A. Proof of Proposition 1

The usual way of writing the CGM distribution is to replace $f(\mathbf{n};\boldsymbol{\theta})$ in Eq. (3) by

$$f'(\mathbf{n};\boldsymbol{\theta}) = \frac{\prod_{C\in\mathcal{C},i_C\in\mathcal{X}^{|C|}} \mu_C(i_C)^{\mathbf{n}_C(i_C)}}{\prod_{S\in\mathcal{S},i_S\in\mathcal{X}^{|S|}} \left(\mu_S(i_S)^{\mathbf{n}_S(i_S)}\right)^{\nu(S)}} \tag{21}$$

We will show that $f(\mathbf{n};\boldsymbol{\theta}) = f'(\mathbf{n};\boldsymbol{\theta})$ for any $\mathbf{n}$ such that $h(\mathbf{n}) > 0$ by showing that both descibe the probability of an ordered sample with sufficient statistics $\mathbf{n}$. Indeed, suppose there exists some ordered sample $\mathbf{X} = (\mathbf{x}^1,\ldots,\mathbf{x}^N)$ with sufficient statistics $\mathbf{n}$. Then it is clear from inspection of Eq. (3) and Eq. (21) that $f(\mathbf{n};\boldsymbol{\theta}) = \prod_{m=1}^N p(\mathbf{x}^m;\boldsymbol{\theta}) = f'(\mathbf{n};\boldsymbol{\theta})$ by the junction tree reparameterization of $p(\mathbf{x};\boldsymbol{\theta})$ (Wainwright & Jordan, 2008). It only remains to show that such an $\mathbf{X}$ exists whenever $h(\mathbf{n}) > 0$. This is exactly what was shown by Sheldon & Dietterich (2011): for junction trees, the hard constraints of Eq. (4), which enforce local consistency on the integer count variables, are equivalent to the global consistency property that there exists some ordered sample $\mathbf{X}$ with sufficient statistics equal to $\mathbf{n}$. (Since these are integer count variables, the proof is quite different from the similar theorem that local consistency implies global consistency for marginal distributions.) We briefly note two interesting corollaries to this argument. First, by the same reasoning, *any* reparameterization of $p(\mathbf{x};\boldsymbol{\theta})$ that factors in the same way can be used to replace $f(\mathbf{n};\boldsymbol{\theta})$ in the CGM distribution. Second, we can see that the base measure $h(\mathbf{n})$ is exactly the *number of different ordered samples* with sufficient statistics equal to $\mathbf{n}$.

B. Proof of Theorem 1: Additional Details

Suppose $\{\mathbf{n}^N\}$ is a sequence of random vectors that converge in distribution to $\mathbf{n}$, and that $\mathbf{n}_A^N$, $\mathbf{n}_B^N$, and $\mathbf{n}_S^N$ are subvectors that satisfy

$$\mathbf{n}_A^N \perp\!\!\!\perp \mathbf{n}_B^N \mid \mathbf{n}_S^N \tag{22}$$

for all N. Let α, β, and γ be measurable sets in the appropriate spaces and define

$$\begin{aligned} z = {} & \Pr(\mathbf{n}_A \in \alpha, \mathbf{n}_B \in \beta \mid \mathbf{n}_S \in \gamma) - \\ & \Pr(\mathbf{n}_A \in \alpha \mid \mathbf{n}_S \in \gamma)\Pr(\mathbf{n}_B \in \beta \mid \mathbf{n}_S \in \gamma) \end{aligned} \tag{23}$$

Also let z^N be the same expression but with all instances of $\mathbf{n}$ replaced by $\mathbf{n}^N$ and note that $z^N = 0$ for all N by the assumed conditional independence property of Eq. (22). Because the sequence $\{\mathbf{n}^N\}$ converges in distribution to $\mathbf{n}$, we have convergence of each term in z^N to the corresponding term in z, which means that

$$z = \lim_{N\to\infty} z^N = \lim_{N\to\infty} 0 = 0,$$

so the conditional independence property of Eq. (22) also holds in the limit.

C. Proof of Theorem 3: Linear Function from $\tilde{\mathbf{I}}$ to $\mathbf{I}$

We need to show $\mathbf{I}_A$ can be recovered from $\tilde{\mathbf{I}}_{A+}$ with a linear function.

Suppose the last indicator variable in $\mathbf{I}_A$ is i_A^0, which corresponds to the setting that all nodes in A take value L. Let $\mathbf{I}'_A$ be a set of indicators which contains all entries in $\mathbf{I}_A$ but the last one i_A^0. Then $\mathbf{I}_A$ can be recovered from $\mathbf{I}'_A$ by the constraint $\sum_{i_A} \mathbf{I}_A(i_A) = 1$.

Now we only need to show that $\mathbf{I}'_A$ can be recovered from $\mathbf{I}_{A+}$ linearly. We claim that there exists an invertible matrix $\mathbb{H}$ such that $\mathbb{H}\,\mathbf{I}'_A = \tilde{\mathbf{I}}_{A+}$.

Showing the existence of $\mathbb{H}$. Let $\tilde{\mathbf{I}}_{A+}(i_D)$ be the i_D entry of $\tilde{\mathbf{I}}_{A+}$, which is for configuration i_D of clique $D, D \subseteq A$.

$$\tilde{\mathbf{I}}_{A+}(i_D) \quad = \quad \sum_{i_{A\setminus D}} \mathbf{I}'_A(i_D, i_{A\setminus D}) \tag{24}$$

Since no nodes in D take value L by definition of $\tilde{\mathbf{I}}_D$, $(i_D, i_{A\setminus D})$ *cannot* be the missing entry i_A^0 of $\mathbf{I}'_A$, and the equation is always valid.

Showing that $\mathbb{H}$ is square. For each D, there are $(L-1)^{|D|}$ entries, and A has $\binom{|A|}{|D|}$ sub-cliques with size $|D|$. So $\tilde{\mathbf{I}}_{A+}$ have overall $L^{|A|} - 1$ entries, which is the same as $\mathbf{I}'_A$. So $\mathbb{H}$ is a square matrix.

We view $\mathbf{I}'_A$ and $\tilde{\mathbf{I}}_{A+}$ as matrices and each row is a indicator function of graph configurations. Since no trivial linear combination of $\tilde{\mathbf{I}}_{A+}$ is a constant by the conclusion in Loh and Wainwright (2013), $\tilde{\mathbf{I}}_{A+}$ has linearly independent columns. Therefore, $\mathbb{H}$ must have full rank and $\mathbf{I}'_A$ must have linearly independent columns.