The Annals of Statistics
2016, Vol. 44, No. 3, 1010–1037
DOI: 10.1214/15-AOS1396
© Institute of Mathematical Statistics, 2016

A NEW PRIOR FOR DISCRETE DAG MODELS WITH A RESTRICTED SET OF DIRECTIONS

BY HÉLÈNE MASSAM[1] AND JACEK WESOŁOWSKI[2]

York University and Politechnika Warszawska

In this paper, we first develop a new family of conjugate prior distributions for the cell probability parameters of discrete graphical models Markov with respect to a set $\mathcal{P}$ of moral directed acyclic graphs with skeleton a given decomposable graph G. This family, which we call the $\mathcal{P}$-Dirichlet, is a generalization of the hyper Dirichlet given in [*Ann. Statist.* **21** (1993) 1272–1317]: it keeps the directed strong hyper Markov property for every DAG in $\mathcal{P}$ but increases the flexibility in the choice of its parameters, that is, the hyper parameters.

Our second contribution is a characterization of the $\mathcal{P}$-Dirichlet, which yields, as a corollary, a characterization of the hyper Dirichlet and a characterization of the Dirichlet also. Like the characterization of the Dirichlet given in [*Ann. Statist.* **25** (1997) 1344–1369], our characterization of the $\mathcal{P}$-Dirichlet is based on local and global independence of the probability parameters and also a separability property explicitly defined here but implicitly used in that paper through the choice of two particular DAGs. Another advantage of our approach is that we need not make the assumption of the existence of a positive density function. We use the method of moments for our proofs.

1. Introduction. The Dirichlet distribution and distributions derived from the Dirichlet are essential ingredients of Bayesian inference in the analysis of discrete data. For high-dimensional data, Dirichlet-type distributions are often used in conjunction with graphical models. Let $V = \{1, \ldots, d\}$ be a finite set indexing the variables. A graphical model for the discrete random variable $\mathbf{X} = (X_v, v \in V)$ is a statistical model where the dependences between X_v taking their value in the finite set $\mathcal{I}_v, v \in V$ are expressed by means of a graph G. We will assume here that the data is gathered under the form of a d-dimensional contingency table and that the cell counts follow a multinomial distribution with cell probabilities $(p(\underline{i}) = \mathbb{P}(\mathbf{X} = \underline{i}), \underline{i} \in \mathcal{I})$ where $\mathcal{I} = \times_{v \in V} \mathcal{I}_v$ is the set of cells in the contingency table.

If the conditional independences between the variables can be expressed by means of a directed acyclic graph (henceforth abbreviated DAG), the usual priors

Received April 2015; revised September 2015.

[1]Support from an NSERC Discovery grant.

[2]Supported in part by NCN Grant 2012/05/B/ST1/00554.

MSC2010 subject classifications. 62H17, 62F15, 62E99.

Key words and phrases. Bayesian learning, directed strong hyper Markov, conjugate priors, hyper Dirichlet distribution, characterization, local and global independence.

are independent Dirichlet prior distributions on

$$\big(\mathbb{P}(X_v = i_v | \mathbf{X}_{\mathfrak{p}_v} = \underline{i}_{\mathfrak{p}_v}), i_v \in \mathcal{I}_v\big), \qquad \underline{i}_{\mathfrak{p}_v} \in \mathcal{I}_{\mathfrak{p}_v}, v \in V, \tag{1.1}$$

where, for $A \subset V$, $\mathcal{I}_A = \times_{v \in A} \mathcal{I}_v$, $\mathfrak{p}_v$ denotes the set of parents of vertex v in the DAG and $\underline{i}_{\mathfrak{p}_v} = (i_l, l \in \mathfrak{p}_v)$ denotes the $\mathfrak{p}_v$-marginal cell. This means that the conditional probability parameters associated with each node are mutually independent (global independence) and for a given node, the parameters associated with various instances of its parents are also mutually independent (local parameter independence).

A characterization of the Dirichlet distribution through local and global independence using two particular complete DAGs with vertex orders $\mathfrak{p} = \{1, 2, \ldots, d-1, d\}$ and $\mathfrak{p}' = \{d, 1, 2, \ldots, d-1\}$ is given in Theorem 3 of [9]. This result has important practical ramifications. Indeed, if we request that a node with the same parents in two distinct DAGs has identical prior distribution on the parameters associated with this node in both structures (this is called parameter modularity) and if we also request local and global independence of the parameters for $\mathfrak{p}$ and $\mathfrak{p}'$, then the priors on the variables in (1.1) must all be derived from one single Dirichlet distribution $\mathrm{Dir}(\alpha(\underline{i}), \underline{i} \in \mathcal{I})$ on $(p(\underline{i}), \underline{i} \in \mathcal{I})$. The hyper parameters $\alpha(\underline{i})$ of such a Dirichlet are of the form

$$\alpha(\underline{i}) = \alpha\theta(\underline{i}), \qquad \underline{i} \in \mathcal{I}, \tag{1.2}$$

where $\theta(\underline{i})$ is the expected value of $p(\underline{i})$ and α, called the *equivalent sample size*, will be added to the actual total cell count $\sum_{\underline{i} \in \mathcal{I}} n(\underline{i})$ in posterior inference. Most of the time, in practice and in the absence of prior expert information, we take $\theta(\underline{i}) = \frac{1}{|\mathcal{I}|}$ where $|\mathcal{I}|$ is the total number of cells. This implies that our flexibility in the choice of the hyper parameters is restricted to the choice of α.

If the conditional independences between the variables can be expressed by means of a decomposable undirected graph G, then the multinomial distribution on the cell counts is Markov with respect to G and the Diaconis–Ylvisaker [7] prior on $(p(\underline{i}), \underline{i} \in \mathcal{I})$ is the hyper Dirichlet defined in [6]. A simple calculation [see equations (2.10)] shows that the parameters of the hyper Dirichlet have an interpretation of the type (1.2) and we are therefore faced with the same lack of flexibility in the choice of the hyper parameters. Moreover (see Section 4) if, for any DAG with skeleton G, Markov equivalent to G, we make the change of variable from $(p(\underline{i}), \underline{i} \in \mathcal{I})$, to the variables in (1.1), then, for each of these DAGs, the distribution induced from the hyper Dirichlet also possesses the property of local and global independence. One can then be led to think that local and global independence in all possible directions imposes strict restrictions on the choice of the hyper parameters of the prior on $(p(\underline{i}), \underline{i} \in \mathcal{I})$.

In this paper, we therefore consider a restricted family $\mathcal{P}$ of moral DAGs with skeleton a decomposable graph G and Markov equivalent to G. For the set of multinomial models Markov with respect to any of the DAGs in $\mathcal{P}$, we build a

new family of prior distributions for the parameters $(p(\underline{i}), \underline{i} \in \mathcal{I})$. This family of priors, called the $\mathcal{P}$-Dirichlet family, is conjugate, possesses the property of local and global independence and also the directed strong hyper Markov property. Using these priors will thus facilitate posterior inference. Moreover, because of the restrictions on the collection of DAGs defined by $\mathcal{P}$, the $\mathcal{P}$-Dirichlet offers more hyper parameter flexibility than the hyper Dirichlet. We show through an example how this flexibility gives us more freedom when inserting prior expert knowledge in the prior distribution. The development and study of this new family of priors is the first contribution of this paper.

The second contribution is a characterization (Theorem 6.1) of the $\mathcal{P}$-Dirichlet family through local and global independence. As corollaries, we obtain a characterization of the hyper Dirichlet and also the characterization of the Dirichlet as given by Geiger and Heckerman in [9] but without assuming the existence of a density for the vector of random probabilities $\mathbf{p} = (\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$.

The remainder of this paper is organized as follows. The $\mathcal{P}$-Dirichlet family of priors for $\mathbf{p}$, is defined in Section 3. In Section 4, Theorem 4.1, we derive a general formula for the moments of the $\mathcal{P}$-Dirichlet which will be used in Section 6 for the proof of its characterization and in Section 5 for the computation of the posterior mean of any $\mathbf{p}(\underline{i})$ or other probabilities derived from the $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$. In Section 5, we show that the dimension of the $\mathcal{P}$-Dirichlet family is always larger than that of the hyper Dirichlet with the same skeleton and we illustrate the interpretation and flexibility of the hyper parameters with an example. We prove that the $\mathcal{P}$-Dirichlet is conjugate and has the directed strong hyper Markov property. In Section 6, we give its characterization, and thus that of the hyper Dirichlet and the Dirichlet. The next section is devoted to preliminaries.

2. Preliminaries.

2.1. *Graph theoretical notions.* We first recall some well-known notions. Let $G = (V, E)$ be an undirected graph, where $V = \{1, \ldots, d\}$ is a finite set of vertices and E is the set of edges. A graph is said to be decomposable if it does not have any chordless cycle of length greater than or equal to 4. A subset D of V is said to be complete if all vertices of D are linked to each other by an edge. A clique is a complete subset maximal with respect to inclusion. A sequence $B_1, \ldots, B_K$ of subsets of V is said to be perfect if, for

$$H_j = \bigcup_{l=1}^{j} B_l, \qquad S_j = B_j \cap H_{j-1} \quad \text{and}$$
$$R_j = B_j \setminus H_{j-1} = B_j \setminus S_j, \qquad j = 1, \ldots, K,$$

(with the convention that $S_1 = \varnothing$) the following two conditions are fulfilled: the sets S_j are complete and for any $j > 1$, there exists an $i < j$ such that $H_{j-1} \cap B_j \subset B_i$. In this paper, we consider in particular perfect orderings $C_1, \ldots, C_K$ of the

cliques of a decomposable graph G. Let $\mathcal{C}$ denote the set of cliques of G. It is well known that, for G decomposable, the set of separators $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$ associated with a perfect ordering of the cliques is independent of the perfect ordering considered and it is the set of minimal separators for the graph G. For a vertex $v \in V$, we write $nb(v)$ to denote the set of neighbours of v in G. A numbering $v_1, \ldots, v_d$ of the vertices of G is said to be perfect if the sequence of sets

$$B_j = \{v_1, \ldots, v_j\} \cap \big(v_j \cup nb(v_j)\big)$$

is perfect. It can be shown (Proposition 2.17 in [14]) that a graph G is decomposable if and only if its set of cliques admits a perfect order and also if and only if its set of vertices admits a perfect numbering. A DAG is said to be moral if when v and v' are both parents of w, then there is an edge between v and v'. In this paper, we will deal with moral DAGs having a given decomposable graph G as a skeleton. We note that not all DAGs with skeleton a decomposable graph are moral. The simplest counter example is of course the DAG $1 \to 2 \leftarrow 3$. However, if a DAG is moral, its skeleton is a decomposable graph: otherwise, there would be a directed cycle $(v_i, v_{i+1}, \ldots, v_{i+k-1}, v_{i+k} = v_i)$ of length $k \geq 4$ without a chord. Since a DAG is acyclic, there exists $v_j \in \{v_i, \ldots, v_{i+k-1}\}$ with $v_{j-1} \to v_j \leftarrow v_{j+1}$ and since the DAG is moral, there must be an edge between v_{j-1} and v_{j+1} contradicting the fact that $(v_i, v_{i+1}, \ldots, v_{i+k-1}, v_{i+k} = v_i)$ is a cycle without chord.

We next introduce some new notions. Note than any DAG with skeleton G is in one-to-one correspondence with the function $\mathfrak{p} : V \to 2^V$ which describes the parents of each vertex or with the function $\mathfrak{c} : V \to 2^V$ which describes the children of each vertex. We will write $\mathfrak{p}_v$ and $\mathfrak{c}_v$ for the set of parents and children of $v \in V$, respectively. The set of descendants of a vertex v denotes the set of children of v, children of children and so on. The set of nondescendants of v will be denoted $\mathfrak{nd}_v$. Additionally, let $\mathfrak{q}_v = \mathfrak{p}_v \cup \{v\}$ for any $v \in V$ and let $\mathfrak{q} : V \to 2^V$ denote the corresponding mapping. We will now define some additional graph-theoretical notions needed in this work.

DEFINITION 2.1. Given a parent function $\mathfrak{p}$, a perfect order $o = (C_1, \ldots, C_K)$ of the cliques is said to be $\mathfrak{p}$-perfect if for any $l = 1, \ldots, K$ there exists a $v \in C_l \setminus S_l$ such that $S_l = \mathfrak{p}_v$.

A perfect ordering of the cliques of a decomposable graph G induces a perfect numbering of the vertices by numbering the vertices in C_1, then in $R_2, R_3, \ldots, R_K$ successively (see Proposition 2.12 in [14]). However, a perfect numbering of the vertices does not necessarily induce a perfect ordering of the cliques. In the next lemma, we start not with a numbering of the vertices but with the parent function $\mathfrak{p}$ of a moral DAG with skeleton G decomposable and we show that given $\mathfrak{p}$, we can always find a $\mathfrak{p}$-perfect ordering of the cliques $C_1, \ldots, C_K$.

LEMMA 2.1. *Let $\mathfrak{p}$ be a parent function of a moral DAG with skeleton a decomposable graph $G = (V, E)$. Then we have the following*:

(1) *There exists a unique vertex* v_0 *such that* $\mathfrak{p}_{v_0} = \varnothing$.

(2) *There exists a* $\mathfrak{p}$*-perfect ordering* $(C_1, \ldots, C_K)$ *of cliques of* G *such that* $v_0 \in C_1$.

