University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 271

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

Laurent Jacob* Pierre Neuvial†

Sandrine Dudoit‡

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

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

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

http://biostats.bepress.com/ucbbiostat/paper271

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

Laurent Jacob, Pierre Neuvial, and Sandrine Dudoit

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.

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

October 29, 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 functional Magnetic Resonance Imaging (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 covariates $\{X_1^1, \ldots, X_{n_1}^1\} \in \mathbb{R}^p$ 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 a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encodes some type of prior information on the expected relationship between the p variables, the putative *location* or *mean shift* $\delta = \mu_1 - \mu_2$ may be expected to be partly "coherent" with $\mathcal{G}$, *e.g.*, each node has a shift which is similar to the shift of the nodes pointing to it. 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 alternative hypotheses where the distribution shift is coherent with the graph.

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 avoiding one-by-one enumeration and testing of these subgraphs and 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 explains how to build a lower-dimension basis in which to apply the multivariate test of means. 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 (without fully enumerating) all the subgraphs of a large graph. Section 5 presents results for synthetic data as well as a breast cancer gene expression dataset analyzed in the light of pathways from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database. Finally, Section 7 summarizes our findings and outlines ongoing work.

2 Graph-based dimensionality reduction

As stated in the introduction, each of the two main paradigms for testing differential expression of a gene set have their limitations. Two-step methods generally do not directly test the existence of a mean shift between two multivariate distributions [Goeman and Bühlmann, 2007]. The second step, which often treats the *genes* as the sampling units, renders the interpretation of p-values problematic and may lead to a large loss of power or Type I error control when sets of genes have correlated expression.

Multivariate statistics, on the other hand, allow a direct formulation of and solution to the testing question: the sampling units are vectors of gene expression measures (*e.g.*, corresponding to patients) and the question is whether two such sets of random vectors are likely to have arisen from distributions with equal means. Figure 1 illustrates another classical advantage of multivariate approaches: genes taken individually may have extremely small mean shifts between two populations, although their joint distributions clearly differ between the two populations. Here, again, this phenomenon typically happens for sets of genes whose expression measures are correlated, which is not unlikely for pathways or annotated gene sets.

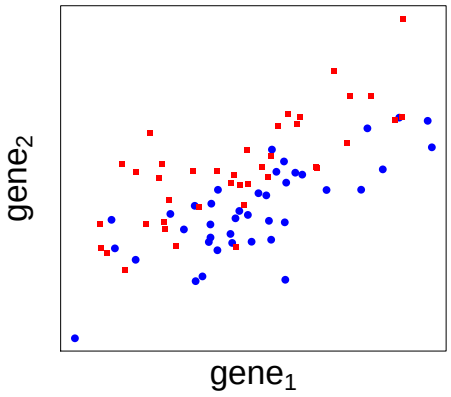

Figure 1: Synthetic example of the joint distribution of the expression measures of two genes in two patient populations. The color and shape of the plotting symbols indicate the patient group and the x- and y-axes correspond to the expression measures of the first and second gene, respectively.

Unfortunately, with moderate sample sizes, multivariate statistics lose power quickly in high dimension. If some type of side information is available regarding particular properties of the expression shift, a possible approach to get the best of both worlds would be to: (1) project the vectors of covariates in a new space of *lower dimension* that preserves the distribution shift, *i.e.*, the distance between the expression measures of the two groups, and (2) apply the multivariate statistic in this new space. One could thus perform the appropriate multivariate test, while avoiding the loss of power caused by the high dimensionality of the original covariate space.

A possible source of information about the expression shift is the growing number of available gene networks. Indeed, while the difference in mean expression between two groups of patients may not be entirely coherent with an existing network (*e.g.*, because of noise in the data, errors in the annotation, or inappropriateness of the chosen network for the biological question of interest), it is reasonable to expect that this shift will not be entirely contradictory with the given graph structure. For example, repressed genes

should be more connected to other repressed genes than to over-expressed genes. Given this assumption, we intend to build a space of lower dimension than the original gene space, but which preserves most of the distribution shift between the two populations.

More precisely, consider a network of p genes, represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = p$ nodes and edge set $\mathcal{E}$. Let $\delta \in \mathbb{R}^p$ denote the mean shift, *i.e.*, the vector of differences in mean expression measures for these p genes between the two populations of interest. Suppose we expect the shift δ to be coherent with the graph $\mathcal{G}$, in the sense that it has low energy $E_{\mathcal{G}}(\delta)$ for a particular energy function $E_{\mathcal{G}}$ defined on $\mathcal{G}$. Then, we wish to build a space of lower dimension $k \ll p$ capturing most of the low energy functions. To this end, we start by finding the function that has the lowest possible energy, then the function that has lowest possible energy in the orthogonal space of the first one, up to the kth function with lowest energy in the orthogonal subspace of the first $k-1$ functions. That is, for each $i \leq k$, we define

$$u_i = \begin{cases} \underset{f \in \mathbb{R}^p}{\arg\min}\, E_{\mathcal{G}}(f) \\ \text{such that } u_i \perp u_j,\ j < i. \end{cases} \tag{1}$$

If $E_{\mathcal{G}}$ is a positive semi-definite quadratic form $E_{\mathcal{G}}(\delta) = \delta^\top Q_{\mathcal{G}} \delta$, for some positive semi-definite matrix $Q_{\mathcal{G}} = U \Lambda U^\top$, where U is an orthogonal matrix and Λ a diagonal matrix with elements λ_i, $i = 1, \ldots, p$, then the solution to Equation (1) is given by the k eigenvectors of $Q_{\mathcal{G}}$ corresponding to the smallest k eigenvalues. It is easy to check that these eigenvalues are the energies of the corresponding functions u_i, *i.e.*, $E_{\mathcal{G}}(u_i) = \lambda_i$.

Different choices of $Q_{\mathcal{G}}$ lead to different notions of coherence of the expression shift with the network. A classical choice is the *graph Laplacian* $\mathcal{L}$. Suppose $\mathcal{G}$ is an undirected graph with adjacency matrix A, with $a_{ij} = 1$ if and only if $(i,j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise, 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 $D_{ii} = d_i$. The Laplacian matrix of $\mathcal{G}$ is then 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, leading to energies $\sum_{i,j \in \mathcal{V}} (\delta_i - \delta_j)^2$ and $\sum_{i,j \in \mathcal{V}} \left(\frac{\delta_i}{\sqrt{d_i}} - \frac{\delta_j}{\sqrt{d_j}} \right)^2$, respectively. Note that, in this case, the Laplacian matrix $\mathcal{L}$, energy E, and basis functions u_i extend the classical Fourier analysis of functions on Euclidean spaces to functions on graphs, by transferring the notions of Laplace operator, Dirichlet energy, and Fourier basis, respectively.

More generally, any positive semi-definite matrix can be chosen. In the case of gene regulation networks, we do not necessarily expect as strong a coherence as that corresponding to the Dirichlet energy defined by the graph Laplacian, since some of the annotated interactions may not be relevant in the studied context and some antagonist interactions may cancel each other. For example, if a gene is activated by two others, one who is under-expressed and the other over-expressed, we may observe no change in the expression of the gene, but a non-zero Dirichlet energy $\sum_{i,j \in \mathcal{V}} (\delta_i - \delta_j)^2$.

