HUB DISCOVERY IN PARTIAL CORRELATION GRAPHS

ALFRED HERO AND BALA RAJARATNAM

Abstract. One of the most important problems in large scale inference problems is the identification of variables that are highly dependent on several other variables. When dependency is measured by partial correlations these variables identify those rows of the partial correlation matrix that have several entries with large magnitudes; i.e., hubs in the associated partial correlation graph. This paper develops theory and algorithms for discovering such hubs from a few observations of these variables. We introduce a hub screening framework in which the user specifies both a minimum (partial) correlation ρ and a minimum degree δ to screen the vertices. The choice of ρ and δ can be guided by our mathematical expressions for the phase transition correlation threshold ρ_c governing the average number of discoveries. They can also be guided by our asymptotic expressions for familywise discovery rates under the assumption of large number p of variables, fixed number n of multivariate samples, and weak dependence. Under the null hypothesis that the dispersion (covariance) matrix is sparse these limiting expressions can be used to enforce familywise error constraints and to rank the discoveries in order of increasing statistical significance. For $n \ll p$ the computational complexity of the proposed partial correlation screening method is low and is therefore highly scalable. Thus it can be applied to significantly larger problems than previous approaches. The theory is applied to discovering hubs in a high dimensional gene microarray dataset.

Keywords

Gaussian graphical models, correlation networks, nearest neighbor dependency, node degree and connectivity, asymptotic Poisson limits, discovery rate phase transitions, p-value trajectories

1. Introduction

This paper treats the problem of screening a p-variate sample for strongly and multiply connected vertices in the partial correlation graph associated with the the partial correlation matrix of the sample. This problem, called hub screening, is important in many applications ranging from network security to computational biology to finance to social networks. In the area of network security, a node that becomes a hub of high correlation with neighboring nodes might signal anomalous activity such as a coordinated flooding attack. In the area of computational biology the set of hubs of a gene expression correlation graph can serve as potential targets for drug treatment to block a pathway or modulate host response. In the area of finance a hub might indicate a vulnerable financial instrument or sector whose collapse might have major repercussions on the market. In the area of social networks a hub of observed interactions between criminal suspects could be an influential ringleader.

The techniques and theory presented in this paper permit scalable and reliable screening for vertices that are singly connected or multiply connected. Unlike the correlations screening problem studied in [11], this paper considers the more challenging problem of partial correlation screening for variables with a specified degree of connectivity. In particular we consider 1) extension to screening for partial correlations exceeding a specified magnitude; 2) extension to screening variables whose vertex degree in the associated partial correlation graph[1] exceeds a specified degree. In contrast to previous asymptotic "$n \ll p$" studies, we do not require n to go to infinity with p. This "sample starved" finite n large p regime is one of the principal features of our approach.

While correlation graphs only reflect the marginal dependency between variables, partial correlation graphs characterize their patterns of conditional dependency. These partial correlations are encoded in the inverse covariance matrix and zeros in this matrix can be reflected via a graph. In the special case when the variables are distributed multivariate Gaussian, no edge exists between a pair of variables implies the two variables are conditionally independent given the remaining ones. A graphical model captures this Markovian conditional independence structure and is instantiated in the partial correlation graph. In applications partial correlation graphs are often considered more useful than correlation graphs as they are more interpretable, and can directly reflect possible underlying Markovian structure.

The hub screening theory presented here can be applied to structure discovery in p-dimensional Gaussian graphical models (GGM), a topic of recent interest to statisticians, computer scientists and engineers working in areas such as gene expression analysis, information theoretic imaging and sensor networks [7], [14], [22]. A GGM uses a $p \times p$ covariance matrix to capture patterns of statistical dependency in the p-variate measurement. The GGM specifies a graph over p vertices where edges are placed between any pair of vertices that corresponds to a non-zero entry of the inverse covariance matrix. As the partial correlation matrix and the inverse covariance matrix are equivalent (modulo a diagonal matrix transformation) a hub vertex in the GGM will also be a hub vertex in the partial correlation graph. Thus a GGM method such as Glasso could also be applied to discover vertices of high degree. However, our proposed hub screening method attacks the hub screening problem directly, is scalable to high dimensions, and comes with asymptotic theory to control false positives. As an example, the authors of [1] propose a Euclidean nearest neighbor graph method for testing independence in a GGM. When specialized to the null hypothesis of spatially independent Gaussian measurements, our results characterize the large p phase transitions and specify a weak (Poisson-like) limit law on the number of highly connected nodes in such nearest neighbor graphs for finite number n of observations.

Many different methods for inferring properties of correlation and partial correlation matrices have been recently proposed [7], [17], [18], [3], [15]. Several of these methods have been contrasted and compared in bioinformatics applications [8], [13], [16] similar to the one we consider in Sec. 5. The above papers address the *covariance selection* problem [6]: to find the non-zero entries in the covariance or inverse covariance matrix or, equivalently, to find the edges in the associated correlation or partial correlation graph.

[1]Partial correlation graphs are often called concentration graphs.

The problem treated and the solution proposed in this paper differ from those of these previous papers in several important ways: 1) as contrasted to covariance selection [6] our objective is to screen for connected vertices in the graph instead of to screen for edges; 2) unlike [3] our objective is to directly control false positives instead of maximizing a likelihood function or minimizing a matrix norm approximation error; 3) our framework is specifically adapted to the case of a finite number of samples and a large number of variables ($n \ll p$); 4) our asymptotic theory provides mathematical expressions for the p-value for each of the variables with respect to a sparse null hypothesis on the covariance; 5) unlike lasso type methods like [15] the hub screening implementation can be directly applied to very large numbers of variables without the need for ancillary prefiltering or variable reduction. Additional relevant literature on correlation based methods can be found in [11].

For specified $\rho \in [0, 1]$ and $\delta \in 1, \ldots, p-1$, a hub is defined broadly as any variable that is correlated with at least δ other variables having magnitude correlation exceeding ρ. Hub screening is accomplished by thresholding the sample correlation matrix or partial correlation matrix and searching for rows with more than δ non-zero entries. We call the former *correlation hub screening* and the latter *partial correlation hub screening*. The reader may question the advantage of a screening method that is based on two variables ρ and δ, but they can serve a useful purpose in the following sense, and can actually be an advantage. In practice, the magnitude of the correlation and the vertex degree are complementary quantities in differentiating salient graph properties. For example, a graph of high degree but low correlation can be of equal or great interest than a hub of low degree with a high correlation. Therefore imposing any simple linear ordering of these variables would be less informative since it would deprive the experimenter of stratifying the analysis into hubs of increasing degree, thus enhancing the interpretability of the results of the screening procedure.

The screening is performed in a computationally efficient manner by exploiting the equivalence between correlation graphs and ball graphs on the set of Z-scores. Specifically, assume that n samples of p variables are available in the form of a data matrix where $n < p$. First the columns of the data matrix are converted to standard n-variate Z-scores. The set of p Z-scores uniquely determine the sample correlation matrix. If partial correlations are of interest, these Z-scores are replaced by equivalent modified Z-scores that characterize the sample partial correlation matrix, defined as the Moore-Penrose pseudo-inverse of the sample correlation matrix. Then an approximate k-nearest neighbor (ANN) algorithm is applied to the Z-scores or the modified Z-scores to construct a ball graph associated with the given threshold ρ. Hub variables are discovered by scanning the graph for those whose vertex degree exceeds δ. The ANN approach only computes a small number of the sample correlations or partial correlations, circumventing the difficult (or impossible) task of computing all entries of the correlation matrix when there are millions (or billions) of variables. State-of-the-art ANN software has been demonstrated on over a billion variables [12] and thus our proposed hub screening procedure has potential application to problems of massive scale. We also note that using the standard Moore-Penrose inverse is well understood to be a sub-optimal estimator of the partial correlation matrix in terms of minimum mean square error [9]. To our knowledge its properties for screening for partial correlations have yet to be investigated. This paper demonstrates through theory and experiment that the Moore-Penrose inverse can be used to screen for hubs in partial correlation graphs.

No screening procedure would be complete without error control. We establish limiting expressions for mean hub discovery rates. These expressions are used to obtain an approximate phase transition threshold ρ_c below which the average number of hub discoveries abruptly increases. When the screening threshold ρ is below ρ_c the discoveries are likely to be dominated by false positives. We show that the probability of false positives can be approximated by a Poisson-like probability of the form $P(N_{\delta,\rho} > 0) = 1 - e^{-\lambda}$. This Poisson-like approximation becomes more accurate in the limit as ρ approaches 1 and p goes to infinity.

In the case of independent identically distributed (i.i.d.) elliptically distributed samples and sparse block diagonal dispersion matrix, the Poisson rate does not depend on the unknown correlations. In this case we can specify asymptotic p-values on hub discoveries of given degree under a sparse dispersion matrix null model. Finite p bounds on the Poisson p-value approximation error are given that decrease at rates determined by p, δ, ρ, and the sparsity factor of the dispersion matrix.

To illustrate the power of the proposed hub screening method we apply it to a public gene expression dataset: the NKI breast cancer data [4]. Each of these datasets contains over twenty thousand variables (genes) but many fewer observations (GeneChips). In addiiton to recapitulating results of previous studies of this data, our screening method reveals interesting and previously unreported dependency structure among the variables. For the purposes of exploring neighborhood structure of the discoveries we introduce a waterfall plot of their approximate p-values that plots the family of degree-indexed p-value curves over the range of partial correlation thresholds. This graphic rendering can provide insight into the structure and significance of the correlation neighborhoods as we sweep the variables over different vertex degree curves in the waterfall plot.

The outline of this paper is as follows. In Sec. 2 we formally define the hub screening problem. In Sec. 2.3 we present the Z-score representation for the pseudo-inverse of the sample correlation matrix. In Sec. 3 we give an overview of the results pertaining to phase transition thresholds and limit theorems for the familywise discovery rates and p-values. This section also describes the proposed hub screening procedure. Section 4 gives the formal statements of the results in the paper. The proofs of these results are given in the appendix. In Sec. 5 we validate the theoretical predictions by simulation and illustrate the proposed hub screening procedure on gene microarray data.

2. Hub screening framework

Let the p-variate $\mathbf{X} = [X_1, \ldots, X_p]^T$ have mean μ and non-singular $p \times p$ dispersion matrix $\mathbf{\Sigma}$. We will often assume that $\mathbf{X}$ has an elliptically contoured density: $f_{\mathbf{X}}(\mathbf{x}) = g\left((\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$ for some non-negative strictly decreasing function g on $\mathbb{R}^+$. This family of densities includes the multivariate Gaussian, in which case $\mathbf{\Sigma}$ is the covariance matrix, and the multivariate Student-t densities as special cases. The correlation matrix and the partial correlation matrix are defined as $\mathbf{\Gamma} = \mathbf{D}_{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{D}_{\Sigma}^{-1/2}$ and $\mathbf{\Omega} = \mathbf{D}_{\Sigma^{-1}}^{-1/2}\mathbf{\Sigma}^{-1}\mathbf{D}_{\Sigma^{-1}}^{-1/2}$, respectively, where for a square matrix $\mathbf{A}$, $\mathbf{D}_A = \mathrm{diag}(\mathbf{A})$ denotes the diagonal matrix obtained from $\mathbf{A}$ by zeroing out all entries not on its diagonal.

Available for observation is a $n \times p$ data matrix $\mathbb{X}$ whose rows are (possibly dependent) replicates of $\mathbf{X}$:

$$\mathbb{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_p] = [\mathbf{X}_{(1)}^T, \cdots, \mathbf{X}_{(n)}^T]^T,$$

where $\mathbf{X}_i = [X_{1i}, \ldots, X_{ni}]^T$ and $\mathbf{X}_{(i)} = [X_{i1}, \ldots, X_{ip}]$ denote the i-th column and row, respectively, of $\mathbb{X}$. Define the sample mean of the i-th column $\overline{X}_i = n^{-1}\sum_{j=1}^n X_{ji}$, the vector of sample means $\overline{\mathbf{X}} = [\overline{X}_1, \ldots, \overline{X}_p]$, the $p \times p$ sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{X}_{(i)} - \overline{\mathbf{X}})^T(\mathbf{X}_{(i)} - \overline{\mathbf{X}})$, and the $p \times p$ sample correlation matrix

$$\mathbf{R} = \mathbf{D}_{\mathbf{S}}^{-1/2}\mathbf{S}\mathbf{D}_{\mathbf{S}}^{-1/2}. \tag{1}$$

For a full rank sample correlation matrix $\mathbf{R}$ the sample partial correlation matrix is defined as

$$\mathbf{P} = \mathbf{D}_{R^{-1}}^{-1/2}\mathbf{R}^{-1}\mathbf{D}_{R^{-1}}^{-1/2}.$$

In the case that $\mathbf{R}$ is not full rank this definition must be modified. Several methods have been proposed for regularizing the inverse of a rank deficient covariance including shrinkage and pseudo-inverse approaches [19]. In this paper we adopt the pseudo-inverse approach and define the sample partial correlation matrix as

$$\mathbf{P} = \mathbf{D}_{\mathbf{R}^\dagger}^{-1/2}\mathbf{R}^\dagger\mathbf{D}_{\mathbf{R}^\dagger}^{-1/2}, \tag{2}$$

where $\mathbf{R}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{R}$.

