Gains in Power from Structured Two-Sample Tests of Means on Graphs

Laurent Jacob, Pierre Neuvial, Sandrine Dudoit

Gains in Power from Structured Two-Sample Tests of Means on Graphs

Laurent Jacob
Department of Statistics
University of California, Berkeley
laurent@stat.berkeley.edu

Pierre Neuvial
Department of Statistics
University of California, Berkeley
pierre@stat.berkeley.edu

Sandrine Dudoit
Division of Biostatistics and Department of Statistics
University of California, Berkeley
sandrine@stat.berkeley.edu

September 27, 2010

Abstract

We consider multivariate two-sample tests of means, where the location shift between the two populations is expected to be related to a known graph structure. An important application of such tests is the detection of differentially expressed genes between two patient populations, as shifts in expression levels are expected to be coherent with the structure of graphs reflecting gene properties such as biological process, molecular function, regulation, or metabolism. For a fixed graph of interest, we demonstrate that accounting for graph structure can yield more powerful tests under the assumption of smooth distribution shift on the graph. We also investigate the identification of non-homogeneous subgraphs of a given large graph, which poses both computational and multiple testing problems. The relevance and benefits of the proposed approach are illustrated on synthetic data and on breast cancer gene expression data analyzed in context of KEGG pathways.

1 Introduction

The problem of testing whether two data generating distributions are equal has been studied extensively in the statistical and machine learning literatures. Practical applications range from speech recognition to fMRI and genomic data analysis.

Parametric approaches typically test for divergence between two distributions using statistics based on a standardized difference of the two sample means, *e.g.*, Student's t-statistic in the univariate case or Hotelling's T^2-statistic in the multivariate case [Lehmann and Romano, 2005]. A variety of non-parametric rank-based tests have also been proposed. More recently, Harchaoui et al. [2007] and Gretton et al. [2007] devised kernel-based statistics for homogeneity tests in a function space.

In several settings of interest, prior information on the structure of the distribution shift is available as a graph on the variables. Specifically, suppose we observe $\{X_1^1, \ldots, X_{n_1}^1\} \in \mathbb{R}^p$ from a first multivariate normal distribution $\mathcal{N}(\mu_1, \Sigma)$ and $\{X_1^2, \ldots, X_{n_2}^2\} \in \mathbb{R}^p$ from a second such distribution $\mathcal{N}(\mu_2, \Sigma)$. In cases where an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encoding some type of covariance information in $\mathbb{R}^p$ is given, the putative *location* or *mean shift* $\delta = \mu_1 - \mu_2$ may be expected to be coherent with $\mathcal{G}$. That is, δ viewed as a function of $\mathcal{G}$ is *smooth*, in the sense that the shifts δ_i and δ_j for two connected nodes v_i and $v_j \in \mathcal{V}$ are similar. Classical tests, such as Hotelling's T^2-test, consider the null hypothesis $\mathbf{H}_0 : \mu_1 = \mu_2$ against the alternative $\mathbf{H}_1 : \mu_1 \neq \mu_2$, without reference to the graph. Our goal is to take into account the graph structure of the variables in order to build a more powerful two-sample test of means under smooth-shift alternatives.

Just as a natural notion of smoothness of functions on a Euclidean space can be defined through the notion of Dirichlet energy and controlled by Fourier decomposition and filtering [Stain and Weiss, 1971], it is well-known [Chung, 1997] that the smoothness of functions on a graph can be naturally defined and controlled through spectral analysis of the graph Laplacian. In particular, the eigenvectors of the Laplacian provide a basis of functions which vary on the graph at increasing frequencies (corresponding to the increasing eigenvalues). In this paper, we propose to compare two populations in terms of the first few components of the graph-Fourier basis or, equivalently, in the original space, after filtering out high-frequency components.

An important motivation for the development of our graph-structured test is the detection of groups of genes whose expression changes between two conditions. For example, identifying groups of genes that are differentially expressed (DE) between patients for which a particular treatment is effective and patients which are resistant to the treatment may give insight into the resistance mechanism and even suggest targets for new drugs. In such a context, expression data from high-throughput microarray and sequencing assays gain much in relevance from their association with graph-structured prior information on the genes, *e.g.*, Gene Ontology (GO; `http://www.geneontology.org`) or Kyoto Encyclopedia of Genes and Genomes (KEGG; `http://www.genome.jp/kegg`). Most approaches to the joint analysis of gene expression data and gene graph data involve two distinct steps. Firstly, tests of differential expression are performed separately for each gene. Then, these univariate (gene-level) testing results are extended to the level of gene sets, *e.g.*, by assessing the over-representation of DE genes in each set based on p-values for Fisher's exact test (or a χ^2 approximation thereof) adjusted for multiple testing [Beissbarth and Speed, 2004] or based on permutation adjusted p-values for weighted Kolmogorov-Smirnov-like statistics [Subramanian et al., 2005]. Another family of methods directly performs multivariate tests of differential expression for groups of genes, *e.g.*, Hotelling's T^2-test [Lu et al., 2005]. It is known [Goeman

and Bühlmann, 2007] that the former family of approaches can lead to incorrect interpretations, as the sampling units for the tests in the second step become the genes (as opposed to the patients) and these are expected to have strongly correlated expression measures. This suggests that direct multivariate testing of gene set differential expression is more appropriate than posterior aggregation of individual gene-level tests. On the other hand, while Hotelling's T^2-statistic is known to perform well in small dimensions, it loses power very quickly with increasing dimension [Bai and Saranadasa, 1996], essentially because it is based on the inverse of the empirical covariance matrix which becomes ill-conditioned. In addition, such direct multivariate tests on unstructured gene sets do not take advantage of information on gene regulation or other relevant biological properties. An increasing number of regulation networks are becoming available, specifying, for example, which genes activate or inhibit the expression of which other genes. As stated before, incorporating such biological knowledge in DE tests is important. Indeed, if it is known that a particular gene in a tested gene set activates the expression of another, then one expects the two genes to have coherent (differential) expression patterns, *e.g.*, higher expression of the first gene in resistant patients should be accompanied by higher expression of the second gene in these patients. Accordingly, the first main contribution of this paper is to propose and validate multivariate test statistics for identifying distribution shifts that are coherent with a given graph structure.

Next, given a large graph and observations from two data generating distributions on the graph, a more general problem is the identification of smaller non-homogeneous subgraphs, *i.e.*, subgraphs on which the two distributions (restricted to these subgraphs) are significantly different. This is very relevant in the context of tests for gene set differential expression: given a large set of genes, together with their known regulation network, or the concatenation of several such overlapping sets, it is important to discover novel gene sets whose expression change significantly between two conditions. Currently-available gene sets have often been defined in terms of other phenomena than that under study and physicians may be interested in discovering sets of genes affecting in a concerted manner a specific phenotype. Our second main contribution is therefore to develop algorithms that allow the exhaustive testing of all the subgraphs of a large graph, while accounting for the multiplicity issue arising from the vast number of subgraphs.

As the problem of identifying variables or groups of variables which differ in distribution between two populations is closely-related to supervised learning, our proposed approach is similar to several learning methods. Rapaport et al. [2007] use filtering in the Fourier space of a graph to train linear classifiers of gene expression profiles whose weights are smooth on a gene network. However, their classifier enforces global smoothness on the large regularization network of all the genes, whereas we are concerned with the selection of gene sets with locally-smooth expression shift between populations. In Jacob et al. [2009], sparse learning methods are used to build a classifier based on a small number of gene sets. While this approach leads in practice to the selection of groups of variables whose distributions differ between the two classes, the objective is to achieve the best classification performance with the smallest possible number of groups. As a result, correlated groups of variables are typically not selected. Other related work includes Fan

and Lin [1998], who proposed an adaptive Neyman test in the Fourier space for time-series. However, as illustrated below in Section 5, direct translation of the adaptive Neyman statistic to the graph case is problematic, as assumptions on Fourier coefficients which are true for time-series do not hold for graphs. In addition, the Neyman statistic converges very slowly towards its asymptotic distribution and the required calibration by bootstrapping renders its application to our subgraph discovery context difficult. By contrast, other methods do not account for shift smoothness and try to address the loss of power caused by the poor conditioning of the T^2-statistic by applying it after dimensionality reduction [Ma and Kosorok, 2009] or by omitting the inverse covariance matrix and adjusting instead by its trace [Bai and Saranadasa, 1996, Chen and Qin, 2010]. Vaske et al. [2010] recently proposed DE tests, where a probabilistic graphical model is built from a gene network. However, this model is used for gene-level DE tests, which then have to be combined to test at the level of gene sets. Several approaches for subgraph discovery, like that of Ideker et al. [2002], are based on a heuristic to identify the most differentially expressed subgraphs and do not amount to testing exactly all the subgraphs. Concerning the discovery of distribution-shifted subgraphs, Vandin et al. [2010] propose a graph Laplacian-based testing procedure to identify groups of interacting proteins whose genes contain a large number of mutations. Their approach does not enforce any smoothness on the detected patterns (smoothness is not necessarily expected in this context) and the graph Laplacian is only used to ensure that very connected genes do not lead to spurious detection. The Gene Expression Network Analysis (GXNA) method of Nacu et al. [2007] detects differentially expressed subgraphs based on a greedy search algorithm and gene set DE scoring functions that do not account for the graph structure.

