More Power via Graph-Structured Tests for Differential Expression of Gene Networks

Laurent Jacob[1], Pierre Neuvial[1,3], Sandrine Dudoit[1,2]

[1]UC Berkeley, Department of Statistics
[2]UC Berkeley, Division of Biostatistics
[3]Laboratoire Statistique et Génome (Univ. Évry, CNRS, INRA)

http://stat.genopole.cnrs.fr/~pneuvial

SMPGD 2012

Outline

1. Multivariate two-sample tests of gene expression
2. Harmonic analysis on graphs
3. Two-sample test on a graph
4. Non-homogeneous subgraph discovery

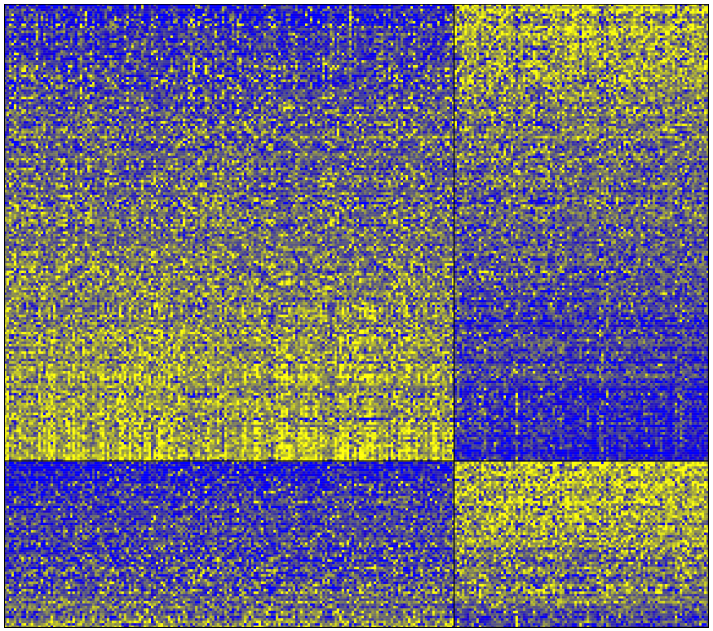
Outline

1. Multivariate two-sample tests of gene expression
2. Harmonic analysis on graphs
3. Two-sample test on a graph
4. Non-homogeneous subgraph discovery

Differential expression analyses

Setting

- Data: a **thin** ($n \times p, n \ll p$) gene expression matrix
- Outcome: two phenotypes with sample size n_1, n_2 such that $n_1 + n_2 = n$
- Goal: Find a subset of genes in $\{1 \dots p\}$ whose mean expression differ between phenotypes



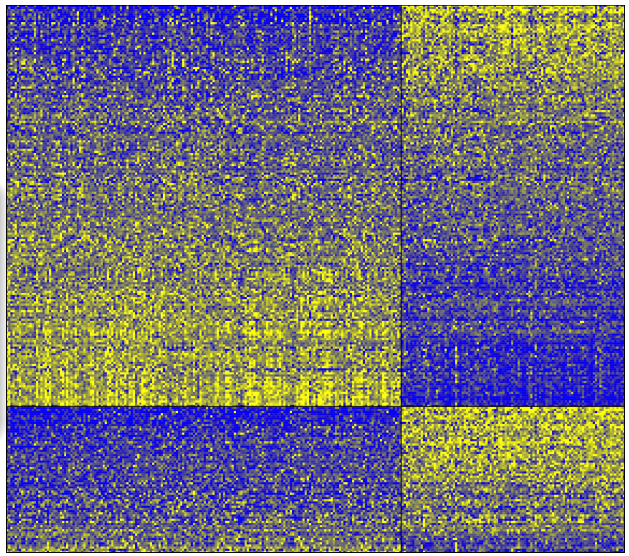
Classical approach

- one test per gene
- multiple testing correction

Problem: **interpretability** of gene lists

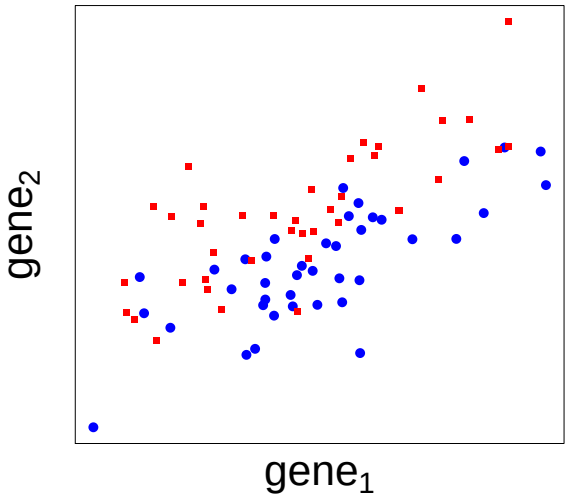
Gene set enrichment analyses

Idea: incorporate biological information through **gene sets**



Two step approaches to gene set enrichment

1. Test differential expression of genes,
2. Test enrichment of gene sets in DE genes.

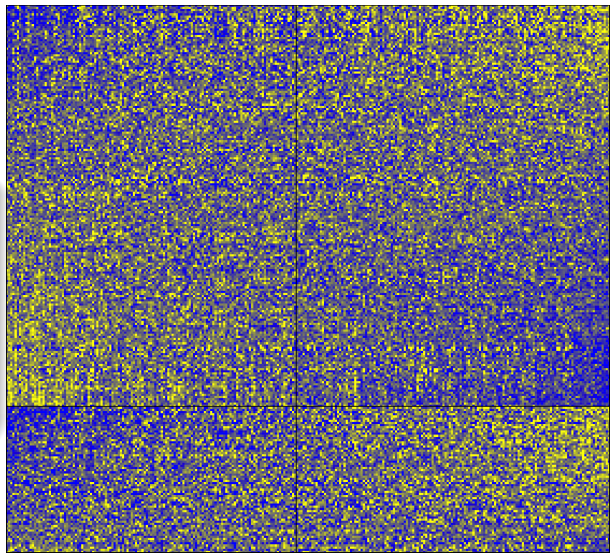


Limitations

- **univariate**: correlation structure between genes lost at step 1
- **pathways** are more than gene "sets"
- unclear interpretation of the **null hypothesis tested**

