Message Passing for Collective Graphical Models

Tao Sun^1 **Daniel Sheldon**^{1,2} **Akshat Kumar**^{3}

^{1} College of Information and Computer Science, University of Massachusetts Amherst
^{2} Department of Computer Science, Mount Holyoke College
^{3} School of Information System, Singapore Management University
{taosun, sheldon}@cs.umass.edu akshatkumar@smu.edu.sg

Abstract

Collective graphical models (CGMs) are a formalism for inference and learning about a population of i.i.d. individuals when only noisy aggregate data are available. Our recent paper [19] highlights a close connection between approximate MAP inference in CGMs and *marginal inference* in standard graphical models. The connection leads us to derive a novel Belief Propagation (BP) style algorithm for collective graphical models. Mathematically, the algorithm is a strict generalization of BP — it can be viewed as an extension to minimize the Bethe free energy plus additional energy terms that are non-linear functions of the marginals. For CGMs, the algorithm is much more efficient than previous approaches to inference. We demonstrate its state-of-the-art performance on a synthetic bird migration benchmark, we also contribute to a novel application about synthetic collective human mobility, where cell phone carriers anonymize the data to protect differential privacy, but we could recover the data to some accuracy level and thus allow for downstream applications with the recovered data.

1 Introduction

In an influential paper, Yedidia et al. [23] showed that the loopy Belief Propagation (BP) algorithm for marginal inference in graphical models can be understood as a fixed-point iteration that attempts to satisfy the first-order optimality conditions of the Bethe free energy, which approximates the true variational free energy. The result shed considerable light on the convergence properties of BP and led to many new ideas for approximate variational inference.

Our recent paper [19] highlights a connection between the Bethe free energy and the objective function for approximate MAP inference in Collective Graphical Models (CGMs) [15], which are models for inference and learning about individuals when only noisy aggregate data are available. We then follow reasoning similar to that of Yedidia et al. to derive a novel message-passing algorithm for CGMs. The algorithm, *non-linear energy belief propagation* (NLBP), has the interesting property that message updates are identical to BP, *with the exception that edge potentials change in each step* based on the gradient of the non-linear "evidence terms" that are present in the CGM objective but not in the Bethe free energy. NLBP is a strict generalization of BP to deal with the presence of these additional non-linear terms.

The new algorithm has significant practical benefits. We show experimentally that, by exploiting the graph structure, NLBP solves the approximate MAP optimization problem for CGMs much faster than a generic optimization solver, and scales significantly better than previous approaches for inference in CGMs. NLBP advances applications of CGMs by significantly reducing the computational burden of inference, which was previously a limiting factor. We demonstrate this point through two synthetic applications. First, we apply CGMs to the problem of modeling bird migration [14, 15, 10], where inference is used to reconstruct bird migration routes, make forecasts of migration, and learn parameters of migration models. Our algorithm lets us scale from small problems to realistic-sized problems. Second, we contribute a novel application for modeling human mobility [2, 8]. In this case, data providers (e.g., cell phone companies) release aggregate statistics about human movements for the purpose of model fitting, but corrupt those statistics with noise to guarantee differential privacy [4, 1, 5, 11]. CGM inference algorithms provide a way to reason about the true sufficient statistics for the purpose of learning. We show that a CGM-based learning algorithm that uses NLBP is much more accurate than a baseline approach that uses noisy statistics directly for parameter estimation.

2 Collective Graphical Models

CGMs compactly describe the distribution of the aggregate statistics of a population sampled independently from a discrete graphical model. Let $G = (V, E)$ be an undirected graph, and consider the following pairwise graphical model over the discrete random vector $\boldsymbol{X} = (X_1, \dots, X_{|V|})$:

$$p(\mathbf{x};\boldsymbol{\theta}) = \Pr(\boldsymbol{X}=\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{(i,j)\in E} \phi_{ij}(x_i, x_j;\boldsymbol{\theta}). \tag{1}$$

Here, $\phi_{ij}(\cdot,\cdot;\boldsymbol{\theta})$ is a local potential defined on the setting of variables (X_i, X_j). The local potentials are controlled by a parameter vector $\boldsymbol{\theta}$, and $Z(\boldsymbol{\theta})$ is the partition function. We assume for simplicity that each variable X_i takes values in the same finite set $\mathcal{X}$. We also assume henceforth that G is a tree. For graphical models that are not trees or have higher-order potentials, our results can be generalized to junction trees, with the usual blowup in space and running-time depending on the clique-width of the junction tree.

