# Monte Carlo MCMC: Efficient Inference by Approximate Sampling

**Sameer Singh**
University of Massachusetts
140 Governor's Drive
Amherst MA
sameer@cs.umass.edu

**Michael Wick**
University of Massachsetts
140 Governor's Drive
Amherst, MA
mwick@cs.umass.edu

**Andrew McCallum**
University of Massachusetts
140 Governor's Drive
Amherst MA
mccallum@cs.umass.edu

## Abstract

Conditional random fields and other graphical models have achieved state of the art results in a variety of tasks such as coreference, relation extraction, data integration, and parsing. Increasingly, practitioners are using models with more complex structure—higher treewidth, larger fan-out, more features, and more data—rendering even approximate inference methods such as MCMC inefficient. In this paper we propose an alternative MCMC sampling scheme in which transition probabilities are approximated by sampling from the set of relevant factors. We demonstrate that our method converges more quickly than a traditional MCMC sampler for both marginal and MAP inference. In an author coreference task with over 5 million mentions, we achieve a 13 times speedup over regular MCMC inference.

## 1 Introduction

Conditional random fields and other graphical models are at the forefront of many natural language processing (NLP) and information extraction (IE) tasks because they provide a framework for discriminative modeling while succinctly representing dependencies among many related output variables. Previously, most applications of graphical models were limited to structures where exact inference is possible, for example linear-chain CRFs (Lafferty et al., 2001). More recently, there has been a desire to include more factors, longer range dependencies, and more sophisticated features; these include skip-chain CRFs for named entity recognition (Sutton and McCallum, 2004), probabilistic DBs (Wick et al., 2010), higher-order models for dependency parsing (Carreras, 2007), entity-wise models for coreference (Culotta et al., 2007; Wick et al., 2009), and global models of relations (Hoffmann et al., 2011). The increasing sophistication of these individual NLP components compounded with the community's desire to model these tasks jointly across cross-document considerations has resulted in graphical models for which inference is computationally intractable. Even popular approximate inference techniques such as loopy belief propagation and Markov chain Monte Carlo (MCMC) may be prohibitively slow.

MCMC algorithms such as Metropolis-Hastings are usually efficient for graphical models because the only factors needed to score a proposal are those touching the changed variables. However, MCMC is slowed in situations where a) the model exhibits variables that have a high-degree (neighbor many factors), b) proposals modify a substantial subset of the variables to satisfy domain constraints (such as transitivity in coreference), or c) evaluating a single factor is expensive, for example when features are based on string-similarity. For example, the seemingly innocuous proposal changing the entity type of a single entity requires examining all its mentions, i.e. scoring a linear number of factors (in the number of mentions of that entity). Similarly, evaluating coreference of a mention to an entity also requires scoring factors to all the mentions of the entity. Often, however, the factors are somewhat *redundant*, for example, not all mentions of the "USA" entity need to be examined to confidently conclude that it is a COUNTRY, or that it is coreferent with "United

*Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1104–1113, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

States of America".

In this paper we propose an approximate MCMC framework that facilitates efficient inference in high-degree graphical models. In particular, we approximate the acceptance ratio in the Metropolis Hastings algorithm by replacing the exact model score with a stochastic approximation that samples from the set of relevant factors. We explore two sampling strategies, a fixed proportion approach that samples the factors uniformly, and a dynamic alternative that samples factors until the method is confident about its estimate of the model score.

We evaluate our method empirically on both synthetic and real-world data. On synthetic classification data, our approximate MCMC procedure obtains the true marginals faster than a traditional MCMC sampler. On real-world tasks, our method achieves 7 times speedup on citation matching, and 13 times speedup on large-scale author disambiguation.

## 2 Background

### 2.1 Graphical Models

Factor graphs (Kschischang et al., 2001) succinctly represent the joint distribution over random variables by a product of factors that make the dependencies between the random variables explicit. A factor graph is a bipartite graph between the variables and factors, where each (log) factor $f \in \mathcal{F}$ is a function that maps an assignment of its neighboring variables to a real number. For example, in a linear-chain model of part-of-speech tagging, transition factors score compatibilities between consecutive labels, while emission factors score compatibilities between a label and its observed token.

The probability distribution expressed by the factor graph is given as a normalized product of the factors, which we rewrite as an exponentiated sum:

$$p(\mathbf{y}) = \frac{\exp \psi(\mathbf{y})}{Z} \tag{1}$$

$$\psi(\mathbf{y}) = \sum_{f \in \mathcal{F}} f(\mathbf{y}_f) \tag{2}$$

$$Z = \sum_{\mathbf{y} \in \mathcal{Y}} \exp \psi(\mathbf{y}) \tag{3}$$

Intuitively, the model favors assignments to the random variables that yield higher factor scores and will assign higher probabilities to such configurations.