Additionally, for applications like structured gene set differential expression detection, one may use negative weights for edges that reflect a 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 $|\delta_i + \delta_j|$. This can be achieved in the same formalism by simply considering a signed version of the adjacency matrix A, *i.e.*, $a_{ij} = 1$ if gene i activates gene j and -1 if it inhibits gene j. A signed version of the graph Laplacian is then $\mathcal{L}_{\text{sign}} = D - A$, where D is still the degree matrix, *i.e.* $D = \text{Diag}\,(|A|\mathbf{1})$, $|A|$ denoting the entry-wise absolute value of A. Note that such a signed Laplacian was used as a penalty for semi-supervised learning in Goldberg [2007].

In the context of this work, we moreover consider directed graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the edge set $\mathcal{E}$ consists of ordered pairs of nodes. The adjacency matrix A is asymmetric, with entries $a_{ij} \neq 0$ if and only if $(i, j) \in \mathcal{E}$, *i.e.*, there is an edge going from node v_i to node v_j. We then use the following energy function:

$$E_{\mathcal{G}}(\delta) = \sum_{i:d_i^- \neq 0}^{p} \left(\delta_i - \frac{1}{d_i^-} \sum_{(j,i)\in\mathcal{E}} a_{ji}\delta_j \right)^2 , \tag{2}$$

where $d_i^- \triangleq \sum_{j=1}^{p} |a_{ji}|$ is the indegree of node v_i, *i.e.*, the number of directed edges that connect any node to v_i. According to this definition, an expression shift will have low energy if the difference in mean expression of any given gene between the two populations is similar to the (signed) average of the differences in mean expression for the genes that either activate or inhibit it.

It is immediate to check that $E_{\mathcal{G}}(\delta) = \delta^\top M_{\mathcal{G}} \delta$, with $M_{\mathcal{G}} \triangleq (\tilde{I} - D_-^{-1}A^\top)^\top(\tilde{I} - D_-^{-1}A^\top)$, where $D_- \triangleq \text{Diag}\left((d_i^-)_{i=1,\ldots,p}\right)$ is the matrix of indegrees, $\tilde{I} \triangleq \text{Diag}\left((\text{I}(d_i^- \neq 0))_{i=1,\ldots,p}\right)$ is a modification of the identity matrix where diagonal elements corresponding to nodes with zero indegree are set to zero, and the value of the indicator function I is 1 if its argument is true and zero otherwise. Note that a very similar function was used in the context of regularized supervised learning by Sandler et al. [2009].

Following our principle to build a lower dimension space, we should use the first few eigenvectors of $M_{\mathcal{G}}$ to obtain orthonormal functions with low energy. As an example, Figure 2 displays the eigenvectors of $M_{\mathcal{G}}$ for a simple four-node graph with

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

where A takes negative values for negative interactions, such as expression inhibition. The first eigenvector, corresponding to the smallest energy (eigenvalue of zero), can be viewed as a "constant" function on the graph, in the sense that its absolute value is identical for all nodes, but nodes connected by an edge with negative weight take on

values of opposite sign. By contrast, the last eigenvector, corresponding to the highest energy, 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. Note that, for this particular example, the adjacency matrix is symmetric, which needs not always be the case. Actually, here, the signed Laplacian would have the same eigen-decomposition (this is not the case for all undirected graphs). For a slightly different graph:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

which is the same as above but with directed edges and only positive interactions to avoid confusion, Figure 3 shows that the two notions of energy lead to two different bases. While the signed Laplacian matrix has only one (constant) eigenvector of null energy, two of energy 1, and one of 4, $M_{\mathcal{G}}$ has three orthogonal vectors of null energy. Note, however, that the first and last eigenvectors are still the same across the two bases.

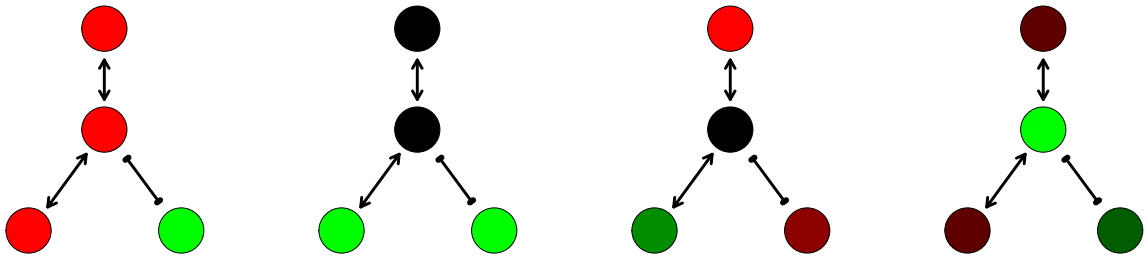

Figure 2: Eigenvectors of $M_{\mathcal{G}}$ and of the signed Laplacian $\mathcal{L}_{\text{sign}}$ for a simple undirected four-node graph. The corresponding eigenvalues are $0, 1, 1, \frac{16}{3}$ and $0, 1, 1, 4$. Nodes are colored according to the value of the eigenvector, where green corresponds to high positive values, red to high negative values, and black to 0. "T"-shaped edges have negative weights.

While we introduce this idea in the context of gene regulation networks and testing for differential expression, the same dimensionality reduction principle applies to any multivariate testing problem for which the variables have a known structure, as represented by a graph.

In the remainder of this paper, we denote by $\tilde{f} = U^\top f$ the coefficients of a vector $f \in \mathbb{R}^{|\mathcal{V}|}$ after projection on a basis U (typically the eigenvectors of a $Q_{\mathcal{G}}$ matrix).

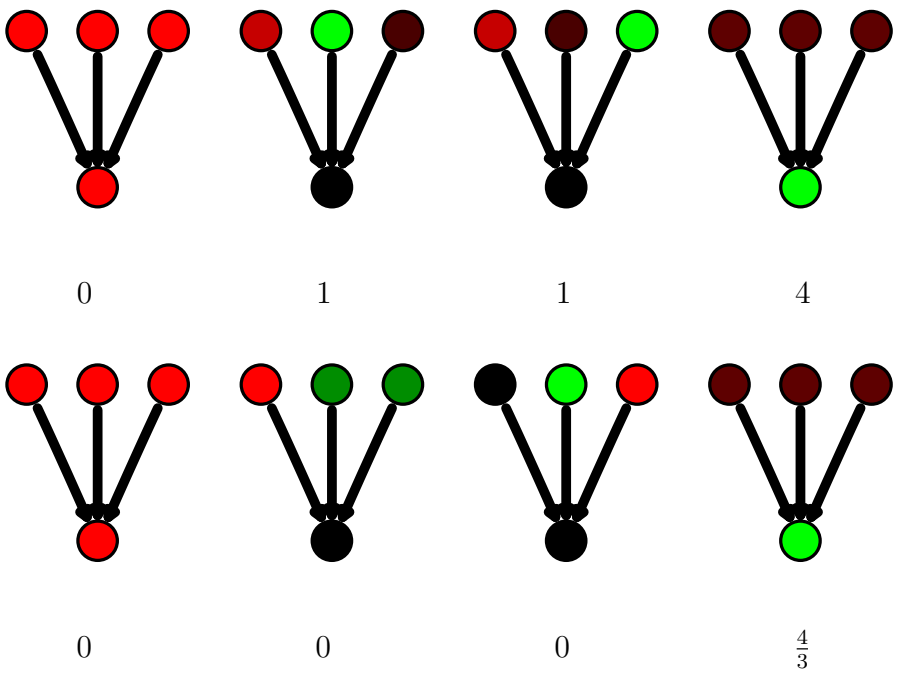

Figure 3: Eigenvectors of the signed Laplacian $\mathcal{L}_{\text{sign}}$ (top) and of $M_{\mathcal{G}}$ (bottom) for a simple directed four-node graph. The corresponding eigenvalues are $0, 1, 1, 4$ for the Laplacian and $0, 0, 0, \frac{4}{3}$ for $M_{\mathcal{G}}$. Nodes are colored according to the value of the eigenvector, where green corresponds to high positive values, red to high negative values, and black to 0.