(3) *There exists a perfect ordering of the vertices defined as follows*:

$$v = \begin{cases} v_{1,1}, & \text{if } v = v_0, \\ v_{1,i}, & \text{if } \mathfrak{p}_v = \mathfrak{q}_{v_{1,i-1}}, i = 2, \ldots, c_1, \\ v_{j,s_j+1}, & \text{if } \mathfrak{p}_v = S_j, j = 2, \ldots, K, \\ v_{j,l}, & \text{if } \mathfrak{p}_v = \mathfrak{q}_{v_{j,l-1}}, l = s_j + 2, \ldots, c_j, j = 2, \ldots, K, \end{cases} \tag{2.1}$$

where c_j *and* s_j *denote the cardinality of* C_j *and* S_j, $j = 1, \ldots, K$, *respectively.*

The proof is given in Section 2.1 of the supplementary material [15]. The unique vertex $v_0 = v_{1,1}$ such that $\mathfrak{p}_v = \varnothing$ is called the source for $\mathfrak{p}$. This $\mathfrak{p}$-perfect order of cliques is not necessarily unique: see Example 3.1 "continued" in Section 4.2.1 of this paper and of the supplementary material [15]. To obtain it, it suffices to follow the construction outlined in the proof, that is, choose $v_{1,1} = v_0$, then $v_{1,i}$ such that $\mathfrak{p}_{v_{1,i}} = \mathfrak{q}_{v_{1,i-1}}$ for $i = 2, \ldots, c_1$ where $\mathfrak{q}_{v_{1,c_1}}$ is a maximal complete subset which is chosen as C_1. Then choose v_{2,s_2+1} such that $\mathfrak{p}_{v_{2,s_2+1}} = S_2$ and so on. This construction is illustrated in Example 2.1 of the supplementary material [15]. Numbering (2.1) is clearly a perfect ordering of the vertices obtained by following $C_1, R_2, \ldots, R_K$.

We now proceed to introduce three sets $\mathfrak{P}$, $\mathfrak{Q}$ and $\mathfrak{R}$ of subsets of V. Sets $\mathfrak{P}$ and $\mathfrak{Q}$ generalize respectively $\mathcal{S}$ and $\mathcal{C}$ while $\mathfrak{R}$ is a new collection of complete subsets.

LEMMA 2.2. *Consider a moral DAG with skeleton* G *defined by a parent function* $\mathfrak{p}$. *Then*

$$\mathcal{C} \subset \mathfrak{q}(V) \quad \text{and} \quad \mathcal{S} \subset \mathfrak{p}(V).$$

Moreover

$$\mathfrak{p}(V) \setminus \mathcal{S} = \mathfrak{q}(V) \setminus \mathcal{C}.$$

PROOF. Let $o = (C_1, \ldots, C_K)$ be a $\mathfrak{p}$-perfect order of the cliques. From (2.1), we have $\mathfrak{p}_{v_{l,s_l+1}} = S_l$ and $\mathfrak{q}_{v_{l,c_l}} = C_l$, $l = 1, \ldots, K$ so that the first statement of the lemma is proved.

For the second part, note that for $l = 1, \ldots, K$ and $j = s_l + 2, \ldots, c_l$, we have $\mathfrak{q}_{v_{l,j-1}} = \mathfrak{p}_{v_{l,j}} \notin \mathcal{C} \cup \mathcal{S}$. That is $\mathfrak{p}(V) \setminus \mathcal{S} = \mathfrak{q}(V) \setminus \mathcal{C} = \bigcup_{l=1}^{K} \bigcup_{j=s_l+2}^{c_l} \mathfrak{p}_{v_{l,j}}$. □

Note that if $v, w \in V$ are distinct then $\mathfrak{q}_v$ and $\mathfrak{q}_w$ are also distinct, but $\mathfrak{p}_v$ and $\mathfrak{p}_w$ may be the same. Consequently, a given set can appear in $\mathfrak{p}(V)$ more than once. For example, $S \in \mathcal{S}$ can be a multiple separator. However, because of the equation

in Lemma 2.2, such a set can appear at most once in $\mathfrak{p}(V) \setminus \mathcal{S}$, (see DAG $\mathfrak{p}$ in Example 2.1 of the supplementary material [15]).

Denote

$$\mathfrak{R}_{\mathfrak{p}} := \mathfrak{p}(V) \setminus \mathcal{S} = \mathfrak{q}(V) \setminus \mathcal{C}.$$

Let $\mathcal{P}$ be a family of moral DAGs with the same skeleton G. Consider the following sets:

$$\mathfrak{Q} = \bigcap_{\mathfrak{p} \in \mathcal{P}} \mathfrak{q}(V) \supset \mathcal{C} \quad \text{and} \quad \mathfrak{P} = \bigcap_{\mathfrak{p} \in \mathcal{P}} \mathfrak{p}(V) \supset \mathcal{S}. \tag{2.2}$$

By Lemma 2.2, it follows that $\mathfrak{Q} = \mathcal{C} \cup \mathfrak{R}$ and $\mathfrak{P} = \mathcal{S} \cup \mathfrak{R}$, where $\mathfrak{R} = \bigcap_{\mathfrak{p} \in \mathcal{P}} \mathfrak{R}_{\mathfrak{p}}$.

DEFINITION 2.2. The separator $S \in \mathcal{S}$ and the clique $C \in \mathcal{C}$ are said to be paired by the perfect ordering $o = (C_1, \ldots, C_K)$ of the cliques if

$$\exists l \in \{1, \ldots, K\}: \qquad S = S_l, \qquad C = C_l$$

and we write $S \overset{o}{\to} C$.

As we shall see in Example 3.1 "continued" in Section 4.2.1, to one $\mathfrak{p}$ may correspond more than one $\mathfrak{p}$-perfect order. In the sequel, for a family $\mathcal{P}$ of moral DAGs, we will use collections

$$O_{\mathcal{P}} = (o_{\mathfrak{p}}, \mathfrak{p} \in \mathcal{P} : o_{\mathfrak{p}} \text{ is } \mathfrak{p}\text{-perfect}) \tag{2.3}$$

of perfect orders where for each $\mathfrak{p} \in \mathcal{P}$, one order $o_{\mathfrak{p}}$ is selected.

LEMMA 2.3. *For a family $\mathcal{P}$ of moral DAGs with skeleton G, let $O_{\mathcal{P}} = (o_{\mathfrak{p}}, \mathfrak{p} \in \mathcal{P})$ be a given collection of $\mathfrak{p}$-perfect orders of cliques. For any clique C, let $\mathfrak{R}_C$ denote a family of these elements of $\mathfrak{R}$ which are contained in C. If there exists $\mathfrak{p} \in \mathcal{P}$ such that $\mathcal{S} \ni S \overset{o_{\mathfrak{p}}}{\to} C$ then all the elements $Q_1^C, \ldots, Q_{j_C-1}^C \in \mathfrak{R}_C$ (it may be empty) can be numbered as follows*:

$$S =: Q_{j_C}^C \subsetneq Q_{j_C-1}^C \subsetneq Q_{j_C-2}^C \subsetneq \cdots \subsetneq Q_2^C \subsetneq Q_1^C \subsetneq Q_0^C := C. \tag{2.4}$$

PROOF. It follows immediately from (2.1) that possible sets Q_i^C, $1 \le i \le j_C - 1$, from $\mathfrak{R}_C$ are of the form $\mathfrak{q}_{v_{l,k}} = \mathfrak{p}_{v_{l,k+1}}$, $k = s_l + 1, \ldots, c_l - 1$. We define $Q_{j_C}^C = S = \mathfrak{p}_{v_{l,s_l+1}} \subsetneq \mathfrak{q}_{v_{l,s_l+1}}$ and $Q_0^C = C = \mathfrak{q}_{v_{l,c_l}} \supsetneq \mathfrak{q}_{v_{l,c_l-1}}$, $l = 1, \ldots, K$ but neither $Q_{j_C}^C$ nor Q_0^C are in $\mathfrak{R}$. □

2.2. *Markov properties and the hyper Dirichlet.* Let $\mathbf{X} = (X_v, v \in V)$ be a discrete random vector with variables X_v taking values in the finite set $\mathcal{I}_v$, $v \in V$. Let $\mathcal{I} = \times_{v\in V}\mathcal{I}_v$ be the set of cells for the corresponding contingency table. The cell probabilities are

$$p(\underline{i}) := \mathbb{P}(\mathbf{X} = \underline{i}) = \mathbb{P}(X_v = i_v, v \in V), \qquad \underline{i} \in \mathcal{I}.$$

We are given a sample $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ from this distribution. The cell counts $(N(\underline{i}), \underline{i} \in \mathcal{I})$ taking values $(n(\underline{i}), \underline{i} \in \mathcal{I})$ are assumed to follow a multinomial distribution $\mathcal{M}(M, p(\underline{i}), \underline{i} \in \mathcal{I})$. Given an undirected graph G, the distribution of $\mathbf{X}$ is said to be Markov with respect to G if X_i and X_j are independent given $\mathbf{X}_{V\setminus\{i,j\}}$ whenever the pair $\{i, j\}$ does not belong to E. Given a DAG $\mathcal{G}$ and its corresponding parent function $\mathfrak{p}$, we say that $\mathbf{X}$ is Markov with respect to $\mathcal{G}$ if, for any $v \in V$, $X_v \perp \mathbf{X}_{\mathfrak{nd}_v} | \mathbf{X}_{\mathfrak{p}_v}$.

In this paper, we consider only moral DAGs $\mathcal{G}$ with a decomposable skeleton G which by Lemma 3.21 of [14] encode the same conditional independences as G. Then, by Proposition 3.28 of [14], we have the following two equivalent factorizations of $p(\underline{i}), \underline{i} \in \mathcal{I}$:

$$\frac{\prod_{l=1}^{K} \mathbb{P}(\mathbf{X}_{C_l} = \underline{i}_{C_l})}{\prod_{l=2}^{K} \mathbb{P}(\mathbf{X}_{S_l} = \underline{i}_{S_l})} = p(\underline{i}) = \prod_{v\in V} \mathbb{P}(X_v = i_v | \mathbf{X}_{\mathfrak{p}_v} = \underline{i}_{\mathfrak{p}_v}), \tag{2.5}$$

according to G and $\mathcal{G}$, respectively.

Since the cell counts $(N(\underline{i}), \underline{i} \in \mathcal{I})$ follow a multinomial $\mathcal{M}(M, p(\underline{i}), \underline{i} \in \mathcal{I})$ distribution, the density of $(N(\underline{i}), \underline{i} \in \mathcal{I})$ with respect to the counting measure is proportional to

$$\begin{aligned}\prod_{\underline{i}\in\mathcal{I}} p(\underline{i})^{n(\underline{i})} &= \frac{\prod_{l=1}^{K} \prod_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}} \mathbb{P}(\mathbf{X}_{C_l} = \underline{i}_{C_l})^{n(\underline{i}_{C_l})}}{\prod_{l=2}^{K} \prod_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}} \mathbb{P}(\mathbf{X}_{S_l} = \underline{i}_{S_l})^{n(\underline{i}_{S_l})}} \\ &= \prod_{v\in V} \prod_{\underline{i}_{\mathfrak{p}_v}\in\mathcal{I}_{\mathfrak{p}_v}} \prod_{i_v\in\mathcal{I}_v} \mathbb{P}(X_v = i_v | \mathbf{X}_{\mathfrak{p}_v} = \underline{i}_{\mathfrak{p}_v})^{n(\underline{i}_{\mathfrak{q}_v})}.\end{aligned}$$

We note that for any $l = 1, \ldots, K$ the marginal counts satisfy

$$\begin{aligned}&\sum_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}} n(\underline{i}_{C_l}) = M, \qquad \sum_{\underline{j}_{C_l}\in\mathcal{I}_{C_l}:\underline{j}_{S_l}=\underline{i}_{S_l}} n(\underline{j}_{C_l}) = n(\underline{i}_{S_l}), \\ &\sum_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}} n(\underline{i}_{S_l}) = M\end{aligned} \tag{2.6}$$

and, for $v \in C_l$ and $\mathfrak{q}_v \subseteq C_l$, $\sum_{\underline{j}_{C_l}\in\mathcal{I}_{C_l}:\underline{j}_{\mathfrak{q}_v}=\underline{i}_{\mathfrak{q}_v}} n(\underline{j}_{C_l}) = n(\underline{i}_{\mathfrak{q}_v})$, $\underline{i}_{\mathfrak{q}_v} \in \mathcal{I}_{\mathfrak{q}_v}$.

In Bayesian inference, we view the cell probabilities parameter of the multinomial distribution as a random vector and accordingly, we will write $\mathbf{p} = (\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ with $\mathbf{p}(\underline{i}) > 0$ and $\sum_{\underline{i}\in\mathcal{I}} \mathbf{p}(\underline{i}) = 1$. Thus, $\mathbf{p}(\underline{i}) = \mathbb{P}(\mathbf{X} = \underline{i} \mid \mathbf{p})$ which, in the

sequel, we will denote $\mathbb{P}_{\mathbf{p}}(\mathbf{X} = \underline{i})$ when we want to emphasize that $\mathbf{p}(\underline{i})$ represents the conditional probability that $\mathbf{X} = \underline{i}$. The random vector $\mathbf{p}$ takes values $p = (p(\underline{i}), \underline{i} \in \mathcal{I})$ and is the random variable of interest in this paper.

As mentioned in the Introduction, for G decomposable, the Diaconis–Ylvisaker conjugate prior on $\mathbf{p}$ is the hyper Dirichlet first identified by [6] with density

$$\frac{\Gamma(\alpha)\prod_{l=2}^{K}\prod_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}}\Gamma(\alpha_{\underline{i}_{S_l}}^{S_l})}{\prod_{l=1}^{K}\prod_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}}\Gamma(\alpha_{\underline{i}_{C_l}}^{C_l})}\frac{\prod_{l=1}^{K}\prod_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}}(p_{\underline{i}_{C_l}}^{C_l})^{\alpha_{\underline{i}_{C_l}}^{C_l}-1}}{\prod_{l=2}^{K}\prod_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}}(p_{\underline{i}_{S_l}}^{S_l})^{\alpha_{\underline{i}_{S_l}}^{S_l}-1}}, \tag{2.7}$$

where $p_{\underline{i}_{C_l}}^{C_l}$ and $p_{\underline{i}_{S_l}}^{S_l}$ are the values of the random variables $\mathbf{p}_{\underline{i}_{C_l}}^{C_l} = \mathbb{P}_{\mathbf{p}}(\mathbf{X}_{C_l} = \underline{i}_{C_l})$ and $\mathbf{p}_{\underline{i}_{S_l}}^{S_l} = \mathbb{P}_{\mathbf{p}}(\mathbf{X}_{S_l} = \underline{i}_{S_l})$, $l = 1, \ldots, K$, respectively, and where the hyper parameters satisfy constraints parallel to (2.6), namely

$$\sum_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}}\alpha_{\underline{i}_{C_l}}^{C_l} = \alpha, \qquad \sum_{\underline{j}_{C_l}\in\mathcal{I}_{C_l}:\underline{j}_{S_l}=\underline{i}_{S_l}}\alpha_{\underline{j}_{C_l}}^{C_l} = \alpha_{\underline{i}_{S_l}}^{S_l}, \qquad \sum_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}}\alpha_{\underline{i}_{S_l}}^{S_l} = \alpha. \tag{2.8}$$

Unlike the Dirichlet on the complete graph, this density is typically defined on a complicated manifold determined not only by summation to one but also by all the conditional independence properties encoded in G. This is one of the reasons our approach is through moments and not densities.