The two common inference problems for graphical models in NLP are *maximum a posterior* (MAP) and marginal inference. For models without latent variables, the MAP estimate is the setting to the variables that has the highest probability under the model:

$$\mathbf{y}_{\text{MAP}} = \operatorname*{argmax}_{\mathbf{y}} p(\mathbf{y}) \tag{4}$$

Marginal inference is the problem of finding marginal distributions over subsets of the variables, used primarily in maximum likelihood gradients and for max marginal inference.

### 2.2 Markov chain Monte Carlo (MCMC)

Often, computing marginal estimates of a model is computationally intractable due to the normalization constant $Z$, while maximum a posteriori (MAP) is prohibitive due to the search space of possible configurations. Markov chain Monte Carlo (MCMC) is important tool for performing sample- and search-based inference in these models. A particularly successful MCMC method for graphical model inference is Metropolis-Hastings (MH). Since sampling from the true model $p(\mathbf{y})$ is intractable, MH instead uses a simpler distribution $q(\mathbf{y}'|\mathbf{y})$ that conditions on a current state $\mathbf{y}$ and proposes a new state $\mathbf{y}'$ by modifying a few variables. This new assignment is then accepted with probability $\alpha$:

$$\alpha = \min\left(1, \frac{p(\mathbf{y}')}{p(\mathbf{y})} \frac{q(\mathbf{y}|\mathbf{y}')}{q(\mathbf{y}'|\mathbf{y})}\right) \tag{5}$$

Computing this acceptance probability is often highly efficient because the partition function cancels, as do all the factors in the model that do not neighbor the modified variables. MH can be used for both MAP and marginal inference.

#### 2.2.1 Marginal Inference

To compute marginals with MH, the variables are initialized to an arbitrary assignment (i.e., randomly or with some heuristic), and sampling is run until the samples $\{\mathbf{y}_i | i = 0, \cdots, n\}$ become independent of the initial assignment. The ergodic theorem provides the MCMC analog to the law-of-large-numbers, justifying the use of the generated samples to compute the desired statistics (such as feature expectations or variable marginals).

### 2.2.2 MAP Inference

Since MCMC can efficiently explore the high density regions for a given distribution, the distribution $p$ can be modified such that the high-density region of the new distribution represents the MAP configuration of $p$. This is achieved by adding a temperature term $\tau$ to the distribution $p$, resulting in the following MH acceptance probability:

$$\alpha = \min\left(1, \left(\frac{p(\mathbf{y}')}{p(\mathbf{y})}\right)^{\frac{1}{\tau}}\right) \tag{6}$$

Note that as $\tau \rightarrow 0$, MH will sample closer to the MAP configuration. If a cooling schedule is implemented for $\tau$ then the MH sampler for MAP inference can be seen as an instance of simulated annealing (Bertsimas and Tsitsiklis, 1993).

## 3 Monte Carlo MCMC

In this section we introduce our approach for approximating the acceptance ratio of Metropolis-Hastings that samples the factors, and describe two sampling strategies.

### 3.1 Stochastic Proposal Evaluation

Although one of the benefits of MCMC lies in its ability to leverage the locality of the proposal, for some information extraction tasks this can become a crucial bottleneck. In particular, evaluation of each sample requires computing the score of all the factors that are *involved* in the change, i.e. all factors that neighbor any variable in the set that has changed. This evaluation becomes a bottleneck for tasks in which a large number of variables is involved in each proposal, or in which the model contains a number of high-degree variables, resulting in a large number of factors, or in which computing the factor score involves an expensive computation, such as string similarity between mention text.

Instead of evaluating the log-score $\psi$ of the model exactly, this paper proposes a Monte-Carlo estimation of the log-score. In particular, if the set of factors for a given proposal $\mathbf{y} \rightarrow \mathbf{y}'$ is $\mathcal{F}(\mathbf{y}, \mathbf{y}')$, we use a sampled subset of the factors $\mathcal{S} \subseteq \mathcal{F}(\mathbf{y}, \mathbf{y}')$ as an approximation of the model score. In the following we use $\mathcal{F}$ as an abbreviation for $\mathcal{F}(\mathbf{y}, \mathbf{y}')$. Formally,

$$\begin{aligned} \psi(\mathbf{y}) &= \sum_{f \in \mathcal{F}} f(\mathbf{y}_f) = |\mathcal{F}| \cdot \mathbb{E}_{\mathcal{F}}\left[f(\mathbf{y}_f)\right] \\ \psi_{\mathcal{S}}(\mathbf{y}) &= |\mathcal{F}| \cdot \mathbb{E}_{\mathcal{S}}\left[f(\mathbf{y}_f)\right] \end{aligned} \tag{7}$$

We use the sample log-score ($\psi_{\mathcal{S}}$) in the acceptance probability $\alpha$ to evaluate the samples. Since we are using a stochastic approximation to the model score, in general we need to take more MCMC samples before we converge, however, since evaluating each sample will be *much* faster ($O(|\mathcal{S}|)$ as opposed to $O(|\mathcal{F}|)$), we expect overall sampling to be faster.

