Learning Graphs from Linear Measurements: Fundamental Trade-offs and Applications

Tongxin Li, *Student Member, IEEE,* Lucien Werner, *Student Member, IEEE,* and Steven H. Low, *Fellow, IEEE*

***Abstract*—We consider a specific graph learning task: reconstructing a symmetric matrix that represents an underlying graph using linear measurements. We present a sparsity characterization for distributions of random graphs (that are allowed to contain *high-degree* nodes), based on which we study fundamental trade-offs between the number of measurements, the complexity of the graph class, and the probability of error. We first derive a necessary condition on the number of measurements. Then, by considering a three-stage recovery scheme, we give a sufficient condition for recovery. Furthermore, assuming the measurements are Gaussian IID, we prove upper and lower bounds on the (worst-case) sample complexity for both noisy and noiseless recovery. In the special cases of the uniform distribution on trees with n nodes and the Erdős-Rényi (n, p) class, the fundamental trade-offs are tight up to multiplicative factors with noiseless measurements. In addition, for practical applications, we design and implement a polynomial-time (in n) algorithm based on the three-stage recovery scheme. Experiments show that the heuristic algorithm outperforms basis pursuit on star graphs. We apply the heuristic algorithm to learn admittance matrices in electric grids. Simulations for several canonical graph classes and IEEE power system test cases demonstrate the effectiveness and robustness of the proposed algorithm for parameter reconstruction.**

***Index Terms*—Graph signal processing, sample complexity, network parameter reconstruction, information theory, sparse recovery**

I. Introduction

A. Background

SYMMETRIC matrices are ubiquitous in graphical models with examples such as the $(0, 1)$ adjacency matrix and the (generalized) Laplacian of an undirected graph. A major challenge in graph learning is inferring graph parameters embedded in those graph-based matrices from historical data or real-time measurements. In contrast to traditional statistical inference methods [1], [2], [3], model-based graph learning, such as physically-motivated models and graph signal processing (GSP) [4], takes advantage of additional data structures offered freely by nature. Among different measurement models for graph learning, linear models have been used and analyzed widely for different tasks, *e.g.,* linear structural equation models (SEMs) [5], [6], linear graph measurements [7], generalized linear cascade models [8], *etc*.

Despite extra efforts required on data collection, processing and storage, model-based graph learning often guarantees provable sample complexity, which is often significantly lower than the empirical number of measurements needed with traditional inference methods. In many problem settings, having computationally efficient algorithms with low sample complexity is important. One reason for this is that the graph parameters may change in a short time-scale, making sample complexity a vital metric to guarantee that the learning can be accomplished with limited measurements. Indeed many applications, such as real-time optimal power flow [9], [10], [11], real-time contingency analysis [12] and frequency control [13] in power systems *etc*., require data about the network that are time-varying. For example, the generations or net loads may change rapidly due to the proliferation of distributed energy resources. The topology and line parameters of the grid may be reconfigured to mitigate cascading failure [14]. Line switching has changed the traditional idea of a power network with a fixed topology, enabling power flow control by switching lines [15], *etc.* Hence analyzing fundamental limits of parameter reconstruction and designing graph algorithms that are efficient in both computational and sample complexity are important.

The number of measurements needed for reconstructing a graph Laplacian can be affected by various system parameters, such as data quality (distribution), physical laws, and graph structures. In particular, existing recovery algorithms often assume the graph to be recovered is in a specific class, *e.g.,* trees [1], sparse graphs [16], graphs with no high-degree nodes [17], with notable exceptions such as [18], which considers an empirical algorithm for topology identification. However, there is still a lack of understanding of sample complexity for learning general undirected graphs that may contain high-degree nodes, especially with measurements constrained naturally by a linear system.

In this work, we consider a general graph learning problem where the measurements and underlying matrix to be recovered can be represented as or approximated by a linear system. A *graph matrix* $\mathbf{Y}(G)$ with respect to an underlying graph G, which may have *high-degree* nodes (see Definition II.1) is defined as an $n \times n$ symmetric matrix with each nonzero (i, j)-th entry corresponding to an edge connecting node i and node j where $n \in \mathbb{N}_+$ is the number of nodes of the underlying *undirected* graph. The diagonal entries can be arbitrary. The measurements are summarized as two $m \times n$ $(1 \leq m \leq n)$ real or complex matrices $\mathbf{A}$ and $\mathbf{B}$ satisfying

$$\mathbf{A} = \mathbf{B}\mathbf{Y}(G) + \mathbf{Z} \tag{1}$$

where $\mathbf{Z}$ denotes additive noise.

We focus on the following problems:

- *Fundamental Trade-offs*. What is the *minimum number* m of linear measurements required for reconstructing the

Li, Werner and Low are with the Computing + Mathematical Sciences Department, California Institute of Technology, Pasadena, CA 91125 USA (e-mails: {tongxin, lwerner, slow}@caltech.edu)

symmetric matrix $\mathbf{Y}(G)$? Is there an algorithm *asymptotically achieving* recovery with the minimum number of measurements? As a special case, can we characterize the sample complexity when the measurements are Gaussian IID[1]?

- *Applications to Electrical Grids*. Do the theoretical guarantees on sample complexity result in a practical algorithm (in terms of both sample and computational complexity) for recovering electric grid topology and parameters?

Some comments about the above model and the results in this paper are as follows.

***Remark* 1.** It has been noted that vectorization and standard compressed sensing techniques do not lead to straightforward results (see [17] for detailed arguments about a similar linear system). This issue is discussed extensively in Section I-B3.

***Remark* 2.** The results in this paper do not assume low-degree nodes as most of existing results do, with notable exceptions such as [18] which gives empirical and data-based subroutines for topology identification.

B. Related Work

1) Graph Learning: Algorithms for learning sparse graphical model structures have a rich tradition in the literature. For general Markov random fields (MRFs), learning the underlying graph structures is known to be NP-hard [19]. However, in the case when the underlying graph is a tree, the classical Chow-Liu algorithm [1] offers an efficient approach to structure estimation. Recent results contribute to an extensive understanding of the Chow-Liu algorithm. The authors in [3] analyzed the error exponent and showed experimental results for chain graphs and star graphs. For pairwise binary MRFs with bounded maximum degree, [20] provides sufficient conditions for correct graph selection. Similar achievability results for Ising models are in [21]. Model-based graph learning has been emerging recently and assuming the measurements form linear SEMs, the authors in [5], [6] showed theoretical guarantees of the sample complexity for learning a directed acyclic graph (DAG) structure, under mild conditions on the class of graphs.

For converse, information-theoretic tools have been widely applied to derive fundamental limits for learning graph structures. For a Markov random field with bounded maximum degree, necessary conditions on the number of samples for estimating the underlying graph structure were derived in [20] using Fano's inequality (see [22]). For Ising models, [23] combines Fano's inequality with the idea of *typicality* to derive weak and strong converse. Similar techniques have also been applied to Gaussian graphical models [24] and Bayesian networks [25]. Fundamental limits for noisy compressed sensing have been extensively studied in [26] under an information-theoretic framework.

2) System Identification in Power Systems: Graph learning has been widely used in electric grids applications, such as state estimation [27], [28] and topology identification [29], [30]. Most of the literature focuses on topology identification or change detection, but there is less work on joint topology and parameter reconstruction, with notable exceptions of [31], [32], [33], [34]. However, the linear system proposed in [32] does not leverage the sparsity of the graph[2]. Thus, in the worst case, the matrix $\mathbf{B}$ needs to have full column rank, implying that $m = \Omega(n)$ measurements are necessary for recovery.

Moreover, there is little exploration on the fundamental performance limits (estimation error and sample complexity) on topology and parameter reconstruction of power networks, with the exception of [35] where a sparsity condition was given for exact recovery of outage lines. Based on single-type measurements (either current or voltage), correlation analysis has been applied for topology identification [36], [37], [38]. Approximating the measurements as normal distributed random variables, the authors of [29] proposed an approach for topology identification with limited measurements. A graphical learning-based approach can be found in [39]. Recently, data-driven methods were studied for parameter estimation [33]. In [32], a similar linear system as (6) was used combined with regression to recover the symmetric graph parameters (which is the admittance matrix in the power network).

3) Compressed Sensing and Sketching: It is well known that compressed sensing ([40], [41]) techniques allow for recovery of a sparse matrix with a limited number of measurements in various applications such as medical imaging [42], wireless communication [43], channel estimation [44] and circuit design [45], *etc*. For electricity grids, in [46], based on these techniques, experimental results have been given for topology recovery. However, nodal admittance matrices (generalized Laplacians) for power systems have two properties for which there are gaps in the sparse recovery literature: 1) the presence of high-degree nodes in a graph (corresponding to dense columns in its Laplacian) and 2) symmetry.

Consider a vectorization of system (1) using tensor product notation, with $\mathbf{a} := \mathrm{vec}(\mathbf{A})$ and $\mathbf{y}(G) := \mathrm{vec}(\mathbf{Y}(G))$. Then linear system (1) is equivalent to $\mathbf{a} = (\mathbf{I} \otimes \mathbf{B})\mathbf{y}(G)$ where $\mathrm{vec}(\cdot)$ produces a column vector by stacking the columns of the input matrix and $\mathbf{I}\otimes\mathbf{B}$ is the Kronecker product of an identity matrix $\mathbf{I} \in \mathbb{R}^{n\times n}$ and $\mathbf{B}$. With the sensing matrix being a Kronecker product of two matrices, traditional compressed sensing analysis works for the case when $\mathbf{y}$ contains only $\mu = \Theta(1)$ non-zeros [47]. For instance, the authors of [48] showed that the restricted isometry constant (see Section III-B for the definition), $\delta_\mu(\mathbf{I}\otimes\mathbf{B})$ is bounded from above by $\delta_\mu(\mathbf{B})$, the restricted isometry constant of $\mathbf{B}$. However, if a column (or row) of $\mathbf{Y}(\mathbf{G})$ is dense, classical restricted isometry-based approach cannot be applied straightforwardly.

Another way of viewing it is that vectorizing $\mathbf{A}$ and $\mathbf{Y}(G)$ and constructing a sensing matrix $\mathbf{I} \otimes \mathbf{B}$ is equivalent to recovering each of the column (or row) of $\mathbf{Y}(G)$ separately from $A_j = \mathbf{B}Y_j(G)$ for $j = 1, \ldots, n$ where A_j's and $Y_j(G)$'s are columns of $\mathbf{A}$ and $\mathbf{Y}(G)$. For a general "sparse" graph G, such as a star graph, some of the columns (or rows) of the graph matrix $\mathbf{Y}(G)$ may be dense vectors consisting of

[1]This means the entries of the matrix $\mathbf{B}$ are IID normally distributed.

[2]With respect to sparsity, we consider not only graphs with bounded degrees, but a broader class of graphs which may contain high-degree nodes. Definition III.1 gives a comprehensive characterization of sparsity.

many non-zeros. The results in [48], [47] give no guarantee for the recovery of the dense columns of $\mathbf{Y}(G)$ (correspondingly, the high-degree nodes in G), and thus they cannot be applied directly to the analysis of sample complexity. This statement is further validated in our experimental results shown in Figure 4 and Figure 5.

The authors of [17] considered the recovery of an unknown sparse matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ (not necessarily symmetric) from an $m \times m$ matrix $\overline{\mathbf{A}} = \overline{\mathbf{B}}\mathbf{M}\overline{\mathbf{C}}^T$ where $\overline{\mathbf{B}} \in \mathbb{R}^{m\times n}$ and $\overline{\mathbf{C}} \in \mathbb{R}^{m\times n}$ with $m \ll n$. By adding a symmetry constraint to their recovery formulation, we obtain the following modified basis pursuit as a convex optimization:

$$\text{minimize } ||\mathbf{Y}(G)||_1 \tag{2}$$
$$\text{subject to } \mathbf{B}\mathbf{Y}(G) = \mathbf{A}, \tag{3}$$
$$\mathbf{Y}(G) \in \mathbb{S}^{n\times n} \tag{4}$$

where $||\mathbf{Y}(G)||_1 = ||\text{vec}(\mathbf{Y}(G))||_1$ is the entry-wise ℓ_1-norm of $\mathbf{Y}(G)$ and $\mathbb{S}^{n\times n}$ denotes the set of all symmetric matrices in $\mathbb{R}^{n\times n}$. However, the approach in [17] does not carry through to our setting for two reasons. First, the analysis of such an optimization often requires stronger assumptions, *e.g.,* the non-zeros are not concentrated in any single column (or row) of $\mathbf{Y}(G)$, as in [17]. Second, having the symmetry property of $\mathbf{Y}$ as a constraint does not explicitly make use of the fact that many columns in $\mathbf{Y}$ are indeed sparse and can be recovered correctly. As a consequence, basis pursuit may produce poor results in certain scenarios where our approach performs well, as demonstration in our experimental results on star graphs in Section VI-B4.

Although the columns of $\mathbf{Y}(G)$ are correlated because of the symmetry, in general there are no constraints on the support sets of the columns. Thus distributed compressed sensing schemes (for instance, [49] requires the columns to share the same support set) are not directly applicable in this situation.

The previous studies and aforementioned issues together motivate us to propose a novel three-stage recovery scheme for the derivation of a sufficient recovery condition, which leads to a practical algorithm that is sample and computationally efficient as well as robust to noise.

C. Our Contributions

We demonstrate that the linear system in (1) can be used to learn the topology and parameters of a graph. Our framework can be applied to perform system identification in electrical grids by leveraging synchronous nodal current and voltage measurements obtained from phasor measurement units (PMUs).

Compared to existing methods and analysis, the main results of this paper are three-fold:

1) *Fundamental Trade-offs*: In Theorem III.1, we derive a general lower bound on the *probability of error* for topology identification (defined in (7)). In Section III-C, we describe a simple three-stage recovery scheme combining ℓ_1-norm minimization with an additional step called *consistency-checking*, rendering which allows us to bound the number of measurements for exact recovery from above as in Theorem III.2.
2) *(Worst-case) Sample Complexity*: We provide sample complexity results for recovering a random graph that may contain *high-degree* nodes. The unknown distribution that the graph is sampled from is characterized based on the definition of "(μ, K, ρ)-*sparsity*" (see Definition III.1). Under the assumption that the matrix $\mathbf{B}$ has Gaussian IID entries, in Section IV, we provide upper and lower bounds on the worst-case sample complexity in Theorem IV.1. We show two applications of Theorem IV.1 for the uniform sampling of trees and the Erdős-Rényi (n, p) model in Corollary IV.1 and IV.2, respectively.
3) *(Heuristic) Algorithm*: Motivated by the three-stage recovery scheme, a heuristic algorithm with polynomial (in n) running-time is reported in Section V, together with simulation results for power system test cases validating its performance in Section VI.

