ISIT'2013 1569718735

On the Difficulty of Learning Power Law Graphical Models

Rashish Tandon
Department of Computer Science
University of Texas at Austin
TX, USA
Email: rashish@cs.utexas.edu

Pradeep Ravikumar
Department of Computer Science
University of Texas at Austin
TX, USA
Email: pradeepr@cs.utexas.edu

***Abstract*—A power-law graph is any graph $G = (V, E)$, whose degree distribution follows a power law *i.e.* the number of vertices in the graph with degree i, y_i, is proportional to $i^{-\beta}$: $y_i \propto i^{-\beta}$. In this paper, we provide information-theoretic lower bounds on the sample complexity of learning such power-law graphical models *i.e.* graphical models whose Markov graph obeys the power law. In addition, we briefly revisit some existing state of the art estimators, and explicitly derive their sample complexity for power-law graphs.**

I. Introduction

Undirected graphical models, also known as Markov Random Fields (MRFs), are useful tools for representing multivariate probability distributions. These models compactly represent a joint distribution using clique-wise functions over an undirected graph which captures the dependencies among the variables. The task of *graphical model selection* is to infer this underlying dependency graph (called the Markov graph) based on data drawn from the corresponding distribution. This task is especially difficult in *high-dimensional* settings where the number of observations, n is typically even smaller than the number of variables p. In this paper, we study the graphical model selection problem in the setting where the underlying graph structure follows a power-law. A power-law graph [1], [2] is any graph $G = (V, E)$, whose degree distribution follows a power law *i.e.* the number of vertices in the graph with degree i, y_i, is proportional to $i^{-\beta}$: $y_i \propto i^{-\beta}$. Here, $\beta > 1$ is some fixed constant, also called the power-law exponent of the graph. Power-law graphs have been seen to occur in several real-world scenarios e.g. internet graphs [3], biological networks [4] and several social networks [5].

The sample complexity of existing sparse model estimators [6], [7], [8], [9] scales at least linearly with the maximum node-degree, and for discrete models it even scales quadratically and cubically. For power-law graphs however, the maximum node degree is not bounded, and could be very large. Motivated by this, there have been a few M-estimators that explicitly target power-law graphical model estimation. [10] propose a novel *non-convex* regularization different from the ℓ_1 norm motivated by the power-law degree distribution, a convex variant of which was considered in [11], but their experimental results while better than ℓ_1 regularization based methods, demonstrated considerable room for improvements. In gist, there seems to be a lack of efficient estimators for power-law graphical model estimation, with clean guarantees as for degree-bounded graphs.

In this paper, we ask the following converse question: how difficult are power-law graphical models to learn? Specifically, our goal is to provide information-theoretic lower bounds on the sample complexity of learning such power-law models. Such bounds are needed not just to characterize difficulty of learning, but also to provide a sample complexity target for practical estimators, and to check whether existing estimators are already optimal.

Our main set of contributions are two sets of lower bounds. The first is for the broad class of power-law graphs, as specified by the standard (α, β) class of graphs [2]. Our second set of lower bounds are for graphs drawn from the Chung Lu model [12]. In addition, we also revisit some of the aforementioned state of the art estimators, and explicitly derive their sample complexity for power-law graphs. These results thus pose the outstanding question of efficient estimators for power-law graphical model selection in sharp relief.

II. Preliminaries

Suppose $X = (X_1, \ldots, X_p)$ is a random vector, with each variable X_s taking values in a set $\mathcal{X}$. Suppose $G = (V, E)$ is an undirected graph over p nodes corresponding to the p variables, and let $\mathcal{C}$ be a set of cliques (fully-connected subgraphs) of the graph G. A graphical model distribution over X, given the graph G, is then specified by a set of clique-wise functions $\{\psi_c(x_c),\ c \in \mathcal{C}\}$, and takes the form

$$\mathbb{P}(x; G) \quad \propto \quad \prod_{c \in \mathcal{C}} \psi_c(x_c)\,. \tag{1}$$

