Rapid Mixing Swendsen-Wang Sampler for Stochastic Partitioned Attractive Models

Sejun Park* **Yunhun Jang*** **Andreas Galanis†** **Jinwoo Shin*** **Daniel Štefankovič‡** **Eric Vigoda§**

*Korea Advanced Institute of Science and Technology
†University of Oxford
‡University of Rochester
§Georgia Institute of Technology

Abstract

The Gibbs sampler is the most popular Markov chain used for learning and inference problems in Graphical Models (GM). These tasks are computationally intractable in general, and the Gibbs sampler often suffers from slow mixing. In this paper, we study the Swendsen-Wang dynamics which is a more sophisticated Markov chain designed to overcome bottlenecks that impede Gibbs sampler. We prove $O(\log n)$ mixing time for attractive binary pairwise GMs (i.e., ferromagnetic Ising models) on stochastic partitioned graphs having n vertices, under some mild conditions including low temperature regions where the Gibbs sampler provably mixes exponentially slow. Our experiments also confirm that the Swendsen-Wang sampler significantly outperforms the Gibbs sampler for learning parameters of attractive GMs.

1 INTRODUCTION

Graphical models (GM) express a factorization of joint multivariate probability distributions in statistics via a graph of relations between variables. GM's have been used successfully in information theory, physics, artificial intelligence and machine learning [37, 12, 22, 21, 1, 9]. For typical learning and inference problems using GM's, marginalizing the joint distribution, or equivalently computing the partition function (normalization factor), is the key computational bottleneck; this sampling/counting problem is computationally intractable in general, more formally, it is NP-hard even to approximate the partition function [5, 38]. Nevertheless, Markov Chain Monte Carlo (MCMC) methods, typically using the Gibbs sampler, are widely-used in learning and inference applications of GM, but they often suffer from slow mixing.

To address the potential slow mixing of the Gibbs sampler, there have been extensive efforts in the literature to establish fast mixing regimes of the Gibbs sampler (also known as the Glauber dynamics). Most of these theoretical works have studied under perspectives of the Ising model and its variants [31, 24, 6]. Given a graph $G = (V, E)$ having n vertices and parameters $\beta = [\beta_{uv} : (u, v) \in E] \in \mathbb{R}^{|E|}, \gamma = [\gamma_v : v \in V] \in \mathbb{R}^n$, the Ising model is a joint probability distribution on all spin configurations $\Omega = \{\sigma : \sigma = [\sigma_v] \in \{-1, 1\}^n\}$ such that

$$\mu(\sigma) \propto \exp\left(\sum_{(u,v)\in E} \beta_{uv}\sigma_u\sigma_v + \sum_{v\in V} \gamma_v\sigma_v\right) \qquad (1)$$

The parameter γ corresponds to the presence of an "external (magnetic) field", and when $\gamma_v = 0$ for all $v \in V$, we say the model has no (or zero) external field. If $\beta_{uv} \geq 0$ for all $(u, v) \in E$ the model is called *ferromagnetic/attractive*, and *anti-ferromagnetic/repulsive* if $\beta_{uv} \leq 0$ for all $(u, v) \in E$. It is naturally expected that the Gibbs sampler mixes slow if interaction strengths of GM are high, i.e., β is large which corresponds to low temperature regimes. For example, for the ferromagnetic Ising models on the complete graph G (which is commonly referred to as the mean-field model) it is known that the mixing-time in the high temperature regime ($\beta < 1$) has $O(n \log n)$, whereas the mixing-time in the low temperature regime ($\beta > 1$) is exponential in n [24].

This paper focuses on ferromagnetic Ising models (FIM), where any pairwise binary attractive GM can be expressed by FIM. We study the Swendsen-Wang

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

dynamics[1] which is a more sophisticated Markov chain designed to overcome bottlenecks that impede Gibbs sampler. Pairwise binary attractive GMs, equivalently FIMs, have gained much attentions in the GM literature because they do not contain frustrated cycles and have several advantages to design good algorithms for approximating the partition function [20, 44, 45, 34, 30, 39]. Furthermore, they have been used for various machine learning applications. For example, the non-negative Boltzmann machine (NNBM) has been used to describe multimodal non-negative data [7]. Non-negative restricted Boltzmann machine (RBM), equivalent to FIM on complete bipartite graphs, has also been studied in the context of unsupervised deep learning models [33], where non-negativity (i.e., ferromagneticity) provides non-negative matrix factorization [23] like interpretable features, which is especially useful for analyzing medical data [41, 26] and document data [33]. FIM is also a popular model for studying strategic diffusion in social networks [35, 29], where in the case β_{uv} represents a friendship or other positive relationships between two individuals u, v.

Motivated by the recent studies on FIM, we prove $O(\log n)$ mixing time of the Swendsen-Wang sampler for FIM on stochastic partitioned graphs[2], which include complete bipartite graphs and social network models (e.g., stochastic block models [19]) as special cases. In particular, we show that the Swendsen-Wang chain mixes fast in low temperature regions where the Gibbs sampler provably mixes exponentially slow. Our experimental results also confirm that the Swendsen-Wang sampler significantly outperforms the Gibbs sampler for learning parameters of attractive GMs. We remark that it has been recently shown that an arbitrary binary pairwise GM can be approximated by FIM of a certain partitioned structure. In conjunction with this, we believe that our results potentially extend to a certain class of non-attractive GMs as well (see Section 6).

Related work. There has been considerable effort on analyzing the mixing times of the Swendsen-Wang and Gibbs samplers for the ferromagnetic Ising model. All of the below theoretical works consider 'uniform' parameters on edges, i.e., all β_{uv}'s are equal, and zero external field, i.e., $\gamma_v = 0$. There are several works showing examples where the Swendsen-Wang dynamics has exponentially slow mixing time [16, 4, 2, 3, 11] for the Potts model which is the generalization of the Ising model to more than two spins; all of these slow mixing results are at the critical point for the associated phase transition. It was very recently shown that the Swendsen-Wang dynamics is rapidly mixing on every graph and at every (positive) temperature [17]; the mixing time is a large polynomial, e.g., $O(n^{10})$ for the complete bipartite graphs, so this general result does not give bounds which are useful in practice. However, the appeal for utilizing this dynamics is that its mixing time is conjectured to be much smaller, as we prove $O(\log n)$ for stochastic partitioned graphs.

