APPENDIX: SUPPLEMENTARY MATERIAL
Bethe and Related Pairwise Entropy Approximations

In this Appendix, we provide:

- Proofs of Theorems 6 and 7, and Lemma 8 from the main text.
- Background on the loop series method (Chertkov and Chernyak, 2006; Sudderth et al., 2007).

Second Derivatives of $\mathcal{F}_A$

Theorem 6. ($H_{ij} = \frac{\partial^2 \mathcal{F}_A}{\partial q_i \partial q_j}$ second derivatives of $\mathcal{F}_A(q_1, \ldots, q_n)$, assuming optimum pairwise marginals ξ_{ij})

$$H_{ij} = \begin{cases} \frac{q_i q_j - \xi_{ij}}{\rho_{ij} T_{ij}} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{if } (i,j) \notin \mathcal{E} \end{cases}, \quad H_{ii} = \frac{c_i}{q_i(1-q_i)} + \sum_{j \in \mathcal{N}(i)} \left(\frac{q_j(1-q_j)}{\rho_{ij} T_{ij}} - \frac{\rho_{ij}}{q_i(1-q_i)} \right),$$

where ξ_{ij} takes its optimum value from Theorem 2, and $T_{ij} = q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \geq 0$, with equality iff q_i or $q_j \in \{0, 1\}$.

Proof. The proof of this result for arbitrary counting numbers extends the earlier approaches of Weller and Jebara (2013) and Korč et al. (2012), which examined only the restricted case of the Bethe approximation. Consider the equation for the free energy approximation $\mathcal{F}_A$ (7). Note that we shall always assume optimum pairwise marginal ξ_{ij} terms to be given implicitly by Theorem 2. We first consider pairwise terms of $\mathcal{F}_A$, then singleton terms, which will be added together to give the result. $T_{ij} > 0$ unless q_i or $q_j \in \{0, 1\}$ follows from (Weller and Jebara, 2013, Lemma 12).

Pairwise terms. Consider an edge $(i,j) \in \mathcal{E}$ and collect its pairwise terms together from $\mathcal{F}_A$ (7), defining

$$f(q_i, q_j) = -W_{ij} \xi_{ij}(q_i, q_j) - \rho_{ij} S_{ij}(q_i, q_j). \tag{12}$$

Let $\mathbf{y} = (y_1, y_2, y_3)$ be one of four possible vectors with components $y_1 = a$, $y_2 = b$ and $y_3 = 1$, where $a, b \in \mathbb{B} = \{0, 1\}$. Note that a third 'dimension' restricted to the value 1 has been added for notational convenience. Let $\pi(\mathbf{y}) = \mu_{ij}(a, b)$, that is the (a, b) element from the μ_{ij} matrix (2), given the values of q_i and q_j. Let $\phi(\mathbf{y}) = W_{ij}$ if $\mathbf{y} = (1, 1, 1)$, or $\phi(\mathbf{y}) = 0$ otherwise. Let $\mathbf{r} = (q_i, q_j, 1)$. Define function h used in entropy calculations as $h(z) = -z \log z$.

Consider (12) and instead of solving for ξ_{ij} (or equivalently for π) explicitly, express f as an optimization problem, minimizing the approximate free energy subject to local consistency and normalization constraints in order to use techniques from convex optimization. We have $f(q_i, q_j) = g(\mathbf{r})$ where

$$\begin{aligned} g(\mathbf{r}) = & \min_{\pi} \sum_{\mathbf{y}} \left[-\phi(\mathbf{y}) \pi(\mathbf{y}) - \rho_{ij} h(\pi(\mathbf{y})) \right] \\ & \text{s.t.} \sum_{\mathbf{y}: y_k = 1} \pi(\mathbf{y}) = r_k \quad k = 1, 2, 3. \end{aligned} \tag{13}$$

