# Global Joint Distribution Factorizes into Local Marginal Distributions on Tree-Structured Graphs

**Xinhua Zhang** XINHUA.ZHANG@ANU.EDU.AU

*Research School of Information Sciences and Engineering*
*The Australian National University, Canberra ACT 0200, Australia*

*Statistical Machine Learning Program*
*National ICT Australia, Canberra, Australia*

## 1. Introduction

In this note, we present a self-contained proof of the following property in tree-structured graphs, including trees, junction trees, and hypertrees:

*The global joint distribution of any tree-structured graph factorizes in terms of the local marginal distributions.*

In tree-structured graphs, this property plays a central role in the proof of many properties, which are unique to trees amongst all arbitrary graph topologies. These properties include, but not limited to:

1. Local consistency guarantees the global consistency (see Proposition 2 and 4 below for proof),
2. Inference on tree-structured graphs can be performed efficiently,
3. The Bethe approximation is exact on tree-structured graphs (see proof in section 4.2.3 of (Wainwright and Jordan, 2003)). In essence, the global entropy decomposes into local entropies.

Many papers quote these results as given. Here we give a self-contained proof, using almost only the fundamental definitions. The style is a little verbose, but we want to highlight some subtle confusions and misunderstandings in this topic. The main objective is for you to avoid the following awkward moments:

| | |
|---|---|
| Katherine: | Hi Jack. |
| | Do you know local consistency on trees implies global consistency? |
| Jack: | Sure, everybody knows. Simple! I use it everyday. |
| Katherine: | Fantastic. Could you show me the proof? |
| Jack: | Obvious, hmmm..., let's see. |
| | (10 minutes) |
| | hmmm... I refer you to Jordan's book. (In fact, I don't know ☺) |
| Katherine: | Then what does global consistency mean? |
| Jack: | Well, I guess you should also read Jordan's. (I am not sure as well.) |

First of all, we introduce some notation. Suppose the graph is $G = \langle V(G), E(G) \rangle$, where $V(G)$ is the set of nodes and $E(G)$ is the set of edges. When $G$ is clear from the context, we just write $V$ and $E$ for simplicity. Associated with each node $s \in V$ is a random variable $x_s$ taking values in some set $\mathcal{X}_s$ called state space, which can be either continuous (e.g., $\mathcal{X}_s \in \mathcal{R}$) or discrete (e.g., $\mathcal{X}_s \in \{1, ..., m\}$). For a subset $A$ of the node set $V$, we define $x_A := \{x_s | s \in A\}$ and use $x$ as a shorthand for $x_V$. We use notations like $p(x_s, x_t)$ and $p\left(x_{\{s,t\}}\right)$ interchangeably.

## 2. Trees

We will start with trees, and then extend to more general tree-structured graphs. In a tree, associated with each edge $(s,t)$ is a non-negative edge potential function $\psi_{st}(x_s, x_t)$, and associated with each node is a non-negative node potential function $\psi_s(x_s)$. The joint distribution is defined by

$$p(x) := \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \tag{1}$$

where $Z$ is the normalization factor (partition function). On trees, the property of factorization can be formally expressed in Proposition 1:

**Proposition 1** *If the graph $T$ is a tree, then*

$$p(x) = \prod_{s \in V(T)} p(x_s) \prod_{(s,t) \in E(T)} \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)}.$$

To prove this result, we actually prove a strengthened result, which places a tree in a more general graph.

**Lemma 1** *For any graph $G$, suppose $T$ is a connected subgraph of $G$. Assume for every pair of nodes in $T$, there is a single path in $\boldsymbol{G}$ connecting them. Then the marginal distribution of $p(x_T)$ factorizes as follows:*

$$p(x_T) = \prod_{s \in V(T)} p(x_s) \prod_{(s,t) \in E(T)} \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)} \tag{2}$$

*where $V(T)$ and $E(T)$ are the set of nodes and edges of $T$, respectively.*

**Remark 1** *1. The precondition of Lemma 1 ensures that $T$ is a tree. For any two different nodes $s, t \in V(T)$, they are singly connected in $\boldsymbol{G}$, which means that there is a unique path between them, namely $sv_1v_2 \cdots v_nt$ ($v_i \in V(G)$). But since $T$ is a connected subgraph and the path is unique in $G$, so all $v_i$ must be in $V(T)$. Figure 1 gives an example. The whole graph $G$ consists of nodes $A$ to $L$, and $G$ is not a tree. The subgraph composed of nodes $A$ to $F$ forms a tree in $G$. But if we add an edge between node $H$ and $L$, then node set $\{A, B, \cdots, F\}$ no longer forms a tree in $G$ because between $A$ and $C$ there are two paths in $G$, namely $AC$ and $AHLC$.*

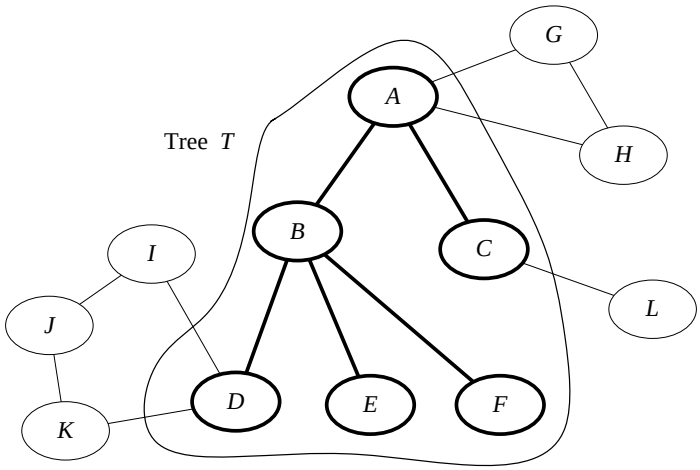



Figure 1: Example of a tree in a general graph $G$. The nodes and edges in bold form the tree $T$.