For the mean-field model (i.e., the complete graph) a detailed analysis of the Swendsen-Wang dynamics was established by [27] who proved that the mixing time is $\Theta(1)$ for $\beta < \beta_c$, $O(n^{1/4})$ for $\beta = \beta_c$ and $O(\log n)$ for $\beta > \beta_c$ where β_c is the inverse critical temperature. For the two-dimensional lattice, [42] established polynomial mixing time of the Swendsen-Wang dynamics for all $\beta > 0$. On the other hand, the mixing time of the Gibbs sampler (also known as the Glauber dynamics or Metropolis-Hastings algorithm) for the complete graph is known to be $\Theta(n \log n)$ for $\beta < \beta_c$, $\Theta(n^{3/2})$ for $\beta = \beta_c$ and $e^{\Omega(n)}$ for $\beta > \beta_c$ [24]. For the Erdős-Rényi random graph $G(n, d/n)$, the mixing time of the Gibbs chain is $O(n^{1+\Theta(1/\log\log n)})$ for $d \tanh \beta < 1$ [32] and $e^{\Omega(n)}$ for $d \tanh \beta > 1$ [13] with high probability over the choice of the graph.

2 PRELIMINARIES

2.1 Swendsen-Wang Sampler

The Swendsen-Wang dynamics [40] is a Markov chain $\{X_t \in \Omega : t = 0, 1, 2, \dots\}$ having μ as its stationary (i.e., invariant) distribution. A step of the Swendsen-Wang dynamics works at a high-level as follows: (i) the current spin configuration X_t is converted into a configuration M in the random-cluster model [8] by taking the monochromatic edges, (ii) then we do a percolation step on M by each edge being deleted with some probability, and finally (iii) each component of the percolated subgraph chooses a random spin and this defines the new spin configuration X_{t+1}. Whereas the traditional Gibbs sampler modifies the spin at one vertex in a step, the Swendsen-Wang dynamics may change the spin at every vertex in a single step. Under the Ising model with no external field, its formal description of transitions from X_t to X_{t+1} is defined as follows:

1. Let M be the set of monochromatic edges in X_t, i.e., $M = \{(u, v) \in E : X_t(u) = X_t(v)\}$.
2. For each edge $(u, v) \in M$, delete it with probability $1 - p_{uv}$, where $p_{uv} = 1 - \exp(-2\beta_{uv})$. Let M' denote the set of monochromatic edges that were not deleted.
3. For each connected component C of the subgraph

[1] The Swendsen-Wang dynamics is formally defined in Section 2.1.

[2] See Section 3 for the formal definition of stochastic partitioned graphs

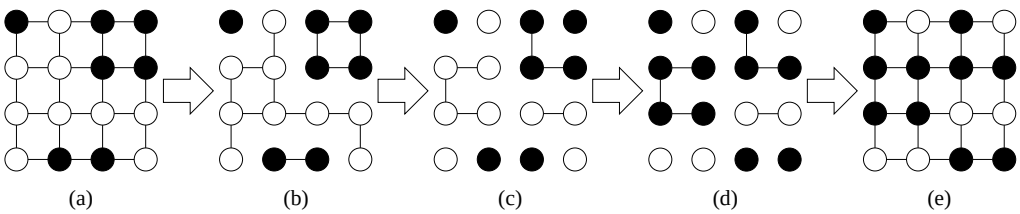


Figure 1: Illustration of a single iteration of the Swendsen-Wang dynamics. Each subfigure represents (a) an input X_t (b) a subgraph induced by the set of monochromatic edges M (c) a subgraph induced by the set of monochromatic edges M' after the step 2 (d) a configuration after the step 3 (e) an output X_{t+1} where black and white imply assignments $-1, +1$ respectively.

$G' = (V, M')$, independently, choose a spin $s \in \{-1, +1\}$ uniformly at random and assign spin s to all vertices in C. Let X_{t+1} denote the resulting spin configuration.

One can generalize the dynamics to a model having external fields by modifying step 3 as follows:

3. For each connected component C of the subgraph $G' = (V, M')$, set

$$s = \begin{cases} +1 & \text{with probability } \frac{\exp\left(2\sum_{v\in V(C)}\gamma_v\right)}{1+\exp\left(2\sum_{v\in V(C)}\gamma_v\right)} \\ -1 & \text{with probability } \frac{1}{1+\exp\left(2\sum_{v\in V(C)}\gamma_v\right)} \end{cases}.$$

Then, assign all vertices in C the chosen spin s and let X_{t+1} denote the resulting spin configuration.

Figure 1 visualizes each step of the Swendsen-Wang dynamics. For completeness, in the supplementary material, we prove the following well-known fact that the above Swendsen-Wang dynamics has μ as the stationary distribution.[3]

Lemma 1 *The stationary distribution of the Swendsen-Wang chain is* (1).

2.2 Mixing Time and Coupling

We use the following popular notion of 'mixing time': given an ergodic Markov chain $\{X_t \in \Omega : t = 0, 1, 2, \dots\}$ having the stationary distribution μ, we define its mixing time T_{mix} as

$$T_{\text{mix}} := \arg\min_t \left(\sup_{X_0\in\Omega, A\subset\Omega} |\Pr(X_t \in A) - \mu(A)| \le \frac{1}{4} \right).$$

A classical technique for bounding the mixing time is the 'coupling' technique [25]. Consider two copies (X_t, Y_t) of the same Markov chain (i.e., X_t, Y_t have the same transition probabilities) defined jointly with property that if $X_t = Y_t$ then $X_{t'} = Y_{t'}$ for all $t' \ge t$. We call such (X_t, Y_t) as a coupling, where X_t, Y_t might be dependent and there can be many ways to design such dependencies. Then, one can observe that

$$\begin{aligned} &\sup_{X_0\in\Omega, A\subset\Omega} |\Pr(X_t \in A) - \mu(A)| \\ &\le \sup_{X_0, Y_0\in\Omega, A\subset\Omega} |\Pr(X_t \in A) - \Pr(Y_t \in A)| \\ &\le \sup_{X_0, Y_0\in\Omega} \Pr(X_t \ne Y_t), \end{aligned}$$

which implies that

$$T_{\text{mix}} \le \arg\min_t \left(\sup_{X_0, Y_0\in\Omega} \Pr(X_t \ne Y_t) \le \frac{1}{4} \right). \quad (2)$$

We will design a coupling for obtaining a bound on mixing time of the Swendsen-Wang chain.

