Network Ranking With Bethe Pseudomarginals

Kui Tang
Columbia University
kt2384@columbia.edu

Adrian Weller
Columbia University
aw2506@columbia.edu

Tony Jebara
Columbia University
tj2008@columbia.edu

Abstract

Network structure often contains information that can be useful for ranking algorithms. We incorporate network structure by formulating ranking as marginal inference in a Markov random field (MRF). Though inference is generally NP-hard, we apply a recently-developed polynomial-time approximation scheme (PTAS) to infer Bethe pseudomarginals. As a case study, we investigate the problem of ranking failing transformers that are physically connected in a network. Compared to independent score-based ranking, the current state of the art, we show superior ranking results. We conclude by discussing an empirical phenomenon of critical parameter regions, with implications for new algorithms.

1 Introduction

Many important data-sets involve networks: people belong to social networks, webpages are joined in a link graph, and power utilities are connected in a grid. Often, these networks describe how influence propagates within the population. For instance, edges may imply conditional dependence and independence of certain nodes which encourages us to model such data-sets as Markov random fields. The goal of this article is to rank nodes in such a network, or Markov random field. Rather than using only node-attributes and ranking each node in isolation, as some state-of-the-art approaches [1, 2], we wish to infer marginal probabilities within the network.

Marginal inference on graphical models with a large tree-width is problem that is NP-hard [3] and even NP-hard to approximate [4]. However, a method was recently developed to efficiently infer Bethe pseudomarginals on a large dense network of binary variables [5]. Previous work developed a polynomial-time approximation scheme using Bethe free energy that applies whenever pairwise interactions are submodular (equivalently associative or ferromagnetic). In this article, we further develop this approach and explore its application to infer which transformers in a power network are most likely to fail. This problem was initially studied by [6], which estimated the probability for each transformer to fail on its own. However, transformers are physically connected to one another. Therefore, the failure of one increases load on its neighbors, causing them in turn to be more likely to fail. Inferring posterior failure probabilities through the network allows us to then rank the transformers to determine which ones are most at risk to then give the list of most high priority transformers that a utility company needs to repair first.

We have a network topology and independent scores for each node, which are outputs of a RankBoost model run at the utility company. We use the PTAS developed in [5] to search for optimal MRF parameters and to predict rankings more accurately than the independent scores. While we developed and tested algorithms on real data, in this paper, we present experiments on *simulated* data designed to mimic the statistics of the real data. Our simulations are on a smaller scale, to facilitate comparisons with exact marginals.

Section 2 describes ranking model; section 3 describes the details of the PTAS; section 4 highlight empirical results, section 5 reviews related work and section 6 concludes.

2 Model

To each item in the dataset we associate a binary variable indicating its relevance. The topology of the MRF with vertices $\mathcal{V}$ and edges $\mathcal{E}$ is the same that of the underlying data points. First consider the overcomplete parameterization:

$$\begin{aligned} p(x|\theta) &= \frac{1}{Z}\exp\{-E(x|\theta)\} \\ E(x|\theta) &= -\sum_{i\in\mathcal{V}}\theta_i(x_i) - \sum_{(i,j)\in\mathcal{E}}\theta_{ij}(x_i,x_j) \quad x_i,x_j\in\{0,1\} \end{aligned} \tag{1}$$

We derive θ_i and θ_{ij} from prespecified scores and model assumptions of pairwise interactions. We have scores s_i for each node, so we set $\theta_i(0) = 0, \theta_i(1) = s_i$. Next, we assume that edge potentials are homogeneous. In our problem, edges are physical wires, and we assume wires throughout the city have similar characteristics. Moreover, if x_i and x_j are neighbors, the events $x_i = 1, x_j = 0$ and $x_i = 0, x_j = 1$ are symmetric: the wire does not care which end fails first. Thus, we write

$$\theta_{ij}(x_i,x_j) = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$$

for each edge potential. Submodularity requires $\alpha + \gamma > 2\beta$. We can further simplify (1) to

$$\begin{aligned} E(x|\theta) &= \alpha - \sum_{i\in\mathcal{V}}(s_i + b)x_i - \sum_{(i,j)\in\mathcal{E}} w x_i x_j \\ b &= \beta - \alpha \\ w &= \alpha + \gamma - 2\beta = \gamma - \beta - b \end{aligned} \tag{2}$$

So α is a constant independent of x. Therefore, to derive probabilities from homogeneous potentials of the form (2), it suffices to specify a bias b and an edge weight w. The higher the value of w, the more highly correlated each pair along an edge will be. For a fixed w, the parameter b trades off weight between α, favoring both nodes to function, and γ, favoring both nodes to fail.