*2. The $p(x_T)$ in Eq (2) is actually given by*

$$p(x_T) = \sum_{x_{V(G) \setminus V(T)}} p(x), \tag{3}$$

*i.e., marginalizing out $x_{V(G)\setminus V(T)}$ from the joint distribution $p(x)$ on $x_{V(G)}$. A common confusion is assuming that*

$$p(x_T) := \frac{1}{Z_T} \prod_{s \in V(T)} \psi_s(x_s) \prod_{(s,t) \in E(T)} \psi_{st}(x_s, x_t), \tag{4}$$

*which is not what we mean here. And in general, the $p(x_T)$ given by Eq (3) does not necessarily has the form of Eq (4), where the $\psi$'s are the potential functions used to define the whole joint distribution $p(x_G)$.*

*3. Proposition 1 is a special case of Lemma 1, by choosing $T = G$.*

*4. Lemma 1 does not say that $p(x_T)$ is independent of all the $x_s$ where $s \in V(G) \backslash V(T)$. Generally speaking, the node potentials $\psi_s(x_s)$ and edge potentials $\psi_{st}(x_s, x_t)$ $(s, t \in V(G) \backslash V(T))$ DO affect $p(x_T)$, as long as there is a path connecting $s$ or $t$ to $T$. For example, in Figure 1, $\psi_H(x_H)$ affects $p(x_T)$. But since they also affect the right-hand side (RHS) of Eq (2), so Eq (2) can still hold.*

*5. Equation (2) can be rewritten as*

$$p(x) = \prod_{s \in V(T)} p(x_s) \prod_{(s,t) \in E(T)} \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)} = \frac{\prod_{(s,t) \in E(T)} p(x_s, x_t)}{\prod_{s \in V(T)} p(x_s)^{d_s - 1}},$$

*where $d_s$ stands for the degree of node $s$ in the tree. For example, in Figure 1, $d_B = 4$ in tree $T$, and $d_A = 2$ in $T$(not counting in the edges linking to node $G$ and $H$).*

**Proof.** We prove Lemma 1 by induction on the number of nodes in $T$. If $|V(T)| = 1$, i.e., the tree $T$ is just a single node, then Eq (2) obviously holds. Suppose Eq (2) holds for any $|V(T)| < k$ $(k > 1)$. Then for an arbitrary subgraph tree $T$ with $|V(T)| = k$, since it is a tree, $T$ must have a leaf node $s$, i.e., whose degree is 1 in $T$ (i.e., only one adjacent node in $T$, though it may have other neighbors in $G \backslash T$). Let that neighbor be $t$ and denote $U = T \backslash \{s, t\}$. So $U \cup \{t\}$ is a tree. Refer to Figure 2.

Then by the graph structure, we have

$$\begin{aligned} p(x_s, x_t, x_U) &= p(x_s, x_t)\, p(x_U | x_s, x_t) \\ &\overset{(a)}{=} p(x_s, x_t)\, p(x_U | x_t) = p(x_s) \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)} p(x_U, x_t), \end{aligned} \tag{5}$$

where (a) is because $x_U \perp\!\!\!\perp x_s | x_t$ due to the topology of $G$, in which the only path connecting node $s$ and nodes in $U$ must be via node $t$.

Since $U \cup \{t\}$ is also a tree, which we denote as $T'$, and its cardinality is $k - 1$, so by the assumption of induction, we have

$$p(x_U, x_t) = \prod_{p \in U \cup \{t\}} p(x_p) \prod_{(p,q) \in E(T')} \frac{p(x_p, x_q)}{p(x_p)\, p(x_q)}, \tag{6}$$

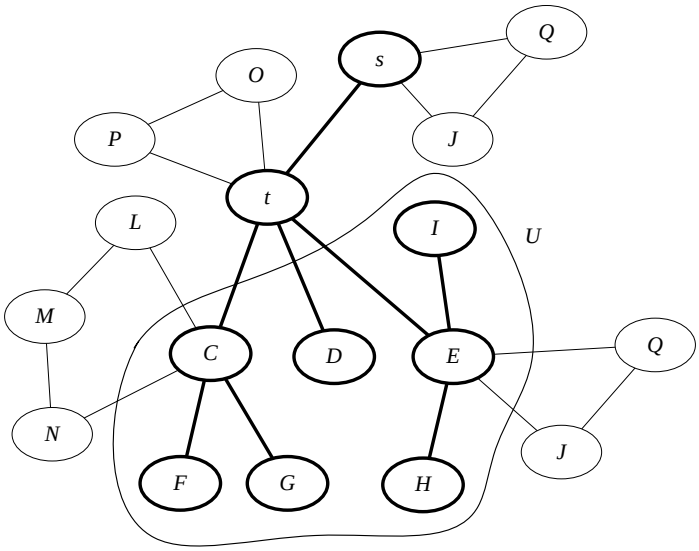



Figure 2: Illustration of Lemma 1 proof.

where the second product is defined as 1 if $T'$ consists of a single node $t$ only (i.e., $U = \emptyset$). Continuing Eq (5), we have

$$p(x_s, x_t, x_U) = p(x_s) \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)} \prod_{p \in U \cup \{t\}} p(x_p) \prod_{(p,q) \in E(T')} \frac{p(x_p, x_q)}{p(x_p)\, p(x_q)}$$

$$= \prod_{p \in V(T)} p(x_p) \prod_{(p,q) \in E(T)} \frac{p(x_p, x_q)}{p(x_p)\, p(x_q)}.$$

So Eq (2) also holds for tree $T$ with $|V(T)| = k$. So by induction, we have proven that Eq (2) holds for all trees. ■

It looks redundant to put a tree in a general graph and prove the strengthened result. However, the reason is that otherwise we will have trouble when invoking Eq 6, since the tree $U \cup \{t\}$ is in a bigger tree $U \cup \{s, t\}$. If we directly prove the Proposition 1 by induction, then the induction assumption does *not* allow us to invoke Eq 6 (at least not directly).

