Machine learning, statistical, and network science approaches for comparing brain graphs within and between modalities

Jonas Richiardi
FINDlab / LabNIC

http://www.stanford.edu/~richiard/

STANFORD UNIVERSITY
Dept. of Neurology & Neurological Sciences

UNIVERSITÉ DE GENÈVE
Dept. of Neuroscience
Dept. of Clinical Neurology

CRM Neuro workshop 24/10/13

Research question and applications

Given two brain graphs, representing "connectivity", how "similar" are they?

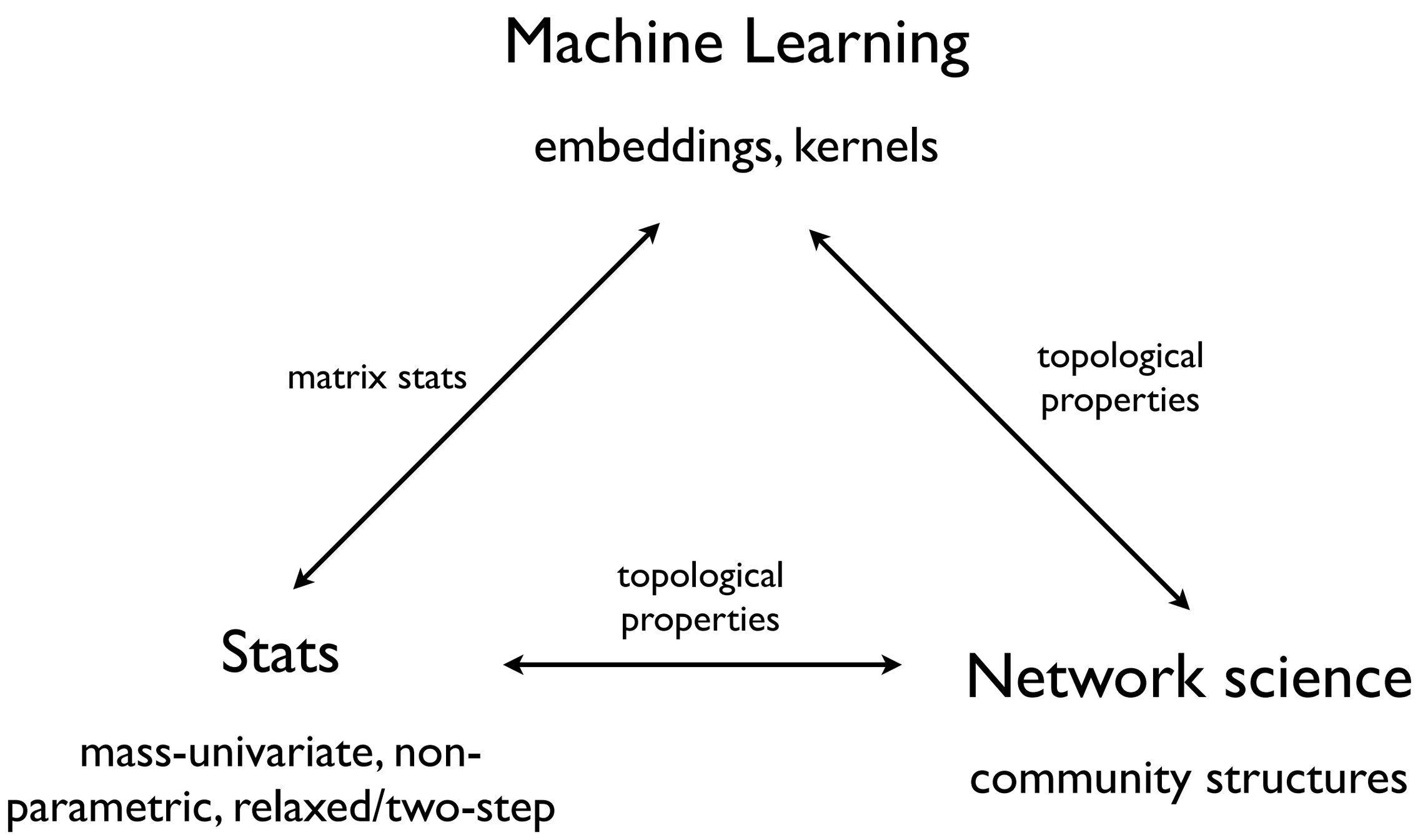
Within subject: *How do the graphs differ between experimental conditions?*