2.1. **Correlation and partial correlation graphs.** Let the non-negative definite symmetric matrix $\mathbf{\Phi} = ((\Phi_{ij}))_{i,j=1}^p$ be generic notation for a correlation-type matrix like $\mathbf{\Gamma}$, $\mathbf{\Omega}$, $\mathbf{R}$, or $\mathbf{P}$. For a threshold $\rho \in [0,1]$ define $\mathcal{G}_\rho(\mathbf{\Phi})$ the undirected graph induced by thresholding $\mathbf{\Phi}$ as follows. The graph $\mathcal{G}_\rho(\mathbf{\Phi})$ has vertex set $\mathcal{V} = \{1, \ldots, p\}$ and edge set $\mathcal{E} = \{e_{ij}\}_{i,j\in\{1,\ldots,p\}:i<j}$, where an edge $e_{ij} \in \mathcal{E}$ exists in $\mathcal{G}_\rho(\mathbf{\Phi})$ if $|\Phi_{ij}| \geq \rho$. The degree of the i-th vertex of $\mathcal{G}_\rho(\mathbf{\Phi})$ is $\text{card}\,\{j \neq i : |\Phi_{ij}| \geq \rho\}$, the number of edges that connect to i. When $\mathbf{\Phi}$ is set to $\mathbf{\Gamma}$ or $\mathbf{\Omega}$ we obtain the population correlation graph $\mathcal{G}_\rho(\mathbf{\Gamma})$ and the population partial correlation graph $\mathcal{G}_\rho(\mathbf{\Omega})$ [5]. Likewise, when $\mathbf{\Phi}$ is set to $\mathbf{R}$ or $\mathbf{P}$ we obtain the sample correlation graph $\mathcal{G}_\rho(\mathbf{R})$ and the sample partial correlation graph $\mathcal{G}_\rho(\mathbf{P})$

A $p \times p$ matrix is said to be row-sparse of degree k, called the *sparsity degree*, if no row contains more than $k+1$ non-zero entries. When $\mathbf{\Phi}$ is row-sparse of degree k its graph $\mathcal{G}_\rho(\mathbf{\Phi})$ has no vertex of degree greater than k. A special case is a block-sparse matrix of degree k; a matrix that can be reduced via row-column permutation to block diagonal form having a single $k \times k$ block.

2.2. **Hub discoveries.** A given vertex i of the correlation graph is declared a hub screening discovery at degree level δ and threshold level ρ if the observed vertex degree δ_i in $\mathcal{G}_\rho(\mathbf{\Phi})$ is at least δ. More specifically

$$\delta_i = \text{card}\,\{j : j \neq i, |\Phi_{ij}| \geq \rho\} \geq \delta, \tag{3}$$

where $\mathbf{\Phi}$ is equal to $\mathbf{R}$ for correlation hub screening or is equal to $\mathbf{P}$ for partial correlation hub screening. We denote by $N_{\delta,\rho} \in \{0, \ldots, p\}$ the total number of hub screening discoveries at degree level δ

$$N_{\delta,\rho} = \text{card}\{i : \delta_i \geq \delta\}.$$

There will generally be false positives among the discoveries and, to be practically useful, these must be predicted as a function of screening parameters ρ and δ in (3). In the sequel we will develop a large p asymptotic analysis to address this prediction problem and establish two results: 1) existence of phase transitions in the mean number of discoveries $E[N_{\delta,\rho}]$; 2) asymptotic expressions for familywise false positive rate $P(N_{\delta,\rho} > 0)$.

2.3. **Z-score representation.** Define the $n \times p$ matrix of Z-scores associated with the data matrix $\mathbb{X}$

$$\mathbb{T} = [\mathbf{T}_1, \ldots, \mathbf{T}_p] = (n-1)^{-1/2}(\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbb{X}\mathbf{D}_{\mathbf{S}}^{-1/2}, \tag{4}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^n$. This Z-score matrix is to be distinguished from the $(n-1) \times p$ Z-score matrices $\mathbb{U}$ and $\mathbb{Y}$, denoted collectively by the notation $\mathbb{Z}$ in the sequel, that are derived from the matrix $\mathbb{T}$.

We exploit the following Z-score representation of the sample correlation matrix

$$\mathbf{R} = \mathbb{T}^T\mathbb{T}, \tag{5}$$

and define a set of equivalent but lower dimensional Z-scores called U-scores. The U-scores lie in the unit sphere S_{n-2} in $\mathbb{R}^{n-1}$ and are obtained by projecting away the rowspace components of $\mathbb{T}$ in the direction of vector $\mathbf{1}$. More specifically, they are constructed as follows. Define the orthogonal $n \times n$ matrix $\mathbf{H} = [n^{-1/2}\mathbf{1}, \mathbf{H}_{2:n}]$. The matrix $\mathbf{H}_{2:n}$ can be obtained by Gram-Schmidt orthogonalization of the columns of $[n^{-1/2}\mathbf{1}, \mathbf{I}_{n,n-1}]$, where $\mathbf{I}_{n,n-1}$ is a matrix consisting of the last $n-1$ columns of $\mathbf{I}_n$. It satisfies the properties

$$\mathbf{1}^T\mathbf{H} = [\sqrt{n}, 0, \ldots, 0], \quad {\mathbf{H}_{2:n}}^T\mathbf{H}_{2:n} = \mathbf{I}_{n-1}.$$

The U-score matrix $\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$ is obtained from $\mathbb{T}$ by the following relation

$$\begin{bmatrix} \mathbf{0}^T \\ \mathbb{U} \end{bmatrix} = \mathbf{H}^T\mathbb{T}. \tag{6}$$

Using representation (6) we obtain the following result.

Lemma 1. *Assume that $n < p$. The Moore-Penrose pseudo-inverse of $\mathbf{R}$ has the representation*

$$\mathbf{R}^\dagger = \mathbb{U}^T[\mathbb{U}\mathbb{U}^T]^{-2}\mathbb{U}. \tag{7}$$

The proof of Lemma 7 simply verifies that $\mathbf{Q} \stackrel{\text{def}}{=} \mathbb{U}^T[\mathbb{U}\mathbb{U}^T]^{-2}\mathbb{U}$ satisfies the Moore-Penrose conditions for $\mathbf{Q}$ to be the unique pseudo-inverse of $\mathbf{R}$: 1) the matrices $\mathbf{QR}$ and $\mathbf{RQ}$ are symmetric; 2) $\mathbf{RQR} = \mathbf{R}$; and 3) $\mathbf{QRQ} = \mathbf{Q}$ [10].

The representation (7) leads to a Z-score representation of the sample partial correlation similar to (5) that allows us to unify the treatment of the sample correlation and sample partial correlation. Specifically, using Lemma 7 the sample partial correlation matrix (2) can be representated as:

$$\mathbf{P} = \mathbb{Y}^T\mathbb{Y}, \tag{8}$$

where $\mathbb{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_p]$ is an $(n-1) \times p$ matrix of partial correlation Z-scores

$$\mathbb{Y} = [\mathbb{U}\mathbb{U}^T]^{-1}\mathbb{U}\mathbf{D}^{-1/2}_{\mathbb{U}^T[\mathbb{U}\mathbb{U}^T]^{-2}\mathbb{U}}, \tag{9}$$

with $\mathbf{Y}_i \in S_{n-2}$.

3. Overview of results

This section provides a high level overview of the technical material presented in Sec. 4 and explains its practical utility for hub screening. Propositions 1, 2, and 3 establish our two main results: 1) the existence of a phase transition, as a function of the applied threshold ρ, in the mean number of discoveries, explained in Sec. 3.1; and 2) an asymptotic Poisson-like limit of the probability of discoveries, explained in Sec. 3.2. In Sec. 5.1 the proposed screening algorithm is defined and we explain how the asymptotic Poisson-like limit can be used to predict the probability of false positives under the hypothesis that the rows of the data matrix $\mathbb{X}$ are i.i.d. with an elliptically contoured distribution and sparse dispersion matrix $\mathbf{\Sigma}$.

3.1. **Phase transitions in the mean number of hub discoveries.** There is a phase transition in the mean number $E[N_{\delta,\rho}]$ of hub discoveries of degree δ that depends on the applied screening threshold ρ. This critical phase transition threshold, which we call $\rho_{c,\delta}$, is such that if the screening threshold ρ decreases below $\rho_{c,\delta}$, the of hub discoveries of degree δ abruptly increases to the maximum p. The mathematical form of the critical phase transition threshold is the same for correlation graphs and partial correlation graphs. An asymptotic expression for the critical threshold is obtained from the limiting form (21) of $E[N_{\delta,\rho}]$ given in Prop. 2

$$\rho_{c,\delta} = \sqrt{1 - (c_{n,\delta}(p-1))^{-2\delta/(\delta(n-2)-2)}}, \tag{10}$$

where $c_{n,\delta} = \mathrm{a}_n \delta J_{p,\delta}$and $\mathrm{a}_n = 2B((n-2)/2, 1/2)$ with $B(i,j)$ denoting the beta function. The unknown dispersion matrix $\mathbf{\Sigma}$ influences $\rho_{c,\delta}$ only through the quantity $J_{p,\delta} = J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}})$, defined in (33), which is a measure of average $(\delta+1)$-order dependency among the Z-scores $\{\mathbf{Z}_i\}_{i=1}^p$.

When the rows of $\mathbb{X}$ are i.i.d. elliptically distributed and $\mathbf{\Sigma}$ is block-sparse of degree k then, from Prop. 3

$$J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}) = 1 + O\left((k/p)^{\gamma_\delta}\right), \tag{11}$$

where $\gamma_\delta = \delta + 1$ for correlation hub screening and $\gamma_\delta = 1$ for partial correlation hub screening. In the extreme case where $\mathbf{\Sigma}$ is diagonal, $k = 0$ and $J_{p,\delta} = 1$. In this case all discoveries are false positives. Otherwise, if k increases more slowly than p then the remainder $O\left((k/p)^{\gamma_\delta}\right)$ goes to zero and the phase transition threshold $\rho_{c,\delta}$ no longer depends on $\mathbf{\Sigma}$, reducing to the false positive phase transition threshold one would obtain if $\mathbf{\Sigma}$ were diagonal.

For large p, $c_{n,\delta}$ depends only weakly on p and the critical threshold increases to 1 at rate $O((p-1)^{-2\delta/(\delta(n-2)-2)})$, which is close to logarithmic in p for large n ($n \gg \log p$) but much faster than logarithmic for small n ($n \ll \log p$). Fig. 1 plots the false positive critical

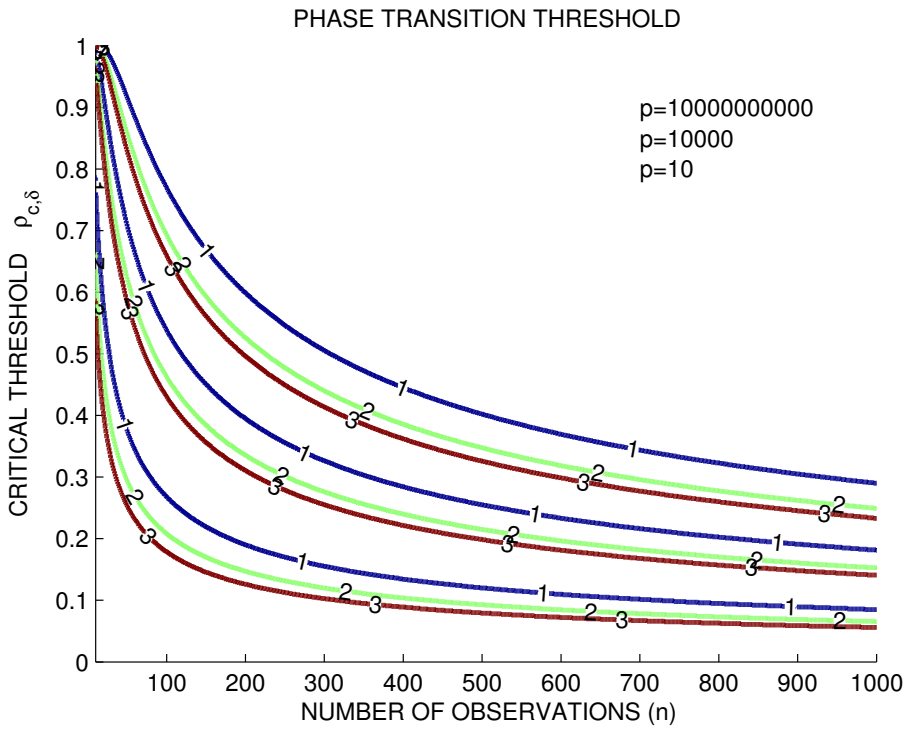

FIGURE 1. Critical phase transition threshold $\rho_{c,\delta}$ governing the mean number of false positives in the number of hub discoveries as a function of number of observations n for various vertex degrees $\delta = 1, 2, 3$ and (from bottom to top curve groups) variable dimensions $p = 10$, $p = 10,000$, and $p = 10 \times 10^9$. When the correlation threshold is greater than $\rho_{c,\delta}$ the number of false positives falls rapidly to zero. The figure shows that the critical threshold decreases as either n or δ increase. The critical threshold is close to one when n is small in which case reliable detection of hubs is impossible. However, a relatively small increase in sample size is sufficient to reduce the critical threshold even for very large p. For example, with $p = 10$ billion variables only $n = 200$ samples are required to bring $\rho_{c,1}$ down to 0.6.

threshold (diagonal $\mathbf{\Sigma}$) as a function of n for several values of δ and p. The critical threshold decreases as either the sample size n increases, the number of variables p decreases, or the vertex degree δ increases. Remarkably, even for ten billion samples (upper curves on the figure) only a relatively small number of samples are necessary for hub screening to be useful. For example, with $n = 200$ one can reliably discover connected vertices ($\delta = 1$ in the figure) having partial correlation greater than $\rho_{c,\delta} = 0.6$.