Based on Proposition 1, we are able to prove the important result about the relationship between local consistency and global consistency on a tree. First of all, we describe in detail the meanings of local and global consistency.

Suppose we are given a set of marginal distributions on all cliques: $\{p_c : c \in \mathcal{C}\}$ where $\mathcal{C}$ is the set of all cliques in the graph in general. In the special case of trees, $\mathcal{C}$ consists of all edges and all nodes. In general, we say $\{p_c : c \in \mathcal{C}\}$ is *locally consistent* if the following two conditions are satisfied:

L1. *Validity* (non-negativity and normalization): For all cliques $c \in \mathcal{C}$, $\sum_{x_c} p_c(x_c) = 1$, and $p_c(x_c) \geq 0$ for all configuration $x_c$.

L2. *Consistency*: For all cliques $s, t \in \mathcal{C}$, if $c := s \cap t \neq \emptyset$, then for all assignment $x_c$:

$$\sum_{x'_s : x'_c = x_c} p_s(x'_s) = \sum_{x'_t : x'_c = x_c} p_t(x'_t).$$

In other words, the marginal distribution of $c$ calculated from clique $s$ must agree with that calculated from clique $t$. Written in function form: $\sum_{x_{s\setminus c}} p_s(x_c, x_{s\setminus c}) = \sum_{x_{t\setminus c}} p_t(x_c, x_{t\setminus c})$ [1].

We say $\{p_c : c \in \mathcal{C}\}$ is *globally consistent* if there exists a global joint distribution $p(x)$ on $x_{V(G)}$, such that the following two conditions are satisfied:

G1. *Validity*: $p(x) \geq 0$ for all configuration $x$, and $\sum_x p(x) = 1$,

G2. *Consistency*: $\sum_{x' : x'_c = x_c} p(x') = p_c(x_c)$ for all clique $c \in \mathcal{C}$ and $x_c$. Written in function form: $\sum_{x_{V(G)\setminus c}} p(x) = p_c(x_c)$ for clique $c \in \mathcal{C}$ (note we are deliberately not saying for all $x_c$, because this equality is already between functions).

It is obvious that for any graph and any distribution, global consistency implies local consistency. But the reverse direction is not necessarily true. A classic example is illustrated in Figure 3.

In Figure 3, suppose all random variables $x_A, x_B, x_C$ are binary ($\{0, 1\}$). Consider the following marginals on nodes and edges:

$p(x_A = 0) = p(x_A = 1) = 0.5$, $p(x_B = 0) = p(x_B = 1) = 0.5$, $p(x_C = 0) = p(x_C = 1) = 0.5$.

1. Normally, people just write $\sum_{x_{s\setminus c}} p_s(x_s) = \sum_{x_{t\setminus c}} p_t(x_t)$, which in appearance, does not explicitly say the variable $x_c$ assumes the same value in the LHS and RHS. However, this assumption is made explicitly if we write $\sum_{x_{s\setminus c}} p_s(x_c, x_{s\setminus c}) = \sum_{x_{t\setminus c}} p_t(x_c, x_{t\setminus c})$. Albeit a standard notation, if you think of it carefully, this new notation does not make immediate mathematical sense, and needs some explanation. Here $p_s(x_s)$ represents a function, just like what we normally write $f(x)$ as a function. However, if I write $f(x_0)$ or $f(\hat{x})$, then chances are that you will feel it is a particular value after applying a function $f$ on $x_0$ or $\hat{x}$. The meaning should of course not depend on the symbol of variable, and that is why I call it a notational confusion. Sometimes people write $f(\cdot)$ to clearly represent a function, or just write $f$. This is particularly useful when one talks about functional spaces (spaces of functions). So now, let us think of $p_s(x_s)$ as a function over $x_s$. Then $\sum_{x_{s\setminus c}} p_s(x_s)$ obviously represents a function over $x_c$. So $\sum_{x_{s\setminus c}} p_s(x_s) = \sum_{x_{t\setminus c}} p_t(x_t)$ is actually an equality between two *functions*! In this case, it means the marginal distribution of $x_c$ (a function of the assignment of $x_c$) is the same. We will use this function form when the notation becomes messy otherwise. In fact, we have used it in Eq (3).

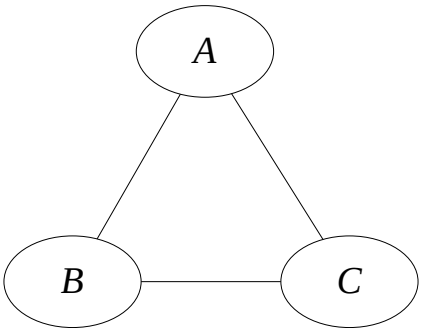



Figure 3: Example of a locally consistent distribution, but not globally consistent.

| $p(x_A x_B)$ | $x_A = 0$ | $x_A = 1$ | $p(x_B x_C)$ | $x_C = 0$ | $x_C = 1$ | $p(x_A x_C)$ | $x_C = 0$ | $x_C = 1$ |
|---|---|---|---|---|---|---|---|---|
| $x_B = 0$ | 0.4 | 0.1 | $x_B = 0$ | 0.4 | 0.1 | $x_A = 0$ | 0.1 | 0.4 |
| $x_B = 1$ | 0.1 | 0.4 | $x_B = 1$ | 0.1 | 0.4 | $x_A = 1$ | 0.4 | 0.1 |