Note that x_c corresponds to the subset of variables in the clique c. We shall denote by P_G some canonical graphical model distribution with markov graph G. Let $\mathcal{G}$ denote the set of graphs over the p nodes, and suppose nature picks a graph G from this set according to some *specified* distribution over $\mathcal{G}$. We as the statistician observe n i.i.d. samples $\mathbf{X} = (X^{(1)}, \ldots, X^{(n)})$ drawn from the graphical model distribution P_G. The goal of the statistician is to come up with an efficient graph structure estimator $\phi : \mathcal{X}^{np} \to \mathcal{G}_p$, which is a function that takes the data as input, and has a graph in $\mathcal{G}$ as the output. Suppose we evaluate the estimator with the 0-1 loss,

$\mathbb{I}[\phi(\mathbf{X}) \neq G]$. The expected loss is then the probability of incorrect graph selection, $P[\phi(\mathbf{X}) \neq G]$. Our goal is to provide information-theoretic lower-bounds on this probability, and in particular provide a lower bound on the number of samples n, so that if the number of samples n is smaller than the lower bound, then this probability of error asymptotically goes to 1.

First, we review some standard random graph distributions used in power-law network analysis.

A. Standard Random Power Law Graph Models

In this section, we describe two *offline* models (*i.e.* models for a fixed number of nodes) for generating power law graphs.

1) Uniform Distribution and the Configuration Model: A natural way to study power-law graphs that satisfy degree requirements strictly is through (α, β)-graphs [13], which are defined using two parameters : α and β. α influences the number of vertices in the graph, and β is the power-law exponent. A graph $G = (V, E)$ is said to be an (α, β)-graph if it satisfies

$$y_i = \left\lfloor \frac{e^\alpha}{i^\beta} \right\rfloor \quad \forall\, i \in \{1, \ldots, \Delta\}, \tag{2}$$

where y_i is the number of vertices in G with degree i and Δ is the maximum degree. Then, $\Delta \leq \lfloor e^{\alpha/\beta} \rfloor$. Note that fixing an α and β fixes the number of nodes, p and the number of edges, m in an (α, β) graph as :

$$\begin{aligned} p &= \sum_{i=1}^{\Delta} y_i = \sum_{i=1}^{\Delta} \left\lfloor \frac{e^\alpha}{i^\beta} \right\rfloor, \\ m &= \frac{1}{2} \sum_{i=1}^{p} d_i = \frac{1}{2} \sum_{i=1}^{\Delta} i.y_i = \frac{1}{2} \sum_{i=1}^{\Delta} i \left\lfloor \frac{e^\alpha}{i^\beta} \right\rfloor. \end{aligned} \tag{3}$$

The following lemma establishes inequalities between p, m, α and β.

Lemma 1. *Given an (α, β)-graph $G = (V, E)$ with $|V| = p$ and $|E| = m$, we have*

$$\begin{aligned} &\textit{For } \beta > 1,\ p \leq \frac{e^\alpha \beta}{\beta - 1}. \\ &\textit{For } \beta > 2,\ m \leq \frac{1}{2} \frac{e^\alpha (\beta - 1)}{\beta - 2}. \end{aligned}$$

Also, $p \geq e^\alpha$ and $m \geq \frac{1}{2} e^\alpha$.

For any graph $G = (V, E)$ with $|V| = p$, a degree sequence $D = (d_1, d_2, \ldots, d_p)$ represents an assignment of degrees to the vertices of G *i.e.* the i^{th} vertex has degree d_i. A degree sequence D is said to be an (α, β)-sequence if any graph that is realized with this degree sequence is an (α, β)-graph. In other words, for D to be an (α, β)-sequence, for every $i \in \{1, \ldots, \Delta\}$, we must have i occurring in y_i $(= \lfloor e^\alpha / i^\beta \rfloor)$ number of positions in D. Note that, for a fixed α and β, the total number of different (α, β)-sequences is

$$\frac{p!}{\prod_{i=1}^{\Delta} (y_i!)}. \tag{4}$$

Let us denote by $\mathcal{G}_{\mathcal{U}}(\alpha, \beta)$, the uniform distribution over all (α, β)-*simple* graphs. Also, for a fixed (α, β)-sequence D, let us denote by $\mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$, the uniform distribution over all (α, β)-*simple* graphs that satisfy the specific degree distribution D. Ideally, we would like a model that generates a graph as per $\mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$ (and by a simple extension, as per $\mathcal{G}_{\mathcal{U}}(\alpha, \beta)$). However, this is not easy to do for any degree sequence D, and a common way to examine random graphs with given degree sequence D is through the Configuration model [14].