3.2. **Poisson-type limits on probability of hub discoveries.** For fixed n, as $p \to \infty$ and $\rho \to 1$ Prop. 2 establishes that, under certain conditions, the probability of hub discoveries $P(N_{\delta,\rho} > 0)$ converges to a Poisson-type limit $1 - \exp(-\Lambda_{\delta,\rho})$ where $\Lambda_{\delta,\rho}$ is the rate parameter

of a related Poisson distributed random variable $N^*_{\delta,\rho}$. A sufficient condition for the Poisson-type limit to hold is that: 1) ρ increases to one with p at a prescribed rate depending on n; 2) the covariance matrix $\mathbf{\Sigma}$ is block sparse of degree k with $k = o(p)$. For correlation graphs ($\mathbf{\Phi} = \mathbf{R}$) the second condition can be relaxed to 2') the covariance matrix is row sparse of degree k with $k = o(p)$

The rate of convergence is provided in Prop. 1 along with a finite p approximation to the Poisson rate parameter $\Lambda_{\delta,\rho}$

$$\Lambda_{\delta,\rho} = \lambda_{\delta,\rho} J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}), \tag{12}$$

with

$$\lambda_{\delta,\rho} = \lim_{p\to\infty} p\binom{p-1}{\delta} P_0(\rho_p, n)^\delta, \tag{13}$$

and $P_0(\rho, n)$ is the spherical cap probability defined in (26). It is important to note that the parameter $\lambda_{\delta,\rho}$ does not depend on the distribution of the data $\mathbb{X}$.

When the rows of $\mathbb{X}$ are i.i.d. elliptically distributed with dispersion matrix $\mathbf{\Sigma}$ block-sparse of degree k, Prop. 3 establishes two things: 1) relation (11), so that $\Lambda_{\delta,\rho} \approx \lambda_{\delta,\rho}$; and 2) $\|\Delta_{p,n,k,\delta}\|_1$ is equal to zero for correlation hub screening and is of order $O(k/p)$ for partial correlation hub screening.

Therefore, Prop. 3 and 2 imply that when the block-sparse model is posed as the null hypothesis the false positive familywise error rate (FWER) can be approximated as

$$P(N_{\delta,\rho} > 0) = 1 - \exp(-\lambda_{\delta,\rho}). \tag{14}$$

The accuracy of the approximation (14) is specified by the bound (20) given in Prop. 1. Corollary 1 provides rates of convergence under the assumptions that $p(p-1)^\delta(1-\rho^2)^{(n-2)/2} = O(1)$ and the rows of $\mathbb{X}$ are i.i.d. with sparse covariance. For example, assume that the covariance is block-sparse of degree k. If k does not grow with p then the rate of convergence of $P(N_{\delta,\rho} > 0)$ to its Poisson limit is no worse than $O(p^{-1/\delta})$ for $\delta > n - 3$. On the other hand, if k grows with rate at least $O(p^{1-\alpha})$, for $\alpha = \min\{(\delta+1)^{-1}, (n-2)^{-1}\}/\delta$, the rate of convergence is no worse than $O\left(k/p\right)$. This latter bound can be replaced by $O\left((k/p)^{\delta+1}\right)$ for correlation hub screening under the less restrictive assumption that the covariance is row-sparse.

The combination of Prop. 1 and the assertions (Prop. 3) that $J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}) = 1 + O\left((k/p)^{\gamma_\delta}\right)$ and $\|\Delta_{p,n,k,\delta}\|_1 \leq O(k/p)$ yields

$$|P(N_{\delta,\rho} > 0) - (1 - \exp(-\lambda_{\delta,\rho}))| \leq \begin{cases} O\left(\max\left\{(k/p)^{\gamma_\delta}, p^{-(\delta-1)/\delta}(k/p)^{\delta-1}, p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right), & \delta > 1 \\ O\left(\max\left\{(k/p)^{\gamma_\delta}, (k/p)^2, p^{-1}, (1-\rho)^{1/2}\right\}\right), & \delta = 1 \end{cases}.$$

The terms $(k/p)^{\gamma_\delta}$, $p^{-(\delta-1)/\delta}(k/p)^{\delta-1}, p^{-1/\delta}$ and $(1-\rho)^{1/2}$ respectively quantify the contribution of errors due to: 1) insufficient sparsity in the covariance or, equivalently, the correlation graph; 2) excessive dependency among neighbor variables in this graph; 3) poor convergence of $E[N_{\delta,\rho}]$; and 4) inaccurate mean-value approximation of the integral representation of $\lim_{p\to\infty} E[N_{\delta,\rho}]$ by (38). One of these terms will dominate depending on the regime of operation. For example, specializing to partial correlation hub screening ($\gamma_\delta = \delta + 1$), if $\delta > 1$

and $O\left(p^{1-(\delta+1)/(2\delta)}\right) \leq k \leq o(p)$ then $(k/p)^{\delta+1} > p^{-(\delta-1)/\delta}(k/p)^{\delta-1}$ and the deficiency in the Poisson probability approximation will not be the determining factor on convergence rate.

3.3. **Asymptotic p-values and waterfall plots.** Under the null hypothesis of block sparse $\mathbf{\Sigma}$ the FWER approximation (14) can be used to assess the statistical significance of each discovery. Let the sample correlations between the Z-score of variable i and all the other $p-1$ variables be ordered as $\rho_i(1) < \rho_i(2) < \ldots < \rho_i(p-1)$, where $\rho_i(k)$ denotes the correlation (or partial correlation) between the i-th variable and its k-th nearest neighbor in the complete correlation (or partial correlation) graph $\mathcal{G}_0(\mathbf{\Phi})$, $\mathbf{\Phi} = \mathbf{R}$ or $\mathbf{P}$. The p-value associated with discovery i at degree level δ is defined as

$$pv_\delta(i) = 1 - \exp(-\lambda_{\delta,\rho_i(\delta)}). \tag{15}$$

where $\lambda_{\delta,\rho}$ is specified in (13). The quantity (15) approximates the probability that at least one hub vertex of degree greater than or equal to δ would be discovered in $\mathcal{G}_\rho(\mathbf{\Phi})$ using a threshold ρ equal to the observed correlation value $\rho_i(\delta)$.

Additional useful information can be gleaned by graphical rendering of the aggregate lists of p-values. Assume that the hub screening procedure generates an associated family of graphs $\{\mathcal{G}_\rho(\mathbf{\Phi})\}_{\rho\in[0,\rho^*]}$, where ρ^* is an intital threshold. We define the *waterfall plot of p-values* as the family of curves, plotted against the thresholds $\rho_i(\delta)$, indexed by $\delta = 1, 2, \ldots$,where the δ-th curve is formed from the (linearly interpolated) ordered list of p-values $\{pv_\delta(i_j)\}_{j=1}^p$, $pv_\delta(i_1) \geq \ldots \geq pv_\delta(i_p)$ (see Fig. 3).

4. Main theorems

The asymptotic theory for hub discovery in correlation and partial correlation graphs is presented in the form of three propositions and one corollary. Prop. 1 gives a general bound on the finite sample approximation error associated with the approximation of the mean and probability of discoveries given in Prop. 2. The results of Props. 1 and 2 apply to general random matrices of the form $\mathbb{Z}^T\mathbb{Z}$ where the p columns of $\mathbb{Z}$ lie on the unit sphere $S_{n-2} \subset \mathbb{R}^{n-1}$ and, in view of (6) and (8), they provide a unified theory of hub screening for correlation graphs and partial correlation graphs. Corollary 1 specializes the bounds presented in Prop. 1 to the case of sparse correlation graphs using Prop. 3.

For $\delta \geq 1$, $\rho \in [0,1]$, and $\mathbf{\Phi}$ equal to the sample correlation matrix $\mathbf{R}$ or the sample partial covariance matrix $\mathbf{P}$ we recall the definition of $N_{\delta,\rho}$ as the number of vertices of degree at least δ in $\mathcal{G}_\rho(\mathbf{\Phi})$. Define $\tilde{N}_{\delta,\rho}$ as the number of subgraphs in $\mathcal{G}_\rho$ that are isomorphic to a star graph with δ edges. In the sequel we will use the key property that $N_{\delta,\rho} = 0$ if and only if $\tilde{N}_{\delta,\rho} = 0$.

For $\delta \geq 1$, $\rho \in [0,1]$, and $n > 2$ define

$$\Lambda = \xi_{p,n,\delta,\rho} J(\overline{f_{\mathbf{Z}_{*1},\ldots,\mathbf{Z}_{*\delta+1}}}) \tag{16}$$

where

$$\xi_{p,n,\delta,\rho} = p\binom{p-1}{\delta} P_0^\delta, \tag{17}$$

$P_0 = P_0(\rho, n)$ is defined in (26), J is given in (33), and $\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}$ is the average joint density given in (30).

We also define the following quantity needed for the bounds of Prop. 1

$$\eta_{p,\delta} = p^{1/\delta}(p-1)P_0. \tag{18}$$

Note that $\xi_{p,n,\delta,\rho}/\eta_{p,\delta}^{\delta} = (\mathrm{a}_n/(n-2))^{\delta}/\delta!$ to order $O\left(\max\{p^{-1}, 1-\rho\}\right)$, where $\mathrm{a}_n = (2\Gamma((n-1)/2))/(\sqrt{\pi}\Gamma((n-2)/2))$. Let $\varphi(\delta)$ be the function equal to 1 for $\delta > 1$ and equal to 2 for $\delta = 1$.

Proposition 1. *Let $\mathbb{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$ be a $(n-1) \times p$ random matrix with $\mathbf{Z}_i \in S_{n-2}$. Fix integers δ and n where $\delta \geq 1$ and $n > 2$. Let the joint density of any subset of the $\mathbf{Z}_i$'s be bounded and differentiable. Then, with Λ defined in (16),*

$$|E[N_{\delta,\rho}] - \Lambda| \leq O\left(\eta_{p,\delta}^{\delta} \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right) \tag{19}$$

Furthermore, let $N_{\delta,\rho}^*$ be a Poisson distributed random variable with rate $E[N_{\delta,\rho}^*] = \Lambda/\varphi(\delta)$. If $(p-1)P_0 \leq 1$ then, for any integer k, $1 \leq k \leq p$,

$$\left|P(N_{\delta,\rho} > 0) - P(N_{\delta,\rho}^* > 0)\right| \leq \tag{20}$$
$$\begin{cases} O\left(\eta_{p,\delta}^{\delta} \max\left\{\eta_{p,\delta}^{\delta} (k/p)^{\delta+1}, Q_{p,k,\delta}, \|\Delta_{p,n,k,\delta}\|_1, p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right), & \delta > 1 \\ O\left(\eta_{p,1} \max\left\{\eta_{p,1} (k/p)^2, \|\Delta_{p,n,k,1}\|_1, p^{-1}, (1-\rho)^{1/2}\right\}\right), & \delta = 1 \end{cases},$$

with $Q_{p,k,\delta} = \eta_{p,\delta}^{\delta-1} p^{-(\delta-1)/\delta} (k/p)^{\delta-1}$ and $\|\Delta_{p,n,k,\delta}\|_1$ defined in (32).

The proof of the above proposition is given in the Appendix. The Poisson-type limit (22) is established by showing that the count $\tilde{N}_{\rho,\delta}$ of the number of groups of δ mutually coincident edges in $\mathcal{G}_\rho$ converges to a Poisson random variable with rate $\Lambda/\varphi(\delta)$.

