Learning Structure of Power-Law Markov Networks

Abhik Kumar Das, Praneeth Netrapalli, Sujay Sanghavi and Sriram Vishwanath
Department of Electrical & Computer Engineering, The University of Texas at Austin, USA
Email: {akdas,praneethn}@utexas.edu, sanghavi@mail.utexas.edu, sriram@austin.utexas.edu

***Abstract*—We consider the problem of learning the underlying graph structure of discrete Markov networks based on power-law graphs, generated using the configuration model. We translate the learning problem into an equivalent channel coding problem and obtain necessary conditions for solvability in terms of problem parameters. In particular, we relate the exponent of the power-law graph to the hardness of the learning problem, and show that more number of samples are required for exact recovery of discrete power-law Markov graphs with small exponent values. We develop an efficient learning algorithm for accurate reconstruction of graph structure of Ising model on power-law graphs. Finally, we show that order-wise optimal number of samples suffice for recovering the exact graph under certain constraints on Ising model parameters and scalings of node degrees.**

***Index Terms*—Markov network, power-law graph, Ising model**

I. Introduction

Markov networks, also known as graphical models, provide a powerful framework for succinctly encoding probability distributions in form of undirected graphs – random variables get mapped to nodes of the undirected graph, while the interdependencies among them get mapped to its edges. As such, Markov networks are widely used for modeling and designing applications in a multitude of settings like social networks [1], [2], image processing [3], [4], and computational biology [5], [6]. With the increasing use of this framework in complex and non-conventional domains, the problem of selecting the most suitable Markov network from among the large space of possible network structures has gained considerable importance. Thus, the domain of successful recovery of Markov graphs using observed samples generated from their probability distributions, also known as the *graphical model selection problem*, is an active area of study and research.

An interesting approach for tackling the learning problem is to interpret it as an equivalent noisy channel coding problem [7], and utilize *achievability* and *converse* techniques to derive the necessary and sufficient conditions related to recovery. While achievability can be associated with designing learning algorithms that can accurately estimate the graph structure and parameters of Markov networks from observed samples, converse characterizes the information-theoretic limits of the learning problem, i.e., necessary conditions on the nature and number of samples that individuate the Markov networks.

Power-law graphs are relatively common across a variety of domains. A power-law graph is one whose degree sequence exhibits a power-law or Pareto probability distribution. The standard property of a power-law graph is as follows – given $\alpha > 1$, the number of nodes with degree k in a power-law graph having exponent α is approximately proportional to $k^{-\alpha}$. Examples of instances where power-law behavior has been observed include social networks [8], protein complex networks [9], gene networks [10] and portions of the internet [11]. Thus, many Markov networks derived from natural or practical setups are typically based on power-law graphs.

In this paper, we consider the problem of learning the underlying graph structure of discrete Markov networks based on power-law graphs. We explore the connection between the power-law exponent and sample complexity of the learning problem, examining it both from achievability and converse perspectives. Understanding the picture concerning sample complexity is critical when designing algorithms for Markov graph recovery, for example, when the minimum node degree scales like a constant, while the maximum node degree scales with the number of nodes [12]. We consider the family of power-law graphs generated using the configuration model [13], and use structural properties of these power-law graphs for designing a learning algorithm ensuring accurate graph recovery for the Ising model (assuming certain constraints are satisfied). Thus, we investigate the relationship between the hardness of learning Markov graphs and their structural properties, providing some partial answers in this regard.

Related Work: There is a significant body of literature related to analysis of the graphical model selection problem, especially for specialized families of Markov networks such as Ising model [14]–[18] and Gaussian model [19]–[21], with their underlying graphs selected from ensembles of degree-bounded graphs [14]–[16], [19], [21], large-girth graphs [17], and sparse random graphs like Erdös-Rényi and small-world graphs [16], [21]. As far as the converse aspect is concerned, strong lower bounds on the probability of error of learning algorithms are derived in [16] and [22] for exact recovery of structure of Ising model based on Erdös-Rényi graphs, and Gaussian model based on degree-bounded graphs respectively. Likewise, lower bounds on number of observed samples are obtained in [12], for ensuring accurate reconstruction of discrete Markov networks based on two ensembles of power-law graphs, namely the configuration model and Chung-Lu model [23], with exponent exceeding 3 for both of them.