The rest of this paper is organized as follows: Section 2 introduces elements of Fourier analysis for graphs which are needed to develop our method. Section 3 presents our graph-structured two-sample test statistic and states results on power gain for smooth-shift alternatives. Section 4 describes procedures for systematically testing all the subgraphs of a large graph. Section 5 presents results for synthetic data as well as breast cancer gene expression and KEGG data. Finally, Section 6 summarizes our findings and outlines ongoing work.

2 Fourier analysis on graphs

The fundamental idea of harmonic analysis for functions defined on a Euclidean space is to build a basis of the function space, such that each basis function varies at a different frequency. The basis functions are typically sinusoids. They were originally obtained in an attempt to solve the heat equation, as the eigenfunctions of the Laplace operator, with corresponding eigenvalues proportional to the frequencies of the sinusoids. Any function can then be decomposed on the basis as a linear combination of sinusoids of increasing frequency. The set of projections of the function on the basis sinusoids gives a dual representation of the function, often referred to as Fourier transform. This representation is useful for filtering functions, by removing or shrinking coefficients associated with high frequencies, as these are

typically expected to reflect noise, and then taking the inverse Fourier transform. The resulting filtered function contains the same signal in the low frequencies as the original function. A related concept is the Dirichlet energy of a function f on an open subspace Ω, defined as $\frac{1}{2}\int_\Omega |\nabla f(x)|^2 dx$ where ∇ is the gradient operator, a measure of variation that is consistent with the Laplace operator. In particular, the Dirichlet energy of the basis functions is proportional to their associated frequencies.

For functions on a Euclidean space, natural notions of smoothness, along with the Dirichlet energy and dual representation in the frequency domain by projection on a Fourier basis, are therefore classically defined from the Laplace operator and its spectral decomposition. Likewise, notions of smoothness for functions on graphs can be defined based on the graph Laplacian. Specifically, consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = p$ nodes, adjacency matrix A, and degree matrix $D = \mathrm{Diag}\,(A\mathbf{1})$, where $\mathbf{1}$ is a unit column-vector, $\mathrm{Diag}(x)$ is the diagonal matrix with diagonal x for any vector x, and $D_{ii} = d_i$. Let $f : \mathbb{R}^{|\mathcal{V}|} \to \mathbb{R}$ denote a function that associates a real value to each node of the graph $\mathcal{G}$. The Laplacian matrix of $\mathcal{G}$ is typically defined as $\mathcal{L} = D - A$ or $\mathcal{L}_{norm} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ for the normalized version. More generally, given any gradient matrix $\nabla \in \mathbb{R}^{|\mathcal{E}|,|\mathcal{V}|}$, defined on $\mathcal{G}$ and associating to each function on the graph its variation on each edge, it is possible to derive a corresponding Laplacian matrix following the classical definition of the Laplace operator, $\mathcal{L} = -\mathrm{div}\nabla = \nabla^\top \nabla$, where div is the divergence operator defined as the negative of the adjoint operator of the gradient [Zhou and Schölkopf, 2005]. Any desired notion of variation may be encoded in a gradient function and thus translated into its associated Dirichlet energy $\frac{1}{2} f^\top \mathcal{L} f$, for a function f defined on the graph $\mathcal{G}$. A common choice of gradient is the finite difference operator $\nabla f = (f_i - f_j)_{i,j\in\mathcal{V}}$. This definition leads to the unnormalized Laplacian above. The corresponding energy function is $\frac{1}{2}\sum_{i,j\in\mathcal{V}} (f_i - f_j)^2$. Let $\mathcal{L} = U\Lambda U^\top$ denote the spectral decomposition of the Laplacian, where Λ is the diagonal matrix of eigenvalues λ_i and the columns of the matrix U are the corresponding eigenvectors u_i. Then, by definition, the eigenvectors of $\mathcal{L}$ are functions of increasing energy, as $u_i^\top \mathcal{L} u_i = \lambda_i$ for all $i = 1, \ldots, p$. In the remainder of this paper, we denote by $\tilde{f} = U^\top f$ the Fourier coefficients of a function f defined on a graph.

If the above two notions of smoothness are not appropriate for a particular application, other gradients, leading to other Laplacian matrices, may be devised to build the function basis. For example, introducing weights on the edges of a graph and using these weights in the normalized version of the finite differences allows the incorporation of prior belief on where a shift in distributions is expected to be smooth. For applications like structured gene set differential expression detection, one may use negative weights for edges that reflect an expected negative correlation between two variables, *e.g.*, a gene i whose expression inhibits the expression of another gene j. In this case, a small variation of the shift on the edge between i and j should correspond to a small $|\delta_i + \delta_j|$. Accordingly, the gradient should be defined as $(f_i - s_{ij} f_j)_{i,j\in\mathcal{V}}$, where s_{ij} is -1 for negative interactions and 1 for positive interactions. The eigenvectors of the corresponding Laplacian $\mathcal{L}_{\text{sign}}$ are functions of increasing $\frac{1}{2}\sum_{i,j\in\mathcal{V}} (f_i - s_{ij} f_j)^2$, an appropriate notion of smoothness for the application at hand. A signed Laplacian can be recovered from the classical definition $\mathcal{L}_{\text{sign}} = D - A_{\text{sign}}$, where A_{sign} is allowed to have negative entries. Note

that such a smoothness function is used as a penalty for semi-supervised learning in Goldberg [2007].

As an example, Figure 1 displays the eigenvectors of the signed Laplacian $\mathcal{L}_{\text{sign}}$ for a simple four-node graph with

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A_{\text{sign}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad \mathcal{L}_{\text{sign}} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

The first eigenvector, corresponding to the smallest frequency (eigenvalue of zero), can be viewed as a "constant" function on the graph, in the sense that its absolute value is identical for all the nodes, but nodes connected by an edge with negative weight take on values of opposite in sign. By contrast, the last eigenvector, corresponding to the highest frequency, is such that nodes connected by positive edges take on values of opposite sign and nodes connected by negative edges take on values of the same sign.

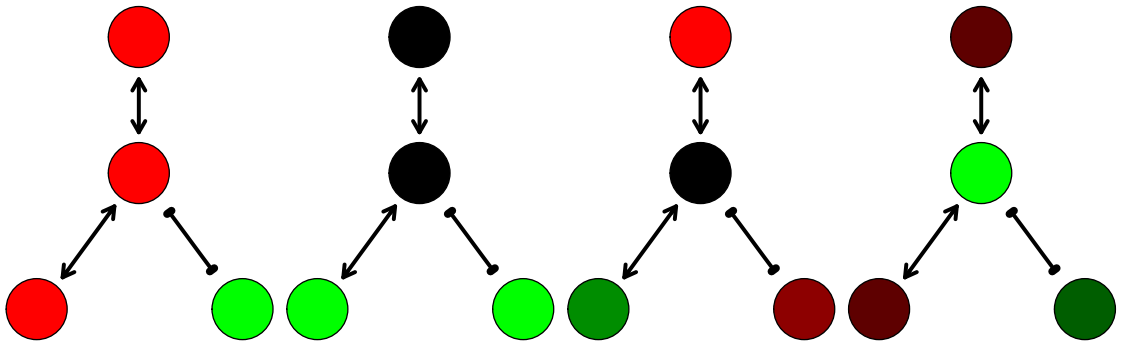

Figure 1: Eigenvectors of the signed Laplacian $\mathcal{L}_{\text{sign}}$ for a simple four-node graph. The corresponding eigenvalues are $0, 1, 1, 4$. Nodes are colored according to the value of the eigenvector, where green corresponds to high positive values, red to high negative values, and black to 0. "T"-shaped edges have negative weights.