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 orthonormal basis U and, in particular, for our graph-based basis, 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 new space are identical. More generally, for $k \leq p$, the statistic in the original space after filtering out dimensions above k is the same as the statistic $\tilde{T}_k^2$ restricted to the first k coefficients in the new space defined by U:

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

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 dimensions 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 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 new 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 in the first few graph-based coefficients, 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 new space restricted to its first k components yields more power than testing in the entire new 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 new components is a *non-invertible* transformation.

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

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

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 new 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. □

Note that the increase in shift $\eta(\alpha, k, l)$ required to maintain power when increasing dimension can be evaluated numerically for any (α, k, l). 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 new 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 new components. Although non-necessary, this condition is plausible when the mean shift lies at the beginning of the spectrum (i.e., has low energy), 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 more general, as testing in the first k new 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 4 illustrates, under different settings, the increase in distribution shift necessary to maintain a given power level against the number of added coefficients.

Under the same block-diagonal covariance assumption, we have the following second corollary which directly relates the energy of the mean shift vector to the gain in power :

Corollary 2. *Consider any positive semi-definite matrix* $Q_{\mathcal{G}} = U\Lambda U^\top$*, with corresponding energy function* $E_{\mathcal{G}}(f) = f^\top Q_{\mathcal{G}} f$*, and mean shift* $\delta = \mu_1 - \mu_2$*, with projection* $\tilde{\delta} = U^\top \delta$ *in the eigenvector basis of* $Q_{\mathcal{G}}$*. Then, at any level* α *and for any* $k < p$*, there exists* $c(\alpha, k) > 0$ *such that*

$$E_{\mathcal{G}}(\delta) \leq c(\alpha, k)\lambda_k s_0 + \sum_{i=1}^{k-1} \lambda_i \tilde{\delta}_i^2 \Rightarrow \beta_{\alpha,k}(\Delta_k^2(\tilde{\delta}, \tilde{\Sigma})) > \beta_{\alpha,p}(\Delta_p^2(\tilde{\delta}, \tilde{\Sigma})),$$

where s_0 *denotes the smallest eigenvalue of the last* $(p-k)$*-block of* $\tilde{\Sigma}$*.*

Proof. From the block-diagonality assumption, we have that

$$\Delta_p^2 - \Delta_k^2 \leq \frac{\sum_{i=k}^p \tilde{\delta}_i^2}{s_0} \leq \frac{\sum_{i=k}^p \lambda_i \tilde{\delta}_i^2}{\lambda_k s_0} = \frac{E_{\mathcal{G}}(\delta) - \sum_{i=1}^{k-1} \lambda_i \tilde{\delta}_i^2}{\lambda_k s_0},$$

so the result directly follows from Lemma 1, with $c(\alpha, k) = d(\alpha, k, p)$. □

Corollary 2 states that if the energy of the mean shift vector δ is small enough, *i.e.*, if the mean shift is coherent enough with the network, then testing in the first k dimensions of the new basis is more powerful than testing in the original space. The corresponding upper bound on the mean shift energy can be quantified for a given generative setting (μ_1, μ_2, Σ), graph $\mathcal{G}$, and level α. Tighter and looser bounds can be straightforwardly derived using the same principle for the diagonal and general covariance cases, respectively.

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 with higher energies 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 of its principal submatrices.

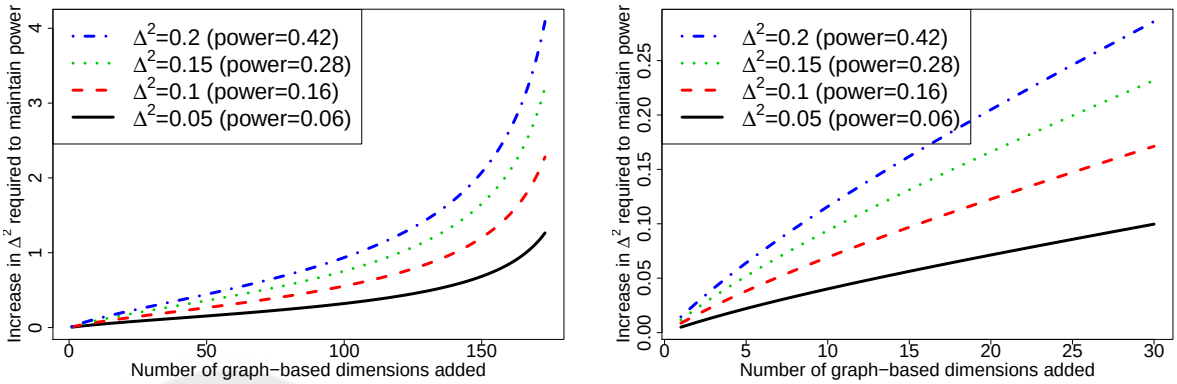

Figure 4: Left: Increase in distribution shift required for Hotelling's T^2-test to maintain a given power when increasing the number of tested new coefficients: $\Delta^2_{k+l} - \Delta^2_k$ vs. l such that $\beta_{\alpha,k+l}(\Delta^2_{k+l}) = \beta_{\alpha,k}(\Delta^2_k)$. Power $\beta_{\alpha,k+l}(\Delta^2_{k+l})$ computed under the non-central F-distribution $F\left(\frac{n_1 n_2}{n_1+n_2}\Delta^2_{k+l}; k+l, n_1+n_2-(k+l)-1\right)$, 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.