Proposition 2. *Let $\rho_p \in [0,1]$ be a sequence converging to one as $p \to \infty$ such that $p^{1/\delta}(p-1)(1-\rho_p^2)^{(n-2)/2} \to e_{n,\delta} \in (0,\infty)$. Then*

$$\lim_{p\to\infty} E[N_{\delta,\rho_p}] = \kappa_{n,\delta} \lim_{p\to\infty} J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}), \tag{21}$$

where $\kappa_{n,\delta} = \left(e_{n,\delta}\mathrm{a}_n/(n-2)\right)^{\delta}/\delta!$. Assume that the weak dependency condition $\lim_{p,k\to\infty} \|\Delta_{p,n,k,\delta}\|_1 = 0$ is satisfied for some $k = o(p)$, where $\|\Delta_{p,n,k,\delta}\|_1$ is defined in (32). Then

$$P(N_{\delta,\rho_p} > 0) \to 1 - \exp(-\Lambda/\varphi(\delta)). \tag{22}$$

The proof of Prop. 2 is an immediate and obvious consequence of Prop. 1 and is omitted. For correlation graphs ($\mathbf{\Phi} = \mathbf{R}$), $\|\Delta_{p,n,k,\delta}\|_1 = 0$ when $\mathbf{\Sigma}$ is row sparse of degree k. Row sparsity is thus a sufficient condition that guarantees the weak dependency condition in Prop. 2. For partial correlation screening a sufficient condition is that the covariance matrix be block sparse of degree k.

Propositions 1 and 2 are general results that apply to both correlation hub and partial correlation hub screening under a wide range of conditions. Corollary 1 specializes these results to the case of sparse covariance and i.i.d. rows of $\mathbb{X}$ having elliptical distribution.

Corollary 1. *In addition to the hypotheses of Prop. 2 assume that $n > 3$ and that the rows of $\mathbb{X}$ are i.i.d. elliptically distributed with a covariance matrix $\mathbf{\Sigma}$ that is row-sparse of degree k. Assume that k grows as $O\left(p^{1-\alpha}\right) \leq k \leq o(p)$ where $\alpha = \min\left\{(\delta+1)^{-1}, (n-2)^{-1}\right\}/\delta$. Then, for correlation hub screening the asymptotic approximation error in the limit (22) is upper bounded by $O\left((k/p)^{\delta+1}\right)$. Under the additional assumption that the covariance is block-sparse, for partial correlation hub screening this error is upper bounded by $O\left(k/p\right)$.*

The proof of Corollary 1 is given in the Appendix. The proposition below specializes these results to sparse covariance.

Proposition 3. *Let $\mathbb{X}$ be a $n \times p$ data matrix whose rows are i.i.d. realizations of an elliptically distributed p-dimensional vector $\mathbf{X}$ with mean parameter μ and covariance parameter $\mathbf{\Sigma}$. Let $\mathbb{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$ be the matrix of correlation Z-scores (6) and $\mathbb{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_p]$ be the matrix of partial correlation Z-scores (9) defined in Sec. 2.3. Assume that the covariance matrix $\mathbf{\Sigma}$ is block-sparse of degree q. Then the pseudo-inverse partial correlation matrix $\mathbf{P} = \mathbb{Y}^T\mathbb{Y}$ has the representation*

$$\mathbf{P} = \mathbb{U}^T\mathbb{U}(1 + O(q/p)). \tag{23}$$

Let $\mathbf{Z}_i$ denote $\mathbf{U}_i$ or $\mathbf{Y}_i$ and assume that for $\delta \geq 1$ the joint density of any distinct set of Z-scores $\mathbf{U}_{i_1}, \ldots, \mathbf{U}_{i_{\delta+1}}$ is bounded and differentiable over $S_{n-2}^{\delta+1}$. Then the $(\delta+1)$-fold average function J (30) and the dependency coefficient $\Delta_{p,n,k,\delta}$ (32) satisfy

$$J(\overline{f_{\mathbf{Z}_{*1},\ldots,\mathbf{Z}_{*\delta+1}}}) = 1 + O\left(\delta q/p\right), \tag{24}$$

$$\|\Delta_{p,n,k,\delta}\|_1 = \begin{cases} O\left(\delta q/p\right), & \varphi = 1 \\ 0, & \varphi = 0 \end{cases} \tag{25}$$

where $\varphi = 0$ and $\varphi = 1$ for correlation and partial correlation hub screening, respectively.

Proof of Proposition 3:

The proof of Proposition 3 is given in the Appendix. □

5. Experiments

In this section we illustrate the hub screening proposed procedure, summarized in Sec. 5.1, for three different datasets. The first two are simulation experiments that have ground truth. These consist of a low dimensional toy example, which illustrates the interpretation of the p-value waterfall plots, and a high dimensional sham dataset that illustrates the fidelity of our error control. The third is a real gene expression dataset with no ground truth but for which the proposed procedure recapitulates results of previously published analysis.

5.1. Hub screening procedure.

We summarize the main steps of the proposed procedure as follows. First an initial correlation threshold ρ^* is selected close to the critical phase transition threshold (10). This threshold is applied to the sample correlation matrix $\mathbf{\Phi} = \mathbf{R}$ (1) or to the sample partial correlation matrix $\mathbf{\Phi} = \mathbf{P}$ (2) to generate the graph $\mathcal{G}_{\rho^*}(\mathbf{\Phi})$. The p-value expression (15) is then used to compute the waterfall plot of p-values on the

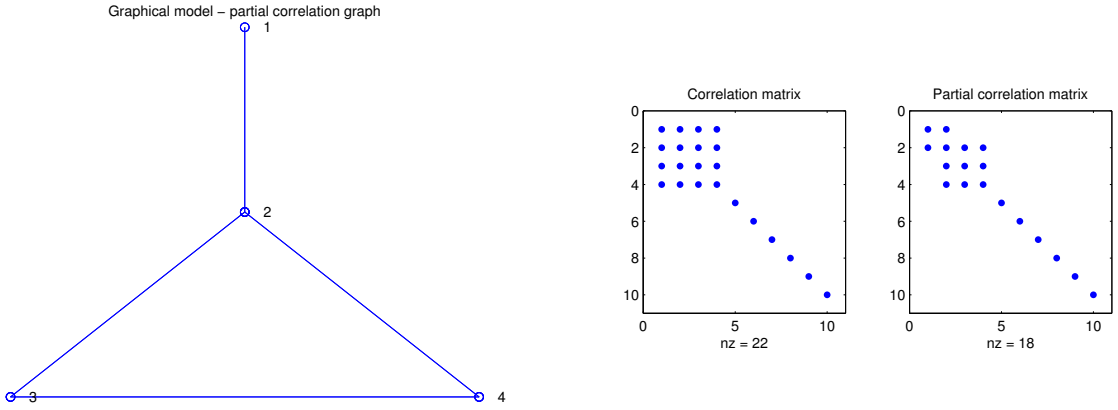

FIGURE 2. Left: diagram of the graphical model governing the first 4 variables among the vector of 1000 variables in the simulation study. The remaining 996 variables are i.i.d and uncorrelated with the first 4 variables. Right: first 10 rows and columns of the sparse correlation and partial correlation matrices. The top 4×4 block in the correlation is a full matrix while the corresponding block in the partial correlation has zero entries corresponding to conditionally independent pairs ('1','3') and ('1','4') variables '2'.

basis of the sample correlations, or partial correlations, of the nearest neighbors of each connected vertex in $\mathcal{G}_{\rho^*}(\mathbf{\Phi})$. The waterfall plot can then be used to identify those vertices having p-values that fall below a user-specified level.

5.2. **Numerical simulation study.** We first considered screening for hubs in a simple synthetic Gaussian graphical model. We drew $n = 40$ samples of a 1000-dimensional zero mean Gaussian random vector to form the 40×1000 data matrix $\mathbb{X}$. Only 4 of the 1000 variables have any dependency; the rest are i.i.d. and are independent of the first 4. Specifically, the first 4 variables have partial correlation matrix equal to

$$\Omega = \begin{bmatrix} 1 & 0.41 & 0 & 0 \\ 0.41 & 1 & -0.52 & -0.82 \\ 0 & -0.52 & 1 & 0.71 \\ 0 & -0.82 & 0.71 & 1 \end{bmatrix}.$$

This partial correlation matrix is represented by the simple graphical model shown in Fig 2. As $\mathbf{\Sigma}$ is block sparse with sparsity factor $k/p = 1/250$, we computed the critical phase transition threshold $\rho_{c,1}$ using (10) with parameter $c_{n,\delta} = \mathrm{a}_n\delta$. For $n = 40$ and $p = 1000$ the critical threshold was found to be equal to 0.593. This value was used as the initial threshold ρ^* resulting in an average $E[N_{1,\rho^*}] = 55$ of the 1000 variables passing the correlation threshold and forming the initial sample partial correlation graph $\mathcal{G}_{\rho^*}(P)$.

The result of one of the simulation runs provides an illustration of the waterfall plot, Fig. 3. For this simulation $N_{1,\rho^*} = 64$ and the p-values of each of these variables were evaluated and plotted as described in Sec. 3.3. Using the threshold $\rho = 0.59$, no vertices in $\mathcal{G}_{\rho^*}(P)$

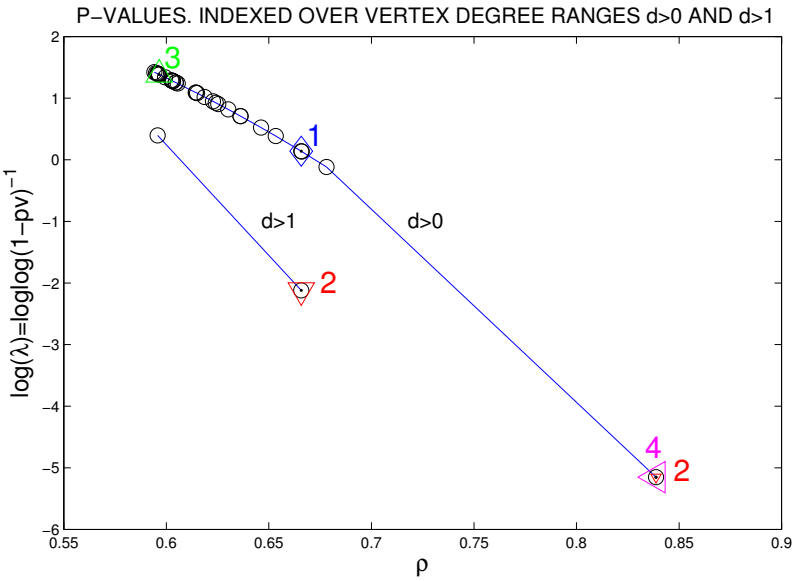

FIGURE 3. Waterfall plot of approximate p-values (plotted in terms of the log of the Poisson rate λ) for partial correlation hub screening. Here there are $p = 1000$ variables in a Gaussian graphical model displayed in Fig. 2 and there are $n = 40$ samples. For these values of n, p the expression (10) gave 0.593 as the critical phase transition threshold $\rho_{c,1}$ governing the mean number of partial correlation discoveries using the proposed procedure. Using this partial correlation threshold value, 64 variables survive the screen and the p-values of these variables are shown as circles on the figure. These variables fall into two groups denoted by the two curves. The "ρ" position of a circle on the top curve is the minimal threshold for which the associated variable would be declared as connected (vertex degree $d > 0$) in the sample partial correlation graph. The $\log(\lambda)$ position of this circle indicates the associated p-value via the formula pv$= 1 - e^{-e^{\log \lambda}}$. The bottom curve is similar except that it designates the connected variables whose vertex degrees are $d > 1$. In particular, the figure shows that only vertices '2' and '4' would pass the hub screen at a false positive level less than $0.13 = 1 - e^{-e^{-2}}$ and only vertex '2' would pass the screen at this level as a hub of degree $d > 1$. No false positives of any degree would occur at this level.

were found to have degree exceeding two and the only only vertex found of degree greater than 1 was vertex '2,' the vertex of degree three in Fig. 2. For ease of visualization, in Fig. 3 we plot the log rate parameters $\log \lambda$ instead of the p-values themselves using the relation: pv$= 1 - e^{-e^{\log \lambda}}$. The exhibited waterfall plot has grouped the variables, denoted as circles, into two groups of vertices in $\mathcal{G}_{\rho^*}(P)$: a group of 2 vertices having degree $d > 1$ and a larger group of connected vertices (having degree $d > 0$). The reader will note that our screening procedure would correctly identify vertices '1', '2' and '4' as being connected and vertex '2' as having degree at least 2, at false positive level less than 0.13, which corresponds to $\log \lambda < -2$.

observed degree	# predicted ($E[N_{\delta,\rho^*}]$)	# actual (N_{δ,ρ^*})
$d_i > \delta = 0$	8531	8354
$d_i > \delta = 1$	1697	1631
$d_i > \delta = 2$	234	240
$d_i > \delta = 3$	24	24
$d_i > \delta = 4$	2	1

TABLE 1. Illustration of agreement between the predicted (mean) number of false positives and the observed number of false positives in a single simulation run of the sham NKI dataset experiment shown in Fig. 4. Only actual observed degrees 1 through 5 appear in the table. No vertices of higher degree than 5 were discovered.