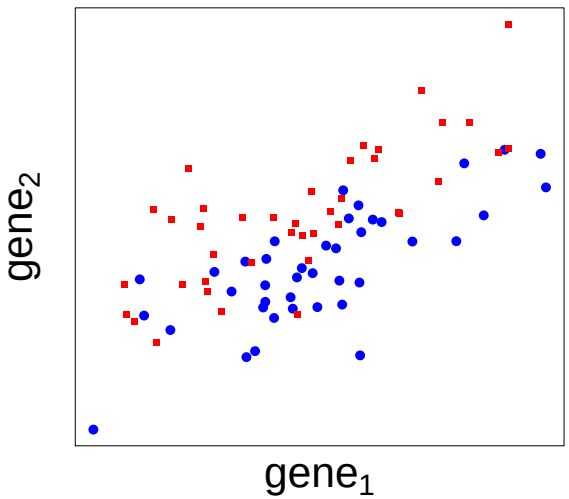
Gene set enrichment analyses

Idea: incorporate biological information through **gene sets**



Two step approaches to gene set enrichment

1. Test differential expression of genes,
2. Test enrichment of gene sets in DE genes.



Limitations

- **univariate**: correlation structure between genes lost at step 1
- **pathways** are more than gene "sets"
- unclear interpretation of the **null hypothesis tested**

Multivariate two sample tests

Generalization of Student's t-test to p-dimensional vectors

Hotelling's T^2 test

Let (n_1, n_2) such that $p < n_1 + n2 - 1$. Assume $(x_{1j})_{1 \leq j \leq n_1}$ are iid $\sim \mathcal{N}_p(\mu_1, \Sigma)$ and $(x_{2j})_{1 \leq j \leq n_2}$ are iid $\sim \mathcal{N}_p(\mu_2, \Sigma)$. Let

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_2 - \bar{x}_1)^T \widehat{\Sigma}^{-1} (\bar{x}_2 - \bar{x}_1)$$

where $\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}$ and $\bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}$. Then T^2 follows **Fisher's distribution** $F(N\Delta^2; p, n1 + n2 - p - 1)$ with non-centrality parameter $N\Delta^2$, where $N = \frac{n_1 n_2}{n_1 + n_2}$ and $\Delta^2 = (\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1)$

Limitations

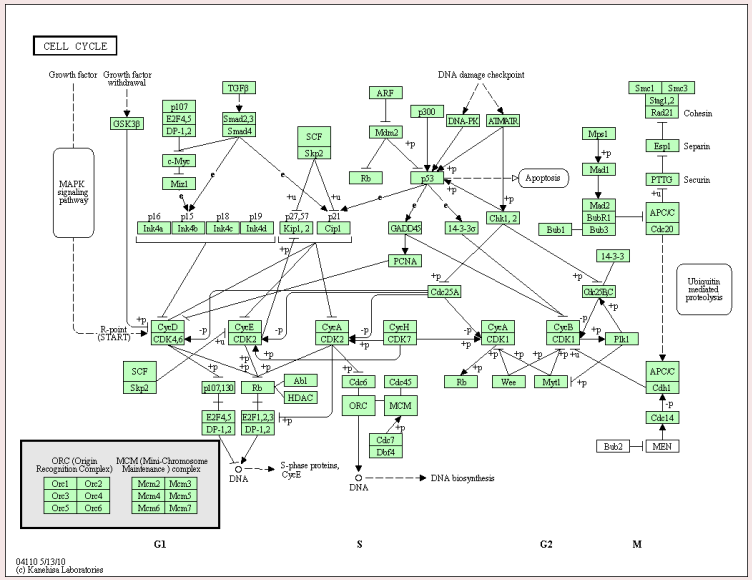
- only applies when $p < n_1 + n2 - 1$
- even then loses power quickly in high dimension due to $\widehat{\Sigma}^{-1}$

Structured Two-Sample Test

Idea

Use prior information on distribution shift to **reduce dimension** and **gain power**.

Gene networks



Possible prior : distribution shift is (partly) coherent with known network.

Outline

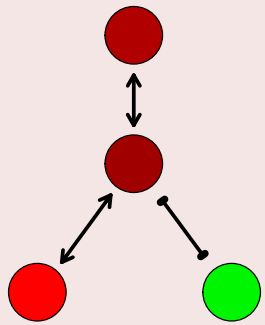
1. Multivariate two-sample tests of gene expression
2. Harmonic analysis on graphs
3. Two-sample test on a graph
4. Non-homogeneous subgraph discovery

Gene profiles as functions on graphs

Idea

- A function on a graph **associates a real value to each of its nodes**
- Any vector $f \in \mathbb{R}^{|\mathcal{V}|}$ may be interpreted as a function on $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- *E.g.* :
  - Gene expressions for the genes in the network (x_i),
  - Average of gene expressions over patients within a phenotype $(\bar{x}_1)$,
  - Difference between the averages within the two phenotypes $(\bar{x}_1 - \bar{x}_2)$.