For any degree sequence D, the configuration model imposes a distribution on the set of all *multi-graphs* that satisfy the degree sequence D (and not just the set of *simple graphs*). For a fixed (α, β)-sequence D, we shall denote this distribution by $\mathcal{G}_{CM}(\alpha, \beta, D)$. The configuration model generates a random graph as follows : Given a degree sequence $D = (d_1, \ldots, d_p)$, consider a set S of $\sum_{i=1}^{p} d_i$ $(= 2m)$ points such that there are d_i distinct points corresponding to vertex i. Now, choose a perfect matching on the points in set S uniformly at random, and for each pair of matched points, draw an edge between their corresponding vertices.

Clearly, a graph constructed by this process may have loops and/or multiple edges, and therefore, can be a *multi-graph*. Note that even though each matching is picked above with equal probability $\left(\text{which is } \frac{m! 2^m}{(2m)!}\right)$, this does not imply a uniform distribution on the set of *multi-graphs*, since a different number of matchings could correspond to the same graph, depending on the multiplicity of the edges in the graph. However, the configuration model does impose the same probability on all simple graphs, as discussed below.

Let $\overline{G} \sim \mathcal{G}_{CM}(\alpha, \beta, D)$ and let N_G denote the number of matchings corresponding to a graph G. If G is simple, it can be seen that $N_G = \prod_{i=1}^{p} (d_i)!$, while for the case where G is not simple, $N_G \leq \prod_{i=1}^{p} (d_i)!$. Then, for a simple graph G,

$$P(\overline{G} = G) = \frac{m! 2^m}{(2m)!} N_G = \frac{m! 2^m}{(2m)!} \prod_{i=1}^{p} (d_i)!. \tag{5}$$

If G is not simple,

$$P(\overline{G} = G) = \frac{m! 2^m}{(2m)!} N_G \leq \frac{m! 2^m}{(2m)!} \prod_{i=1}^{p} (d_i)!. \tag{6}$$

2) Chung-Lu Model: The Chung-Lu model [12] is an extension of the Erdős-Rényi model [15] for producing random graphs with given *expected* degree sequences. An appropriate assignment of these expected degrees can promote power-law behaviour in the random graphs generated by this model. Note that this model is different from the Uniform distribution discussed in Section II-A1 in the sense that, this is a distribution over the space of *all* possible graphs, which simply encourages graphs with degree sequence close to the given *expected* degree sequence by assigning higher probabilities to those graphs. However, even among graphs with the same degree sequence, different graphs may be assigned different probabilities.

Given a set of labelled vertices V with $|V| = p$ and a sequence of expected degrees $\mathbf{w} = (w_1, w_2, \ldots, w_p)$, where the i^{th} vertex has expected degree w_i, the Chung-Lu model generates a random graph $G = (V, E)$ as follows: for any two nodes $i, j \in V$, the edge (i, j) occurs independently with

probability $\frac{w_i w_j}{\rho}$, where $\rho := \sum_k w_k$. It is assumed that $w_{\max}^2 \leq \rho$, for the probabilities to be valid. To enforce power-law behaviour, the expected degrees are taken as:

$$w_i = \alpha(i + \Delta_{\min} - 1)^{-\frac{1}{\beta-1}} \quad \forall\, i \in [p], \text{ with}$$
$$\alpha := \frac{(\beta-2)}{(\beta-1)} \overline{w} p^{\frac{1}{\beta-1}}, \qquad \Delta_{\min} = p\left(\frac{\overline{w}(\beta-2)}{\Delta_{\max}(\beta-1)}\right)^{\beta-1}, \tag{7}$$

where $\overline{w}$ is the average degree, $\Delta_{\min}$ is the minimum degree, $\Delta_{\max}$ is the maximum degree and $\beta > 2$ is the power-law exponent. $\Delta_{\min}$ is usually set to 1. For a fixed $p, \overline{w}$ and β, this enforces a restriction on $\Delta_{\max}$ as $\Delta_{\max} = \frac{\overline{w}(\beta-2)}{(\beta-1)} p^{1/(\beta-1)}$, and our expected degrees become

$$w_i = \alpha i^{-\frac{1}{\beta-1}} \quad \forall\, i \in [p]. \tag{8}$$

We shall denote by $\mathcal{G}_{CL}(p, \overline{w}, \beta)$, the distribution on *all* graphs for the Chung-Lu random graph model with $\Delta_{\min} = 1$. Note that in the extreme case of $\beta = \infty$ and $\overline{w} = c$ (constant), the Chung-Lu model, $\mathcal{G}_{CL}(p, \overline{w}, \beta)$ transforms to the Erdős-Rényi model [15], $\mathcal{G}_{ER}(p, c/p)$.