Now, consider an ordered sample $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}$ of random vectors drawn independently from the graphical model. We also refer to this sample as a *population*. Define the contingency tables $\mathbf{n}_i = (n_i(x_i) : x_i \in \mathcal{X})$ over nodes of the model and $\mathbf{n}_{ij} = (n_{ij}(x_i, x_j) : x_i, x_j \in \mathcal{X})$ over edges of the model, whose entries count the number of times particular variable settings occur in the population:

$$n_i(x_i) = \textstyle\sum_{m=1}^{M} \mathbb{I}\big(X_i^{(m)} = x_i\big), \quad n_{ij}(x_i, x_j) = \sum_{m=1}^{M} \mathbb{I}\big(X_i^{(m)} = x_i,\ X_j^{(m)} = x_j\big).$$

Here, $\mathbb{I}(\cdot)$ is an indicator function. Define the vector $\mathbf{n}$ to be the concatenation of all edge-based contingency tables $\mathbf{n}_{ij}$ together with all node-based contingency tables $\mathbf{n}_i$. This is a random vector that depends on the entire population and comprises sufficient statistics of the population, which can be seen by writing the joint probability:

$$p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)};\boldsymbol{\theta}) = g(\mathbf{n}, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})^M} \prod_{(i,j)\in E} \prod_{x_i, x_j} \phi_{ij}(x_i, x_j;\boldsymbol{\theta})^{n_{ij}(x_i,x_j)}. \tag{2}$$

In CGMs, one makes noisy observations $\mathbf{y}$ of some subset of the sufficient statistics $\mathbf{n}$ and then seeks to answer queries about the sufficient statistics given $\mathbf{y}$ (e.g., for the purpose of learning the parameters $\boldsymbol{\theta}$) through the conditional distribution $p(\mathbf{n}\,|\,\mathbf{y};\boldsymbol{\theta}) \propto p(\mathbf{n};\boldsymbol{\theta})p(\mathbf{y}\,|\,\mathbf{n})$. The first term in this product, $p(\mathbf{n};\boldsymbol{\theta})$, is the prior distribution over the sufficient statistics or the *CGM distribution*. In Section 2.2, we will describe how the CGM distribution is derived from the individual model (1). We refer to the second term, $p(\mathbf{y}\,|\,\mathbf{n})$, as the *noise model* or the *CGM evidence term*. It is often assumed that $p(\mathbf{y}\,|\,\mathbf{n})$ is log-concave in $\mathbf{n}$, which makes the negative log-likelihood convex in $\mathbf{n}$, though most of the results of this paper do not rely on that assumption.

Example. For modeling bird migration, assume that $X = (X_1, \dots, X_T)$ is the sequence of discrete locations (e.g. map grid cells) visited by an individual bird, and that the graphical model $p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{t=1}^{T-1} \phi_t(x_t, x_{t+1};\boldsymbol{\theta})$ is a chain model governing the migration of an individual, where the parameter vector $\boldsymbol{\theta}$ controls how different relevant factors (distance, direction, time of year, etc.) influence the affinity $\phi_t(x_t, x_{t+1};\boldsymbol{\theta})$ between two locations x_t and x_{t+1}. In the CGM, M birds of a given species independently migrate from location to location according to the chain model. The node-table entries $n_t(x_t)$ indicate how many birds are in location x_t at time t. The edge-table entries $n_{t,t+1}(x_t, x_{t+1})$ count how many birds move from location x_t to location x_{t+1} from time t to time $t + 1$. A reasonable model for eBird data [12, 18] is that the number of birds of the target species counted by a birdwatcher is a Poisson random variable with mean proportional to the true number of birds $n_t(x_t)$, or $y_t(x_t)\,|\,n_t(x_t) \sim \text{Pois}(\alpha n_t(x_t))$, where α is the detection rate. Given only the noisy eBird counts and the prior specification of the Markov chain, the goal is to answer queries about the distribution $p(\mathbf{n} \mid \mathbf{y}; \boldsymbol{\theta})$ to inform us about migratory transitions made by the population. Because the vector $\mathbf{n}$ consists of sufficient statistics, these queries also provide all the relevant information for learning the parameters $\boldsymbol{\theta}$ from this data.