Learning algorithms (for achievability aspect) can broadly be classified into three categories – search-based, optimization-based, and greedy techniques. Search-based algorithms find the smallest set of nodes through exhaustive search, conditioned on which a node is independent of others [14], [16], [21]. Optimization-based algorithms frame the learning problem as a convex optimization problem, but require a strong

incoherence assumption to ensure exact recovery [20]. The algorithms that use greedy methods, discover the neighborhoods of nodes by minimizing some function of the random variables, like conditional entropy, in a greedy fashion [17], [18]. [12] examines the performance of these algorithms for learning power-law Markov graphs and observes that sample complexity scales poorly with the number of nodes if the variation in degrees of nodes is large; it also states that the task of designing efficient and near-optimal learning algorithms for such Markov networks is an outstanding open problem.

Main Results: We examine the impact of power-law exponent on the information-theoretic limits of sample complexity, and use converse arguments to show that learning algorithms require greater number of samples (in order-wise sense) for exact recovery of Markov graphs having small exponent values. Moreover, a sharp transition in the sample complexity requirement is observed at exponent value of 2. A major issue faced while designing algorithms for power-law Markov networks is the absence of reasonable guarantees on number of samples required for exact recovery [12]; for example, if the degrees of multiple nodes scale with the number of nodes. We design a learning algorithm, motivated by the one in [21], and show that the graph structure of ferromagnetic Ising model based on power-law graphs with p random variables and exponent greater than 3 can be accurately learnt using $\Omega(\log_2 p)$ samples, that is order-wise optimal if the minimum degree scales like a constant. In case the exponent lies between 2 and 3, we obtain a sample complexity requirement that is poly-log in the number of nodes ($\Omega((\log_2 p)^3)$ samples, to be precise), under certain constraints on scalings of degrees.

Due to limitation of space, we omit the proofs of results presented in this paper. The audience can refer to [24] for a more detailed version of the paper, along with the proofs.

II. Preliminaries

We consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and E is the set of edges. A Markov network is obtained by associating a random variable X_i to $i \in V$, that takes values from some alphabet set $\mathcal{A}$, and specifying a joint probability distribution $f(\cdot)$ over vector $X = (X_1, X_2, \ldots, X_p)$ that possesses the following property:

$$f(x_A, x_B | x_C) = f(x_A | x_C) f(x_B | x_C),$$

where A, B, C are any disjoint subsets of V such that every path between a node in A and a node in B passes through at least one node in C, and x_A, x_B, x_C denote the restrictions of $(x_1, \ldots, x_p) \in \mathcal{A}^p$ to indices in A, B, C respectively. Note that $f(\cdot)$ denotes the probability mass function (p.m.f.) for the case of discrete Markov networks (i.e., $\mathcal{A}$ is a finite set).

Ising Model: This is a well-known family of discrete Markov networks, studied in diverse fields like statistical physics, computer vision and game theory. An Ising model, with G as its underlying graph, is obtained by setting $\mathcal{A} = \{-1, 1\}$, assigning node potentials $h_i \in \mathbb{R}$ to $i \in V$ and edge potentials $\theta_{ij} \in \mathbb{R}$ to $(i, j) \in E$. Then the p.m.f. of X satisfies

$$f(x) \propto \exp\left(\sum_{i \in V} h_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right).$$

A special case of the Ising model is the *ferromagnetic Ising model*, where $\theta_{ij} > 0$ for all $(i, j) \in E$. Note that the normalization constant associated with the p.m.f. is affected by the graph topology and values of node/edge potentials.

A. Learning Algorithm and Error Criterion

We denote the set of undirected graphs on p nodes by $\mathcal{U}_p$. We consider a family of discrete Markov networks, comprising of p random variables that take values from $\mathcal{A}$, and an ensemble of undirected graphs $\mathcal{G} \subseteq \mathcal{U}_p$. We choose $G \in \mathcal{G}$ uniformly at random, select a Markov network, with G as its underlying graph, from the family, and obtain n i.i.d. vector samples $\mathbf{x}^n = (x^{(1)}, \ldots, x^{(n)})$ from this distribution. The problem of learning Markov graphs is to reconstruct G from $\mathbf{x}^n$. A learning algorithm is any mapping $\phi : \mathcal{A}^{np} \to \mathcal{U}_p$ that enables us to generate a graph estimate $\hat{G} = \phi(\mathbf{x}^n)$. We define the error event as $\{\hat{G} \neq G\}$ (i.e., we focus on exact graph recovery); thus, the probability of error of ϕ is given by