III. Lower Bounds

In this section, we present lower bounds on the sample complexity for graph estimation given samples from a discrete graphical model whose underlying graph is drawn from two specific distributions on power-law graphs : a Uniform distribution and the Chung-Lu model. The proofs of all Lemmas, Theorems and Propositions can be found in [16].

A. Uniform Distribution and the Configuration Model

A graph counting argument. One way to obtain a lower bound for estimation of a graph drawn from $\mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$ is to lower bound the total number of graphs in the space of $\mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$ and then, follow the approach in [17]. The space of graphs for $\mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$ is the set of all (α, β)-simple graphs that satisfy the degree sequence D. Let us denote this as $\mathcal{S}_{D,\alpha,\beta}$.

First, we present the main theorem for the lower bound. Recall that we let P_G to be some discrete graphical model distribution over $\mathcal{X}^p$ with Markov graph G and this gives a family of distributions $\{P_G \,|\, G \in \mathcal{S}_{D,\alpha,\beta}\}$. Also, note that we use $H(\cdot)$ to denote entropy in this and all subsequent sections.

Theorem 1. *Let $\beta > 3$. Let $G \sim \mathcal{G}_{\mathcal{U}}(\alpha, \beta, D)$ and let $\mathbf{X} = (X^{(1)}, \ldots, X^{(n)})$ be n i.i.d. samples drawn from P_G. Let $\phi : \mathcal{X}^{np} \to \mathcal{G}_p$ be any estimator. Then, there exists a constant $\epsilon > 0$, such that for $n \leq \epsilon \frac{\beta-1}{\beta \log|\mathcal{X}|} \log\left(\frac{(\beta-1)p}{\beta}\right)$*

$$P(\phi(\mathbf{X}) \neq G) \to 1 \text{ as } p \to \infty. \tag{9}$$

Proof Summary for Theorem 1: Using the approach of [17], we can show

$$P(\phi(X) \neq G) \geq 1 - \frac{|\mathcal{X}|^{np}}{|\mathcal{S}_{D,\alpha,\beta}|}. \tag{10}$$

So, for any constant δ s.t. $0 < \delta < 1$, if $n \leq \delta \frac{\log|\mathcal{S}_{D,\alpha,\beta}|}{p \log|\mathcal{X}|}$, then $P(\phi(X) \neq G) \to 1$ as $p \to \infty$. Now, if we have lower bound for $|\mathcal{S}_{D,\alpha,\beta}|$, we may substitute it here to get the bound for n. The approach for computing this lower bound is now stated below.

The lower bound on $|\mathcal{S}_{D,\alpha,\beta}|$, required for the above theorem, is obtained indirectly from the configuration model as follows. Consider $\overline{G} \sim \mathcal{G}_{CM}(\alpha, \beta, D)$. Let,

$$W = \begin{cases} 1 & \text{if } \overline{G} \text{ is simple} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Then, from Equation (5), for a simple graph $G \in \mathcal{S}_{D,\alpha,\beta}$, we have,

$$\begin{aligned} P(\overline{G} = G \,|\, W = 1) &= \frac{1}{|\mathcal{S}_{D,\alpha,\beta}|} = \frac{P(\overline{G} = G)}{P(W = 1)} \\ &= \frac{1}{P(W=1)} \frac{m! 2^m}{(2m)!} \prod_{i=1}^{p} (d_i)!. \end{aligned} \tag{12}$$

Thus, if we can lower bound $P(W = 1)$, then we also lower bound $|\mathcal{S}_{D,\alpha,\beta}|$. Luckily, from a result in [18], we know that if the maximum degree in D is $o(M_1(D)^{1/3})$, then,

$$P(W = 1) = e^{-O\left(\frac{M_2(D)^2}{M_1(D)^2}\right)}, \tag{13}$$

where $M_1(D)$ and $M_2(D)$ are the first and second moments of D respectively, *i.e.*,

$$M_1(D) = \sum_{i=1}^{p} d_i, \quad M_2(D) = \sum_{i=1}^{p} d_i^2. \tag{14}$$

For an (α, β)-sequence D, Lemma 1 relates $M_1(D)(= 2m)$ to α and β. Additionally, we have the following lemma that relates $M_2(D)$ to α and β.