Between subjects: *How do the graphs differ between disease states ?*

Between modalities: *Are some aspects of the graph's topology preserved across modalities?*

Across spatial scales: *Are the differences over the whole graph, or localised in a subgraph, or limited to single edge or vertex?*

Overview of approaches



[Richiardi et al., IEEE Sig. Proc. Mag., 2013] [Richiardi & Ng, GlobalSIP, 2013]

Labelled graphs

"Brain graphs" can be expressed formally as labelled graphs.

Labelled graphs are written: $g = (V, E, \alpha, \beta)$

V: the *set of vertices* (voxels, ROIs, ICA components, sources...)

E: the *set of edges*

α: *vertex labelling function* (returns a scalar or vector for each vertex)

β: *edge labelling function* (returns a scalar, or vector for each edge)

...but comparing such graphs includes the weighted graph matching problem which is maybe NP-complete

A useful restriction

Brain graphs obtained from a fixed vertex-to-space mapping (e.g. functional or structural atlasing in fMRI) can be modelled by **graphs with fixed-cardinality vertex sequences**[1], a subclass of Dickinson et al.'s *graphs with unique node labels*[2]:

Fixed number of vertices for all graph instances:	$\forall i \ \lvert V_i \rvert = M$
Fixed ordering of the set (sequence) V:	$V = (v_1, v_2, \ldots, v_M)$
Scalar edge labelling functions:	$\beta : (v_i, v_j) \mapsto \mathbb{R}$
(optional) Undirected:	$\mathbf{A}^T = \mathbf{A}$

This is a very restricted (but still expressive) class of graphs

This limits the effectiveness of many classical methods for comparing general graphs (based on *graph matching*).

1 [Richiardi et al., ICPR, 2010] 2 [Dickinson et al., IJPRAI, 2004]

Undesirability of (exact) graph matching

Graphs G, H are isomorphic iff there exists a permutation matrix **P** s.t. $\mathbf{P}\mathbf{A}_g\mathbf{P}^T = \mathbf{A}_h$

Goal: recover an optimal permutation matrix $\hat{\mathbf{P}}$ to transform one graph into the other (map nodes).

Discrete optimisation[1]: search algorithm (A*, branch-and-bound...) + cost function (typically graph edit distance)

Continuous optimisation[2,3]: write $||\mathbf{P}\mathbf{A}_g\mathbf{P}^T - \mathbf{A}_h||_F$, relax constraints on **P**, optimise, then do credit assignment

The remaining cost after optimisation is a measure of distance between graphs

But we already know $\hat{\mathbf{P}} = \mathbf{I}$

To compare noisy brain graphs we're more interested in other techniques...

1 e.g. [Gregory and Kittler, SSPR, 2002]
2 e.g. [Zaslavskiy et al., ICISP, 2008]

3 interesting upcoming work by Josh Vogelstein (http://jovo.me)

Overview of approaches

Machine Learning

embeddings, kernels

matrix stats

topological properties

Stats

mass-univariate, non-parametric, relaxed/two-step

topological properties

Network science

community structures

Graph embedding

Graph embedding maps graphs to points in $\mathbb{R}^D$

With G a set of graphs, a graph embedding $\varphi : G \to \mathbb{R}^D$ maps graphs to D-dimensional vectors:

$$\varphi(g) = (x_1 \ \ldots \ x_D)^T$$

For brain graphs, we are generally interested in preserving edge label information