Introducing dual variables $\boldsymbol{\lambda}$, the Lagrangian can be written as

$$L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = \sum_{\mathbf{y}} [(-\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda} \rangle) \pi(\mathbf{y}) - \rho_{ij} h(\pi(\mathbf{y}))] + \langle \mathbf{r}, \boldsymbol{\lambda} \rangle,$$

with derivative

$$\frac{\partial L_{\mathbf{r}}(\pi, \boldsymbol{\lambda})}{\partial \pi} = -\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda} \rangle + \rho_{ij}(1 + \log \pi),$$

which yields a minimum at

$$\pi^*_{\boldsymbol{\lambda}}(\mathbf{y}) = \exp\left(\frac{\phi(\mathbf{y}) + \langle \mathbf{y}, \boldsymbol{\lambda} \rangle}{\rho_{ij}} - 1 \right). \tag{14}$$

Since the minimization problem in (13) is convex and satisfies the weak Slater's condition (the constraints are affine), strong duality applies and $g(\mathbf{r}) = \max_{\boldsymbol{\lambda}} G(\mathbf{r}, \boldsymbol{\lambda}) = G(\mathbf{r}, \boldsymbol{\lambda}^*(\mathbf{r}))$ where the dual is

$$G(\mathbf{r}, \boldsymbol{\lambda}) = \min_{\pi} L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = -\rho_{ij} \sum_{\mathbf{y}} \pi^*_{\boldsymbol{\lambda}}(\mathbf{y}) + \langle \mathbf{r}, \boldsymbol{\lambda} \rangle. \tag{15}$$

Hence, $\frac{\partial g}{\partial r_k} = \frac{\partial G}{\partial r_k}\Big|_{\lambda^*} = \lambda^*_k$. Our aim is to obtain second derivatives of f via $\frac{\partial^2 g}{\partial r_l \partial r_k} = \frac{\partial \lambda^*_k}{\partial r_l}$, which we shall derive in terms of a 3×3 matrix C, where we define

$$C_{kl} := \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} = \frac{\partial D_k}{\partial \lambda_l}, \quad k, l = 1, 2, 3$$

with

$$D_k(\mathbf{r}, \boldsymbol{\lambda}) := \frac{\partial G(\mathbf{r}, \boldsymbol{\lambda})}{\partial \lambda_k} = -\sum_{\mathbf{y}} y_k \pi^*_{\boldsymbol{\lambda}}(\mathbf{y}) + r_k, \quad \text{using (15).} \tag{16}$$

Now $D_k(\mathbf{r}, \boldsymbol{\lambda}^*) = 0$ for $k = 1, 2, 3$. Differentiating this with respect to r_l,

$$\begin{aligned} 0 = \frac{dD_k(\mathbf{r}, \boldsymbol{\lambda}^*)}{dr_l} = & \quad \frac{\partial D_k}{\partial r_l} + \sum_{p=1}^{3} \frac{\partial D_k}{\partial \lambda_p} \frac{\partial \lambda^*_p}{\partial r_l}, & k, l = 1, 2, 3 \\ = & \quad \delta_{kl} + \sum_{p} C_{kp} \frac{\partial^2 g}{\partial r_l \partial r_p}, & \text{using (16) and definition of } C. \end{aligned}$$

Hence, $\frac{\partial^2 g}{\partial r_l \partial r_k} = -[C^{-1}]_{kl}$. Using its definition and (16), we have

$$\begin{aligned} C_{kl} = \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} &= \frac{\partial}{\partial \lambda_l} \left(-\sum_{\mathbf{y}} y_k \pi^*_{\boldsymbol{\lambda}}(\mathbf{y}) + r_k \right) \\ &= -\frac{1}{\rho_{ij}} \sum_{\mathbf{y}} y_k y_l \pi^*_{\boldsymbol{\lambda}}(\mathbf{y}) = -\frac{1}{\rho_{ij}} \sum_{\mathbf{y}: y_k = y_l = 1} \pi^*_{\boldsymbol{\lambda}}(\mathbf{y}). \end{aligned}$$

Thus, using shorthand $\mu_{ab} = \mu_{ij}(a, b)$,

$$C = -\frac{1}{\rho_{ij}} \begin{pmatrix} \mu_{10} + \mu_{11} & \mu_{11} & \mu_{10} + \mu_{11} \\ \mu_{11} & \mu_{01} + \mu_{11} & \mu_{01} + \mu_{11} \\ \mu_{10} + \mu_{11} & \mu_{01} + \mu_{11} & 1 \end{pmatrix}. \tag{17}$$

Recall constraints $\mu_{00} + \mu_{01} + \mu_{10} + \mu_{11} = 1$, $\mu_{01} + \mu_{11} = q_j$, $\mu_{10} + \mu_{11} = q_i$.