Furthermore none of the 996 non-dependent vertices would be discovered at any reasonable false positive level (the lowest circle on the waterfall plot is at a level $\log \lambda > 0$ corresponding to the extremely weak p-value of 0.63). The relatively low statistical significance of the true discoveries '1','2','3','4' is due to the small value of n chosen for this illustration. The level of significance improves substantially if one doubles the number n of samples (results not shown here).

To illustrate the scalability of our proposed method we created a sham dataset of high dimension. Figure 4 shows the waterfall plot of partial correlation hub p-values for a sham measurement matrix with i.i.d. normal entries that emulates the NKI experimental data presented in the next subsection. There are $n = 266$ samples and $p = 24,481$ variables in this sham. For these values of n, p the critical phase transition threshold on discoveries with positive vertex degree was determined to be $\rho_{c,1} = 0.296$. For purposes of illustration of the fidelity of our theoretical predictions we used an initial screening threshold equal to $\rho^* = 0.26$. As this is a sham, all discoveries are false positives.

The waterfall plots of p-values (15) are shown in Fig. 4. The way we have parameterized the p-value curves on the left means that the leftmost point of each curve in the left waterfall plot should occur at approximately $(\rho^*, E[N_{\delta,\rho^*}])$, as can be verified by comparing the second and third columns of Table 1. This table demonstrates good agreement between the predicted (mean) number of partial correlation hub discoveries and the actual number of discoveries for a single realization of the data matrix. The realization shown in the table is representative of the several simulations runs observed on this sham dataset.

5.3. **Parcor screening of NKI dataset.** The Netherlands Cancer Institute (NKI) dataset [4] contains data from Affymetrix GeneChips collected from 295 subjects who were diagnosed with early stage breast cancer. The dataset was collected and analyzed by van de Vijver *et al* [21] to discover gene expression signatures that can differentiate between good and poor prognosis. Several other groups have used the NKI dataset to illustrate various types of statistical algorithms. Notably, Peng *et al* [15] used the dataset to illustrate their graphical lasso method for covariance selection that they subsequently used to identify hubs in the partial correlation graph. Here we illustrate the proposed partial correlation hub screening method on this NKI dataset and compare to results of Peng *et al*.

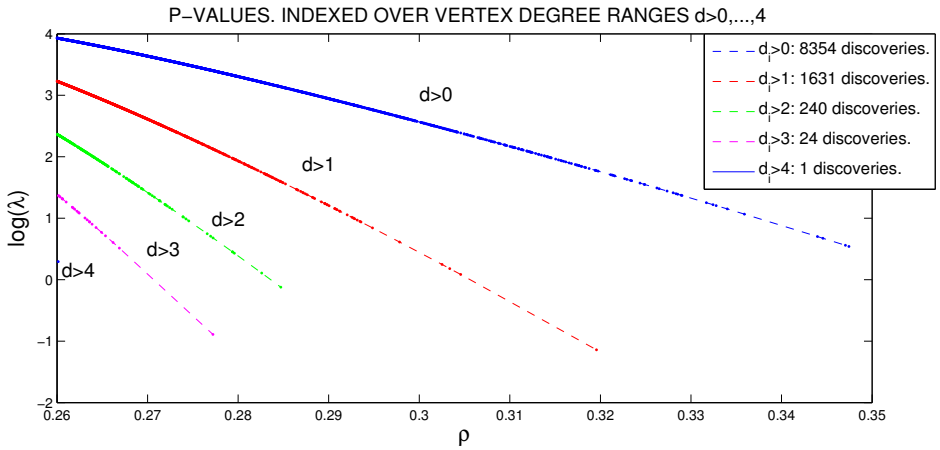

FIGURE 4. Waterfall plot of p-values for partial correlation hub screening of a sham version of the NKI dataset [4] plotted in terms of log λ as in Fig. 2. The data matrix $\mathbb{X}$ has $n = 266$ rows and $p = 24,481$ columns and is populated with i.i.d. zero mean and unit variance Gaussian variables ($\mathbf{\Sigma} = \mathbf{I}$). The predicted mean number of discoveries of any degree is very close to the actual number of discoveries of that degree (see Table I). The level of significance of any of the discoveries indicated on the waterfall plot for this sham dataset is, as expected, very low (p-value greater than 0.3).

As in Peng *et al* [15], we only used a subset of the available GeneChip samples. Specifically, since 29 of the 295 GeneChips had variables with missing values, only 266 of the them were used in our analysis. Each GeneChip sample in the NKI dataset contains the expression levels of 24,481 genes. Peng *et al* [15] applied univariate Cox regression to reduce the number of variables to 1,217 genes prior to applying their sparse partial correlation estimation (`space`) method. In contrast, we applied our partial correlation hub screening procedure directly to all 24,481 variables.

An initial threshold $\rho^* = 0.35 > \rho_{c,1} = 0.296$ was selected. Figure 5 illustrates the waterfall plot of p-values of all discovered variables. Note in particular the very high level of significance of certain variables at the lower extremities of the p-value curves. According to NCBI Entrez several of the most statistically significant discovered genes on these strands have been related to breast cancer, lymphoma, and immune response. The p-value trajectories (colored labels) across different values of δ of three of these genes are illustrated in the figure (ARRB2 (*Arrestin, Beta 2*), CTAG2 (*Cancer/testis antigen*) and IL14 (*Interleukin*)). Note that some genes are highly statistically significant only at low vertex degree (CTAG2) or at high vertex degree (IL14), while others retain high statistical significance across all vertex degrees (ARRB2). Figure 6 is the same plot with the trajectories of the 6 unambiguously annotated hub genes given in Table 4 of Peng *et al* [15]. While these 6 genes do not have nearly as high p-values, or as high partial correlation, as compared to other genes shown in Fig. 5 their predicted p-values are still very small.

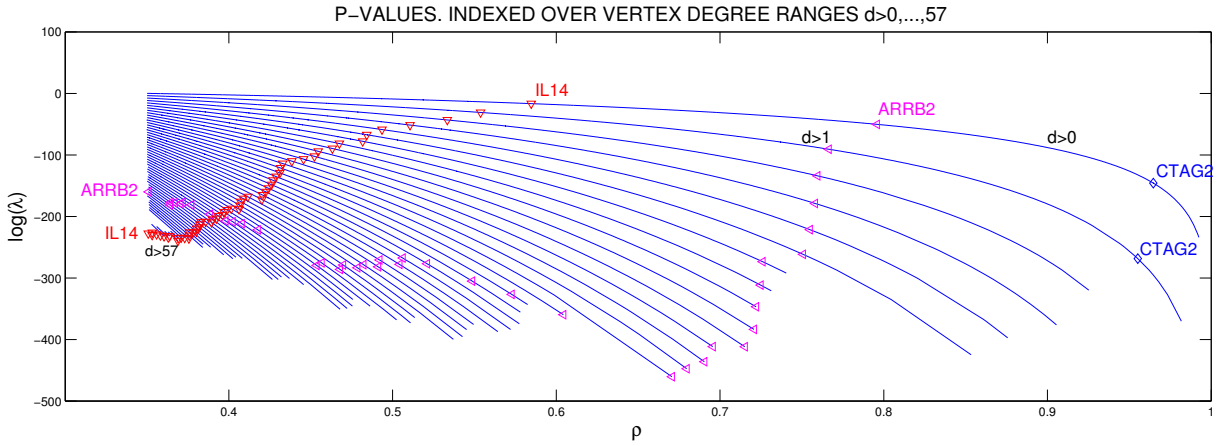

FIGURE 5. Waterfall plot of partial correlation p-values for NKI gene expression dataset of [4] plotted in terms log Poisson rate. Each curve indexes the p-values for a particular degree threshold δ and a gene is on the curve if its degree d_i in the initial graph is greater than or equal to δ. The discovered vertex degree ranges from 1 to 58 (last dot labeled IL14 at bottom left). The p-value trajectories of several genes of interest are indicated over the range of vertex degree δ in the partial correlation graph. The cancer-related CTAG2 gene does not appear on any curve except for $d > 0$ and $d > 1$, but has low p-value on these two curves - it has high statistical significance as a connected gene of low degree (< 3) but not as a hub gene of high degree. In contrast, the gene ARBB2 appears with low p-values on most of the curves suggesting that it is a significant hub genes of high vertex degree. Interestingly, IL14 does not have high significance as a connected vertex of low degree but has high significance as a hub of high degree.

6. CONCLUSIONS

We treated the problem of screening for variables that are strongly correlated hubs in a correlation or partial correlation graph when $n \ll p$ and p is large. The proposed hub screening procedure thresholds the sample correlation or the pseudo-inverse of the sample correlation matrix using Z-score representations of the correlation and partial correlation matrices. For large p and finite n asymptotic limits that specify the probability of false hub discoveries were established. These limits were used to obtain expressions for phase transition thresholds and p-values under the assumption of a block-sparse covariance matrix. To illustrate the applicability and computational scalability of our hub screening algorithm we applied it to the NKI breast cancer gene expression dataset. With different levels of false positive control, the proposed algorithm recapitulated genes previously reported in a graphical lasso study [15] of the same dataset while discovering cancer-related genes not reported in [15] having higher statistical significance.

The screening algorithms introduced in this paper apply under the hypothesis that the $(\delta+1)$-order dependency function J defined in (12) is close to one. Proposition 3 establishes

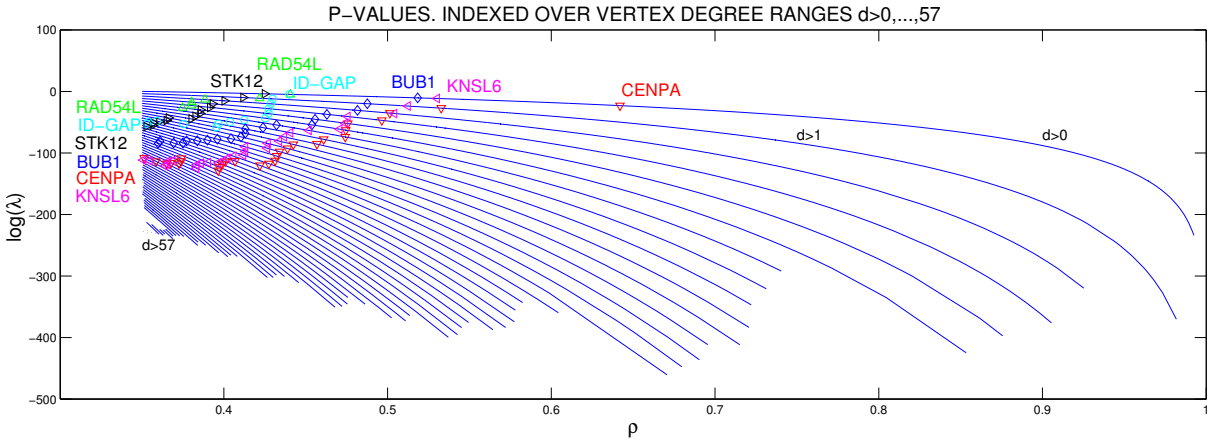

FIGURE 6. Same as Fig. 5 but showing the p-value trajectories of 6 of the hub genes ('BUB1' 'CENPA' 'KNSL6' 'STK12' 'RAD54L' 'ID-GAP') reported in Peng *et al* [15]. These are the genes reported in Table 4 of [15] that have unambiguous annotation on the GeneChip array.

this hypothesis holds when p is large and the dispersion matrix Σ is sparse. In [20] an entropy-estimation approach was proposed for empirically testing this hypothesis, but it was restricted to the case $\delta = 1$. A general theory of empirical estimation and testing of J remains to be developed.

Acknowledgements

Alfred Hero was supported in part by DIGITEO and National Science Foundation grant CCF 0830490. Bala Rajaratnam was supported in part by NSF grants DMS-05-05303, DMS-09-06392 and grant SUFSC08-SUSHSTF09-SMSCVISG0906.

References

[1] A. Anandkumar, L. Tong, and A. Swami. Detection of Gauss–Markov random fields with nearest-neighbor dependency. *Information Theory, IEEE Transactions on*, 55(2):816–827, 2009.

[2] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5(4):403–424, 1990.

[3] P.J. Bickel and E. Levina. Covariance regularization via thresholding. *Annals of Statistics*, 34(6):2577–2604, 2008.

[4] H.Y. Chang, D.S.A. Nuyten, J.B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie, H. Dai, Y.D. He, L.J. Van't Veer, H. Bartelink, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences*, 102(10):3738, 2005.

[5] D.R. Cox and N. Wermuth. *Multivariate dependencies: models, analysis and interpretation*. Chapman & Hall/CRC, 1996.

[6] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

[8] R. Gill, S. Datta, and S. Datta. A statistical framework for differential network analysis from microarray data. *BMC bioinformatics*, 11(1):95, 2010.

[9] M. Goldstein and A. F. M. Smith. Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):pp. 284–291, 1974.

[10] G. H. Golub and C. F. Van Loan. *Matrix Computations (2nd Edition)*. The Johns Hopkins University Press, Baltimore, 1989.