In the next sections we describe several alternative strategies for sampling the set of factors $\mathcal{S}$. The primary restriction on the set of samples $\mathcal{S}$ is that their mean should be an unbiased estimator of $\mathbb{E}_{\mathcal{F}}[f]$. Further, time taken to obtain the set of samples should be negligible when compared to scoring all the factors in $\mathcal{F}$. Note that there is an implicit minimum of 1 to the number of the sampled factors.

### 3.2 Uniform Sampling

The most direct approach for subsampling the set of $\mathcal{F}$ is to perform uniform sampling. In particular, given a proportion parameter $0 < p \leq 1$, we select a random subset $\mathcal{S}_p \subseteq \mathcal{F}$ such that $|\mathcal{S}_p| = p \cdot |\mathcal{F}|$. Since this approach is agnostic as to the actual factors scores, $\mathbb{E}_{\mathcal{S}}[f] \equiv \mathbb{E}_{\mathcal{F}}[f]$. A low $p$ leads to fast evaluation, however it may require a large number of samples due to the substantial approximation. On the other hand, although a higher $p$ will converge with fewer samples, evaluating each sample is slower.

### 3.3 Confidence-Based Sampling

Selecting the best value for $p$ is difficult, requiring analysis of the graph structure, and statistics on the distribution of the factors scores; often a difficult task in real-world applications. Further, the same value for $p$ can result in different levels of approximation for different proposals, either unnecessarily accurate or problematically noisy. We would prefer a strategy that adapts to the distribution of the scores in $\mathcal{F}$.

Instead of sampling a fixed proportion of factors, we can sample until we are confident that the current set of samples $\mathcal{S}_c$ is an accurate estimate of the true mean of $\mathcal{F}$. In particular, we maintain a running count of the sample mean $\mathbb{E}_{\mathcal{S}_c}[f]$ and variance

$\sigma_{\mathcal{S}_c}$, using them to compute a confidence interval $I_{\mathcal{S}}$ around our estimate of the mean. Since the number of sampled factors $\mathcal{S}$ could be a substantial fraction of the set of factors $\mathcal{F}$,[1] we also incorporate *finite population control (fpc)* in our sample variance computation. We compute the confidence interval as follows:

$$\sigma^2_{\mathcal{S}} = \frac{1}{|\mathcal{S}|-1}\sum_{f\in\mathcal{S}}\left(f - \mathbb{E}_S\left[f\right]\right)^2 \tag{8}$$

$$I_{\mathcal{S}} = 2z\frac{\sigma_{\mathcal{S}}}{\sqrt{|\mathcal{S}|}}\sqrt{\frac{|\mathcal{F}|-|\mathcal{S}|}{|\mathcal{F}|-1}} \tag{9}$$

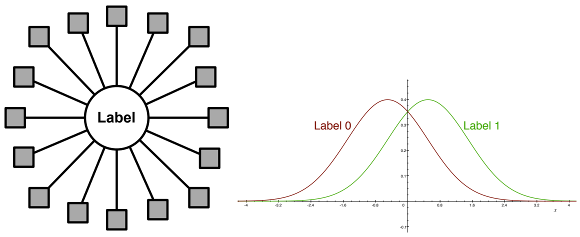
where we set the $z$ to $1.96$, i.e. the $95\%$ confidence interval. This approach starts with an empty set of samples, $\mathcal{S} = \{\}$, and iteratively samples factors without replacement to add to $\mathcal{S}$, until the confidence interval around the estimated mean falls below a user specified maximum interval width threshold $i$. As a result, for proposals that contain high-variance factors, this strategy examines a large number of factors, while proposals that involve similar factors will result in fewer samples. Note that this user-specified threshold is agnostic to the graph structure and the number of factors, and instead directly reflects the score distribution of the relevant factors.

## 4 Experiments

In this section we evaluate our approach for both marginal and MAP inference.