Lemma 2. *If D is an (α, β)-sequence, then*

$$\text{For } \beta > 3,\ M_2(D) \leq \frac{e^{\alpha}(\beta-2)}{\beta-3}.$$

Also, $M_2(D) \geq e^{\alpha}$.

Note that the requirement of $\Delta = o(M_1(D)^{1/3})$ is also satisfied for $\beta > 3$, and thus in this case,

$$P(W = 1) = e^{-O\left(\frac{(\beta-2)^2}{(\beta-3)^2}\right)}. \tag{15}$$

Now, we have the following proposition.

Proposition 1. *For $\beta > 3$, there exists a constant $c > 0$ s.t.*

$$\log|\mathcal{S}_{D,\alpha,\beta}| \geq e^{\alpha}\alpha - \frac{e^{\alpha}[(2\beta-3)(\beta-2)+2]}{2(\beta-2)^2} - c\frac{(\beta-2)^2}{(\beta-3)^2}. \tag{16}$$

Writing the result in terms of p and ignoring lower order terms, we get

$$\log|\mathcal{S}_{D,\alpha,\beta}| = \Omega\left(\frac{(\beta-1)p}{\beta} \log\left(\frac{(\beta-1)p}{\beta}\right)\right). \tag{17}$$

Generalizing over all (α, β)-sequences. To obtain a lower bound for $\mathcal{G}_{\mathcal{U}}(\alpha, \beta)$, we need to lower bound the cardinality of $\mathcal{S}_{\alpha,\beta}$, the set of *all* simple (α, β)-graphs *i.e.*, (α, β)-graphs

over all possible degree sequences. However, obtaining this bound is easy once we have lower bounded $\mathcal{S}_{D,\alpha,\beta}$ since each (α,β)-sequence D gives a distinct set of (α,β)-graphs. Letting $\mathcal{D}$ to be the set of all (α,β)-sequences, we have :

$$|\mathcal{S}_{\alpha,\beta}| = \sum_{D\in\mathcal{D}} |\mathcal{S}_{\alpha,\beta,D}|. \tag{18}$$

Note that $|\mathcal{D}|$ is as specified in Equation (4). We obtain the same asymptotic lower bound as $|\mathcal{S}_{\alpha,\beta,D}|$, given in the following proposition.

Proposition 2. *For $\beta > 3$,*

$$\log|\mathcal{S}_{\alpha,\beta}| = \Omega\left(\frac{(\beta-1)p}{\beta}\log\left(\frac{(\beta-1)p}{\beta}\right)\right). \tag{19}$$

Thus, the lower bound on the sample complexity is the same, and is restated here for completeness.

Theorem 2. *Let $\beta > 3$. Let $G \sim \mathcal{G}_{\mathcal{U}}(\alpha,\beta)$ and let $\mathbf{X} = (X^{(1)},\ldots,X^{(n)})$ be n i.i.d. samples drawn from P_G. Let $\phi : \mathcal{X}^{np} \to \mathcal{G}_p$ be any estimator. Then, there exists a constant $\epsilon > 0$, such that for $n \le \epsilon\frac{\beta-1}{\beta\log|\mathcal{X}|}\log\left(\frac{(\beta-1)p}{\beta}\right)$,*

$$P(\phi(X) \neq G) \to 1 \text{ as } p \to \infty. \tag{20}$$

Argument using Fano's Lemma. A straightforward use of Fano's lemma [19] also gives us a converse result on the lower bound on sample complexity required to guarantee that the probability of error is upper bounded by some fraction. This is presented in the following theorem.