3 Graph-structured two-sample test of means under smooth-shift alternatives

For multivariate normal distributions, Hotelling's T^2-test, a classical test of location shift, is known to be uniformly most powerful invariant against global-shift alternatives. The test statistic is based on the squared *Mahalanobis norm* of the sample mean shift and is given by $T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top \hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_2)$, where n_i, $\bar{x}_i$, and $\hat{\Sigma}^{-1}$ denote, respectively, the sample sizes, means, and pooled covariance matrix, for random samples drawn from two p-dimensional Gaussian distributions, $\mathcal{N}(\mu_i, \Sigma)$, $i = 1, 2$. Under the null hypothesis $\mathbf{H}_0 \ : \mu_1 = \mu_2$ of equal means, NT^2 follows a (central) F-distribution $F_0(p, n_1 + n_2 - p - 1)$, where $N = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p}$. In general,

NT^2 follows a non-central F-distribution $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2(\delta,\Sigma); p, n_1+n_2-p-1)$, where the non-centrality parameter is a function of the Mahalanobis norm of the mean shift δ, $\Delta^2(\delta,\Sigma) = \delta^\top \Sigma^{-1}\delta$, which we refer to as *distribution shift*. In the remainder of this paper, unless otherwise specified, T^2-statistics are assumed to follow the nominal F-distribution, *e.g.*, for critical value and power calculations.

For any graph-Fourier basis U, direct calculation shows that $T^2 = \tilde{T}^2 \triangleq \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top U \left(U^\top \hat{\Sigma} U\right)^{-1} U^\top(\bar{x}_1 - \bar{x}_2)$, *i.e.*, the statistic T^2 in the original space and the statistic $\tilde{T}^2$ in the graph-Fourier space are identical. More generally, for $k \leq p$, the statistic in the original space after filtering out frequencies above k is the same as the statistic $\tilde{T}_k^2$ restricted to the first k coefficients in the graph-Fourier space:

$$\begin{aligned}\tilde{T}_k^2 &\triangleq \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top U_{[k]} \left(U_{[k]}^\top \hat{\Sigma} U_{[k]}\right)^{-1} U_{[k]}^\top(\bar{x}_1 - \bar{x}_2) \\ &= \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)^\top U 1_k U^\top \left(U 1_k U^\top \hat{\Sigma} U 1_k U^\top\right)^{+} U 1_k U^\top(\bar{x}_1 - \bar{x}_2),\end{aligned}$$

where A^+ denotes the generalized inverse of a matrix A, the $p \times k$ matrix $U_{[k]}$ denotes the restriction of U to its first k columns, and 1_k is a $p \times p$ diagonal matrix, with ith diagonal element equal to one if $i \leq k$ and zero otherwise. Note that as retaining the first k Fourier components is a *non-invertible* transformation, this filtering indeed has an effect on the test statistic, that is, we have $\tilde{T}_k^2 \neq \tilde{T}^2$ in general. As the Mahalanobis norm is invariant to linear invertible transformations, using an invertible filtering (such as weighting each Fourier component according to its corresponding eigenvalue) would have no impact on the test statistic.

Hotelling's T^2-test is known to behave poorly in high dimension; the following lemma shows that gains in power can be achieved by filtering. Specifically, let $\tilde{\delta} = U^\top \delta$ and $\tilde{\Sigma} = U^\top \Sigma U$ denote, respectively, the mean shift and covariance matrix in the graph-Fourier space. Given $k \leq p$, let $\Delta_k^2(\delta,\Sigma) = \delta_{[k]}^\top \left(\Sigma_{[k]}\right)^{-1} \delta_{[k]}$ denote the distribution shift restricted to the first k dimensions of δ and Σ, *i.e.*, based on only the first k elements of δ, $(\delta_i : i \leq k)$, and the first $k \times k$ diagonal block of Σ, $(\sigma_{ij} : i,j \leq k)$. Under the assumption that the distribution shift is smooth, *i.e.*, lies mostly at the beginning of the graph spectrum, so that $\Delta_k^2(\tilde{\delta},\tilde{\Sigma})$ is nearly maximal for a small value of k, Lemma 1 states that performing Hotelling's test in the graph-Fourier space restricted to its first k components yields more power than testing in the full graph-Fourier space. Equivalently, the test is more powerful in the original space after filtering than in the original unfiltered space. Note that this result holds because retaining the first k Fourier components is a *non-invertible* transformation.

Lemma 1. *For any level α and any $1 < l \leq p-k$, there exists $d(\alpha,k,l) > 0$ such that*

$$\Delta_{k+l}^2(\tilde{\delta},\tilde{\Sigma}) - \Delta_k^2(\tilde{\delta},\tilde{\Sigma}) < d(\alpha,k,l) \Rightarrow \beta_{\alpha,k}(\Delta_k^2(\tilde{\delta},\tilde{\Sigma})) > \beta_{\alpha,k+l}(\Delta_{k+l}^2(\tilde{\delta},\tilde{\Sigma})),$$

where $\beta_{\alpha,k}(\Delta^2)$ is the power of Hotelling's T^2-test at level α in dimension k for a distribution shift Δ^2, according to the nominal F-distribution $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2; k, n_1 + n_2 - k - 1)$.

Proof. This lemma is a direct application of Corollary 2.1 in Das Gupta and Perlman [1974] to Hotelling's T^2-test in the graph-Fourier space. The bottom line of the proof of Das Gupta and Perlman [1974]'s result is that $\beta_{\alpha,k}$ can be shown to be a continuous and strictly decreasing function of k, so that a strictly positive increase in the non-centrality parameter Δ^2 of the F-distribution is necessary to maintain power when increasing dimension. □

In particular, a direct application of Lemma 1 yields the following Corollary:

Corollary 1. *If* $\forall\ 1 < l \leq p - k,\ \Delta_k^2(\tilde{\delta}, \tilde{\Sigma}) = \Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma})$*, then*

$$\beta_{\alpha,k}(\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})) > \beta_{\alpha,k+l}(\Delta_{k+l}^2(\tilde{\delta}, \tilde{\Sigma})).$$

According to Corollary 1, if the distribution shift lies in the first k Fourier coefficients, then testing in this subspace yields strictly more power than using additional coefficients. In particular, if there exists $k < p$ such that $\tilde{\delta}_j = 0\ \forall\ j > k$ (*i.e.*, the mean shift is smooth) and $\tilde{\Sigma}$ is block-diagonal such that $\tilde{\sigma}_{ij} = 0\ \forall\ i < k, j > k$, then gains in power are obtained by testing in the first k Fourier components. Although non-necessary, this condition is plausible when the mean shift lies at the beginning of the spectrum, as the coefficients which do not contain the shift are not expected to be correlated with the ones that do contain it.

Note that the result in Lemma 1 is even more general, as testing in the first k Fourier components can increase power even when the distribution shift partially lies in the remaining components, as long as the latter portion is below a certain threshold. Figure 2 illustrates, under different settings, the increase in distribution shift necessary to maintain a given power level against the number of added coefficients.

If for some reason one expects that the mean shift δ is smooth (rather than the distribution shift Δ), *i.e.*, $\tilde{\delta}$ lies at the beginning of the spectrum, and that the covariance between coefficients that contain the shift and those that do not is non-zero, then one should use test statistics based on estimators of the unstandardized *Euclidean norm* $\|\delta\|$ of this shift, *e.g.*, Z [Bai and Saranadasa, 1996][Equation (4.5)] or T_n [Chen and Qin, 2010]. Results similar to Lemma 1 can be derived for these statistics. Namely, the corresponding tests gain asymptotic power when applied at the beginning of the spectrum, provided the Euclidean norm of δ only increases moderately as coefficients for higher frequencies are added. The results follow from Bai and Saranadasa [1996][Theorem 4.1] and Chen and Qin [2010][Equations (3.11)–(3.12)], using the fact that, by Cauchy's interlacing theorem, the trace of the square of any positive semi-definite matrix is larger than the trace of the square of any principal submatrix.

4 Non-homogeneous subgraph discovery

A systematic approach for discovering non-homogeneous subgraphs, *i.e.*, subgraphs of a large graph that exhibit a significant shift in means, is to test all of them. In practice, however, this can represent an intractable number of tests, so it is important to be able to rapidly identify sets of subgraphs that all satisfy the null