$$P_e^{(n)}(\phi) = P(\hat{G} \neq G) = P(\phi(\mathbf{x}^n) \neq G).$$

The converse aspect of the learning problem is concerned with obtaining a strong lower bound on $P_e^{(n)}(\phi)$ in terms of n and graph ensemble parameters for any ϕ. On the other hand, the achievability aspect of the learning algorithm is concerned with designing ϕ having the property that $P_e^{(n)}(\phi)$ becomes arbitrarily small as the problem and sample sizes increase.

III. Configuration Model

We consider a degree sequence $\mathbf{d} = (d_1, d_2, \ldots, d_p)$ for an undirected graph on p nodes, and a set of configuration points $W = \{1, 2, \ldots, 2m\}$, where $2m = \sum_{i=1}^{n} d_i$. We define $W_k = \left\{ \sum_{i=1}^{k-1} d_i + 1, \sum_{i=1}^{k-1} d_i + 2 \ldots, \sum_{i=1}^{k} d_i \right\}$, for $k = 1, 2, \ldots, p$ (we set $d_0 = 1$). Thus, $\{W_k : 1 \leq k \leq p\}$ forms a partition of W with $|W_k| = d_k$. Next, we define a mapping $\psi : W \to \{1, 2, \ldots, p\}$ such that $\psi(x) = k$ for $x \in W_k$. Then, given a (perfect) matching $\mathcal{F}$ for W (i.e., a partition of W into m pairs $\{x, y\}$), one can obtain a multi-graph $G(\mathcal{F}) = (V, E)$ with $V = \{1, 2, \ldots, p\}$ and $(\psi(x), \psi(y)) \in E$ for each $\{x, y\} \in \mathcal{F}$. Therefore, choosing a matching $\mathcal{F}$ for W uniformly at random results in the generation of a multi-graph $G(\mathcal{F})$, where $i \in V$ has degree d_i. We refer to this as the *configuration model* and designate the ensemble by $\mathcal{G}(\mathbf{d})$.

The number of distinct matchings $\mathcal{F}$ of the $2m$ points in W is given by $N_{2m} = \frac{(2m)!}{m! 2^m}$. We call a multi-graph *simple* if it has no self-loops or multiple edges between nodes. An interesting point to note is that the number of matchings corresponding to each simple graph in $\mathcal{G}(\mathbf{d})$ is the same, i.e., simple graphs are equiprobable in the space of multi-graphs. We refer to the subset of simple graphs as $\mathcal{G}_s(\mathbf{d}) \subset \mathcal{G}(\mathbf{d})$. We define $d_{\min} = \min_{i \in V} d_i$, $d_{\max} = \max_{i \in V} d_i$, and assume that

$d_{\max} = o(p^{\frac{1}{3}})$, $d_{\min} = o(d_{\max})$ – under these scalings, it is known that the probability of $G(\mathcal{F})$ being simple for uniformly chosen $\mathcal{F}$ approaches $q_s = \exp(-\frac{\nu}{2} - \frac{\nu^2}{4})$ as $p \to \infty$, where $\nu = \frac{\sum_i d_i^2}{\sum_i d_i} - 1$ [25]. Using this fact gives the following result:

Lemma III.1. *Given* $\mathbf{d}$ *and large enough values of* p*, we have*

$$|\mathcal{G}_s(\mathbf{d})| \geq \frac{q_s}{2} \frac{N_{2m}}{\prod_{i \in V} d_i!} \geq \frac{q_s}{2} \prod_{i \in V} \left(\frac{m^{\frac{1}{2}}}{2d_i} \right)^{d_i}.$$