Some comments about the above results are as follows:

D. Outline of the Paper

The remaining content is organized as follows. In Section II, we specify our models. In Section III-A, we present the converse result as fundamental limits for recovery. The achievability is provided in III-C. We present our main result as the worst-case sample complexity for Gaussian IID measurements in Section IV. A heuristic algorithm together with simulation results are reported in Section V and VI.

II. Model and Definitions

A. Notation

Let $\mathbb{F}$ denote a field that can either be the set of real numbers $\mathbb{R}$, or the set of complex numbers $\mathbb{C}$. The set of all symmetric $n\times n$ matrices whose entries are in $\mathbb{F}$ is denoted by $\mathbb{S}^{n\times n}$. The imaginary unit is denoted by j. Throughout the work, let $\log(\cdot)$ denote the binary logarithm with base 2 and let $\ln(\cdot)$ denote the natural logarithm with base e. We use $\mathbb{E}[\cdot]$ to denote the expectation of random variables. The mutual information is denoted by $\mathbb{I}(\cdot)$. The entropy function (either differential or discrete) is denoted by $\mathbb{H}(\cdot)$ and in particular, we reserve $h(\cdot)$ for the binary entropy function. To distinguish random variables and their realizations, we follow the convention and denote the former by capital letters (*e.g.,* A) and the latter by lower case letters (*e.g.,* a). The symbol C is used to designate a constant.

Matrices are denoted in boldface (*e.g.,* $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Y}$). The i-th row, the j-th column and the (i, j)-th entry of a matrix $\mathbf{A}$ are denoted by $A^{(i)}$, A_j and $A_{i,j}$ respectively. For notational convenience, let $\mathcal{S}$ be a subset of $\mathcal{V}$. Denote by $\overline{\mathcal{S}} := \mathcal{V}\backslash\mathcal{S}$ the complement of $\mathcal{S}$ and by $\mathbf{A}_{\mathcal{S}}$ a sub-matrix consisting of $|\mathcal{S}|$ columns of the matrix $\mathbf{A}$ whose indices are chosen from $\mathcal{S}$. The notation $\top$ denotes the transpose of a matrix, $\det(\cdot)$ calculates its determinant. For the sake of notational simplicity, we use big O notation (o,ω,O,Ω,Θ) to quantify asymptotic behavior.

B. Graphical Model

Denote by $\mathcal{V} = \{1, \dots, n\}$ a set of n nodes and consider an *undirected* graph $G = (\mathcal{V}, \mathcal{E})$ (with no self-loops) whose edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains the desired topology information. The degree of each node j is denoted by d_j. The connectivity between the nodes is unknown and our goal is to determine it by learning the associated *graph matrix* using linear measurements.

***Definition* II.1** (Graph matrix)**.** Provided with an underlying graph $G = (\mathcal{V}, \mathcal{E})$, a *symmetric* matrix $\mathbf{Y}(G) \in \mathbb{S}^{n\times n}$ is called a *graph matrix* if the following conditions hold:

$$Y_{i,j}(G) = \begin{cases} \neq 0 & \text{if } i \neq j \text{ and } (i,j) \in \mathcal{E} \\ 0 & \text{if } i \neq j \text{ and } (i,j) \notin \mathcal{E} \\ \text{arbitrary} & \text{otherwise} \end{cases}.$$

***Remark* 3.** Our theorems can be generalized to recover a broader class of symmetric matrices, as long as the matrix to be recovered satisfies (1) Knowing $\mathbf{Y}(G) \in \mathbb{F}^{n\times n}$ gives the full knowledge of the topology of G; (2) The number of non-zero entries in a column of $\mathbf{Y}(G)$ has the same order as the degree of the corresponding node, *i.e.,* $|\mathrm{supp}(Y_j)| = O(d_j)$. for all $j \in \mathcal{V}$. To have a clear presentation, we consider specifically the case $|\mathrm{supp}(Y_j)| = d_j$.

In this work, we employ a probabilistic model and assume that the graph G is chosen randomly from a *candidacy set* $\mathsf{C}(n)$ (with n nodes), according to some distribution $\mathcal{G}_n$. Both the candidacy set $\mathsf{C}(n)$ and distribution $\mathcal{G}_n$ are not known to the estimator. For simplicity, we often omit the subscripts of $\mathsf{C}(n)$ and $\mathcal{G}_n$.

***Example* II.1.** We exemplify some possible choices of the candidacy set and distribution:

(a) *(Mesh Network)* When G represents a transmission (mesh) power network and no prior information is available, the corresponding candidacy set $\mathsf{G}(n)$ consisting of all graphs with n nodes and G is selected uniformly at random from $\mathsf{G}(n)$. Moreover, $|\mathsf{G}(n)| = 2^{\binom{n}{2}}$ in this case.

(b) *(Radial Network)* When G represents a distribution (radial) power network and no other prior information is available, then the corresponding candidacy set $\mathsf{T}(n)$ is a set containing all spanning trees of the complete graph with n buses (nodes) and G is selected uniformly at random from $\mathsf{T}(n)$; the cardinality is $|\mathsf{T}(n)| = n^{n-2}$ by Cayley's formula.

(c) *(Radial Network with Prior Information)* When $G = (\mathcal{V}, \mathcal{E})$ represents a distribution (radial) power network, and we further know that some of the buses cannot be connected (which may be inferred from locational/geographical information), then the corresponding candidacy set $\mathsf{T}_H(n)$ is a set of spanning trees of a subgraph $H = (\mathcal{V}, \mathcal{E}_H)$ with n buses. An edge $e \notin \mathcal{E}_H$ if and only if we know $e \notin \mathcal{E}$. The size of $\mathsf{T}_H(n)$ is given by Kirchhoff's matrix tree theorem (*c.f.* [50]).

(d) *(Erdős-Rényi (n,p) model)* In a more general setting, G can be a random graph chosen from an ensemble of graphs according to a certain distribution. When a graph G is sampled according to the Erdős-Rényi (n,p) model, each edge of G is connected IID with probability p. We denote the corresponding graph distribution for this case by $\mathcal{G}_{\mathrm{ER}}(n,p)$.

The next section is devoted to describing available measurements.

C. Linear System of Measurements

Suppose the measurements are sampled discretely and indexed by the elements of the set $\{1, \dots, m\}$. As a general framework, the measurements are collected in two matrices $\mathbf{A}$ and $\mathbf{B}$ and defined as follows.

***Definition* II.2** (Generator and measurement matrices)**.** Let m be an integer with $1 \leq m \leq n$. The *generator matrix* $\mathbf{B}$ is an $m \times n$ *random* matrix and the *measurement matrix* $\mathbf{A}$ is an $m \times n$ matrix with entries selected from $\mathbb{F}$ that satisfy the linear system (1):

$$\mathbf{A} = \mathbf{B}\mathbf{Y}(G) + \mathbf{Z}$$

where $\mathbf{Y}(G) \in \mathbb{S}^{n\times n}$ is a graph matrix to be recovered, with an underlying graph G and $\mathbf{Z} \in \mathbb{F}^{m\times n}$ denotes the random *additive noise*. We call the recovery *noiseless* if $\mathbf{Z} = \mathbf{0}$. Our goal is to resolve the matrix $\mathbf{Y}(G)$ based on given matrices $\mathbf{A}$ and $\mathbf{B}$.

In the remaining contexts, we sometime simplify the matrix $\mathbf{Y}(G)$ as $\mathbf{Y}$ if there is no confusion.

D. Applications to Electrical Grids

Various applications fall into the framework in (1). Here we present two examples of the graph identification problem in power systems. The measurements are modeled as time series data obtained via nodal sensors at each node, *e.g.,* PMUs, smart switches, or smart meters.

1) Example 1: Nodal Current and Voltage Measurements: We assume data is obtained from a short time interval over which the unknown parameters in the network are *time-invariant*. $\mathbf{Y} \in \mathbb{C}^{n\times n}$ denotes the *nodal admittance matrix* of the network and is defined

$$Y_{i,j} := \begin{cases} -y_{i,j} & \text{if } i \neq j \\ y_i + \sum_{k\neq i} y_{i,k} & \text{if } i = j \end{cases} \tag{5}$$

where $y_{i,j} \in \mathbb{C}$ is the admittance of line $(i,j) \in \mathcal{E}$ and y_i is the self-admittance of bus i. Note that if two buses are not connected then $Y_{i,j} = 0$.

The corresponding generator and measurement matrices are formed by simultaneously measuring both current (or equivalently, power injection) and voltage at each node and at each time step. For each $t = 1, \dots, m$, the nodal current injection is collected in an n-dimensional random vector $I_t = (I_{t,1}, \dots, I_{t,n})$. Concatenating the I_t into a matrix we get $\mathbf{I} := [I_1, I_2, \dots, I_m]^\top \in \mathbb{C}^{m\times n}$. The generator matrix $\mathbf{V} := [V_1, V_2, \dots, V_m]^\top \in \mathbb{C}^{m\times n}$ is constructed analogously. Each pair of measurement vectors (I_t, V_t) from $\mathbf{I}$ and $\mathbf{V}$ must satisfy Kirchhoff's and Ohm's laws,

$$I_t = \mathbf{Y}V_t, \quad t = 1, \dots, m. \tag{6}$$

In matrix notation (6) is equivalent to $\mathbf{I} = \mathbf{VY}$, which is a noiseless version of the linear system defined in (1).

Compared with only obtaining one of the current, power injection or voltage measurements (for example, as in [36], [3], [37]), collecting simultaneous current-voltage pairs doubles the amount of data to be acquired and stored. There are benefits however. First, exploiting the physical law relating voltage and current not only enables us to identify the topology of a power network but also recover the parameters of the admittance matrix. Furthermore, dual-type measurements significantly reduce the sample complexity for learning the graph, compared with the results for single-type measurements.

2) Example 2: Nodal Power Injection and Phase Angles: Similar to the previous example, at each time $t = 1, \ldots, m$, denote by $P_{t,j}$ and $\theta_{t,j}$ the active nodal power injection and the phase of voltage at node j respectively. The matrices $\mathbf{P} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{\theta} \in \mathbb{R}^{m\times n}$ are constructed in a similar way by concatenating the vectors $P_t = (P_{t,1}, \ldots, P_{t,n})$ and $\theta_t = (\theta_{t,1}, \ldots, \theta_{t,n})$. The matrix representation of the DC power flow model can be expressed as a linear system $\mathbf{P} = \boldsymbol{\theta}\mathbf{CSC}^\top$, which belongs to the general class represented in (1). Here, the diagonal matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ is the susceptence matrix whose e-th diagonal entry represents the susceptence on the e-th edge in $\mathcal{E}$ and $\mathbf{C} \in \{-1, 0, 1\}^{n\times|\mathcal{E}|}$ is the node-to-link incidence matrix of the graph. The vertex-edge incidence matrix[3] $\mathbf{C} \in \{-1, 0, 1\}^{n\times|\mathcal{E}|}$ is defined as

$$C_{j,e} := \begin{cases} 1, & \text{if bus } j \text{ is the source of } e \\ -1, & \text{if bus } j \text{ is the target of } e \\ 0, & \text{otherwise} \end{cases}.$$

Note that $\mathbf{CSC}^\top$ specifies both the network topology and the susceptences of power lines.

E. Probability of Error as the Recovery Metric

We define the error criteria considered in this paper. We refer to finding the edge set $\mathcal{E}$ of G via matrices $\mathbf{A}$ and $\mathbf{B}$ as the *topology identification problem* and recovering the graph matrix $\mathbf{Y}$ via matrices $\mathbf{A}$ and $\mathbf{B}$ as the *parameter reconstruction problem*.

***Definition* II.3.** Let f be a function or algorithm that returns an estimated graph matrix $\mathbf{X} = f(\mathbf{A}, \mathbf{B})$ given inputs $\mathbf{A}$ and $\mathbf{B}$. The *probability of error for topology identification* ε_{T} is defined to be the probability that the estimated edge set is not equal to the correct edge set:

$$\varepsilon_{\mathrm{T}} := \mathbb{P}\left(\exists\ i \neq j \ \middle|\ \operatorname{sign}(X_{i,j}) \neq \operatorname{sign}\left(Y_{i,j}(G)\right)\right) \tag{7}$$

where the probability is taken over the randomness in $G, \mathbf{B}$ and $\mathbf{Z}$. The *probability of error for parameter reconstruction* $\varepsilon_{\mathrm{P}}(\eta)$ is defined to be the probability that the Frobenius norm of the difference between the estimate $\mathbf{X}$ and the original graph matrix $\mathbf{Y}(G)$ is larger than $\eta > 0$:

$$\varepsilon_{\mathrm{P}}(\eta) := \sup_{\mathbf{Y}\in\mathsf{Y}(G)} \mathbb{P}\left(||\mathbf{X} - \mathbf{Y}(G)||_{\mathrm{F}} > \eta\right) \tag{8}$$

[3]Although the underlying network is a directed graph, when considering the fundamental limit for topology identification, we still refer to the recovery of an undirected graph G.

where $||\cdot||_{\mathrm{F}}$ denotes the Frobenius norm, $\eta > 0$ and $\mathsf{Y}(G)$ is the set of all graph matrices $Y(G)$ that satisfy Definition II.1 for the underlying graph G, and the probability is taken over the randomness in G, $\mathbf{B}$ and $\mathbf{Z}$. Note that for noiseless parameter reconstruction, *i.e.,* $\mathbf{Z} = 0$, we always consider exact recovery and set $\eta = 0$ and abbreviate the probability of error as ε_{P}.

III. Fundamental Trade-offs

We discuss fundamental trade-offs of the parameter reconstruction problem defined in Section II-B and II-C. The converse result is summarized in Theorem III.1 as an inequality involving the probability of error, the distributions of the underlying graph, generator matrix and noise. Next, in Section III-C, we focus on a particular three-stage scheme, and show in Theorem III.2 that under certain conditions, the probability of error is asymptotically zero (in n).