We now want to recall the expression of the moments of the hyper Dirichlet. For simplicity, let us write

$$\alpha(G,\mathcal{I}) = \bigl(\alpha, \alpha_{\underline{k}}^{C_l}, \underline{k} \in \mathcal{I}_{C_l}, l = 1, \ldots, K, \alpha_{\underline{k}}^{S_l}, \underline{k} \in \mathcal{I}_{S_l}, l = 2, \ldots, K\bigr).$$

For any d-dimensional table $(r_{\underline{i}}, \underline{i} \in \mathcal{I})$ of nonnegative integers, we write the E-marginal sums as

$$r_{\underline{e}}^{E} = \sum_{\underline{i}:\underline{i}_E=\underline{e}} r(\underline{i}). \tag{2.9}$$

In particular, we will use $r_{\underline{i}_{C_l}}^{C_l}$, $r_{\underline{i}_{S_l}}^{S_l}$. By analogy with $\alpha(G,\mathcal{I})$, we will also use the notation $r(G,\mathcal{I})$. The normalizing constant of the hyper Dirichlet is

$$\mathcal{Z}\bigl(\alpha(G,\mathcal{I})\bigr) = \frac{\prod_{l=1}^{K}\prod_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}}\Gamma(\alpha_{\underline{i}_{C_l}}^{C_l})}{\Gamma(\alpha)\prod_{l=2}^{K}\prod_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}}\Gamma(\alpha_{\underline{i}_{S_l}}^{S_l})}.$$

The moments are then equal to

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \mathbb{E}\frac{\prod_{l=1}^{K}\prod_{\underline{i}_{C_l}\in\mathcal{I}_{C_l}}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{C_l} = \underline{i}_{C_l})^{r_{\underline{i}_{C_l}}^{C_l}}}{\prod_{l=2}^{K}\prod_{\underline{i}_{S_l}\in\mathcal{I}_{S_l}}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{S_l} = \underline{i}_{S_l})^{r_{\underline{i}_{S_l}}^{S_l}}} = \frac{\mathcal{Z}(\alpha(G,\mathcal{I}) + r(G,\mathcal{I}))}{\mathcal{Z}(\alpha(G,\mathcal{I}))}.$$

In particular,

$$\mathbb{E}\mathbf{p}(\underline{i}) = \frac{\prod_{l=1}^{K}\Gamma(\alpha_{\underline{i}_{C_l}}^{C_l}+1)}{\Gamma(\alpha+1)\prod_{l=2}^{K}\Gamma(\alpha_{\underline{i}_{S_l}}^{S_l}+1)}\frac{\Gamma(\alpha)\prod_{l=2}^{K}\Gamma(\alpha_{\underline{i}_{S_l}}^{S_l})}{\prod_{l=1}^{K}\Gamma(\alpha_{\underline{i}_{C_l}}^{C_l})} = \frac{\prod_{l=1}^{K}\alpha_{\underline{i}_{C_l}}^{C_l}}{\alpha\prod_{l=2}^{K}\alpha_{\underline{i}_{S_l}}^{S_l}}.$$

Together with the constraints (2.8), this shows that we can write

$$\alpha_{\underline{i}_{C_l}}^{C_l} = \alpha\theta(\underline{i}_{C_l}), \qquad \alpha_{\underline{i}_{S_l}}^{S_l} = \alpha\theta(\underline{i}_{S_l}), \tag{2.10}$$

where $\theta(\underline{i}_{C_l}) = \mathbb{E}\mathbf{p}(\underline{i}_{C_l})$ and $\theta(\underline{i}_{S_l}) = \mathbb{E}\mathbf{p}(\underline{i}_{S_l})$, a relationship similar to (1.2) as mentioned in the Introduction.

3. $\mathfrak{p}$-Dirichlet and $\mathcal{P}$-Dirichlet distributions. Let G a decomposable graph and $\mathbf{X} = (X_v, v \in V)$ Markov with respect to G be as in Section 2.2. Let $\mathfrak{p}$ be the parent function of a moral DAG Markov equivalent to G. We introduce the random variables

$$\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v} := \mathbb{P}_{\mathbf{p}}(X_v = m|\mathbf{X}_{\mathfrak{p}_v} = \underline{k}) = \frac{\mathbf{p}_{(\underline{k},m)}^{\mathfrak{q}_v}}{\mathbf{p}_{\underline{k}}^{\mathfrak{p}_v}}, \qquad m \in \mathcal{I}_v, \underline{k} \in \mathcal{I}_{\mathfrak{p}_v}, v \in V,$$

where $\mathbf{p}_{\underline{n}}^{D} = \sum_{\underline{j}\in\mathcal{I}:\underline{j}_D=\underline{n}}\mathbf{p}(\underline{j})$ for $\underline{n} \in \mathcal{I}_D$, $D \subset V$ is the vector of D-marginal probability for $\underline{n} \in \mathcal{I}_D$. We thus have

$$\mathbf{p}(\underline{i}) = \prod_{v\in V}\mathbf{p}_{i_v|\underline{i}_{\mathfrak{p}_v}}^{v|\mathfrak{p}_v}, \qquad \underline{i} \in \mathcal{I}. \tag{3.1}$$

Since $\mathbf{X} = (X_1, \ldots, X_d)$ is Markov with respect to G, we know from the first equality in (2.5) that $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$, factorizes also with respect to G. When the random vector $\mathbf{p}(\underline{i})$ satisfies the first equality in (2.5), we say that the random vector $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ is associated with the graph G.

DEFINITION 3.1. The random vector $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ associated with the graph G has a $\mathfrak{p}$-Dirichlet distribution if the random vectors $(\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v}, m \in \mathcal{I}_v)$, $\underline{k} \in \mathcal{I}_{\mathfrak{p}_v}$, $v \in V$, in representation (3.1) are independent and follow (classical) Dirichlet distributions.

Recall that if a random vector $(\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v}, m \in \mathcal{I}_v)$ has a classical Dirichlet distribution $\mathrm{Dir}(\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v}, m \in \mathcal{I}_v)$, its density is

$$f_{\underline{k}}^{v|\mathfrak{p}_v}(\underline{x}) = \frac{\Gamma(\sum_{m\in\mathcal{I}_v}\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})}{\prod_{m\in\mathcal{I}_v}\Gamma(\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})}\prod_{m\in\mathcal{I}_v}x_m^{\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v}-1}I_{T_{|\mathcal{I}_v|}}(\underline{x}),$$

where $\sum_{m\in\mathcal{I}_v}x_m = 1$ and $T_{n+1} = \{(x_1, \ldots, x_n) \in (0,1)^n : \sum_{i=1}^{n}x_i < 1\}$. Moreover, since we assume that the random vectors $(\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v}, m \in \mathcal{I}_v)$, $\underline{k} \in \mathcal{I}_{\mathfrak{p}_v}$, $v \in V$, are independent the joint density of

$$\left(\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v}, m \in \mathcal{I}_v, \underline{k} \in \mathcal{I}_{\mathfrak{p}_v}, v \in V\right)$$

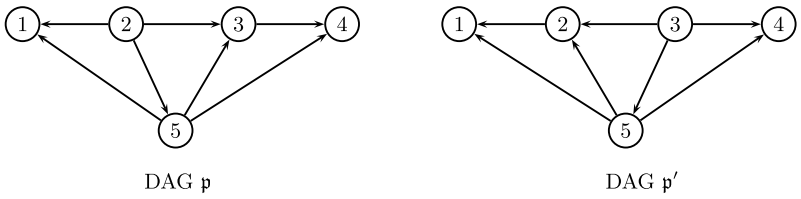

FIG. 1. *The two DAGs in Example 3.1.*

has the form

$$f\big(\underline{x}^{v,\underline{k}}, \underline{k} \in \mathcal{I}_{\mathfrak{p}_v}, v \in V\big) = \prod_{v \in V} \prod_{\underline{k} \in \mathcal{I}_{\mathfrak{p}_v}} \frac{\Gamma\big(\sum_{m \in \mathcal{I}_v} \alpha_{m|\underline{k}}^{v|\mathfrak{p}_v}\big)}{\prod_{m \in \mathcal{I}_v} \Gamma\big(\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v}\big)} \prod_{m \in \mathcal{I}_v} \big(x_m^{v,\underline{k}}\big)^{\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v} - 1},$$

where the support is a Cartesian product of unit simplexes $\times_{v \in V} T_{|\mathcal{I}_v|}^{\times |\mathcal{I}_{\mathfrak{p}_v}|}$, that is $\underline{x}^{v,\underline{k}} \in T_{|\mathcal{I}_v|}$ and $\sum_{m \in \mathcal{I}_v} x_m^{v,\underline{k}} = 1$, $\underline{k} \in \mathcal{I}_{\mathfrak{p}_v}$, $v \in V$.

Now we define a $\mathcal{P}$-Dirichlet distribution, where $\mathcal{P}$ is a family of moral DAGs with skeleton G.

DEFINITION 3.2. Let $\mathcal{P}$ be a family of moral DAGs with skeleton G. The random vector $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ associated with the graph G has the $\mathcal{P}$-Dirichlet distribution iff it has a $\mathfrak{p}$-Dirichlet distribution for any $\mathfrak{p} \in \mathcal{P}$.

Of course, this definition implies that some consistency conditions for the parameters of the Dirichlet distributions defining the $\mathfrak{p}$-Dirichlet distributions for the various $\mathfrak{p} \in \mathcal{P}$, have to be satisfied. This issue will be conveniently treated by looking at the moments of the $\mathfrak{p}$-Dirichlet and $\mathcal{P}$-Dirichlet laws.

EXAMPLE 3.1. Let $G = (V, E)$ be the decomposable graph with $V = \{1, 2, 3, 4, 5\}$ and cliques $\{1, 2, 5\}$, $\{2, 3, 5\}$ and $\{3, 4, 5\}$. Let $\mathcal{P} = \{\mathfrak{p}, \mathfrak{p}'\}$ with $\mathfrak{p}$ and $\mathfrak{p}'$ as in Figure 1. We have

$$\mathfrak{p}_1 = \{2, 5\}, \qquad \mathfrak{p}_2 = \varnothing, \qquad \mathfrak{p}_3 = \{2, 5\}, \qquad \mathfrak{p}_4 = \{3, 5\}, \qquad \mathfrak{p}_5 = \{2\}$$

and

$$\mathfrak{p}'_1 = \{2, 5\}, \qquad \mathfrak{p}'_2 = \{3, 5\}, \qquad \mathfrak{p}'_3 = \varnothing, \qquad \mathfrak{p}'_4 = \{3, 5\}, \qquad \mathfrak{p}'_5 = \{3\}.$$

Then the $\mathfrak{p}$-Dirichlet distribution is defined up to a multiplicative constant through the following independent Dirichlets:

$$\mathbf{p}^2 \sim \mathrm{Dir}\big(\alpha_m^2, m \in \mathcal{I}_2\big), \qquad \mathbf{p}_k^{5|2} \sim \mathrm{Dir}\big(\alpha_{m|k}^{5|2}, m \in \mathcal{I}_5\big), \qquad k \in \mathcal{I}_2,$$

$$\mathbf{p}_{\underline{k}}^{1|25} \sim \mathrm{Dir}\big(\alpha_{m|\underline{k}}^{1|25}, m \in \mathcal{I}_1\big), \qquad \underline{k} \in \mathcal{I}_{25},$$

$$\mathbf{p}_{\underline{k}}^{3|25} \sim \mathrm{Dir}\big(\alpha_{m|\underline{k}}^{3|25}, m \in \mathcal{I}_3\big), \qquad \underline{k} \in \mathcal{I}_{25},$$
$$\mathbf{p}_{\underline{k}}^{4|35} \sim \mathrm{Dir}\big(\alpha_{m|\underline{k}}^{4|35}, m \in \mathcal{I}_4\big), \qquad \underline{k} \in \mathcal{I}_{35}$$

and the $\mathfrak{p}'$-Dirichlet is defined through

$$\mathbf{p}^{3} \sim \mathrm{Dir}\big(\beta_{m}^{3}, m \in \mathcal{I}_3\big), \qquad \mathbf{p}_{k}^{5|3} \sim \mathrm{Dir}\big(\beta_{m|k}^{5|3}, m \in \mathcal{I}_5\big), \qquad k \in \mathcal{I}_3,$$
$$\mathbf{p}_{\underline{k}}^{1|25} \sim \mathrm{Dir}\big(\beta_{m|\underline{k}}^{1|25}, m \in \mathcal{I}_1\big), \qquad \underline{k} \in \mathcal{I}_{25},$$
$$\mathbf{p}_{\underline{k}}^{2|35} \sim \mathrm{Dir}\big(\beta_{m|\underline{k}}^{2|35}, m \in \mathcal{I}_2\big), \qquad \underline{k} \in \mathcal{I}_{35},$$
$$\mathbf{p}_{\underline{k}}^{4|35} \sim \mathrm{Dir}\big(\beta_{m|\underline{k}}^{4|35}, m \in \mathcal{I}_4\big), \qquad \underline{k} \in \mathcal{I}_{35}.$$

We will come back to this example later in Section 4.2.1 and see how the constraints on the hyper parameters, that is, the parameters of the $\mathcal{P}$-Dirichlet come about.