It is easy to check that the marginals are locally consistent. However, one can prove that there doesn't exist any global distribution $p(x_A x_B x_C)$ which yields such a marginal distribution. In fact a quick proof of the non-existence is that if they were globally consistent, then $\mathbb{E}_{p(x_A x_B x_C)}[\![yy^\top]\!]$ would be positive semi-definite, where $y = (1, x_A, x_B, x_C)^\top$. However,

$$\mathbb{E}[\![yy^\top]\!] = \begin{pmatrix} 1 & p(x_A = 1) & p(x_B = 1) & p(x_C = 1) \\ p(x_A = 1) & p(x_A = 1) & p(x_A = x_B = 1) & p(x_A = x_C = 1) \\ p(x_B = 1) & p(x_A = x_B = 1) & p(x_B = 1) & p(x_B = x_C = 1) \\ p(x_C = 1) & p(x_A = x_C = 1) & p(x_B = x_C = 1) & p(x_C = 1) \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.4 & 0.1 \\ 0.5 & 0.4 & 0.5 & 0.4 \\ 0.5 & 0.1 & 0.4 & 0.5 \end{pmatrix},$$

and the last matrix turns out not to be positive semi-definite. In fact, the determinant of the first, second, third and fourth principal minors are 1, 0.25, 0.04, –0.008, respectively. The negativity of the determinant of the matrix alone is enough to disprove the positive semi-definiteness of the matrix.

Fortunately, on trees, it is well-known (though its proof is much less well-known) that local consistency is sufficient to guarantee global consistency. We state it formally in Proposition 2.

**Proposition 2** *On any tree $T$, local consistency implies global consistency. Formally, suppose we are given a set of marginal distributions $\{p_s(\cdot) : s \in V(T)\}$ and $\{p_{st}(\cdot,\cdot) : (s,t) \in E(T)\}$ which satisfy the above two conditions L1 and L2. On trees, L1 and L2 mean:*

*(a1) For any node* $s \in V(T)$, $\sum_{x_s} p_s(x_s) = 1$, *and* $p_s(x_s) \geq 0$ *for all* $x_s \in \mathcal{X}_s$.

*(a2) For any edge* $(s,t) \in E(T)$, $\sum_{x_s,x_t} p_{st}(x_s,x_t) = 1$, *and* $p_{st}(x_s,x_t) \geq 0$ *for all* $(x_s,x_t) \in \mathcal{X}_s \times \mathcal{X}_t$.

*(b) For any edge* $(s,t) \in E(T)$, $\sum_{x_s} p_{st}(x_s,x_t) = p_t(x_t)$ *for all* $x_t \in \mathcal{X}_t$, *and* $\sum_{x_t} p_{st}(x_s,x_t) = p_s(x_s)$ *for all* $x_s \in \mathcal{X}_s$. *In fact, (b) and (a2) implies (a1).*

*Then there must exist a global joint distribution* $\tilde{p}(x)$[2] *satisfying G1 and G2. More specifically,*

*(A)* $\tilde{p}(x) \geq 0$ *for all* $x$, *and* $\sum_x \tilde{p}(x) = 1$,

*(B1)* $\sum_{x':x'_s=x_s} \tilde{p}(x') = p_s(x_s)$ *for all* $x_s \in \mathcal{X}_s$, *and node* $s \in V(T)$,

*(B2)* $\sum_{x':x'_s=x_s,x'_t=x_t} \tilde{p}(x') = p_{st}(x_s,x_t)$ *for all* $(x_s,x_t) \in \mathcal{X}_s \times \mathcal{X}_t$ *and edge* $(s,t) \in E(T)$.

**Proof.** We prove by construction, i.e., by showing that the following global joint distribution $p(x)$ (simply according to Proposition 1) satisfies the above three conditions (A), (B1), and (B2):

$$\tilde{p}(x) = \prod_{s \in V(T)} p_s(x_s) \prod_{(s,t) \in E(T)} \frac{p_{st}(x_s,x_t)}{p_s(x_s)\, p_t(x_t)}. \tag{7}$$

Now we check (A), (B1), and (B2). Obviously $p(x) \geq 0$ for all $x$. Since $p_s(\cdot)$ and $p_{st}(\cdot,\cdot)$ are locally consistent by assumption, it suffices to check (B2)

$$\sum_{x':x'_s=x_s,x'_t=x_t} \tilde{p}(x') = p_{st}(x_s,x_t), \tag{8}$$

for all $(x_s,x_t) \in \mathcal{X}_s \times \mathcal{X}_t$ and $(s,t) \in E(T)$, which implies (B1) in conjunction with local consistency (b), and implies $\sum_x \tilde{p}(x) = 1$ in conjunction with (a2). We prove Eq (8) by induction on the number of nodes in the tree. As a basis, if the tree has only two nodes $s$ and $t$, then $\tilde{p} = p_{st}$ trivially satisfies Eq (8). Suppose Eq (8) holds for any $|V(T)| < k$ ($k > 2$), i.e., marginalizing the joint distribution defined by Eq (7) into every edge recovers the prescribed edge marginal. Then for an arbitrary tree $T$ with $|V(T)| = k$, since it is a tree, $T$ must have a leaf node $s$. Denote its unique neighbor as $t$. Then $T' := T \setminus \{u\}$ is a