2.1 CGM Distribution

We now describe the form of the CGM distribution $p(\mathbf{n}; \boldsymbol{\theta})$ and basic aspects of inference in this distribution. Sundberg [20] originally described the form of this distribution for a graphical model that is decomposable (i.e., its cliques are the nodes of some junction tree), in which case its probabilities can be written in closed form in terms of the marginal probabilities of the original model. Liu et al. [10] refined this result to be written in terms of the original potentials instead of marginal probabilities. Applied to our tree-structured model, this gives the following CGM distribution:

$$p(\mathbf{n}; \boldsymbol{\theta}) = M! \frac{\prod_{i \in V} \prod_{x_i \in \mathcal{X}} \left(n_i(x_i)!\right)^{\nu_i - 1}}{\prod_{(i,j) \in E} \prod_{x_i, x_j \in \mathcal{X}} n_{ij}(x_i, x_j)!} \cdot g(\mathbf{n}, \boldsymbol{\theta}) \cdot \mathbb{I}(\mathbf{n} \in \mathbb{L}_M^{\mathbb{Z}}). \tag{3}$$

The first term is a base measure (it does not depend on the parameters) that counts the number of different ordered samples that give rise to the sufficient statistics $\mathbf{n}$; in this term, ν_i is the degree of node i. The second term, $g(\mathbf{n}, \boldsymbol{\theta})$, is the joint probability of any ordered sample with sufficient statistics $\mathbf{n}$ as defined in Eq. (2). The final term is a hard constraint that restricts the support of the distribution to vectors $\mathbf{n}$ that are valid sufficient statistics of some ordered sample. Sheldon and Dietterich [16] showed that, for trees or junction trees, this requirement is satisfied if and only if $\mathbf{n}$ belongs to the *integer-valued scaled local polytope* $\mathbb{L}_M^{\mathbb{Z}}$ defined by the following constraints:

$$\mathbb{L}_M^{\mathbb{Z}} = \Big\{ \mathbf{n} \in \mathbb{Z}_+^{|\mathbf{n}|} \;\Big|\; M = \sum_{x_i} n_i(x_i) \;\; \forall i \in V, n_i(x_i) = \sum_{x_j} n_{ij}(x_i, x_j) \;\; \forall i \in V, x_i \in \mathcal{X}, j \in N(i) \Big\}, \tag{4}$$

where $N(i)$ is the set of neighbors of i. The reader will recognize that $\mathbb{L}_M^{\mathbb{Z}}$ is equivalent to the standard local polytope of a graphical model [21] except for two differences: (1) the marginals, which in our case represent counts instead of probabilities, are scaled to sum to the population size M instead of summing to one, and (2) these counts are constrained to be integers. The set $\mathbb{L}_M^{\mathbb{Z}}$ is the true support of the CGM distribution. Let $\mathbb{L}_M$ be the relaxation of $\mathbb{L}_M^{\mathbb{Z}}$ obtained by removing the integrality constraint, i.e., the set of real-valued vectors with non-negative entries that satisfy the same constraints.

2.2 Approximate MAP Inference

The MAP inference problem for CGMs is to find $\mathbf{n} \in \mathbb{L}_M^{\mathbb{Z}}$ to maximize $p(\mathbf{n} \mid \mathbf{y}; \boldsymbol{\theta})$. Henceforth, we will suppress the dependence on $\boldsymbol{\theta}$ to simplify notation when discussing inference with respect to fixed parameters. Unfortunately, exact MAP inference is intractable [15], but by relaxing the feasible set from $\mathbb{L}_M^{\mathbb{Z}}$ to $\mathbb{L}_M$ (i.e., removing the integrality requirement), taking the negative log of the objective, and using Stirling's approximation, Sheldon et al. [15] arrived at the following convex relaxation of the MAP problem:

$$\min_{\mathbf{z} \in \mathbb{L}_M} F_{\text{CGM}}(\mathbf{z}) := E_{\text{CGM}}(\mathbf{z}) - H_B(\mathbf{z}). \tag{5}$$

$$E_{\text{CGM}}(\mathbf{z}) = - \sum_{(i,j) \in E} \sum_{x_i, x_j} z_{ij}(x_i, x_j) \log \phi_{ij}(x_i, x_j) - \log p(\mathbf{y} \mid \mathbf{z}),$$

$$H_B(\mathbf{z}) = - \sum_{(i,j) \in E} \sum_{x_i, x_j} z_{ij}(x_i, x_j) \log z_{ij}(x_i, x_j) + \sum_{i \in V} (\nu_i - 1) \sum_{x_i} z_i(x_i) \log z_i(x_i).$$

We write $\mathbf{z}$ in place of $\mathbf{n}$ to emphasize that the contingency tables are now real-valued. The quantity $H_B(\mathbf{z})$ is the *Bethe entropy*. It is well known that the Bethe entropy is concave over the local polytope of a tree [7]. We have grouped the remaining terms into the *CGM energy function* $E_{\text{CGM}}(\mathbf{z})$ for comparison with the free energies we will discuss below. If the noise model $p(\mathbf{y} \mid \mathbf{n})$ is log-concave then the overall problem is convex and can be solved by off-the-shelf solvers [15]. This inference approach is extremely accurate and much faster than the previous method of Gibbs sampling, but it is still not efficient enough for large-scale problems.