### 4.1 Marginal Inference on Synthetic Data

Consider the task of classifying entities into a set of types, for example, POLITICIAN, VEHICLE, CITY, GOVERMENT-ORG, etc. For knowledge base construction, this prediction often takes place on the entity-level, as opposed to the mention-level common in traditional NLP. To evaluate the type at the entity-level, the scored factors examine features of all the entity mentions of the entity, along with the labels of all relation mentions for which it is an argument. See Yao et al. (2010) and Hoffmann et al. (2011) for examples of such models. Since a subset of the mentions can be sufficiently informative for the model, we expect our stochastic MCMC approach to work well.



(a) Binary Classification Model ($n = 100$) (b) Distribution of Factor scores

Figure 1: Synthetic Model for Classification

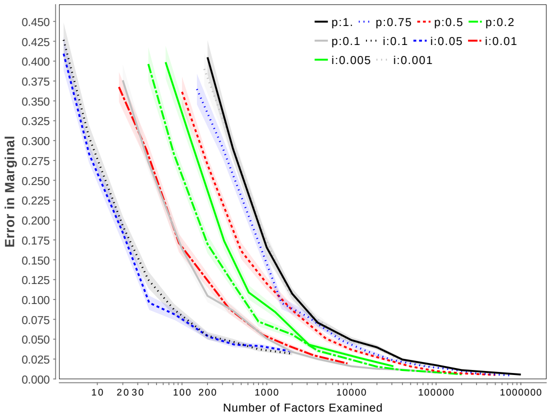


Figure 2: Marginal Inference Error for Classification on Synthetic Data

We use synthetic data for such a model to evaluate the quality of marginals returned by the Gibbs sampling form of MCMC. Since the Gibbs algorithm samples each variable using a fixed assignment of its neighborhood, we represent generating a single sample as classification. We create star-shaped models with a single unobserved variable (entity type) that neighbors many unary factors, each representing a single entity- or a relation-mention factor (See Figure 1a for an example). We generate a synthetic dataset for this model, creating 100 variables consisting of 100 factors each. The scores of the factors are generated from gaussians, $\mathcal{N}(0.5, 1)$ for the positive label, and $\mathcal{N}(-0.5, 1)$ for the negative label (note the overlap between the weights in Figure 1b). Although each structure contains only a single variable, and no cycles, it is a valid benchmark to test our sampling approach since the effects of the setting of burn-in period and the thinning samples are not a concern.

We perform standard Gibbs sampling, and com-

[1] Specifically, the fraction may be higher than $> 5\%$

pare the marginals obtained during sampling with the true marginals, computed exactly. We evaluate the previously described uniform sampling and confidence-based sampling, with several parameter values, and plot the $L_1$ error to the true marginals as more factors are examined. Note that here, and in the rest of the evaluation, we shall use the number of factors scored as a proxy for running time, since the effects of the rest of the steps of sampling are relatively negligible. The error in comparison to regular MCMC ($p = 1$) is shown in Figure 2, with standard error bars averaging over 100 models. Initially, as the sampling approach is made more stochastic (lowering $p$ or increasing $i$), we see a steady improvement in the running time needed to obtain the same error tolerance. However, the amount of relative improvements slows as stochasticity is increased further; in fact for extreme values ($i = 0.05, p = 0.1$) the chains perform worse than regular MCMC.

### 4.2 Entity Resolution in Citation Data

To evaluate our approach on a real world dataset, we apply stochastic MCMC for MAP inference on the task of citation matching. Given a large number of citations (that appear at the end of research papers, for example), the task is to group together the citations that refer to the same paper. The citation matching problem is an instance of entity resolution, in which observed mentions need to be partitioned such that mentions in a set refer to the same underlying entity. Note that neither the identities, or the number of underlying entities is known.

In this paper, the graphical model of entity resolution consists of observed mentions ($m_i$), and pairwise binary variables between all pairs of mentions ($y_{ij}$) which represent whether the corresponding observed mentions are coreferent. There is a local factor for each coreference variable $y_{ij}$ that has a high score if the underlying mentions $m_i$ and $m_j$ are similar. For the sake of efficiency, we only instantiate and incorporate the variables and factors when the variable is true, i.e. if $y_{ij} = 1$. Thus, $\psi(\mathbf{y}) = \sum_e \sum_{m_i, m_j \in e} f(y_{ij})$. The set of possible worlds consists of all settings of the $\mathbf{y}$ variables that are consistent with transitivity, i.e. the binary variables directly represent a valid clustering over the mentions. An example of the model defined over 5 mentions is given in Figure 3. This representation is equivalent to Model 2 as introduced in McCallum and Wellner (2004). As opposed to belief propagation and other approximate inference techniques, MCMC is especially appropriate for the task as it can directly enforce transitivity.