A. Generating Power-Law Graphs

Given $\alpha > 1$, a power-law graph with exponent α has the property that the number of nodes with degree k is proportional to $k^{-\alpha}$. For p nodes and given values of $d_{\min}, d_{\max}$, we define $\zeta(\alpha) = (\sum_{k=d_{\min}}^{d_{\max}} k^{-\alpha})^{-1}$. Then the number of nodes with degree k approximately equals $p\zeta(\alpha)k^{-\alpha}$, where $d_{\min} \leq k \leq d_{\max}$. For the sake of simplicity, we assume that $p\zeta(\alpha)k^{-\alpha}$, $d_{\min} \leq k \leq d_{\max}$, are integers. Note that this imposes the constraint $d_{\max} \leq (p\zeta(\alpha))^{\frac{1}{\alpha}}$, since there is at least one node with degree $d_{\max}$. Therefore, we assume $d_{\max} = o(p^{\min(\frac{1}{3}, \frac{1}{\alpha})})$, and define the degree sequence $\mathbf{d}$ as

$$d_j = l, \quad p\zeta(\alpha) \left(\sum_{k=d_{\min}}^{l-1} k^{-\alpha} \right) < j \leq p\zeta(\alpha) \left(\sum_{k=d_{\min}}^{l} k^{-\alpha} \right),$$

for $d_{\min} \leq l \leq d_{\max}$. We denote the resulting ensemble of power-law graphs by $\mathcal{G}_\alpha$, and its subset of simple graphs by $\mathcal{G}_{s,\alpha}$. We also define the following quantities, dependent on $\mathbf{d}$:

$$\bar{d} = \begin{cases} \frac{(\alpha-1)}{(2-\alpha)} d_{\max}^{2-\alpha} d_{\min}^{\alpha-1} & , \ 1 < \alpha < 2, \\ \frac{(\alpha-1)}{(\alpha-2)} d_{\min} & , \ \alpha > 2, \end{cases}$$

$$\tilde{d} = \begin{cases} \frac{(2-\alpha)}{(3-\alpha)} d_{\max} & , \ 1 < \alpha < 2, \\ \frac{(\alpha-2)}{(3-\alpha)} d_{\max}^{3-\alpha} d_{\min}^{\alpha-2} & , \ 2 < \alpha < 3, \\ \frac{(\alpha-2)}{(\alpha-3)} d_{\min} & , \ \alpha > 3. \end{cases}$$

One can check that $\bar{d}$ is within a constant factor of average degree of the power-law graph (i.e., $\frac{\sum_i d_i}{p}$), and $\tilde{d}$ is within a constant factor of the ratio of average squared degree to average degree (i.e., $\frac{\sum_i d_i^2}{\sum_i d_i}$). We have the following lemma:

Lemma III.2. *Given* $\alpha > 1$ *and large enough values of* p*, there exists a positive constant* c_0 *s.t. for* $d_{\min} \geq c_0$*, we have*

$$\log_2 |\mathcal{G}_{s,\alpha}| \geq \frac{p\bar{d}}{9} \log_2 \left(\frac{(\alpha-1)}{|\alpha-2|} p \right).$$

Note that we inherently assume that $d_{\min}$ is larger than some suitable constant and p is chosen sufficiently large in all the subsequent results concerning the graphs in $\mathcal{G}_\alpha$ and $\mathcal{G}_{s,\alpha}$.

Next, we state a structural property of power-law graphs in $\mathcal{G}_{s,\alpha}$ with $\alpha > 2$, that facilitates the process of learning discrete Markov networks based on them. For this, we make some additional stronger assumptions on the scalings of $d_{\min}, d_{\max}$:

(A1) $d_{\min} = \Theta(1)$, $d_{\max} = o((\log_2 p)^{\frac{1}{2(3-\alpha)}})$, $2 < \alpha < 3$,

(A2) $d_{\min} = \Theta(1)$, $d_{\max} = o(p^{\min(\frac{1}{8}, \frac{1}{\alpha})})$, $\alpha > 3$.

These restrictions on scalings ensure that $\nu = o((\log_2 p)^{\frac{1}{2}})$; therefore, the probability of getting a simple graph scales as $\exp(-o(\log_2 p))$. This implies that as long as some structural property is satisfied for a uniformly generated graph from $\mathcal{G}_\alpha$ with probability $\geq 1 - p^{-\Theta(1)}$, it also holds for a uniformly selected graph from $\mathcal{G}_{s,\alpha}$ with probability $\geq 1 - p^{-\Theta(1)}$.