Example



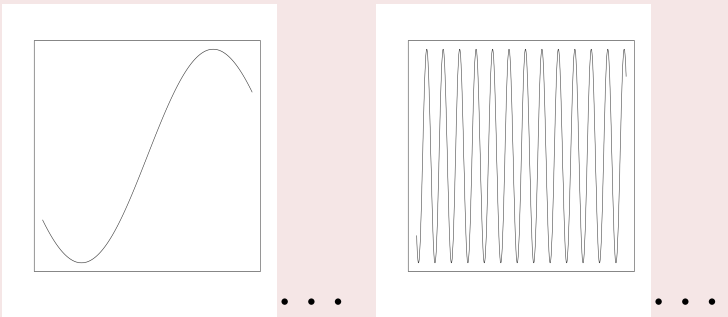
Energy on a graph

Hilbert space

- Gradient operator ∇.
- Laplace operator $\mathcal{L} = -\text{div}\nabla$.
- Dirichlet energy of function f :

$$\frac{1}{2}\int |\nabla f(x)|^2 dx.$$

- Eigenfunctions of $\mathcal{L}$: sinusoids with increasing energy :

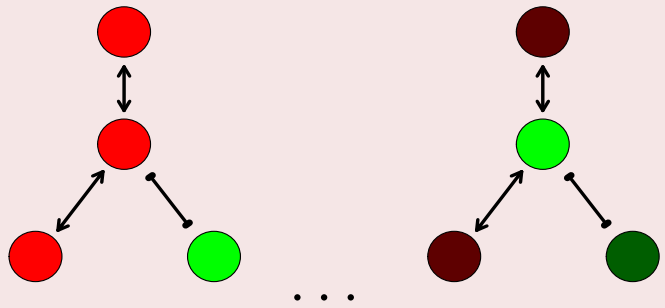


Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Gradient matrix $\nabla \in \mathbb{R}^{|\mathcal{E}|,|\mathcal{V}|}$.
- Laplacian matrix $\mathcal{L} = \nabla^\top \nabla$.
- Dirichlet energy of $f \in \mathbb{R}^{|\mathcal{V}|}$:

$$\frac{1}{2} f^\top \mathcal{L} f = \frac{1}{2}\|\nabla f\|^2.$$

- Eigenvectors of $\mathcal{L}$: vectors with increasing energy :



Energy is defined by the **graph topology** (regardless of expression values)

"Graph-Fourier" decomposition

- **Graph-Fourier** coefficients $\tilde{f}_i$ are projections of f on eigenvectors of $\mathcal{L}$:

$$\tilde{f}_i \triangleq u_i^\top f, \; i = 1, \ldots, |\mathcal{V}|.$$

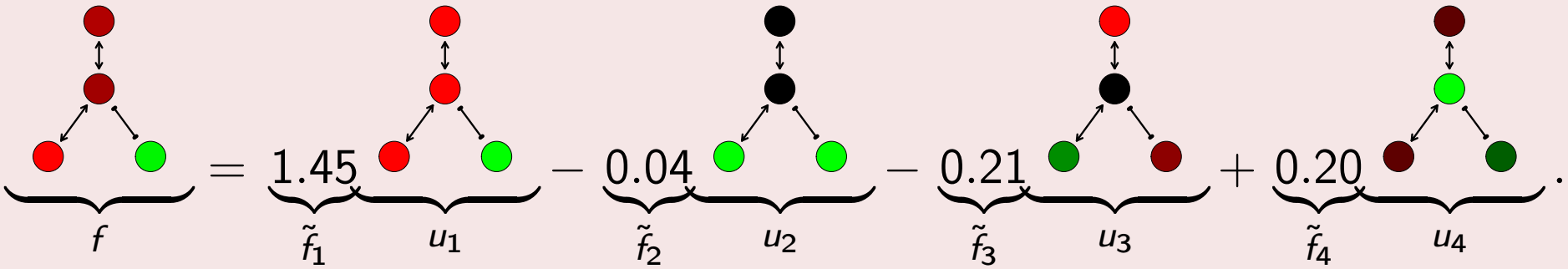
- Inverse transform :

$$f = \sum_{i=1}^{|\mathcal{V}|} \tilde{f}_i u_i.$$

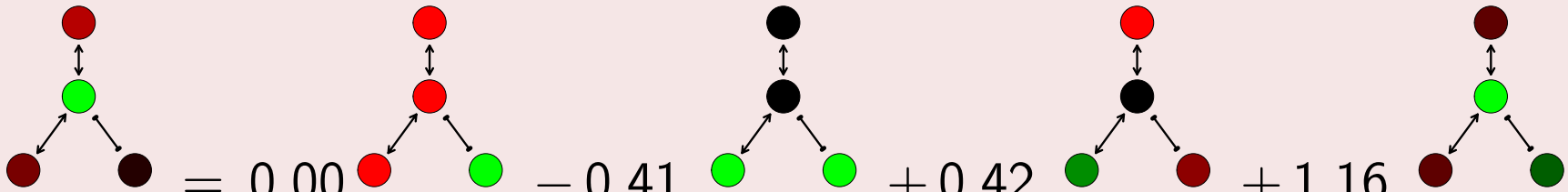
- f and $\tilde{f}$ are **two dual ways of writting the same function** :
  - As node values f_i (*e.g.* expression shifts),
  - As graph-Fourier coefficients $\tilde{f}_i$.