3 Discrete Bethe Pseudomarginals

Algorithm 1, which also describes the method in [5] outlines the PTAS to find the global minimum of the Bethe free energy up to additive error ϵ. We use the parameterization in [11] which requires only singleton (not pairwise) marginals:

$$F_B(q) = \sum_{i\in\mathcal{V}} -\theta_i q_i + (\delta_i - 1)S_1(q_i) + \sum_{(i,j)\in\mathcal{E}} -(W_{ij}\xi_{ij} + S_2(q_i,q_j)) \tag{3}$$

Algorithm 1 Discrete optimization of the Bethe free energy

```
Compute bounds $\{A_i, B_i\}_{i=1}^{N}$ away from 0, 1 of Bethe pseudomarginals using Eqs. (23) to (25) in [7].
for $i := 1 : N$ do                                                          ▷ Place mesh points
    Compute mesh size $\gamma = f(n, W, \epsilon)$ using max-eigenvalue bound in [5].
    Put $K_i := \left\lceil \frac{1-B_i-A_i}{\gamma} \right\rceil$
    Compute mesh points $\{q_i(k) := A_i + k\gamma\}_{k=1}^{K_i}$
    Alternatively, compute adaptive mesh $\{q_i(k)\}_{k=1}^{K_i}$ using [8], with $K_i$ determined by the algorithm.
end for
                                  ▷ Construct multi-label discrete MRF; decompose terms of $F_B$
Put $\eta_i(k) := -\theta_i q_i(k) + (\delta_i - 1)S_1(q_i(k))$ for $i \in \mathcal{V}, k = 1, \ldots, K_i$
Put $\eta_{ij}(k, \ell) := -W_{ij}\xi_{ij}(q_i(k), q_j(\ell)) - S_2(q_i(k), q_j(\ell))$ for $(i, j) \in \mathcal{E}, k = 1, \ldots, K_i, \ell = 1, \ldots, K_j$
Reduce the MRF $\{\eta_i(\cdot), \eta_{ij}(\cdot)\}$ to a binary MRF using [9].
Solve the binary MRF with a graph cut using [10].
Decode the multi-label solution, and the corresponding mesh point $q_i(k)$ for each $i \in \mathcal{V}$ from the cut.
```

where δ_i is degree of node i, q_i is a singleton pseudomarginal, θ_i and W_{ij} are singleton and pairwise weights, S_1 and S_2 are singleton and pairwise entropy functions, and each ξ_{ij} is the unique root of a quadratic function of $W_i j$, q_i, and q_j. For details on this notation, see [11].

We first bound the minimizer of the Bethe free energy in each dimension, using an iteration derived in [7]. We then place a grid between these bounds such that the minimum of F_B on the grid points is no more than ϵ from the global minimum. Conceptually, there are exponentially many grid points. But we can find the grid point with lowest energy in polynomial time by solving a submodular multi-label min-cut (negated MAP) problem. To do this, we decompose the terms of (3) along the nodes and edges of G and evaluating the singleton entropies and pairwise entropies along edges, resulting in a multi-label MRF. If the original model is submodular, then this discrete MRF is as shown in [5]. We can solve the final problem as a graph cut [10].

In work under review [8], a more efficient adaptive mesh was derived. In all, with the mesh points, discretization, and final MAP inference, the worst case runtime of this algorithm is $O((W|\mathcal{V}||\mathcal{E}|/\epsilon)^3)$ where W is the maximum weight on an edge.