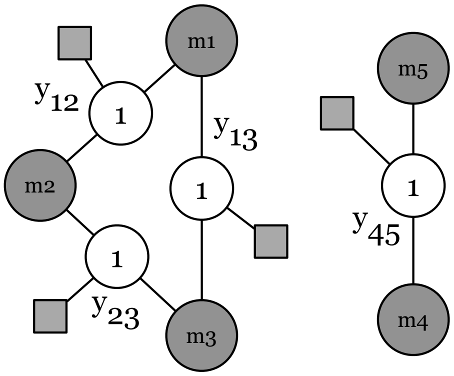


Figure 3: **Graphical Model for Entity Resolution:** defined over 5 mentions, with the setting of the variables resulting in 2 entities. For the sake of brevity, we've only included variables set to 1; binary variables between mentions that are not coreferent have been omitted.

When performing MCMC, each sample is a setting to all the $\mathbf{y}$ variables that is consistent with transitivity. To maintain transitivity during sampling, Metropolis Hastings is used to change the binary variables in a way that is consistent with moving individual mentions. Our proposal function selects a random mention, and moves it to a random entity, changing all the pairwise variables with mentions in its old entity, and the pairwise variables with mentions in its new entity. Thus, evaluation of such a proposal function requires scoring a number of factors linear in the size of the entities, which, for large datasets, can be a significant bottleneck. In practice, however, these set of factors are often highly redundant, as many of the mentions that refer to the same entity contain redundant information and features, and entity membership may be efficiently determined by observing a subset of its mentions.

We evaluate on the Cora dataset (McCallum et al., 1999), used previously to evaluate a number of information extraction approaches (Pasula et al., 2003), including MCMC based inference (Poon and Domingos, 2007; Singh et al., 2009). The dataset

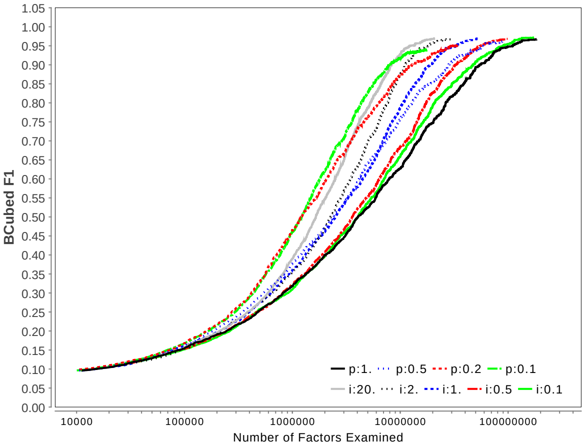


Figure 4: Citation Resolution Accuracy Plot for uniform and variance-based sampling compared to regular MCMC ($p = 1$)

consists of 1295 mentions, that refer to 134 true underlying entities. We use the same features for our model as (Poon and Domingos, 2007), using true *author*, *title*, and *venue* segmentation for features. Since our focus is on evaluating scalability of inference, we combine all the three folds of the data, and train the model using Samplerank (Wick et al., 2011).

We run MCMC on the entity resolution model using the proposal function described above, running our approach with different parameter values. Since we are interested in the MAP configuration, we use a temperature term for annealing. As inference progresses, we compute $BCubed$[2] F1 of the current sample, and plot it against the number of scored factors in Figure 4. We observe consistent speed improvements as stochasticity is improved, with uniform sampling and confidence-based sampling performing competitively. To compute the speedup, we measure the number of factors scored to obtain a desired level of accuracy (90% F1), shown for a diverse set of parameters in Table 1. With a very large confidence interval threshold ($i = 20$) and small proportion ($p = 0.1$), we obtain up to 7 times speedup over regular MCMC. Since the average entity size in this data set is $< 10$, using a small proportion (and a wide interval) is equivalent to picking a single mention to compare against.

| **Method** | **Factors Examined** | **Speedup** |
|---|---|---|
| Baseline | 57,292,700 | 1x |
| Uniform Sampling | | |
| $p = 0.75$ | 34,803,972 | 1.64x |
| $p = 0.5$ | 28,143,323 | 2.04x |
| $p = 0.3$ | 17,778,891 | 3.22x |
| $p = 0.2$ | 12,892,079 | 4.44x |
| $p = 0.1$ | 7,855,686 | 7.29x |
| Variance-Based Sampling | | |
| $i = 0.001$ | 52,522,728 | 1.09x |
| $i = 0.01$ | 51,547,000 | 1.11x |
| $i = 0.1$ | 47,165,038 | 1.21x |
| $i = 0.5$ | 32,828,823 | 1.74x |
| $i = 1$ | 18,938,791 | 3.02x |
| $i = 2$ | 11,134,267 | 5.14x |
| $i = 5$ | 9,827,498 | 5.83x |
| $i = 10$ | 8,675,833 | 6.60x |
| $i = 20$ | 8,295,587 | 6.90x |