Given an integer r, we define the r-neighborhood of a node in a graph as the subgraph comprising of nodes that are reachable from it via at most r edges. We define $r_0 = \frac{1}{3} \frac{\log_2 p}{\log_2(80\tilde{d})}$ and assume it is an integer. Then the following result holds:

Lemma III.3. *Given* $\alpha > 2$ *and assumptions (A1), (A2) hold, if a graph is selected uniformly at random from* $\mathcal{G}_{s,\alpha}$*, at most one cycle exists in the* r_0*-neighborhood of any node (i.e., the graph is locally tree-like) with probability* $\geq 1 - p^{-\Theta(1)}$.

In other words, there are a few paths of length $\leq r_0$ between any two nodes in a graph of $\mathcal{G}_{s,\alpha}$ with high probability. This implies that for discrete Markov networks based on graphs in $\mathcal{G}_{s,\alpha}$, any two random variables are near-independent, conditioned on a small number of random variables. This motivates the design of search-based learning algorithm for recovering the graph structure of these Markov networks.

IV. Learning Algorithm: Analysis and Design

The description of the setup for analysis in Section II-A enables the interpretation of graphical model selection problem as a channel coding problem, where the Markov graph, set of observed samples and learning algorithm can be treated as the transmitted message, received signal and message decoder respectively. This enables the application of converse and achievability techniques for a better understanding of limitations and performance of the learning problem framework.

A. Lower Bounds on Sample Complexity (Converse)

First, we examine the converse aspect and derive lower bounds on number of samples required for any algorithm to accurately learn the graph structure of a discrete Markov network with its graph in $\mathcal{G}_{s,\alpha}$. To be precise, we obtain a threshold value such that if the number of samples, n, is less than the threshold value, the probability of error of any learning algorithm is bounded away from zero. For this, we select $\mathcal{G} = \mathcal{G}_{s,\alpha}$ and consider any family of discrete Markov networks with graphs in $\mathcal{G}_{s,\alpha}$. Then the following result holds:

Theorem IV.1. *Given any learning algorithm* ϕ*, if we have*

$$n < \frac{\bar{d}}{10 \log_2 |\mathcal{A}|} \log_2 \left(\frac{(\alpha-1)}{|\alpha-2|} p \right),$$

then its probability of error satisfies $\lim_{p \to \infty} P_e^{(n)}(\phi) = 1$.

Thus, Theorem IV.1 indicates a sample complexity requirement of $n = \Omega(d_{\max}^{2-\alpha} d_{\min}^{\alpha-1} \log_2 p)$ for $1 < \alpha < 2$, and $n = \Omega(d_{\min} \log_2 p)$ for $\alpha > 2$, to ensure exact recovery of the Markov graph. Note that the sample complexity result for $\alpha > 3$ matches the one derived in [12] in order-wise sense, where a slightly modified version of configuration model is used and $d_{\min}$ is set as 1. If $d_{\max}$ scales with p, Theorem IV.1 implies that a learning algorithm needs more number of samples (in order-wise sense) to reconstruct the underlying

graph of a discrete Markov graph when α is less than 2, as compared to when it is greater than 2. Also, there is a sharp transition in sample complexity requirement observed at $\alpha = 2$ – a potential reason for this phenomenon is that the fraction of high degree nodes decreases as α increases. In other words, it is inherently difficult to learn (in terms of sample complexity) power-law graph-based discrete Markov networks with lower exponents (less than 2, to be precise); moreover, this issue aggravates as the exponent decreases in value from 2 to 1.

B. Learning Algorithm for Ising Model (Achievability)

Next, we examine the achievability aspect of the learning problem and design an algorithm for learning the graph structure of Ising model based on graphs in $\mathcal{G}_{s,\alpha}$. In particular, we focus on the ferromagnetic Ising model family with bounded node potentials. As observed in Section IV-A, the sample complexity requirement for exact recovery tends to be large if the exponent is less than 2, since then the average degree depends on the maximum degree, that may scale with number of nodes. Therefore, we restrict ourselves to the regime $\alpha > 2$, and also allow assumptions (A1), (A2) to hold. As pointed out in [12], one of the major challenges that a learning algorithm faces in the context of power-law Markov networks is to tackle the large variation in degrees of nodes. We use a learning algorithm that resembles the empirical conditional variation distance thresholding-based algorithm, studied in [16], [17], [26] for reconstructing the graph structure of Ising model having degree-bounded, large-girth or Erdös-Rényi graphs.