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 one-by-one. 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 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 new 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 $k \leq q$ new 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 $\tilde{T}^2_k$)**.** *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*

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

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 new 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

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

This inequality can then be used in the procedure described in 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 $\tilde{T}^2_k(g) > T^2_{\alpha,k}$ 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 $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2 < \theta$,*

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

*where $T^2_{\alpha,k}$ is the level-α critical value for $\tilde{T}^2_k$ (*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 $\hat{\tilde{\Sigma}}_{[k]}(g) = U_{[k]}\hat{\Sigma}(g)U^\top_{[k]}$.*

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 $\tilde{T}^2_k(g)$ but not by the Euclidean criterion $\|\hat{\tilde{\delta}}_{[k]}(g)\|^2$, the sample covariance matrix in the restricted new space (after filtering) has an eigenvalue below a certain threshold. This implies that such false negative subgraphs (from the Euclidean approximation to the exact algorithm) have a small mean shift in the new 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 significant for these genes, they correspond to small effects that may not be 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 new 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 hypothesis 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 $\mathcal{G}$ with $p = 20$ nodes. Next, we generate $10,000$ datasets in the space corresponding to the basis U defined by the eigenvectors of the $Q_{\mathcal{G}}$ matrix for the graph $\mathcal{G}$; an inverse transformation is applied to random vectors generated is this new space. Each dataset comprises $n_1 = n_2 = 20$ Gaussian random vectors in $\mathbb{R}^p$, with null mean shift δ for $5,000$ datasets and non-null mean shift δ for the remaining $5,000$. For the latter datasets, the non-zero shift is built in the first $k_0 = 3$ graph-based coefficients (the shift being zero for the remaining $p - k_0$ coefficients): $\tilde{\delta}_i \neq 0$ if and only if $i \leq k_0$ and $\Delta^2(\delta, \Sigma) = \Delta^2(\tilde{\delta}, \tilde{\Sigma}) = \tilde{\delta}^\top \tilde{\Sigma}^{-1} \tilde{\delta} = 1$. We consider two covariance settings. In the first one, the covariance matrix in the new space is diagonal, with diagonal elements equal to $\frac{1}{\sqrt{p}}$. In the second one, correlation is introduced between the shifted coefficients only. Specifically, for $i, j \leq k_0$, $\tilde{\Sigma}_{ij} = \frac{0.5}{\sqrt{p}}$ if $i \neq j$, $\tilde{\Sigma}_{ii} = \frac{0.9}{\sqrt{p}}$ otherwise.

Figure 5 displays receiver operator characteristic (ROC) curves for mean shift detection by the standard Hotelling T^2-test, T^2 in the first k_0 graph-based 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 fourth 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 $\mathbf{H}_1$.

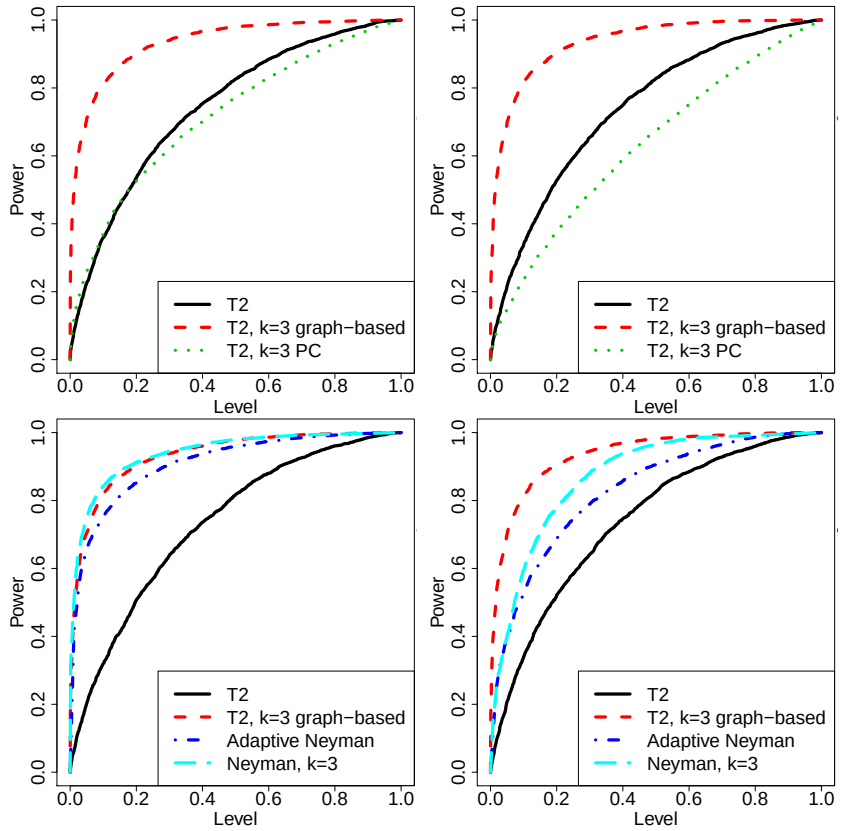

Figure 5: ROC curves for the detection of a smooth shift using various test statistics. Left: Diagonal covariance structure. Right: Block-diagonal covariance structure. Top: Comparison of tests based on the standard Hotelling T^2-statistic in the original space, T^2-statistic in the first k_0 graph-based coefficients, and T^2-statistic in the first k_0 principal components. Bottom: Comparison with the adaptive Neyman statistics of Fan and Lin [1998].

The first important comparison is between the classical Hotelling T^2-test versus the T^2-test in the new graph-based space (top two plots of Figure 5). 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 6 shows the power of the test in the new 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 improvement in performance 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 graph-based test when the shift lies in the first principal components rather than graph-based 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 5 for the sake of readability), which illustrates that working in the new 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] (bottom two plots of Figure 5), 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 of the Neyman test 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.

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 7). 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 $k_0 = 3$ 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.

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

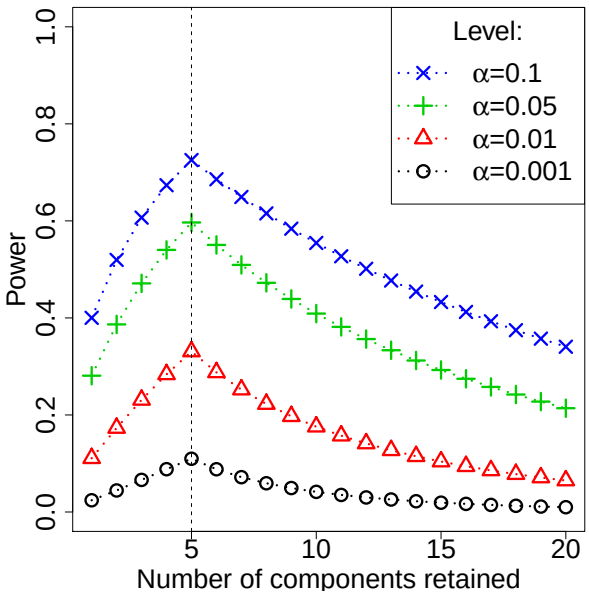

Figure 6: Power of the T^2-test in the first k graph-based coefficients for a graph of 20 nodes, when the actual distribution shift $\Delta^2 = 1$ is evenly distributed among the first $k_0 = 5$ graph-based coefficients and with $n_1 = n_2 = 20$.

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 gene expression data

We also validated our methods using the microarray dataset of Loi et al. [2008], which comprises the expression measures of 15,737 genes for 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 is 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 new graph-based space retaining only the first 20% 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. The Benjamini and Hochberg [1995] procedure was then applied to control the false discovery rate (FDR) at level 0.05.

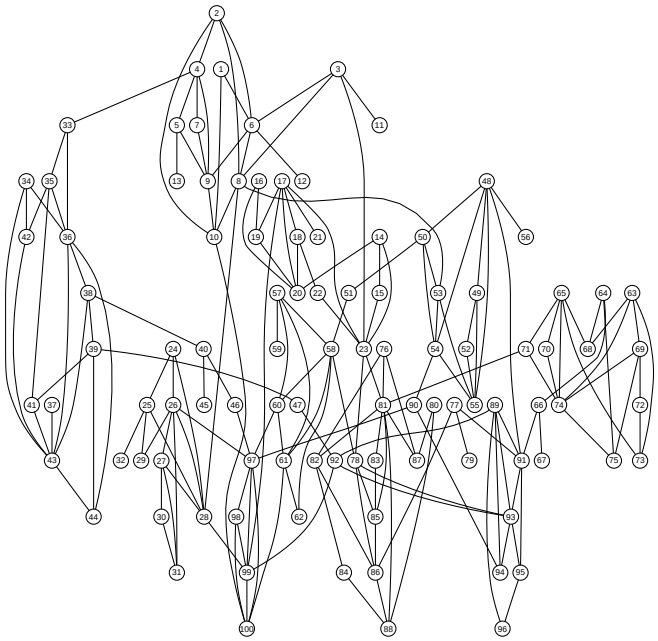

Figure 7: Random graph used in the evaluation of the pruning procedure.