[11] A.O. Hero and B. Rajaratnam. Large scale correlation screening. *Journ. of American Statistical Association*, 106(496):1540–1552, Dec 2011. Available as Arxiv preprint arXiv:1102.1204.

[12] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, January 2011.

[13] N. Krämer, J. Schäfer, and A.-L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(384):1–24, 2009.

[14] Y. Liu, V. Chandrasekaran, A. Anandkumar, and A.S. Willsky. Feedback message passing for inference in gaussian graphical models. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1683–1687. IEEE, 2010.

[15] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[16] V. Pihur, S. Datta, and S. Datta. Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, 24(4):561, 2008.

[17] B. Rajaratnam, H. Massam, and C.M. Carvalho. Flexible covariance estimation in graphical gaussian models. *Annals of Statistics*, 36:2818–2849, 2008.

[18] A. J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[19] J. Schaefer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4(32), 2005.

[20] K. Sricharan, A.O. Hero and B. Rajaratnam, A local dependence measure and its application to screening for high correlations in large data sets. *Proc. 14th Intl. Conf. on Inform. Fusion*, July, 2011.

[21] M.J. van de Vijver, Y.D. He, L.J. van 't Veer, H. Dai, A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[22] A. Wiesel, Y.C. Eldar, and A.O. Hero. Covariance estimation in decomposable Gaussian graphical models. *Signal Processing, IEEE Transactions on*, 58(3):1482–1492, 2010.

7. Appendix

This appendix contains two subsections. Section 7.1 gives the necessary definitions. Section 7.2 gives proofs of the theory given in Sec. 4 .

7.1. Notation, Preliminaries and Definitions.

- $\mathbb{X}$: $n \times p$ matrix of observations.
- $\mathbb{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$: $(n-1) \times p$ matrix of correlation (or partial correlation) Z-scores $\{\mathbf{Z}_i\}_i$ associated with $\mathbb{X}$.
- $\mathbb{Z}^T\mathbb{Z}$: $p \times p$ sample correlation matrix $\mathbf{R}$ (if $\mathbb{Z} = \mathbb{U}$) or sample partial correlation matrix $\mathbf{P}$ (if $\mathbb{Z} = \mathbb{Y}$) associated with $\mathbb{X}$.
- $\mathbf{Z}_i^T\mathbf{Z}_j$: sample correlation (or partial correlation) coefficient, the i,j-th element of $\mathbb{Z}^T\mathbb{Z}$.
- $\rho \in [0,1]$: screening threshold applied to matrix $\mathbb{Z}^T\mathbb{Z}$.
- $r = \sqrt{2(1-\rho)}$: spherical cap radius parameter.
- S_{n-2}: unit sphere in $\mathbb{R}^{n-1}$.
- $a_n = |S_{n-2}|$: surface area of S_{n-2}.
- $\mathcal{G}_0(\mathbf{\Phi})$: graph associated with population correlation matrix $\mathbf{\Phi} = \mathbf{\Gamma}$ or partial correlation matrix $\mathbf{\Phi} = \mathbf{\Omega}$. An edge in $\mathcal{G}_0(\mathbf{\Phi})$ corresponds to a non-zero entry of $\mathbf{\Phi}$.
- $\mathcal{G}_\rho = \mathcal{G}_\rho(\mathbf{\Phi})$: graph associated with thresholded sample correlation matrix $\mathbf{\Phi} = \mathbf{R}$ or partial correlation matrix $\mathbf{\Phi} = \mathbf{P}$. Specifically, the edges of $\mathcal{G}_\rho(\mathbf{\Phi})$ are specified by the non-diagonal entries of $\mathbb{Z}^T\mathbb{Z}$ whose magnitudes exceed level ρ.
- d_i: observed degree of vertex i in $\mathcal{G}_\rho(\mathbf{\Phi})$, $\mathbf{\Phi} \in \{\mathbf{R}, \mathbf{P}\}$.
- δ: screening threshold for vertex degrees in $\mathcal{G}_\rho(\mathbf{\Phi})$, $\mathbf{\Phi} \in \{\mathbf{R}, \mathbf{P}\}$.
- k: upper bound on vertex degrees of $\mathcal{G}_0(\mathbf{\Phi})$, $\mathbf{\Phi} \in \{\mathbf{\Gamma}, \mathbf{\Omega}\}$.
- $N_{\delta,\rho}$: generic notation for the number of correlation hub discoveries ($N_{\delta,\rho}(\mathbf{R})$) or partial correlation hub discoveries ($N_{\delta,\rho}(\mathbf{P})$) of degree $d_i \geq \delta$ in $\mathcal{G}_\rho(\mathbf{R})$, or $\mathcal{G}_\rho(\mathbf{P})$, respectively.
- $\tilde{N}_{\delta,\rho}$ counts the number of subsets of δ mutually coincident edges in $\mathcal{G}_\rho$.
- $A(r, \mathbf{z})$: the union of two anti-polar spherical cap regions in S_{n-2} of radii $r = \sqrt{2(1-\rho)}$ centered at points $-\mathbf{z}$ and $\mathbf{z}$.
- P_0: probability that a uniformly distributed vector $\mathbf{U} \in S_{n-2}$ falls in $A(r, \mathbf{z})$

$$
\begin{aligned}
P_0 = \ P_0(\rho, n) &= \mathrm{a}_n \int_\rho^1 \left(1-u^2\right)^{\frac{n-4}{2}} du \\
&= (n-2)^{-1}\mathrm{a}_n(1-\rho^2)^{(n-2)/2}(1+O(1-\rho^2)),
\end{aligned} \tag{26}
$$

where $\mathrm{a}_n = 2B((n-2)/2, 1/2)$ and $B(l,m)$ is the beta function.

For given integer k, $0 \leq k < p$, and Φ either the population correlation matrix $\mathbf{\Gamma}$ or the population partial correlation matrix $\mathbf{\Omega}$ define

$$
\mathcal{N}_k(i) = \mathrm{argmax}_{j_1 \neq \cdots \neq j_{\min(k,\bar{d}_i)}} \sum_{l=1}^{\min(k,\bar{d}_i)} |\Phi_{ij_l}|, \tag{27}
$$

where $\overline{d}_i$ denotes the degree of vertex i in $\mathcal{G}_0(\mathbf{\Phi})$ and the maximization is over the range of distinct $j_l \in \{1,\ldots,p\}$ that are not equal to i. When $k \geq \overline{d}_i$ these are the indices of the $\overline{d}_i$ neighbors of vertex i in $\mathcal{G}_0(\mathbf{\Phi})$. When $k < \overline{d}_i$ these are the subset of the k-nearest neighbors (k-NN) of vertex i. For the sequel it will be convenient to define the following vector valued indexing variable: $\vec{i} = (i_0,\ldots,i_\delta)$, where $0 < \delta \leq p$ and $i_0,\ldots,i_\delta$ are distinct integers in $\{1,\ldots,p\}$. With this index denote by $\mathbf{Z}_{\vec{i}}$ the set of $\delta+1$ Z-scores $\{\mathbf{Z}_{i_j}\}_{j=0}^{\delta}$.

Define the set of complementary k-NN's of $\mathbf{Z}_{\vec{i}}$ as $\mathbf{Z}_{A_k(\vec{i})} = \{\mathbf{Z}_l : l \in A_k(\vec{i})\}$, where

$$A_k(\vec{i}) = \left(\cup_{l=0}^{\delta} \mathcal{N}_k(i_l)\right)^c - \{\vec{i}\}, \tag{28}$$

with A^c denoting set complement of set A. The complementary k-NN's include vertices outside of the k-nearest-neighbor regions of the set of points $\mathbf{Z}_{\vec{i}}$.

Define the δ-fold leave-one-out average of the density, a function of i, $f_{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_\delta},\mathbf{Z}_i}$:

$$\begin{aligned}\overline{f_{\mathbf{Z}_{*_1-i},\ldots,\mathbf{Z}_{*_\delta-i},\mathbf{Z}_i}}(\mathbf{z}_1,\ldots,\mathbf{z}_\delta,\mathbf{z}_i) \\ = 2^{-d} \sum_{s_1,\ldots,s_\delta\in\{-1,1\}} \binom{p-1}{\delta}^{-1} \sum_{i_1\neq\cdots\neq i_\delta\neq i}^{p} f_{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_\delta},\mathbf{Z}_i}(s_1\mathbf{z}_1,\ldots,s_\delta\mathbf{z}_\delta,\mathbf{z}_i),\end{aligned} \tag{29}$$

where in the inner summation, indices $i_1,\ldots,i_\delta$ range over $\{1,\ldots,p\}$. Also define the $(\delta+1)$-fold average of the same density

$$\begin{aligned}\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}(\mathbf{z}_1,\ldots,\mathbf{z}_\delta,\mathbf{z}_i) \\ = p^{-1}\sum_{i=1}^{p}\left(\tfrac{1}{2}\overline{f_{\mathbf{Z}_{*_1-i},\ldots,\mathbf{Z}_{*_\delta-i},\mathbf{Z}_i}}(\mathbf{z}_1,\ldots,\mathbf{z}_\delta,\mathbf{z}_i) + \tfrac{1}{2}\overline{f_{\mathbf{Z}_{*_1-i},\ldots,\mathbf{Z}_{*_\delta-i},\mathbf{Z}_i}}(\mathbf{z}_1,\ldots,\mathbf{z}_\delta,-\mathbf{z}_i)\right).\end{aligned} \tag{30}$$

For any data matrix $\mathbb{Z}$ define the dependency coefficient between the columns $\mathbf{Z}_{\vec{i}}$ and their complementary k-NN's

$$\Delta_{p,n,k,\delta}(\vec{i}) = \left\| (f_{\mathbf{Z}_{\vec{i}}|\mathbf{Z}_{A_k(\vec{i})}} - f_{\mathbf{Z}_{\vec{i}}})/f_{\mathbf{Z}_{\vec{i}}} \right\|_\infty, \tag{31}$$

and the average of these coefficients is

$$\|\Delta_{p,n,k,\delta}\|_1 = \left(p\binom{p-1}{\delta}\right)^{-1} \sum_{i_0=1}^{p} \sum_{i_1<\ldots<i_\delta} \Delta_{p,n,k,\delta}(\vec{i}). \tag{32}$$

where the second sum is indexed over $i_1,\ldots,i_\delta \neq i_0$.

The coefficient (32) quantifies weak dependence of the Z-scores. If, for all i, $\mathbf{Z}_{\vec{i}}$ and its complementary k-NN neighborhood variables are independent then $\|\Delta_{p,n,k,\delta}\|_1 = 0$. When the rows of $\mathbb{X}$ are i.i.d. and elliptically distributed, and $\mathbb{Z} = \mathbb{U}$ are the standard correlation Z-scores, then a sufficient condition for $\|\Delta_{p,n,k,\delta}\|_1 = 0$ is that $\mathcal{G}_0(\mathbf{\Phi})$ have no vertex of degree greater than k or, equivalently, that the dispersion matrix $\mathbf{\Sigma}$ be row sparse of degree k.

Finally, for arbitrary joint density $f_{\mathbf{Z}_1,\ldots,\mathbf{Z}_\delta}(\mathbf{z}_1,\ldots,\mathbf{z}_\delta)$ on $S_{n-2}^{\delta} = \times_{i=1}^{\delta} S_{n-2}$, define

$$J(f_{\mathbf{Z}_1,\ldots,\mathbf{Z}_\delta}) = |S_{n-2}|^{\delta-1} \int_{S_{n-2}} f_{\mathbf{Z}_1,\ldots,\mathbf{Z}_\delta}(\mathbf{z},\ldots,\mathbf{z}) d\mathbf{z}. \tag{33}$$

7.2. **Proofs of theorems.** *Proof of Prop. 1:*

The proof of (19) uses similar arguments to those used to establish Lemma 1 and Prop. 1 in [11].