A. Necessary Conditions

The following theorem states the fundamental limit.

***Theorem* III.1** (Converse)**.** *The probability of error for topology identification ε_{T} is bounded from below as*

$$\varepsilon_{\mathrm{T}} \geq 1 - \frac{\mathbb{H}(\mathbf{A}) - \mathbb{H}(\mathbf{Z}) + \ln 2}{\mathbb{H}(\mathcal{G}_n)} \tag{9}$$

where $\mathbb{H}(\mathbf{A})$, $\mathbb{H}(\mathbf{Z})$ are differential entropy (in base e) functions of the random variables $\mathbf{A}$, $\mathbf{Z}$ respectively and $\mathbb{H}(\mathcal{G}_n)$ is the entropy (in base e) of the probability distribution $\mathcal{G}_n$.

***Remark* 4.** It can be inferred from the theorem that $\varepsilon_{\mathrm{T}} = 1 - O(mn/\mathbb{H}(\mathcal{G}_n))$, given that the generator matrix $\mathbf{B}$ has Gaussian IID entries and the noise $\mathbf{Z}$ is additive white Gaussian (see Lemma 3). Therefore, the structure of the graphs reflected in the corresponding entropy of the graph distribution determines the number of samples needed. Consider the four cases listed in Example II.1. The number of samples must be at least linear in n (size of the graph) to ensure a small probability of error, given that the graph, as a mesh network, is chosen uniformly at random from $\mathsf{C}(n)$ (see Example II.1 (a)) since $\mathbb{H}(\mathcal{U}_{\mathsf{G}(n)}) = \binom{n}{2}$. On the other hand, as corollaries, under the assumptions of Gaussian IID measurements, $m = \Omega(\log n)$ is *necessary* for making the probability of error less or equal to $1/2$, if the graph is chosen uniformly at random from $\mathsf{T}(n)$; $m = \Omega(nh(p))$ is *necessary* if the graph is sampled according to $\mathcal{G}_{\mathrm{ER}}(n, p)$, as in Examples II.1 (b) and (c), respectively. The theorem can be generalized to complex measurements by adding additional multiplicative constants.

Note that $\varepsilon_{\mathrm{P}} \geq \varepsilon_{\mathrm{T}}$ for any fixed noiseless parameter reconstruction algorithm, the necessary conditions work for both topology and (noiseless) parameter reconstruction. The proof is postponed to Appendix A and the key steps are first applying the generalized Fano's inequality (see [22], [26]) and then bounding the mutual information $\mathbb{I}(G; \mathbf{A}|\mathbf{B})$ from above by $\mathbb{H}(\mathbf{A}) - \mathbb{H}(\mathbf{Z})$. The general converse stated in Theorem III.1 is used in asserting the results on worst-case

sample complexity in Theorem IV.1. Next, we analyze the sufficient condition for recovering a graph matrix $\mathbf{Y}(G)$. Before proceeding to the results, we introduce a novel characterization of the distribution $\mathcal{G}_n$, from which a graph G is sampled. In particular, the graph G is allowed to have high-degree nodes.

B. Characterization of Graph Distributions

Let $d_j(G)$ denote the degree of node $j \in \mathcal{V}$ in G. Denote by $\mathcal{V}_{\text{Large}}(\mu) := \{j \in \mathcal{V} \mid d_j(G) > \mu\}$ the set of nodes having degrees greater than the *threshold parameter* $0 \leq \mu \leq n-2$ and $\mathcal{V}_{\text{Small}}(\mu) := \mathcal{V} \backslash \mathcal{V}_{\text{Large}}(\mu)$ the set of nodes for all μ-sparse column vectors of $\mathbf{Y}$. With a *counting parameter* $0 \leq K \leq n$, we define a set of graphs wherein each graph consists of no more than K nodes with degree larger than μ, denoted by $\mathsf{C}(n, \mu, K) := \{G \in \mathsf{C}(n) \mid |\mathcal{V}_{\text{Large}}(\mu)| \leq K\}$. The following definition characterizes graph distributions.

***Definition* III.1** ((μ, K, ρ)-sparse distribution)**.** A graph distribution $\mathcal{G}_n$ is said to be (μ, K, ρ)-*sparse* if assuming that G is distributed according to $\mathcal{G}_n$, then the probability that G belongs to $\mathsf{C}(n, \mu, K)$ is larger than $1-\rho$, *i.e.,*

$$\mathbb{P}_{\mathcal{G}_n}\left(G \notin \mathsf{C}(n, \mu, K)\right) \leq \rho. \tag{10}$$

1) Uniform Sampling of Trees:

Based on the definition above, for particular graph distributions, we can find the associated parameters. We exemplify by considering two graph distributions introduced in Example II.1. Denote by $\mathcal{U}_{\mathsf{T}(n)}$ the uniform distribution on the set $\mathsf{T}(n)$ of all trees with n nodes.

***Lemma* 1.** *For any $\mu \geq 1$ and $K > 0$, the distribution $\mathcal{U}_{\mathsf{T}(n)}$ is $(\mu, K, 1/K)$-sparse.*

2) Erdős-Rényi (n, p) model:

Denote by $\mathcal{G}_{\text{ER}}(n, p)$ the graph distribution for the Erdős-Rényi (n, p) model. Similarly, the lemma below classifies $\mathcal{G}_{\text{ER}}(n, p)$ into a (μ, K, ρ)-sparse distribution with appropriate parameters.

***Lemma* 2.** *For any $\mu(n, p) \geq 2nh(p)/(\ln 1/p)$ and $K > 0$, the distribution $\mathcal{G}_{\text{ER}}(n, p)$ is $(\mu, K, n\exp(-nh(p))/K)$-sparse.*

The proofs of Lemmas 1 and 2 are in Appendix D.

***Remark* 5.** It is worth noting that the (μ, K, ρ)-*sparsity* is capable of characterizing *any* arbitrarily chosen distribution. The interesting part is that for some of the well-known distributions, such as $\mathcal{G}_{\text{ER}}(n, p)$, this sparsity characterization offers a method that can be used in the analysis and moreover, it leads to an *exact characterization* of sample complexity for the noiseless case. Therefore, for the particular examples presented in Lemma 1 and Lemma 2, the selected threshold and counting parameters for both of them are "tight" (up to multiplicative factors), in the sense that the corresponding sample complexity matches (up to multiplicative factors) the lower bounds derived from Theorem III.1. This can be seen in Corollary IV.1 and IV.2.

Data: Matrices of measurements $\mathbf{A}$ and $\mathbf{B}$
Result: Estimated graph matrix $\mathbf{X}$

Step (a): Recovering columns independently:

---

for $j \in \mathcal{V}$ **do**
  Solve the following ℓ_1-minimization and obtain an optimal $\mathbf{X}$:

$$\begin{aligned} \text{minimize} \quad & \left|\left|X_j\right|\right|_1 \\ \text{subject to} \quad & ||\mathbf{B}X_j - A_j||_2 \leq \gamma, \\ & X_j \in \mathbb{F}^n. \end{aligned}$$

end

Step (b): Consistency-checking:

---

for $\mathcal{S} \subseteq \mathcal{V}$ *with* $|\mathcal{S}| = n - K$ **do**
  for $i, j \in \mathcal{S}$ **do**
    if $|X_{i,j} - X_{j,i}| \leq 2\gamma$ **then**
      break;
    end
    Declare an error;
  end
  for $j \in \overline{\mathcal{S}}$ **do**

Step (c): Resolving unknown entries:

Update $X_j^{\overline{\mathcal{S}}}$ by solving the linear system:

$$\mathbf{B}_{\overline{\mathcal{S}}} X_j^{\overline{\mathcal{S}}} = A_j - \mathbf{B}_{\mathcal{S}} X_j^{\mathcal{S}}.$$

  end
  return $\mathbf{X} = (X_1, \ldots, X_n)$;
end

Algorithm 1: A three-stage recovery scheme. The first stage focuses on solving each column of the matrix $\mathbf{Y}$ independently using ℓ_1-minimization. In the second stage, the recovery correctness of the first stage is further verified via *consistency-checking*, which utilizes the fact that the matrix to be recovered $\mathbf{Y}$ is *symmetric*. The parameter γ is set to zero for the analysis of noiseless parameter reconstruction.

C. Sufficient Conditions

In this subsection, we consider the sufficient conditions (achievability) for parameter reconstruction. The proofs rely on constructing a three-stage recovery scheme (Algorithm 1), which contains three steps – *column-retrieving*, *consistency-checking* and *solving unknown entries*. The worst-case running time of this scheme depends on the underlying distribution $\mathcal{G}_n$[4]. The scheme is presented as follows.

1) Three-stage Recovery Scheme:

Step (a): Retrieving columns. In the first stage, using ℓ_1-norm minimization, we recover each column of $\mathbf{Y}$ based on

[4]Although for certain distributions, the computational complexity is not polynomial in n, the scheme still provides insights on the fundamental trade-offs between the number of samples and the probability of error for recovering graph matrices. Furthermore, motivated by the scheme, a polynomial-time heuristic algorithm is provided in Section V and experimental results are reported in Section VI.

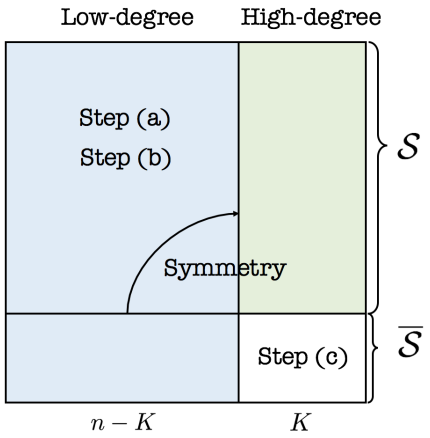

Fig. 1. The recovery of a graph matrix $\mathbf{Y}$ using the three-stage scheme in Algorithm 1. The $n-K$ columns of $\mathbf{Y}$ colored by gray are first recovered via the ℓ_1-minimization (11a)-(11c) in step (a), after they are accepted by passing the consistency check in step (b). Then, symmetry is used for recovering the entries in the matrix marked by green. Leveraging the linear measurements again, in step (c), the remaining K^2 entries in the white symmetric sub-matrix are solved using Equation (12).

(1):

$$\begin{aligned} \text{minimize} \quad & \left|\left|X_j\right|\right|_1 && \text{(11a)} \\ \text{subject to} \quad & ||\mathbf{B}X_j - A_j||_2 \leq \gamma, && \text{(11b)} \\ & X_j \in \mathbb{F}^n. && \text{(11c)} \end{aligned}$$

Let $X_j^{\mathcal{S}} := (X_{i,j})_{i\in\mathcal{S}}$ be a length-$|\mathcal{S}|$ column vector consisting of $|\mathcal{S}|$ coordinates in X_j, the j-th retrieved column. We do not restrict the methods for solving the ℓ_1-norm minimization in (11a)-(11c), as long as there is a unique solution for sparse columns with fewer than μ non-zeros (provided enough number of measurements and the parameter $\mu > 0$ is defined in Definition III.1).

Step (b): Checking consistency.

In the second stage, we check for error in the decoded columns $X_1, \ldots, X_n$ using the symmetry property (perturbed by noise) of the graph matrix $\mathbf{Y}$. Specifically, we fix a subset $\mathcal{S} \subseteq \mathcal{V}$ with a given size $|\mathcal{S}| = n - K$ for some integer[5] $0 \leq K \leq n$. Then we check if $|X_{i,j} - X_{j,i}| \leq 2\gamma$ for all $i, j \in \mathcal{S}$. If not, we choose a different set $\mathcal{S}$ of the same size. This procedure stops until either we find such a subset $\mathcal{S}$ of columns, or we go through all possible subsets without finding one. In the latter case, an error is declared and the recovery is unsuccessful. It remains to recover the vectors X_j for $j \in \overline{\mathcal{S}}$.

Step (c): Resolving unknown entries. In the former case, for each vector $X_j, j \in \mathcal{S}$, we accept its entries $X_{i,j}, i \in \overline{\mathcal{S}}$, as correct and therefore, according to the symmetry assumption, we know the entries $X_{i,j}, i \in \mathcal{S}, j \in \overline{\mathcal{S}}$ (equivalently $\{X_j^{\mathcal{S}} : j \in \overline{\mathcal{S}}\}$), which are used together with the sub-matrices $\mathbf{B}_{\mathcal{S}}$ and $\mathbf{B}_{\overline{\mathcal{S}}}$ to compute the other entries $X_{i,j}, i \in \overline{\mathcal{S}}$, of X_j using (11b):

$$\mathbf{B}_{\overline{\mathcal{S}}} X_j^{\overline{\mathcal{S}}} = A_j - \mathbf{B}_{\mathcal{S}} X_j^{\mathcal{S}}, \quad j \in \overline{\mathcal{S}}. \tag{12}$$

Note that to avoid being over-determined, in practice, we solve

$$\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}} X_j^{\overline{\mathcal{S}}} = A_j^{\mathcal{K}} - \mathbf{B}_{\mathcal{S}}^{\mathcal{K}} X_j^{\mathcal{S}}, \quad j \in \overline{\mathcal{S}}$$

where $\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}}$ is a $K \times K$ matrix whose rows are selected from $\mathbf{B}_{\overline{\mathcal{S}}}$ corresponding to $\mathcal{K} \subseteq \mathcal{V}$ with $|\mathcal{K}| = K$ and $\mathbf{B}_{\mathcal{S}}^{\mathcal{K}}$ selects the rows of $\mathbf{B}_{\mathcal{S}}$ in the same way. We combine $X_j^{\mathcal{S}}$ and $X_j^{\overline{\mathcal{S}}}$ to obtain a new estimate X_j for each $j \in \overline{\mathcal{S}}$. Together with the columns X_j, $j \in \mathcal{S}$, that we have accepted, they form the estimated graph matrix $\mathbf{X}$. We illustrate the three-stage scheme in Figure 1. In the sequel, we analyze the sample complexity of the three-stage scheme based on the (μ, K, ρ)-sparse distributions defined in Definition III.1.