Given $0 < \theta_{\min} \leq \theta_{\max}$, we consider the family of ferromagnetic Ising model on p random variables, with bounded node potentials and edge potentials lying in $[\theta_{\min}, \theta_{\max}]$. We choose any Ising model from this family with p.m.f. $f(\cdot)$ and assume $G = (V, E) \in \mathcal{G}_{s,a}$ to be its underlying graph. We define the following mappings for $i, j \in V$, that can be thought of some distance measure between random variables X_i, X_j:

$$\rho(i,j) = \min_{U \subseteq V : |U| \leq 2} \max_{x_i, x_i', x_U} |f(x_i|x_j, x_U) - f(x_i|x_j', x_U)|,$$

This allows us to define the empirical p.m.f., based on $\mathbf{x}^n$, as

$$\hat{f}(x) = \hat{f}(x_1, x_2, \ldots, x_p) = \frac{1}{n} \sum_{l=1}^{n} \mathbb{I}(x_i = x_i^{(l)},\, 1 \leq i \leq p),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Note that $\hat{f}(\cdot)$ can be used to compute the empirical marginal and conditional p.m.f.'s of $f(\cdot)$. This allows us to define the empirical version of $\rho(i,j)$:

$$\hat{\rho}(i,j) = \min_{U \subseteq V : |U| \leq 2} \max_{x_i, x_i', x_U} |\hat{f}(x_i|x_j, x_U) - \hat{f}(x_i|x_j', x_U)|.$$

The learning algorithm ϕ^* for obtaining $\hat{G}$, the estimate of G, is tabulated in form of Algorithm 1; the choice of threshold $\zeta_{n,p}$ influences the sample complexity requirement of ϕ^*. The motivation behind this learning algorithm comes from the observation that $\rho(i,j)$ tends to be larger when an edge exists between i and j than when the edge does not exist. In other words, the influence of X_j on X_i is more when i and j are neighbors versus when they are not neighbors. Also, ρ and $\hat{\rho}$ are close to each other in value if the number of samples n is large enough – we corroborate all these facts below.

Algorithm 1 Learning algorithm ϕ^* to obtain $\hat{G}$ from $\mathbf{x}^n$

Require: $V = \{1, 2, \ldots, p\}$, $\hat{E} = \emptyset$
  for all $i, j \in V$ **do**
    if $\hat{\rho}(i,j) > \zeta_{n,p}$ **then**
      $\hat{E} \leftarrow \hat{E} \cup \{(i,j)\}$
    end if
  end for
Ensure: $\hat{G} = (V, \hat{E})$

Non-neighboring Nodes: If $i, j \in V$ are non-neighboring nodes in G, by Lemma III.3 there exists at most two short paths of length at most r_0 connecting i to j with high probability. We define the l-separator set for two nodes as the minimum number of nodes that need to be removed for eliminating paths of length at most l between them. This implies the r_0-separator set size for i, j is at most two with high probability. Then a strong correlation decay result, related to separator sets, derived in [16] can be used to show that $\rho(i,j)$ becomes arbitrarily small as problem size increases.

Theorem IV.2. *Consider a ferromagnetic Ising model based on a uniformly selected graph from $\mathcal{G}_{s,\alpha}$, where $\alpha > 2$, assumptions (A1), (A2) hold, and $\tanh\theta_{\max} < (80\tilde{d})^{-2}$. Then $\rho(i,j) = o(p^{-\kappa})$ for some constant $\kappa > 0$, with probability $\geq 1 - p^{-\Theta(1)}$, for any pair of non-neighboring nodes i, j.*

Neighboring Nodes: If $i, j \in V$ are neighboring nodes, the following result shows that $\rho(i,j)$ is bounded away from zero:

Theorem IV.3. *Consider a ferromagnetic Ising model based on a uniformly selected graph from $\mathcal{G}_{s,\alpha}$, where $\alpha > 2$. Then $\rho(i,j) \geq \frac{1}{16}(1 - e^{-4\theta_{\min}})$ for any neighboring nodes i, j.*

Performance Analysis: We select $\zeta_{n,p} = \frac{1}{32}(1 - e^{-4\theta_{\min}})$. We also assume $\theta_{\min} = \Omega((\log_2 p)^{-r})$ for some constant $r > 0$. Then the following result describes the performance of ϕ^*:

Theorem IV.4. *Consider a ferromagnetic Ising model based on a uniformly selected graph from $\mathcal{G}_{s,\alpha}$, where $\alpha > 2$, assumptions (A1), (A2) hold, $\tanh\theta_{\max} < (80\tilde{d})^{-2}$ and $\theta_{\min} = \Omega((\log_2 p)^{-r})$ for some constant $r > 0$. Suppose $\zeta_{n,p} = \frac{1}{32}(1 - e^{-4\theta_{\min}})$, and number of i.i.d. samples n satisfies*

$$n > \frac{2^{18}}{(1 - e^{-4\theta_{\min}})^2 f_{\min}^2} \log_2 \left(\frac{p}{3}\right).$$

Then ϕ^* recovers the correct graph structure of the ferromagnetic Ising model with probability $\geq 1 - p^{-\Theta(1)}$. Moreover, the computational complexity for executing ϕ^* is $O(p^4)$.

Consequences of Theorem IV.4: Since we have the restriction that $\tanh\theta_{\min} \leq \tanh\theta_{\max} \leq (80\tilde{d})^{-2} \leq 80^{-2}$, we can approximate $1 - e^{-4\theta_{\min}} \approx 4\theta_{\min}$. Also, it can be shown that $f_{\min}$ is greater than some constant (or bounded away from zero), along the lines of [26] (the result in [26] is demonstrated for Erdos-Renyi graphs, but the same proof

technique can be used to prove that $f_{\min}$ is bounded in our case). Thus, it is sufficient to have a sample complexity of $n = \Omega(\theta_{\min}^{-2} \log_2 p)$ for ϕ^* to recover the correct graph structure. For the scaling $\theta_{\min} = \Theta(\tilde{d}^{-2})$, this translates to a sample complexity requirement of $n = \Omega(\tilde{d}^4 \log_2 p)$ – this reduces to $n = \Omega((\log_2 p)^3)$ with $\theta_{\min} = \Theta((\log p)^{-(3-\alpha)\delta_1})$ for $2 < \alpha < 3$, and $n = \Omega(d_{\min}^4 \log_2 p)$ with $\theta_{\min} = \Theta(d_{\min}^{-2})$ for $\alpha > 3$. Keeping in mind the fact that $n = \Omega(d_{\min} \log_2 p)$ is the information-theoretic lower bound on sample complexity for $\alpha > 2$ to ensure accurate recovery, one can note that the constraints are more restrictive and the sample complexity result is worse for the case $2 < \alpha < 3$ compared to the case $\alpha > 3$. However, since assumption (A2) makes $d_{\min}$ a constant, the sample complexity associated with the converse and achievability aspects match in an order-wise sense for $\alpha > 3$. A probable reason for the relatively poor performance of the learning algorithm for $2 < \alpha < 3$ could be the structural nature of power-law graphs in that regime – they tend to have a big core with many high degree nodes residing in it [23]. So, there is scope for designing improved learning algorithms with better performance (in both sample and computational complexity) in this regime of power-law exponent value.

Comparison with Previous Results: The statistical guarantees provided by some well-known algorithms in the context of learning power-law graphical models are examined in [12] – two generative models of power-law graphs are considered, the configuration model and Chung-Lu model [23]. The ℓ_1-regularization based learning algorithm [27] needs a sample complexity of $n = \Omega(d_{\max}^3 \log_2 p)$ for both configuration and Chung-Lu power-law graphs (the average degree is assumed to be $\Theta(1)$). The greedy algorithm, described in [28], performs slightly better and guarantees accurate recovery with $n = \Omega(d_{\max}^2 \log_2 p)$ samples. The performance analysis of the conditional variation distance thresholding estimator [16], the motivation behind learning algorithm ϕ^*, exhibits a trade-off in the case of learning Chung-Lu power-law graph-based Ising model – restricting the algorithm to run in polynomial time shoots up the sample complexity requirement to $\Omega(\text{poly}(p) \log_2 p)$. In contrast, by performing a careful analysis, we show that learning algorithm ϕ^* performs reasonably well for $\alpha > 3$. On the other hand, there is an additional $(\log_2 p)^2$ factor in the sample complexity requirement result for the regime $2 < \alpha < 3$ when $d_{\max} = \Theta(\text{poly}(\log_2 p))$.