2. Perhaps $\tilde{p}$ makes you feel more comfortable than $p$.

tree with $k-1$ nodes. Refer to Figure 2 for illustration. Now we define a joint distribution

$$\begin{aligned}\tilde{p}_T(x) &:= \prod_{u\in V(T)} p_u(x_u) \prod_{(u,v)\in E(T)} \frac{p_{uv}(x_u,x_v)}{p_u(x_u)\,p_v(x_v)} \\ &= p_s(x_s)\,\frac{p_{st}(x_s,x_t)}{p_s(x_s)\,p_t(x_t)} \prod_{u\in V(T')} p_u(x_u) \prod_{(u,v)\in E(T')} \frac{p_{uv}(x_u,x_v)}{p_u(x_u)\,p_v(x_v)}.\end{aligned}$$

By induction assumption,

$$\tilde{q}\left(x_{V(T')}\right) := \prod_{u\in V(T')} p_u(x_u) \prod_{(u,v)\in E(T')} \frac{p_{uv}(x_u,x_v)}{p_u(x_u)\,p_v(x_v)}$$

is a valid global joint distribution on $T'$. So $\tilde{p}_T(x) = \frac{p_{st}(x_s,x_t)}{p_t(x_t)}\tilde{q}\left(x_{V(T')}\right)$. Now observe

i)

$$\sum_{x':x'_s=x_s,x'_t=x_t} \tilde{p}_T(x') = \sum_{x':x'_s=x_s,x'_t=x_t} \frac{p_{st}(x'_s,x'_t)}{p_t(x'_t)}\tilde{q}\left(x'_{V(T')}\right) = \frac{p_{st}(x_s,x_t)}{p_t(x_t)} \sum_{x'_{V(T')}:x'_t=x_t} \tilde{q}\left(x'_{V(T')}\right).$$

Since $\tilde{q}\left(x'_{V(T')}\right)$ is consistent with the local marginals by induction assumption, we have $\sum_{x'_{V(T')}:x'_t=x_t} \tilde{q}\left(x'_{V(T')}\right) = p_t(x_t)$. So

$$\sum_{x':x'_s=x_s,x'_t=x_t} \tilde{p}_T(x') = \frac{p_{st}(x_s,x_t)}{p_t(x_t)} p_t(x_t) = p_{st}(x_s,x_t).$$

ii) Since

$$\sum_{x_s} \tilde{p}_T(x) = \sum_{x_s} \frac{p_{st}(x_s,x_t)}{p_t(x_t)}\tilde{q}\left(x_{V(T')}\right) = \frac{p_t(x_t)}{p_t(x_t)}\tilde{q}\left(x_{V(T')}\right) = \tilde{q}\left(x_{V(T')}\right),$$

so for any edge $(\alpha,\beta)\in E(T')$ and any assignment $(x_\alpha,x_\beta)$, we have

$$\sum_{x':x'_\alpha=x_\alpha,x'_\beta=x_\beta} \tilde{p}_T(x') = \sum_{x':x'_\alpha=x_\alpha,x'_\beta=x_\beta} \sum_{x'_s} \tilde{p}_T(x') = \sum_{x'_{V(T')}:x'_\alpha=x_\alpha,x'_\beta=x_\beta} \tilde{q}\left(x'_{V(T')}\right) = p_{\alpha\beta}(x_\alpha,x_\beta).$$

Combining i) and ii), we have shown that Eq (8) holds for all edges in $T$. So by induction, we have shown that Eq (8) holds for all trees and hence local consistency implies global consistency on all trees. ■

## 3. Junction Trees

One immediate generalization of trees is the junction tree. In a junction tree, each node is a collection of the original nodes, i.e., a subset of $V(G)$, and are called *clique nodes*. The original graph $G$ need not be a tree. The edges in the junction tree ensure that the topology is a tree and the *running intersection property* is satisfied:

*For every pair of clique nodes $V$ and $W$, all clique nodes on the unique path between $V$ and $W$ contain $V \cap W$.*

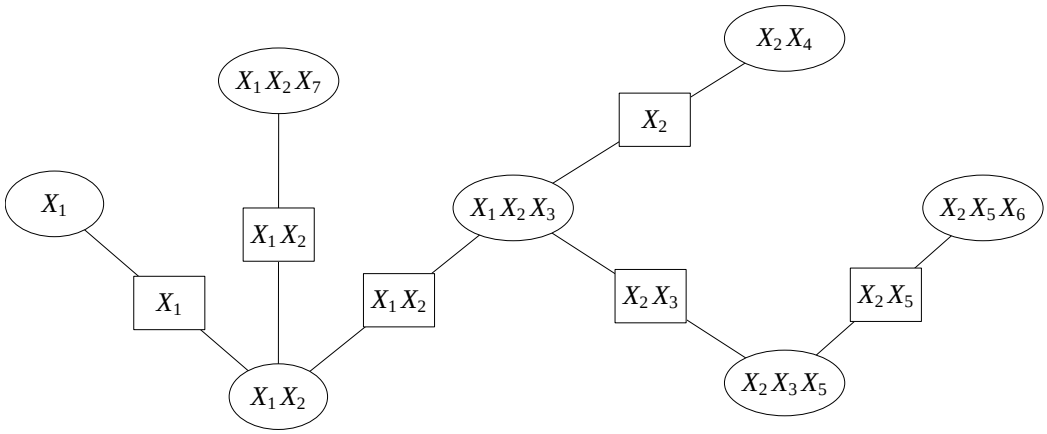


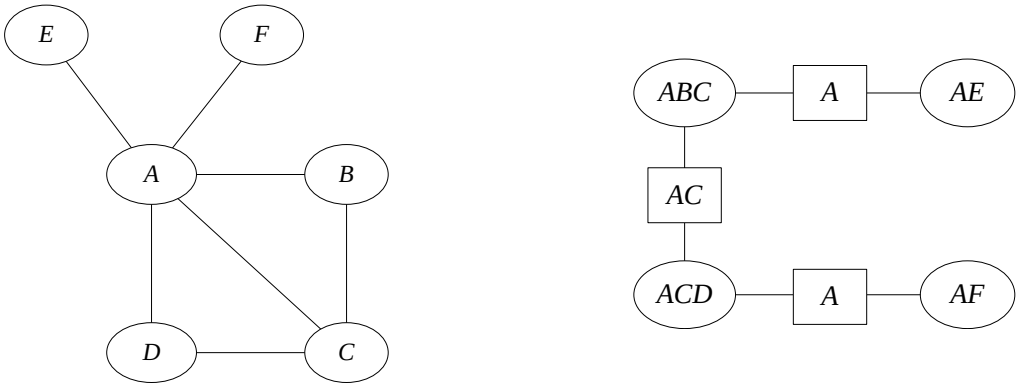

Figure 4: A junction tree example. Clique nodes ($C(J)$) are ellipses and sepset nodes ($S(J)$) are rectangles.