2) Analysis of the Scheme:

Let $\mathbb{F} \equiv \mathbb{R}$ for the simplicity of representation and analysis. We now present another of our main theorems. Consider the models defined in Section II-B and II-C. The Γ-*probability of error* is defined to be the maximal probability that the ℓ_2-norm of the difference between the estimated vector $X \in \mathbb{R}^n$ and the original vector $Y \in \mathbb{R}^n$ (satisfying $A = \mathbf{B}Y + Z$ and both A and $\mathbf{B}$ are known to the estimator) is larger than $\Gamma > 0$:

$$\bar{\varepsilon}_{\mathrm{P}}(\Gamma) := \sup_{Y \in \Upsilon(\mu)} \mathbb{P}\left(||X - Y||_2 > \Gamma\right)$$

where $\Upsilon(\mu)$ is the set of all μ-sparse vectors in $\mathbb{R}^n$ and the probability is taken over the randomness in the generator matrix $\mathbf{B}$ and the additive noise Z. Given a generator matrix $\mathbf{B}$, the corresponding *restricted isometry constant* denoted by δ_μ is the smallest positive number with

$$(1 - \delta_\mu)\,||\mathbf{x}||_2^2 \leq ||\mathbf{B}_{\mathcal{S}}\mathbf{x}||_2^2 \leq (1 + \delta_\mu)\,||\mathbf{x}||_2^2 \tag{13}$$

for all subsets $\mathcal{S} \subseteq \mathcal{V}$ of size $|\mathcal{S}| \leq \mu$ and all $\mathbf{x} \in \mathbb{R}^{|\mathcal{S}|}$. Below we state a sufficient condition[6] derived form the three-stage scheme for parameter reconstruction.

***Theorem* III.2** (Achievability)**.** *Suppose the generator matrix satisfies that $\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}} \in \mathbb{R}^{K\times K}$ is invertible for all $\overline{\mathcal{S}} \subseteq \mathcal{V}$ and $\mathcal{K} \subseteq \mathcal{V}$ with $|\overline{\mathcal{S}}| = |\mathcal{K}| = K$. Let the distribution $\mathcal{G}_n$ be (μ, K, ρ)-sparse. If the three-stage scheme in Algorithm 1 is used for recovering a graph matrix $\mathbf{Y}(G_n)$ of G_n that is sampled according to $\mathcal{G}_n$, then the probability of error satisfies $\varepsilon_{\mathrm{P}}(\eta) \leq \rho + (n-K)\bar{\varepsilon}_{\mathrm{P}}(\Gamma)$ with η greater or equal to*

$$2\left(n\Gamma + \frac{\Gamma||\mathbf{B}||_2 + \gamma}{1 - \delta_{2K}}\right)\left(2(n-K) + K\xi(\mathbf{B})\right)$$

where δ_{2K} is the corresponding restricted isometry constant of $\mathbf{B}$ with $\mu = 2K$ defined in (13) *and*

$$\xi(\mathbf{B}) := \max_{\mathcal{S},\mathcal{K}\subseteq\mathcal{V},|\overline{\mathcal{S}}|=|\mathcal{K}|=K} \left|\left|\mathbf{B}_{\mathcal{S}}\right|\right|_2 \left|\left|\left(\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}}\right)^{-1}\right|\right|_2.$$

The proof is in Appendix B. The theory of classical compressed sensing (see [40], [41], [51]) implies that for

[5]The choice of K depends on the structure of the graph to be recovered and more specifically, K is the counting parameter in Definition III.1. In Theorem III.2 and Corollary III.1, we analyze the sample complexity of this three-stage recovery scheme by characterizing an arbitrary graph into the classes specified by Definition III.1 with a fixed K.

[6]Note that γ cannot be chosen arbitrarily and Γ depends on γ; otherwise the probability of error $\bar{\varepsilon}_{\mathrm{P}}(\Gamma)$ will blow up. Theorem IV.2 indicates that for Gaussian ensembles setting $\Gamma = O(\gamma) = O(\sqrt{n}\sigma_{\mathrm{N}})$ is a valid choice where σ_{N} is the standard deviation of each independent $Z_{i,j}$ in $\mathbf{Z}$.

noiseless parameter reconstruction, if the generator matrix $\mathbf{B}$ has restricted isometry constants $\delta_{2\mu}$ and $\delta_{3\mu}$ satisfying $\delta_{2\mu}+\delta_{3\mu}<1$, then all columns Y_j with $j\in\mathcal{V}_{\rm Small}$ are correctly recovered using the minimization in (11a)-(11c). Denote by $\mathrm{spark}(\mathbf{B})$ the smallest number of columns in the matrix $\mathbf{B}$ that are linearly dependent (see [52] for the requirements on the spark of the generator matrix to guarantee desired recovery criteria). The following corollary is an improvement of Theorem III.2 for the noiseless case. The proof is postponed to Appendix C.

***Corollary* III.1.** *Let $\mathbf{Z}=0$ and suppose the generator matrix $\mathbf{B}$ has restricted isometry constants $\delta_{2\mu}$ and $\delta_{3\mu}$ satisfying $\delta_{2\mu}+\delta_{3\mu}<1$ and furthermore, $\mathrm{spark}(\mathbf{B})>2K$. If the distribution $\mathcal{G}_n$ is (μ,K,ρ)-sparse, then the probability of error for the three-stage scheme to recover the parameters of a graph matrix $\mathbf{Y}(G_n)$ of G_n that is sampled according to $\mathcal{G}_n$ satisfies $\varepsilon_{\rm P}\le\rho$.*

IV. Gaussian IID Measurements

In this section, we consider a special regime when the measurements in the matrix $\mathbf{B}$ are Gaussian IID random variables. Utilizing the converse in Theorem III.1 and the achievability in Theorem III.2, the Gaussian IID assumption allows the derivation of explicit expressions of sample complexity as upper and lower bounds on the number of measurements m. Combining with the results in Lemma 1 and 2, we are able to show that for the corresponding lower and upper bounds match each other for graphs distributions $\mathcal{U}_{\mathsf{T}(n)}$ and $\mathcal{G}_{\rm ER}(n,p)$ (with certain conditions on p and n).

For the convenience of presentation, in the remainder of the paper, we restrict that the measurements are chosen from $\mathbb{R}$, although the theorems can be generalized to the complex measurements. In realistic scenarios, for instance, a power network, besides the measurements collected from the nodes, nominal state values, *e.g.,* operating current and voltage measurements are known to the system designer a priori. Representing the nominal values at the nodes by $\overline{A}\in\mathbb{R}^n$ and $\overline{B}\in\mathbb{R}^n$ respectively, the measurements in $\mathbf{A}$ and $\mathbf{B}$ are centered around $m\times n$ matrices $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ defined as

$$\overline{\mathbf{A}}:=\begin{bmatrix}\cdots & \overline{A} & \cdots\\ \cdots & \overline{A} & \cdots\\ & \vdots & \\ \cdots & \overline{A} & \cdots\end{bmatrix},\quad \overline{\mathbf{B}}:=\begin{bmatrix}\cdots & \overline{B} & \cdots\\ \cdots & \overline{B} & \cdots\\ & \vdots & \\ \cdots & \overline{B} & \cdots\end{bmatrix}.$$

The rows in $\mathbf{A}$ and $\mathbf{B}$ are the same, because the graph parameters are time-invariant, so are the nominal values. Without system fluctuations and noise, the nominal values satisfy the linear system in (1), *i.e.,*

$$\overline{\mathbf{A}}=\overline{\mathbf{B}}\mathbf{Y}. \tag{14}$$

Knowing $\overline{A}$ and $\overline{B}$ is not sufficient to infer the network parameters (the entries in the graph matrix $\mathbf{Y}$), since the rank of the matrix $\overline{B}$ is one. However, measurement fluctuations can be used to facilitate the recovery of $\mathbf{Y}$. The deviations from the nominal values are denoted by additive perturbation matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ such that $\mathbf{A}=\overline{\mathbf{A}}+\widetilde{\mathbf{A}}$. Similarly, $\mathbf{B}=\overline{\mathbf{B}}+\widetilde{\mathbf{B}}$ where $\widetilde{\mathbf{B}}$ is an $m\times n$ matrix consisting of additive perturbations. Therefore, considering the original linear system in (1), the equations above imply that $\overline{\mathbf{A}}+\widetilde{\mathbf{A}}=\mathbf{BY}+\mathbf{Z}=\overline{\mathbf{B}}\mathbf{Y}+\widetilde{\mathbf{B}}\mathbf{Y}+\mathbf{Z}$ leading to $\widetilde{\mathbf{A}}=\widetilde{\mathbf{B}}\mathbf{Y}+\mathbf{Z}$ where we have made use of (14) and extracted the perturbation matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$. We specifically consider the case when the additive perturbations $\widetilde{\mathbf{B}}$ is a matrix with Gaussian IID entries. Without loss of generality, we suppose the mean of the Gaussian random variable is zero and the standard deviation is $\sigma_{\rm S}$. We consider additive white Gaussian noise (AWGN) with mean zero and standard deviation $\sigma_{\rm N}$. For simplicity, in the remainder of this paper, we slightly abuse the notation and replace the perturbation matrices $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ by $\mathbf{A}$ and $\mathbf{B}$ (we assume that $\mathbf{B}$ is Gaussian IID), if the context is clear. Under the assumptions above, the following lemma can be inferred from Theorem III.1 and the proof is in Appendix F.

***Lemma* 3.** *Consider the linear model $\mathbf{A}=\mathbf{BY}+\mathbf{Z}$. Suppose $B_{i,j}\sim\mathcal{N}(0,\sigma_{\rm S}^2)$ and $Z_{i,j}\sim\mathcal{N}(0,\sigma_{\rm N}^2)$ are mutually independent Gaussian random variables for all $i,j\in\mathcal{V}$. The probability of error for topology identification $\varepsilon_{\rm T}$ is bounded from below as*

$$\varepsilon_{\rm T}\ge 1-\frac{nm\ln\left(1+\frac{\sigma_{\rm S}^2}{\sigma_{\rm N}^2}\overline{Y}\right)}{2\mathbb{H}(\mathcal{G}_n)} \tag{15}$$

where $\overline{Y}:=\max_{i,j}|Y_{i,j}|$ denotes the maximal absolute value of the entries in the graph matrix $\mathbf{Y}$. In particular, if $\mathbf{Z}=0$, then for parameter reconstruction,

$$\varepsilon_{\rm P}\ge 1-\frac{nm\ln\left(2\pi e\overline{Y}\sigma_{\rm S}^2\right)}{2\mathbb{H}(\mathcal{G}_n)}. \tag{16}$$

A. Sample Complexity for Sparse Distributions

We consider the worst-case sample complexity for recovering graphs generated according to a sequence of sparse distributions, defined similarly as Definition III.1 to characterize asymptotic behavior of graph distributions.

***Definition* IV.1** (Sequence of sparse distributions)**.** A sequence $\{\mathcal{G}_n\}$ of graph distributions is said to be (μ,K)-*sparse* if assuming a sequence of graphs $\{G_n\}$ is generated according to $\{\mathcal{G}_n\}$, the sequences $\{\mu(n)\}$ and $\{K(n)\}$ guarantee that

$$\lim_{n\to\infty}\mathbb{P}_{\mathcal{G}_n}\left(G_n\notin\mathsf{C}(n)(\mu(n),K(n))\right)=0. \tag{17}$$

In the remaining contexts, we write $\mu(n)$ and $K(n)$ as μ and K for simplicity if there is no confusion. Based on the sequence of sparse distributions we defined above, we show the following theorem, which provides upper and lower bounds on the worst-case sample complexity, with Gaussian IID measurements.

***Theorem* IV.1** (Noiseless worst-case sample complexity)**.** *Let $\mathbf{Z}=0$. Suppose that the generator matrix $\mathbf{B}$ has Gaussian IID entries with mean zero and variance one and assume $\mu<n^{-3/\mu}(n-K)$ and $K=o(n)$. For any sequence of distributions that is (μ,K)-sparse, the three-stage scheme guarantees that $\lim_{n\to\infty}\varepsilon_{\rm P}=0$ using $m=O\left(\mu\log(n/\mu)+K\right)$ measurements. Conversely, there exists a (μ,K)-sparse sequence of distributions such that the number of measurements*

must satisfy $m = \Omega\left(\mu \log(n/\mu) + K/n^{3/\mu}\right)$ *to make the probability of error* ε_{P} *less than* $1/2$ *for all* n.

The proof is postponed to Appendix G.

***Remark* 6.** The upper bound on m that we are able to show differs from the lower bound by a sub-linear term $n^{3/\mu}$. In particular, when the term $\mu \log(n/\mu)$ dominates K, the lower and upper bounds become tight up to a multiplicative factor.

B. Applications of Theorem IV.1

1) Uniform Sampling of Trees:

As one of the applications of Theorem IV.1, we characterize the sample complexity of the uniform sampling of trees.

***Corollary* IV.1.** *Let* $\mathbf{Z} = 0$. *Suppose that the generator matrix* $\mathbf{B}$ *has Gaussian IID entries with mean zero and variance one and assume* G_n *is distributed according to* $\mathcal{U}_{\mathsf{T}(n)}$. *There exists an algorithm that guarantees* $\lim_{n\to\infty} \varepsilon_{\mathrm{P}} = 0$ *using* $m = O(\log n)$ *measurements. Conversely, the number of measurements must satisfy* $m = \Omega(\log n)$ *to make the probability of error* ε_{P} *less than* $1/2$.

Proof. The achievability follows from combining Theorem IV.1 and Lemma 1, by setting $K(n) = \log n$. Substituting $\mathbb{H}(\mathcal{U}_{\mathsf{T}(n)}) = \Omega(n \log n)$ into (16) yields the desired result for converse. □

2) Erdős-Rényi (n, p) *model:*

Similarly, recalling Lemma 2, the sample complexity for recovering a random graph generated according to the Erdős-Rényi (n, p) model is obtained.

***Corollary* IV.2.** *Let* $\mathbf{Z} = 0$. *Assume* G_n *is a random graph sampled according to* $\mathcal{G}_{\mathrm{ER}}(n, p)$ *with* $1/n \le p \le 1 - 1/n$. *Under the same conditions in Corollary IV.1, there exists an algorithm that guarantees* $\lim_{n\to\infty} \varepsilon_{\mathrm{P}} = 0$ *using* $m = O(nh(p))$ *measurements. Conversely, the number of measurements must satisfy* $m = \Omega(nh(p))$ *to make the probability of error* ε_{P} *less than* $1/2$.