4. Moments.

4.1. *The* $\mathfrak{p}$*-Dirichlet distribution*. If the vector of random probabilities $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ follows the $\mathfrak{p}$-Dirichlet distribution, then for any nonnegative integers $r_{\underline{i}}$, $\underline{i} \in \mathcal{I}$,

$$\begin{aligned}\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} &= \prod_{v\in V}\prod_{\underline{k}\in\mathcal{I}_{\mathfrak{p}_v}} \mathbb{E}\prod_{m\in\mathcal{I}_v}[\mathbf{p}_{m|\underline{k}}^{v|\mathfrak{p}_v}]^{r_{(\underline{k},m)}^{\mathfrak{q}_v}} \\ &= \prod_{v\in V}\prod_{\underline{k}\in\mathcal{I}_{\mathfrak{p}_v}} \frac{\Gamma(\sum_{m\in\mathcal{I}_v}\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})}{\prod_{m\in\mathcal{I}_v}\Gamma(\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})} \frac{\prod_{m\in\mathcal{I}_v}\Gamma(r_{(\underline{k},m)}^{\mathfrak{q}_v}+\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})}{\Gamma(r_{\underline{k}}^{\mathfrak{p}_v}+\sum_{m\in\mathcal{I}_v}\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})},\end{aligned}$$

where for $(\underline{k}, m) \in \mathcal{I}_{\mathfrak{p}_v} \times \mathcal{I}_v = \mathcal{I}_{\mathfrak{q}_v}$, $r_{(\underline{k},m)}^{\mathfrak{q}_v}$ and $r_{\underline{k}}^{\mathfrak{p}_v}$ are as defined in (2.9). That is,

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \prod_{v\in V}\prod_{\underline{k}\in\mathcal{I}_{\mathfrak{p}_v}} \frac{\prod_{m\in\mathcal{I}_v}(\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v})^{r_{(\underline{k},m)}^{\mathfrak{q}_v}}}{(\tilde{\alpha}_{\underline{k}}^{\mathfrak{p}_v})^{r_{\underline{k}}^{\mathfrak{p}_v}}}, \tag{4.1}$$

where we write $(\alpha)^r = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)} = \alpha(\alpha+1)\cdots(\alpha+r-1)$ for the rising factorial power and

$$\tilde{\alpha}_{\underline{k}}^{\mathfrak{p}_v} = \sum_{m\in\mathcal{I}_v}\alpha_{m|\underline{k}}^{v|\mathfrak{p}_v}. \tag{4.2}$$

Note that since the $\mathfrak{p}$-distribution has a bounded support it is uniquely determined by the moments as given in (4.1).

4.2. *The $\mathcal{P}$-Dirichlet distribution.* If a vector of random probabilities $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ follows the $\mathcal{P}$-Dirichlet distribution then the formula for moments (4.1) holds for all $\mathfrak{p} \in \mathcal{P}$.

Since its support is bounded, the $\mathcal{P}$-Dirichlet distribution can be defined through the form of its moments. The equality of the moments for the different representations of the $\mathcal{P}$-Dirichlet by the $\mathfrak{p}$-Dirichlet, $\mathfrak{p} \in \mathcal{P}$, will impose equality constraints on the parameters of the $\mathfrak{p}$-Dirichlet's, and consequently those of the $\mathcal{P}$-Dirichlet. Before developing the theory let us illustrate the mechanism using Example 3.1 above.

4.2.1. *Example 3.1 continued.* Using (2.9) with $r^{\varnothing} = r$, the equality of the moments obtained from (4.1) for both $\mathfrak{p}$ and $\mathfrak{p}'$ yields

$$\begin{aligned}
\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} &= \frac{\prod_{m\in\mathcal{I}_2}(\alpha_m^2)^{r_m^2}}{(\tilde{\alpha})^r}\prod_{n\in\mathcal{I}_2}\frac{\prod_{m\in\mathcal{I}_5}(\alpha_{m|n}^{5|2})^{r_{(n,m)}^{25}}}{(\tilde{\alpha}_n^2)^{r_n^2}}\prod_{\underline{n}\in\mathcal{I}_{25}}\frac{\prod_{m\in\mathcal{I}_1}(\alpha_{m|\underline{n}}^{1|25})^{r_{(m,\underline{n})}^{125}}}{(\tilde{\alpha}_{\underline{n}}^{25,1})^{r_{\underline{n}}^{25}}}\\
&\quad\times\prod_{\underline{n}\in\mathcal{I}_{25}}\frac{\prod_{m\in\mathcal{I}_3}(\alpha_{m|\underline{n}}^{3|25})^{r_{(m,\underline{n})}^{325}}}{(\tilde{\alpha}_{\underline{n}}^{25,3})^{r_{\underline{n}}^{25}}}\prod_{\underline{n}\in\mathcal{I}_{35}}\frac{\prod_{m\in\mathcal{I}_4}(\alpha_{m|\underline{n}}^{4|35})^{r_{(m,\underline{n})}^{435}}}{(\tilde{\alpha}_{\underline{n}}^{35})^{r_{\underline{n}}^{35}}}\\
&= \frac{\prod_{m\in\mathcal{I}_3}(\beta_m^3)^{r_m^3}}{(\tilde{\beta})^r}\prod_{n\in\mathcal{I}_3}\frac{\prod_{m\in\mathcal{I}_5}(\beta_{m|n}^{5|3})^{r_{(n,m)}^{35}}}{(\tilde{\beta}_n^3)^{r_n^3}}\prod_{\underline{n}\in\mathcal{I}_{35}}\frac{\prod_{m\in\mathcal{I}_2}(\beta_{m|\underline{n}}^{2|35})^{r_{(m,\underline{n})}^{235}}}{(\tilde{\beta}_{\underline{n}}^{35,2})^{r_{\underline{n}}^{35}}}\\
&\quad\times\prod_{\underline{n}\in\mathcal{I}_{35}}\frac{\prod_{m\in\mathcal{I}_4}(\beta_{m|\underline{n}}^{4|35})^{r_{(m,\underline{n})}^{435}}}{(\tilde{\beta}_{\underline{n}}^{35,4})^{r_{\underline{n}}^{35}}}\prod_{\underline{n}\in\mathcal{I}_{25}}\frac{\prod_{m\in\mathcal{I}_1}(\beta_{m|\underline{n}}^{1|25})^{r_{(m,\underline{n})}^{125}}}{(\tilde{\beta}_{\underline{n}}^{25})^{r_{\underline{n}}^{25}}}.
\end{aligned}$$

Since there are no factorial powers in r_m^2 on the right-hand side of the equation above the terms in r_m^2 on the left-hand side must cancel out, that is, $\alpha_m^2 = \tilde{\alpha}_m^2$. Similarly, $\beta_m^3 = \tilde{\beta}_m^3$. The factorial power $r_{\underline{n}}^{125}$ on the right- and left-hand side must be the same and, therefore, $\alpha_{m|\underline{n}}^{1|25} = \beta_{m|\underline{n}}^{1|25}$. Similarly, $\alpha_{m|\underline{n}}^{3|25} = \beta_{m|\underline{n}}^{2|35}$, $\alpha_{m|\underline{n}}^{4|35} = \beta_{m|\underline{n}}^{4|35}$ and also $\tilde{\alpha} = \tilde{\beta}$. For the factorial powers in $r_{\underline{n}}^{25}$ we observe that on the left-hand side there is one power in the numerator and two in the denominator, while on the right-hand side there is only one power in the denominator. Therefore, the factorial power in the numerator must cancel with one of the two factorial powers of $\tilde{\alpha}_{\underline{n}}^{25,3}$ or of $\tilde{\alpha}_{\underline{n}}^{25,1}$ in the denominator. This means that for any $\underline{n} \in \mathcal{I}_{25}$:

- either $\alpha_{\underline{n}}^{5|2} = \tilde{\alpha}_{\underline{n}}^{25,3}$ and $\tilde{\beta}_{\underline{n}}^{25} = \tilde{\alpha}_{\underline{n}}^{25,1}$, which as we will see later, corresponds to the $\mathfrak{p}$-perfect order of cliques $o_{\mathfrak{p}}^{(1)} = (C_1 = \{2,3,5\}, C_2 = \{1,2,5\}, C_3 = \{3,4,5\})$,
- or $\alpha_{\underline{n}}^{5|2} = \tilde{\alpha}_{\underline{n}}^{25,1}$ and $\tilde{\beta}_{\underline{n}}^{25} = \tilde{\alpha}_{\underline{n}}^{25,3}$, which as we will see later, corresponds to the $\mathfrak{p}$-perfect order of cliques $o_{\mathfrak{p}}^{(2)} = (C_1 = \{1,2,5\}, C_2 = \{2,3,5\}, C_3 = \{3,4,5\})$.

Similarly, for the factorial powers $r_{\underline{n}}^{35}$ one can choose to cancel the factorial power of $r_{\underline{n}}^{35}$ in the numerator with factorial powers of either $\tilde{\beta}_{\underline{n}}^{35,2}$ or $\tilde{\beta}_{\underline{n}}^{35,4}$. Then, for any $\underline{n} \in \mathcal{I}_{35}$:

- either $\beta_{\underline{n}}^{5|3} = \tilde{\beta}_{\underline{n}}^{35,2}$ and $\tilde{\alpha}_{\underline{n}}^{35} = \tilde{\beta}_{\underline{n}}^{35,4}$, corresponding to the $\mathfrak{p}'$-perfect order of cliques $o_{\mathfrak{p}'}^{(1)} = (C_1 = \{2,3,5\}, C_2 = \{3,4,5\}, C_3 = \{1,2,5\})$,
- or $\beta_{\underline{n}}^{5|3} = \tilde{\beta}_{\underline{n}}^{35,4}$ and $\tilde{\alpha}_{\underline{n}}^{35} = \tilde{\alpha}_{\underline{n}}^{35,2}$, corresponding to the $\mathfrak{p}'$-perfect order of cliques $o_{\mathfrak{p}'}^{(2)} = (C_1 = \{3,4,5\}, C_2 = \{2,3,5\}, C_3 = \{1,2,5\})$.

From any of these cancelation possibilities, we obtain four $\mathcal{P}$-Dirichlet families with the same formula of moments

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \frac{\prod_{\underline{n}\in\mathcal{I}_{125}}(\nu_{\underline{n}}^{125})^{r_{\underline{n}}^{125}}\prod_{\underline{n}\in\mathcal{I}_{235}}(\nu_{\underline{n}}^{235})^{r_{\underline{n}}^{235}}\prod_{\underline{n}\in\mathcal{I}_{345}}(\nu_{\underline{n}}^{345})^{r_{\underline{n}}^{345}}}{(\mu)^r\prod_{\underline{m}\in\mathcal{I}_{25}}(\mu_{\underline{m}}^{25})^{r_{\underline{m}}^{25}}\prod_{\underline{m}\in\mathcal{I}_{35}}(\mu_{\underline{m}}^{35})^{r_{\underline{m}}^{35}}},$$

but with different sets of parameter constraints depending on the choice of pairs $O_{\mathcal{P}} = (o_{\mathfrak{p}}^{(i)}, o_{\mathfrak{p}'}^{(j)})$, $i, j = 1, 2$. See Section 4.2.1 of the supplementary material [15] for further details.

In the remainder of this section, we will show that to each collection $O_{\mathcal{P}}$ of orders, as defined in (2.3), corresponds a family of $\mathcal{P}$-Dirichlet distributions.

4.2.2. *The moment formula.* In the following theorem giving us the moments of the $\mathcal{P}$-Dirichlet distribution, we use the notions of $\mathfrak{P}$, $\mathfrak{Q}$, $\mathfrak{R}$ and Lemmas 2.2 and 2.3 and the notation $O_{\mathcal{P}}$ and $Q_1^C, \ldots, Q_{j_C-1}^C$, $C \in \mathcal{C}$ of (2.3) and (2.4), respectively.

THEOREM 4.1. *A vector of random probabilities* $(\mathbf{p}(\underline{i}), i \in \mathcal{I})$ *associated with the graph* G *follows a* $\mathcal{P}$*-Dirichlet distribution iff there exists a collection* $O_{\mathcal{P}}$ *as defined in* (2.3), *such that for any nonnegative integers* $r_{\underline{i}}$, $\underline{i} \in \mathcal{I}$,

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \frac{\prod_{A\in\mathfrak{Q}}\prod_{\underline{m}\in\mathcal{I}_A}(\nu_{\underline{m}}^A)^{r_{\underline{m}}^A}}{\prod_{B\in\mathfrak{P}}\prod_{\underline{n}\in\mathcal{I}_B}(\mu_{\underline{n}}^B)^{r_{\underline{n}}^B}}, \tag{4.3}$$

where $\nu_{\underline{m}}^A$, $\underline{m} \in \mathcal{I}_A$, $A \in \mathfrak{Q}$, *and* $\mu_{\underline{n}}^B$, $\underline{n} \in \mathcal{I}_B$, $B \in \mathfrak{P}$, *are positive numbers satisfying*

$$\mu_{\underline{n}}^B = \sum_{\underline{k}\in\mathcal{I}_{A\setminus B}} \nu_{(\underline{n},\underline{k})}^A \qquad \forall \underline{n} \in \mathcal{I}_B \tag{4.4}$$

whenever there exist $\mathcal{S} \ni S \subset C \in \mathcal{C}$ *and* $o_{\mathfrak{p}} \in O_{\mathcal{P}}$ *such that* $S \overset{o_{\mathfrak{p}}}{\to} C$ *and* $B = Q_i^C \subsetneq A = Q_{i-1}^C$ *for some* $i \in \{1, \ldots, j_C\}$.