Harmonic analysis on graphs

Example 1 (smooth shift)



Example 2 (non-smooth shift)



Remark

Smooth functions have **large** coefficients at the **beginning** of the spectrum.

Outline

1. Multivariate two-sample tests of gene expression
2. Harmonic analysis on graphs
3. Two-sample test on a graph
4. Non-homogeneous subgraph discovery

Statistic in (graph-)frequency domain

Test statistic

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

where $U_{[k]}$ is the restriction of U to its first k columns

Remarks

- Equivalent to test in frequency and graph domain ($T^2 = \tilde{T}^2$).
- More generally :

T^2 computed after **filtering out** frequencies above k

$=$

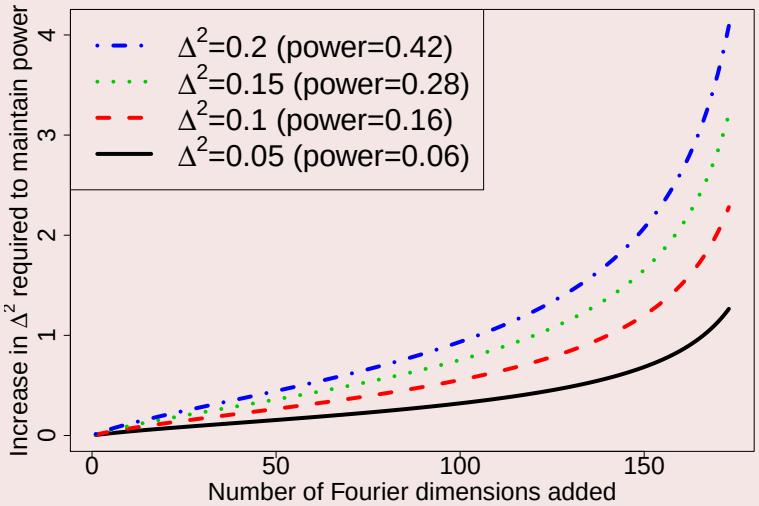
$\tilde{T}_k^2$ computed in frequency domain restricted to the first k coefficients.

