decomposable graphs $\mathcal{G}$, Lauritzen [7] shows that $\mathcal{M}_E(\mathcal{G}) = \mathcal{M}_F(\mathcal{G})$. Because the graphs corresponding to examples **(D3)** and **(D4)** are decomposable, however, $\mathcal{M}_C(\mathcal{G}) \neq \mathcal{M}_E(\mathcal{G})$ for this class of graphs. An interesting set of questions is whether there are non-trivial classes of graphs and/or distributions for which $\mathcal{M}_C(\mathcal{G})$ is equal to one or more of $\mathcal{M}_L(\mathcal{G})$, $\mathcal{M}_G(\mathcal{G})$, $\mathcal{M}_E(\mathcal{G})$, or $\mathcal{M}_F(\mathcal{G})$.

5. ACKNOWLEDGMENTS

We thank Steffen Lauritzen for useful discussions and the reviewers and editor for helpful and detailed comments on earlier versions of this manuscript. Thomas Richardson was supported by the U.S. National Science Foundation grant CNS-0855230 and U.S. National Institutes of Health grant R01 AI032475.

(Received November 2, 2011)

REFERENCES

[1] A. Agresti: Categorical Data Analysis. Wiley and Sons, New York 1990.

[2] B. C. Arnold, E. Castillo, and J. Sarabia: Conditional Specification of Statistical Models. Springer-Verlag, New York 1999.

[3] M. S. Bartlett: An Introduction to Stochastic Processes. University Press, Cambridge 1955.

[4] J. Besag: Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. Ser. B *36* (1974), 192–236.

[5] D. Brook: On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbor systems. Biometrika *51* (1964), 481–483.

[6] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie: Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res. *1* (2000), 49–75.

[7] S. L. Lauritzen: Graphical Models. Clarendon Press, Oxford 1996.

[8] P. Lévy: Chaînes doubles de Markoff et fonctions aléatoires de deux variables. C. R. Académie des Sciences, Paris *226* (1948), 53–55.

[9] J. Moussouris: Gibbs and Markov random systems with constraints. J. Statist. Phys. *10* (1974), 11–33.

[10] F. Matúš and M. Studený: Conditional independence among four random variables I. Combin. Probab. Comput. *4* (1995), 269–78.

[11] J. R. Norris: Markov Chains. Cambridge University Press, Cambridge 1997.

[12] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu: Graphical Models via Generalized Linear Models. In: Advances in Neural Information Processing Systems *25* (2013), Cambridge.

David Heckerman, Microsoft Research, 1100 Glendon Ave, PH1, Los Angeles, CA 90024. U. S. A.
e-mail: heckerma@microsoft.com

Christopher Meek, Microsoft Research, One Microsoft Way, Redmond, WA 98056. U. S. A.
e-mail: meek@microsoft.com

Thomas Richardson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195. U. S. A.
e-mail: thomasr@u.washington.edu