3 Non-Linear Energy Belief Propagation (NLBP)

We generalize the analysis by Yedidia et al. [23] to derive a belief propagation (BP) algorithm for *arbitrary* non-linear energies $E(\mathbf{z})$ such as $E_{\text{CGM}}(\mathbf{z})$ in the CGM MAP objective. Interested reader should refer to [19] for derivation. Given

$$\min_{\mathbf{z} \in \mathbb{L}_M} F(\mathbf{z}) := E(\mathbf{z}) - H_B(\mathbf{z}), \tag{9}$$

the energy function $E(\mathbf{z})$ need not to be linear with respect to node and edge marginals. Algorithm 1 shows Non-Linear Belief Propagation (NLBP). Note that the *only* difference from standard BP is that we replace the edge potential $\phi_{ij}(x_i, x_j)$ by the exponentiated negative gradient of $E(\mathbf{z})$. For standard linear energy $E_B(\mathbf{z})$, this is always equal to the original edge potential, and we recover standard BP. For non-linear energies, the gradient is not constant with respect to $\mathbf{z}$, so, unlike in standard BP, we must track the value of the marginals $\mathbf{z}$ (normalized to sum to M) in each iteration so we can use them to update the current edge potentials. Note that the algorithm stores the current edge potentials $\widehat{\phi}_{ij}$ as separate variables, which is not necessary but will add useful flexibility in ordering updates.

```
Algorithm 1: Non-Linear Belief Propagation
Input: Graph $G = (V, E)$, (non-linear) energy function
       $E(\mathbf{z})$, population size $M$
Init : $m_{ij}(x_j) = 1$, $\widehat{\phi}_{ij}(x_i, x_j) = \phi_{ij}(x_i, x_j)$,
       $z_{ij}(x_i, x_j) \propto \phi_{ij}(x_i, x_j)$,  $\forall (i, j) \in E, x_i, x_j$.
while $\neg$ converged do
    Execute the following updates in any order:
    $\widehat{\phi}_{ij}(x_i, x_j) = \exp\Big\{-\frac{\partial E(\mathbf{z})}{\partial z_{ij}(x_i, x_j)}\Big\}$                                  (6)
    $m_{ij}(x_j) \propto \sum_{x_i} \widehat{\phi}_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i)$                           (7)
    $z_{ij}(x_i, x_j) \propto \widehat{\phi}_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i) \prod_{l \in N(j) \setminus i} m_{lj}(x_j)$
                                                                          (8)
Extract node marginals: $z_i(x_i) \propto \prod_{k \in N(i)} m_{ki}(x_i)$
```

Theorem 1. *Suppose the NLBP message passing updates converge and the resulting vector $\mathbf{z}$ has strictly positive entries. Then $\mathbf{z}$ is a constrained stationary point of $F(\mathbf{z})$ in Problem (9) with respect to the set $\mathbb{L}_M$. If G is a tree and $E(\mathbf{z})$ is convex, then $\mathbf{z}$ is a global minimum.*

Our proof in [19] does not rely on convexity of the noise term except to guarantee that a global minimum is reached in the case of tree-structured models. Also note that NLBP maintains positive marginals as long as the gradient of $E(\mathbf{z})$ is finite (which is analogous to the assumption of positive potentials in the linear case), so the assumption of positivity is not overly restrictive. Unlike standard BP, which is guaranteed to converge in one pass for trees, in NLBP the edge potentials change with each iteration so it is an open question whether convergence is guaranteed even for trees. In practice, we find it is necessary to damp the updates to messages [6] and marginals $\mathbf{z}$, and that sufficient damping always leads to convergence in our experiments. See Algorithm 2 for details of damping.

3.1 Edge Evidence vs. Node Evidence

In our applications we consider two primary types of CGM observations, one where noisy edge counts are observed and one where noisy node counts are observed. In both cases, we assume the table entries are corrupted independently by a univariate noise model $p(y \mid z)$:

$$\begin{aligned} p_{\text{edge}}(\mathbf{y} \mid \mathbf{z}) &= \prod_{(i,j) \in E, x_i, x_j} p(y_{ij}(x_i, x_j) \mid z_{ij}(x_i, x_j)), \\ p_{\text{node}}(\mathbf{y} \mid \mathbf{z}) &= \prod_{i, x_i} p(y_i(x_i) \mid z_i(x_i)). \end{aligned} \tag{10}$$

The first model occurs in our human mobility application: a data provider wishes to release *sufficient* statistics (edge tables) but must add noise to those statistics to maintain privacy. The second model occurs in our bird migration application: birdwatchers submit counts that provide evidence only about the locations of birds at a particular time, and not about the migratory transitions they make.