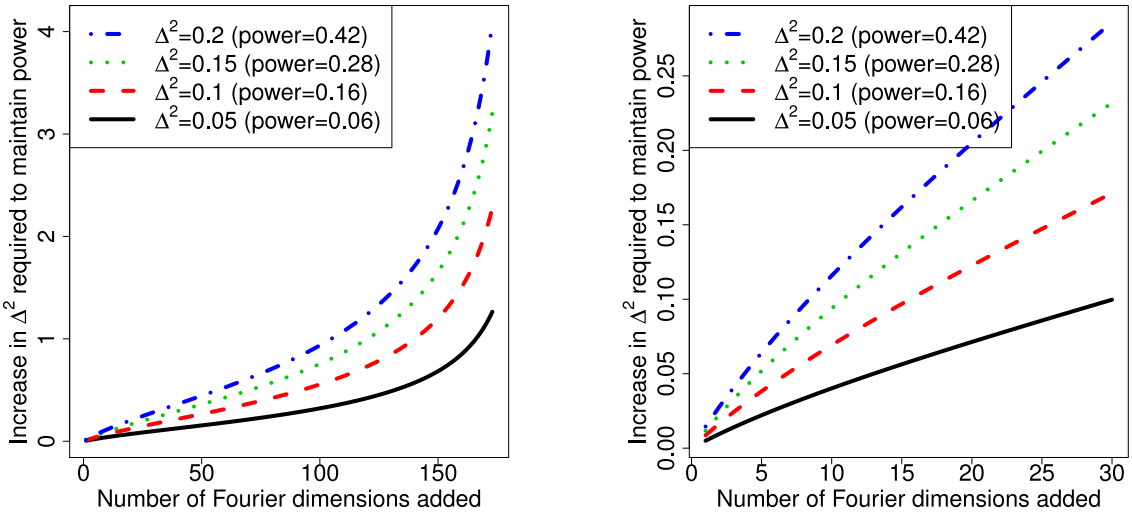

Figure 2: Left: Increase in distribution shift required for Hotelling's T^2-test to maintain a given power when increasing the number of tested Fourier coefficients: $\Delta^2_{k+l} - \Delta^2_k$ vs. l such that $\beta_{\alpha,k+l}(\Delta^2_{k+l}) = \beta_{\alpha,k}(\Delta^2_k)$. Power $\beta_{\alpha,k+l}(\Delta^2_{k+l})$ computed under the non-central F-distribution $F(\frac{n_1 n_2}{n_1+n_2}\Delta^2_{k+l}; k+l, n_1+n_2-(k+l)-1)$, for $n_1 = n_2 = 100$ observations, $k = 5$, and $\alpha = 10^{-2}$. Each line corresponds to the fixed shift Δ^2_k and power $\beta_{\alpha,k}(\Delta^2_k)$ pair indicated in the legend. Right: Zoom on the first 30 dimensions.

hypothesis of equal means. To this end, we devise a pruning approach based on an upper bound on the value of the test statistic for any subgraph containing a given set of nodes.

4.1 Exact algorithm

Given a large graph $\mathcal{G}$ with p nodes, we adopt the following classical branch-and-bound-like approach to test subgraphs of size $q \leq p$ at level α. We start by checking, for each node in $\mathcal{G}$, whether the Hotelling T^2-statistic in the first k graph-Fourier components of any subgraph of size q containing this node can be guaranteed to be below the level-α critical value $T^2_{\alpha,k}$ (*e.g.*, $(1-\alpha)$-quantile of $F_0(k, n_1+n_2-k-1)$ distribution). If this is the case, the node is removed from the graph. We then repeat the procedure on the edges of the remaining graph and, iteratively, on the subgraphs up to size $q-1$, at which point we test all the remaining subgraphs of size q.

Specifically, for a subgraph g of $\mathcal{G}$ of size $q \leq p$, Hotelling's T^2-statistic in the first $k \leq q$ graph-Fourier components of g is defined as

$$\tilde{T}^2_k(g) = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1(g) - \bar{x}_2(g))^\top U_{[k]} \left(U_{[k]}^\top \hat{\Sigma}(g) U_{[k]}\right)^{-1} U_{[k]}^\top (\bar{x}_1(g) - \bar{x}_2(g)),$$

where $U_{[k]}$ is the $q \times k$ restriction of the matrix of q eigenvectors of the Laplacian of g to its first k columns (*i.e.*, $U_{[k]}(g)$, where we omit g to ease notation) and $\bar{x}_i(g)$, $i = 1, 2$, and $\hat{\Sigma}(g)$ are, respectively, the empirical means and pooled covariance matrix restricted to the nodes in g. We make use of the following upper bound on $\tilde{T}^2_k(g)$.

Lemma 2 (Upper bound on $\tilde{T}_k^2$)**.** *For any subgraph g of $\mathcal{G}$ of size $q \leq p$, any subgraph g' of g of size $q' \leq q$, and any $k \leq q$, then*

$$\tilde{T}_k^2(g) \leq T^2(\nu(g', q - q'))\,,$$

where $\nu(g', r)$ is the r-neighborhood of g', that is, the union of the nodes of g' and the nodes whose shortest path to a node of g' is less than or equal to r.

The proof involves the following result:

Lemma 3 (Bessel inequality for Mahalanobis norm)**.** *Let $\Sigma \in \mathbb{R}^{p,p}$ be an invertible matrix and $P \in \mathbb{R}^{p,k}$, $k \leq p$, be a matrix with orthonormal columns. For any $x \in \mathbb{R}^p$,*

$$x^\top \Sigma^{-1} x \geq x^\top P \left(P^\top \Sigma P\right)^{-1} P^\top x.$$

Proof. First note that, by orthonormality of the columns of P, $P^\top \Sigma P$ is indeed invertible, and that

$$\Sigma^{-1} - P \left(P^\top \Sigma P\right)^{-1} P^\top = \Sigma^{-\frac{1}{2}} \left(I - \Sigma^{\frac{1}{2}} P \left(P^\top \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} P\right)^{-1} P^\top \Sigma^{\frac{1}{2}}\right) \Sigma^{-\frac{1}{2}},$$

where $\Sigma^{\frac{1}{2}} P \left(P^\top \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} P\right)^{-1} P^\top \Sigma^{\frac{1}{2}}$ is an orthogonal projection, with eigenvalues either 0 or 1. Thus, $I - \Sigma^{\frac{1}{2}} P \left(P^\top \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} P\right)^{-1} P^\top \Sigma^{\frac{1}{2}}$ is positive-semi-definite, as its eigenvalues are also either 0 or 1. The result follows from properties of products of positive-semi-definite matrices. □

We can now prove Lemma 2.

Proof. By Lemma 3,

$$\begin{aligned}
\tilde{T}_k^2(g) &\leq \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1(g) - \bar{x}_2(g))^\top U \left(U^\top \hat{\Sigma}(g) U\right)^{-1} U^\top (\bar{x}_1(g) - \bar{x}_2(g)) \\
&= \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1(g) - \bar{x}_2(g))^\top \hat{\Sigma}(g)^{-1} (\bar{x}_1(g) - \bar{x}_2(g)) = T^2(g).
\end{aligned}$$

As $g \subset \nu(g', q - q')$, applying Lemma 3 a second time with the compression from $\nu(g', q - q')$ to the nodes of g yields the result. □

Note that the bound takes into account the fact that the T^2-statistic is eventually computed in the first few components of a basis which is not known beforehand : at each step, for each potential subgraph g' which would include the subgraph g which we consider for pruning, the $\tilde{T}_k^2(g')$ that we need to upper bound depends on the graph Laplacian of g'.

4.2 Mean-shift approximation

For "small-world" graphs above a certain level of connectivity and q large enough, the $(q-s)$-neighborhood of g', $\nu(g', q-s)$, tends to be large, at least at the beginning of the above exact algorithm, and the number of tests actually performed won't decrease much compared to the total number of possible tests. One can, however, identify much more efficiently the subgraphs whose sample mean shift in the first k components of the graph-Fourier space has Euclidean norm $\|\hat{\tilde{\delta}}_{[k]}(g)\| = \|U_{[k]}^\top(\bar{x}_1(g) - \bar{x}_2(g))\|$ above a certain threshold. Indeed, it is straightforward to see that

$$\begin{aligned}\|U_{[k]}^\top(\bar{x}_1(g) - \bar{x}_2(g))\|^2 &\leq \|U^\top(\bar{x}_1(g) - \bar{x}_2(g))\|^2 \\ &= \|\bar{x}_1(g) - \bar{x}_2(g)\|^2 \\ &\leq \|\bar{x}_1(g') - \bar{x}_2(g')\|^2 \\ &+ \max_{v_1,\ldots,v_{q-s}\in\nu(g',q-s)} \|\bar{x}_1(v_1,\ldots,v_{q-s}) - \bar{x}_2(v_1,\ldots,v_{q-s})\|^2.\end{aligned}$$

This inequality can then be used in the procedure of Section 4.1, to identify all subgraphs for which the Euclidean norm of the sample mean shift exceeds a given threshold: $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2 > \theta$. For any α, if this threshold θ is low enough, all the subgraphs with $\tilde{T}_k^2(g) > T_{\alpha,k}^2$ are included in this set. Performing the actual T^2-test on these pre-selected subgraphs yields exactly the set of subgraphs that would have been identified using the exact procedure of Section 4.1. More precisely, we have the following result:

Lemma 4. *For any threshold $\theta > 0$, $k \leq q \leq p$, and any subgraph g of size q such that $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2 < \theta$,*

$$N\tilde{T}_k^2(g) > T_{\alpha,k}^2 \Rightarrow \lambda_{min}(\hat{\tilde{\Sigma}}_{[k]}(g)) < \frac{Nn_1n_2\theta}{(n_1+n_2)T_{\alpha,k}^2},$$