Applying the result above and Cramer's rule,

$$\begin{aligned} \frac{\partial^2 f}{\partial q_i^2} = \frac{\partial^2 g}{\partial r_1^2} &= -\frac{1}{\rho_{ij}^2 \det C} (\mu_{01} + \mu_{11})(\mu_{00} + \mu_{10}) = \frac{q_j(1 - q_j)}{-\rho_{ij}^2 \det C} \\ \frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} &= \frac{\partial^2 g}{\partial r_1 \partial r_2} = \frac{(\mu_{01}\mu_{10} - \mu_{00}\mu_{11})}{-\rho_{ij}^2 \det C} \\ \frac{\partial^2 f}{\partial q_j^2} = \frac{\partial^2 g}{\partial r_2^2} &= -\frac{1}{\rho_{ij}^2 \det C} (\mu_{10} + \mu_{11})(\mu_{00} + \mu_{01}) = \frac{q_i(1 - q_i)}{-\rho_{ij}^2 \det C}. \end{aligned}$$

From (17), after simplifying, $-\rho_{ij} \det C = \mu_{00}\mu_{10}\mu_{11} + \mu_{10}\mu_{11}\mu_{01} + \mu_{11}\mu_{10}\mu_{00} + \mu_{01}\mu_{00}\mu_{10} \geq 0$ (all products of three terms of the pairwise pseudomarginal matrix (2)). Substituting in terms from (2) and simplifying establishes $-\rho_{ij} \det C = T_{ij}$ from the statement of the theorem, and $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_i q_j - \xi_{ij}$.

Hence,

$$\frac{\partial^2 f}{\partial q_i^2} = \frac{q_j(1 - q_j)}{\rho_{ij} T_{ij}}, \quad \frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{q_i q_j - \xi_{ij}}{\rho_{ij} T_{ij}}, \quad \frac{\partial^2 f}{\partial q_j^2} = \frac{q_i(1 - q_i)}{\rho_{ij} T_{ij}}. \tag{18}$$

Singleton terms. Let $f_i(q_i)$ be the singleton terms from (7) for X_i. The only non-zero derivatives are with respect to q_i.

$$f_i(q_i) = -\theta_i q_i + S_i(q_i)\Big(- c_i + \sum_{j \in \mathcal{N}(i)} \rho_{ij}\Big),$$
$$\frac{\partial f_i}{\partial q_i} = -\theta_i - [\log q_i - \log(1-q_i)]\Big(- c_i + \sum_{j \in \mathcal{N}(i)} \rho_{ij}\Big),$$
$$\frac{\partial^2 f_i}{\partial q_i^2} = \frac{c_i - \sum_{j \in \mathcal{N}(i)} \rho_{ij}}{q_i(1-q_i)}.$$

Adding pairwise and singleton terms gives the result. □

Submodularity of $\mathcal{F}_A$

Here we consider $\mathcal{F}_A(q_1, \ldots, q_n)$ with pairwise marginals given by Theorem 2, and show that for any discrete mesh $\mathcal{M} = \prod_{i=1}^n M_i$, where M_i is a finite set of points for q_i in $[0,1]$, and for any counting numbers (provided all $\rho_{ij} \neq 0$), then the discrete optimization to find the point in $\mathcal{M}$ with lowest $\mathcal{F}_A$ is submodular for any attractive model (hence can be solved efficiently). We follow the same reasoning used by Weller and Jebara (2013) for the Bethe approximation.