Proof. Taking $K = nh(p)/\log n$ and $\mu = 2nh(p)/(\ln 1/p)$, we check that $\mu < n^{-3/\mu}(n - K)$ and $K = o(n)$. The assumptions on $h(p)$ guarantee that $h(p) \ge \log n/n$, whence $nh(p) = \omega(\log(n/K))$. The choices of $\{\mu(n)\}$ and $\{K(n)\}$ make sure that the sequence of distributions is $(\mu(n), K(n))$-sparse. Theorem IV.1 implies that $m = O(nh(p))$ is sufficient for achieving a vanishing probability of error. For the second part of the corollary, substituting $\mathbb{H}(\mathcal{G}_{\mathrm{ER}}(n, p)) = h(p)\binom{n}{2} = \Omega\left(n^2 h(p)\right)$ into (16) yields the desired result. □

C. Measurements corrupted by AWGN

The results on sample complexity can be extended to the case with noisy measurements. The following theorem is proved by combining Theorem III.2 and Lemma 3. The details can be found in Appendix H.

***Theorem* IV.2** (Noisy worst-case sample complexity)**.** *Suppose that* $\mathbf{B}$ *and* $\mathbf{Z}$ *are defined as in Lemma 3. Let* $\mu < n^{-3/\mu}(n - K)$ *and* $K = o(n)$. *Conversely, there exists a* (μ, K)*-sparse sequence of distributions such that the number of measurements must satisfy*

$$m = \Omega\left(\frac{\mu \log(n/\mu) + K/n^{3/\mu}}{\log(1 + \sigma_{\mathrm{S}}^2/\sigma_{\mathrm{N}}^2)}\right)$$

to make the probability of error ε_{T} *less than* $1/2$ *for all* n. *Moreover, if* $\sigma_{\mathrm{N}} = o(1/n^{5/2})$, $\sigma_{\mathrm{S}} = 1/\sqrt{m}$ *and* $K \le \mu$, *then for any sequence of distributions that is* (μ, K)*-sparse, the three-stage scheme guarantees that* $\lim_{n\to\infty} \varepsilon_{\mathrm{T}} = 0$ *using* $m = O(\mu \log(n/\mu))$ *measurements. Moreover,* $\lim_{n\to\infty} \varepsilon_{\mathrm{P}}(\eta) = 0$ *with* $\eta = o(1)$.

***Remark* 7.** The proof of Theorem IV.2 implies that $\eta = O(n^{5/2}\sigma_{\mathrm{N}})$. Therefore, if we consider the normalized Frobenius norm of $(1/n^2)||\mathbf{Y} - \mathbf{X}||_{\mathrm{F}}$ where $\mathbf{X}$ and $\mathbf{Y}$ are the recovered and original graph matrices respectively, then $\sigma_{\mathrm{N}} = o(1/\sqrt{n})$ guarantees that the normalized Frobenius norm vanishes. For topology identification, we need to consider the Frobenius norm bound, η, to rule out the worst-case situation and the sufficient condition becomes $\sigma_{\mathrm{N}} = o(1/n^{5/2})$. Another implication is that the choice of γ in (11b) satisfying $\gamma = O(\sqrt{n}\sigma_{\mathrm{N}})$ (used in the proof) guarantees the reconstruction criteria and its effectiveness is also validated in our experiments in Section VI-B5.

V. Heuristic Algorithm

We present in this section an algorithm motivated by the consistency-checking step in the proof of achievability (see Section III-C). Instead of checking the consistency of each subset of $\mathcal{V}$ consisting of $n - K$ nodes, as the three-stage scheme does and which requires $O(n^K)$ operations, we compute an estimate X_j for each column of the graph matrix independently and then assign a score to each column based on its symemtric consistency with respect to the other columns in the matrix. The lower the score, the closer the estimate of the matrix column X_j is to the ground truth Y_j. Using a scoring function we rank the columns, select a subset of them to be "correct", and then eliminate this subset from the system. The size of the subset determines the number of iterations. Heuristically, this procedure results in a polynomial-time algorithm to compute an estimate $\mathbf{X}$ of the graph matrix $\mathbf{Y}$.

The algorithm proceeds in four steps.

1) Step 1. *Initialization:* Let matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{m\times n}$ be given and set the number of columns fixed in each iteration to be an integer s such that $1 \le s \le n$. For the first iteration, set $\mathcal{S}(0) \leftarrow \mathcal{V}$, $\mathbf{A}(0) \leftarrow \mathbf{A}$, and $\mathbf{B}(0) \leftarrow \mathbf{B}$.

For each iteration $r = 0, \ldots, \lceil n/s \rceil - 1$, we perform the remaining three stages. The system dimension is reduced by s after each iteration.

2) Step 2. *Independent* ℓ_1*-minimization:* For all $j \in \mathcal{S}(r)$, we solve the following ℓ_1-minimization:

$$X_j(r) = \underset{x\in\mathbb{F}^{n-sr}}{\arg\min} \quad \big|\big|x\big|\big|_1 \tag{18}$$

$$\text{subject to} \quad ||\mathbf{B}(r)x - A_j(r)||_2 \le \gamma, \quad x \in \mathcal{X}_j(r). \tag{19}$$

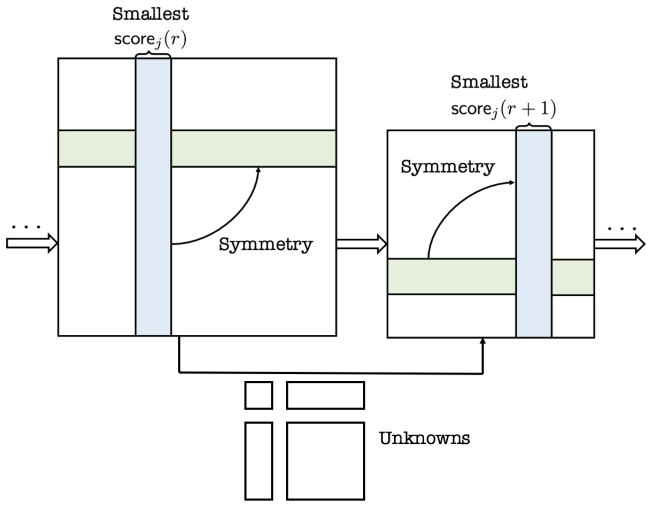

Fig. 2. Iterative dimension reduction of the heuristic algorithm. At step r, the s columns with the smallest scores defined in (20) are assumed to be "correct" and eliminated from the linear system. The dimension of variables is reduced by s and this procedure is repeated until the $\lceil n/s \rceil$ iterations are complete.

Constraint (18) is optional; the set $\mathcal{X}_j(r)$ may encode additional constraints on the form of x such as entry-wise positivity or negativity (*e.g.,* Section VI). The forms of reduced matrix $\mathbf{B}(r)$ and reduced vector $A_j(r)$ are specified in Step 4.

3) Step 3. *Column scoring:* We rank the *symmetric consistency* of the independently solved columns. For all $j \in \mathcal{S}(r)$, let

$$\mathsf{score}_j(r) := \sum_{i=1}^{n-sr} |X_{i,j}(r) - X_{j,i}(r)|. \tag{20}$$

Note that if $\mathsf{score}_j(r) = 0$ then $X_j(r)$ and its partner symmetric row in $\mathbf{X}(r)$ are identical. Otherwise there will be some discrepancies between the entries and the sum will be positive. The subset of the $X_j(r)$ corresponding to the s smallest values of $\mathsf{score}_j(r)$ is deemed "correct". Call this subset of correct indices $\mathcal{S}'(r)$.

4) Step 4. *System dimension reduction:* Based on the assumption that s of the previously computed columns $X_j(r)$ are correct, the dimension of the linear system is reduced by s. We set $\mathcal{S}(r+1) \leftarrow \mathcal{S}(r) \backslash \mathcal{S}'(r)$. For all $i, j \in \mathcal{S}'(r)$, we fix

$$X_{i,j} = X_{i,j}(r), \;\; X_{j,i} = X_{i,j}(r). \tag{21}$$

The measurement matrices are reduced to

$$\begin{aligned} \mathbf{B}(r+1) &\leftarrow \underline{\mathbf{B}}_{\mathcal{S}(r+1)}, \\ A_j(r+1) &\leftarrow \underline{A}_j(r) - \sum_{i \in \mathcal{S}'(r)} \underline{B}_i X_{i,j}. \end{aligned}$$

When $r \leq n - m$, $\underline{\mathbf{B}}_{\mathcal{S}(r+1)} = \mathbf{B}_{\mathcal{S}(r+1)}$, $\underline{A}_j(r) = A_j(r)$ and $\underline{B}_i = B_i$. When $r > n - m$, to avoid making the reduced matrix $\mathbf{B}(r+1)$ over-determined, we set $\mathbf{B}(r+1)$ to be an $(n-r) \times (n-r)$ sub-matrix of $\mathbf{B}_{\mathcal{S}(r+1)}$ by selecting $n - r$ rows of $\mathbf{B}_{\mathcal{S}(r+1)}$ uniformly at random. A new length-$(n-r)$ vector $\underline{A}_j(r)$ is formed by selecting the corresponding entries from $A_j(r)$. Once the $\lceil n/s \rceil$ iterations complete, an estimate $\mathbf{X}$ is returned using (21). The algorithm requires at most $\lceil n/s \rceil$ iterations and in each iteration, the algorithm solves an ℓ_1-minimization and updates a linear system. Solving an ℓ_1-minimization can be done in polynomial time (*c.f.* [53]). Thus, the heuristic algorithm is a polynomial-time algorithm.

VI. Applications in Electric Grids

Experimental results for the heuristic algorithm are given here for both synthetic data and IEEE standard power system test cases. The algorithm was implemented in Matlab; simulated power flow data was generated using Matpower 7.0 [54] and CVX 2.1 [55] with the Gurobi solver [56] was used to solve the sparse optimization subroutine.

A. Scalable Topologies and Error Criteria

We first demonstrate our results using synthetic data and two typical graph ensembles – stars and chains. For both topologies, we increment the graph size from $n = 5$ to $n = 300$ and record the number of samples required for accurate recovery of parameters and topology. For each simulation, we generate a complex-valued random admittance matrix $\mathbf{Y}$ as the ground truth. Both the real and imaginary parts of the line impedances of the network are selected uniformly and IID from $[-100, 100]$. A valid electrical admittance matrix is then constructed using these impedances. The real components of the entries of $\mathbf{B}$ are distributed IID according to $\mathcal{V}(1,1)$ and the imaginary components according to $\mathcal{V}(0,1)$. $\mathbf{A} = \mathbf{YB}$ gives the corresponding complex-valued measurement matrix. The parameter γ in (19) is 0 since we consider noiseless reconstruction here.

Given data matrices $\mathbf{A}, \mathbf{B}$ the algorithm returns an estimate $\mathbf{X}$ of the ground truth $\mathbf{Y}$. We set $s = \lceil n/2 \rceil$ for each graph. If an entry of $\mathbf{X}$ has magnitude $|X_{i,j}| < 10^{-5}$, then we fix it to be 0. Following this, if $\mathrm{supp}(\mathbf{X}) = \mathrm{supp}(\mathbf{Y})$ then the topology identification is deemed exact. The criterion for accurate parameter reconstruction is $||\mathbf{Y} - \mathbf{X}||_{\mathrm{F}}/n^2 < 10^{-6}$. The number of samples m (averaged over repeated trials) required to meet both of these criteria is designated as the sample complexity for accurate recovery. The sample complexity trade-off displayed in Figure 3 shows approximately logarithmic dependence on graph size n for both ensembles.

B. IEEE Test Cases

We also validate the heuristic algorithm on 17 IEEE standard power system test cases ranging from 5 to 200 buses. The procedure for determining sample complexity for accurate recovery is the same as above, but the data generation is more involved.

1) Power flow data generation: A sequence of time-varying loads is created by scaling the nominal load values in the test cases by a times series of Bonneville Power Administration's aggregate load on 02/08/2016, 6am to 12pm [57]. For each test case network, we perform the following steps to generate a set of measurements:

a) Interpolate the aggregate load profile to 6-second intervals, extract a length-m random consecutive subsequence,

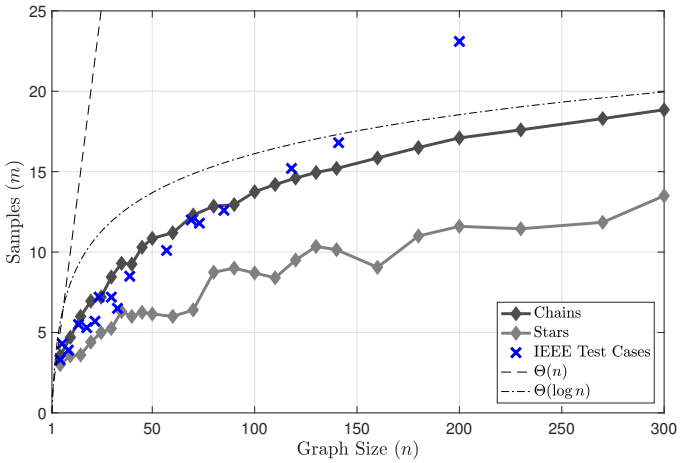

Fig. 3. The number of samples required to accurately recover the nodal admittance matrix is shown on the vertical axis. Results are averaged over 20 independent simulations. Star and chain graphs are scaled in size between 5 and 300 nodes. IEEE test cases ranged from 5 to 200 buses. In the latter case, there are no assumptions on the random IID selection of the entries of $\mathbf{Y}$ (in contrast to the star/chain networks). Linear and logarithmic (in n) reference curves are plotted as dashed lines.