*where $T_{\alpha,k}^2$ is the level-α critical value for $\tilde{T}_k^2$ (*e.g.*, $(1-\alpha)$-quantile of $F_0(k, n_1+n_2-k-1)$), $N = \frac{n_1+n_2-k-1}{(n_1+n_2-2)k}$ and $\lambda_{min}(\hat{\tilde{\Sigma}}_{[k]}(g))$ denotes the smallest eigenvalue of $\hat{\tilde{\Sigma}}_{[k]}(g) = U_{[k]}\hat{\Sigma}(g)U_{[k]}^\top$.*

Proof. As $I - (\hat{\tilde{\Sigma}}_{[k]}(g))^{-1}\lambda_{min}(\hat{\tilde{\Sigma}}_{[k]}(g)) \succeq 0$, it follows that, for any x,

$$x^\top(\hat{\tilde{\Sigma}}_{[k]}(g))^{-1}x \leq \frac{\|x\|^2}{\lambda_{min}(\hat{\tilde{\Sigma}}_{[k]}(g))}.$$

□

Lemma 4 states that for any subgraph which would be detected by Hotelling's T^2-statistic $\tilde{T}_k^2(g)$ but not by the Euclidean criterion $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2$, the sample covariance matrix in the restricted graph-Fourier space has an eigenvalue below a certain threshold. This implies that such false negative subgraphs (from the Euclidean approximation to the exact algorithm) always have a small mean shift in

the graph-Fourier space, but in a direction of small variance. In context of gene expression, this is related to the well-known issue of the detection of DE genes by virtue of their small variances. Even though the differences in expression appear to be highly significant for these genes, they correspond to small effects that are not interesting from a practical point of view (*i.e.*, biologically insignificant). Methods for addressing this problem are proposed in Lönnstedt and Speed [2001]. Note that $\lambda_{min}(\hat{\Sigma}(g))) \leq \lambda_{min}(\hat{\tilde{\Sigma}}_{[k]}(g)))$; thus, the remark on variances holds for both the graph-Fourier and original spaces. However, if q is large, we expect $\lambda_{min}(\hat{\Sigma}(g))$ to be very small, while filtering somehow controls the conditioning of the covariance matrix.

4.3 Multiple testing

Testing for homogeneity over the potentially large number of subgraphs investigated as part of the above algorithms immediately raises the issue of multiple testing. However, the present multiplicity problem is unusual, in the sense that one does not know in advance the total number of tests and which tests will be performed specifically. Standard multiple testing procedures, such as those in Dudoit and van der Laan [2008], are therefore not immediately applicable.

In an attempt to address the multiplicity issue, we apply a permutation procedure to control the number of false positive subgraphs under the complete null hypothesis of identical distributions in the two populations. Specifically, one permutes the class/population labels (1 or 2) of the $n_1 + n_2$ observations and applies the non-homogeneous subgraph discovery algorithm to the permuted data to yield a certain number of false positive subgraphs. Repeating this procedure a sufficiently large number of times produces an estimate of the distribution of the number of Type I errors under the complete null hypothesis of identical distributions.

5 Results

We evaluate the empirical behavior of the procedures proposed in Sections 3 and 4, first on synthetic data, then on breast cancer microarray data analyzed in context of KEGG pathways.

5.1 Synthetic data

The performance of the graph-structured test is assessed in cases where the distribution shift Δ^2 satisfies the smoothness assumptions described in Section 3. We first generate a connected random graph with $p = 20$ nodes. Next, we generate $10,000$ datasets, each comprising $n_1 = n_2 = 20$ Gaussian random vectors in $\mathbb{R}^p$, with null mean shift δ for $5,000$ datasets and mean shift $\delta = 1$ for the remaining $5,000$. For the latter datasets, the non-zero shift is built in the first $k_0 = 3$ Fourier coefficients (the shift being zero for the remaining $p - k_0$ coefficients) and an inverse Fourier transformation is applied to random vectors generated in the graph-Fourier space. We consider two covariance settings: in the first one, the covariance matrix

in the graph-Fourier space is diagonal with diagonal elements at $\frac{1}{\sqrt{p}}$. In the second one, correlation is introduced between the shifted coefficients only. Specifically, for $i, j \leq k_o$, $\Sigma_{ij} = \frac{0.5}{\sqrt{p}}$ if $i \neq j$, $\Sigma_{ii} = \frac{0.9}{\sqrt{p}}$ otherwise.

Figure 3 displays receiver operator characteristic (ROC) curves for mean shift detection by the standard Hotelling T^2-test, T^2 in the first k_0 Fourier coefficients, T^2 in the first k_0 principal components (PC), the adaptive Neyman test of Fan and Lin [1998], and a modified version of this test where the correct value of k_0 is specified. Note that we do not consider sparse learning approaches [Jacob et al., 2009, Jenatton et al., 2009], but it would be straightforward to design a realistic setting where such approaches are outperformed by testing, *e.g.*, by adding correlation between some of the functions under $\mathbf{H}_1$.

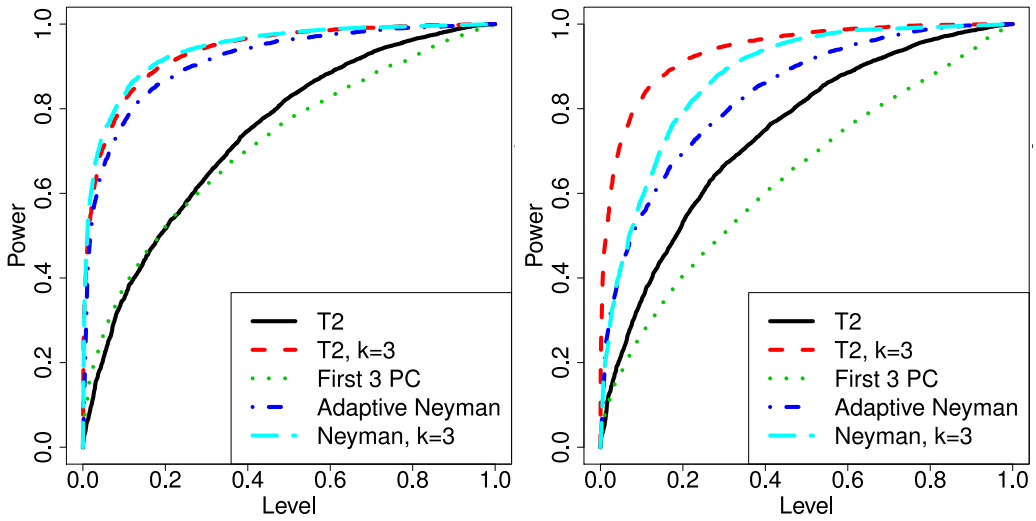

Figure 3: ROC curves for the detection of a smooth shift using various test statistics. Left: Diagonal covariance structure. Right: Block-diagonal covariance structure.

The first important comparison is between the classical Hotelling T^2-test versus the T^2-test in the graph-Fourier space. As expected from Lemma 1, testing in the restricted space where the shift lies performs much better than testing in the full space which includes irrelevant coefficients. The difference can be made arbitrarily large by increasing the dimension p and keeping the shift unchanged. The graph-structured test retains a large advantage even for moderately smooth shifts, *e.g.*, when $k_0 = 3$ and $p = 5$. Of course, this corresponds to the optimistic case where the number of shifted coefficients k_0 is known. Figure 4 shows the power of the test in the graph-Fourier space for various choices of k. Even when missing some coefficients ($k < k_0$) or adding a few non-relevant ones ($k > k_0$), the power of the graph-structured test is higher than that of the T^2-test in the full space. The principal component approach is shown because it was proposed for the application which motivated our work [Ma and Kosorok, 2009] and as it illustrates that the performance improvement originates not only from dimensionality reduction, but also from the fact that this reduction is in a direction that does not decrease the shift. We emphasize that power entirely depends on the nature of the shift

and that a PC-based test would outperform our Fourier-based test when the shift lies in the first principal components rather than Fourier coefficients. The statistics of Bai and Saranadasa [1996] and Chen and Qin [2010] are also largely outperformed by our graph-structured statistic (ROC curves not shown in Figure 3 for the sake of readability), which illustrates that working in the graph-Fourier space solves the problem of high-dimensionality for which these statistics were designed. Here again, for a non-smooth shift, the comparison would be less favorable. Finally, we consider the adaptive Neyman test of Fan and Lin [1998], which takes advantage of smoothness assumptions for time-series. This test differs from our graph-structured test, as Fourier coefficients for stationary time-series are known to be asymptotically independent and Gaussian. For graphs, the asymptotics would be in the number of nodes, which is typically small, and necessary conditions such as stationarity are more difficult to define and unlikely to hold for data like gene expression measurements. In the uncorrelated setting, the modified version of the Fan and Lin [1998] statistic based the true number of non-zero coefficients performs approximately as well as the graph-structured T^2. However, for correlated data, it loses power and both versions can have arbitrarily degraded performance. This, together with the need to use the bootstrap to calibrate this test, illustrates that direct transposition of the Fan and Lin [1998] test to the graph context is not optimal.