For each edge in the junction tree, we introduce a set called separator set (sepset) defined as the intersection of the two end clique nodes. For clarity, in a junction tree $J$, we call the set of clique nodes as $C(J)$, and the set of sepset as $S(J)$. For example, in Figure 4, we have:

$$C(J) = \{\{X_1\}, \{X_1X_2\}, \{X_1X_2X_3\}, \{X_2X_4\}, \{X_2X_3X_5\}, \{X_2X_5X_6\}, \{X_1X_2X_7\}\},$$
$$S(J) = \{\{X_1\}, \{X_1X_2\}, \{X_2\}, \{X_2X_3\}, \{X_2X_5\}\}.$$

Note that the sepset $\{X_1X_2\}$ appears twice in $J$. However, since the definition of set does not allow duplicate elements, we use $d_c$ to denote ***one plus*** how many times a sepset node $c \in S(J)$ appears in $J$. For example, $d_{\{X_1X_2\}} = 3$. This "one plus" is to comply with the common notation, e.g., (Wainwright and Jordan, 2003), which (on page 15) claims $d_c$ is the number of maximal cliques to which $c$ is adjacent. However, even though we have the running intersection property, the following claim still does *not* necessarily hold:

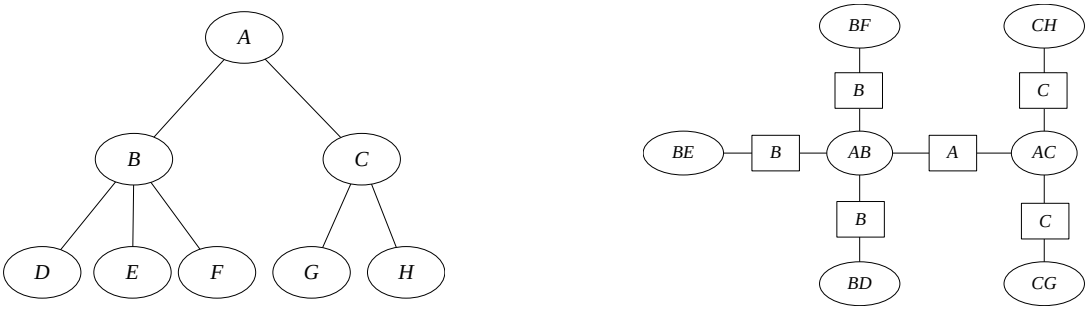
*If a sepset node $c$ appears for $x$ times in a junction tree, then $c$ must be adjacent to $x+1$ clique nodes.*




(a) Original graph

(b) Corresponding junction tree

Figure 5: A counter-example of *#adjacent max clique* $= 1 +$ *#occurrance*.




(a) Original tree

(b) Corresponding junction tree

Figure 6: An example of a tree and its corresponding junction tree.

A counter-example is given in Figure 5, where the sepset node $\{A\}$ appears twice but is adjacent to 4 clique nodes. *So in this note, we stick to the definition of one plus the multiplicity.*

When the original original graph is a tree $T$, then $d_{\{s\}} = d_s$, where the second $d_s$ is the degree of node $s$ in $T$ (ref. point 5 of Remark 1). For example, in Figure 6b, we have $d_{\{B\}} = 3 + 1 = 4$ (since $\{B\}$ appears for three times as a sepset), while in Figure 6a, $d_B = 4$ as well. In this sense, we call $d_c$ the *degree* of $c$ for $c \in S(J)$. When $s$ is a leaf in $T$, $d_{\{s\}} = 1$ which is consistent with the fact that $\{s\}$ does not appear in the junction tree.

Now the factorization property can be expressed mathematically in Proposition 3.

**Proposition 3** *In a junction tree $J$, let $C(J)$ be the set of clique nodes, and let $S(J)$ be the set of sepset, then the joint distribution $p(x)$ factorizes as*

$$p(x) = \frac{\prod_{c \in C(J)} p(x_c)}{\prod_{c \in S(J)} p(x_c)^{d_c - 1}}. \tag{9}$$

**Remark 2** *Although Proposition 1 and Proposition 3 are expressed in slightly different ways, the former can be easily derived from the latter. If the original graph $T$ is a tree, then every clique node in its corresponding junction tree $J$ corresponds to an edge in $T$, and the sepset of two neighboring clique nodes $\{a, b\}$ and $\{a, c\}$ is $\{a\}$, which corresponds to the common node in the two original edges $(a, b)$ and $(a, c)$. So by using Eq (9), we have*

$$p(x) = \frac{\prod_{c \in C(J)} p(x_c)}{\prod_{c \in S(J)} p(x_c)^{d_c - 1}} \overset{(a)}{=} \frac{\prod_{(s,t) \in E(T)} p(x_s, x_t)}{\prod_{s \in V(T)} p(x_s)^{d_s - 1}} = \prod_{s \in V(T)} p(x_s) \prod_{(s,t) \in E(T)} \frac{p(x_s, x_t)}{p(x_s)\, p(x_t)},$$

*where $d_s$ stands for the degree of node $s$ in $T$. Equality (a) is because for each node $s \in T$, the sepset $\{s\}$ appears for $d_s - 1$ times in $J$.*

Before proving Proposition 3, we need a lemma which essentially tells us how to read off conditional independence relations from a junction tree, in analogy to the conditional independence relations in a tree graph.