Table 1: Speedups on Cora to obtain 90% $B^3$ F1

### 4.3 Large-Scale Author Coreference

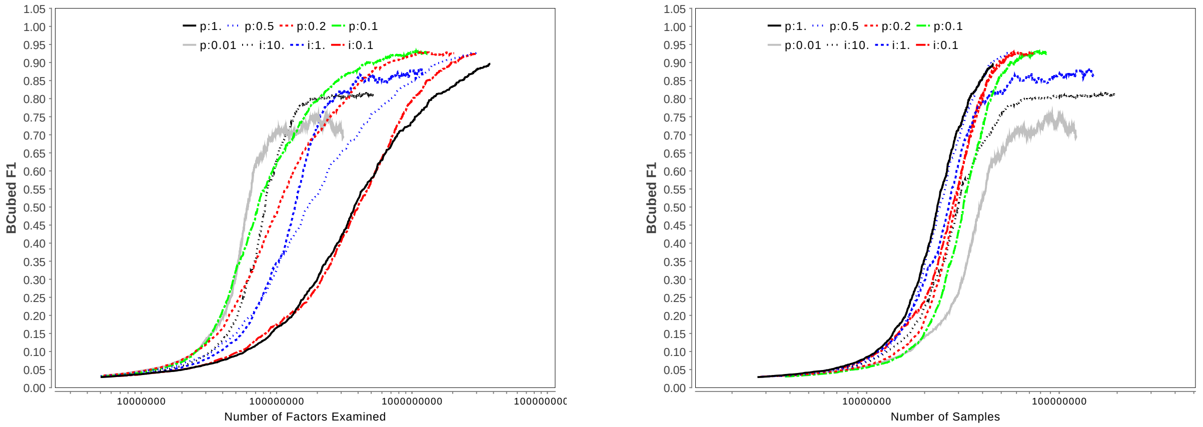
As the body of published scientific work continues to grow, author coreference, the problem of clustering mentions of research paper authors into the real-world authors to which they refer, is becoming an increasingly important step for performing meaningful bibliometric analysis. However, scaling typical pairwise models of coreference (e.g., McCallum and Wellner (2004)) is difficult because the number of factors in the model grows quadratically with the number of mentions (research papers) and the number of factors evaluated for every MCMC proposal scales linearly in the size of the clusters. For author coreference, the number of author mentions and the number of references to an author entity can often be in the millions, making the evaluation of the MCMC proposals computationally expensive.

We use the publicly available DBLP dataset[3] of BibTex entries as our unlabeled set of mentions, which contains nearly 5 million authors. For evaluation of accuracy, we also include author mentions from the Rexa corpus[4] that contains 2, 833 mentions

[2] $B^3$ is a coreference evaluation metric, introduced by Bagga and Baldwin (1998)

[3] http://www.informatik.uni-trier.de/~ley/db/

[4] http://www2.selu.edu/Academics/Faculty/aculotta/data/rexa.html



(a) Accuracy versus Number of Factors scored

(b) Accuracy versus Number of Samples

Figure 5: Performance of Different Sampling Strategies and Parameters for coreference over 5 million mentions. Plot with $p$ refer to uniform sampling with proportion $p$ of factors picked, while plots with $i$ sample till confidence intervals are narrower than $i$.

labeled for coreference.

We use the same Metropolis-Hastings scheme that we employ in the problem of citation matching. As before, we initialize to the singleton configuration and run the experiments for a fixed number of samples, plotting accuracy versus the number of factors evaluated (Figure 5a) as well as accuracy versus the number of samples generated (Figure 5b). We also tabulate the relative speedups to obtain the desired accuracy level in Table 2. Our proposed method achieves substantial savings on this task: speedups of 13.16 using the variance sampler and speedups of 9.78 using the uniform sampler. As expected, when we compare the performance using the number of generated samples, the approximate MCMC chains appear to converge more slowly; however, the overall convergence for our approach is substantially faster because evaluation of each sample is significantly cheaper. We also present results on using extreme approximations (for example, $p = 0.01$), resulting in convergence to a low accuracy.