Due to runtime, we report pseudomarginals as the minimizer of F_B. A method with guaranteed error bounds is instead to compute $\mu(x_i = 1) = \exp\{-\min F_B(x_i = 1) + \min F_B\}$ where $-F_B(x_i = 1)$ is the Bethe free energy conditioned on $x_i = 1$. This requires computing $N + 1$ partition functions, but has the benefit of inheriting ϵ-close guarantees. Note that this is still not a bound from the *true* marginals, since the Bethe free energy can be far from the true (Gibbs) free energy.

4 Results

We designed a simulation that reproduced summary statistics of our real data. We drew independent node scores from a mixture of Gaussians and a scale free network (100 nodes, 200 edges) from the Barabsi-Albert model [12]. We parameterized the MRF (2) with $w = 4, 8, 12$ and $b = 0, -4, -8$ accordingly to compensate so that the overall failure probabilities remained constant. Since we have a generative model, we evaluated algorithms on *expected* AUC under the true distribution, instead of the more common one-sample AUC. We used a simple Monte Carlo approximation with 10^5 samples, which was sufficient for error to be negligible

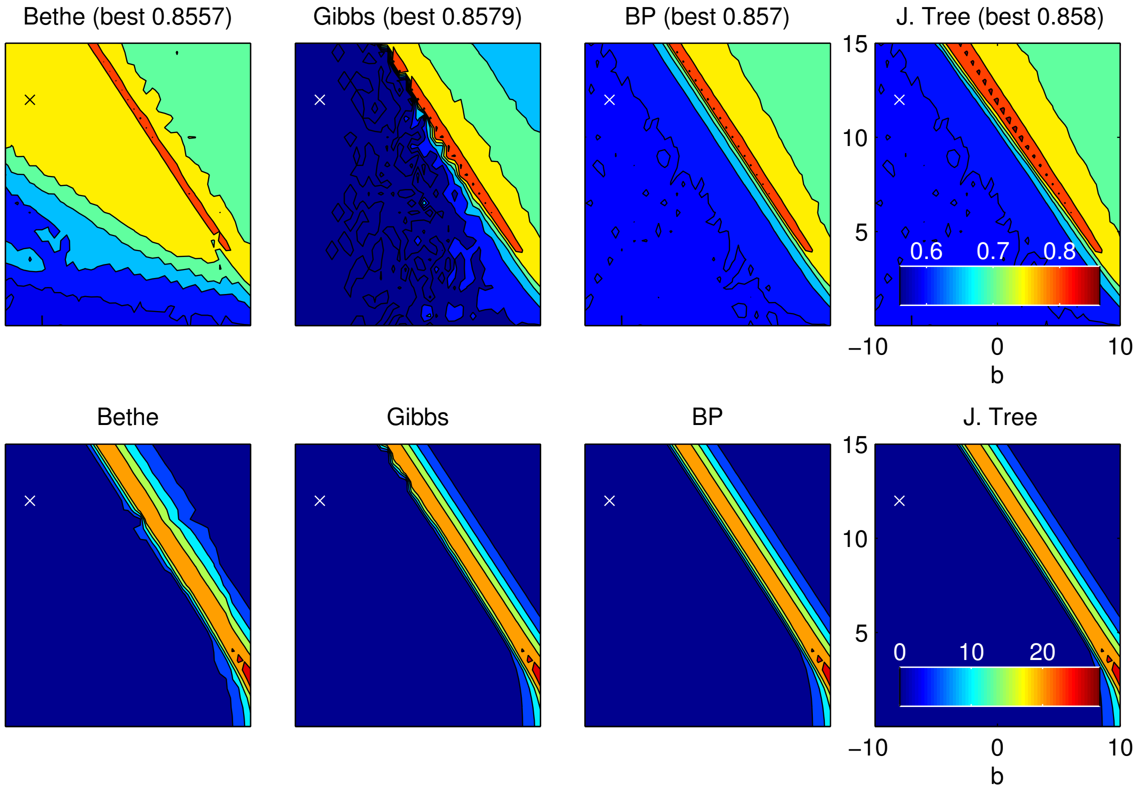

Figure 1: *Top:* Expected AUCs. *Bottom:* Entropies of singleton marginals. *Both:* Horizontal axis is b, vertical b. Redder means higher values. A phase transition characteristic of Ising models appears as a diagonal line. An × marks the parameters of the true distribution prior to rejection sampling. See text for details.