With noisy edge counts, it is clear how to update the edge potentials within NLBP. Let $\ell(z \mid y) = -\log p(y \mid z)$. Eq (6) becomes $\widehat{\phi}_{ij}(x_i, x_j) = \phi_{ij}(x_i, x_j) \exp\Big\{\ell'\big(z_{ij}(x_i, x_j) \mid y_{ij}(x_i, x_j)\big)\Big\}$, where ℓ' is the partial derivative with respect to the marginal. With noisy node counts, we must rewrite

$p(\mathbf{y}\,|\,\mathbf{z})$ using only the edge tables. We choose to write $z_i(x_i) = \frac{1}{\nu_i}\sum_{j\in N(i)}\sum_{x_j} z_{ij}(x_i,x_j)$ as the average of the marginal counts obtained from all incident edge tables. This leads to symmetric updates in Eq (6): $\widehat{\phi}_{ij}(x_i,x_j) = \phi_{ij}(x_i,x_j)\exp\Big\{\frac{1}{\nu_i}\ell'\big(y_i(x_i)\,|\,z_i(x_i)\big)+\frac{1}{\nu_j}\ell'\big(y_j(x_j)\,|\,z_j(x_j)\big)\Big\}$, where $z_i(x_i)$ and $z_j(x_j)$ are marginal counts of $\mathbf{z}_{ij}$.

3.2 Update Schedules and Feasibility Preservation

The NLBP algorithm is a fixed-point iteration that allows updating of edge potentials, messages, and the marginals in any order. We first considered a *naive schedule*, where message updates are sequenced as in standard BP (for trees, in a pass from leaves to root and then back). When message m_{ij} is scheduled for update, the operations are performed in the order listed in Algorithm 1: first the edge potential is updated, then the message is updated, and then all marginals that depend on m_{ij} are updated. Unlike BP, this algorithm does not achieve convergence in one round, so the entire process is repeated until convergence. In our initial experiments, we discovered that the naive schedule can take many iterations to achieve a solution that satisfies the consistency constraints among marginals (Eq. (4)).

We devised a second *feasibility-preserving schedule* (Algorithm 2) that always maintains feasibility and has the appealing property that it can be implemented as a simple wrapper around standard BP. This algorithm specializes NLBP to alternate between two phases. In the first phase, edge potentials are frozen while messages and marginals are updated in a full pass through the tree. This is equivalent to one call to the standard BP algorithm, which, for trees, is guaranteed to converge in one pass and return feasible marginals. In the second phase, only edge potentials are updated. Algorithm 2 maintains the property that it's current iterate $\mathbf{z}$ is always a convex combination of feasible marginals returned by standard BP, so $\mathbf{z}$ is also feasible.

Algorithm 2: Feasibility Preserving NLBP

```
Input same as Algorithm 1, damping parameter $\alpha \geq 0$
Init : $\mathbf{z} \leftarrow$ STANDARD-BP$\big(\{\boldsymbol{\phi}_{ij}\}\big)$
while $\neg$ converged do
    $\widehat{\phi}_{ij}(x_i,x_j) \leftarrow \exp\Big\{-\frac{\partial E(\mathbf{z})}{\partial z_{ij}(x_i,x_j)}\Big\}, \forall (i,j) \in E$
    $\mathbf{z}^{\text{new}} \leftarrow$ STANDARD-BP$\big(\{\widehat{\boldsymbol{\phi}}_{ij}\}\big)$
    $\mathbf{z} \leftarrow (1-\alpha)\mathbf{z} + \alpha\mathbf{z}^{\text{new}}$ ;        // damped updates
```

4 Evaluation

4.1 Speed of Inference: Synthetic Bird Migration

At first we evaluate the extent to which NLBP accelerates CGM inference and learning for a benchmark synthetic bird migration problem [15, 10]. We compared the speed and accuracy of NLBP both as a standalone inference method and as a subroutine for learning versus the baselines of using MATLAB's interior-point algorithm to solve the approximate MAP problem [15] and inference in the Gaussian approximation of CGMs [10].

We leave the detailed experiments setup and analysis in our paper [19]. The results show that NLBP is not only highly scalable and much faster than all competitors, it also maintains high accuracy across a wide range of parameter settings.

4.2 Collective Human Mobility

Next we turn to a novel application of CGMs. We address the problem of learning the parameters of a chain-structured graphical model for human mobility, where, unlike the bird migration model, we have access to *transition counts* (edge counts) instead of node counts. Transition counts are sufficient statistics for the model, so learning with *exact* transition counts would be straightforward. However, we assume the available data are corrupted by noise to maintain privacy of individuals. The problem becomes one of learning with noisy sufficient statistics.