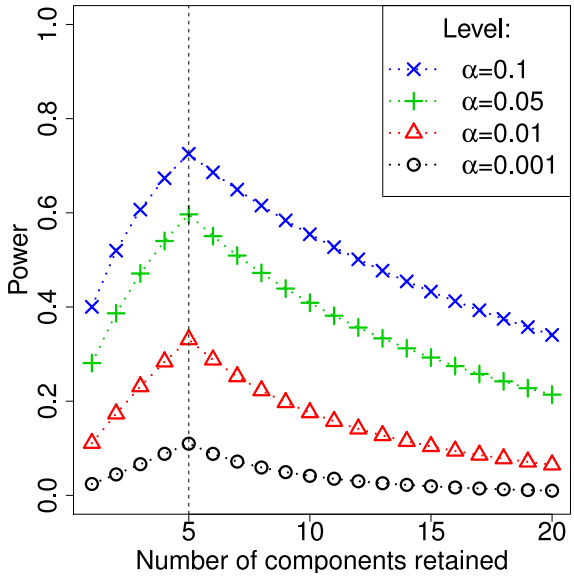

Figure 4: Power of the T^2-test in the graph-Fourier space with an actual mean shift evenly distributed among the first $k_0 = 5$ coefficients.

To evaluate the performance of the subgraph discovery algorithms proposed in Section 4, we generated a graph of 100 nodes formed by tightly-connected hubs of sizes sampled from a Poisson distribution with parameter 10 and only weak connections between these hubs (Figure 5). Such a graph structure mimics the typical topology of gene regulation networks. We randomly selected one subgraph of 5 nodes to be non-homogeneous, with smooth shift in the first $k_0 = 3$ Fourier coefficients. The mean shift was set to zero on the rest of the graph. We set the

norm of the mean shift to 1 and the covariance matrix to identity, so that detecting the shifted subgraph is impossible by just looking at the mean shift on the graph.

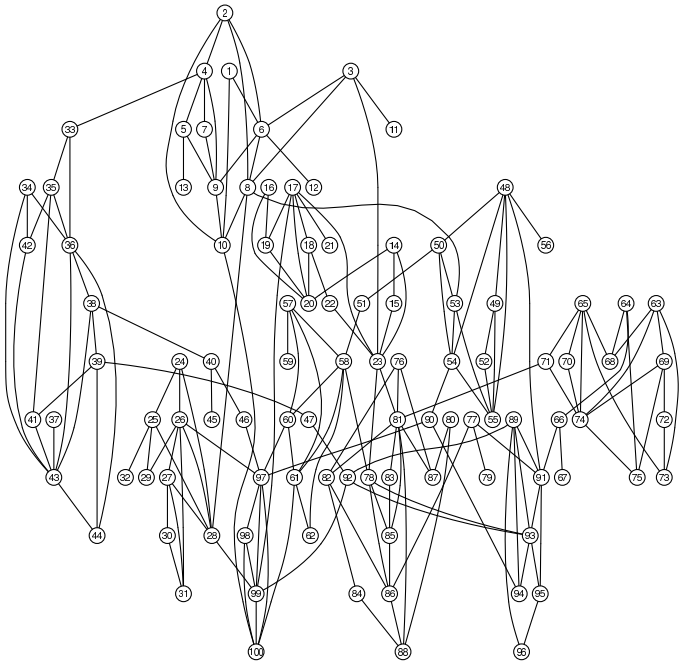

Figure 5: Random graph used in the evaluation of the pruning procedure.

We evaluated run-time for full enumeration, the exact branch-and-bound algorithm based on Lemma 2 (Section 4.1), and the approximate algorithm based on the Euclidean norm (Section 4.2). We also examined run-time on data with permuted class labels, as the subgraph discovery procedure is to be run on such data to evaluate the number of false positives and adjust for multiple testing. Averaging over 20 runs, the full enumeration procedure took 732 ± 9 seconds per run and the exact branch-and-bound 627 ± 59 seconds on the non-permuted data and 578 ± 100 seconds on permuted data. Over 100 runs, the approximation at $\theta = 0.5$ ($\lambda_{min} = 0.52$) took 204 ± 86 seconds (129 ± 91 on permuted data) and the approximation at $\theta = 1$ ($\lambda_{min} = 1.04$) took 183 ± 106 seconds (40 ± 60 on permuted data). The latter approximation missed the non-homogeneous subgraph in 5% of the runs.

While neither the exact nor the approximate bounds are efficient enough to allow systematic testing on huge graphs for which the exact approach would be impossible, they allow a significant gain in speed, especially for permuted data, and will thus prove to be very useful for multiple testing adjustment.

5.2 Breast cancer expression data

We also validated our methods using the microarray dataset of Loi et al. [2008], which comprises expression measures for 15,737 genes in 255 patients treated with tamoxifen. Using distant metastasis free survival as a primary endpoint, 68 patients are labeled as resistant to tamoxifen and 187 are labeled as sensitive to tamoxifen.

Our goal was to detect structured groups of genes which are differentially expressed between resistant and sensitive patients.

We first tested individually 323 connected components from 89 KEGG pathways corresponding to known gene regulation networks, using the classical Hotelling T^2-test and the T^2-test in the graph-Fourier space retaining only the first 20% Fourier coefficients ($k = 0.2p$). For each of the 323 graphs, (unadjusted) p-values were computed under the nominal F-distributions $F_0(p, n_1+n_2-p-1)$ and $F_0(k, n_1+n_2-k-1)$, respectively. Figure 6 shows the pathway for which the ratio of graph-Fourier to full space p-values is the lowest (*i.e.*, most significant for graph-structured test relative to classical test) and the pathway for which it is the highest. As expected, the former corresponds to a shift which appears to be coherent with the network (even on edges corresponding to inhibition), while the latter is a small network with non-smooth shift. More generally, the classical approach tends to select very small networks. The coherent pathway selected by our graph-structured test corresponds to *Leukocyte transendothelial migration*. To the best of our knowledge, this pathway is not specifically known to be involved in tamoxifen resistance. However, its role in resistance is plausible, as leukocyte infiltration was recently found to be involved in breast tumor invasion [Man, 2010]; more generally, the immune system and inflammatory response are closely-related to the evolution of cancer.

We then ran our branch-and-bound non-homogeneous subgraph discovery procedure on the cell cycle pathway, which, after restriction to edges of known sign (inhibition or activation), has 86 nodes and 442 edges. Specifically, we sought to detect differentially expressed subgraphs of size $q = 5$, after pre-selecting those for which the squared Euclidean norm of the empirical shift exceeded $\theta = 0.1$; for a test in the first $k = 3$ Fourier components at level $\alpha = 10^{-4}$, this corresponded to $\lambda_{min} < 0.23$ and to an expected removal of 95% of the subgraphs under the approximation that the squared Euclidean norm of the subgraphs follows a χ^2_5-distribution.

For $\alpha = 10^{-4}$, none of the 50 runs on permuted data gave any positive subgraph and 31 overlapping subgraphs (Figure 7) were detected on the original data, corresponding to a connected subnetwork of 22 genes. Some of these genes have large individual differential expression, namely TP53 whose mutation has been long-known to be involved in tamoxifen resistance [Andersson et al., 2005, Fernandez-Cuesta et al., 2010]. E2F1, whose expression level was recently shown to be involved in tamoxifen resistance [Louie et al., 2010], is also part of the identified network, as well as CCND1 [Barnes, 1997, Musgrove and Sutherland, 2009]. Some other genes in the network have quite low t-statistics and would not have been detected individually. This is the case of CCNE1 and CDK2, which were also described in [Louie et al., 2010] as part of the same mechanism as E2F1. Similarly, CDKN1A was recently found to be involved in anti-œstrogene treatment resistance [Musgrove and Sutherland, 2009] and in ovarian cancer which is also a hormone-dependent cancer [Cunningham et al., 2009]. The networks also contains RB1, a tumor suppressor whose expression or loss is known to be correlated to tamoxifen resistance [Musgrove and Sutherland, 2009]. RB1 is inhibited by CDK4, whose inhibition has been described in Sutherland and Musgrove [2009] as acting synergistically with tamoxifen or trastuzumab. More generally, a large part of the network displayed on Figure 2A of Musgrove and Sutherland [2009] is included in our network, along

with other known actors of tamoxifen resistance. Our system-based approach to pathway discovery therefore directly identifies a set of interacting important genes and may therefore prove to be more efficient than iterative individual identification of single actors.