3 MAIN RESULTS

In this section, we state the main results of this paper that the Swendsen-Wang chain mixes fast for a class of stochastic partitioned graphs. To this end, we first formally define the notion of stochastic partitioned graphs. Given a positive integer $r \in \mathbb{Z}_+$, a vector $[\alpha_i] \in (0,1)^r$ with $\sum_i \alpha_i = 1$ and a matrix $[p_{ij}] \in [0,1]^{r\times r}$, a stochastic partitioned graph $(V, E) = G(n, [\alpha_i], [p_{ij}])$ on n vertices and r partitions (or communities) of size $\alpha_1 n, \dots, \alpha_r n$ is a random graph model such that

$$V = \bigcup_i V_i, \quad |V_i| = \alpha_i n \quad \text{and} \quad V_i \cap V_j = \emptyset, \text{ for } i \ne j.$$

An edge between any pair of vertices $u \in V_i, v \in V_j$ exists with probability p_{ij} independently. For example, if $p_{ii} = 0$ for all i and $p_{ij} = 1$ for all $i \ne j$, then the stochastic partitioned graph is the complete r-partite graph. One can also check that the stochastic block model [19] is a special case of the stochastic partitioned graph. In particular, if $r = 1$, $p_{11} = p$

[3] We did not find a version of the Swendsen-Wang dynamics for the Ising model with external fields in the literature, and hence present it with the formal proof for completeness.

for some $p \in [0,1]$, we say it is the Erdős-Rényi random graph and use the notation $G(n,p)$ to denote it. Similarly, if $r = 2$, $p_{11} = p_{22} = 0$ and $p_{12} = p_{21} = p$ for some $p \in [0,1]$, we say it is the bipartite Erdős-Rényi random graph and use the notation $G(n,m,p) = (V_L, V_R, E)$ to denote it, where n, m are sizes of partitions V_L, V_R. We say a graph (V,E) has size n if $|V| = n$ and bipartite graph (V_L, V_R, E) has size (n,m) if $|V_L| = n$ and $|V_R| = m$.

3.1 $O(\log n)$ Mixing in Low Temperatures

We first establish the following rapid mixing property of the Swendsen-Wang chain in low temperature regimes, i.e., $\beta_{uv} = \Omega(1)$. These are in particular most interesting regimes since the Gibbs chain (provably) mixes slower as β_{uv} grows. Moreover, these are also reasonable in practical applications, e.g., in social networks, β_{uv} represents a positive interaction strength between two individuals u, v and it is independently of the network size n.

Theorem 2 *The mixing time T_{mix} of Swendsen-Wang chain on graph $G(n, [\alpha_i], [p_{ij}])$ is*

$$T_{mix} = O(\log n)$$

almost surely if

- $\alpha_i = \Omega(1)$ *for all* i
- $\gamma_v \geq 0$ *(or* $\gamma_v \leq 0$*) for all* $v \in V$

and either a) or b) holds

a) $p_{ii} = \Omega(1)$ *for all* i, $\beta_{uv} = \Omega(1)$ *for all* $u, v \in V_i$

b) $p_{ij} = \Omega(1)$ *and* $\beta_{uv} = \Omega(1)$ *for all* $u \in V_i, v \in V_j$ *with* $i \neq j$.

The proof of Theorem 2 is presented in Section 4.2, where we will show the existence of a good coupling of the Swendsen-Wang chain. Theorem 2 implies that the Swendsen-Wang chain mixes fast as long as positive parameters $[p_{ij}]$ and $[\beta_{uv}]$ are not 'too small' (e.g., $p_{ij} = \Omega(1)$ and $\beta_{uv} = \Omega(1)$) and all external fields $[\gamma_v]$ are positive or negative. We believe that the restriction on positive (or negative) external fields is inevitable since it is known that approximating the partition function of ferromagnetic Ising model under mixed external fields is known to be #P-hard [15]. Despite the worst-case theoretical barrier, the Swendsen-Wang chain still works well under mixed external fields in our experiments (see Section 5).

3.2 $O(\log n)$ Mixing in High Temperatures

The restriction on parameters $[\beta_{uv}]$ in Theorem 2 is merely for technical reasons in our proof techniques, and we believe that it is not necessary. This is because it is natural to expect that a Markov chain mixes faster for higher temperatures. To support the conjecture, as stated in the following theorem, we prove that the Swendsen-Wang chain mixes fast even for small parameters $[\beta_{uv}]$ under the complete bipartite graphs, where its proof is much harder than that of Theorem 2.

Theorem 3 *Given any constant $k > 0$, the mixing time T_{mix} of Swendsen-Wang chain on the complete bipartite graph (V_L, V_R, E) of size (n, kn) is*

$$T_{mix} = O(\log n)$$

if $\beta_{uv} = -\frac{1}{2}\log\left(1 - \frac{B}{n\sqrt{k}}\right)$ *for all* $(u,v) \in E$ *for some non-negative constant* $B \neq 2$ *and* $\gamma_v = 0$ *for all* $v \in V$.

In the above theorem, we consider the scenario $\beta_{uv} = o(1)$, i.e.,

$$\beta_{uv} = -\frac{1}{2}\log\left(1 - \frac{B}{n\sqrt{k}}\right) \approx \frac{B}{2n\sqrt{k}}.$$

The proof of Theorem 3 is presented in Section 4.3, where we will also show the existence of a good coupling of the Swendsen-Wang chain using a similar strategy to that in [10]. The authors of [10] establish the rapid mixing property of the Swendsen-Wang chain for the complete graphs by analyzing a one-dimensional function, the so-called simplified Swendsen-Wang (see Section B.1 in the supplementary material), and utilizing known properties of Erdős-Rényi random graphs. In the case of the complete bipartite graphs, the simplified Swendsen-Wang becomes a two-dimensional function, which makes harder to analyze. Furthermore, the proof of Theorem 3 requires properties of the bipartite Erdős-Rényi random graph $G(n,m,p)$ which are less studied compared to the popular 'non-bipartite' Erdős-Rényi random graph $G(n,p)$. In this paper, we also establish necessary properties of $G(n,m,p)$ for the proof of Theorem 3. We believe that the conclusion of Theorem 3 holds for general stochastic partitioned graphs. However, in this case, there exist technical challenges handling more randomness in graphs, and we do not explore further in this paper.