The proof is given in Section 4.2.2 of the supplementary material [15].

It follows from the expression (4.3) of the moments that the $\mathcal{P}$-Dirichlet distribution has a density which, following a given perfect order $o_{\mathfrak{p}} \in O_{\mathcal{P}}$ of the cliques, can be expressed as the product of independent classical Dirichlet distributions as given in (4.5) below. Using the notation $R_i^{C_l} = Q_i^{C_l} \setminus Q_{i+1}^{C_l}$, $i \in \{0, \dots, j_{C_l} - 1\}$, $l \in \{1, \dots, K\}$, we have the following.

COROLLARY 4.2. *Let $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ be a random vector having $\mathcal{P}$-Dirichlet distribution with constraints (4.4) governed by a family $O_{\mathcal{P}}$. Consider any perfect order $o_{\mathfrak{p}} = (C_1, \dots, C_K) \in O_{\mathcal{P}}$ of the cliques. There exist independent vectors $(\mathbf{p}_{\underline{m}|\underline{k}}^{R_i^{C_l}|Q_{i+1}^{C_l}}, \underline{m} \in \mathcal{I}_{R_i^{C_l}})$ having classical Dirichlet distributions $\mathrm{Dir}(\nu_{(\underline{k},\underline{m})}^{Q_i^{C_l}}, \underline{m} \in \mathcal{I}_{R_i^{C_l}})$, $\underline{k} \in \mathcal{I}_{Q_{i+1}^{C_l}}$, $i \in \{0, 1, \dots, j_{C_l} - 1\}$, $l \in \{1, \dots, K\}$, such that for any $\underline{i} \in \mathcal{I}$*

$$\mathbf{p}(\underline{i}) = \prod_{l=1}^{K} \prod_{i=0}^{j_{C_l}-1} \mathbf{p}_{\underline{m}|\underline{k}}^{R_i^{C_l}|Q_{i+1}^{C_l}} \qquad \textit{where } \underline{m} = \underline{i}_{R_i^{C_l}} \textit{ and } \underline{k} = \underline{i}_{Q_{i+1}^{C_l}}.$$

Thus, the density of the $\mathcal{P}$-Dirichlet distribution can be written as

$$\prod_{l=1}^{K} \prod_{i=0}^{j_{C_l}-1} \prod_{\underline{k} \in \mathcal{I}_{Q_{i+1}^{C_l}}} \mathrm{Dir}\big(\nu_{(\underline{k},\underline{m})}^{Q_i^{C_l}}, \underline{m} \in \mathcal{I}_{R_i^{C_l}}\big). \tag{4.5}$$

We note that the well-known decomposition of the hyper Dirichlet density (2.7) into

$$C \prod_{l=1}^{K} \prod_{\underline{k} \in \mathcal{I}_{S_l}} \prod_{\underline{m} \in \mathcal{I}_{R_l}} (p_{\underline{m}|\underline{k}}^{R_l|S_l})^{\alpha_{(\underline{k},\underline{m})}^{C_l} - 1},$$

where C is the normalizing constant, is a special case of (4.5) where $j_{C_l} = 1$ and $Q_1^{C_l}$ and $Q_0^{C_l}$ are respectively replaced by S_l and C_l, and $\nu_{\underline{k},\underline{m}}^{Q_0}$ is replaced by $\alpha_{\underline{k},\underline{m}}^{C_l}$.

4.2.3. *Interpretation of the hyper parameters.* The interpretation of the hyper parameters of the $\mathcal{P}$-Dirichlet follows from Corollary 4.2. Since for each $\underline{k} \in \mathcal{I}_{Q_{i+1}^{C_l}}$, the vector $(\mathbf{p}_{\underline{m}|\underline{k}}^{R_i^{C_l}|Q_{i+1}^{C_l}}, \underline{m} \in \mathcal{I}_{R_i^{C_l}})$ is Dirichlet, we have the expected value

$$\begin{aligned} \mathbb{E}\mathbf{p}_{\underline{m}|\underline{k}}^{R_i^{C_l}|Q_{i+1}^{C_l}} &= \frac{\nu_{(\underline{k},\underline{m})}^{Q_i^{C_l}}}{\sum_{\underline{n} \in \mathcal{I}_{R_i^{C_l}}} \nu_{(\underline{k},\underline{n})}^{Q_i^{C_l}}} \\ &= \frac{\nu_{(\underline{k},\underline{m})}^{Q_i^{C_l}}}{\mu_{\underline{k}}^{Q_{i+1}^{C_l}}}, \qquad \underline{m} \in \mathcal{I}_{R_i^{C_l}}, i = 0, \dots, j_{C_l} - 1, \end{aligned} \tag{4.6}$$

where the last equality is due to (4.4). The ratios on the right-hand side of equation (4.6) above can therefore be interpreted as the cell probabilities in a fictive $|R_i^{C_l}|$-dimensional contingency table for $\mathbf{X}_{R_i^{C_l}}$ given $\mathbf{X}_{Q_{i+1}^{C_l}} = \underline{k}$. Equivalently, the hyper parameters $(\nu^{Q_i^{C_l}}_{(\underline{k},\underline{m})}, \underline{m} \in \mathcal{I}_{R_i^{C_l}})$ can be viewed as the cell counts in the same fictive table with total count (or equivalent sample size) $\mu^{Q_{i+1}^{C_l}}_{\underline{k}}$.

To better see how the hyper parameters of the $\mathcal{P}$-Dirichlet differ from that of the hyper Dirichlet, let us now give the expectation of $\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{Q_i^{C_l}} = \underline{m})$. For the sake of clarity, we will write $\underline{m}_{j,l}$ for $\underline{m}_{Q_j^{C_l}}$. Using (4.6) and the global independences inherent to the $\mathcal{P}$-Dirichlet, we have that if $\mathbf{p}$ follows the $\mathcal{P}$-Dirichlet, for $i = 0, \ldots, j_{C_l} - 1$,

$$
\begin{aligned}
\mathbb{E}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{Q_i^{C_l}} = \underline{m}) &= \mathbb{E}\left(\mathbf{p}^{Q_{j_c}^{C_l}}_{\underline{m}_{j_{C_l},l}} \prod_{j=i}^{j_{C_l}-1} \mathbf{p}^{R_j^{C_l}|Q_{j+1}^{C_l}}_{\underline{m}_{R_j^{C_l}}|\underline{m}_{j+1,l}} \right) \\
&= \left(\mathbb{E}\mathbf{p}^{Q_{j_c}^{C_l}}_{\underline{m}_{j_{C_l},l}}\right) \prod_{j=i}^{j_{C_l}-1} \mathbb{E}\mathbf{p}^{R_j^{C_l}|Q_{j+1}^{C_l}}_{\underline{m}_{R_j^{C_l}}|\underline{m}_{j+1,l}} \qquad (4.7) \\
&= \left(\mathbb{E}\mathbf{p}^{Q_{j_c}^{C_l}}_{\underline{m}_{j_{C_l},l}}\right) \prod_{j=i}^{j_{C_l}-1} \frac{\nu^{Q_j^{C_l}}_{\underline{m}_{j,l}}}{\mu^{Q_{j+1}^{C_l}}_{\underline{m}_{j+1,l}}},
\end{aligned}
$$

where $\mathbb{E}\mathbf{p}^{Q_{j_c}^{C_l}}_{\underline{m}_{j_{C_l},l}}$ is computed by summing the expectation of all the $\mathbf{p}(\underline{i})$ making up $\mathbf{p}^{Q_{j_c}^{C_l}}_{\underline{m}_{j_{C_l},l}}$. The $\mu^{Q_{j+1}^{C_l}}_{\underline{m}_{j+1,l}}$ are the sums of $\nu^{Q_j^{C_l}}_{\underline{m}_{j,l}}$ as in (4.4) but are otherwise unrelated to any other hyper parameters.

In contrast, if $\mathbf{p}$ follows a hyper Dirichlet with hyper parameters $(\nu^{C_l}_{\underline{m}}, \underline{m} \in \mathcal{I}_{C_l}, l = 1, \ldots, K)$, then $\mathfrak{Q} = \mathcal{C}$, $Q_0^{C_l} = C_l$, $Q_1^{C_l} = S_l$ and (4.7) reduces to

$$
\mathbb{E}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{C_l} = \underline{m}) = \frac{\nu^{C_l}_{\underline{m}}}{\mu}, \qquad (4.8)
$$

where

$$
\sum_{\underline{k} \in \mathcal{I}_{C_l \setminus S_l}} \nu^{C_l}_{(\underline{n},\underline{k})} = \mu^{S_l}_{\underline{n}} \quad \text{and} \quad \sum_{\underline{m} \in \mathcal{I}_{C_l}} \nu^{C_l}_{\underline{m}} = \mu.
$$

We see from (4.7) and (4.8) that if we want to choose the hyper parameters of the $\mathcal{P}$-Dirichlet according to expert prior knowledge on $\mathbb{E}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{Q_i^{C_l}} = \underline{m}_{i,l})$, we have much more flexibility if $\mathbf{p}$ follows a $\mathcal{P}$-Dirichlet than if $\mathbf{p}$ follows a hyper Dirichlet.

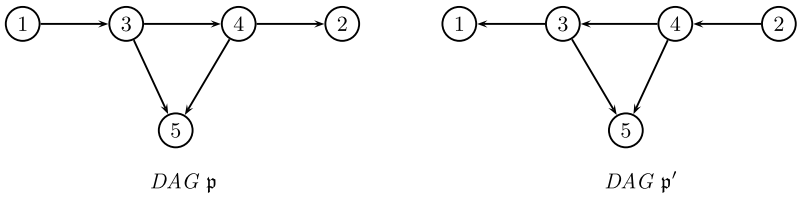

FIG. 2. *The two DAGs in Example 4.1.*

The interpretation and flexibility of the hyper parameters in the $\mathcal{P}$-Dirichlet will be formalized and illustrated in Section 5.1.

4.2.4. *More examples and the hyper Dirichlet as a special case of the $\mathcal{P}$-Dirichlet.* We now give a few examples of the $\mathcal{P}$-Dirichlet distribution. We start with an example in which $\mathfrak{R} \neq \varnothing$.

EXAMPLE 4.1. Let $G = (V, E)$ be a graph with

$$V = \{1, 2, 3, 4, 5\} \quad \text{and} \quad E = \big\{\{1, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\big\}.$$

Then $\mathcal{C} = \{\{1, 3\}, \{3, 4, 5\}, \{2, 4\}\}$ and $\mathcal{S} = \{\varnothing, \{3\}, \{4\}\}$. Let $\mathcal{P} = \{\mathfrak{p}, \mathfrak{p}'\}$ be as given in Figure 2. We have

$$\mathfrak{p}_1 = \varnothing, \qquad \mathfrak{p}_2 = \{4\}, \qquad \mathfrak{p}_3 = \{1\}, \qquad \mathfrak{p}_4 = \{3\}, \qquad \mathfrak{p}_5 = \{3, 4\}$$

and

$$\mathfrak{p}'_1 = \{3\}, \qquad \mathfrak{p}'_2 = \varnothing, \qquad \mathfrak{p}'_3 = \{4\}, \qquad \mathfrak{p}'_4 = \{2\}, \qquad \mathfrak{p}'_5 = \{3, 4\}.$$

Then

$$\mathfrak{Q} = \mathcal{C} \cup \{3, 4\}, \qquad \mathfrak{P} = \mathcal{S} \cup \{3, 4\},$$
$$\mathfrak{R}_{\{3,4,5\}} = \{3, 4\}, \qquad \mathfrak{R}_{\{1,3\}} = \mathfrak{R}_{\{2,4\}} = \varnothing.$$

Moreover, there is only one available collection of orders $O_{\mathcal{P}} = \{o, o'\}$, where the $\mathfrak{p}$-perfect order o is $(C_1 = \{1, 3\}, C_2 = \{3, 4, 5\}, C_3 = \{2, 4\})$ and the $\mathfrak{p}'$-perfect order o' is $(C'_1 = \{2, 4\}, C'_2 = \{3, 4, 5\}, C'_3 = \{1, 3\})$.

Then formula (4.3) for moments becomes

$$\mathbb{E} \prod_{\underline{i} \in \mathcal{I}} [\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \frac{\prod_{\underline{m} \in \mathcal{I}_{13}} (\nu^{13}_{\underline{m}})^{r^{13}_{\underline{m}}} \prod_{\underline{m} \in \mathcal{I}_{24}} (\nu^{24}_{\underline{m}})^{r^{24}_{\underline{m}}} \prod_{\underline{m} \in \mathcal{I}_{345}} (\nu^{345}_{\underline{m}})^{r^{345}_{\underline{m}}} \prod_{\underline{m} \in \mathcal{I}_{34}} (\nu^{34}_{\underline{m}})^{r^{34}_{\underline{m}}}}{(\mu)^r \prod_{m \in \mathcal{I}_3} (\mu^3_m)^{r^3_m} \prod_{m \in \mathcal{I}_4} (\mu^4_m)^{r^4_m} \prod_{\underline{m} \in \mathcal{I}_{34}} (\mu^{34}_{\underline{m}})^{r^{34}_{\underline{m}}}}$$

with the following consistency conditions:

$$\mu = \sum_{\underline{k} \in \mathcal{I}_{13}} \nu^{13}_{\underline{k}} = \sum_{\underline{k} \in \mathcal{I}_{24}} \nu^{24}_{\underline{k}}; \tag{4.9}$$

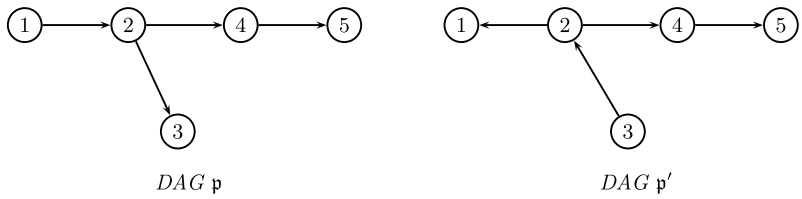

FIG. 3. *DAGs for Example 4.2.*

$$\mu_n^3 = \sum_{k\in\mathcal{I}_4} \nu_{(n,k)}^{34} = \sum_{k\in\mathcal{I}_1} \nu_{(k,n)}^{13}, \qquad n \in \mathcal{I}_3; \tag{4.10}$$

$$\mu_n^4 = \sum_{k\in\mathcal{I}_3} \nu_{(k,n)}^{34} = \sum_{k\in\mathcal{I}_2} \nu_{(k,n)}^{24}, \qquad n \in \mathcal{I}_4; \tag{4.11}$$

$$\mu_{\underline{n}}^{34} = \sum_{k\in\mathcal{I}_5} \nu_{(\underline{n},k)}^{345}, \qquad \underline{n} \in \mathcal{I}_{34}. \tag{4.12}$$

Consequently, combining the above equations, we also get

$$\mu = \sum_{n\in\mathcal{I}_3} \mu_n^3 = \sum_{n\in\mathcal{I}_4} \mu_n^4 = \sum_{\underline{k}\in\mathcal{I}_{34}} \nu_{\underline{k}}^{34}.$$

We will now look at the case where $\mathfrak{R} = \varnothing$. We will see that the formula for moments simplifies and becomes closer to the moment formula for the hyper Dirichlet. Let us first consider an example.