| **Method** | **Factors Examined** | **Speedup** |
|---|---|---|
| Baseline | 1,395,330,603 | 1x |
| Uniform | | |
| $p = 0.5$ | 689,254,134 | 2.02x |
| $p = 0.2$ | 327,616,794 | 4.26x |
| $p = 0.1$ | 206,157,705 | 6.77x |
| $p = 0.05$ | 152,069,987 | 9.17x |
| $p = 0.02$ | 142,689,770 | 9.78x |
| Variance | | |
| $i = 0.00001$ | 1,442,091,344 | 0.96x |
| $i = 0.0001$ | 1,419,110,724 | 0.98x |
| $i = 0.001$ | 1,374,667,077 | 1.01x |
| $i = 0.1$ | 1,012,321,830 | 1.38x |
| $i = 1$ | 265,327,983 | 5.26x |
| $i = 10$ | 179,701,896 | 7.76x |
| $i = 100$ | 106,850,725 | 13.16x |

Table 2: Speedups on DBLP to reach 80% $\mathrm{B}^3$ F1

## 5 Discussion and Related Work

MCMC is a popular method for inference amongst researchers that work with large and dense graphical models (Richardson and Domingos, 2006; Poon and Domingos, 2006; Poon et al., 2008; Singh et al., 2009; Wick et al., 2009). Some of the probabilistic programming packages popular amongst NLP practitioners also rely on MCMC for inference and learning (Richardson and Domingos, 2006; McCallum et al., 2009). Although most of these methods apply MCMC directly, the rate of convergence of MCMC has become a concern as larger and more densely-factored models are being considered, motivating the need for more efficient sampling that uses parallelism (Singh et al., 2011; Gonzalez et al., 2011) and domain knowledge for blocking (Singh et al., 2010). Thus we feel providing a method to speed up MCMC inference can have a significant impact.

There has also been recent work in designing scalable approximate inference techniques. Belief propagation has, in particular, has gained some recent interest. Similar to our approach, a number of researchers propose modifications to BP that perform inference without visiting all the factors. Recent work introduces dynamic schedules to prioritize amongst the factors (Coughlan and Shen, 2007; Sutton and McCallum, 2007) that has been used to only visit a small fraction of the factors (Riedel and Smith, 2010). Gonzalez et al. (2009) utilize these schedules to facilitate parallelization.

A number of existing approaches in statistics are also related to our contribution. Leskovec and Faloutsos (2006) propose techniques to sample a graph to compute certain graph statistics with associated confidence. Christen and Fox (2005) also propose an approach to efficiently evaluate a proposal, however, once accepted, they score all the factors. Murray and Ghahramani (2004) propose an approximate MCMC technique for Bayesian models that estimates the partition function instead of computing it exactly.

Related work has also applied such ideas for robust learning, for example Kok and Domingos (2005), based on earlier work by Hulten and Domingos (2002), uniformly sample the groundings of an MLN to estimate the likelihood.

## 6 Conclusions and Future Work

Motivated by the need for an efficient inference technique that can scale to large, densely-factored models, this paper considers a simple extension to the Markov chain Monto Carlo algorithm. By observing that many graphical models contain substantial redundancy among the factors, we propose *stochastic* evaluation of proposals that subsamples the factors to be scored. Using two proposed sampling strategies, we demonstrate improved convergence for marginal inference on synthetic data. Further, we evaluate our approach on two real-world entity resolution datasets, obtaining a 13 times speedup on a dataset containing 5 million mentions.

Based on the ideas presented in the paper, we will consider additional sampling strategies. In particular, we will explore *dynamic* sampling, in which we sample fewer factors during the initial, burn-in phase, but sample more factors as we get close to convergence. Motivated by our positive results, we will also study the application of this approach to other approximate inference techniques, such as belief propagation and variational inference. Since training is often a huge bottleneck for information extraction, we will also explore its applications to parameter estimation.

## Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by ARFL under prime contract number is FA8650-10-C-7059, and the University of Massachusetts gratefully acknowledges the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. The U.S. Government is authorized to reproduce and distribute reprint for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

## References

[Bagga and Baldwin1998] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *International Conference on Language Resources and Evaluation (LREC) Workshop on Linguistics Coreference*, pages 563–566.

[Bertsimas and Tsitsiklis1993] D. Bertsimas and J. Tsitsiklis. 1993. Simulated annealing. *Statistical Science*, pages 10–15.

[Carreras2007] Xavier Carreras. 2007. Experiments with a higher-order projective dependency parser. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 957–961.

[Christen and Fox2005] J. Andrés Christen and Colin Fox. 2005. Markov chain monte carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):pp. 795–810.

[Coughlan and Shen2007] James Coughlan and Huiying Shen. 2007. Dynamic quantization for belief propagation in sparse spaces. *Computer Vision and Image Understanding*, 106:47–58, April.