4 PROOFS OF THEOREMS

4.1 Notation

Before we start the proof of Theorem 2 and Theorem 3, we define some notations about configurations

of the Ising model on a stochastic partitioned graph. Given a spin configuration σ, denote $V_-(\sigma), V_+(\sigma)$ as sets of vertices of spin $-1, +1$, respectively. In particular, given the Ising model on a bipartite graph (V_L, V_R, E) with partitions of vertices V_L, V_R, edges $E \subset \{(u,v) : u \in V_L, v \in V_R\}$ and a spin configuration $\sigma \in \{-1,1\}^{|V_L \cup V_R|}$, we say the configuration σ has the 'phase' $\alpha(\sigma) = (\alpha_L, \alpha_R)$ if a larger spin class, say $s \in \{-,+\}$ with $V_s(\sigma) \geq (|V_L| + |V_R|)/2$, of σ satisfies

$$(\alpha_L, \alpha_R) = \left(\frac{V_s(\sigma) \cap V_L}{V_L}, \frac{V_s(\sigma) \cap V_R}{V_R} \right).$$

One can define the induced probability on the phase (α_L, α_R) under the Ising model as

$$\Pr(\alpha_L, \alpha_R) = \sum_{\sigma \,:\, \alpha(\sigma) = (\alpha_L, \alpha_R)} \mu(\sigma).$$

4.2 Proof of Theorem 2

In this section, we only present the proof of Theorem 2 for $\gamma_v \geq 0$ since the proof for the case $\gamma_v \leq 0$ is identical. We will first show that in $O(\log n)$ iterations of the Swendsen-Wang chain, all spins are same with probability $\Theta(1)$. Using this fact, we will bound the mixing time via the coupling technique. To this end, we introduce the following key lemmas on the (bipartite) Erdős-Rènyi random graph. Proofs of Lemma 4-7 are presented in the supplementary material.

Lemma 4 *If $p = \Omega(1)$, then every induced subgraphs of $G(n,p)$ of size cn, $c = \Omega(1)$, contain a component of size $\geq cn - O(1)$ with probability $1 - e^{-\Omega(n)}$.*

Lemma 5 *If $p = \Omega(1)$, then every induced subgraphs of size $n - O(\sqrt{n})$ of $G(n,p)$ is connected with probability $1 - e^{-\Omega(n)}$.*

Lemma 6 *If $k = \Theta(1)$ and $p = \Omega(1)$, then every induced subgraphs of $G(n, kn, p)$ of size $(c_L n, c_R kn)$, $c_L, c_R = \Omega(1)$, contain a component of size $\geq \big(c_L n - O(1), c_R kn - O(1)\big)$ with probability $1 - e^{-\Omega(n)}$.*

Lemma 7 *If $k = \Theta(1)$ and $p = \Omega(1)$, then every induced subgraphs of size $\big(n - O(\sqrt{n}), kn - O(\sqrt{n})\big)$ of $G(n, kn, p)$ is connected with probability $1 - e^{-\Omega(n)}$.*

Now, we provide the proof of Theorem 2 assuming the condition a) while the proof for the condition b) is almost identical. Consider the Swendsen-Wang chain $\{X_t : t = 0, 1, \dots\}$ on $G(n, [\alpha_i], [p_{ij}])$ under the condition a). One can easily observe that using the assumption $\gamma_v \geq 0$, after running a single iteration of the Swendsen-Wang chain from any initial state X_0, there exists a spin class $s \in \{-,+\}$ in X_1 such that

$$|V_s(X_1) \cap V_i| \geq \frac{\alpha_i}{2} n \tag{3}$$

for all $i = 1, 2, \dots, r$ with probability $\Theta(1)$. This is because we assume the number of partitions r and $\alpha_i = |V_i|/n$ are constants. Assume that the event (3) occurs at X_1. After the step 2 of the Swendsen-Wang dynamics starting from X_1, the resulting graph on vertices $V_s(X_1) \cap V_i$ follows the distribution

$$G\big(|V_s(X_1) \cap V_i|, [1], [p_{ii}(1 - \exp(-2\beta_{uv}))]\big).$$

Since $V_s(X_1)$ has a constant fraction of vertices of each partition, $p_{ii} = \Omega(1)$ for all i and $\beta_{uv} = \Omega(1)$ for all $u, v \in V_i$, Lemma 4 implies that $\geq 1 - O(n^{-1})$ fraction of $V_s(X_1)$ is still connected with probability $1 - e^{-\Omega(n)}$ after the step 2 of the Swendsen-Wang dynamics starting from X_1, i.e. almost all spins in $V_s(X_1)$ are identical at X_2. We call this connected $\geq 1 - O(n^{-1})$ fraction in Lemma 4 as a 'giant component'. Note that an edge between different partitions only increase the size of the giant component in a single partition.We define the event $\mathcal{E}_1$ that after the step 2 of the Swendsen-Wang dynamics, $1 - O(n^{-1})$ fraction of vertices of $V_s(X_i)$ are connected for $T_1 = \frac{1}{2}\log_2 n + 1$ iterations of the Swendsen-Wang chain, i.e. giant components exists for T_1 iterations. By Lemma 4, the event $\mathcal{E}_1$ occurs with probability $1 - e^{-\Omega(n)}$. Conditioning on the event $\mathcal{E}_1$, let N_t be the number of vertices having a different spin from a giant component after t-th iteration of the Swendsen-Wang chain. Since $\gamma_v \geq 0$ and each component receives the spin + with probability $\geq 1/2$ in the step 3 of the Swendsen-Wang dynamics, the expectation of N_{t+1} given $N_t = \Omega(\sqrt{n})$ and $\mathcal{E}_1$ becomes

$$E[N_{t+1} | N_t, \mathcal{E}_1] \leq \left(\frac{1}{2} + O(n^{-1/2}) \right) N_t \tag{4}$$

for $t \leq T_1$. By using (4), one can bound the expectation of N_{T_1} as

$$E[N_{T_1} | \mathcal{E}_1] \leq \frac{n}{2} \left(\frac{1}{2} + O(n^{-1/2}) \right)^{T_1 - 1}.$$