Since there is no gold standard regarding which pathways are actually involved in endocrine resistance, practical validation of the entire set of detected pathways requires advanced biological expertise and further experiments and is the subject of ongoing collaborations. Nonetheless, inspection of our list reveals several pathways which would not have been detected (or would have been farther down in the list) without accounting for the network structure and which have recently been shown to be central in tamoxifen resistance. Many of these pathways involve the Ras/Raf-1/MAPK cascade [McGlynn et al., 2009], like one of the connected components of the prostate cancer pathway shown in Figure 8 and one connected component of the GnRH pathway shown in Figure 9. The former also involves the over-expressed FGFR1, whose amplification was very recently implicated in endocrine therapy resistance by Turner et al. [2010]. The latter pathway involves over-expressed src, which is also a well-studied target when trying to prevent tamoxifen resistance [Herynk et al., 2006]. Both pathways have a much smaller p-value when accounting for their graph structure than when testing in the original gene space : 10^{-4} versus 0.02 for the prostate cancer pathway and 10^{-3} versus 0.11 for the GnRH signaling pathway. This is because the differences in expression of individual genes are insufficient to be significant in 36 and 19 dimensions, respectively, while the expression shift projected in the first 8 and 4 graph-based directions, respectively, is significant. Note that the corresponding p-values for the hypergeometric enrichment test are 0.15 and 0.31. The complete gene lists of

the two components are reported in Tables 1 and 2, respectively. Using a system-based approach like our proposed graph-based test therefore allows to recover several known results (which may not have been obvious from the same data when looking at each gene individually) and may give insight regarding other resistance mechanisms by highlighting connections between these results.

Another example of a network selected only when accounting for graph structure is *Leukocyte transendothelial migration*, shown in Figure 10. 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. Here again, the p-value of the hypergeometric test is extremely high (0.31). The entire list of genes in this component is reported in Table 3.

We then ran our branch-and-bound non-homogeneous subgraph discovery procedure on the cell cycle pathway, whose largest connected component, 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 exceeds $\theta = 0.1$; for a test in the first $k = 3$ components at level $\alpha = 10^{-4}$, this corresponds 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 $\alpha = 10^{-4}$, over 100 runs on permuted data, only 9 rejected the null hypothesis for at least one subgraph. More precisely, 4 of these 9 runs detected 1 subgraph and the others detected 3, 6, 6, 21, and 26 subgraphs. In contrast, 41 overlapping subgraphs (Figure 11) were detected on the original data, corresponding to a connected subnetwork of 25 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]. Accordingly, its negative regulator MDM2 is over-expressed and its positive regulator CREBBP is under-expressed. 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]. Interestingly, RBX1, a gene coding for a RING-domain E3 ligase known to be involved in degradation of estrogen receptor α (ERα) [Ohtake et al., 2007], appears to be over-expressed in resistant patients. This may suggest that some of the resistant ER+ patients had fewer receptors and, as a result, their tumors were relying less on estrogen for their growth; hence, the limited effect of selective estrogen receptor modulator (SERM) like tamoxifen. The networks also contains 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 in 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 an important set of interacting genes and may therefore prove to be more efficient than iterative individual identification of single actors.

6 Software implementation

The graph-structured test of Section 3 is implemented in the R software package `DEGraph`, released through the Bioconductor Project (release 2.7). Instructions for download and installation are available at `http://bioconductor.org/help/bioc-views/2.7/bioc/html/DEGraph.html`. Note that implementations of the branch-and-bound algorithms are not part of the package yet, but are available upon request.

7 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, without enumerating and testing these subgraphs one-by-one. The first algorithm is exact and reduces the number of explicitly tested subgraphs. The second is one 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 large 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-based 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 structured 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-based 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 inter-*

national 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`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, 57:289–300, 1995.

Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Stat.*, 38(2):808–835, Feb 2010.

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.

Matthew H Herynk, Amanda R Beyer, Yukun Cui, Heidi Weiss, Elizabeth Anderson, Tim P Green, and Suzanne A W Fuqua. Cooperative action of tamoxifen and c-src inhibition in preventing the growth of estrogen receptor-positive human breast cancer cells. *Mol Cancer Ther*, 5(12):3023–3031, Dec 2006. doi: 10.1158/1535-7163.MCT-06-0394. URL http://dx.doi.org/10.1158/1535-7163.MCT-06-0394.

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.

Liane M McGlynn, Tove Kirkegaard, Joanne Edwards, Sian Tovey, David Cameron, Chris Twelves, John M S Bartlett, and Timothy G Cooke. Ras/raf-1/mapk pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients. *Clin Cancer Res*, 15(4):1487–1495, Feb 2009. doi: 10.1158/1078-0432.CCR-07-4967. URL `http://dx.doi.org/10.1158/1078-0432.CCR-07-4967`.

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.

Fumiaki Ohtake, Atsushi Baba, Ichiro Takada, Maiko Okada, Kei Iwasaki, Hiromi Miki, Sayuri Takahashi, Alexander Kouzmenko, Keiko Nohara, Tomoki Chiba, Yoshiaki Fujii-Kuriyama, and Shigeaki Kato. Dioxin receptor is a ligand-dependent e3 ubiquitin ligase. *Nature*, 446(7135):562–566, Mar 2007. doi: 10.1038/nature05683. URL `http://dx.doi.org/10.1038/nature05683`.

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

Ted Sandler, John Blitzer, Partha Talukdar, and Fernando Pereira. Regularized learning with networks of features. In *Neural Information Processing Systems*, Cambridge, MA, 2009. MIT Press.

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.

Nicholas Turner, Alex Pearson, Rachel Sharpe, Maryou Lambros, Felipe Geyer, Maria A Lopez-Garcia, Rachael Natrajan, Caterina Marchio, Elizabeth Iorns, Alan Mackay, Cheryl Gillett, Anita Grigoriadis, Andrew Tutt, Jorge S Reis-Filho, and Alan Ashworth. Fgfr1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. *Cancer Res*, 70(5):2085–2094, Mar 2010. doi: 10.1158/0008-5472.CAN-09-3746. URL `http://dx.doi.org/10.1158/0008-5472.CAN-09-3746`.

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.

A Gene lists

Table 1: *Prostate cancer pathway.* Gene list with univariate t-statistics and corresponding nominal p-values. The unadjusted p-values for the T^2-statistics in the original space and in the graph-based reduced space are p-value(Hotelling)=0.019 and p-value(netHotelling)=0.00014, respectively. The p-value for the standard hypergeometric enrichment test is p-value(hyper)=0.15.