Regarding the expression for H_{ij} from Theorem 6 together with Lemma 3, observe that provided $\rho_{ij} \neq 0$ and $q_i, q_j \in (0,1)$, $W_{ij} \geq 0 \Leftrightarrow \frac{\partial^2 \mathcal{F}_A}{\partial q_i \partial q_j} \leq 0$ (whatever the sign of ρ_{ij}).

We first show that third derivatives of $\mathcal{F}_A$ exist and are finite. Recall that by definition, $\alpha_{ij} = \exp(W_{ij}/\rho_{ij}) - 1 > -1$, with the same sign as W_{ij}/ρ_{ij}.

Lemma 9 (Finite 3rd derivatives)**.** *If $q_i, q_j \in (0,1)$ and $\rho_{ij} \neq 0 \; \forall (i,j) \in \mathcal{E}$, then all third derivatives exist and are finite.*

Proof. Using Theorem 6 and noting $T_{ij} > 0$ strictly given our conditions, it is sufficient to show that any $\frac{\partial \xi_{ij}}{\partial q_k}$ is finite. We may assume $k \in \{i, j\}$ else the derivative is 0 and by symmetry need only check $\frac{\partial \xi_{ij}}{\partial q_i}$. Differentiating (8),

$$\frac{\partial \xi_{ij}}{\partial q_i} = \frac{\alpha_{ij}(q_j - \xi_{ij}) + q_j}{1 + \alpha_{ij}(q_i - \xi_{ij} + q_j - \xi_{ij})}.$$

Recalling (2), $q_i - \xi_{ij}$ and $q_j - \xi_{ij}$ are elements of the edge pseudomarginal and hence are nonnegative. For $\alpha_{ij} > 0$, it is clear that the denominator is positive. If $\alpha_{ij} < 0$ then note that $\alpha_{ij} \in (-1, 0)$, hence it is sufficient to show that $(q_i - \xi_{ij} + q_j - \xi_{ij}) \leq 1$. This follows immediately from other constraints ensuring that elements of the pseudomarginal are valid, i.e. $\xi_{ij} \geq 0$ and $1 + \xi_{ij} - q_i - q_j \geq 0$. □

Next we show a stronger version of Lemma 3. This will simplify the subsequent proof of Theorem 7.

Lemma 10 (Better lower bound for ξ_{ij}, Lemma 14 in Weller and Jebara, 2013)**.** *If $\alpha_{ij} > 0$, then $\xi_{ij} \geq q_i q_j + \alpha_{ij} q_i q_j (1 - q_i)(1 - q_j)/[1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)]$, equality only possible at an edge, i.e. one or both of $q_i, q_j \in \{0, 1\}$.*

Proof. Write $\xi_{ij} = q_i q_j + y$ and substitute into (8) to give

$$\alpha_{ij} y^2 - y[1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)] + \alpha_{ij} q_i q_j (1 - q_i)(1 - q_j) = 0.$$

This is a convex parabola which at $y = 0$ is above the abscissa (unless q_i or $q_j \in \{0, 1\}$), with negative gradient.[7] Hence, all roots are at $y \geq 0$, and given convexity we can bound below using the tangent at $y = 0$, which yields the result. □

Now we prove the main result of this Section.

Theorem 7 For any counting numbers with $\rho_{ij} \neq 0 \; \forall (i,j) \in \mathcal{E}$, and any discretization, an attractive model yields a submodular discrete optimization problem to estimate $\log Z_A$.

[7] Observe that $q_i + q_j - 2q_i q_j = \frac{1}{2} - 2(q_i - \frac{1}{2})(q_j - \frac{1}{2})$, hence $\in (0,1)$ for $q_i, q_j \in (0,1)$.