If we use the Markov inequality, it follows that

$$\Pr(N_{T_1} < \sqrt{n} | \mathcal{E}_1) \geq 1 - \sqrt{n} \left(\frac{1}{2} \right)^{T_1} = \Theta(1) \tag{5}$$

and $\Pr(N_{T_1} < \sqrt{n}) \geq \Pr(N_{T_1} < \sqrt{n} | \mathcal{E}_1) \Pr(\mathcal{E}_1) \geq \Theta(1)$. Given the event $N_{T_1} < \sqrt{n}$, by Lemma 5, the largest component of X_{T_1+1} remains to be connected with probability $1 - e^{-\Omega(n)}$. We define the event $\mathcal{E}_2$ that given $N_{T_1} < \sqrt{n}$, after the step 2 of the Swendsen-Wang dynamics, the largest component is still connected for $T_2 = \frac{1}{2}\log_2 n + 1$ iterations of the Swendsen-Wang chain starting from X_{T_1}. By Lemma 5, the event $\mathcal{E}_2$ occurs with probability $1 - e^{-\Omega(n)}$. Using the same technique that we used for (5), one can obtain the following inequality

$$\Pr(N_{T_1+T_2} = 0) \geq \Theta(1),$$

i.e. $X_{T_1+T_2}$ consists of a single spin with probability $\Theta(1)$. Since all events so far occur with probability $\Theta(1)$, for two independent copies $\{X_t : t = 0, 1, \dots\}$ and $\{Y_t : t = 0, 1, \dots\}$ of the Swendsen-Wang chain, all spins of $X_{T_1+T_2+1}$ and $Y_{T_1+T_2+1}$ are same, i.e. $X_{T_1+T_2+1} = Y_{T_1+T_2+1}$ with probability $\Theta(1)$. Thus, there exists $T = O(\log n)$ and a trivial coupling (X_t, Y_t) such that $\Pr(X_T \neq Y_T) \leq 1/4$. This completes the proof of Theorem 2.

4.3 Proof of Theorem 3

In this section, we present the proof of Theorem 3. We provide the proof outlines for the cases $B > 2$ and $B < 2$, where proofs of key lemmas are given in the supplementary material. We first define

$$(\alpha_L^*, \alpha_R^*) := \lim_{n\to\infty} \arg \max_{(\alpha_L, \alpha_R)} \Pr(\alpha_L, \alpha_R),$$

where such (α_L^*, α_R^*) uniquely exists as we stated in Lemma 17 in the supplementary material.

Rapid mixing proof for $B > 2$. In this case, we will show that the Swendsen-Wang chain moves within the constantly small distance from (α_L^*, α_R^*) for any starting state in $O(1)$ iterations with probability $\Theta(1)$. Then, we will show that the Swendsen-Wang chain moves within $O(n^{-1/2})$ distance from (α_L^*, α_R^*) in $O(\log n)$ iterations with probability $\Theta(1)$. Finally, using this fact, we will bound the mixing time via the coupling technique. More formally, we introduce the following key lemmas.

Lemma 8 *Let $\{X_t : t = 0, 1, \dots\}$ be the Swendsen-Wang chain on a complete bipartite graph of size (n, kn) with any constants $k \geq 1, B > 2$ and any starting state X_0. For any constant $\delta > 0$, there exists $T = O(1)$ such that $\|\alpha(X_T) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq \delta$ with probability $\Theta(1)$.*

Lemma 9 *Let $\{X_t : t = 0, 1, \dots\}$ be the Swendsen-Wang chain on a complete bipartite graph of size (n, kn) with any constants $k \geq 1, B > 2$. There exist constants $\delta, L > 0$ such that the following statement holds. Suppose that we start at state X_0 such that $\|\alpha(X_0) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq \delta$. Then, in $T = O(\log n)$ iterations, the Swendsen-Wang chain moves to $X_{T'}$ such that $\|\alpha(X_T) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq Ln^{-1/2}$ with probability $\Theta(1)$.*

Lemma 10 *Let $\{X_t : t = 0, 1, \dots\}$, $\{Y_t : t = 0, 1, \dots\}$ be Swendsen-Wang chains on a complete bipartite graph of size (n, kn) with any positive constants $k \geq 1, B \neq 2$. Let X_0, Y_0 be a pair of configurations satisfying*

$$\|\alpha(X_0) - (\alpha_L^*, \alpha_R^*)\|_\infty, \|\alpha(Y_0) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq Ln^{-1/2}$$

for some constant $L > 0$. Then, there exists a coupling for (X_t, Y_t) such that $\alpha(X_1) = \alpha(Y_1)$ with probability $\Theta(1)$.

Lemma 11 *Let $\{X_t : t = 0, 1, \dots\}$, $\{Y_t : t = 0, 1, \dots\}$ be Swendsen-Wang chains on a complete bipartite graph of size (n, kn) with any constants $k \geq 1, B > 0$. For any constant $\varepsilon > 0$, there exist $T = O(\log n)$ and a coupling for (X_t, Y_t) such that $\Pr[X_T \neq Y_T \mid \alpha(X_0) = \alpha(Y_0)] \leq \varepsilon$.*

The proofs of the above lemmas are presented in the supplementary material. Since the proof of Lemma 11 is identical to that of Lemma 9 in [10], we omit it. Now, we are ready to complete the proof of Theorem 3 for $B > 2$.

Consider two copies X_t, Y_t under the Swendsen-Wang chain. We will show that for some $T = O(\log n)$, there exists a coupling such that $\Pr[X_T \neq Y_T] \leq 1/4$. Let δ, L be as in Lemma 8 and Lemma 9. Then, for some $T_1 = O(1)$ with probability $\Theta(1)$, we have that

$$\|\alpha(X_{T_1}) - (\alpha_L^*, \alpha_R^*)\|_\infty, \|\alpha(Y_{T_1}) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq \delta.$$

Furthermore, for some $T_2 = O(\log n)$ with probability $\Theta(1)$, we have that

$$\begin{aligned} \|\alpha(X_{T_1+T_2}) - (\alpha_L^*, \alpha_R^*)\|_\infty &\leq Ln^{-1/2} \\ \|\alpha(Y_{T_1+T_2}) - (\alpha_L^*, \alpha_R^*)\|_\infty &\leq Ln^{-1/2}. \end{aligned} \tag{6}$$

Conditioning on (6) and using Lemma 10, there exists a coupling that $\alpha(X_{T_1+T_2+1}) = \alpha(Y_{T_1+T_2+1})$ holds with probability $\Theta(1)$. Conditioning on $\alpha(X_{T_1+T_2+1}) = \alpha(Y_{T_1+T_2+1})$ and using Lemma 11, for any constant $\varepsilon' > 0$, there exists $T_3 = O(\log n)$ and another coupling such that $\Pr(X_{T_1+T_2+T_3+1} \neq Y_{T_1+T_2+T_3+1}) \leq \varepsilon'$. Since all events so far occur with probability $\Theta(1)$, there exists small enough constant ε' so that $\Pr(X_T \neq Y_T) \leq 1/4$ for some $T = O(\log n)$ under some coupling. This completes the proof of Theorem 3 for the case $B > 2$.