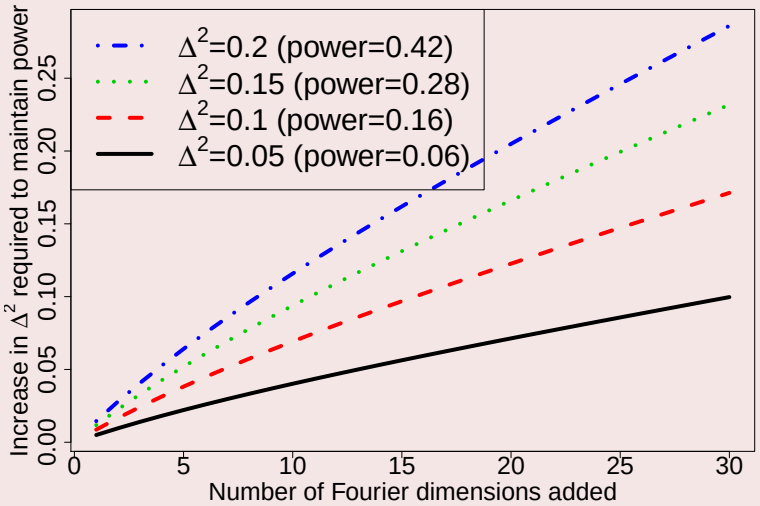
Gain in power

Lemma

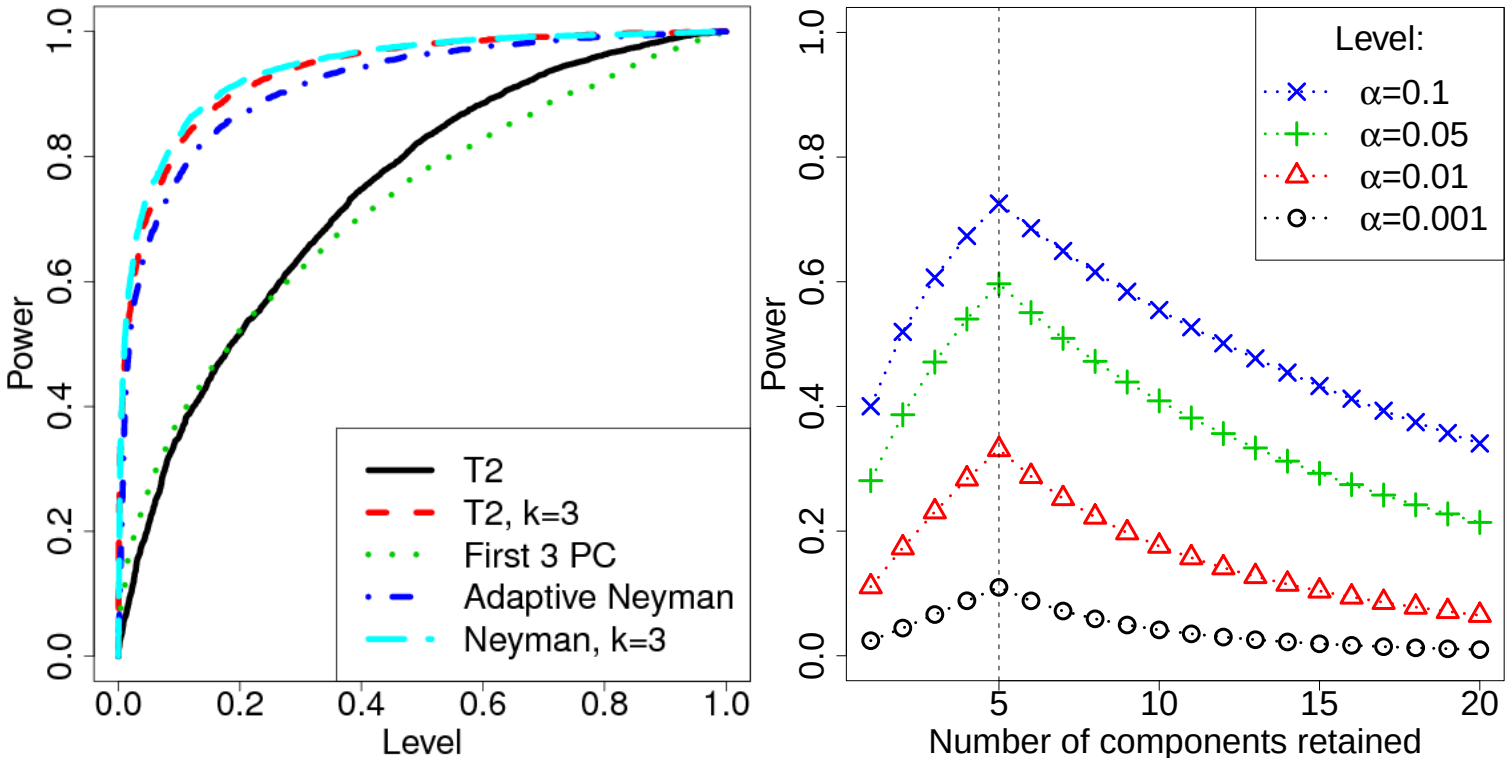
For any level and any number of Fourier coefficients, ***maintaining the power*** *of the* T^2 *test in the Fourier space after adding a coefficient requires a* ***strictly positive increase of distribution shift****.*

Illustration





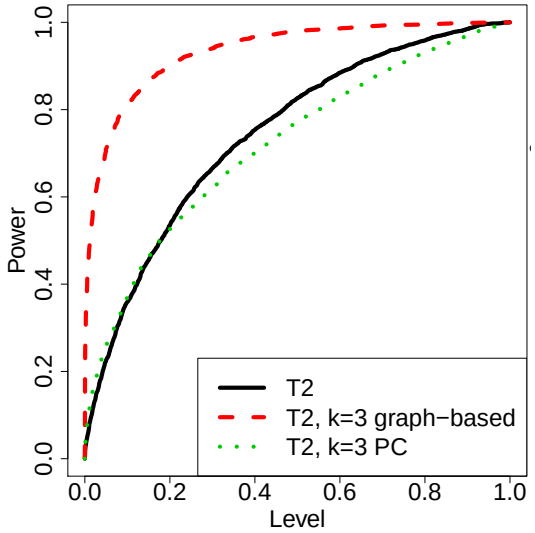
Synthetic data, gain in power



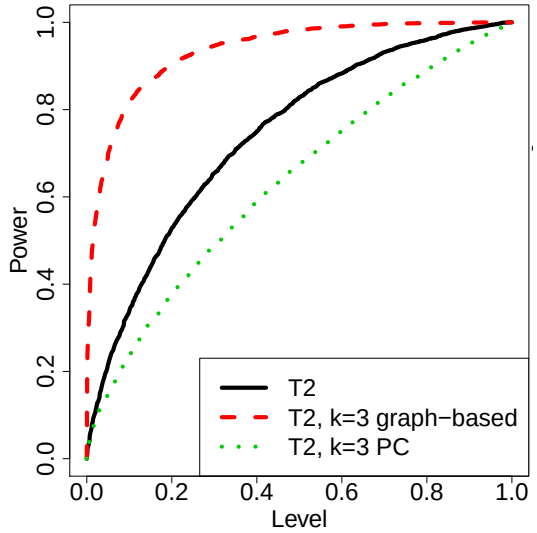
- Left : ROC curves for the detection of a smooth shift for various test statistics, with diagonal covariance structure.
- Right : Power of the T^2-test in the graph-Fourier space with shift evenly distributed among the first $k = 5$ coefficients.