In particular, our application simulates the following situation: a mobile phone carrier uses phone records to collect information about the transitions of people among a discrete set of regions, for example, the areas closest to each mobile tower, which form a Voronoi tesselation of space [17, 3, 1].

Data is aggregated into discrete time steps to provide hourly counts of the number of people that move between each pair of regions. The provider wishes to release this aggregate data to inform public policy and scientific research about human mobility. However, to maintain the privacy of their customers, they choose to release data in a way that maintains the privacy guarantees of *differential privacy* [4, 11, 1, 5]. In particular, they follow the Laplace mechanism and add independent Laplace noise to each aggregate count [4].

Ground-Truth Model. We are interested in fitting models of daily commuting patterns from aggregate data of this form. We formulate a synthetic version of this problem where people migrate among the grid cells of a 15×14 rectangular map. We simulate movement from home destinations to work destinations across a period of $T = 10$ time steps (e.g., half-hour periods covering the period from 6:00 a.m. to 11:00 a.m.). We parameterize the joint probability of the movement sequence for each individual as:

$$p(\mathbf{x}_{1:10}) = \frac{1}{Z} \cdot \phi_1(x_1) \cdot \left(\prod_{t=1}^{9} \psi(x_t, x_{t+1}) \right) \cdot \phi_{10}(x_{10}).$$

The potentials ϕ_1 and ϕ_{10} represent preferences for home and work locations, respectively, while ψ is a pairwise potential that scores transitions as more or less preferred. For the ground truth model, we use compact parameterizations for each potential: ϕ_1 and ϕ_{10} are discretized Gaussian potentials (that is, $\phi(x_t)$ is the value of a Gaussian density over the map measured at the center of grid cell x_t) centered around a "residential area" (top right of the map) and "commercial area" (bottom left). For the transition potential, we set $\phi(x_t, x_{t+1})$ proportional to $\exp\big(- ||v_t - v_{t+1}||^2/(2\sigma^2)\big)$, where v_t and v_{t+1} are the centers of grid cells x_t and x_{t+1}, to prefer short transitions over long ones.

Data Generation. To generate data, we simulated $M = 1$ million trajectories from the ground truth model, computed the true transition counts, and then added independent Laplace noise to each true count n to generate the noisy count y. The Laplace noise is controlled by a scale parameter b:

$$p(y \mid n) = \text{Laplace}(b; n) = \frac{1}{2b} \exp\left\{ -\frac{|y-n|}{b} \right\}.$$

To explore the relative power of edge counts versus node counts for model fitting, we also performed a version of the experiments where we marginalized the noisy transition counts to give only noisy node counts $y_t(x_t) = \sum_{x_{t+1}} y_{t,t+1}(x_t, x_{t+1})$ as evidence.

Parameters and Evaluation. We wish to compare the abilities of CGM-based algorithms and a baseline algorithm to recover the true mobility model. When *fitting* models, it would be a severe oversimplification to assume the simple parametric form used to generate data. Instead, we use a fully parameterized model with parameters $\boldsymbol{\theta} = (\log \phi_1, \log \phi_{10}, \log \psi)$. Here $\log \phi_1$ and $\log \phi_{10}$ are arbitrary $L \times 1$ vectors, and $\log \psi$ is an arbitrary $L \times L$ table. Note that this parameterization is over-complete, and hence not identifiable. To evaluate fitted models, we will compare their *pairwise marginal distributions* to those of the ground truth model: unlike the potentials, the pairwise marginals uniquely identify the joint distribution. The pairwise MAE is defined as the mean absolute error among all $L^2 \times (T-1)$ entries of the pairwise marginals. We also considered node MAE, which is the mean error among the $L \times T$ entries of the node marginals. Note that these *do not* uniquely identify the distribution, but node MAE is an interesting metric for comparing the ability to learn with node evidence vs. edge evidence.

Algorithms. Is it possible to estimate parameters of a graphical model given only noisy sufficient statistics? An "obvious" approach is to ignore the noise and perform maximum-likelihood estimation using the noisy sufficient statistics $\mathbf{y}$ in place of the true ones $\mathbf{n}$. To the best of our knowledge, this is the only previously available approach, and we use it as a baseline. The approach has been criticized in the context of general multidimensional contingency tables [22]. To maximize the likelihood with respect to our parameters, we use a gradient-based optimizer with message passing as a subroutine to compute the likelihood and its gradient [9].

For the CGM-based approach, we treat the true sufficient statistics as hidden variables and use EM to maximize the likelihood. The overall EM approach is the same as in the bird migration model. When the evidence is noisy edge counts, we first run the baseline algorithm and use those parameters to initialize EM. When the evidence is noisy node counts, the baseline algorithm does not apply and we initialize the parameters randomly.