EXAMPLE 4.2. Consider a tree $G = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{2, 4\}, \{4, 5\}\}$. Let $\mathcal{P} = \{\mathfrak{p}, \mathfrak{p}'\}$ as in Figure 3. We have

$$\mathfrak{p}_1 = \varnothing, \qquad \mathfrak{p}_2 = \{1\}, \qquad \mathfrak{p}_3 = \{2\}, \qquad \mathfrak{p}_4 = \{2\}, \qquad \mathfrak{p}_5 = \{4\}$$

and

$$\mathfrak{p}'_1 = \{2\}, \qquad \mathfrak{p}'_2 = \{3\}, \qquad \mathfrak{p}'_3 = \varnothing, \qquad \mathfrak{p}'_4 = \{2\}, \qquad \mathfrak{p}'_5 = \{4\}.$$

Then

$$\mathfrak{Q} = \mathcal{C} = E, \qquad \mathfrak{P} = \big\{\{2\}, \{2\}, \{4\}\big\} = \mathcal{S}.$$

Moreover, $\mathfrak{R}_C = \varnothing$ for any $C \in \mathcal{C}$. Formula (4.3) for moments is

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}} [\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \frac{\prod_{\underline{m}\in\mathcal{I}_{12}} (\nu_{\underline{m}}^{12})^{r_{\underline{m}}^{12}} \prod_{\underline{m}\in\mathcal{I}_{23}} (\nu_{\underline{m}}^{23})^{r_{\underline{m}}^{23}} \prod_{\underline{m}\in\mathcal{I}_{24}} (\nu_{\underline{m}}^{24})^{r_{\underline{m}}^{24}} \prod_{\underline{m}\in\mathcal{I}_{45}} (\nu_{\underline{m}}^{45})^{r_{\underline{m}}^{45}}}{(\mu)^r \prod_{m\in\mathcal{I}_2} (\mu_m^{2(1)})^{r_m^2} \prod_{m\in\mathcal{I}_2} (\mu_m^{2(2)})^{r_m^2} \prod_{m\in\mathcal{I}_4} (\mu_m^4)^{r_m^4}},$$

with the following consistency conditions. If we consider the pair $(o_{\mathfrak{p}}, o_{\mathfrak{p}'})$, where $o_{\mathfrak{p}} = (C_1 = \{1, 2\}, C_2 = \{2, 3\}, C_3 = \{2, 4\}, C_4 = \{4, 5\})$ and $o_{\mathfrak{p}'} = (C'_1 =$

$\{2,3\}, C_2' = \{1,2\}, C_3' = \{2,4\}, C_4' = \{4,5\}$) then the constraints are

$$\mu = \sum_{\underline{k} \in \mathcal{I}_{12}} \nu_{\underline{k}}^{12} = \sum_{\underline{k} \in \mathcal{I}_{23}} \nu_{\underline{k}}^{23},$$

$$\mu_n^{2(1)} = \sum_{k \in \mathcal{I}_1} \nu_{(k,n)}^{12} = \sum_{k \in \mathcal{I}_3} \nu_{(n,k)}^{23}, \qquad n \in \mathcal{I}_2,$$

$$\mu_n^{2(2)} = \sum_{k \in \mathcal{I}_4} \nu_{(n,k)}^{24}, \qquad n \in \mathcal{I}_2,$$

$$\mu_n^{4} = \sum_{k \in \mathcal{I}_5} \nu_{(n,k)}^{45}, \qquad n \in \mathcal{I}_4.$$

If we keep $o_{\mathfrak{p}}$ as above and take $o_{\mathfrak{p}'} = (C_1' = \{2,3\}, C_2' = \{2,4\}, C_3' = \{4,5\}, C_4' = \{1,2\})$, then we obtain the same restrictions on the parameters as in the preceding situation.

The above example falls under a more general setting which follows immediately from Theorem 4.1 and which we formalize as follows.

COROLLARY 4.3. *Let the random vector* $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$, *associated with the graph* G, *have a* $\mathcal{P}$*-Dirichlet distribution for a family* $\mathcal{P}$ *of moral DAGs and a collection* $O_{\mathcal{P}}$ *of* $\mathfrak{p}$*-perfect orders. If* $\mathcal{P}$ *has the property* $\mathfrak{Q} = \mathcal{C}$ (*equivalently* $\mathfrak{P} = \mathcal{S}$ *or* $\mathfrak{R} = \varnothing$) *then*

$$\mathbb{E} \prod_{\underline{i} \in \mathcal{I}} [\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \frac{\prod_{C \in \mathcal{C}} \prod_{\underline{m} \in \mathcal{I}_C} (\nu_{\underline{m}}^C)^{r_{\underline{m}}^C}}{(\mu)^r \prod_{S \in \mathcal{S}} \prod_{\underline{m} \in \mathcal{I}_S} (\mu_{\underline{m}}^S)^{r_{\underline{m}}^S}}, \tag{4.13}$$

with the following consistency conditions: *whenever* $S \overset{o_{\mathfrak{p}}}{\to} C$ *for some* $\mathfrak{p} \in \mathcal{P}$,

$$\sum_{\underline{m} \in C : \underline{m}_S = \underline{n}} \nu_{\underline{m}}^C = \mu_{\underline{n}}^S, \qquad \underline{n} \in \mathcal{I}_S. \tag{4.14}$$

The hyper Dirichlet distribution is, of course, uniquely defined by moments of exactly the same form as (4.13) but with stronger consistency conditions. Indeed, (4.14) is satisfied for any $\mathcal{S} \ni S \subset C \in \mathcal{C}$. Therefore, a $\mathcal{P}$-Dirichlet distribution for the family $\mathcal{P}$ of *all* moral DAGs with skeleton $G = (V, E)$ is a hyper Dirichlet distribution. Actually, it follows directly from Corollary 4.3 that a considerably smaller family $\mathcal{P}$ forces the $\mathcal{P}$-Dirichlet to be hyper Dirichlet. We state this as follows.

THEOREM 4.4. *Let* $\mathcal{L}$ *be a* $\mathcal{P}$*-Dirichlet distribution for a family* $\mathcal{P}$ *of moral DAGs and a collection* $O_{\mathcal{P}}$ *as defined in* (2.3). *If*

$$\mathfrak{R} = \varnothing \tag{4.15}$$

and

$$(4.16)\qquad \forall (S \in \mathcal{S}, C \in \mathcal{C}) \qquad \text{if } S \subset C \text{ then } \exists o_{\mathfrak{p}} \in O_{\mathcal{P}} \text{ such that } S \overset{o_{\mathfrak{p}}}{\to} C$$

then $\mathcal{L}$ is a hyper Dirichlet distribution.

At this point, it is important to note that any $\mathcal{P}$-Dirichlet family contains a hyper Dirichlet distribution, which is obtained by imposing the following additional constraints:

- $\mu_{\underline{m}}^{Q} = \nu_{\underline{m}}^{Q}$ for any $\underline{m} \in \mathcal{I}_Q$ and any $Q \in \mathfrak{R}$,
- for any $B \in \mathcal{S}$ contained in $A \in \mathcal{C}$ condition (4.4) holds.

Yet, conditions (4.15) and (4.16) may not be satisfied, as can readily be seen from Example 4.1 if we choose $\mu_{\underline{m}}^{34} = \nu_{\underline{m}}^{34}$ for any $\underline{m} \in \mathcal{I}_{34}$.

5. The $\mathcal{P}$-Dirichlet as a prior distribution. In this section, we look at the properties of the $\mathcal{P}$-Dirichlet as a prior distribution. We first compute the dimension of the $\mathcal{P}$-Dirichlet family for a given $\mathcal{P}$ and show that it is always larger than the dimension of the hyper Dirichlet family with the same skeleton G [unless the set $\mathcal{P}$ is so large that the corresponding $\mathfrak{R}$ is empty and (4.16) holds]. We then show that the $\mathcal{P}$-Dirichlet has the two desirable properties of being conjugate and directed strong hyper Markov for every $\mathfrak{p} \in \mathcal{P}$. In Section 5.3 we make a remark on the possible extension of the $\mathcal{P}$-Dirichlet to families of immoral graphs represented by a given essential graph.

5.1. *Dimension of the $\mathcal{P}$-Dirichlet family.* We are now going to show that the dimension of the parameter space of the $\mathcal{P}$-Dirichlet distribution is always greater than or equal to that of the hyper Dirichlet. This means, of course, that when choosing the $\mathcal{P}$-Dirichlet as a prior rather than the hyper Dirichlet, we gain flexibility in our choice of the hyper parameters. The dimensions of both families are given in the following theorem.

If $\mathcal{P}$ is a collection of moral DAGs with skeleton G and $O_{\mathcal{P}}$ is a collection of $\mathfrak{p}$-perfect orders, $\mathfrak{p} \in \mathcal{P}$, for $S \in \mathcal{S}$ given, we denote by N_S be the number of cliques C such that

$$\text{if } S \subset C \text{ then } \exists \mathfrak{p} \in \mathcal{P} \text{ such that } S \overset{o_{\mathfrak{p}}}{\to} C.$$

THEOREM 5.1. *For G and $\mathcal{P}$ as given above, the dimension of the parameter space of the $\mathcal{P}$-Dirichlet family of distributions is*

$$(5.1)\qquad \mathcal{N}_{\mathcal{P}} = \sum_{Q \in \mathfrak{Q}} \prod_{v \in Q} |\mathcal{I}_v| - \sum_{S \in \mathcal{S}} (N_S - 1) \prod_{v \in S} |\mathcal{I}_v|.$$

The dimension of the parameter space of the hyper Dirichlet family of distributions with the same skeleton G is equal to

$$(5.2)\qquad \mathcal{N}_{HP} = \sum_{C \in \mathcal{C}} \prod_{v \in C} |\mathcal{I}_v| - \sum_{S \in \mathcal{S}} (N_S - 1) \prod_{v \in S} |\mathcal{I}_v|.$$

Moreover, if the $\mathcal{P}$-Dirichlet is not identical to the hyper Dirichlet, we always have

$$\mathcal{N}_{\mathcal{P}} > \mathcal{N}_{HP}. \tag{5.3}$$

The proof is given in Section 5.1 of the supplementary material [15].

Let us illustrate Theorem 5.1 by deriving the dimension of the $\mathcal{P}$-Dirichlet and hyper Dirichlet families respectively when G and $\mathcal{P}$ are as given in Example 4.1. Following the computations of the dimension of the two families, we will consider the problem of translating prior knowledge given by an expert into specific values of the hyper parameters. We will see that, of course, we have much more flexibility with the $\mathcal{P}$-Dirichlet than with the hyper Dirichlet.

EXAMPLE 5.1 (Example 4.1 continued). Let us assume that all variables are binary. From (4.3), we see that we have 20 $\nu^A_{\underline{m}}$ parameters, 4 for $(\nu^{13}_{\underline{m}}, \underline{m} \in \mathcal{I}_{13})$, 4 for $(\nu^{34}_{\underline{m}}, \underline{m} \in \mathcal{I}_{34})$, 4 for $(\nu^{24}_{\underline{m}}, \underline{m} \in \mathcal{I}_{24})$ and 8 for $(\nu^{345}_{\underline{m}}, \underline{m} \in \mathcal{I}_{345})$. The separators are $\{3\}$ and $\{4\}$. We have $\{3\} \overset{o'}{\to} \{1,3\}$ and $\{3\} \overset{o}{\to} \{3,4,5\}$ and, therefore, $N_{\{3\}} = 2$. Similarly, $\{4\} \overset{o}{\to} \{2,4\}$ and $\{4\} \overset{o'}{\to} \{3,4,5\}$ and, therefore, $N_{\{4\}} = 2$. According to (5.1), the dimension of the $\mathcal{P}$-Dirichlet family is

$$\mathcal{N}_{\mathcal{P}} = 20 - 2 - 2 = 16.$$

The dimension of the hyper Dirichlet family is

$$\mathcal{N}_{HP} = 16 - 2 - 2 = 12.$$

We will now continue with this example to illustrate how we can use the flexibility of the $\mathcal{P}$-Dirichlet prior. Let us assume that we have expert knowledge which tells us that one expects that $\mathbf{X}_{345}$ will take the value $(0,1,1)$ more often than the value $(1,1,1)$. We translate this with the inequality

$$\mathbb{EP}_{\mathbf{p}}\big(\mathbf{X}_{345} = (1,1,1)\big) \le \mathbb{EP}_{\mathbf{p}}\big(\mathbf{X}_{345} = (0,1,1)\big). \tag{5.4}$$

Following (4.7), we obtain after some simplifications the inequality

$$\frac{\nu^{345}_{111}\nu^{34}_{11}}{\nu^{345}_{111} + \nu^{345}_{110}} \le \frac{\nu^{345}_{011}\nu^{34}_{01}}{\nu^{345}_{011} + \nu^{345}_{010}}. \tag{5.5}$$

The constraints on the hyper parameters are given by (4.9)–(4.12). Constraint (4.12) is nothing but the definition of $\mu^{34}_{\underline{n}}$. Constraints (4.9)–(4.11) do not involve $\nu^{345}_{\underline{m}}$. So there are no constraints on the parameters $\nu^{345}_{\underline{m}}, \underline{m} \in \{0,1\}^3$: we can freely choose these values to reflect prior expert knowledge.