Vertex labels can be dropped because of the correspondence

Once we have vectors we can use any ML algorithm we want

“Direct” embedding

Use the upper-triangular part of the adjacency matrix[1,2,3]

$$\begin{pmatrix} (1,1) & \dots & (1,|V_i|) \\ & \ddots & \\ & & (|V_i|,|V_i|) \end{pmatrix} \longrightarrow \begin{pmatrix} (1,2) \\ \vdots \\ (|V_i|-1,|V_i|) \end{pmatrix}$$

$$\mathbf{A}_i \in \mathbb{R}^{|V_i|\times|V_i|} \qquad \mathbf{a}_i \in \mathbb{R}^{\binom{|V_i|}{2}\times 1}$$

“Cursed” representation, but generally a competitive baseline (at least with ~100 vertices, fMRI)

Combines whole-brain (global) and regional (local) aspects

Decision is on the full graph

Each edge has a weight: discriminative information content of edges can be localised and it is easy to show brain-space maps

1 [Wang et al., MICCAI, 2006]
2 [Craddock et al., MRM, 2009]

3 [Richiardi et al., ISBI 2010]
[Richiardi et al., ICPR 2010]
[Richiardi et al., NeuroImage, 2011+12]

Application: fMRI/MS diagnosis

Can resting-state functional connectivity serve as a surrogate marker of MS ?

Data: **14 HC, 22 MS**, 450 volumes @ TR 1.1s, 3T scanner

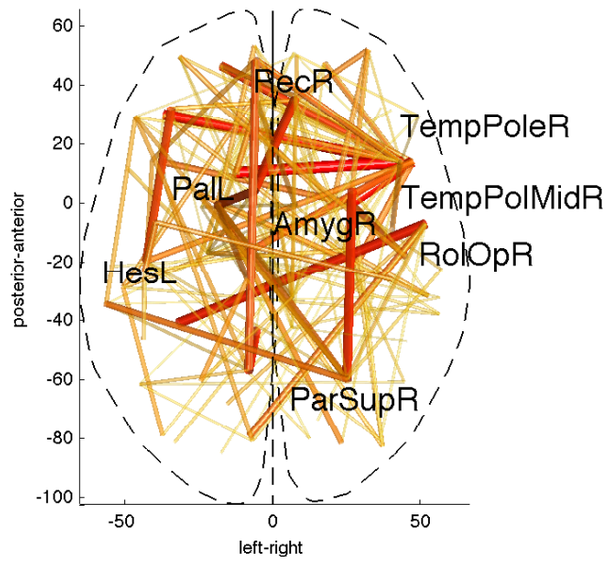
Graph: **AAL 90**, 0.06-0.11 Hz, winsorising 95 % , **Pearson correlation**