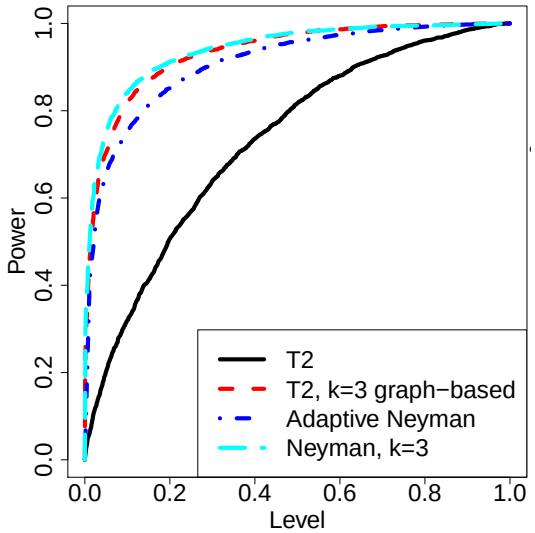
Synthetic data, gain in power

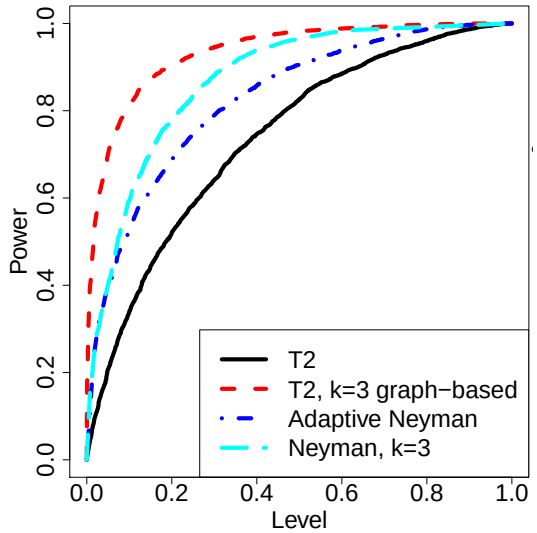
Diagonal covariance



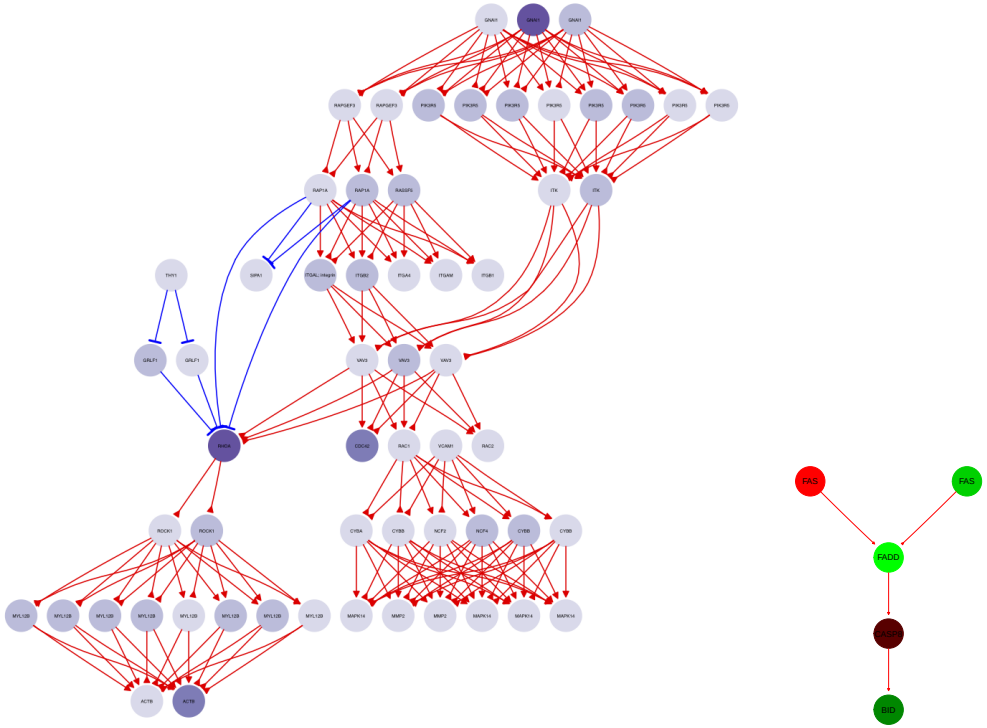
Block-diagonal covariance







Breast cancer and KEGG data, known pathways



Difference in sample mean expression measures between tamoxifen-resistant and non-resistant patients, for genes in two KEGG regulation networks.

- Left : Regulation network (Leukocyte transendothelial migration) with the lowest ratio of graph-Fourier to full space p-values.
- Right : Regulation network (Alzheimer's disease) with the highest ratio of graph-Fourier to full space p-values.

Outline

1. Multivariate two-sample tests of gene expression
2. Harmonic analysis on graphs
3. Two-sample test on a graph
4. Non-homogeneous subgraph discovery

Non-homogeneous subgraph discovery

Motivation

- Relevant **pathways** for the studied phenomenon may be subgraphs of known networks.
- Search for subgraphs with location shift.
- Strategy : apply test to all subgraphs of size q.

Algorithm

Use a branch-and-bound like strategy :

1. Check, for each $v \in \mathcal{V}$, whether $\tilde{T}_k^2$ of any subgraph of size q containing v can be **guaranteed to be below the critical value**.
2. If this is the case, v is removed from $\mathcal{G}$.
3. Repeat the procedure on the edges of the remaining graph and, iteratively, on the subgraphs up to size $q-1$.
4. Test all remaining subgraphs of size q.

Upper bound on $\tilde{T}_k^2$

Potential issue

"$\tilde{T}_k^2$ *of any subgraph of size* q *containing* v" depends on the **Laplacian of the subgraph** (not only on node values).

Lemma

For any subgraph g *of* $\mathcal{G}$ *of size* $q \leq p$*, any subgraph* g' *of* g *of size* $s \leq q$*, and any* $k \leq q$*, then*

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

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*.*

Euclidean approximation

Limit

- If the large graph is very connected, the exact bound can be loose.
- Use a bound on the Euclidean norm.