Gene Entrez	Gene symbol	t-statistic	p-value
369	ARAF	0.25	0.81
673	BRAF	0.87	0.39
1950	EGF	-1.3	0.19
1956	EGFR	-0.22	0.82
2064	ERBB2	1.1	0.25
2260	FGFR1	1.7	0.098
2263	FGFR2	-2.4	0.019
2885	GRB2	1.9	0.056
3265	HRAS	2.7	0.0085
3479	IGF1	-3.5	0.00058
3630	INS	0.29	0.77
3645	INSRR	0.88	0.38
3845	KRAS	0.76	0.45
4893	NRAS	0.45	0.65
5154	PDGFA	-0.81	0.42
5155	PDGFB	3.2	0.0021
5156	PDGFRA	-3.1	0.0024
5159	PDGFRB	-0.88	0.38
5290	PIK3CA	0.38	0.71
5291	PIK3CB	-0.23	0.82
5293	PIK3CD	-2.6	0.0092
5294	PIK3CG	-0.59	0.56
5295	PIK3R1	-1.8	0.07
5296	PIK3R2	1.4	0.17
5594	MAPK1	0.44	0.66
5595	MAPK3	0.99	0.32
5604	MAP2K1	1.1	0.26
5605	MAP2K2	0.84	0.4
5894	RAF1	0.73	0.47
6654	SOS1	0.73	0.47
6655	SOS2	0.28	0.78
7039	TGFA	3.3	0.0015

8503	PIK3R3	-0.35	0.72
23533	PIK3R5	0.71	0.48
56034	PDGFC	-1.8	0.071
80310	PDGFD	-2.8	0.0068

Table 2: *GnRH signaling pathway.* Gene list with univariate t-statistics and corresponding nominal p-values. The unadjusted p-values for the T^2-statistics in the original space and in the graph-based reduced space are p-value(Hotelling)=0.11 and p-value(netHotelling)=0.0012, respectively. The p-value for the standard hypergeometric enrichment test is p-value(hyper)=0.31.

Gene Entrez	Gene symbol	t-statistic	p-value
1839	HBEGF	-0.8	0.43
1956	EGFR	-0.22	0.82
2002	ELK1	1.6	0.1
2885	GRB2	1.9	0.056
3265	HRAS	2.7	0.0085
3845	KRAS	0.76	0.45
4313	MMP2	-1.9	0.066
4893	NRAS	0.45	0.65
5578	PRKCA	-0.97	0.33
5579	PRKCB1	-1.8	0.075
5580	PRKCD	-0.74	0.46
5594	MAPK1	0.44	0.66
5595	MAPK3	0.99	0.32
5604	MAP2K1	1.1	0.26
5605	MAP2K2	0.84	0.4
5894	RAF1	0.73	0.47
6654	SOS1	0.73	0.47
6655	SOS2	0.28	0.78
6714	SRC	2.6	0.012

Table 3: *Leukocyte transendothelial migration pathway.* Gene list with univariate t-statistics and corresponding nominal p-values. The unadjusted p-values for the T^2-statistics in the original space and in the graph-based reduced space are p-value(Hotelling)=0.073 and p-value(netHotelling)=1.5e-05, respectively. The p-value for the standard hypergeometric enrichment test is p-value(hyper)=0.31.

Gene Entrez	Gene symbol	t-statistic	p-value
60	ACTB	-0.64	0.53
71	ACTG1	1.4	0.15
387	RHOA	0.067	0.95
394	ARHGAP5	-0.58	0.56
998	CDC42	0.94	0.35
1432	MAPK14	1.9	0.057
1535	CYBA	-1.1	0.27
1536	CYBB	-1	0.31
2770	GNAI1	0.066	0.95
2771	GNAI2	-0.87	0.39
2773	GNAI3	1.1	0.3
2909	GRLF1	-1.8	0.072
3676	ITGA4	-2.1	0.037
3683	ITGAL	-1.6	0.12
3684	ITGAM	-2.2	0.03
3688	ITGB1	-1.1	0.27
3689	ITGB2	-1.9	0.056
3702	ITK	-1	0.32
4313	MMP2	-1.9	0.066
4318	MMP9	1.2	0.22
4633	MYL2	0.72	0.47
4636	MYL5	0.63	0.53
4688	NCF2	-0.12	0.9
4689	NCF4	-0.85	0.4
5290	PIK3CA	0.38	0.71
5291	PIK3CB	-0.23	0.82
5293	PIK3CD	-2.6	0.0092
5294	PIK3CG	-0.59	0.56
5295	PIK3R1	-1.8	0.07
5296	PIK3R2	1.4	0.17
5600	MAPK11	0.79	0.43
5603	MAPK13	2	0.052

5879	RAC1	0.45	0.65
5880	RAC2	-2.4	0.02
5906	RAP1A	-1.1	0.27
5908	RAP1B	0.25	0.8
6093	ROCK1	-1.7	0.098
6300	MAPK12	2.1	0.038
6494	SIPA1	-1.4	0.18
7070	THY1	0.083	0.93
7294	TXK	0.85	0.4
7409	VAV1	-1.6	0.11
7410	VAV2	0.69	0.49
7412	VCAM1	-1.2	0.23
8503	PIK3R3	-0.35	0.72
9475	ROCK2	0.53	0.6
10398	MYL9	-0.92	0.36
10411	RAPGEF3	-2.1	0.04
10451	VAV3	-1.3	0.19
10627	MRCL3	0.96	0.34
11069	RAPGEF4	-0.74	0.46
23533	PIK3R5	0.71	0.48
27035	NOX1	-0.034	0.97
29895	MYLPF	0.64	0.53
50508	NOX3	1.1	0.25
58498	MYL7	0.14	0.89
83593	RASSF5	-5.2	8.6e-07
93408	MYLC2PL	-0.87	0.39
103910	MRLC2	0.36	0.72

