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 highlights 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 \( V \) and edges \( E \) is the same that of the underlying data points. First consider the overcomplete parameterization:

\[
p(x|\theta) = \frac{1}{Z} \exp \{-E(x|\theta)\}
\]

\[
E(x|\theta) = -\sum_{i \in V} \theta_i(x_i) - \sum_{(i,j) \in E} \theta_{ij}(x_i, x_j) \quad x_i, x_j \in \{0, 1\}
\]  

We derive \( \theta_i \) and \( \theta_{ij} \) 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

\[
E(x|\theta) = \alpha - \sum_{i \in V} (s_i + b)x_i - \sum_{(i,j) \in E} wx_i x_j
\]

\[
b = \beta - \alpha
\]

\[
w = \alpha + \gamma - 2\beta = \gamma - \beta - b
\]

So \( \alpha \) 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 \( \alpha \), favoring both nodes to function, and \( \gamma \), 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 \( \epsilon \). We use the parameterization in [11] which requires only singleton (not pairwise) marginals:

\[
F_B(q) = \sum_{i \in V} -\theta_i q_i + (\delta_i - 1)S_1(q_i) + \sum_{(i,j) \in E} - (W_{ij} \xi_{ij} + S_2(q_i, q_j))
\]
Due to runtime, we report pseudomarginals as the minimizer of $W$ is the maximum weight on an edge. where $\delta_i$ is degree of node $i$, $q_i$ is a singleton pseudomarginal, $\theta_i$ and $W_{ij}$ are singleton and pairwise weights, $S_1$ and $S_2$ are singleton and pairwise entropy functions, and each $\xi_{ij}$ is the unique root of a quadratic function of $W_{ij}, q_i, q_j$. 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 $W$ is far from the true (Gibbs) free energy. 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 $\epsilon$-close guarantees. Note that this is still not a bound 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 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 $w$. Redder means higher values. A phase transition characteristic of Ising models appears as a diagonal line. An $\times$ marks the parameters of the true distribution prior to rejection sampling. See text for details.

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

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.
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 \( \times \) 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 \( \ell_1 \) 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 \( 5 \times 10^7 \) 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