	$w = 4, b = 0$	$w = 8, b = -4$	$w = 12, b = -8$
Independent	0.5761	0.6352	0.5984
Bethe	0.8573, $b = -6, w = 10$	0.8683, $b = -3.5, w = 11.5$	0.8557, $b = 1, w = 11$
Gibbs	0.8575, $b = -6, w = 10$	0.8693, $b = 0.5, w = 7.5$	0.8579, $b = 0.5, w = 11.5$
J. Tree	0.8575, $b = -6, w = 10$	0.8693, $b = 0.5, w = 7.5$	0.8580, $b = 1, w = 11$
BP	0.8576, $b = -6, w = 10$	0.8693, $b = 0.5, w = 7.5$	0.8567, $b = 1, w = 11$

Table 1: Maximum expected AUC for an algorithm and a data distribution. Marginals outperform ranking by independent scores. The location of the maximum is generally stable across algorithms, though in one case our Bethe approximation found a different solution.

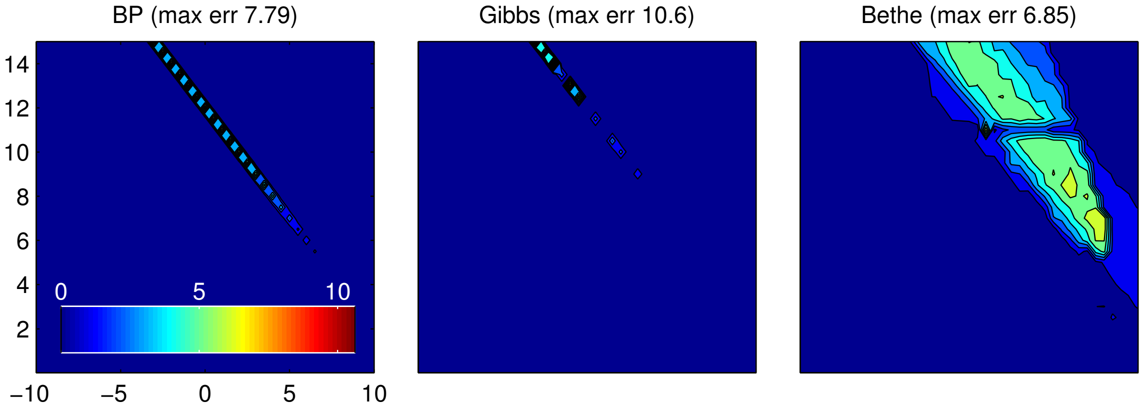

Figure 2: Errors compared to true marginals. Our method has lower maximum error.

for our purposes. To compute the expected AUC, we reject samples which do not contain failures. This step is required for a well-defined AUC and represents a realistic data model: the model is useful only if we predict at least one transformer to fail.

With approximate inference, maximum likelihood learning may no longer be consistent [13]. Moreover, the maximum likelihood model need not be the best ranker. We instead swept the two parameters w and b and plot the resulting expected AUC in the top row of Figure 1. At each point, we inferred marginals and plotted the expected AUC of those marginals. We show the trial for $w = 12, b = -8$, though the others were qualitatively similar. Table 1 shows values and locations of maximum expected AUC for every trial. An × marks the location of the true distribution.

The plots exhibit a mode and phase transition along a line $w + c_1 b = c_0$. The same phenomena occur in the other trials and in real data. For parameters far below this line, the marginals approach zero, making it difficult to discern a ranking. Far above the line, the marginals approach one, with similar problems. We observe this phenomenon quantitatively in the bottom row of Figure 1, which shows the entropy of the collection of estimated singleton marginals for each w and b. The regions with highest AUC are also those with highest entropy, supporting the intuition that the best ranking obtains with evenly-spread marginal probabilities—not necessarily the true ones.