Rapid mixing proof for $B < 2$. In this case, we will show that $\alpha(X_t)$ moves within $O(n^{-1/2})$ distance from (α_L^*, α_R^*) in $O(1)$ iterations. Then, we will bound the mixing time via the coupling technique as before. More formally, we introduce the following key lemmas.

Lemma 12 *Let $\{X_t : t = 0, 1, \dots\}$ be the Swendsen-Wang chain on a complete bipartite graph of size (n, kn) with any constants $k \geq 1, B < 2$. There exists a constant L such that for any starting state X_0 after $T = O(1)$, the Swendsen-Wang chain moves to state X_T such that $\|\alpha(X_T) - (\alpha_L^*, \alpha_R^*)\|_\infty \leq Ln^{-1/2}$ with probability $\Theta(1)$.*

The proof of Lemma 12 is presented in the supplementary material. By combining Lemmas 10-12 and

using same arguments used for the case $B > 2$, one can complete the proof of Theorem 3 for $B < 2$.

5 EXPERIMENTS

In this section, we compare the empirical performances of the Swendsen-Wang and the Gibbs chains for learning parameters of ferromagnetic Ising models. We construct models on real world social graphs and synthetic stochastic partitioned graphs by assigning random parameters $[\beta_{uv}], [\gamma_v]$ on graphs. For the choice of learning algorithm, we use the popular contrastive divergence (CD) algorithm [18] which uses a Markov chain as its subroutine.

Data sets. For each model, we generate a data set of 1000 samples by running the Swendsen-Wang chain. To construct a model, we use two real world social graphs which are known to have certain partitioned structures, e.g., see [14]. The first social graph is a Facebook graph consisting of 4039 nodes and 88234 edges, originally used in [28]. Each node of the graph corresponds to an account of Facebook and each edge of the graph corresponds to a 'friendship' in Facebook. The second social graph is a UCI graph created from an online community consisting of 1899 nodes and 13838 edges, originally used in [36]. Each node in the graph corresponds to a student at the University of California, Irvine and each edge in the graph corresponds to the message log from April to October 2004, i.e. edge (u, v) exists if u sent message to v or vice versa. For the real world social graphs, we assign $\gamma_v \sim \text{Unif}(0, 0.1), \text{Unif}(-0.1, 0.1)$, i.e., both positive and mixed external field, and $\beta_{uv} \sim \text{Unif}(0, x)$ where $x \in [0.01, 1]$.[4] For given x, we sample 10 i.i.d. $[\beta_{uv}]$ to obtain 10 different models.

Our synthetic stochastic partitioned graphs are bipartite random graphs of 100 to 1000 vertices with two partitions of same size, i.e. $|V_1| = |V_2|$. We set the inter-partition edge probability $p_{11} = p_{22} = 0.007$ and the intra-partition edge probability $p_{12} = 0.003$. For each graph size, we sample 10 bipartite random graphs. For synthetic graphs, we assign $\gamma_v \sim \text{Unif}(0, 0.1), \text{Unif}(-0.1, 0.1)$ and $\beta_{uv} \sim \text{Unif}(0, 1)$.

Contrastive divergence learning. Given a data set, the most standard way to estimate/recover parameters of a 'hidden' model is the log-likelihood maximization. To this end, it is known [43] that gradients of a graphical model requires computations of marginal probabilities, e.g., $E[\sigma_u \sigma_v]$ and $E[\sigma_v]$, where one can run a Markov chain to estimate them. However, this is not efficient since the Markov chain has to be run for large enough iterations until it mixes. To address the issue, the contrastive divergence (CD) learning algorithm [18] suggests that it suffices to run a Markov chain for a fixed number of iterations to approximate each gradient. The underlying intuition under CD learning is that it is not necessary to wait for mixing for each gradient update since parameters are slowly changing and mixing effects are amortized over iterations. The detailed procedure of the algorithm is presented in Algorithm 1.

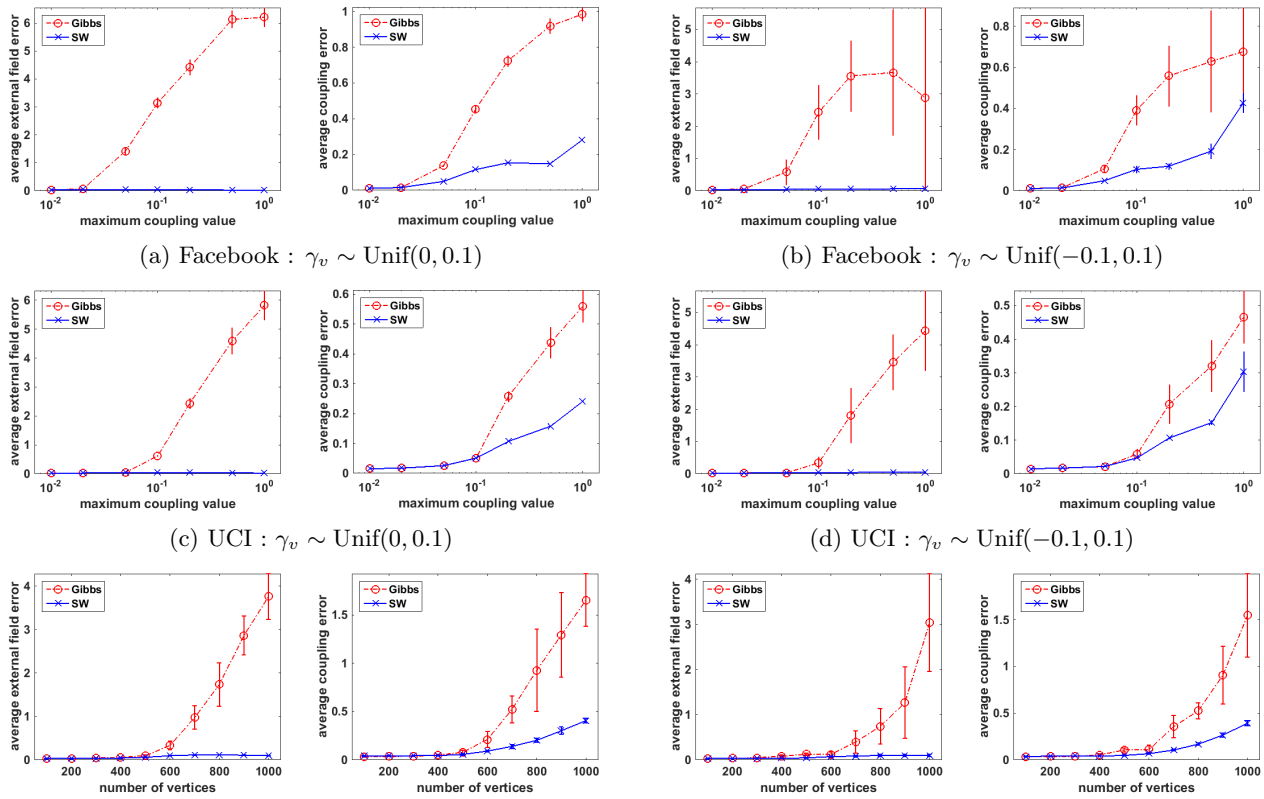
Algorithm 1 Contrastive Divergence Learning