Theorem 3. *Let $\beta > 3$. Let $G \sim \mathcal{G}_{\mathcal{U}}(\alpha,\beta,D)$ and let $\mathbf{X} = (X^{(1)},\ldots,X^{(n)})$ be n i.i.d. samples drawn from P_G. Let $\phi : \mathcal{X}^{np} \to \mathcal{G}_p$ be any estimator. If $P(\phi(\mathbf{X}) \neq G) \le \delta$, we must have*

$$n = \Omega\left((1-\delta)\frac{\beta-1}{\beta\log|\mathcal{X}|}\log\left(\frac{(\beta-1)p}{\beta}\right)\right). \tag{21}$$

Proof Summary for Theorem 3: Using Fano's Lemma, we have

$$1 + P(\phi(\mathbf{X}) \neq G)\log|\mathcal{S}_{D,\alpha,\beta}| \ge H(G) - np\log|\mathcal{X}|, \tag{22}$$

so that $n \ge \frac{\log|\mathcal{S}_{D,\alpha,\beta}|}{p\log|\mathcal{X}|}(1-P(\phi(\mathbf{X}) \neq G))$. Now, the restriction on $P(\phi(\mathbf{X}) \neq G))$ and Proposition 1 gives the result.

This result can also be easily generalized to all (α,β)-sequences by considering $\mathcal{G}_{\mathcal{U}}(\alpha,\beta)$ instead of $\mathcal{G}_{\mathcal{U}}(\alpha,\beta,D)$.

B. Chung-Lu Model

$\mathcal{G}_{CL}(p,\overline{w},\beta)$ is a distribution over the set of all graphs $\{G^1,\ldots,G^M\}$, where $M = 2^{\binom{p}{2}}$. As earlier, we let P_G to be some discrete graphical model over $\mathcal{X}^p$ with Markov graph G. So, we have a family of distributions $\{P_{G^1},\ldots,P_{G^M}\}$. Now, the following theorem presents the lower bound on sample complexity for estimating G given samples from P_G :

Theorem 4. *Let $\beta > 2$. Let $G \sim \mathcal{G}_{CL}(p,\overline{w},\beta)$ and let $\mathbf{X} = (X^{(1)},\ldots,X^{(n)})$ be n i.i.d. samples drawn from P_G. Let $\phi : \mathcal{X}^{np} \to \mathcal{G}_p$ be any estimator. Then, if*

$$P_e = P(\phi(\mathbf{X}) \neq \theta) \le \frac{1}{p},$$

we must have

$$n = \Omega\left(\frac{\overline{w}}{\log|\mathcal{X}|}\log\left(\frac{(\beta-1)^2}{(\beta-2)^2\,\overline{w}}\right) + \left(\frac{\beta-2}{\beta-1}\right)\frac{\overline{w}}{\log|\mathcal{X}|}\log p\right).$$

Proof Summary for Theorem 4: Using Fano's Lemma, we have

$$1 + P_e\log|\mathcal{G}_p| \ge H(G) - np\log|\mathcal{X}|, \tag{23}$$

where $\mathcal{G}_p$ is the set of all graphs on p nodes and $|\mathcal{G}_p| = 2^{\binom{p}{2}}$. Now, if we can lower bound the entropy of $\mathcal{G}_{CL}(p,\overline{w},\beta)$, then this combined with the restriction on P_e gives us a lower bound as $n = \Omega\left(\frac{H(G)}{p\log|\mathcal{X}|}\right)$.

The lower bound on entropy is presented below in Proposition 3.

Proposition 3. *Let $\beta > 2$. Let $G \sim \mathcal{G}_{CL}(p,\overline{w},\beta)$. Then,*

$$H(G) = \Omega\left(\overline{w}p\log\left(\frac{(\beta-1)^2}{(\beta-2)^2\,\overline{w}}\right) + \left(\frac{\beta-2}{\beta-1}\right)\overline{w}p\log p\right). \tag{24}$$

Stating the lower bound in a simplified manner, we get that n must scale as $\Omega\left(\frac{\overline{w}}{\log|\mathcal{X}|}\log\left[p^{\frac{(\beta-2)}{(\beta-1)}}\right]\right)$, where $\overline{w}$ could possibly depend on p. In the extreme case of $\beta = \infty$ and $\overline{w} = c$ (constant), our lower bound becomes, $n = \Omega(\frac{c}{\log|\mathcal{X}|}\log p)$, which is the same as the lower bound for $\mathcal{G}_{ER}(p,c/p)$ reported in [9].