Proof. For any edge (i,j), let f be the pairwise terms from $\mathcal{F}_A$ given in (12), and note the submodularity requirement from §2.3. Let $x=(x_1,x_2)$, $y=(y_1,y_2)$ be any points in $[0,1]^2$. Define $s(x,y)=(s_1,s_2)=(\min(x_1,y_1),\min(x_2,y_2))$, and $t(x,y)=(t_1,t_2)=(\max(x_1,y_1),\max(x_2,y_2))$. Let $g(x,y)=f(s_1,s_2)+f(t_1,t_2)-f(s_1,t_2)-f(s_2,t_1)$, and call this the submodularity of the rectangle defined by x,y. We must show $g(x,y)\leq 0$. Note f is continuous in $[0,1]^2$, hence so also is g. We shall show that $\forall (x,y)\in(0,1)^2,\ g(x,y)<0$ then the result follows by continuity.

Assume $x,y\in(0,1)^2$. Consider derivatives of f in the compact set $R=[s_1,t_1]\times[s_2,t_2]$. Using (9) and bounded pseudomarginal entries (see Weller and Jebara, 2013 for details), first derivatives exist and are bounded. By Theorem 6 and Lemma 9, the same holds for second and third derivatives. Further, Theorem 6 and Lemma 10 show that $\frac{\partial^2 f}{\partial q_i \partial q_j}=\frac{\partial^2 f}{\partial q_j \partial q_i}<0$.

If a rectangle is sliced fully along each dimension so as to be subdivided into sub-rectangles then summing the submodularities of all the sub-rectangles, internal terms cancel and we obtain the submodularity of the original rectangle.

Hence there exists an ϵ such that if we subdivide the rectangle defined by x,y into sufficiently small sub-rectangles with sides $<\epsilon$ and apply Taylor's theorem up to second order with the remainder expressed in terms of the third derivative evaluated in the interval, then the second order terms dominate and the submodularity of each small sub-rectangle <0. Summing over all sub-rectangles yields the result. □

Effect of Approximate Entropy on Marginals

Lemma 8. For a symmetric homogeneous d-regular model on n vertices, let H be the Hessian of the approximate free energy at $q_i=\frac{1}{2}\ \forall i\in\mathcal{V}$, using uniform counting numbers $c_i=c\ \forall i\in\mathcal{V}, \rho_{ij}=\rho\ \forall (i,j)\in\mathcal{E}$, then $\mathbf{1}^T H\mathbf{1}=n\left[4(c-d\rho)+\frac{d}{\rho\xi}\right]$, where $\xi=\frac{1}{2}\sigma\left(\frac{W}{2\rho}\right)$ is the uniform optimum edge marginal term, and $\sigma(u)=\frac{1}{1+e^{-u}}$ is the standard sigmoid function.

Proof. Using (9), it is straightforward to show that there is a stationary point at $q_i=\frac{1}{2}\ \forall i$. By Theorem 2, all optimum pairwise marginal terms are $\xi_{ij}=\xi=\frac{1}{2}\sigma\left(\frac{W}{2\rho_{ij}}\right)$, where $\sigma(u)=\frac{1}{1+e^{-u}}$ is the standard sigmoid function. Now using Theorem 6, all $T_{ij}=T=\frac{1}{16}-\left(\xi-\frac{1}{4}\right)^2=\xi\left(\frac{1}{2}-\xi\right)$, and

$$\begin{aligned}\mathbf{1}^T H\mathbf{1} &= n\left[4c+d\left(\frac{1}{4\rho T}-4\rho\right)+\frac{d}{\rho T}\left(\frac{1}{4}-\xi\right)\right]\\ &= n\left[4(c-d\rho)+\frac{d}{\rho T}\left(\frac{1}{2}-\xi\right)\right]\\ &= n\left[4(c-d\rho)+\frac{d}{\rho\xi}\right]\end{aligned}$$

□

Background on the Loop Series Method

The loop series expansion of Chertkov and Chernyak (2006) provides an expression for the ratio of the true partition function Z to the Bethe approximation Z_B. Here we provide brief background, following the presentation in Sudderth et al. (2007).