Eculidean bound

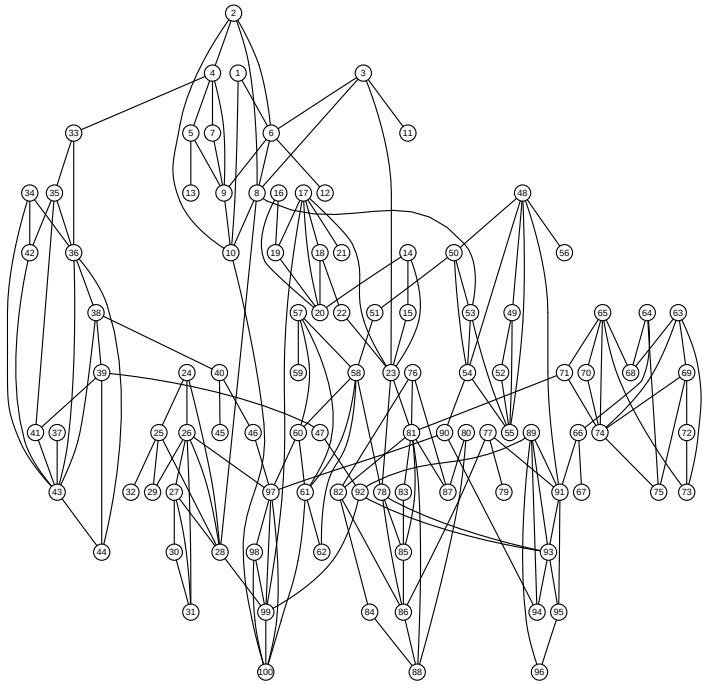
$$\|U_{[k]}^{\top}(\bar{x}_1(g) - \bar{x}_2(g))\|^2 \leq \|\bar{x}_1(g') - \bar{x}_2(g')\|^2 + \max_{\mathbf{v}=v_1,\ldots,v_{q-s}\in\nu(g',q-s)} \|\bar{x}_1(\mathbf{v}) - \bar{x}_2(\mathbf{v})\|^2.$$

Note on the type of false negatives

- Subgraphs missed by the Euclidean approximation are those with a **small shift** in a direction of **small variance**.
- An **upper bound** on this variance can be written.
- Those are **classically filtered out**.

Synthetic data, discovery algorithm

- Artificial graph of 100 nodes, 177 edges, non-zero mean shift on one 5-node subgraph in its first 3 Fourier coefficients.
- Full enumeration : $\mathbf{732 \pm 9}$ seconds per run .

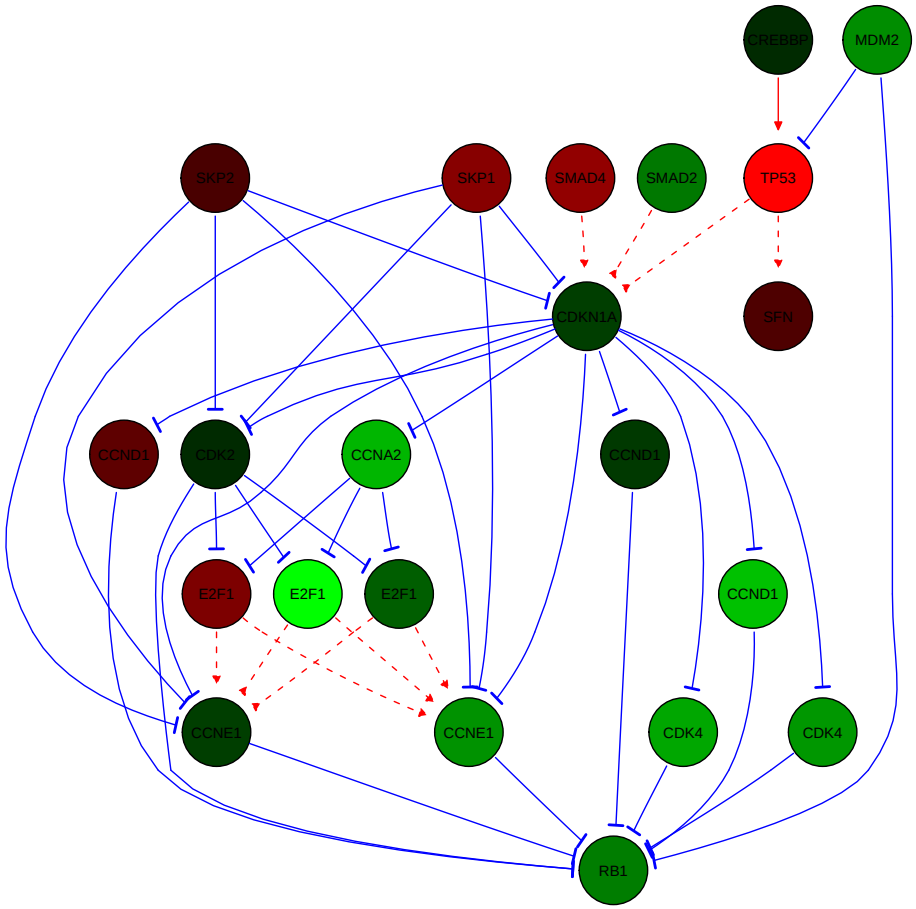


- Exact algorithm : $\mathbf{627 \pm 59}$ seconds ($\mathbf{578 \pm 100}$ on permuted data).
- Approximation at $\theta = 0.5$ ($\lambda_{min} \leq 0.52$) : $\mathbf{204 \pm 86}$ seconds ($\mathbf{129 \pm 91}$ on permuted data).
- Approximation at $\theta = 1$ ($\lambda_{min} \leq 1.04$) : $\mathbf{183 \pm 106}$ seconds ($\mathbf{40 \pm 60}$ on permuted data). Missed the non-homogeneous subgraph in 5% of the runs.