With $\phi_i = I(d_i \geq \delta)$ we have $N_{\delta,\rho} = \sum_{i=1}^p \phi_i$. Define $\phi_{ij} = I\left(\mathbf{Z}_j \in A(r, \mathbf{Z}_i)\right)$ the indicator of the presence of an edge in $\mathcal{G}_\rho(\mathbf{\Phi})$ between vertices i and j, where $A(r, \mathbf{Z}_i)$ is the union of two antipolar caps in S_{n-2} of radius $r = \sqrt{2(1-\rho)}$ centered at $\mathbf{Z}_i$ and $-\mathbf{Z}_i$, respectively. Then ϕ_i and ϕ_{ij} have the explicit relation

$$\phi_i = \sum_{l=\delta}^{p-1} \sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,l)} \prod_{j=1}^{l} \phi_{ik_j} \prod_{m=l+1}^{p-1} (1 - \phi_{ik_m}) \tag{34}$$

where we have defined the index vector $\vec{k} = (k_1, \ldots, k_{p-1})$ and the set

$$\check{\mathcal{C}}_i(p-1,l) = \{\vec{k} : k_1 < \ldots < k_l, k_{l+1} < \ldots < k_{p-1}\ k_j \in \{1, \ldots, p\} - \{i\}, k_j \neq k_{j'}\}.$$

The inner summation in (34) simply sums over the set of distinct indices not equal to i that index all $\binom{p-1}{l}$ different types of products $\prod_{j=1}^{l} \phi_{ik_j} \prod_{m=l+1}^{p-1}(1 - \phi_{ik_m})$. Subtracting $\sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j}$ from both sides of (34)

$$\phi_i - \sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta} \phi_{ik_j} \tag{35}$$

$$= \sum_{l=\delta+1}^{p-1} \sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,l)} \prod_{j=1}^{l} \phi_{ik_j} \prod_{m=l+1}^{p-1} (1 - \phi_{ik_m}) \tag{36}$$

$$+ \sum_{\vec{k} \in \check{\mathcal{C}}_i(p-1,l)} \sum_{m=l+1}^{p-1} (-1)^{m-l} \sum_{k_{l+1} < \ldots < k_m} \prod_{j=1}^{l} \phi_{ik_j} \prod_{n=l+1}^{m} \phi_{ik_n} \tag{37}$$

where, in the last line we have used the expansion

$$\prod_{m=l+1}^{p-1} (1 - \phi_{ik_m}) = 1 + \sum_{m=l+1}^{p-1} (-1)^{m-l} \sum_{k_{l+1} < \ldots < k_m} \prod_{n=l+1}^{m} \phi_{ik_n}.$$

The following simple asymptotic representation will be useful in the sequel. For any $i_1, \ldots, i_k \in \{1, \ldots, p\}$, $i_1 \neq \cdots \neq i_k \neq i$, $k \in \{1, \ldots, p-1\}$,

$$E\left[\prod_{j=1}^{k} \phi_{ii_j}\right] = \int_{S_{n-2}} d\mathbf{v} \int_{A(r,\mathbf{v})} d\mathbf{u}_1 \cdots \int_{A(r,\mathbf{v})} d\mathbf{u}_k\ f_{\mathbf{U}_{i_1},\ldots,\mathbf{U}_{i_k},\mathbf{U}_i}(\mathbf{u}_1, \cdots, \mathbf{u}_k, \mathbf{v}) \tag{38}$$

$$\leq P_0^k a_n^k M_{k|1} \tag{39}$$

with $P_0 = P_0(\rho, n)$ defined in (26), $a_n = |S_{n-2}|$, and

$$M_{k|1} = \max_{i_1 \neq \cdots \neq i_{k+1}} \left\| f_{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_k}|\mathbf{Z}_{i_{k+1}}} \right\|_\infty, \tag{40}$$

The following simple generalization of (39) to arbitrary product indices ϕ_{ij} will also be needed

$$E\left[\prod_{l=1}^{m}\phi_{i_l j_l}\right] \leq P_0^m a_n^m M_{|Q|}, \tag{41}$$

where Q =unique($\{i_l, j_l\}_{l=1}^m$) is the set of unique indices among the distinct pairs $\{(i_l, j_l)\}_{l=1}^m$ and $M_{|Q|}$ is a bound on the joint density of $\mathbf{Z}_Q$.

Define the random variable

$$\theta_i = \binom{p-1}{\delta}^{-1} \sum_{\vec{k}\in\check{\mathcal{C}}_i(p-1,\delta)} \prod_{j=1}^{\delta}\phi_{ik_j}. \tag{42}$$

We show below that for sufficiently large p

$$\left|E[\phi_i] - \binom{p-1}{\delta}E[\theta_i]\right| \;\leq\; \gamma_{p,\delta}((p-1)P_0)^{\delta+1}, \tag{43}$$

where $\gamma_{p,\delta} = \max_{\delta+1\leq l<p}\{a_n^l M_{l|1}\}\left(e - \sum_{l=0}^{\delta}\frac{1}{l!}\right)(1+(\delta!)^{-1})$ and $M_{l|1}$ is a least upper bound on any l-dimensional joint density of the variables $\{\mathbf{Z}_i\}_{j\neq i}^p$ conditioned on $\mathbf{Z}_i$.

To show inequality (43) take expectations of (37) and apply the bound (39) to obtain

$$\begin{aligned}
&\left|E[\phi_i] - \binom{p-1}{\delta}E[\theta_i]\right| \\
&\quad\leq \left|\sum_{l=\delta+1}^{p-1}\binom{p-1}{l}P_0^l a^l M_{l|1} + \binom{p-1}{\delta}\sum_{l=1}^{p-1-\delta}\binom{p-1-\delta}{l}P_0^{\delta+l}a_n^{\delta+l}M_{\delta+l|1}\right| \\
&\quad\leq A(1+(\delta!)^{-1}), \qquad (44)
\end{aligned}$$

where

$$A = \sum_{l=\delta+1}^{p-1}\binom{p-1}{l}P_0^l a^l M_{l|1}.$$

The line (44) follows from the identity $\binom{p-1-\delta}{l}\binom{p-1}{\delta} = \binom{p-1}{l+\delta}(\delta!)^{-1}$ and a change of index in the second summation on the previous line. Since $(p-1)P_0 < 1$

$$\begin{aligned}
|A| \;&\leq\; \max_{\delta+1\leq l<p}\{a_n^l M_{l|1}\}\sum_{l=\delta+1}^{p-1}\binom{p-1}{l}((p-1)P_0)^l \\
&\leq\; \max_{\delta+1\leq l<p}\{a_n^l M_{l|1}\}\left(e - \sum_{l=0}^{\delta}\frac{1}{l!}\right)((p-1)P_0)^{\delta+1}.
\end{aligned}$$

Application of the mean-value-theorem to the integral representation (38) yields

$$\left|E[\theta_i] - P_0^{\delta}J(\overline{f_{\mathbf{Z}_{*_1-i},\ldots,\mathbf{Z}_{*_\delta-i},\mathbf{Z}_i}})\right| \;\leq\; \tilde{\gamma}_{p,\delta}((p-1)P_0)^{\delta}r, \tag{45}$$

where $\tilde{\gamma}_{p,\delta} = 2a_n^{\delta+1}\dot{M}_{\delta+1|1}/\delta!$ and $\dot{M}_{\delta+1|1}$ is a bound on the norm of the gradient

$$\nabla_{\mathbf{z}_{i_1},\ldots,\mathbf{z}_{i_\delta}}\overline{f_{\mathbf{Z}_{*_1-i},\ldots,\mathbf{Z}_{*_\delta-i}|\mathbf{Z}_i}(\mathbf{z}_{i_1},\ldots,\mathbf{z}_{i_\delta}|\mathbf{z}_i)}.$$

Combining (43)-(45) and the relation $r = O((1-\rho)^{1/2})$,

$$\left| E[\phi_i] - \binom{p-1}{\delta} P_0^\delta J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}) \right| \leq O\left(((p-1)P_0)^\delta \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right). \quad (46)$$

Summing over i and recalling the definitions (17) and (18) of $\xi_{p,n,\delta,\rho}$ and $\eta_{p,\delta}$,

$$\begin{aligned} \left| E[N_{\delta,\rho}] - \xi_{p,n,\delta,\rho} J(\overline{f_{\mathbf{Z}_{*_1},\ldots,\mathbf{Z}_{*_{\delta+1}}}}) \right| &\leq O\left(p((p-1)P_0)^\delta \max\left\{(p-1)P_0, (1-\rho)^{1/2}\right\}\right) \\ &= O\left(\eta_{p,\delta}^\delta \max\left\{\eta_{p,\delta} p^{-1/\delta}, (1-\rho)^{1/2}\right\}\right). \end{aligned} \quad (47)$$

This establishes the bound (19).

For the bound (20) we use the Chen-Stein method [2]. The part of the bound (20) that holds for $\delta = 1$ was derived in the course of proof of Prop. 1 in [11]. The key idea used to generalize to the case $\delta > 1$ is to come up with a related counting random variable that is asymptotically Poisson for which the probability of zero counts is asymptotically identical to $P(N_{\delta,\rho} = 0)$. This related random variable is the number of subgraphs in $\mathcal{G}_\rho$ that are isomorphic to a star graph having δ edges. We denote this random variable as $\tilde{N}_{\delta,\rho}$ and it has the representation:

$$\tilde{N}_{\delta,\rho} = \sum_{i_0=1}^{p} \sum_{i_1<\ldots<i_\delta} \prod_{j=1}^{\delta} \phi_{i_0 i_j}, \quad (48)$$

where the second sum is indexed over $i_1, \ldots, i_\delta \neq i_0$. For $\vec{i} \stackrel{\text{def}}{=} (i_0, i_1, \ldots, i_\delta)$ define the index set $B_{\vec{i}} = B_{i_0,i_1,\ldots,i_\delta} = \{(j_0, j_1, \ldots, j_\delta) : j_l \in \mathcal{N}_k(i_l) \cup \{i_l\}, l = 0, \ldots, \delta\} \cap \mathcal{C}^<$ where $\mathcal{C}^< = \{(j_0, \ldots, j_\delta) : j_0 \in \{1, \ldots, p\},\ 1 \leq j_1 < \cdots < j_\delta \leq p, j_1, \ldots, j_\delta \neq j_0\}$. These index the distinct sets of points $\mathbf{Z}_{\vec{i}} = \{\mathbf{Z}_{i_0}, \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_\delta}\}$ and their respective k-NN's. Note that $|B_{\vec{i}}| \leq k^{\delta+1}$. Identifying $\tilde{N}_{\delta,\rho} = \sum_{\vec{i}\in\mathcal{C}^<} \prod_{l=1}^{\delta} \phi_{i_0 i_l}$ and $N^*_{\delta,\rho}$ a Poisson distributed random variable with rate $E[\tilde{N}_{\delta,\rho}]$, the Chen-Stein bound [2, Thm. 1] is

$$2 \max_A |P(\tilde{N}_{\delta,\rho} \in A) - P(N^*_{\delta,\rho} \in A)| \leq b_1 + b_2 + b_3, \quad (49)$$

where

$$b_1 = \sum_{\vec{i}\in\mathcal{C}^<} \sum_{\vec{j}\in B_{\vec{i}}} E\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l}\right] E\left[\prod_{m=1}^{\delta} \phi_{j_0 j_m}\right],$$

$$b_2 = \sum_{\vec{i}\in\mathcal{C}^<} \sum_{\vec{j}\in B_{\vec{i}}-\{\vec{i}\}} E\left[\prod_{l=1}^{\delta} \phi_{i_0 i_l} \prod_{m=1}^{\delta} \phi_{j_0 j_m}\right],$$

and, for $p_{\vec{i}} = E[\prod_{l=1}^{\delta} \phi_{i_0 i_l}]$,

$$b_3 = \sum_{\vec{i}\in\mathcal{C}^<} E\left[E\left[\left.\prod_{l=1}^{\delta} \phi_{i_0 i_l} - p_{\vec{i}}\right| \phi_{\vec{j}} : \vec{j} \notin B_{\vec{i}}\right]\right].$$

Over the range of indices in the sum of b_1 $E[\prod_{l=1}^{\delta} \phi_{i i_l}]$ is of order $O(P_0^\delta)$, by (41), and therefore

$$b_1 \leq O\left(p^{\delta+1} k^{\delta+1} P_0^{2\delta}\right) = O\left(\eta_{p,\delta}^{2\delta} (k/p)^{\delta+1}\right),$$

which follows from definition (18).

Similarly to the proof of Prop. 1 in [11] more care is needed to bound b_2 due to the symmetry relation $\phi_{ij} = \phi_{ji}$. If in the summations defining b_2, $i_0 = j_m$ and $j_0 = i_l$ occur for some l, m then there will be a match and $\phi_{i_0 i_l}\phi_{j_0 j_m} = \phi_{i_0 i_l}$. In such case the summand of b_2 will be of lower order than $O(P_0^{2\delta})$. For example, for the case that $l, m = 1$ a match implies $\phi_{i_0 i_1} = \phi_{j_0 j_1}$ and, from (41),

$$E[\prod_{l=1}^{\delta}\phi_{i_0 i_l}\prod_{m=1}^{\delta}\phi_{j_0 j_m}] = E[\prod_{l=1}^{\delta}\phi_{j_1 i_l}\prod_{m=2}^{\delta}\phi_{i_1 j_m}] = O\left(P_0^{2\delta-1}\right).$$

Over $\mathcal{C}^<$ and $B_{\vec{i}} - \{i\}$ there can be no more than a single match in b_2's summand. For a given match there are at most $p^{\delta+1}k^{\delta-1}$ summands of reduced order. We conclude that

$$\begin{aligned} b_2 &\leq O\left(p^{\delta+1}k^{\delta+1}P_0^{2\delta}\right) + O\left(p^{\delta+1}k^{\delta-1}P_0^{2\delta-1}\right) \\ &= O\left(\eta_{p,\delta}^{2\delta}(k/p)^{\delta+1}\right) + O\left(\eta_{p,\delta}^{2\delta-1}(k/p)^{\delta-1}p^{-(\delta-1)/\delta}\right), \end{aligned}$$

which follows from the relation $p^{2\delta}P_0^{2\delta-1} = (p^{\delta+1}P_0^{\delta})^{2-1/\delta}/p^{(\delta-1)/\delta}$.