Modified Chung-Lu. The Chung-Lu model fixes the *expected* degree for each node *i.e.* the i^{th} node has expected degree w_i, and so, graphs that have actual degree close to w_i have a greater probability of occurrence. One may want to avoid this by picking a permutation $\sigma : [p] \to [p]$ uniformly randomly and then, picking the graph according to the distribution $\mathcal{G}^{\sigma}_{CL}(p,\overline{w},\beta)$, where the i^{th} vertex gets the weight $w_{\sigma(i)}$, and the weights are defined as in (8). Let us call this model the *average* Chung-Lu model, $\mathcal{G}^{A}_{CL}(p,\overline{w},\beta)$.

In this case, however, the same lower bound for entropy and therefore, the sample complexity, will hold. This can be seen as follows. Let p^A be the probability mass function for $\mathcal{G}^{A}_{CL}(p,\overline{w},\beta)$ and let p^{σ} be the probability mass function for $\mathcal{G}^{\sigma}_{CL}(p,\overline{w},\beta)$. Let S_p be the set of all permutations from $[p] \to [p]$. Then, for any graph G,

$$p^A(G) = \sum_{\sigma\in S_p}\frac{1}{p!}\,p^{\sigma}(G). \tag{25}$$

Note that $H(p^{\sigma})$ would be the same for all $\sigma \in S_p$. Also, Proposition 3 holds for each $\sigma \in S_p$. Now, by the concavity of entropy,

$$H(p^A) \ge \sum_{\sigma\in S_p}\frac{1}{p!}\,H(p^{\sigma}) = H(p^{\sigma}) \text{ (for any } \sigma). \tag{26}$$

Thus, the same lower bound on entropy holds for p^A (and therefore, the same lower bound for sample complexity for $\mathcal{G}^{A}_{CL}(p,\overline{w},\beta)$).

IV. Estimator Guarantees

In this section, we study the statistical guarantees available for certain state of the art estimators, for this problem of learning power-law graphical models. We examine guarantees for two estimators : the node-wise ℓ_1 regularization of [6] and, the thresholding based estimator of [9].

A. The ℓ_1-estimator

Neighbourhood selection via ℓ_1 based logistic regression [6] guarantees exact graph recovery for Ising models if the sample complexity scales as $n = O(d^3 \log p)$, where d is the maximum degree of the graph, and if the graphical model satisfies certain irrepresentability conditions. However, as seen below, this sample complexity is unsuitable for learning power-law graphical models in the high-dimensional setting.

For a graph G drawn from $\mathcal{G}_{\mathcal{U}}(\alpha, \beta)$, $G \sim \mathcal{G}_{\mathcal{U}}(\alpha, \beta)$, the maximum degree is exactly $\lfloor e^{\alpha/\beta} \rfloor$. Since $e^{\alpha} \leq p$, the maximum degree $d_{\max}(G)$ scales as $O(p^{1/\beta})$. Thus, the ℓ_1-estimator would require $O(p^{3/\beta} \log p)$ to work.

For a graph G drawn from $\mathcal{G}_{CL}(\bar{w}, p, \beta)$, $G \sim \mathcal{G}_{CL}(\bar{w}, p, \beta)$, it is possible to show (see [2]) that almost surely (with probability atleast $1 - 2p^{-0.2}$) every vertex $i \in [p]$ satisfies

$$|d_i - w_i| \leq 2(\sqrt{w_i \log p} + \log p). \tag{27}$$

So, almost surely, $d_{\max}(G) = \Theta\left(w_{max}\right) = \Theta\left(\overline{w} p^{\frac{1}{\beta-1}}\right)$. Thus, for most graphs drawn from the Chung-Lu model, the ℓ_1-estimator would require $O(\overline{w}^3 p^{\frac{3}{\beta-1}} \log p)$.

Recently, a greedy algorithm for neighbourhood estimation has also been proposed [7], that has an improved sample complexity of $O(d^2 \log p)$: for a power-law graph, this would mean sample complexities of $O(p^{2/\beta} \log p)$ and $O(\overline{w}^2 p^{\frac{2}{\beta-1}} \log p)$ for $\mathcal{G}_{\mathcal{U}}(\alpha, \beta)$ and $\mathcal{G}_{CL}(\bar{w}, p, \beta)$ respectively.