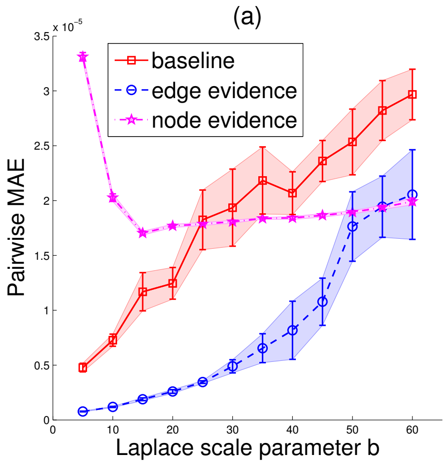

Figure 1: Pairwise / Node MAE vs Laplace scale parameter b after 250 EM iterations. Shaded regions shows 95% confidence intervals computed from 10 trials for each setting.

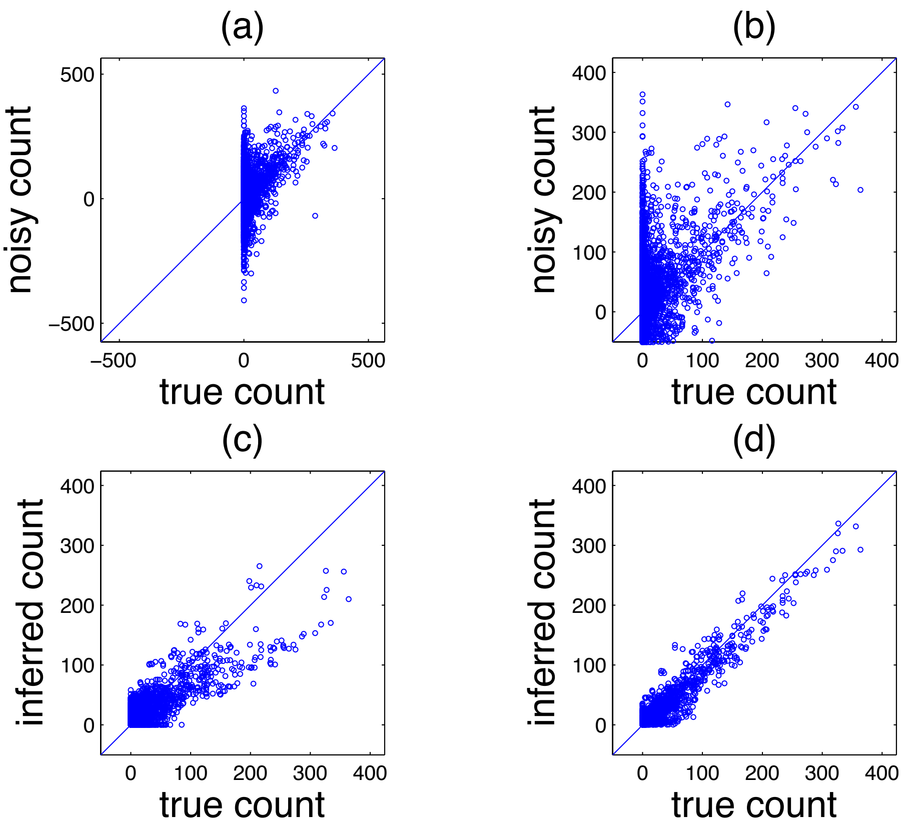

Figure 2: Scatter plots of approximate vs. true edge counts for a small problem ($L = 4 \times 7, T = 5, M = 10000, b = 50$): (a) original noisy edge counts, (b) shown only in the same range as (c-d) for better comparison, (c) reconstructed counts after 1 EM iteration, (d) reconstructed counts after EM convergence.

Results. Figure 1(a) shows the quality of the fitted models (measured by pairwise MAE) vs. the scale of the Laplace noise. For the CGM-based algorithms, we ran 250 EM iterations, which was enough for convergence in almost all cases. Initializing EM with the baseline parameters helped achieve faster convergence (not shown). The results demonstrate that the CGM algorithm with edge evidence improves significantly over the baseline for all values of b. As expected, the node evidence version of the CGM algorithm performs worse, since it has access to less information. However, it is interesting that the CGM with only node evidence outperforms the baseline (which has access to more information) for larger values of b.

Figure 1(b) shows *node* MAE vs b for the same fitted models. In other words, it measures the ability of the methods to find models that match the ground truth on single time-step marginals. We see that both CGM algorithms are substantially better than the baseline, and the CGM algorithm with *less information* (node counts only) performs slightly better. We interpret this as follows: node evidence alone provides enough information to match the ground truth model on node marginals; the additional information of the noisy edge counts helps narrow the model choices to one that also matches the ground truth edge marginals. However, this does not explain why the node evidence performs *better* than edge evidence for node MAE. We leave a deeper investigation of this for future work—it may be a form of implicit regularization.