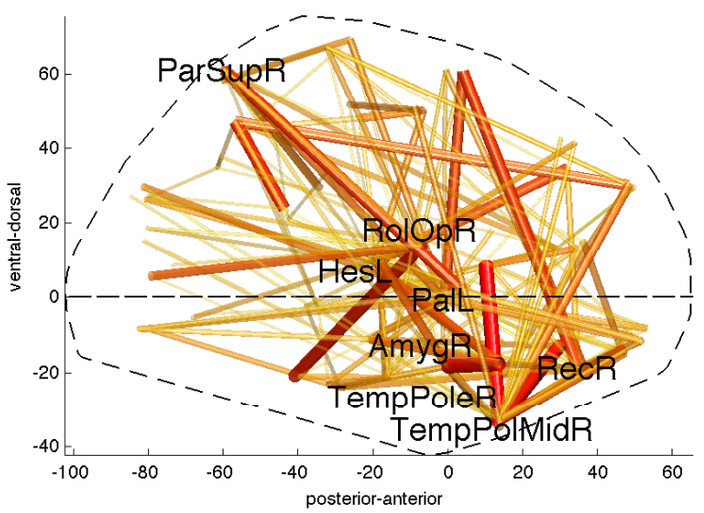
Embedding: **direct,** no FS

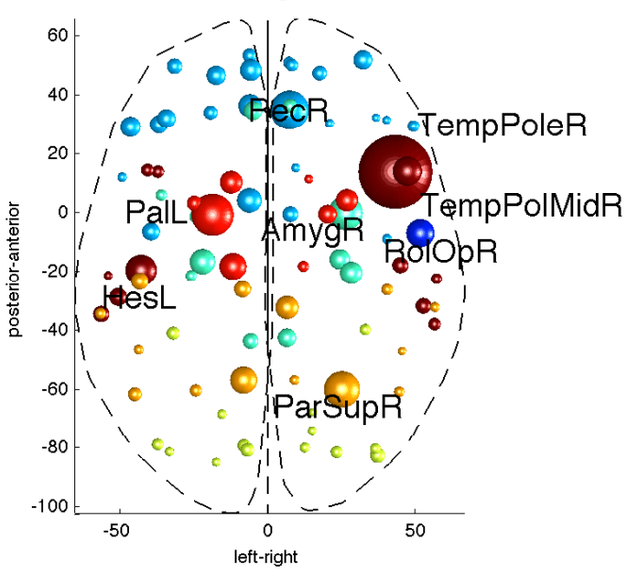
Classifier: FT forest

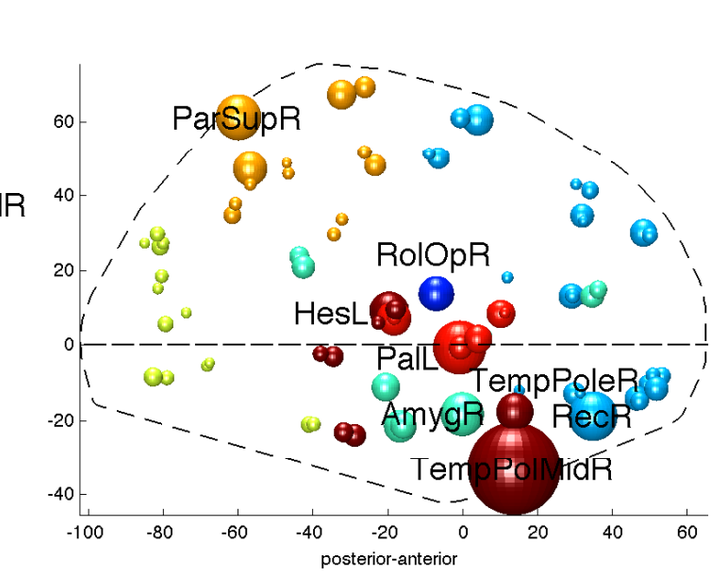
Performance: LOO
CV: **82% sens** (CI 62-93%),
86% spec (CI 60-96%)

Mapping: Label permutation testing: 4% of all edges significantly discriminative









[Richiardi et al., NeuroImage, 2012]