6 Discussion

We developed a graph-structured two-sample test of means, for problems in which the distribution shift is assumed to be smooth on a given graph. We proved quantitative results on power gains for such smooth-shift alternatives and devised branch-and-bound algorithms to systematically apply our test to all the subgraphs of a large graph. The first algorithm is exact and reduces the number of explicitly tested subgraphs. The second is approximate, with no false positives and a quantitative result on the type of false negatives (with respect to the exact algorithm). The non-homogeneous subgraph discovery method involves performing a larger number of tests, with highly-dependent test statistics. However, as the actual number of tested hypotheses is unknown, standard multiple testing procedures are not directly applicable. Instead, we use a permutation procedure to estimate the distribution of the number of false positive subgraphs. Such resampling procedures (bootstrap or permutation) are feasible due to the manageable run-time of the pruning algorithms of Section 4. Results on synthetic data illustrate the good power properties of our graph-structured test under smooth-shift alternatives, as well as the good performance of our branch-and-bound-like algorithms for subgraph discovery. Very promising results are also obtained on the drug resistance microarray dataset of Loi et al. [2008].

Future work should investigate the use of other bases, such as graph-wavelets [Hammond et al., 2009], which would allow the detection of shifts with spatially-located non-smoothness, for example, to take into account errors in existing networks. More systematic procedures for cutoff selection should also be considered, *e.g.*, two-step method proposed in Das Gupta and Perlman [1974] or adaptive approaches as in Fan and Lin [1998]. The pruning algorithm would naturally benefit from sharper bounds. Such bounds could be obtained by controlling the condition number of all covariance matrices, using, for example, regularized statistics which still have known non-asymptotic distributions, such as those of Tai and Speed [2008]. Concerning multiple testing, procedures should be devised to exploit the dependence structure between the tested subgraphs and to deal with the unknown number of tests. The proposed approach could also be enriched to take into account different types of data, *e.g.*, copy number for the detection of DE gene pathways. More subtle notions of smoothness, *e.g.*, "and" and "or" logical relations [Vaske et al., 2010], could also be included. An interesting alternative application would be to explore the list of pathways which are known to be differentially expressed (or detected by the classical T^2-test), but which are not detected by the graph-Fourier approach, to infer possible mis-annotation in the network. Other applications of two-sample tests with smooth-shift on a graph include fMRI and eQTL association studies.

Finally, it would be of interest to compare our testing approach with struc-

tured sparse learning, for the purpose of identifying expression signatures that are predictive of drug resistance. Methods should be compared in terms of prediction accuracy and stability of the selected genes across different datasets, a central and difficult problem in the design of such signatures [Ein-Dor et al., 2005, He and Yu, 2010, Haury et al., 2010]. The comparison should also take into account the merits of the sparsity-inducing norm over the hypothesis testing-based selection, as well as the influence of the smoothness assumption. The latter could indeed also be integrated in a sparsity-inducing penalty by applying, *e.g.*, Jacob et al. [2009] to the reduced graph-Fourier representation of the pathways, yielding a special case of multiple kernel learning [Bach et al., 2004].

Acknowledgments

The authors thank Zaïd Harchaoui, Nourredine El Karoui, and Terry Speed for very helpful discussions and suggestions, and the UC Berkeley Center for Computational Biology Genentech Innovation Fellowship and The Cancer Genome Atlas Project for funding.

References

J. Andersson, L. Larsson, S. Klaar, L. Holmberg, J. Nilsson, M. Ingans, G. Carlsson, J. Ohd, C-M. Rudenstam, B. Gustavsson, and J. Bergh. Worse survival for tp53 (p53)-mutated breast cancer patients receiving adjuvant cmf. *Ann Oncol*, 16(5): 743–748, May 2005. doi: 10.1093/annonc/mdi150. URL `http://dx.doi.org/10.1093/annonc/mdi150`.

F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 6, New York, NY, USA, 2004. ACM. doi: http://doi.acm.org/10.1145/1015330.1015424.

Zhidong Bai and Hewa Saranadasa. Effect of high dimension : by an example of a two sample problem. *Statistica Sinica*, 6:311,329, 1996.

D. M. Barnes. Cyclin d1 in mammary carcinoma. *J Pathol*, 181(3):267–269, Mar 1997. doi: 3.0.CO;2-X. URL `http://dx.doi.org/3.0.CO;2-X`.

Tim Beissbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, Jun 2004. doi: 10.1093/bioinformatics/bth088. URL `http://dx.doi.org/10.1093/bioinformatics/bth088`.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.*, 38(arXiv:1002.4547. IMS-AOS-AOS716):808–835, Feb 2010.

F. R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series*. American Mathematical Society, Providence, 1997.

J. M. Cunningham, R. A. Vierkant, T. A. Sellers, C. Phelan, D. N. Rider, M. Liebow, J. Schildkraut, A. Berchuck, F. J. Couch, X. Wang, B. L. Fridley, Ovarian Cancer Association Consortium, A. Gentry-Maharaj, U. Menon, E. Hogdall, S. Kjaer, A. Whittemore, R. DiCioccio, H. Song, S. A. Gayther, S. J. Ramus, P. D P Pharaoh, and E. L. Goode. Cell cycle genes and ovarian cancer susceptibility: a tagsnp analysis. *Br J Cancer*, 101(8):1461–1468, Oct 2009. doi: 10.1038/sj.bjc.6605284. URL `http://dx.doi.org/10.1038/sj.bjc.6605284`.

Somesh Das Gupta and Michael D. Perlman. Power of the noncentral F test : effect of additional variates on hotelling's t^2-test. *Journal of the American Statistical Association*, 69(345):174–180, Mar 1974.

S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer, New York, 2008.

Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2): 171–178, 2005.

Jianqing Fan and Sheng-kuei Lin. Test of significance when data are curves. *J. Am. Statist. Assoc*, 93:1007–1021, 1998.

Lynnette Fernandez-Cuesta, Suresh Anaganti, Pierre Hainaut, and Magali Olivier. p53 status influences response to tamoxifen but not to fulvestrant in breast cancer cell lines. *Int J Cancer*, Jun 2010. doi: 10.1002/ijc.25512. URL `http://dx.doi.org/10.1002/ijc.25512`.

J J Goeman and P Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, April 2007. doi: 10.1093/bioinformatics/btm051. URL `http://www.ncbi.nlm.nih.gov/pubmed/17303618`.

Andrew B. Goldberg. Dissimilarity in graph-based semisupervised classification. In *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schlkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *CoRR*, abs/0912.3848, 2009. URL `http://dblp.uni-trier.de/db/journals/corr/corr0912.html#abs-0912-3848`. informal publication.

Zaïd Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. MIT Press, 2007.

A.-C. Haury, L. Jacob, and J.-P. Vert. Increasing stability and interpretability of gene expression signatures. *ArXiv e-prints*, January 2010.

Zengyou He and Weichuan Yu. Stable feature selection for biomarker discovery. *CoRR*, abs/1001.0887, 2010.

Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. In *ISMB*, pages 233–240, 2002.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: http://doi.acm.org/10.1145/1553374.1553431.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. Research report, WILLOW - INRIA, 2009. URL `http://hal.inria.fr/inria-00377732/en/`.

E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

S. Loi, B. Haibe-Kains, C. Desmedt, P. Wirapati, F. Lallemand, A.M. Tutt, C. Gillet, P. Ellis, K. Ryder, J.F. Reid, et al. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9(1):239, 2008.

Ingrid Lönnstedt and Terry Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2001.

Maggie C Louie, Ashley McClellan, Christina Siewit, and Lauren Kawabata. Estrogen receptor regulates e2f1 expression to mediate tamoxifen resistance. *Mol Cancer Res*, 8(3):343–352, Mar 2010. doi: 10.1158/1541-7786.MCR-09-0395. URL `http://dx.doi.org/10.1158/1541-7786.MCR-09-0395`.

Yan Lu, Peng-Yuan Liu, Peng Xiao, and Hong-Wen Deng. Hotelling's t2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14):3105–3113, Jul 2005. doi: 10.1093/bioinformatics/bti496. URL `http://dx.doi.org/10.1093/bioinformatics/bti496`.

Shuangge Ma and Michael R. Kosorok. Identification of differential gene pathways with principal component analysis. *Bioinformatics*, 25(7):882–889, 2009. ISSN 1367-4803. doi: http://dx.doi.org/10.1093/bioinformatics/btp085.

Yan-Gao Man. Aberrant leukocyte infiltration: a direct trigger for breast tumor invasion and metastasis. *Int J Biol Sci*, 6(2):129–132, 2010.

Elizabeth A Musgrove and Robert L Sutherland. Biological determinants of endocrine resistance in breast cancer. *Nat Rev Cancer*, 9(9):631–643, Sep 2009. doi: 10.1038/nrc2713. URL `http://dx.doi.org/10.1038/nrc2713`.