**Lemma 2 (The lemma 1 in Jordan's book, Chapter 17)** *Let $C$ be a leaf in a junction tree for a graph with node set $V$. Let $S$ be the associated sepset. Let $R := C \backslash S$ be the set of nodes in $C$ but not in the sepset, and let $U := V \backslash C$ be the set of nodes in $V$ but not in $C$. Then $R \perp\!\!\!\perp U | S$.*

**Proof.** We prove by contradiction. Refer to Figure 7 for illustration. For any arbitrary node $a \in R$, suppose it has a neighboring node $b \in U$ in the original graph. Since $a$ and $b$ are adjacent, there must be a maximal clique node, which contains both $a$ and $b$. This clique node can't be $C$ because $b \notin C$. But $a$ can't be in any clique other than $C$ because otherwise $a$ must belong to $S$ by the running intersection property. Hence no such $b$ exists, and therefore $S$ must separate $a$ from $U$. Since $a \in R$ is arbitrary, $S$ separates $R$ from $U$. ■

Now we turn to proving Proposition 3.

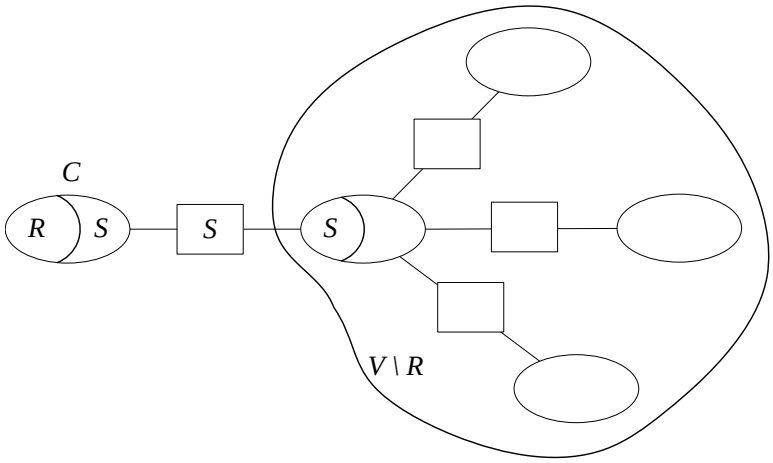



Figure 7: Illustration for Lemma 2.

**Proof.** We prove Proposition 3 by induction. If the junction tree has only one clique node, then Proposition 3 is obviously true (defining $\prod_{c\in S(J)} p(x_c)$ to be 1 if $S(J) = \emptyset$). If $|C(J)| = 2$, it is also easy to verify. Suppose Eq (9) holds for all junction trees $J$ with $|C(J)| < k$ $(k > 1)$. Then for an arbitrary junction trees $J$ with $|C(J)| = k$, since $J$ a tree, $J$ must have a leaf clique node $C$ whose degree is 1 in $J$. Using the same notation as in Lemma 2, we have

$$\begin{aligned} p(x) &= p(x_U|x_R, x_S)\, p(x_R, x_S) \overset{(a)}{=} p(x_U|x_S)\, p(x_R, x_S) \\ &= \frac{p(x_{U\cup S})}{p(x_S)} p(x_R, x_S) = p(x_{U\cup S}) \frac{p(x_C)}{p(x_S)}, \end{aligned} \tag{10}$$

where (a) is by Lemma 2. Observe that after deleting $C$ and $S$ from $J$, the rest of the graph is still a junction tree because $C$ is a leaf clique node. Denote the (smaller) junction tree as $J'$ and $|C(J')| = k - 1$. By induction assumption,

$$P(x_{U\cup S}) = \frac{\prod_{c\in C(J')} p(x_c)}{\prod_{c\in S(J')} p(x_c)^{d'_c - 1}},$$

where $d'_c$ is the degree of sepset node $c$ in $J'$. So Eq (10) continues as

$$p(x) = \frac{\prod_{c\in C(J')} p(x_c)}{\prod_{c\in S(J')} p(x_c)^{d'_c - 1}} \cdot \frac{p(x_C)}{p(x_S)} = \frac{\prod_{c\in C(J)} p(x_c)}{\prod_{c\in S(J)} p(x_c)^{d_c - 1}}.$$

So Eq (9) also holds for any arbitrary junction tree $J$ with $|C(J)| = k$. By induction, we have proven that Eq (9) holds for all junction trees. ■

In analogy to Proposition 2, it is also true that in a junction tree, local consistency implies the global consistency, after a proper re-definition of local and global consistency. In contrast to the ground symbol form used in Proposition 2 and its proof, now we will use the function form (ref. footnote 1).

Suppose a junction tree $J$ has clique node set $C(J)$ and sepset set $S(J)$. We say the marginals $\{p_c : c \in C(J) \cup S(J)\}$ are *locally consistent* if all the following two conditions are satisfied:

JL1. *Validity*: For any clique node or sepset node $c \in C(J) \cup S(J)$, $\sum_{x_c} p_c(x_c) = 1$, and $p_c(x_c) \geq 0$ for all $x_c$;

JL2. *Consistency*: For any clique node $c \in C(J)$, and for any sepset node $s$ adjacent to $c$ and assignment $x_s$, $\sum_{x_{c \setminus s}} p_c(x_c) = p_s(x_s)$. In other words, the marginal distribution calculated from the clique node must agree with the marginal distribution of its associated sepsets.