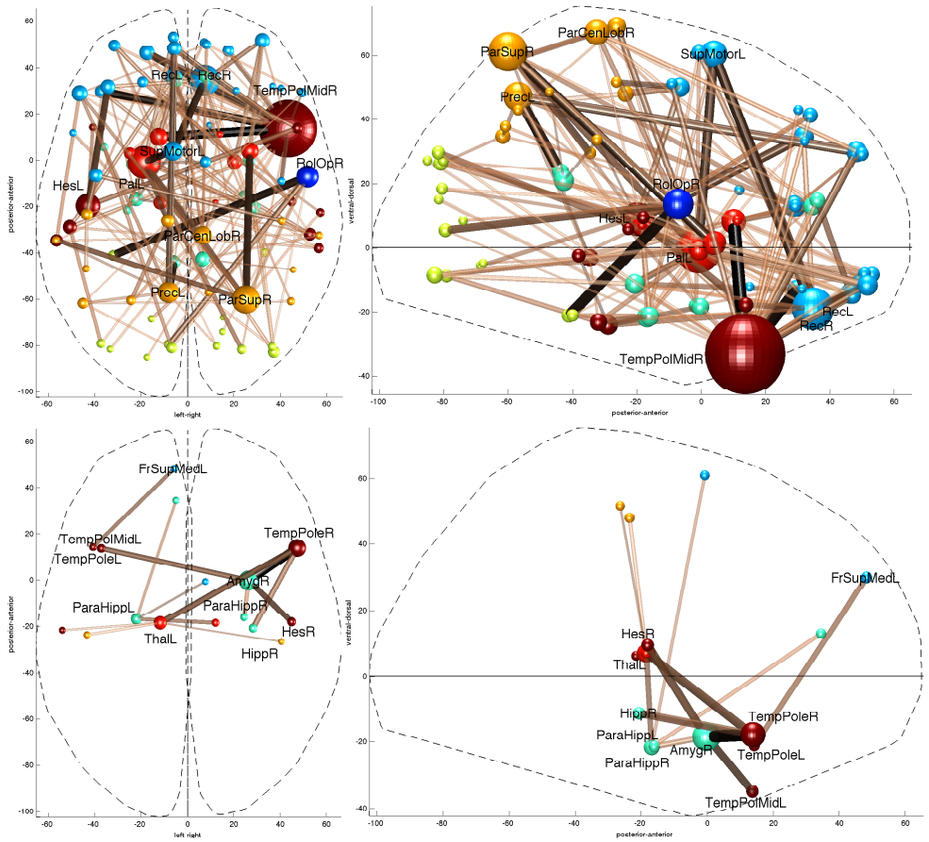
MS(2): Link with structure

Connectivity alterations relate to WM lesions

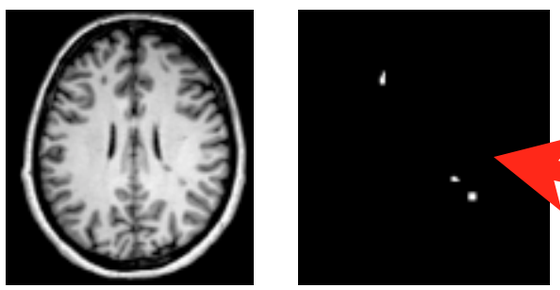
Split discriminative graph in reduced (C+) and increased (C-) connectivity

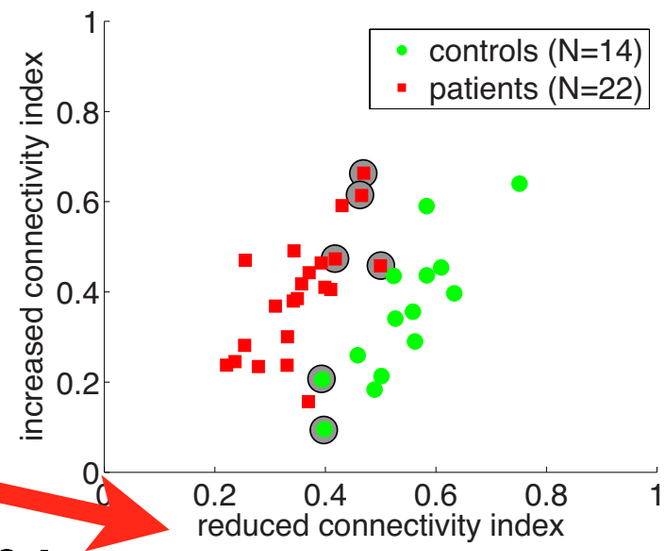
For each subject compute summary index of discriminatively reduced connectivity

$$\mathrm{nRCI}^s = \frac{1}{||\boldsymbol{\rho}^s||_1} \sum_{i \in C_-} w_i^s \rho_i^s$$