Input: n_i, $\eta(\cdot)$, k, n_s, $\texttt{MC}(\cdot, \cdot)$, μ_{uv}, μ_v
Output: Estimated parameters $[\hat{\beta}_{uv}]$, $[\hat{\gamma}_v]$
Initialization: $i, \hat{\beta}_{uv}, \hat{\gamma}_v \leftarrow 0$ and randomly initialize states $\sigma^1, \ldots, \sigma^{n_s}$ of Ising model
while $i < n_i$ **do**
  $s \leftarrow 0$
  while $s < n_s$ **do**
    $\sigma^s \leftarrow \texttt{MC}(\sigma^s, k)$
    $s \leftarrow s + 1$
  end while
  $\hat{\mu}_{uv} \leftarrow \frac{1}{n_s} \sum_{s=1}^{n_s} \sigma_u^s \sigma_v^s$
  $\hat{\mu}_v \leftarrow \frac{1}{n_s} \sum_{s=1}^{n_s} \sigma_v^s$
  $\hat{\beta}_{uv} \leftarrow \hat{\beta}_{uv} + \eta(i)(\mu_{uv} - \hat{\mu}_{uv})$ for all $(u, v) \in E$
  $\hat{\gamma}_v \leftarrow \hat{\gamma}_v + \eta(i)(\mu_v - \hat{\mu}_v)$ for all $v \in V$
  $i \leftarrow i + 1$
end while

In Algorithm 1, we denote $\texttt{MC}(\sigma, k)$ as a state of the Ising model generated from running k iterations of Markov chain starting from the state σ, and $\mu_{uv} = E[\sigma_u \sigma_v], \mu_v = E[\sigma_v]$ are empirical marginals from the data set. In addition, n_i, $\eta(\cdot)$, k, n_s the number of gradient updates, the step size (or learning rate), the number of samples and the number of MC updates, respectively, which are hyper parameters of the CD algorithm. Since the Swendsen-Wang chain takes $O(|V|)$ times longer per each iteration, we use $k = 1$ and $k = |V|$ for the Swendsen-Wang chain and the Gibbs chain, respectively, for fair comparisons.

Experimental results. In our experiments, we observe that the Swendsen-Wang chain outperforms the Gibbs chain, where the gap is significant as β_{uv} or a graph size are large. Our experimental results on real world graphs are reported in Figure 2a, 2c, 2b, 2d, which show that the Swendsen-Wang chain outperforms the Gibbs chain for both errors on $[\gamma_v]$ and $[\beta_{uv}]$. One can observe that the error difference of the Swendsen-Wang chain and the Gibbs chain grows as interaction strength $[\beta_{uv}]$ increases, which is because the Gibbs chain mixes slower at low temperatures. Furthermore, the variance of errors of the Gibbs chain increases while the variance of the Swendsen-Wang chain remains small. Our experimental results using

[4] $\text{Unif}(a, b)$ denotes the random variable chosen in the interval $[a, b]$ uniformly at random.



(a) Facebook : $\gamma_v \sim \text{Unif}(0, 0.1)$

(b) Facebook : $\gamma_v \sim \text{Unif}(-0.1, 0.1)$

(c) UCI : $\gamma_v \sim \text{Unif}(0, 0.1)$

(d) UCI : $\gamma_v \sim \text{Unif}(-0.1, 0.1)$

(e) Synthetic : $\gamma_v \sim \text{Unif}(0, 0.1)$

(f) Synthetic : $\gamma_v \sim \text{Unif}(-0.1, 0.1)$

Figure 2: x-axis value x of (a), (b), (d), (e) is a range that β_{uv} is sampled from, i.e. $\beta_{uv} \sim \text{Unif}(0, x)$, and x-axis value of (c), (f) is a number of vertices in a graph. y-axis of external field error is a normalized external field error $\sum_{v \in V} |\gamma_v - \hat{\gamma}_v|/|V|$ and y-axis of coupling error is a normalized coupling error $\sum_{(u,v) \in E} |\beta_{uv} - \hat{\beta}_{uv}|/|E|$. Each point is an average of 10 independent Ising models while each Ising model is learnt by 1000 data samples.

synthetic graphs are similar to those of the real world social graphs. Figure 2e, 2f show that the Swendsen-Wang chain also outperforms the Gibbs chain as a graph size grows. We observe that the external field error of the Gibbs chain increases as a graph size increases while that of the Swendsen-Wang chain tends to stay.

6 Conclusion

Despite rich expressive powers of graphical models, their expensive inference tasks have been the key bottleneck for their large-scale applications. In this paper, we prove that the Swendsen-Wang sampler mix fast for stochastic partitioned attractive GMs, where our mixing bound $O(\log n)$ is quite practical for large-scale instances. We believe that our findings have much more potentials even for general (not necessarily, attractive) GMs if one can approximate a non-attractive model by an attractive one; it was recently shown that any binary pairwise GM can be approximated by an attractive binary pairwise one on the so-called 2-cover graph having two partitions [39]. For example, one can use the Swendsen-Wang sampler to learn parameters of the 2-cover attractive model and further fine-tune them using the Gibbs sampler on the original model. This is an interesting future research direction.