Breast cancer and KEGG data, pathway discovery

Discovery procedure on **cell cycle pathway** (86 nodes, 442 edges)

Search for subgraphs of size 5, $k = 3$, $\theta = 0.1$, $\alpha = 10^{-4}$ ($\lambda_{min} \leq 0.23$).
31 overlapping subgraphs detected :



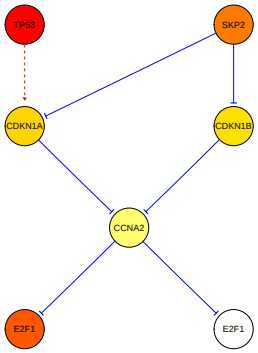
- E2F1 : very recently discovered to play a central role in tamoxifen resistance,
- CDKNA1-2 : low individual t-scores, recently found to be involved in ovarian cancer.

Breast cancer and KEGG data, pathway discovery

Discovery procedure on **cell cycle pathway** (86 nodes, 442 edges)

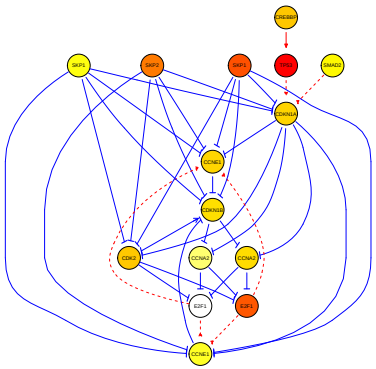
Search for subgraphs of size 5, $k = 3$, $\theta = 0.1$.

- At $\alpha = 10^{-4}$ ($\lambda_{min} \leq 0.23$), two overlapping subgraphs detected :



  - E2F1 : very recently discovered to play a central role in tamoxifen resistance,
  - CDKNA1-2 : low individual t-scores, recently found to be involved in ovarian cancer.
  - No positive detection on 50 permutations.

- At $\alpha = 2.10^{-4}$ ($\lambda_{min} \leq 0.24$), 15 overlapping subgraphs detected.



Only two of 50 permuted runs selected 2 subgraphs each.

Summary

- **Graph-structured two-sample test** of means, for problems in which the distribution shift is assumed to be smooth on a given graph.
- Proved quantitative results on **power gains** for such smooth-shift alternatives.
- Devised branch-and-bound algorithms to systematically apply our test to **all the subgraphs of a large graph**:
  1. exact algorithm: reduces the number of explicitly tested subgraphs.
  2. approximate algorithm: quantitative result on the type of missed subgraphs.
- Promising results on drug resistance microarray dataset.

Acknowledgements and references

Thanks to **Laurent Jacob** and Sandrine Dudoit !

Funding

- UC Berkeley Center for Computational Biology **Genentech** Innovation Fellowship.
- The Cancer Genome Atlas Project (**TCGA**).

References

L. Jacob, P. Neuvial and S. Dudoit.
More Power via Graph-Structured Tests for Differential Expression of Gene Networks.
AoAS (to appear) `http://hal.archives-ouvertes.fr/hal-00521097/en`

- Bioconductor R packages: `DEGraph` and `NCIgraph`

Gain in power

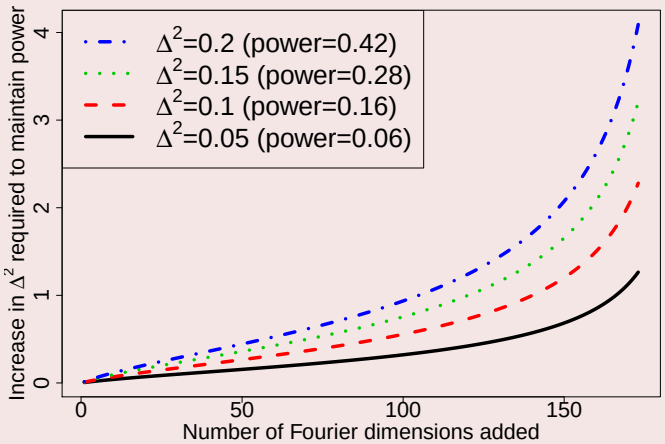
Lemma

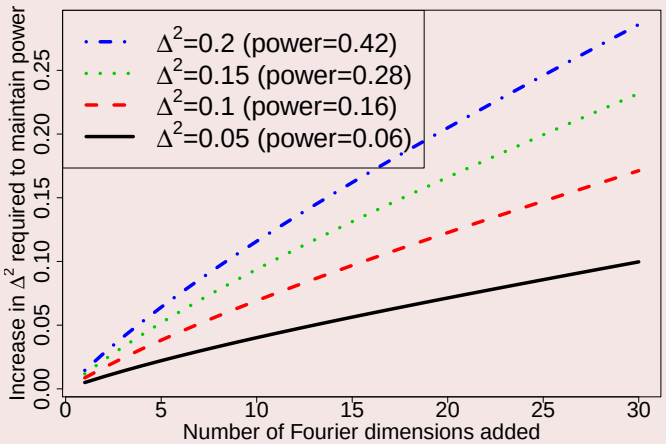
For any level α *and any* $1 < l \leq p - k$*, there exists* $d(\alpha, k, l) > 0$ *such that*

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

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*.*

Illustration





Note on the type of false negatives

Subgraphs missed by the Euclidean approximation are those with a **small shift** in a direction of **small variance** :

Lemma (Characterization of missed subgraphs)

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

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

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

Those are **classically filtered out** because not interesting from a practical point of view.