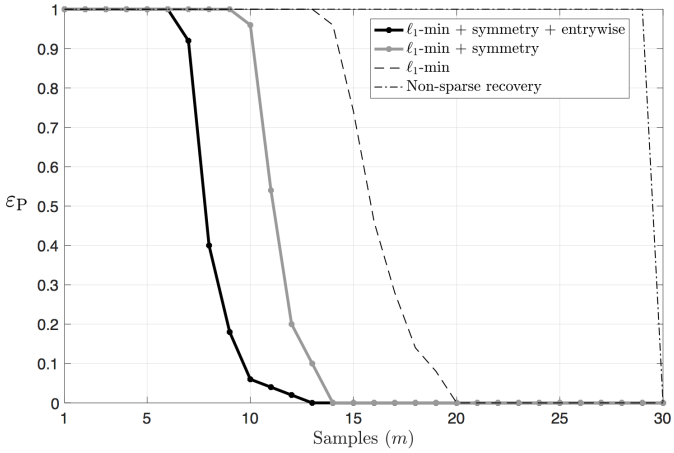

Fig. 4. Probability of error for parameter reconstruction ε_{P} for the IEEE 30-bus test case is displayed on the vertical axis. Probability is taken over 50 independent trials. The horizontal axis shows the number of samples used to compute the estimate $\mathbf{X}$. The probability of error for independent recovery of all X_j via ℓ_1-norm minimization (double dashed line) and full rank non-sparse recovery (dot dashed line) are shown for reference. Adding the symmetry score function (second-to-left) improves over the naive column-wise scheme. Adding entry-wise positivity/negativity constraints on the entries of $\mathbf{X}$ (left-most curve) reduces sample complexity even further ($\approx 1/3$ samples needed compared to full rank recovery).

and then scale the real parts of bus power injections by the load factors in the subsequence.

b) Compute optimal power flow in Matpower for the network at each time step to determine bus voltage phasors.

c) Add a small amount of Gaussian random noise ($\sigma^2 = 0.001$) to the voltage measurements and generate corresponding current phasor measurements using the known admittance matrix.

2) Sample complexity for recovery of IEEE test cases: Figure 3 shows the sample complexity for accurate recovery of the IEEE test cases. The procedure and criteria for determining the necessary number of samples for accurate recovery of the admittance matrix are the same as for the synthetic data case. Unlike the previous setting, here we have no prior assumptions about the structure of the IEEE networks: networks have both mesh and radial topologies. However, because power system topologies are typically highly sparse, the heuristic algorithm was able to achieve accurate recovery with a comparable (logarithmic) dependence on graph size.

3) Influence of structure constraints on recovery: There are structural properties of the nodal admittance matrix for power systems—symmetry, sparsity, and entry-wise positivity/negativity—that we exploit in the heuristic algorithm to improve sample complexity for accurate recovery. The score function $\mathsf{score}_j(r)$ rewards symmetric consistency between columns in $\mathbf{X}$; the use of ℓ_1-minimization promotes sparsity in the recovered columns; and the constraint set $\mathcal{X}_j$ in (18) forces $\mathrm{Re}(X_{i,j}) \leq 0,\ \mathrm{Im}(X_{i,j}) \geq 0$ for $i \neq j$ and $\mathrm{Re}(X_{i,j}) \geq 0$ for $i = j$. These entry-wise properties are commonly found in power system admittance matrices. In Figure 4 we show the results of an experiment on the IEEE 30-bus test case that quantify the effects of the structure constraints on the probability of error. In Figure 5 we show that the score function and the constraints are effective across a range of IEEE test cases, compared with the standard compressed sensing recovery discussed in Section I-B3. Furthermore, this demonstrates the heuristic algorithm is robust to noise for a broad range of real-world graph structures with respect to Frobenius norm error.

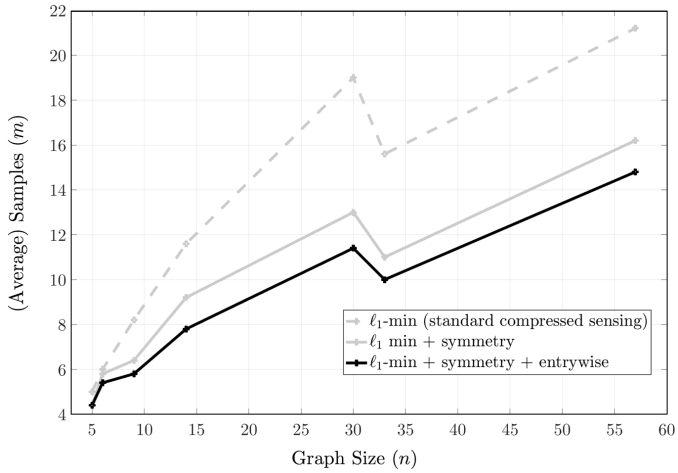

Fig. 5. Sample complexity for accurate recovery is shown for a selection of IEEE power system test cases ranging from 5 to 57 buses. The number of samples for accurate recovery is obtained by satisfying the criterion $||\mathbf{X} - \mathbf{Y}||_{\mathrm{F}}/n^2 < 10^{-4}$. The noise $\mathbf{Z}$ is an IID Gaussian matrix with zero mean and standard deviation 0.01. The parameter γ in (19) is set to be 10^{-4}. As a benchmark, the number of measurements required for separately reconstructing every column of $\mathbf{Y}$ (standard compressed sensing) is also given.

4) Comparison with basis pursuit on star graphs: In Figure 6, we consider star graphs and compare our heuristic algorithm with the modified basis pursuit subroutine in (2)-(4) with noiseless measurements. For a star graph with $n = 24$ nodes, the iterative recovery scheme with $s = 12$ outperforms

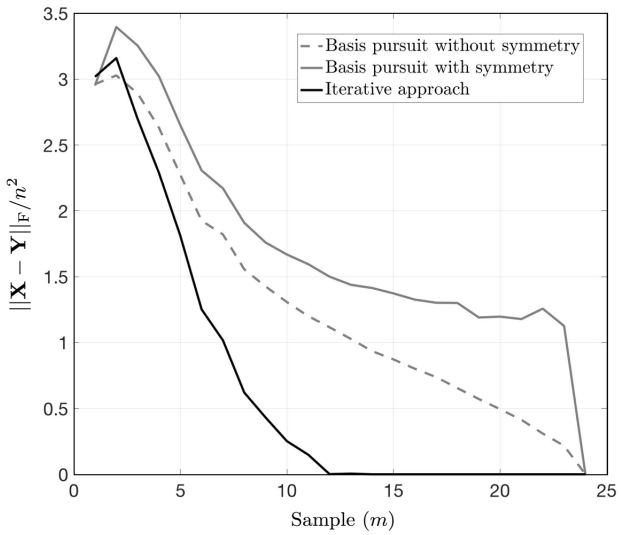

Fig. 6. A comparison between our iterative heuristic and basis pursuit. The Frobenius norm error plotted is averaged over 250 independent trials. The underlying graph is a star graph with $n = 24$. The solid and dotted gray curves are results for basis pursuit with and without a constraint emphasizing symmetry, respectively.

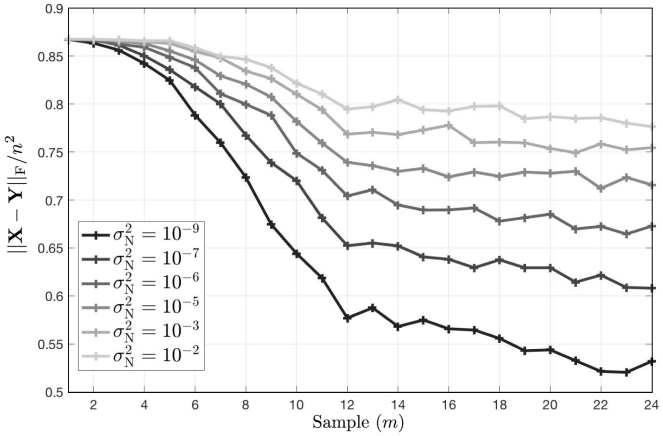

Fig. 7. The impact of measurement noise on sample complexity for recovery of the IEEE 24-bus RTS test case is demonstrated. Trajectories correspond to increasing noise levels from dark (least) to light (most). From left to right, we observe—as expected—that for each variance value, the normalized Frobenius error of the recovered matrix decreases as the number of samples used for recovery increases. From bottom to top, we observe that the error increases (for every value of m) as variance of the additive noise $\mathbf{Z}$ increases.

the basis pursuit, with or without a symmetry constraint. The solid and dotted gray curves show the normalized Frobenius error for cases where $\mathbf{Y}(G)$ is constrained to be symmetric and where it is not, respectively. Our experiments show that convex optimization-based approach breaks down if there are highly dense columns in $\mathbf{Y}$. The star graph contains a high-degree node (degree $n-1$), hindering the standard compressed sensing (basis pursuit without the symmetry constraint) from recovering the whole matrix until the number of measurements reaches n. Surprisingly, adding the symmetry constraint suggests basis pursuit performs less well than basis pursuit without the symmetry condition. This is evidence to support the assumption made in [17]. There, the non-zeros in the matrix to be recovered should not be concentrated in any single column (or row) of $\mathbf{Y}(G)$.

5) Effects of noise and selection of γ: In this section, we consider noisy measurements and fix the additive noise $\mathbf{Z}$ be IID Gaussian with mean zero and variance $\sigma_{\mathrm{N}}^2 \in [10^{-9}, 10^{-2}]$. We set $\gamma = \sqrt{n}\sigma_{\mathrm{N}}$ in (19), as indicated in Remark 7. Due to the presence of noise, there is error in the recovered matrix $\mathbf{X}$. However, the mean absolute percentage error is small.

APPENDIX

A. Proof of Theorem III.1

Proof. The graph G is chosen from a discrete set $\mathsf{C}(n)$ according to some probability distribution $\mathcal{G}_n$. Fano's inequality [22] plays an important role in deriving fundamental limits. We especially focus on its extended version. Similar generalizations appear in many places, *e.g.,* [26], [20] and [58]. We repeat the lemma here for the sake of completion:

***Lemma* 4** (Generalized Fano's inequality)**.** *Let G be a random graph and let $\mathbf{A}$ and $\mathbf{B}$ be matrices defined in Section II-B and II-C. Suppose the original graph G is selected from a nonempty candidacy set $\mathsf{C}(n)$ according to a probability distribution $\mathcal{G}_n$. Let $\hat{G}$ denote the estimated graph. Then the conditional probability of error for estimating G from $\mathbf{A}$ given $\mathbf{B}$ is always bounded from below as*

$$\mathbb{P}\left(\hat{G} \neq G \middle| \mathbf{B}\right) \geq 1 - \frac{\mathbb{I}\left(G; \mathbf{A} \middle| \mathbf{B}\right) + \ln 2}{\mathbb{H}\left(\mathcal{G}_n\right)} \tag{22}$$

where the randomness is over the selections of the original graph G and the estimated graph $\hat{G}$.

In (22), the term $\mathbb{I}\left(G; \mathbf{B} \middle| \mathbf{A}\right)$ denotes the conditional mutual information (base e) between G and $\mathbf{B}$ conditioned on $\mathbf{A}$. Furthermore, the conditional mutual information $\mathbb{I}\left(G; \mathbf{A}|\mathbf{B}\right)$ is bounded from above by the differential entropies of $\mathbf{A}$ and $\mathbf{B}$. It follows that

$$\begin{aligned}
\mathbb{I}\left(G; \mathbf{A}|\mathbf{B}\right) &= \mathbb{H}\left(\mathbf{A}|\mathbf{B}\right) - \mathbb{H}\left(\mathbf{A}|G, \mathbf{B}\right) && (23)\\
&\leq \mathbb{H}\left(\mathbf{A}|\mathbf{B}\right) - \mathbb{H}\left(\mathbf{A}|\mathbf{Y}, \mathbf{B}\right) && (24)\\
&= \mathbb{H}\left(\mathbf{A}|\mathbf{B}\right) - \mathbb{H}\left(\mathbf{Z}\right) && (25)\\
&\leq \mathbb{H}\left(\mathbf{A}\right) - \mathbb{H}\left(\mathbf{Z}\right). && (26)
\end{aligned}$$

Here, Eq. (23) follows from the definitions of mutual information and differential entropy. Moreover, knowing $\mathbf{Y}$, the graph G can be inferred. Thus, $\mathbb{H}\left(\mathbf{A}|G, \mathbf{B}\right) \geq \mathbb{H}\left(\mathbf{A}|\mathbf{Y}, \mathbf{B}\right)$ yields (24). Recalling the linear system in (1), we obtain (25). Furthermore, (26) holds since $\mathbb{H}\left(\mathbf{A}\right) \geq \mathbb{H}\left(\mathbf{A}|\mathbf{B}\right)$.

Plugging (26) into (22),

$$\begin{aligned}
\varepsilon_{\mathrm{T}} =& \mathbb{E}_{\mathbf{B}}\left[\mathbb{P}\left(\hat{G} \neq G \middle| \mathbf{B}\right)\right]\\
\geq& 1 - \frac{\mathbb{H}\left(\mathbf{A}\right) - \mathbb{H}\left(\mathbf{Z}\right) + \ln 2}{\mathbb{H}\left(\mathcal{G}_n\right)},
\end{aligned}$$

which yields the desired (9). □

B. Proof of Theorem III.2

Conditioning on that no less than $n - K$ many columns are recovered with respect to the Γ-probability of error, *i,e.,*

for each entry, the absolute value of the difference between the recovered one and the original one is bounded from above by γ, the union bound ensures the desired bound on the probability of error for noisy parameter reconstruction. It remains to show that the consistency-check in our scheme gives the expression for η. First, if no less than $n-K$ many columns are recovered, there must be a subset $\mathcal{S} \subseteq \mathcal{V}$ passing through the consistency-check. Let us consider the vectors that are not μ-sparse. For any such vector $Y^* \in \mathbb{R}^n$, denote by $e = Y^* - Y'$ the difference of Y^* and the original vector Y'. It follows that e can be decomposed as a summation of a $2K$-sparse vector $\overline{e} \in \mathbb{R}^n$ and a vector $f \in \mathbb{R}^n$ that satisfies $|f_i| \leq 2\Gamma$ for all $i \in \mathcal{V}$. Therefore, the definition of restricted isometry constant ensures the following:

$$\begin{aligned}||e||_2 \leq& ||\overline{e}||_2 + ||f||_2 \\ \leq& \frac{1}{1-\delta_{2K}}||\mathbf{B}\overline{e}||_2 + 2n\Gamma \\ \leq& \frac{1}{1-\delta_{2K}}||\mathbf{B}e||_2 + \left(2n + \frac{2||\mathbf{B}||_2}{1-\delta_{2K}}\right)\Gamma\end{aligned}$$

which can be further bounded by noting that

$$||\mathbf{B}e||_2 = ||(\mathbf{B}Y^* - A) - (\mathbf{B}Y' - A)||_2 \leq 2\gamma$$

since both Y' and Y^* satisfy (11b) where A is a column of $\mathbf{A}$. Thus, the consistency-check guarantees that for each j in the set $\mathcal{S} \subseteq \mathcal{V}$ that passes the check,

$$||X_j - Y_j||_2 \leq 2\left(n + \frac{||\mathbf{B}||_2}{1-\delta_{2K}}\right)\Gamma + \frac{2\gamma}{1-\delta_{2K}}.$$

Consider the reduced linear system in (12). For each j in the set $\overline{\mathcal{S}} \subseteq \mathcal{V}$,

$$\begin{aligned}||X_j^{\overline{\mathcal{S}}} - Y_j^{\overline{\mathcal{S}}}||_2 \leq& \left|\left|\left(\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}}\right)^{-1}\right|\right|_2 \left|\left|\mathbf{B}_{\mathcal{S}}(X_j^{\mathcal{S}} - Y_j^{\mathcal{S}})\right|\right|_2 \\ \leq& \left|\left|\left(\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}}\right)^{-1}\right|\right|_2 ||\mathbf{B}_{\mathcal{S}}||_2 \left|\left|X_j^{\mathcal{S}} - Y_j^{\mathcal{S}}\right|\right|_2.\end{aligned}$$

Summing up the bounds on the ℓ_2 norms for each column and considering the worst case of the invertible matrix $\mathbf{B}_{\overline{\mathcal{S}}}^{\mathcal{K}}$, the bound η on the Frobenius norm follows by arranging the terms.