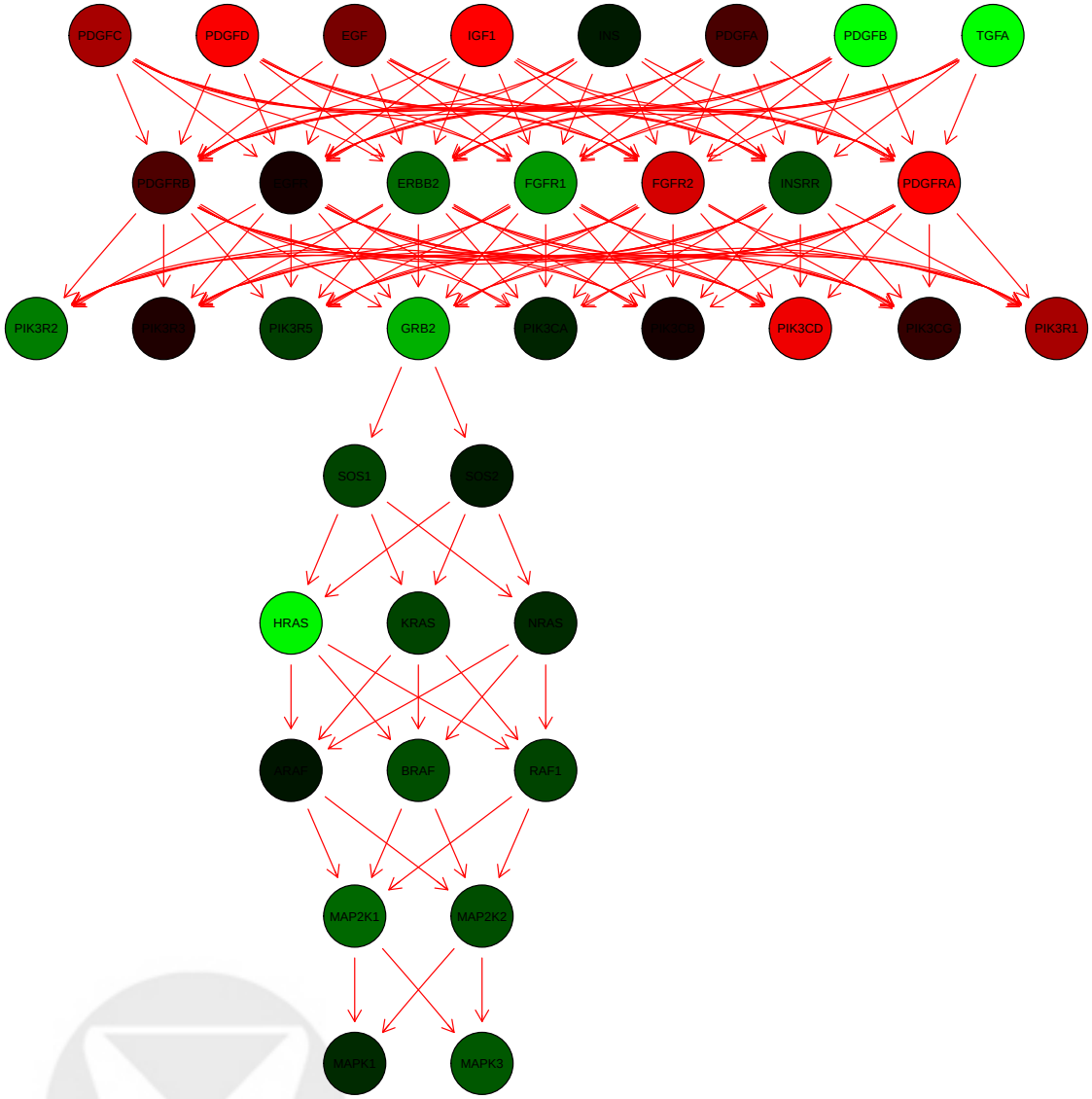

Figure 8: Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG prostate cancer pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activation.

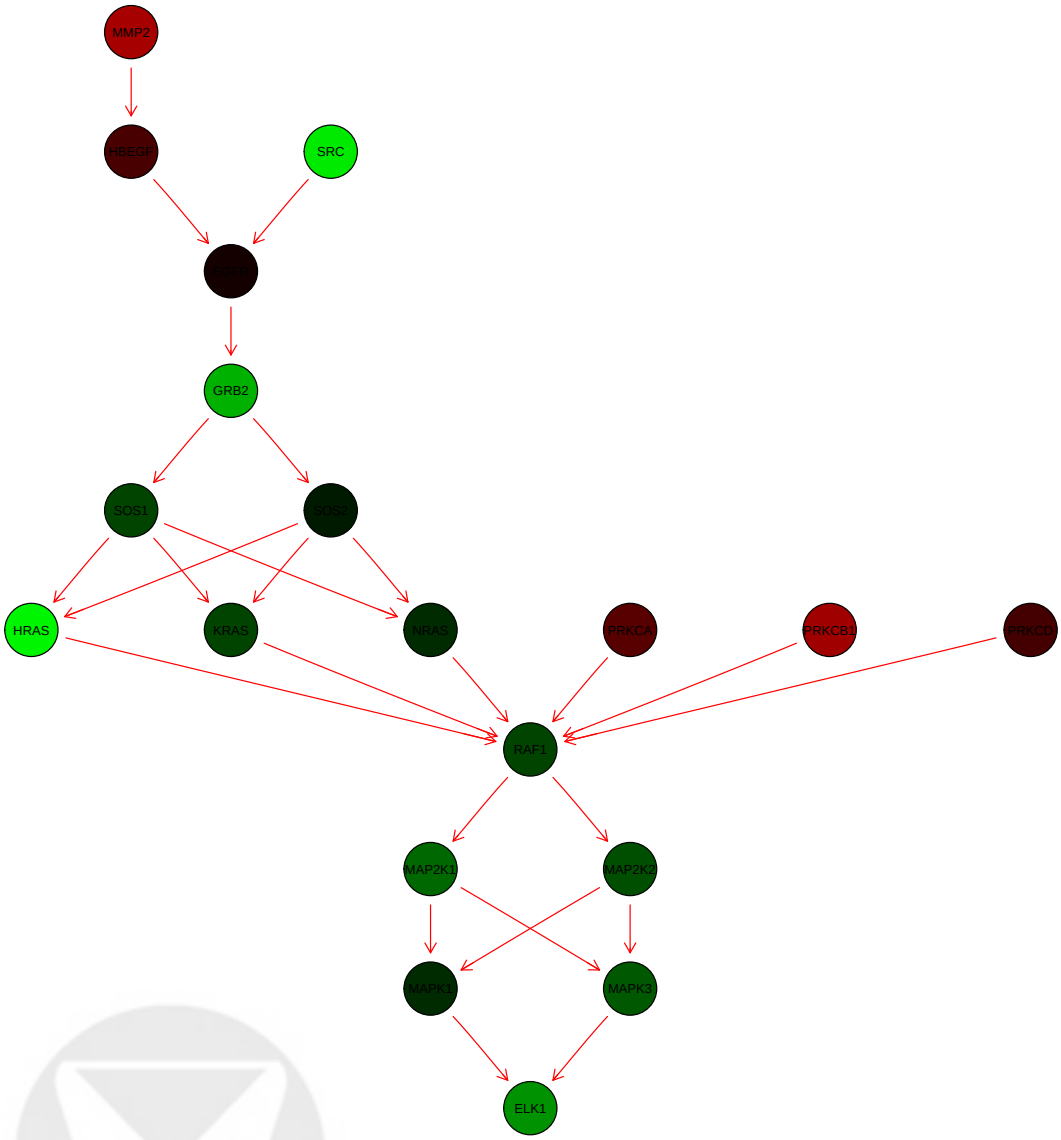

Figure 9: Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG GnRH signaling pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activation.