We say the marginals $\{p_c : c \in C(J) \cup S(J)\}$ are *globally consistent* if there exists a global joint distribution $\tilde{p}(x)$, such that the following two conditions are satisfied:

JG1. *Validity*: $\tilde{p}(x) \geq 0$ for all $x$, and $\sum_x \tilde{p}(x) = 1$;

JG2. *Consistency*: $\sum_{x_{V(J) \setminus c}} \tilde{p}(x) = p_c(x_c)$ for all $c \in C(J) \cup S(J)$.

It is again obvious that for any junction tree and any distribution, global consistency implies local consistency. But the reverse implication is not clear. The following Proposition 4 says the reverse direction also holds.

**Proposition 4** *On any junction tree $J$, local consistency implies global consistency.*

**Proof.** The proof is largely similar to Proposition 2. Suppose we are given a set of marginals $\{p_c : c \in C(J) \cup S(J)\}$ which are locally consistent. With hint from Proposition 3, we construct a global joint distribution

$$\tilde{p}(x) := \frac{\prod_{t \in C(J)} p_t(x_t)}{\prod_{t \in S(J)} p_t(x_t)^{d_t - 1}}. \tag{11}$$

Then we show that $\tilde{p}(x)$ satisfies JG1 and JG2. Obviously, $\tilde{p}(x) \geq 0$ for all $x$. Since $\{p_c : c \in C(J) \cup S(J)\}$ satisfies JL1 and JL2, it suffices to check that for all $c \in C(J)$,

$$\sum_{x_{V(J)\backslash c}} \tilde{p}(x) = p_c(x_c). \tag{12}$$

Again, we prove Eq (12) by induction on the number of clique nodes in a junction tree. If $|C(J)| = 1$, Eq (12) obviously holds if we define $\prod_{t\in\emptyset} p_t(x_t) = 1$. If $|C(J)| = 2$, it is also simple to verify. Now suppose Eq (12) holds for any $|C(J)| < k$ $(k > 2)$, i.e., marginalizing the joint distribution defined by Eq (11) into every clique node recovers the prescribed clique marginal. Then for any arbitrary junction tree $J$ with $|C(J)| = k$, since it is a junction tree, $J$ must have a leaf clique node $C$. Denote its unique adjacent sepset as $S$ (hence $S \subseteq C$), and its unique adjacent clique node as $W$. Then the graph formed by removing $C$ and $S$ from $J$ is still a junction tree, which we call $J'$. $|C(J')| = k - 1$ and

$$V(J') = (V(J)\backslash C) \cup S = V(J)\backslash(C\backslash S). \tag{13}$$

By induction assumption, we have

$$\tilde{q}\left(x_{V(J')}\right) := \frac{\prod_{t\in C(J')} p_t(x_t)}{\prod_{t\in S(J')} p_t(x_t)^{d'_t-1}}$$

is a valid joint distribution on $J'$ which is consistent with $\{p_c : c \in C(J') \cup S(J')\}$. Also notice $\tilde{p}(x) = \frac{p_C(x_C)}{p_S(x_S)}\tilde{q}\left(x_{V(J')}\right)$. So

i)

$$\begin{aligned}\sum_{x_{V(J)\backslash C}} \tilde{p}(x) &= \sum_{x_{V(J)\backslash C}} \frac{p_C(x_C)}{p_S(x_S)}\tilde{q}\left(x_{V(J')}\right) \overset{(a)}{=} \frac{p_C(x_C)}{p_S(x_S)} \sum_{x_{V(J')\backslash S}} \tilde{q}\left(x_{V(J')}\right) \\ &\overset{(b)}{=} \frac{p_C(x_C)}{p_S(x_S)} p_S(x_S) = p_C(x_C),\end{aligned}$$

where (a) is because $S \subseteq C$ and Eq (13), and (b) is because $\tilde{q}$ is consistent with the prescribed marginals on $W$ and hence consistent with $S$ due to JL2.

ii) Since $C$ is a leaf clique node of $J$, so no variable in $C\backslash S$ appears in $J'$. Hence,

$$\sum_{x_{C\backslash S}} \tilde{p}(x) = \sum_{x_{C\backslash S}} \frac{p_C(x_C)}{p_S(x_S)}\tilde{q}\left(x_{V(J')}\right) = \frac{\tilde{q}\left(x_{V(J')}\right)}{p_S(x_S)} \sum_{x_{C\backslash S}} p_C(x_C) = \frac{\tilde{q}\left(x_{V(J')}\right)}{p_S(x_S)} p_S(x_S) = \tilde{q}\left(x_{V(J')}\right),$$

so any further marginalization onto clique nodes in $J'$ will be equal to the prescribed marginals, as guaranteed by the induction assumption on $\tilde{q}\left(x_{V(J')}\right)$.

Combining i) and ii), we obtain that Eq (12) holds for all $c \in C(J)$, and by induction, local consistency implies global consistency in all junction trees. ■

## 4. Hypertrees

Finally, similar properties can be derived for hypertrees. The explanation requires too much effort, so we refer to the Equation 84 and 85 in (Wainwright and Jordan, 2003). The proof of factorization is similar to Proposition 1 and 3, i.e., first identify the conditional independence relationship, and then prove by induction.

## Acknowledgements

The author wishes to thank Dmitry Kamenetsky for his constructive comments, and for pointing out a number of errors and typos in the preliminary versions.

## References

Martin J. Wainwright, and Michael I. Jordan. *Graphical models, exponential families, and variational inference*. UC Berkeley, Dept. of Statistics, Technical Report 649. September, 2003.

Michael I. Jordan: *An Introduction to Probabilistic Graphical Models*, Unpublished.