The sum $w + c_1 b$ represents the (scaled) total positive bias in the model. In the parameterization (2), both b and w bias the marginals toward one: a high b increases the marginals on each node, and a high w increases the marginals for both nodes on a pair. For the Bethe plot (top-left of Figure 1), we fit a least-squares estimate on the fifty points with highest AUC, all of which are contained in the modal region. This gave $c_1 = 0.98$ and $c_0 = 12.23$ ($R^2 = 0.994, p = 5.2 \times 10^{-55}$). Thus, w and b trade off almost one-to-one in order to maintain a total positive bias of around 12, the modularity of the true model

To check approximation quality, we plotted ℓ_1 distances to true marginals in Figure 2. While our Bethe approximation errs more often, BP makes worse errors, and Gibbs makes the worst errors of all, even with 5×10^7 samples with an equally-long burn-in. The errors of the Bethe approximation do not seem to hurt the AUC.

5 Related Work

The work [6] and references therein were the first to apply ranking algorithms for transformer failures. Modeling network effects as MRFs was considered in [14] in a social marketing setting. Their work only studies local optimization algorithms and maximizes expected profit rather than the AUC. Work in [15] is motivated by MRFs, but their algorithm ultimately embeds the network in a vector space, which they argue to be useful for link prediction. Our work focuses on inference, which is more general than link prediction. [16] incorporates network structure by propagating independent scores through a similarity network, which resembles our problem. Although their algorithm converges, they give no other theoretical properties. [17] derive a random-walk model for ranking authors and documents, which combines social, citation, and authorship networks. However, their model requires many parameter choices, while ours requires just two.

6 Discussion

We have formulated the network ranking problem as marginal inference in a particular edge-homogeneous, attractive binary Markov random field. We demonstrated that such accounting of network structure improves the AUC compared to independent scores. We have applied a recently-developed PTAS for marginal inference to a real-world problem, with encouraging results.

The empirical results uncover some phenomenon and suggest directions for further work. A consistent feature is a critical line where both AUC and entropy are high. The statistical physics literature has studied *phase transitions* in the sense of varying temperature [18]; to our knowledge, a similar phase transition with respect to a varying external field (our b parameter) has not been studied. Future work could characterize the location of the critical line, which may generally offer better ranking performance.

Acknowledgments

We thank Nicholas Ruozzi for helpful and insightful discussions. This material is based upon work supported by the National Science Foundation under Grant No. 1117631. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, December 2003.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Edmonton, Alberta, Canada, 2002, pages 133–142.

[3] G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, 1990.

[4] P. Dagum and M. Luby. Approximate probabilistic reasoning in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

[5] A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *Sixteenth International Conference on Artificial Intelligence and Statistics*, 2013.

[6] C. Rudin, D. Waltz, R. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, S. Ierome, D. Isaac, A. Kressner, R. Passonneau, A. Radeva, and L. Wu. Machine learning for the new york city power grid. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):328–345, 2012.

[7] J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, December 2007.

[8] A. Anonymous. Approximating the Bethe partition function. In *Submitted to AISTATs 2014*, 2014.

[9] D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report (TUD-FI06-01). Dresden University of Technology, 2006.

[10] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1124–1137, September 2004.

[11] M. Welling and Y. W. Teh. Belief optimization for binary networks: a stable alternative to loopy belief propagation. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., Seattle, Washington, 2001, pages 554–561.

[12] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1–2):173–187, October 1, 1999.

[13] U. Heinemann and A. Globerson. What cannot be learned with Bethe approximations. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*. AUAI Press, Corvallis, Oregon, 2011, pages 319–326.

[14] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, 2001, pages 57–66.

[15] J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explor. Newsl.*, 7(2):23–30, December 2005.

[16] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble. Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6559–6563, 2004.

[17] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, Saint Petersburg, Russia, 2009, pages 565–576.

[18] T. Lee and C.-N. Yang. Statistical theory of equations of state and phase transitions. II. Lattice gas and Ising model. *Physical Review*, 87(3):410–419, 1952.