V. Conclusion

We study the problem of learning the graph structure of discrete Markov networks based on power-law graphs generated using the configuration model and show that the learning problem presents a sharp increase in sample complexity, for ensuring exact graph recovery, when the power-law exponent is less than 2. Thereafter, we design a algorithm for learning the structure of power-law graph-based ferromagnetic Ising model, subject to certain constraints on node and edge potentials. Our learning algorithm is order optimal when the minimum degree scales as a constant and the power-law exponent is greater than 3. However, the sample complexity of our algorithm is sub-optimal when the power-law exponent lies between 2 and 3, and our future work will focus on improving the sample complexity requirement results in this range.

References

[1] A. Grabowski and R. Kosinski, "Ising-based model of opinion formation in a complex network of interpersonal interactions," *Physica A: Statistical Mechanics and its Applications*, vol. 361, pp. 651–664, 2006.

[2] F. Vega-Redondo, *Complex social networks*. Cambridge Press, 2007.

[3] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society Series B*, vol. 48, pp. 259–279, 1986.

[4] M. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchial context on a large database of object categories," in *IEEE CVPR*, 2010.

[5] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, Feb 2004.

[6] A. Ahmedy, L. Song, and E. P. Xing, "Time-varying networks: Recovering temporally rewiring genetic networks during the life cycle of drosophila melanogaster," tech. rep., 2008. arXiv.

[7] T. Cover and J. Thomas, *Elements of Info. Theory*. Wiley Interscience, 2006.

[8] M. Jackson, *Social and economic Networks*. Princeton Univ. Press, 2008.

[9] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, pp. 644–652, 2008.

[10] S. Wu and X. Gu, "Gene network: Model, dynamics and simulation," *Computing and Combinatorics, 2005*, vol. 3595, pp. 12–21, 2005.

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *ACM SIGCOMM*, 1999.

[12] R. Tandon and P. Ravikumar, "On the difficulty of learning power law graphical models," in *IEEE ISIT*, 2013.

[13] B. Bollobas, *Random graphs*. Cambridge Studies in Advanced Mathematics, 2001.

[14] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," in *APPROX*, pp. 343–356, 2008.

[15] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *arXiv*, 2009.

[16] A. Anandkumar, V. Y. F. Tan, and A. Willsky, "High-dimensional structure learning of Ising models: Tractable graph families," *arXiv Preprint*, 2011.

[17] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai, "Greedy learning of markov network structure," in *IEEE Allerton*, 2010.

[18] A. Ray, S. Sanghavi, and S. Shakkottai, "Greedy learning of graphical models with small girth," in *IEEE Allerton*, 2012.

[19] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *IEEE ISIT*, 2010.

[20] P. Ravikumar and M. Wainwright, "High-dimensional ising model selection using l1-regularized logistic regression," *Annals of Statistics*, vol. 38, pp. 1287–1319.

[21] A. Anandkumar, V. Y. F. Tan, and A. Willsky, "High-dimensional Gaussian graphical model selection: Tractable graph families," *arXiv Preprint*, 2011.

[22] I. Mitliagkas and S. Vishwanath, "Strong information-theoretic limits for source/model recovery," in *Proc. of Allerton Conf. on Communication, Control and Computing, Monticello, USA*, 2010.

[23] F. Chung and L. Lu, *Complex graphs and networks*. American Mathematical Society, 2004.

[24] A. K. Das, P. Netrapalli, S. Sanghavi, and S. Vishwanath, "Learning structure of power-law Markov networks," *http://uts.cc.utexas.edu/∼akdas/plmrf.pdf*, 2014.

[25] M. Abdullah, C. Cooper, and A. Frieze, "Cover time of a random graph with given degree sequence," *Discrete Mathematics*, Nov 2012.

[26] R. Wu, R. Srikant, and J. Ni, "Learning graph structure in discrete Markov random fields," in *IEEE NetSciCom*, 2012.

[27] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using l1-regularized logistic regression," *Annals of Statistics*, 2008.

[28] A. Jalali, C. C. Johnson, and P. Ravikumar, "On learning discrete graphical models using greedy methods," in *NIPS*, 2012.