C. Proof of Corollary III.1

Proof. Conditioned on $G \in \mathsf{C}(n)(\mu, K)$ and the assumption $\delta_{3\mu} + 3\delta_{4\mu} < 2$, there are no less than $n-K$ many columns correctly recovered. Therefore, any such set $\mathcal{S}$ with $|\mathcal{S}| = n-K$ must contain at least $n-2K$ many corresponding indexes of the correctly recovered columns. The consistency-checking verifies that if the collection of an arbitrary set of nodes $\mathcal{S}$ of cardinality $n-K$ satisfies the symmetry property as the true graph $\mathbf{Y}$ must obey. If the consistency-checking fails, it is necessary that there exist two distinct length-n vectors Y' and Y^* in $\mathbb{F}^n$ such that Y^* is the minimizer of the ℓ_1-minimization (11a)-(11c) that differs from the correct answer Y', *i.e.,* $Y' \neq Y^*$ where $A = \mathbf{B}Y'$ and

$$\begin{aligned}Y^* = \arg\min_{Y} \ & ||Y||_1 \\ \text{subject to } & A = \mathbf{B}Y \\ & Y \in \mathbb{F}^n\end{aligned}$$

for some $A \in \mathbb{F}^m$ and furthermore, the vectors Y' and Y^* can have at most $2K$ distinct coordinates,

$$|\mathrm{supp}\,(Y' - Y^*)| \leq 2K.$$

However, the constraints $\mathbf{B}Y' = A$ and $\mathbf{B}Y^* = A$ imply that $\mathbf{B}(Y' - Y^*) = 0$, contradicting to $\mathrm{spark}(\mathbf{B}) > 2K$. Therefore, $n-K$ many columns can be successfully recovered if the decoded solution passes the consistency-checking. Moreover, since $\mathrm{spark}(\mathbf{B}) > 2K$ and number of unknown coordinates in each length-K vector $X_j^{\overline{\mathcal{S}}}$ (for $j = 1, \ldots, |\overline{\mathcal{S}}|$) to be recovered is K, the solution of the system (12) is guaranteed to be unique. Thus, Algorithm 1 always recovers the correct columns $Y_1, \ldots, Y_N$ conditioned on $\mathrm{spark}(\mathbf{B}) > 2K$. It follows that $\varepsilon_{\mathrm{P}} \leq 1 - \mathbb{P}_{\mathcal{G}}(G \in \mathsf{C}(n, \mu, K))$ provided $\mathrm{spark}(\mathbf{B}) > 2K$. In agreement with the assumption that the distribution $\mathcal{G}$ is (μ, K, ρ)-sparse, (10) must be satisfied. Therefore, the probability of error must be less than ρ. □

D. Proof of Lemma 1

Proof. Consider the following function

$$F(\mathcal{E}) = \sum_{j=1}^{n} f(d_j(G))$$

where $d_j(G)$ denotes the degree of the j-th node and consider the following indicator function:

$$f(d_j(G)) := \begin{cases} 1 & \text{if } d_j(G) > \mu \\ 0 & \text{otherwise} \end{cases}.$$

Applying the Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(G \notin \mathsf{T}(n)(\mu, K)\right) &= \mathbb{P}_{\mathcal{U}_{\mathsf{T}(n)}}\left(F(\mathcal{E}) \geq K\right) \\ &\leq \frac{\mathbb{E}_{\mathcal{U}_{\mathsf{T}(n)}}\left[F(\mathcal{E})\right]}{K}. \qquad (27)\end{aligned}$$

Continuing from (27), the expectation $\mathbb{E}_{\mathcal{U}_{\mathsf{T}(n)}}\left[F(\mathcal{E})\right]$ can be further expressed and bounded as

$$\begin{aligned}\mathbb{E}_{\mathcal{U}_{\mathsf{T}(n)}}\left[F(\mathcal{E})\right] &= \sum_{j=1}^{n} \mathbb{E}_{\mathcal{U}_{\mathsf{T}(n)}}\left[f(d_j(G))\right] \\ &= \sum_{j=1}^{n} \mathbb{P}_{\mathcal{U}_{\mathsf{T}(n)}}\left(d_j(G) > \mu\right). \qquad (28)\end{aligned}$$

Since G is chosen uniformly at random from $\mathsf{T}(n)$, it is equivalent to selecting its corresponding Prüfer sequence (by choosing $n-2$ integers independently and uniformly from the set $\mathcal{V}$, *c.f.* [59]) and the number of appearances of each $j \in \mathcal{V}$ equals to $d_j(G) - 1$. Therefore, for any fixed node $j \in \mathcal{V}$, the Chernoff bound implies that

$$\mathbb{P}_{\mathcal{U}_{\mathsf{T}(n)}}\left(d_j(G) > \mu\right) \leq \exp\left(-(n-2)\mathbb{D}_{\mathrm{KL}}\left(\frac{\mu}{n-2}||\frac{1}{n}\right)\right) \qquad (29)$$

where $\mathbb{D}_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback-Leibler divergence and

$$\mathbb{D}_{\mathrm{KL}}\left(\frac{\mu}{n-2}||\frac{1}{n}\right) \geq \frac{\mu}{n-2}\ln n. \qquad (30)$$

Therefore, substituting (30) back into (29) and combining (27) and (28), setting $\mu \geq 1$ leads to

$$\mathbb{P}\left(G \notin \mathsf{T}(n)(\mu, K)\right) \leq \frac{n \exp(-\mu \ln n)}{K} \leq \frac{1}{K}.$$

□

E. Proof of Lemma 2

Proof. For any fixed node $j \in \mathcal{V}$, applying the Chernoff bound,

$$\mathbb{P}_{\mathcal{G}_{\mathrm{ER}}(n,p)}\left(d_j(G) > \mu\right) \leq \exp\left(-n\mathbb{D}_{\mathrm{KL}}\left(\frac{\mu}{n}||p\right)\right).$$

Continuing from (27), the expectation $\mathbb{E}_{\mathcal{G}_{\mathrm{ER}}(n,p}\left[F(\mathcal{E})\right]$ can be further expressed and bounded as

$$\mathbb{E}_{\mathcal{G}_{\mathrm{ER}}(n,p)}\left[F(\mathcal{E})\right] \leq n \cdot \exp\left(-n\mathbb{D}_{\mathrm{KL}}\left(\frac{\mu}{n}||p\right)\right) \tag{31}$$

where the probability p satisfies $0 < p \leq \mu/n < 1$. Note that

$$\mathbb{D}_{\mathrm{KL}}\left(\frac{\mu}{n}||p\right) = \frac{\mu}{n}\ln\frac{1}{p} + \left(1-\frac{\mu}{n}\right)\ln\frac{1}{1-p} - h(p) \tag{32}$$

where the binary entropy $h(p)$ is in base e. Taking $\mu \geq 2nh(p)/(\ln 1/p) \geq 2np$, substituting (32) into (31) leads to

$$\mathbb{E}_{\mathcal{G}_{\mathrm{ER}}(n,p)}\left[F(\mathcal{E})\right] \leq n\exp\left(-nh(p)\right).$$

Therefore, (27) gives

$$\mathbb{P}\left(G \notin \mathsf{C}(n)(\mu, K)\right) \leq \frac{n\exp\left(-nh(p)\right)}{K}.$$

□

F. Proof of Lemma 3

Proof. Continuing from Theorem III.1,

$$\begin{aligned}
&\mathbb{H}(\mathbf{A}) - \mathbb{H}(\mathbf{Z}) \\
=&\sum_{i=1}^{m}\left[\mathbb{H}\left(A^{(i)}\right) - \mathbb{H}\left(Z^{(i)}\right)\right] \\
\overset{(a)}{\leq}&\sum_{i=1}^{m}\frac{n}{2}\left[\ln\left(2\pi e\frac{\mathrm{Tr}\left(\Sigma_{\mathbf{A}^{(i)}}\right)}{n}\right) - \ln(2\pi e\sigma_{\mathrm{N}}^2)\right]
\end{aligned} \tag{33}$$

where $\mathrm{Tr}\left(\Sigma_{\mathbf{A}^{(i)}}\right)$ is the trace of the covariance matrix of $\mathbf{A}^{(i)}$ and we have used the fact that normal distributions maximize entropy and the inequality $\det(\Sigma_{\mathbf{A}^{(i)}}) \leq \left(\mathrm{Tr}\left(\Sigma_{\mathbf{A}^{(i)}}\right)/n\right)^n$ to obtain (a). Note that because of the assumption of independence, the trace is bounded from above by $n\sigma_{\mathrm{S}}^2\overline{Y} + n\sigma_{\mathrm{N}}^2$ where $\overline{Y} := \max_{i,j}|Y_{i,j}|$. Substituting this into (33) completes the proof. The special case when $\mathbf{Z} = 0$ follows similarly. □

G. Proof of Theorem IV.1

Proof. The first part is based on Corollary III.1. Under the assumption of the generator matrix $\mathbf{B}$, using Gordon's escape-through-the-mesh theorem, Theorem 4.3 in [41] implies that for any columns Y_j with $j \in \mathcal{V}_{\mathrm{Small}}$ are correctly recovered using the minimization in (11a)-(11c) with probability at least $1 - 2.5\exp\left(-(4/9)\mu\log(n/\mu)\right)$, as long as the number of measurements satisfies $m \geq 48\mu\left(3 + 2\log(n/\mu)\right)$, and $n/\mu > 2, \mu \geq 4$ (if $\mu \leq 3$, the multiplicative constant increases but our theorem still holds). Similar results were first proved by Candes, *et al.* in [40] (see their Theorem 1.3). Therefore, applying the union bound, the probability that all the μ-sparse columns can be recovered simultaneously is at least $1 - 2.5n\exp\left(-(4/9)\mu\log(n/\mu)\right)$. On the other hand, conditioned on that all the μ-sparse columns are recovered, Corollary III.1 indicates that $\mathrm{spark}(\mathbf{B}) > 2K$ is sufficient for the three-stage scheme to succeed. Since each entry in $\mathbf{B}$ is an IID Gaussian random variable with zero mean and variance one, if $m \geq 48\mu\left(3 + 2\log(n/\mu)\right) + 2K$, with probability one that the spark of $\mathbf{B}$ is greater than $2K$, verifying the statement.

The converse follows by applying Lemma 3 with $\mathbf{Z} = 0$. Consider the uniform distribution $\mathcal{U}_{\mathsf{C}(n)(\mu,K)}$ on $\mathsf{C}(n)(\mu, K)$. Then $\mathbb{H}\left(\mathcal{U}_{\mathsf{C}(n)(\mu,K)}\right) = \ln|\mathsf{C}(n)(\mu, K)|$. Let $0 \leq \alpha, \beta \leq 1$ be parameters such that $\mu < \beta(n - \alpha K)$. To bound the size of $\mathsf{C}(n)(\mu, K)$, we partition $\mathcal{V}$ into $\mathcal{V}_1$ and $\mathcal{V}_2$ with $|\mathcal{V}_1| = n - \alpha K$ and $|\mathcal{V}_2| = \alpha K$. First, we assume that the nodes in $\mathcal{V}_1$ form a $\mu/2$-regular graph. For each node in $\mathcal{V}_2$, construct $\beta(n - \alpha K) \in \mathbb{N}_+$ edges and connect them to the other nodes in $\mathcal{V}$ with uniform probability. A graph constructed in this way always belongs to $\mathsf{C}(n)(\mu, K)$, unless the added edges create more than K nodes with degrees larger than μ. Therefore, as $n \to \infty$,

$$|\mathsf{C}(n)(\mu, K)| \geq \rho \cdot \frac{e^{1/4}\binom{N-1}{\phi}^N\binom{\binom{N}{2}}{\phi N/2}}{\binom{N(N-1)}{\phi N}} \cdot \binom{n-1}{M}^{\alpha K} \tag{34}$$

where $N := n - \alpha K$, $M := \beta(n - \alpha K)$ and $\phi := \mu/2$. The first term ρ denotes the fraction of the constructed graphs that are in $\mathsf{C}(n)(\mu, K)$. The second term in (34) counts the total number of ϕ-regular graphs [60], and the last term is the total number of graphs created by adding new edges for the nodes in $\mathcal{V}_2$. If $K = O(\mu)$, there exists a constant $\alpha > 0$ small enough such that $\rho = 1$. If $\mu = o(K)$, for any fixed node in $\mathcal{V}_1$, the probability that its degree is larger than μ is

$$\begin{aligned}
&\sum_{i=\phi+1}^{\alpha K}\binom{\alpha K}{i}\beta^i(1-\beta)^{\alpha K - i} \\
\leq &\sum_{i=\phi+1}^{\alpha K}\alpha K h\left(\frac{i}{\alpha K}\right)\beta^i \leq (\alpha K)^2\beta^{\phi+1}
\end{aligned}$$

where $h(i/\alpha K)$ is in base e. Take $\beta = n^{-3/\mu}$ and $\alpha = 1/2$. The condition $\mu < n^{-3/\mu}(n - K)$ guarantees that $\mu < \beta(n - \alpha K)$. Letting $F(n) := 1/n$ be the assignment function for each node in $\mathcal{V}_1$, we check that

$$(\alpha K)^2\beta^{\phi+1} \leq \frac{1}{4n} \leq F(n)\cdot\left(1 - \frac{1}{F(n)}\right)^N \leq \frac{1}{en}.$$