Figure 2 provides some insight into the EM algorithm and it's ability to reconstruct edge counts. The original, noisy counts have considerable noise and sometimes take negative values (panels (a) and (b)). After one EM iteration (panel (c)), the reconstructed counts are now feasible, so they can no longer be negative, and they are closer to the original counts. After EM converges, the reconstructed counts are much more accurate (panel (d)).

Challenges and future directions. We generated the noisy observation $y_{ij}(x_i, x_j)$ and $y_i(x_i)$ from $z_{ij}(x_i, x_j)$ and $z_i(x_i)$ accordingly as in Eq. (10). This could be an over-simplification. Some recent works [13, 1] perform Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT) to the aggregated time series data, and then perturb the coefficients with noise before releasing their data. In CGMs, this corresponds to a more general noise model, for example, each y_i could link to all z_i. Another challenge is that we need to generalize our model to deal with multiple "frequent routes" [5] simultaneously. The third challenge comes from data availability. To the best of our knowledge there is still no Call Detail Records data that is publicly available.

5 Conclusion

Our paper [19] highlights a close connection between the problems of approximate MAP inference in collective graphical models (CGMs) and marginal inference in standard graphical models. Inspired by this connection, we derived the non-linear belief propagation (NLBP) algorithm and presented a feasibility-preserving version of NLBP that can be implemented as a simple wrapper around standard BP. By applying NLBP to a synthetic benchmark problem for bird migration modeling, we showed that NLBP runs significantly faster than a generic solver and is significantly more accurate than inference in the Gaussian approximation of CGMs when the grid size or parameter values are large. The feasibility-preserving version of NLBP is twice as fast as the naive NLBP. We then demonstrated the utility of the NLBP algorithm by contributing a novel application of CGMs for modeling human mobility. In this application, CGMs provide a way to fit graphical models when the available sufficient statistics have been corrupted by noise to maintain the privacy of individuals.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1125228, and by the National Research Foundation Singapore under its Corp Lab@University scheme.

References

[1] G. Acs and C. Castelluccia. A case study: privacy preserving release of spatio-temporal density in paris. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1679–1688. ACM, 2014.

[2] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[3] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[4] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[5] O. Görnerup, N. Dokoohaki, and A. Hess. Privacy-preserving mining of frequent routes in cellular network data. In *IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2015.

[6] T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. *Advances in Neural Information Processing Systems (NIPS)*, 15:359–366, 2003.

[7] T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26(1):153–190, 2006.

[8] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. ACM, 2012.

[9] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[10] L.-P. Liu, D. Sheldon, and T. G. Dietterich. Gaussian approximation of collective graphical models. In *International Conference on Machine Learning (ICML)*, volume 32, pages 1602–1610, 2014.

[11] D. J. Mir, S. Isaacman, R. Caceres, M. Martonosi, and R. N. Wright. Dp-where: Differentially private modeling of human mobility. In *Big Data, 2013 IEEE International Conference on*, pages 580–588. IEEE, 2013.

[12] A. M. Munson, K. Webb, D. Sheldon, D. Fink, W. M. Hochachka, M. Iliff, M. Riedewald, D. Sorokina, B. Sullivan, C. Wood, and S. Kelling. The ebird reference dataset, version 5.0. Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY, January 2013.

[13] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 735–746. ACM, 2010.

[14] D. Sheldon, M. A. S. Elmohamed, and D. Kozen. Collective inference on Markov models for modeling bird migration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1321–1328, 2007.

[15] D. Sheldon, T. Sun, A. Kumar, and T. G. Dietterich. Approximate inference in collective graphical models. In *International Conference on Machine Learning (ICML)*, volume 28, pages 1004–1012, 2013.

[16] D. R. Sheldon and T. G. Dietterich. Collective graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1161–1169, 2011.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327 (5968):1018–1021, 2010.

[18] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

[19] T. Sun, D. Sheldon, and A. Kumar. Message passing for collective graphical models. In D. Blei and F. Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 853–861. JMLR Workshop and Conference Proceedings, 2015.

[20] R. Sundberg. Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scandinavian Journal of Statistics*, 2(2): 71–79, 1975.

[21] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[22] X. Yang, S. E. Fienberg, and A. Rinaldo. Differential privacy for protecting multi-dimensional contingency table data: Extensions and applications. *Journal of Privacy and Confidentiality*, 4(1):5, 2012.

[23] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.