S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7):850, 2007.

F. Rapaport, A. Zynoviev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007.

Elias M. Stain and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, 1971.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43): 15545–15550, Oct 2005. doi: 10.1073/pnas.0506580102. URL `http://dx.doi.org/10.1073/pnas.0506580102`.

Robert L Sutherland and Elizabeth A Musgrove. Cdk inhibitors as potential breast cancer therapeutics: new evidence for enhanced efficacy in er+ disease. *Breast Cancer Res*, 11(6):112, 2009. doi: 10.1186/bcr2454. URL `http://dx.doi.org/10.1186/bcr2454`.

Yu Chan Tai and Terry Speed. On gene ranking using replicated microarray time course data. *Biometric*, 65(1):40–51, June 2008.

Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Algorithms for detecting significantly mutated pathways in cancer. In Bonnie Berger, editor, *RECOMB*, volume 6044 of *Lecture Notes in Computer Science*, pages 506–521. Springer, 2010. ISBN 978-3-642-12682-6.

Charles Vaske, Stephen Benz, Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. In *ISMB*, 2010.

Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In Walter G. Kropatsch, Robert Sablatnig, and Allan Hanbury, editors, *DAGM-Symposium*, volume 3663 of *Lecture Notes in Computer Science*, pages 361–368. Springer, 2005. ISBN 3-540-28703-5.

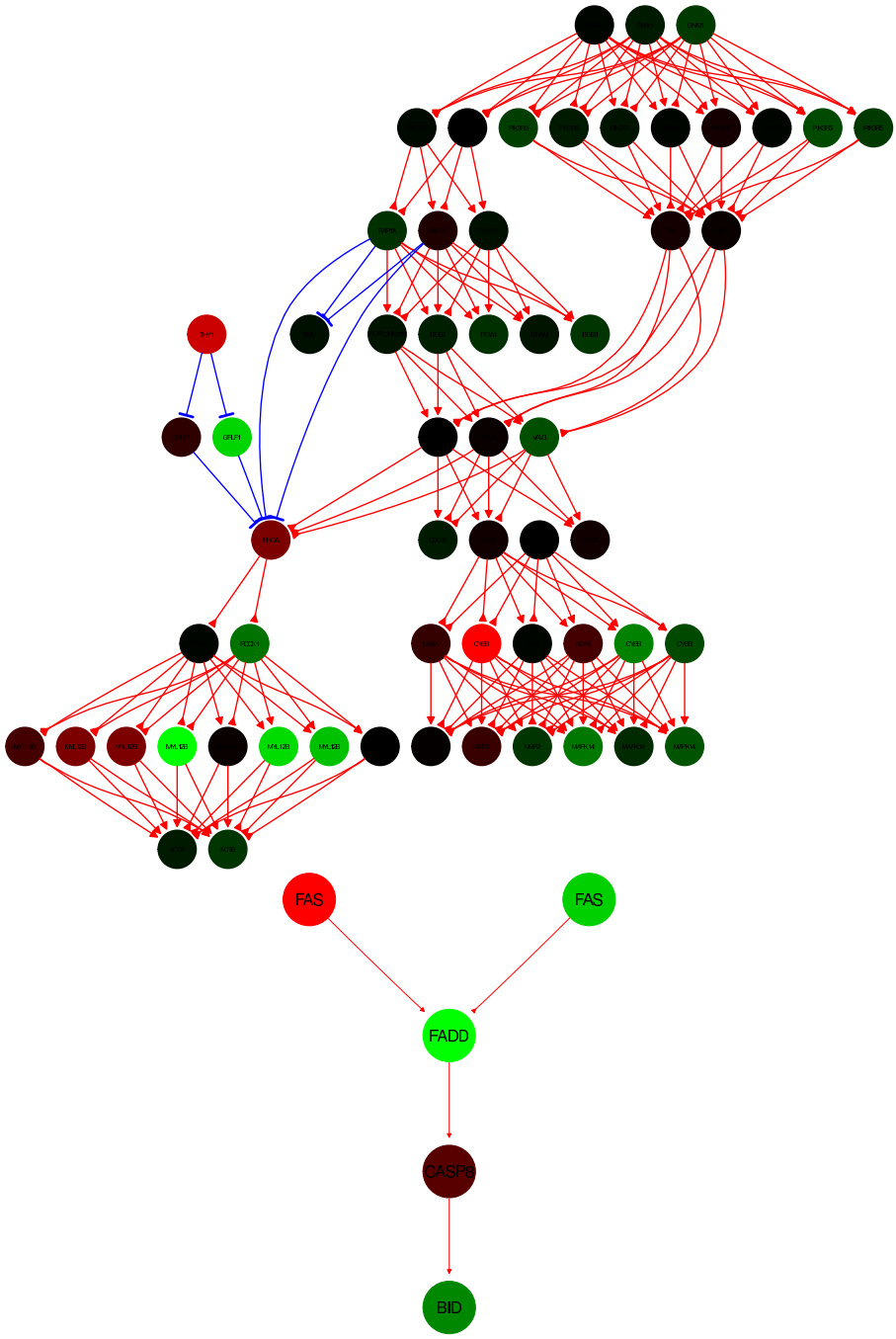

Figure 6: Difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in two KEGG regulation networks. Top: Regulation network (Leukocyte transendothelial migration) with the lowest ratio of graph-Fourier to full space p-values. Bottom: Regulation network (Alzheimer's disease) with the highest ratio of graph-Fourier to full space p-values. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activations, blue arrows inhibition.

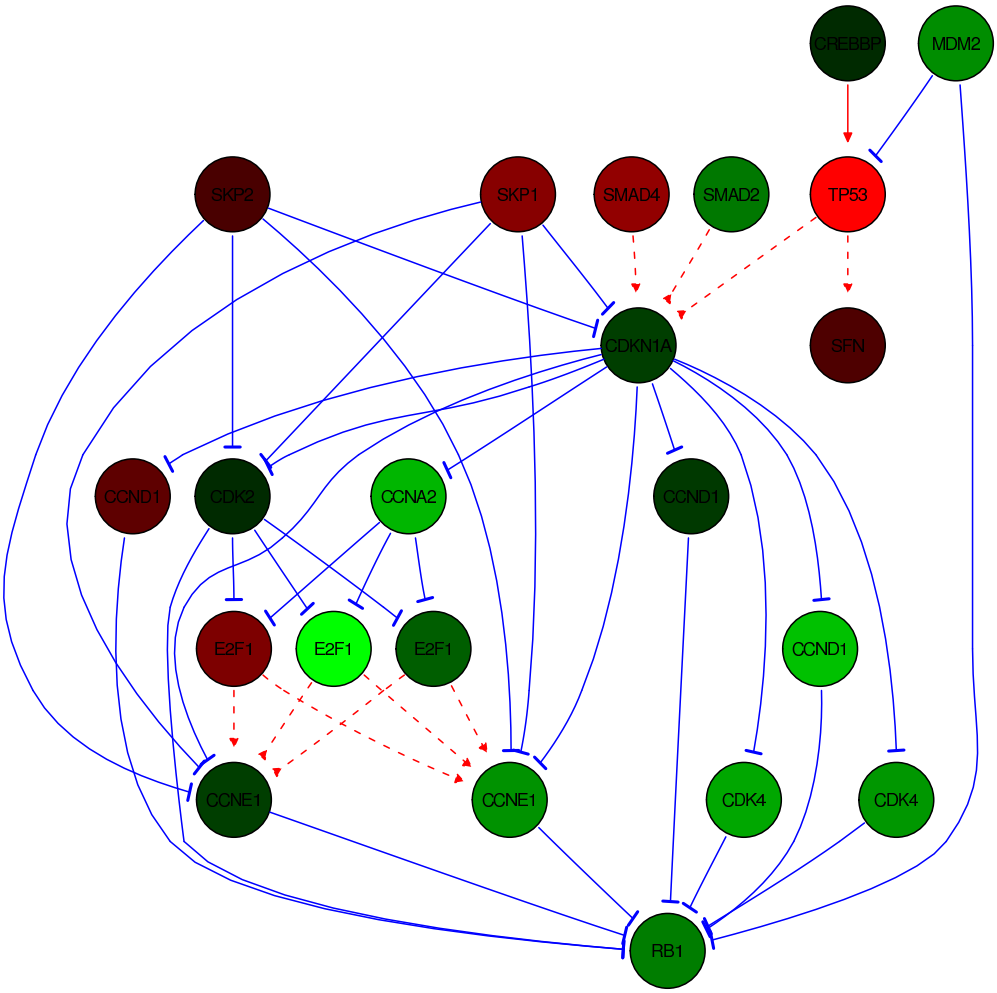

Figure 7: Difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in the two overlapping subgraphs detected at $\alpha = 10^{-4}$. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activations, blue arrows inhibition.