Acknowledgement. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (NO.R0132-17-1005), Content visual browsing technology in the online and offline environments. The first author, Sejun Park, was supported in part by the Bloomberg Data Science Research Grant. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 334828. The paper reflects only the authors' views and not the views of the ERC or the European Commission. The European Union is not liable for any use that may be made of the information contained therein. Research supported in part by NSF grant CCF-1318374. Research supported in part by NSF grant CCF-1217458.

References

[1] Rodney J Baxter. *Exactly solved models in statistical mechanics*. Courier Corporation, 2007.

[2] Christian Borgs, Jennifer T Chayes, Alan M Frieze, Jeong Han Kim, Prasad Tetali, Eric Vigoda, and Van H Vu. Torpid mixing of some Monte Carlo Markov Chain algorithms in statistical physics. In *40th Annual Symposium on Foundations of Computer Science, (FOCS)*, pages 218–229. IEEE, 1999.

[3] Christian Borgs, Jennifer T Chayes, and Prasad Tetali. Tight bounds for mixing of the Swendsen–Wang algorithm at the Potts transition point. *Probability Theory and Related Fields*, 152(3):509–557, 2010.

[4] Colin Cooper and Alan M Frieze. Mixing properties of the Swendsen-Wang process on classes of graphs. *Random Structures & Algorithms*, 15(3-4):242–261, 1999.

[5] Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405, 1990.

[6] Paul Cuff, Jian Ding, Oren Louidor, Eyal Lubetzky, Yuval Peres, and Allan Sly. Glauber dynamics for the mean-field Potts model. *Journal of Statistical Physics*, 149(3):432–477, 2012.

[7] Oliver B Downs, David JC MacKay, and Daniel D Lee. The nonnegative Boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*, pages 428–434, 2000.

[8] Robert G Edwards and Alan D Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and monte carlo algorithm. *Physical Review D*, 38(6):2009, 1988.

[9] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[10] Andreas Galanis, Daniel Stefankovic, and Eric Vigoda. Swendsen-Wang algorithm on the Mean-Field Potts model. In *Proceedings of RANDOM*, pages 815–828, 2015.

[11] Andreas Galanis, Daniel Stefankovic, Eric Vigoda, and Linji Yang. Ferromagnetic potts model: Refined #bis-hardness and related results. In *Proceedings of RANDOM*, pages 677–691, 2014.

[12] Robert Gallager. Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1):21–28, 1962.

[13] A. Gerschenfeld and A. Monianari. Reconstruction for models on random graphs. In *48th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 194–204. IEEE, 2007.

[14] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[15] Leslie A Goldberg and Mark Jerrum. The complexity of ferromagnetic Ising with local fields. *Combinatorics, Probability and Computing*, 16(01):43–61, 2007.

[16] Vivek K Gore and Mark R Jerrum. The Swendsen–Wang process does not always mix rapidly. *Journal of Statistical Physics*, 97(1):67–86, 1999.

[17] Heng Guo and Mark Jerrum. Random cluster dynamics for the ising model is rapidly mixing. *arXiv preprint arXiv:1605.00139*, 2016.

[18] Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade*, pages 599–619. Springer, 2012.

[19] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[20] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

[21] Michael I. Jordan. *Learning in Graphical Models:[proceedings of the NATO Advanced Study Institute...: Ettore Mairona Center, Erice, Italy, September 27-October 7, 1996]*, volume 89. Springer Science & Business Media, 1998.

[22] Frank R Kschischang and Brendan J Frey. Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Journal on Selected Areas in Communications*, 16(2):219–230, 1998.

[23] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[24] David A Levin, Malwina J Luczak, and Yuval Peres. Glauber dynamics for the mean-field Ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146(1-2):223–265, 2010.

[25] David A Levin, Yuval Peres, and Elizabeth L Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[26] Hui Li, Xiaoyi Li, Xiaowei Jia, Murali Ramanathan, and Aidong Zhang. Bone disease prediction and phenotype discovery using feature representation over electronic health records. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 212–221. ACM, 2015.

[27] Yun Long, Asaf Nachmias, Weiyang Ning, and Yuval Peres. A power law of order 1/4 for critical mean field Swendsen-Wang dynamics. *Memoirs of the AMS*, 232(1092), 2014.

[28] Julian J McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 548–556, 2012.

[29] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

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

[31] Elchanan Mossel and Allan Sly. Rapid mixing of Gibbs sampling on graphs that are sparse on average. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 238–247. Society for Industrial and Applied Mathematics, 2008.

[32] Elchanan Mossel and Allan Sly. Exact thresholds for Ising–Gibbs samplers on general graphs. *The Annals of Probability*, 41(1):294–328, 2013.

[33] Tu Dinh Nguyen, Truyen Tran, Dinh Q Phung, and Svetha Venkatesh. Learning parts-based representations with nonnegative restricted Boltzmann machine. In *Asian Conference on Machine Learning (ACML)*, pages 133–148, 2013.

[34] TAGA Nobuyuki and MASE Shigeru. On the convergence of loopy belief propagation algorithm for different update rules. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 89(2):575–582, 2006.

[35] Jungseul Ok, Youngmi Jin, Jinwoo Shin, and Yung Yi. On maximizing diffusion speed in social networks: impact of random seeding and clustering. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 301–313. ACM, 2014.

[36] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[37] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[38] Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1):273–302, 1996.

[39] Nicholas Ruozzi and Tony Jebara. Making pairwise binary graphical models attractive. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1772–1780, 2014.

[40] Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2):86, 1987.

[41] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of biomedical informatics*, 54:96–105, 2015.

[42] Mario Ullrich. *Rapid mixing of Swendsen-Wang dynamics in two dimensions*. PhD thesis, Universität Jena, Germany, 2012. arXiv preprint arXiv:1212.4908.

[43] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[44] Adrian Weller and Tony Jebara. Bethe bounds and approximating the global optimum. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 618–631, 2013.

[45] Adrian Weller and Tony Jebara. Approximating the bethe partition function. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.