At any stationary point $\hat{\mu}$ of the Bethe free energy $\mathcal{F}_B$, specified by our usual singleton $\{q_i : i\in\mathcal{V}\}$ and edge $\{\xi_{ij} : (i,j)\in\mathcal{E}\}$ marginal terms,

$$\frac{Z}{Z_B(\hat{\mu})}=1+\sum_{\emptyset\neq F\subseteq\mathcal{E}}\beta_F\prod_{i\in\mathcal{V}}\mathbb{E}_{q_i}\left[(X_i-q_i)^{d_i(F)}\right], \tag{19}$$

where $\beta_F=\prod_{(i,j)\in F}\beta_{ij}$, $\quad\beta_{ij}=\dfrac{\xi_{ij}-q_iq_j}{q_i(1-q_i)q_j(1-q_j)}$, $\quad$ and $d_i(F)$ is the degree of i in the subgraph induced by F.

We write $Z_B(\hat{\mu})$ to mean $\exp\left[-\mathcal{F}_B(\hat{\mu})\right]$. Note that $Z_B=\max_{\hat{\mu}} Z_B(\hat{\mu})$ and that both $Z, Z_B\geq 0$.

Observe that (19) is a sum over (the potentially large set of) all non-empty edge subsets. However, for any subset F such that $d_i(F) = 1$ for any $i \in \mathcal{V}$, then $\mathbb{E}_{q_i}\left[(X_i - q_i)^{d_i(F)}\right] = 0$, hence the term for this subset is zero and all such subsets may be ignored. This leaves all subsets F such that $d_i(F) \neq 1\ \forall i \in \mathcal{V}$. These remaining subsets are called *generalized loops*. Examples include a single cycle, two disjoint cycles, or two cycles connected by a path between them.

A related concept is the *core* of a graph, which is defined as the (unique) graph which remains after repeatedly removing any nodes with degree 1. It is easy to see that no generalized loop can exist outside the core.

Regarding (19), Sudderth et al. (2007) sought sufficient conditions such that all terms in the sum were nonnegative, in which case clearly $Z_B \leq Z$. One case is if (i) all $\beta_F \geq 0$, and (ii) all $\mathbb{E}_{q_i}\left[(X_i - q_i)^{d_i(F)}\right] \geq 0$. The first condition holds for an attractive model since by Lemma 3, each β_{ij} takes the sign of W_{ij} (all $\rho_{ij} = 1$ for the Bethe approximation). The second condition clearly holds for any i with $d_i(F)$ even (since then we have the expectation of a non-negative quantity), or $d_i(F) = 1$ (in which case it is 0 as noted above). Hence, we must worry only about generalized loops containing variables with odd degree > 1.

Using a standard result for moments of Bernoulli random variables,

$$\mathbb{E}_{q_i}\left[(X_i - q_i)^d\right] = q_i(1 - q_i)\left[(1 - q_i)^{d-1} + (-1)^d q_i^{d-1}\right].$$

For d odd, this is nonnegative provided $(1 - q_i) \geq q_i \Leftrightarrow q_i \leq \frac{1}{2}$. Hence, if this is true for all variables in the core with degree ≥ 3, then this is sufficient to show that $Z_B \leq Z$. Using a slight variant of the same argument, Sudderth et al. (2007) show that it is also sufficient if instead all such variables have $q_i \geq \frac{1}{2}$.

Our new observations. For our first result in §6.3, we apply the same analysis and observe that if a model contains exactly one cycle with edge set C and it is frustrated, then there is only one generalized loop $F = C$: this has $\beta_F \leq 0$ and all $d_i(F) = 2$, hence by (19), $Z/Z_B(\hat{\mu}) \leq 1\ \forall \hat{\mu}$, and thus in particular, $Z_B \geq Z$.[8]

Similarly, we can conclude more generally that $Z_B \geq Z$ for any model such that every generalized loop contains an odd number of repulsive edges (this is a sort of generalized frustrated cycle), and the Bethe optimum marginals for every variable that has an odd degree ≥ 3 in any generalized loop, are either all $\leq \frac{1}{2}$ or all $\geq \frac{1}{2}$.

[8] In fact, for models with exactly one cycle, it is known that the Bethe free energy is convex (Pakzad and Anantharam, 2002), hence there is only one stationary point.