Therefore, applying the Lovász local lemma, the probability that all the nodes in $\mathcal{V}_1$ have degree less than or equal to μ can be bounded from below by $\left(1 - F(n)\right)^N \geq 1/4$ if $n \geq 2$,

which furthermore is a lower bound on ρ. Therefore, taking the logarithm,

$$\mathbb{H}\left(\mathcal{U}_{\mathsf{C}(n)(\mu,K)}\right) \geq \frac{(N-1)^2}{2}h(\varepsilon) - O(N\ln\mu) + \frac{K}{2}\left((n-1)h\left(\frac{M}{n-1}\right) - O(\ln n)\right) - O(1) \quad (35)$$

$$= \Omega\left(n^2 h(\varepsilon) + n^{1-3/\mu}K\right) \quad (36)$$

where $\varepsilon := \phi/(N-1) \leq 1/2$. In (35), we have used Stirling's approximation and the assumption that $K = o(n)$. Continuing from (36), since $2nh(\varepsilon) \geq \mu\ln(n/\mu)$, for sufficiently large n,

$$\mathbb{H}\left(\mathcal{U}_{\mathsf{C}(n)(\mu,K)}\right) = \Omega\left(n\mu\log\frac{n}{\mu} + n^{1-3/\mu}K\right). \quad (37)$$

Substituting (37) into (16), when $n \to \infty$, it must hold that

$$m = \Omega\left(\mu\log(n/\mu) + K/n^{3/\mu}\right)$$

to ensure that ε_{P} is smaller than $1/2$. □

H. Proof of Theorem IV.2

The structure of the proof is the same as Theorem IV.1. The converse follows directly by putting the bounds in (37) and (15) together. For proving the achievability, it is sufficient to show that with high probability (in n), $|Y_{i,j} - X_{i,j}| = o(1)$ for all $i, j \in \mathcal{V}$ where $X_{i,j}$ and $Y_{i,j}$ are the recovered and original (i,j)-th entry of the graph matrix. For the Gaussian IID ensemble considered, the ℓ_2-norm of the inverse matrix $(\mathbf{B}^{\mathcal{K}}_{\mathcal{S}})^{-1}$, equivalently, the minimal singular value of $\mathbf{B}^{\mathcal{K}}_{\mathcal{S}}$ is strictly positive with probability $o(1)$ (see the proof of Lemma III-9 in [26]). Using the Chernoff bound, with high probability,

$$||\mathbf{B}||_2^2 \leq ||\mathbf{B}||_{\mathrm{F}}^2 \leq C_1 nm\sigma_{\mathrm{S}}^2, \quad (38)$$

$$||Z_j||_2^2 \leq C_2 n\sigma_{\mathrm{N}}^2, \text{ for all } j \in \mathcal{V} \quad (39)$$

for some positive constants C_1 and C_2. Noting that if $K \leq \mu$, then $\delta_{2K} < 1$ with high probability, the bound in (38) and the bound on the ℓ_2-norm of the inverse matrix $(\mathbf{B}^{\mathcal{K}}_{\mathcal{S}})^{-1}$ imply $\eta = O(n^2\gamma)$, by applying our Theorem III.2. Moreover, with Gaussian measurements, for each μ-sparse vector Y_j in $\mathbb{R}^n$, $||X_j - Y_j||_2 \leq C_3||Z_j||_2$ for some constant $C_3 > 0$ (*cf.* Theorem 1 in [61]) where Y_j satisfies $\mathbf{B}Y_j + Z_j = A_j$ and X_j is the optimal solution of (11a)-(11c) (with $\mathbb{F} \equiv \mathbb{R}$). Therefore, $\Gamma = O(\gamma)$ and $\gamma = O(\sqrt{n}\sigma_{\mathrm{N}})$ using (39). Since $\eta = O(n^2\gamma)$, the condition $\sigma_{\mathrm{N}} = o(1/n^{5/2})$ guarantees that $\eta = o(1)$, whence $|Y_{i,j} - X_{i,j}| = o(1)$ for all $i, j \in \mathcal{V}$ and the proof is complete.

REFERENCES

[1] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[2] C. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions (corresp.)," *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 369–371, 1973.

[3] V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning gaussian tree models: Analysis of error exponents and extremal structures," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2701–2714, 2010.

[4] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 44–63, 2019.

[5] A. Ghoshal and J. Honorio, "Learning linear structural equation models in polynomial time and sample complexity," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1466–1475.

[6] ——, "Learning identifiable gaussian bayesian networks in polynomial time and sample complexity," in *Advances in Neural Information Processing Systems*, 2017, pp. 6457–6466.

[7] K. J. Ahn, S. Guha, and A. McGregor, "Analyzing graph structure via linear measurements," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2012, pp. 459–467.

[8] J. Pouget-Abadie and T. Horel, "Inferring graphs from cascades: A sparse recovery framework," *arXiv preprint arXiv:1505.05663*, 2015.

[9] J. A. Momoh, R. Adapa, and M. El-Hawary, "A review of selected optimal power flow literature to 1993. i. nonlinear and quadratic programming approaches," *IEEE transactions on power systems*, vol. 14, no. 1, pp. 96–104, 1999.

[10] S. H. Low, "Convex relaxation of optimal power flowpart i: Formulations and equivalence," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 15–27, 2014.

[11] Y. Tang, K. Dvijotham, and S. Low, "Real-time optimal power flow," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2963–2973, 2017.

[12] A. Mittal, J. Hazra, N. Jain, V. Goyal, D. P. Seetharam, and Y. Sabharwal, "Real time contingency analysis for power grids," in *European Conference on Parallel Processing*. Springer, 2011, pp. 303–315.

[13] R. Horta, J. Espinosa, and J. Patiño, "Frequency and voltage control of a power system with information about grid topology," in *Automatic Control (CCAC), 2015 IEEE 2nd Colombian Conference on*. IEEE, 2015, pp. 1–6.

[14] L. Guo, C. Liang, A. Zocca, S. H. Low, and A. Wierman, "Failure localization in power systems via tree partitions," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 6832–6839.

[15] L. Zhao and B. Zeng, "Vulnerability analysis of power grids with line switching," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2727–2736, 2013.

[16] S. P. Chepuri, S. Liu, G. Leus, and A. O. Hero, "Learning sparse graphs under smoothness prior," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6508–6512.

[17] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Sketching sparse matrices, covariances, and graphs via tensor products," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1373–1388, 2015.

[18] E. Belilovsky, K. Kastner, G. Varoquaux, and M. B. Blaschko, "Learning to discover sparse graphical models," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 440–448.

[19] A. Bogdanov, E. Mossel, and S. Vadhan, "The complexity of distinguishing markov random fields," in *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*. Springer, 2008, pp. 331–342.

[20] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions." *IEEE Trans. Information Theory*, vol. 58, no. 7, pp. 4117–4134, 2012.

[21] A. Anandkumar, V. Tan, and A. Willsky, "High dimensional structure learning of ising models on sparse random graphs," *arXiv preprint arXiv:1011.0129*, 2010.

[22] R. M. Fano and D. Hawkins, "Transmission of information: A statistical theory of communications," *American Journal of Physics*, vol. 29, pp. 793–794, 1961.

[23] A. Anandkumar, V. Y. Tan, F. Huang, A. S. Willsky *et al.*, "High-dimensional structure estimation in ising models: Local separation criterion," *The Annals of Statistics*, vol. 40, no. 3, pp. 1346–1375, 2012.

[24] A. Anandkumar, V. Tan, and A. S. Willsky, "High-dimensional graphical model selection: tractable graph families and necessary conditions," in *Advances in Neural Information Processing Systems*, 2011, pp. 1863–1871.

[25] A. Ghoshal and J. Honorio, "Information-theoretic limits of bayesian network structure learning," in *Artificial Intelligence and Statistics*, 2017, pp. 767–775.

[26] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5111–5130, 2010.

[27] G. Cavraro, V. Kekatos, and S. Veeramachaneni, "Voltage analytics for power distribution network topology verification," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 1058–1067, 2017.

[28] Y. C. Chen, T. Banerjee, A. D. Dominguez-Garcia, and V. V. Veeravalli, "Quickest line outage detection and identification," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 749–758, 2015.

[29] Y. Sharon, A. M. Annaswamy, A. L. Motto, and A. Chakraborty, "Topology identification in distribution network with limited measurements," in *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*. IEEE, 2012, pp. 1–6.

[30] D. Deka, S. Backhaus, and M. Chertkov, "Structure learning in power distribution networks," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1061–1074, 2017.

[31] X. Li, H. V. Poor, and A. Scaglione, "Blind topology identification for power systems," in *2013 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2013, pp. 91–96.

[32] Y. Yuan, O. Ardakanian, S. Low, and C. Tomlin, "On the inverse power flow problem," *arXiv preprint arXiv:1610.06631*, 2016.

[33] J. Yu, Y. Weng, and R. Rajagopal, "Patopa: A data-driven parameter and topology joint estimation framework in distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4335–4347, 2017.

[34] S. Park, D. Deka, and M. Chcrtkov, "Exact topology and parameter estimation in distribution grids with minimal observability," in *2018 Power Systems Computation Conference (PSCC)*. IEEE, 2018, pp. 1–6.

[35] H. Zhu and G. B. Giannakis, "Sparse overcomplete representations for efficient identification of power line outages," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2215–2224, 2012.

[36] V. Y. Tan and A. S. Willsky, "Sample complexity for topology estimation in networks of lti systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 9079–9084, 2011.

[37] Y. Liao, Y. Weng, M. Wu, and R. Rajagopal, "Distribution grid topology reconstruction: An information theoretic approach," in *North American Power Symposium (NAPS), 2015*. IEEE, 2015, pp. 1–6.

[38] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 1659–1664.

[39] D. Deka, S. Backhaus, and M. Chertkov, "Estimating distribution grid topologies: A graphical learning based approach," in *Power Systems Computation Conference (PSCC), 2016*. IEEE, 2016, pp. 1–7.

[40] E. Candes, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," in *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*. IEEE, 2005, pp. 668–681.

[41] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[42] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," *IEEE signal processing magazine*, vol. 25, no. 2, p. 72, 2008.

[43] S. Li, L. Da Xu, and X. Wang, "Compressed sensing signal and data acquisition in wireless sensor networks and internet of things," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2177–2186, 2012.

[44] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 164–174, 2010.

[45] T. Li, M. Bakshi, and P. Grover, "Fundamental limits and achievable strategies for low energy compressed sensing with applications in wireless communication," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2016, pp. 1–6.

[46] M. Babakmehr, M. G. Simões, M. B. Wakin, and F. Harirchi, "Compressive sensing-based topology identification for smart grids," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 532–543, 2016.

[47] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 494–504, 2011.

[48] S. Jokar and V. Mehrmann, "Sparse solutions to underdetermined kronecker product systems," *Linear Algebra and its Applications*, vol. 431, no. 12, pp. 2437–2447, 2009.

[49] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," 2005.

[50] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.

[51] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE transactions on information theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[52] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[53] D. Ge, X. Jiang, and Y. Ye, "A note on the complexity of lp minimization," *Mathematical programming*, vol. 129, no. 2, pp. 285–299, 2011.

[54] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2011.

[55] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0," Aug. 2012. [Online]. Available: http://cvxr.com/cvx

[56] Gurobi Optimization, LLC, "Gurobi optimizer reference manual," 2018. [Online]. Available: http://www.gurobi.com

[57] *Bonneville Power Administration*, accessed on Oct. 2016. [Online]. Available: https://transmission.bpa.gov/ Business/Operations/Wind/

[58] A. Ghoshal and J. Honorio, "Information-theoretic limits of bayesian network structure learning," *arXiv preprint arXiv:1601.07460*, 2016.

[59] H. Kajimoto, "An extension of the prüfer code and assembly of connected graphs from their blocks," *Graphs and Combinatorics*, vol. 19, no. 2, pp. 231–239, 2003.

[60] A. Liebenau and N. Wormald, "Asymptotic enumeration of graphs by degree sequence, and the degree sequence of a random graph," *arXiv preprint arXiv:1702.08373*, 2017.

[61] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

Tongxin Li (S'2014) Tongxin Li is a doctoral student in Computing & Mathematical Sciences at the California Institute of Technology (Caltech). His research interest focuses primarily on problems related to graph learning, causal inference and data-driven control. Prior to joining Caltech in 2017, he received B.Sc. and B.Eng. in mathematics and information engineering from The Chinese University of Hong Kong.

Lucien Werner (S'2018) Lucien Werner completed undergraduate studies at Harvard University in 2013 and obtained an MS degree from Northwestern University in applied mathematics in 2015. Currently he is working towards a PhD at the California Institute of Technology in Computing & Mathematical Sciences, advised by Professor Steven Low and funded by the Resnick Graduate Fellowship. His interests are in power system and electricity market design under high renewable penetration.

Steven H. Low (F'2008) is the Gilloon Professor of the Department of Computing & Mathematical Sciences and the Department of Electrical Engineering at Caltech and an honorary professor of the University of Melbourne, Australia. He has also held honorary and professorship in Australia, China and Taiwan. Before that, he was with AT&T Bell Laboratories, Murray Hill, NJ, and the University of Melbourne, Australia. He was a co-recipient of IEEE best paper awards and is an IEEE Fellow. His research on communication networks has been accelerating more than 1TB of Internet traffic every second since 2014. His research on smart electric vehicle charging is providing large scale, cost effective and sustainable EV charging, to workplaces, from high schools to Fortune Global 500 companies. He was a member of the Networking and Information Technology Technical Advisory Group for the US Presidents Council of Advisors on Science and Technology (PCAST) in 2006. He received his B.S. from Cornell and PhD from Berkeley, both in EE. Biography text here.