[Culotta et al.2007] Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*.

[Gonzalez et al.2009] Joseph Gonzalez, Yucheng Low, and Carlos Guestrin. 2009. Residual splash for optimally parallelizing belief propagation. In *Artificial Intelligence and Statistics (AISTATS)*.

[Gonzalez et al.2011] Joseph Gonzalez, Yucheng Low, Arthur Gretton, and Carlos Guestrin. 2011. Parallel gibbs sampling: From colored fields to thin junction trees. In *Artificial Intelligence and Statistics (AISTATS)*, Ft. Lauderdale, FL, May.

[Hoffmann et al.2011] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550, Portland, Oregon, USA, June. Association for Computational Linguistics.

[Hulten and Domingos2002] Geoff Hulten and Pedro Domingos. 2002. Mining complex models from arbitrarily large databases in constant time. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 525–531, New York, NY, USA. ACM.

[Kok and Domingos2005] Stanley Kok and Pedro Domingos. 2005. Learning the structure of markov logic networks. In *International Conference on Machine Learning (ICML)*, pages 441–448, New York, NY, USA. ACM.

[Kschischang et al.2001] Frank R. Kschischang, Brendan J. Frey, and Hans Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions of Information Theory*, 47(2):498–519, Feb.

[Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

[Leskovec and Faloutsos2006] Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 631–636, New York, NY, USA. ACM.

[McCallum and Wellner2004] Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Neural Information Processing Systems (NIPS)*.

[McCallum et al.1999] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 1999. A machine learning approach to building domain-specific search engines. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[McCallum et al.2009] Andrew McCallum, Karl Schultz, and Sameer Singh. 2009. FACTORIE: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

[Murray and Ghahramani2004] Iain Murray and Zoubin Ghahramani. 2004. Bayesian learning in undirected graphical models: Approximate MCMC algorithms. In *Uncertainty in Artificial Intelligence (UAI)*.

[Pasula et al.2003] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. 2003. Identity uncertainty and citation matching. In *Neural Information Processing Systems (NIPS)*.

[Poon and Domingos2006] Hoifung Poon and Pedro Domingos. 2006. Sound and efficient inference with probabilistic and deterministic dependencies. In *AAAI Conference on Artificial Intelligence*.

[Poon and Domingos2007] Hoifung Poon and Pedro Domingos. 2007. Joint inference in information extraction. In *AAAI Conference on Artificial Intelligence*, pages 913–918.

[Poon et al.2008] Hoifung Poon, Pedro Domingos, and Marc Sumner. 2008. A general method for reducing the complexity of relational inference and its application to MCMC. In *AAAI Conference on Artificial Intelligence*.

[Richardson and Domingos2006] Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136.

[Riedel and Smith2010] Sebastian Riedel and David A. Smith. 2010. Relaxed marginal inference and its application to dependency parsing. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, pages 760–768.

[Singh et al.2009] Sameer Singh, Karl Schultz, and Andrew McCallum. 2009. Bi-directional joint inference for entity resolution and segmentation using imperatively-defined factor graphs. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science) and European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 414–429.

[Singh et al.2010] Sameer Singh, Michael L. Wick, and Andrew McCallum. 2010. Distantly labeling data for large scale cross-document coreference. *Computing Research Repository (CoRR)*, abs/1005.4298.

[Singh et al.2011] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011.

Large-scale cross-document coreference using distributed inference and hierarchical models. In *Association for Computational Linguistics: Human Language Technologies (ACL HLT)*.

[Sutton and McCallum2004] Charles Sutton and Andrew McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. Technical Report TR#04-49, University of Massachusetts, July.

[Sutton and McCallum2007] Charles Sutton and Andrew McCallum. 2007. Improved dynamic schedules for belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*.

[Wick et al.2009] Michael Wick, Aron Culotta, Khashayar Rohanimanesh, and Andrew McCallum. 2009. An entity-based model for coreference resolution. In *SIAM International Conference on Data Mining (SDM)*.

[Wick et al.2010] Michael Wick, Andrew McCallum, and Gerome Miklau. 2010. Scalable probabilistic databases with factor graphs and mcmc. *International Conference on Very Large Databases (VLDB)*, 3:794–804, September.

[Wick et al.2011] Michael Wick, Khashayar Rohanimanesh, Kedar Bellare, Aron Culotta, and Andrew McCallum. 2011. Samplerank: Training factor graphs with atomic gradients. In *International Conference on Machine Learning (ICML)*.

[Yao et al.2010] Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Empirical Methods in Natural Language Processing (EMNLP)*.