In contrast, if $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ follows the hyper Dirichlet distribution with parameters $(\nu^{C_l}_{\underline{m}}, \underline{m} \in \mathcal{I}_{C_l})$, $l = 1,2,3$, then (5.4) would simply translate as $\nu^{345}_{111} \le \nu^{345}_{011}$

with constraints of the type (2.8), that is, with constraints (4.9)–(4.12) of the $\mathcal{P}$-Dirichlet and the additional constraints

$$\mu^{34}_{\underline{m}} = \nu^{34}_{\underline{m}}, \qquad \underline{m} \in \mathcal{I}_{34}.$$

With these constraints, both $\mu^{34}_{\underline{m}}$ and $\nu^{34}_{\underline{m}}$ disappear from the moment formula (4.3) and, as we know, disappear as parameters of the hyper Dirichlet, resulting in much more coercive constraints on $\nu^{345}_{\underline{m}}$, $\underline{m} \in \mathcal{I}_{345}$.

We note that the set of equations (4.9)–(4.11) defines a hyper Dirichlet distribution for the probabilities of $\mathbf{X}_{1234}$. Therefore, we could say that the $\mathcal{P}$-Dirichlet in our example is made up of a hyper Dirichlet for the probabilities of $\mathbf{X}_{1234}$ combined with, for each value of $\mathbf{X}_{34} = \underline{m}_{34}$, an independent Dirichlet distribution for the probabilities of X_5 with parameters $(\nu^{345}_{(\underline{m}_{34}, m_5)}, m_5 \in \mathcal{I}_5)$. We can then interpret the parameters of the $\mathcal{P}$-Dirichlet appearing in (4.9)–(4.12) as follows. Consider a fictive table for $\mathbf{X}_{1234}$ with total count (equivalent sample size) μ. The vectors

$$\big(\nu^{13}_{\underline{k}}, \underline{k} \in \mathcal{I}_{13}\big), \qquad \big(\nu^{34}_{\underline{k}}, \underline{k} \in \mathcal{I}_{34}\big), \qquad \big(\nu^{24}_{\underline{k}}, \underline{k} \in \mathcal{I}_{24}\big),$$

are the counts in the $\mathbf{X}_{13}$-, $\mathbf{X}_{34}$-, $\mathbf{X}_{24}$-marginal tables, respectively obtained from the $\mathbf{X}_{1234}$ table. The vectors $(\mu^3_m, m \in \mathcal{I}_3)$ and $(\mu^4_m, m \in \mathcal{I}_4)$ are the counts for the X_3- and X_4-marginal tables respectively obtained from the same $\mathbf{X}_{1234}$ table or equivalently from the $\mathbf{X}_{13}$-, $\mathbf{X}_{34}$-, $\mathbf{X}_{24}$-marginal tables. However, for each $\underline{m} \in \mathcal{I}_{34}$, the vector $(\nu^{345}_{(\underline{m},k)}, k \in \mathcal{I}_5)$ is a free vector of counts for a fictive X_5-table.

5.2. *Conjugacy and directed strong hyper Markov property.* We will now emphasize the properties of the $\mathcal{P}$-Dirichlet that make it a useful prior for Bayesian inference. In the following theorem, we state that for any $\mathfrak{p} \in \mathcal{P}$, the $\mathcal{P}$-Dirichlet is conjugate and directed strong hyper Markov with respect to any $\mathfrak{p} \in \mathcal{P}$. We now recall the definition of the directed strong hyper Markov property. Let $\mathbf{X} = (X_v, v \in V)$ be a random variable Markov with respect to a DAG given by a parent function $\mathfrak{p}$, with distribution parameterized by $\theta \in \mathbb{R}^k$ for some k, which itself follows a law $\mathcal{L}$. Let $\theta_{\mathfrak{p}_v}$, $\theta_{v|\mathfrak{p}_v}$ and $\theta_{\mathfrak{nd}_v}$ denote the parameters of the marginal distribution of $\mathbf{X}_{\mathfrak{p}_v}$, the conditional distribution of X_v given $\mathbf{X}_{\mathfrak{p}_v}$ and the marginal distribution of the non-descendants of v, respectively. Then the law $\mathcal{L}$ is said to be directed strong hyper Markov if we have the conditional independences

$$\theta_{v|\mathfrak{p}_v} \perp \theta_{\mathfrak{nd}_v} | \theta_{\mathfrak{p}_v}, \qquad v \in V.$$

With this definition, we see that the directed strong hyper Markov property of the $\mathcal{P}$-Dirichlet follows by construction. Additionally, the $\mathcal{P}$-Dirichlet is conjugate with respect to the multinomial. These two properties are stated in the following theorem. Using the family $\mathcal{P}$ of Example 4.1 again, we will show how, with the strong directed hyper Markov and conjugacy properties, one can easily obtain the posterior mean of the contingency table cell probabilities.

THEOREM 5.2. *Let the conditional distribution of cell counts* $\mathbf{N} = (N(\underline{i}), \underline{i} \in \mathcal{I})$ *for* $\mathbf{X} = (X_v, v \in V)$ *given* $\mathbf{p} = (\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ *be multinomial* $\mathcal{M}(M, \mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ *Markov with respect to the decomposable graph* G. *Let* $\mathbf{p}$ *follow a* $\mathcal{P}$*-Dirichlet distribution with hyper parameters*

$$\nu^A_{\underline{m}}, \underline{m} \in \mathcal{I}_A, A \in \mathfrak{Q} \quad \text{and} \quad \mu^B_{\underline{k}}, \underline{k} \in \mathcal{I}_B, B \in \mathfrak{P}$$

as given in (4.3) *and* (4.4). *Then the posterior distribution of* $\mathbf{p}$ *given* $\mathbf{N} = (n(\underline{i}), \underline{i} \in \mathcal{I})$ *is* $\mathcal{P}$*-Dirichlet with hyper parameters*

$$n^A_{\underline{m}} + \nu^A_{\underline{m}}, \underline{m} \in \mathcal{I}_A, A \in \mathfrak{Q} \quad \text{and} \quad n^B_{\underline{k}} + \mu^B_{\underline{k}}, \underline{k} \in \mathcal{I}_B, B \in \mathfrak{P},$$

where $n^A_{\underline{m}}$ *is the* A*-marginal count for* $\underline{i}_A = \underline{m}$.

Moreover, for any $\mathfrak{p} \in \mathcal{P}$, *the* $\mathcal{P}$*-Dirichlet is directed strong hyper Markov.*

The proof is given in Section 5.2 of the supplementary material [15]. We now illustrate how we can readily estimate a cell probability with its posterior mean.

EXAMPLE 5.2 (Example 4.1 continued). With the $\mathcal{P}$-Dirichlet for $\mathbf{p}$ as in Example 4.1, we know from (5.5) that $\mathbb{E}\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{345} = (111)) = \frac{\nu^{345}_{111}\nu^{34}_{11}}{\mu^{34}_{11}}$. Given data $(n(i), i \in \mathcal{I})$, by Theorem 5.2, the posterior mean of $\mathbb{P}_{\mathbf{p}}(\mathbf{X}_{345} = (111))$ is

$$\mathbb{E}\big(\mathbb{P}_{\mathbf{p}}\big(\mathbf{X}_{345} = (111)\big) \mid \mathbf{N} = \big(n(i), i \in \mathcal{I}\big)\big) = \frac{(\nu^{345}_{111} + n^{345}_{111})(\nu^{34}_{11} + n^{34}_{11})}{\mu^{34}_{11} + n^{34}_{11}}.$$

We see from this example and also from the theorem above that, when $A \subset V$ belongs to both $\mathfrak{P}$ and $\mathfrak{Q}$, the data counts $n^A_{\underline{m}}$ are added to both $\nu^A_{\underline{m}}$ and $\mu^A_{\underline{m}}$. This is because in that case, the $\nu^A_{\underline{m}}$ and $\mu^A_{\underline{m}}$ refer to two different fictive tables.

Let us note that our priors could also find applications in model selection when the class of models to choose from contains only a few models as was done, for example, in [17], Section 7, for continuous Gaussian variables. In that paper, the authors studied the quality of the various estimates of the covariance matrix. The strength of a flexible prior with multiple hyper parameters generally lies in estimation.

5.3. *Arbitrary DAGs.* In this paper, we have only studied families of moral DAGs with the same decomposable skeleton G. They all belong to the same Markov equivalence class with essential graph $G^* = G$; see [1] or [20]. We think that our approach can be extended to families of arbitrary DAGs in any given Markov equivalence class. In particular, one could consider the family of all DAGs with a given essential graph G^*. This has potential application to model selection and will be the subject of further research. We give two examples of this possible extended $\mathcal{P}$-Dirichlet in Section 5.3 of the supplementary material [15].

6. Characterization by local and global independence.

6.1. *The $\mathcal{P}$-Dirichlet and the hyper Dirichlet.* We now briefly recall the definition of local and global independence and combine them into what we call *parameter independence*.

DEFINITION 6.1. Let $\mathbf{p}$ be a vector of random probabilities associated with the graph G. We say that local parameter independence holds for $\mathbf{p}$ with respect to a DAG with parent function $\mathfrak{p}$ if for any $v \in V$ the random vectors

$$(\mathbf{p}_{l|\underline{k}}^{v|\mathfrak{p}_v}, l \in \mathcal{I}_v), \qquad \underline{k} \in \mathcal{I}_{\mathfrak{p}_v},$$

are independent (nondegenerate) and we say that global parameter independence holds for $\mathbf{p}$ if the random vectors

$$((\mathbf{p}_{l|\underline{k}}^{v|\mathfrak{p}_v}, l \in \mathcal{I}_v), \underline{k} \in \mathcal{I}_{\mathfrak{p}_v}), \qquad v \in V$$

are independent (nondegenerate).

The two properties can be combined: we say that parameter independence holds for $\mathbf{p}$ if the random vectors

$$(\mathbf{p}_{l|\underline{k}}^{v|\mathfrak{p}_v}, l \in \mathcal{I}_v), \qquad \underline{k} \in \mathcal{I}_{\mathfrak{p}_v}, v \in V$$

are independent (nondegenerate). Analogously, we say that parameter independence holds for $\mathbf{p}$ with respect to a family $\mathcal{P}$ of DAGs if it holds for any DAG in $\mathcal{P}$.

We immediately note that under the condition of parameter independence with respect to a DAG having parent function $\mathfrak{p}$, for any d-dimensional table $(r_{\underline{i}}, \underline{i} \in \mathcal{I})$ of nonnegative integers and $r_{(\underline{k},l)}^{\mathfrak{q}_v}$ defined in (2.9), we have, as in Section 4.1,

$$\mathbb{E}\prod_{\underline{i}\in\mathcal{I}}[\mathbf{p}(\underline{i})]^{r_{\underline{i}}} = \prod_{v\in V}\prod_{\underline{k}\in I_{\mathfrak{p}_v}} \mathbb{E}\prod_{l\in\mathcal{I}_v}[\mathbf{p}_{l|\underline{k}}^{v|\mathfrak{p}_v}]^{r_{(\underline{k},l)}^{\mathfrak{q}_v}}. \tag{6.1}$$

This property of parameter independence was crucial for the characterization of the Dirichlet distribution (for a complete graph) given in [9]. We will now extend this characterization to the $\mathcal{P}$-Dirichlet distribution. For particular choices of $\mathcal{P}$, it will yield a characterization of the hyper Dirichlet law. Families $\mathcal{P}$ of DAGs will need to have the separation property which we define now.

DEFINITION 6.2. A family $\mathcal{P}$ of DAGs with skeleton G is called *separating* if

$$\forall v \in V \ \exists \mathfrak{p}, \mathfrak{p}' \in \mathcal{P} \text{ such that } \mathfrak{p}_v \neq \mathfrak{p}'_v. \tag{6.2}$$

Our main result, Theorem 6.1 below, shows that for separating families of moral DAGs parameter independence characterizes the $\mathcal{P}$-Dirichlet distribution.

THEOREM 6.1. *Let* $\mathbf{p}$ *be a vector of random probabilities associated with the graph* G. *Let* $\mathcal{P}$ *be a separating family of moral DAGs with skeleton a decomposable graph* $G = (V, E)$.

If parameter independence for $\mathbf{p}$ *with respect to* $\mathcal{P}$ *holds, then* $\mathbf{p}$ *follows a* $\mathcal{P}$*-Dirichlet distribution.*

The proof is given in Section 6.1 of the supplementary material [15].

We emphasize that if $\mathcal{P}$ is not a separating family then parameter independence alone does not imply that $\mathbf{p}$ follows a $\mathcal{P}$-Dirichlet distribution. For example, consider the three-chain with cliques $\{1, 2\}$ and $\{2, 3\}$. Let $\mathcal{P} = \{\mathfrak{p} \equiv 1 \to 2 \to 3, \mathfrak{p}' \equiv 1 \leftarrow 2 \to 3\}$ and let us assume that $(\mathbf{p}_{\underline{m}}^{12}, \underline{m} \in \mathcal{I}_{12})$ follows a classical Dirichlet distribution while $(\mathbf{p}_{n|k}^{3|2}, n \in \mathcal{I}_3)$, $k \in \mathcal{I}_2$, has an arbitrary distribution and is independent of $(\mathbf{p}_{\underline{m}}^{12}, \underline{m} \in \mathcal{I}_{12})$. Clearly, parameter independence for $\mathfrak{p}$ and $\mathfrak{p}'$ is satisfied. Yet, $\mathcal{P}$ is not a separating family since $\mathfrak{p}_3 = \mathfrak{p}'_3$.