Next we bound the term b_3 in (49). The set $A_k(\vec{i}) = B_{\vec{i}}^c - \{\vec{i}\}$ indexes the complementary k-NN's of $\mathbf{Z}_{\vec{i}}$ so that, using the representation (41),

$$\begin{aligned} b_3 &= \sum_{\vec{i}\in\mathcal{C}^<} E\left[E\left[\prod_{l=1}^{\delta}\phi_{i_0 i_l} - p_{\vec{i}}\middle|\mathbf{Z}_{A_k(\vec{i})}\right]\right] \\ &= \sum_{\vec{i}\in\mathcal{C}^<}\int_{S_{n-2}^{|A_k(\vec{i})|}} d\mathbf{z}_{A_k(\vec{i})}\left(\prod_{l=1}^{\delta}\int_{S_{n-2}} d\mathbf{z}_{i_0}\int_{A(r,\mathbf{z}_{i_0})} d\mathbf{z}_{i_l}\right)\left(\frac{f_{\mathbf{Z}_{\vec{i}}|\mathbf{Z}_{A_k}}(\mathbf{z}_{\vec{i}}|\mathbf{z}_{A_k(\vec{i})}) - f_{\mathbf{Z}_{\vec{i}}}(\mathbf{z}_{\vec{i}})}{f_{\mathbf{Z}_{\vec{i}}}(\mathbf{z}_{\vec{i}})}\right) f_{\mathbf{Z}_{\vec{i}}}(\mathbf{z}_{\vec{i}}) f_{\mathbf{Z}_{A_k(\vec{i})}}(\mathbf{z}_{A_k(\vec{i})}) \\ &\leq O\left(p^{\delta+1}P_0^{\delta}\|\Delta_{p,n,k,\delta}\|_1\right) = O\left(\eta_{p,\delta}^{\delta}\|\Delta_{p,n,k,\delta}\|_1\right). \end{aligned}$$

Observe that, with $\Lambda = E[N_{\delta,\rho}]$

$$\begin{aligned} |P(N_{\delta,\rho} > 0) - (1 - \exp(-\Lambda))| &\leq \left|P(\tilde{N}_{\delta,\rho} > 0) - P(N_{\delta,\rho} > 0)\right| \\ &\quad + \left|P(\tilde{N}_{\delta,\rho} > 0) - \left(1 - \exp(-E[\tilde{N}_{\delta,\rho}])\right)\right| \\ &\quad + \left|\exp(-E[\tilde{N}_{\delta,\rho}]) - \exp(-\Lambda)\right| \\ &\leq b_1 + b_2 + b_3 + O\left(\left|E[\tilde{N}_{\delta,\rho}] - \Lambda\right|\right). \end{aligned} \tag{50}$$

Combining the above inequalities on b_1, b_2 and b_3 yields the first three terms in the argument of the "max" on the right side of (20).

It remains to bound the term $|E[\tilde{N}_{\delta,\rho}] - \Lambda|$. Application of the mean value theorem to the multiple integral (41) gives

$$\left|E\left[\prod_{l=1}^{\delta}\phi_{i i_l}\right] - P_0^{\delta}J\left(f_{\mathbf{Z}_{i_1},\ldots,\mathbf{Z}_{i_\delta},\mathbf{Z}_i}\right)\right| \leq O\left(P_0^{\delta}r\right). \tag{51}$$

Applying relation (48)

$$\left| E[\tilde{N}_{\delta,\rho}] - p\binom{p-1}{\delta} P_0^\delta J\left(\overline{f_{\mathbf{Z}_{*1},\ldots,\mathbf{Z}_{*\delta+1}}}\right)\right| \;\leq\; O\left(p^{\delta+1}P_0^\delta r\right) = O\left(\eta_{p,\delta}^\delta r\right). \tag{52}$$

Combine this with (50) to obtain the bound (20). This completes the proof of Prop. 1. □

Proof of Cor. 1:

For correlation hub screening ($\mathbb{Z} = \mathbb{U}$) $\|\Delta_{p,n,k,\delta}\|_1 = 0$ so it suffices to consider the other arguments of "max" in the bound (20). As in the proof of Prop 2, $(1-\rho_p)^{1/2} = O\left(p^{-(\delta+1)/((n-2)\delta)}\right)$ and we can merge the last two terms in (20) into the single term $p^{-\alpha(\delta+1)} = \max\left\{p^{-1/\delta}, (1-\rho_p)^{1/2}\right\}$, with α defined in the Corollary statement. Finally, note that $\eta_{p,\delta} = O(1)$ and $(k/p)^{\delta+1} \geq Q_{p,k,\delta} = (k/p)^{\delta-1}p^{-(\delta-1)/\delta}$ when $k/p \geq p^{-(\delta-1)/(2\delta)}$. Therefore, as $\alpha \leq (\delta-1)/(2\delta)$ when $n > 3$, we conclude that if $k/p \geq p^{-\alpha}$ all arguments of "max" in the bound (20) are dominated by $(k/p)^{\delta+1}$.Turning to partial correlation hub screening ($\mathbb{Z} = \mathbb{Y}$), under the block-sparse covariance assumption Prop. 3 asserts that $\|\Delta_{p,n,k,\delta}\|_1 = O(k/p)$ which dominates $(k/p)^{\delta+1}$. This completes the proof of Cor. 1. □

Proof of Proposition 3:

By block-sparsity, the matrix $\mathbb{U}$ of Z-scores can be partitioned as $\mathbb{U} = [\widetilde{\mathbb{U}}, \overline{\mathbb{U}}]$, where $\widetilde{\mathbb{U}} = [\widetilde{\mathbb{U}}_1, \ldots, \widetilde{\mathbb{U}}_q]$ and $\overline{\mathbb{U}} = [\overline{\mathbb{U}}_1, \ldots, \overline{\mathbb{U}}_{p-q}]$ are the dependent and independent columns of $\mathbb{U}$, respectively. Since the columns of $\overline{\mathbb{U}}$'s are i.i.d. and uniform over the unit sphere S_{n-2}, as $p \to \infty$ we have

$$(p-q)^{-1}\overline{\mathbb{U}}\;\overline{\mathbb{U}}^T \to E[\overline{\mathbb{U}}_i\overline{\mathbb{U}}_i^T] = (n-1)^{-1}\mathbf{I}_{n-1}.$$

Furthermore, as the entries of the matrix $q^{-1}\widetilde{\mathbb{U}}\widetilde{\mathbb{U}}^T$ are bounded by 1,

$$p^{-1}\widetilde{\mathbb{U}}\widetilde{\mathbb{U}}^T = \mathbf{O}(q/p),$$

where $\mathbf{O}(u)$ is an $(n-1)\times(n-1)$ matrix whose entries are of order $O(u)$. Hence, as $\mathbb{U}\mathbb{U}^T = \overline{\mathbb{U}}\overline{\mathbb{U}}^T + \widetilde{\mathbb{U}}\widetilde{\mathbb{U}}^T$, the pseudo-inverse of $\mathbf{R}$ has the asymptotic large p representation

$$\mathbf{R}^\dagger = \left(\frac{n-1}{p}\right)^2 \mathbb{U}^T[\mathbf{I}_{n-1} + \mathbf{O}(q/p)]^{-2}\mathbb{U} = \left(\frac{n-1}{p}\right)^2 \mathbb{U}^T\mathbb{U}(1 + O(q/p)), \tag{53}$$

which establishes (23).

Define the partition $\mathcal{C} = \mathcal{Q} \cup \mathcal{Q}^c$ of the index set $\mathcal{C} = \{(i_0, \ldots, i_\delta) : 1 \leq i_0 \neq \cdot \neq i_\delta \leq p\}$ where $\mathcal{Q} = \{(i_0, \ldots, i_\delta) : 1 \leq i_l \leq q, 1 \leq l \leq \delta\}$ is the set of $(d+1)$-tuples restricted to the dependent columns $\widetilde{\mathbb{U}}$ of $\mathbb{U}$. The summation representations (30) and (32) of J and $\|\Delta_{p,n,k,\delta}\|_1$ yield

$$J(\overline{f_{\mathbf{Z}_{*1},\ldots,\mathbf{Z}_{*\delta+1}}}) = |\mathcal{C}|^{-1}\left(\sum_{\vec{i}\in\mathcal{Q}} + \sum_{\vec{i}\notin\mathcal{Q}}\right) J(f_{\mathbf{Z}_{i_0},\ldots,\mathbf{Z}_{i_\delta}}), \tag{54}$$

and

$$\|\Delta_{p,n,k,\delta}\|_1 = |\mathcal{C}|^{-1} \left(\sum_{\vec{i} \in \mathcal{Q}} + \sum_{\vec{i} \notin \mathcal{Q}} \right) \Delta_{p,n,k,\delta}(\vec{i}). \tag{55}$$

For correlation hub screening ($\mathbb{Z} = \mathbb{U}$) $\Delta_{p,n,k,\delta}(\vec{i}) = 0$ for all $\vec{i} \in \mathcal{C}$ while, as the set $\{\mathbf{U}_{i_0}, \ldots, \mathbf{U}_{i_\delta}\}$'s are i.i.d. uniform for $\vec{i} \in \mathcal{Q}$, $J(f_{\mathbf{Z}_{i_0},\ldots,\mathbf{Z}_{i_\delta}}) = 1$ for $\vec{i} \in \mathcal{Q}$. As $J(f_{\mathbf{Z}_{i_0},\ldots,\mathbf{Z}_{i_\delta}})$ is bounded and $|\mathcal{Q}^c|/|\mathcal{C}| = O\left(\delta(q/p)\right)$ the relations (24) and (25) are established for the case of correlation screening.

For partial correlation hub screening ($\mathbb{Z} = \mathbb{Y}$) then, as $\mathbb{Y} = [I_{n-1} + \mathbf{O}(q/p)]^{-1}\mathbb{U}$, the joint densities of $\mathbb{Y}$ and $\mathbb{U}$ are related by $f_{\mathbb{Y}} = (1+O(q/p))f_{\mathbb{U}}$. Therefore, over the range $\vec{i} \notin \mathcal{Q}$, the J and $\Delta_{p,n,k,\delta}$ summands in (54) and (55) are of order $1 + O(q/p)$ and $O(q/p)$, respectively, which establishes (24) and (25) for partial correlation screening. This completes the proof of Prop. 3. □

Author biographies

Alfred Hero received the Bachelor of Science (summa cum laude) from Boston University (1980) and the Ph.D from Princeton University (1984), both in Electrical Engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Williams Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. In 2008 he was awarded the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Superieure d'Electricite, Gif-sur-Yvette, France. He has held other visiting positions at organizations including: LIDS Massachusetts Institute of Technology (2006), Boston University (2006), I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Lucent Bell Laboratories (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Superieure des Techniques Avancees (ENSTA), Ecole Superieure d'Electricite, Paris (1990), and M.I.T. Lincoln Laboratory (1987 - 1989).

Alfred Hero is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). In 2011 he was awarded a University of Michigan Distinguished Faculty Achievement Award. He has been plenary and keynote speaker at several workshops and conferences. He has received several best paper awards including: a IEEE Signal Processing Society Best Paper Award (1998), the Best Original Paper Award from the Journal of Flow Cytometry (2008), and the Best Magazine Paper Award from the IEEE Signal Processing Society (2010). He received an IEEE Signal Processing Society Meritorious Service Award (1998), an IEEE Third Millenium Medal (2000) and an IEEE Signal Processing Society Distinguished Lecturership (2002). He was President of the IEEE Signal Processing Society (2006-2007). He has also served on the Board of Directors of IEEE (2009-2011) as Director Division IX (Signals and Applications).

Bala Rajaratnam received his Bachelor of Science with Honors in Mathematical Statistics and a Masters degree (with distinction recommendation) from the University of the Witwatersrand, Johannesburg, South Africa. He obtained his MS/PhD at Cornell University under the supervision of Michael Nussbaum in the Department of Mathematics and Martin Wells in the Department of Statistical Science. His postdoctoral research have been undertaken at SAMSI/Duke University, Stanford and Cambridge University. He is currently an assistant professor in the Department of Statistics at Stanford University. He also has a joint appointment in the Department of Environmental Earth System Science, and an affiliated faculty appointment at the Woods Institute for the Environment at Stanford University.

Bala Rajaratnam was awarded the DARPA young faculty award, DARPA's distinguished career award, in Mathematics in 2011. He was also awarded the mathematical geosciences award by the National Science Foundation, and was ranked first in the nation for the collaborative mathematical geosciences awards. He has also been the recipient of several other

awards and recognitions. He has been a keynote and invited speaker at various workshops/conferences/seminars.

Departments of EECS, BME and Statistics, University of Michigan - Ann Arbor, MI 48109-2122, U.S.A

E-mail address: hero@umich.edu

Department of Statistics, Stanford, CA 94305-4065, U.S.A.

E-mail address: brajarat@stanford.edu