B. The CVDT estimator

Another recent state of the art estimator based on thresholding has been proposed by [9], and which they call the conditional variation distance thresholding (CVDT) estimator. Their estimator is targeted to estimating Ising models drawn from the graph ensemble satisfying the (η, γ)-local paths property *i.e.* between any two vertices, the number of paths of length atmost γ is atmost η.

The computational complexity of their procedure scales as $O(p^{\eta+2})$ and the sample complexity scales as $O((\frac{1}{\alpha})^{\gamma} \log p)$, where $\alpha < 1$ is a quantity that depends on the particular graph ensemble under consideration. In addition, for consistency, they require the assumption that $\alpha^{\gamma} = o(1/J_{\min})$, where $J_{\min}$ is a term that depends on the maximum node-degree through their assumptions. In the case of power law graphs, we can show that this latter condition would entail that $\gamma = \omega(\log p/\beta)$. Now, [9] have also shown that $\mathcal{G}_{CL}(\overline{w}, p, \beta)$ satisfies the (η, γ)-local paths property with high probability when $\overline{w} = o(p^{\frac{\eta-1}{2\eta\gamma} - \frac{2}{\beta-1}})$. However, assuming constant η to obtain a polynomial time complexity, the requirement of $\gamma = \omega(\log p)$ enforces $\overline{w} = o(1)$, and also leads to a sample complexity of $O(poly(p) \log p)$.

V. Conclusion

In this paper, we provided information-theoretic lower bounds on the sample complexity of learning the graph for two classes of power-law graphs, described through (α, β)-graphs and the Chung-Lu model. In addition, we looked at guarantees for two state-of-the-art estimators. For both of these, we observed that the sample complexity scales poorly with the number of nodes. In the light of our information-theoretic lower-bound results, this suggests that the task of deriving efficient methods for power-law structured graphical model selection is an outstanding open problem.

Acknowledgment

We acknowledge the support of NSF via IIS-1149803, and DoD via W911NF-12-1-0390.

References

[1] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[2] F. Chung and L. Lu, *Complex Graphs and Networks*. American Mathematical Society, Aug. 2006.

[3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 251–262, Aug. 1999.

[4] S. N. Dorogovtsev and J. F. F. Mendes, "Scaling properties of scale-free evolving networks: continuous approach," *Phys. Rev. E*, vol. 63, Apr 2001.

[5] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[6] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[7] A. Jalali, C. C. Johnson, and P. K. Ravikumar, "On learning discrete graphical models using greedy methods," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 1935–1943.

[8] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, vol. 34, pp. 1436–1462, 2006.

[9] A. Anandkumar, V. Y. F. Tan, and A. Willsky, "High-Dimensional Structure Learning of Ising Models : Local Separation Criterion," *Preprint*, June 2011.

[10] Q. Liu and A. T. Ihler, "Learning scale free networks by reweighted l1 regularization," *Journal of Machine Learning Research - Proceedings Track*, vol. 15, pp. 40–48, 2011.

[11] A. Defazio and T. Caetano, "A convex formulation for learning scale-free networks via submodular relaxation," in *Advances in Neural Information Processing Systems 24*, 2012.

[12] F. Chung and L. Lu, "Connected components in random graphs with given expected degree sequences," *Annals of Combinatorics*, vol. 6, no. 2, pp. 125–145, 2002.

[13] W. Aiello, F. Chung, and L. Lu, "A random graph model for power law graphs," *Experimental Math*, vol. 10, pp. 53–66, 2000.

[14] B. Bollobas, *Random Graphs*, W. Fulton, A. Katok, F. Kirwan, P. Sarnak, B. Simon, and B. Totaro, Eds. Cambridge University Press, 2001.

[15] P. Erdös and A. Rényi, "On the evolution of random graphs," *Evolution*, vol. 5, no. 1, pp. 17–61, 1960.

[16] R. Tandon and P. Ravikumar, "On the difficulty of learning power law graphical models : Proofs." [Online]. Available: http://www.cs.utexas.edu/~rashish/plgm_proofs.pdf

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

[18] B. D. McKay and N. C. Wormald, "Asymptotic enumeration by degree sequence of graphs with degress o(n$^{1/2}$)," *Combinatorica*, vol. 11, no. 4, pp. 369–382, 1991.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.