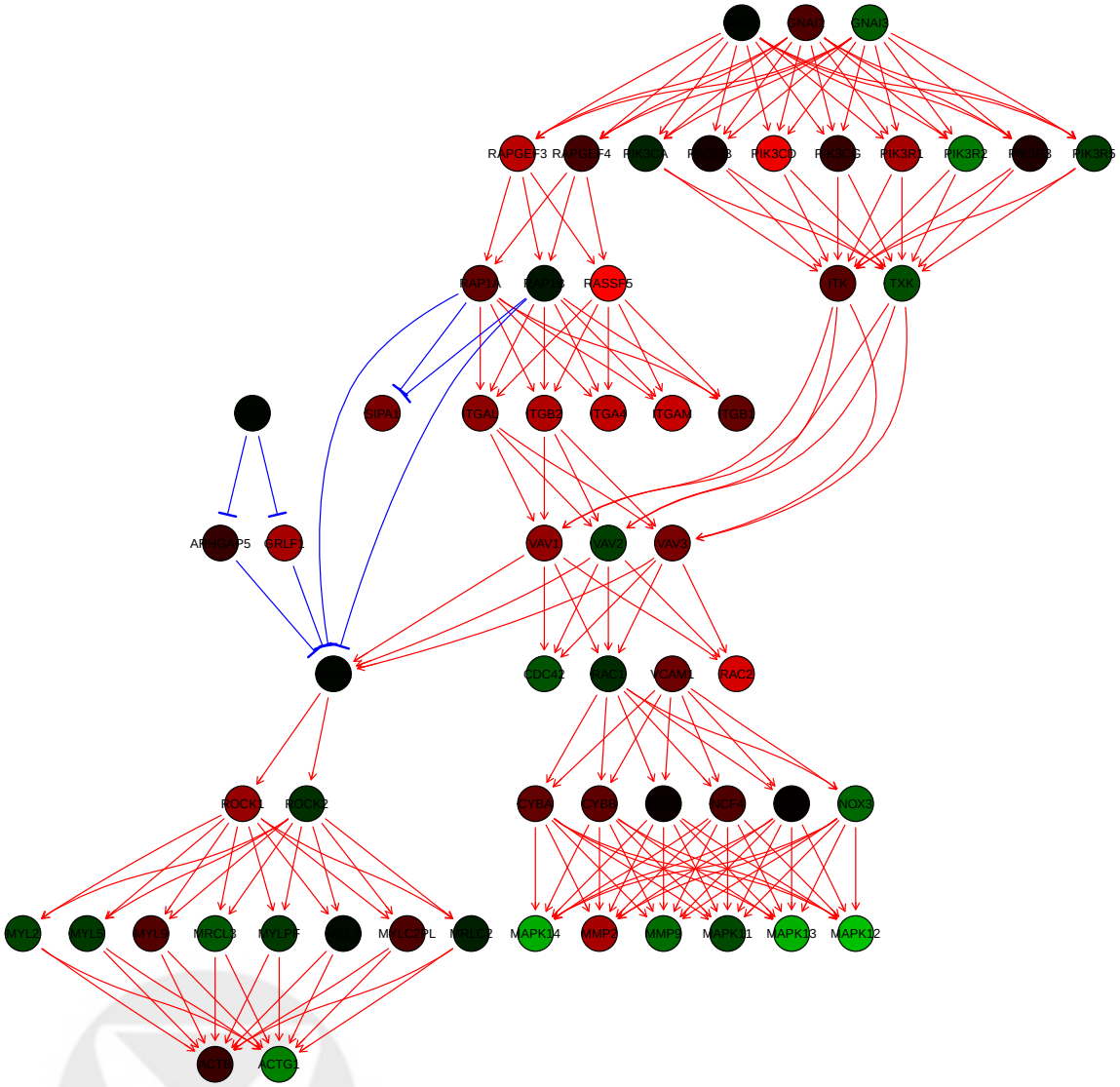

Figure 10: Scaled difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in one component of the KEGG leukocyte transendothelial migration pathway. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activation, blue arrows inhibition.

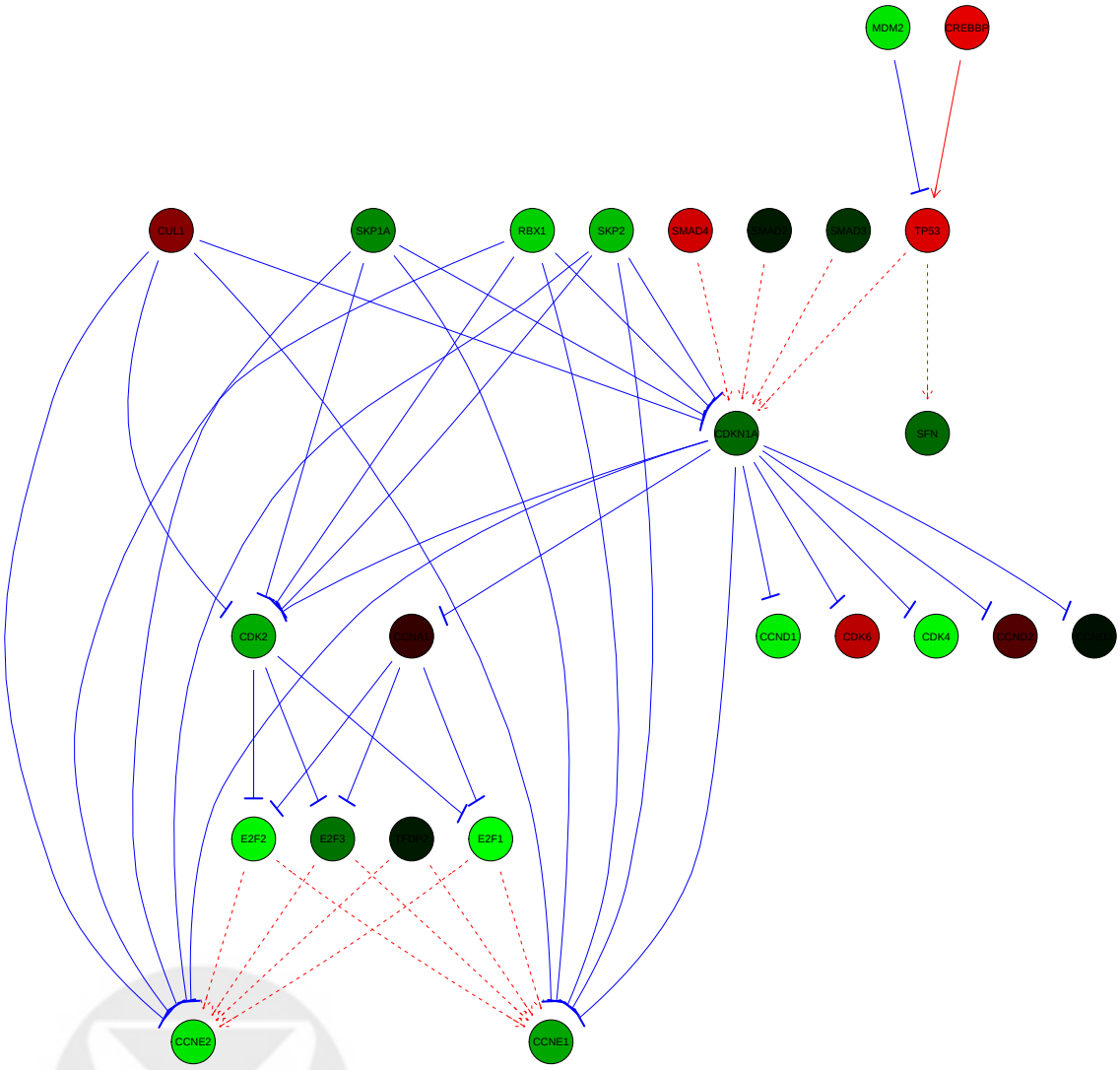

Figure 11: Difference in sample mean expression measures between tamoxifen-resistant and sensitive patients, for genes in the two overlapping subgraphs detected at $\alpha = 10^{-4}$. Nodes are colored according to the value of the difference in means, with green corresponding to high positive values, red to high negative values, and black to 0. Red arrows denote activation, blue arrows inhibition.