The hyper Dirichlet distribution is characterized by parameter independence with respect to a $\mathcal{P}$ which is rich enough. We state this result more precisely in the theorem below which is an immediate consequence of Theorems 6.1 and 4.4.

THEOREM 6.2. *Let* $\mathbf{p}$ *be a vector of random probabilities associated with the graph* G. *Let* $\mathcal{P}$, *a separating family of moral DAGs with skeleton* G, *satisfy* (4.15) *and* (4.16).

If parameter independence for $\mathbf{p}$ *with respect to* $\mathcal{P}$ *holds then* $\mathbf{p}$ *has a hyper Dirichlet distribution.*

6.2. *Special cases.* We now consider some particular graphs.

(a) *The chain and the hyper Dirichlet distribution.* Let $G = (V, E)$ be a chain with vertices $V = \{1, \ldots, d\}$ and edges $E = \{\{i, i + 1\}, i = 1, \ldots, d - 1\}$.

Consider the family $\mathcal{P} = \{\mathfrak{p}, \mathfrak{p}'\}$, where $\mathfrak{p}$ is a DAG following the chain from 1 to d and $\mathfrak{p}'$ is the DAG with reversed arrows. Then $\mathcal{P}$ is separating, $\mathfrak{R}_{\mathfrak{p}} = \mathfrak{p}(V) \setminus \mathcal{S} = \{\{1\}\}$ and $\mathfrak{R}_{\mathfrak{p}'} = \mathfrak{p}'(V) \setminus \mathcal{S} = \{\{d\}\}$ so that $\mathfrak{R} = \varnothing$, that is, (4.15) holds. Moreover, condition (4.16) is clearly satisfied. Thus, from Theorem 6.2 we conclude the following result.

COROLLARY 6.3. *Assume that the random vectors*

$$\big(\mathbf{p}_{l|k}^{j|j-1}, l \in \mathcal{I}_j\big), \qquad k \in \mathcal{I}_{j-1}, j = 1, 2, \ldots, d \tag{6.3}$$

are jointly independent.

Assume also that the random vectors

$$\big(\mathbf{p}_{l|m}^{j|j+1}, l \in \mathcal{I}_j\big), \qquad m \in \mathcal{I}_{j+1}, j = 1, 2, \ldots, d \tag{6.4}$$

are jointly independent.

Then $\mathbf{p}$ *has a hyper Dirichlet distribution with respect to* G.

In the assumptions above, for ease of notation, we used the convention that $X_0 = X_{d+1} \equiv 0$ and $\mathcal{I}_0 = \mathcal{I}_{d+1} = \{0\}$.

Note that the family $\mathcal{P}$ that we defined for the chain is the unique minimal separating family. That is any other family of DAGs is either nonseparating or it contains $\mathcal{P}$ as a proper subset.

Note also that for the two-chain (that is when $d = 2$) we obtain the characterization of the classical Dirichlet distribution given in Theorem 2 of [3]. At the same time, we extend the characterization given in [9], Theorem 2, where additionally to parameter independences (6.3) and (6.4) it was assumed that densities exist and are sufficiently regular. Some of the regularity assumptions were considerably weakened in [12]. More recently, the entire Chapter 23 of the monograph [13] was devoted to this issue.

(b) *The tree and the hyper Dirichlet distribution.* Let $G = (V, E)$ be a tree. As in the case of the chain the set of cliques $\mathcal{C}$ is equal to E and $\mathcal{S} = \{\varnothing, \{v\}, v \in V \setminus L\}$, where $L \subset V$ is the set of leaves, that is, those vertices which belong to exactly one edge. Any DAG can be uniquely defined by choosing a vertex v such that $\mathfrak{p}_v = \varnothing$. We will denote this DAG by $\mathfrak{p}^{(v)}$. Note that for any such $\mathfrak{p}^{(v)}$, for any $w \in V \setminus \{v\}$, the set $\mathfrak{p}_w^{(v)}$ contains exactly one element.

Consider the family $\mathcal{P} = \{\mathfrak{p}^{(v)}, v \in L\}$. Note that each vertex on the unique chain connecting v and w in L has different parents under $\mathfrak{p}^{(v)}$ and $\mathfrak{p}^{(w)}$. Therefore $\mathcal{P}$ is a separating family. Since any separator consists of only one vertex and any clique of only two vertices condition (4.16) follows from the same observation. Since $\mathfrak{p}^{(v)}(V) = \mathcal{S} \cup \{\{v\}\}$, it follows that $\mathfrak{R}_{\mathfrak{p}^{(v)}} = \{\{v\}\}$. Consequently, (4.15) is satisfied. From Theorem 6.2 we have the following result for trees.

COROLLARY 6.4. *Assume that for every leaf $v \in L$ parameter independence with respect to $\mathfrak{p}^{(v)}$ holds. Then $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ follows the hyper Dirichlet distribution with respect to G.*

(c) *The complete graph and the classical Dirichlet distribution.* Consider a complete graph $G = (V, E)$. Then $\mathcal{C} = \{V\}$ and $\mathcal{S} = \{\varnothing\}$. Consider two DAGs with parent functions $\mathfrak{p}$ and $\mathfrak{p}'$. Let $v_1, \ldots, v_d$ be the numbering corresponding to $\mathfrak{p}$, that is, $\mathfrak{p}_{v_i} = \mathfrak{q}_{v_{i-1}}$ with $\mathfrak{p}_{v_1} = \varnothing$ and similarly for $v'_1, \ldots, v'_d$ corresponding to $\mathfrak{p}'$. Let us assume moreover that

$$\forall j = 1, \ldots, d, \qquad \mathfrak{p}_{v_j} \neq \mathfrak{p}'_{v'_j}. \tag{6.5}$$

Since the cardinality of $\mathfrak{p}_{v_j}$ is equal to $j - 1$, $\mathfrak{p}_{v_j} = \mathfrak{p}_{v'_k}$ implies $j = k$. But this is forbidden for $j = 1, \ldots, d-1$ by condition (6.5). Therefore, the family $\mathcal{P} = \{\mathfrak{p}, \mathfrak{p}'\}$ is separating. A similar cardinality argument combined with (6.5) implies that $\mathfrak{R}_{\mathfrak{p}} = \mathfrak{p}(V) \setminus \mathcal{S} = \{\mathfrak{p}_{v_j}, j = 2, \ldots, d\}$ and $\mathfrak{R}_{\mathfrak{p}'} = \mathfrak{p}'(V) \setminus \mathcal{S} = \{\mathfrak{p}'_{v'_j}, j = 2, \ldots, d\}$ cannot have a common element. Thus, (4.15) is satisfied and we have the following result.

COROLLARY 6.5. *If for a complete graph $G = (V, E)$ parameter independence holds for any two DAGs satisfying (6.5) then $(\mathbf{p}(\underline{i}), \underline{i} \in \mathcal{I})$ has a classical Dirichlet distribution.*

In Theorem 7 of [10] (see also [9], Theorem 3), the authors considered parameter independence for $\mathcal{P} = (\mathfrak{p}, \mathfrak{p}')$, where

$$\mathfrak{p}_1 = \varnothing, \qquad \mathfrak{p}_i = \{1, \ldots, i-1\}, \qquad i = 2, \ldots, d$$

and

$$\mathfrak{p}'_i = \{d\} \cup \mathfrak{p}_i, \qquad i = 1, \ldots, d-1, \qquad \mathfrak{p}'_d = \varnothing.$$

With such a choice of $\mathcal{P}$ the case $d > 2$ could be reduced to the case $d = 2$ discussed already in Section 6.2(a). Clearly, $\mathfrak{p}$ and $\mathfrak{p}'$ above satisfy (6.5) and the characterization of [10] follows from Corollary 6.5 without the assumption of existence and smoothness of densities made in that paper.

For $d = 2$, Corollary 6.5 can be found in [3]. Because this paper was focused on neutralities with respect to partitions the authors assumed independence of

$$(\mathbf{p}_i^l, i \in \mathcal{I}_l), \qquad (\mathbf{p}_{i_{V\setminus\{l\}|i}}^{V\setminus\{l\}|l}, \underline{i}_{V\setminus\{l\}} \in \mathcal{I}_{V\setminus\{l\}}), \qquad i \in \mathcal{I}_l$$

for $l = 1, \ldots, d$. For related characterizations of the classical Dirichlet the reader is referred to [2, 4, 5, 8, 11, 18, 19] and the monograph [16], Chapter 2.6.

7. Conclusion. This paper makes two contributions. One contribution is the development of a new family of flexible conjugate prior distributions with the directed strong hyper Markov property, for a family $\mathcal{P}$ of moral DAGs with the same decomposable skeleton. This family, the $\mathcal{P}$-Dirichlet, generalizes the hyper Dirichlet and the Dirichlet. The development of this family shows that the more elements $\mathcal{P}$ contains, the more restrictions we will have on the set of hyper parameters of the $\mathcal{P}$-Dirichlet. With this family, we have also defined the new objects $\mathfrak{P}$ and $\mathfrak{Q}$ generalizing the notion of cliques and separators in a decomposable graph G.

The other contribution is a characterization of this new family of distributions and, as a consequence, of the hyper Dirichlet and the classical Dirichlet. This characterization is based on local and global parameter independence and does not require the assumption of existence and smoothness of the density. We have also shed light on the choice of the two DAGs on a complete graph used in [9] for the characterization of the Dirichlet: we have done so by emphasizing that these two DAGs form a separating family and that this particular choice of two DAGs, important for the method of proof in [9], is only one of many possible choices.

The $\mathcal{P}$-Dirichlet finds its application in Bayesian estimation when we can easily build in expert prior knowledge into our prior. We have also indicated a possible extension of the $\mathcal{P}$-Dirichlet to a distribution where $\mathcal{P}$ would be the family of immoral DAGs represented by an essential graph. This development will be the subject of future research.

Acknowledgments. We wish to thank the Associate Editor and the referee for constructive and helpful comments. We are particularly grateful for the suggestion to consider essential graphs.

SUPPLEMENTARY MATERIAL

Proofs and some detailed examples for "A new prior for discrete DAG models with a restricted set of directions" (DOI: 10.1214/15-AOS1396SUPP; .pdf). Supplement A contains proofs and examples. We provide the proof of Lemma 2.1 and give a simple example of $\mathfrak{p}$-perfect ordering of the cliques and vertices as given in (2.1). We also provide the proofs of Theorems 4.1, 5.1, 5.2 and 6.1. We give the details of the derivation of the four $\mathcal{P}$-Dirichlet families in Example 3.1 *continued*. We also illustrate, with two examples, a possible extension of the $\mathcal{P}$-Dirichlet distribution to arbitrary DAGs through the use of essential graphs.

REFERENCES

[1] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

[2] BOBECKA, K. and WESOŁOWSKI, J. (2007). The Dirichlet distribution and process through neutralities. *J. Theoret. Probab.* **20** 295–308. MR2324532

[3] BOBECKA, K. and WESOŁOWSKI, J. (2009). Moments method approach to characterizations of Dirichlet tables through neutralities. *Publ. Math. Debrecen* **74** 321–339. MR2521378

[4] CHANG, W.-Y., GUPTA, R. D. and RICHARDS, D. ST. P. (2010). Structural properties of the generalized Dirichlet distributions. In *Algebraic Methods in Statistics and Probability II* (M. A. G. Viana and H. P. Wynn, eds.). *Contemp. Math.* **516** 109–124. Amer. Math. Soc., Providence, RI. MR2730744

[5] DARROCH, J. N. and RATCLIFF, D. (1971). A characterization of the Dirichlet distribution. *J. Amer. Statist. Assoc.* **66** 641–643. MR0292213

[6] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

[7] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[8] FABIUS, J. (1973). Two characterizations of the Dirichlet distribution. *Ann. Statist.* **1** 583–587. MR0353531

[9] GEIGER, D. and HECKERMAN, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Ann. Statist.* **25** 1344–1369. MR1447755

[10] HECKERMAN, GEIGER, D. and CHICKERING, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.* **20** 197–243.

[11] JAMES, I. R. and MOSIMANN, J. E. (1980). A new characterization of the Dirichlet distribution through neutrality. *Ann. Statist.* **8** 183–189. MR0557563

[12] JÁRAI, A. (1998). Regularity property of the functional equation of the Dirichlet distribution. *Aequationes Math.* **56** 37–46. MR1628311

[13] JÁRAI, A. (2005). *Regularity Properties of Functional Equations in Several Variables*. *Advances in Mathematics* (*Springer*) **8**. Springer, New York. MR2130441

[14] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

[15] MASSAM, H. and WESOŁOWSKI, J. (2015). Supplement to "A new prior for discrete DAG models with a restricted set of directions." DOI:10.1214/15-AOS1396SUPP.

[16] NG, K. W., TIAN, G.-L. and TANG, M.-L. (2011). *Dirichlet and Related Distributions*. Wiley, Chichester. MR2830563

[17] RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014

[18] RAMAMOORTHI, R. V. and SANGALLI, L. M. (2007). On a characterization of the Dirichlet distribution. In *Bayesian Statistics and Its Applications* (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) 385–397. Anamaya Publishers, New Delhi.

[19] SAKOWICZ, A. and WESOŁOWSKI, J. (2014). Dirichlet distribution through neutralities with respect to two partitions. *J. Multivariate Anal.* **129** 1–15. MR3215975

[20] STUDENÝ, M. and VOMLEL, J. (2009). A reconstruction algorithm for the essential graph. *Internat. J. Approx. Reason.* **50** 385–413. MR2514506

DEPARTMENT OF MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, M3J1P3
CANADA
E-MAIL: massamh@yorku.ca

WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH
POLITECHNIKA WARSZAWSKA
WARSZAWA
POLAND
E-MAIL: J.Wesolowski@mini.pw.edu.pl