Correlate with WM lesion load





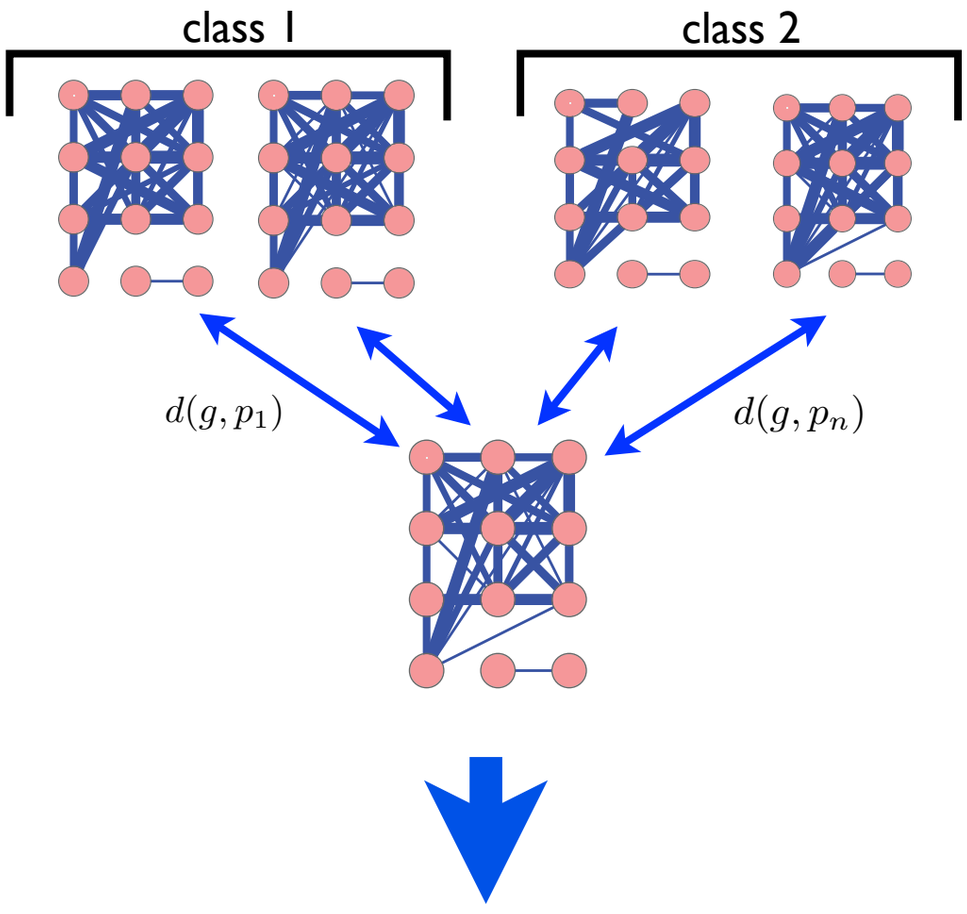
r=0.61, p < 0.001

[Richiardi et al., NeuroImage, 2012]

Pairwise graph (dis)similarity

We can also define dissimilarity functions[1] d(g,h) or kernels k(g,h) operating on graphs, that return a scalar.

Principle



Embedding vector

$$\varphi_n^{\mathcal{P}}(g) = (d(g, p_1), \ldots, d(g, p_n)) \in \mathbb{R}^n$$

Example dissimilarity function - penalised edge label dissimilarity

(special case of weighted Graph Edit Distance (wGED))

Edge label disssimilarity

$$\delta(e_{ij}, e'_{ij}) = \begin{cases} |\beta(i,j) - \beta'(i,j)| & e_{ij} \in E, e'_{ij} \in E' \\ K & otherwise \end{cases}$$

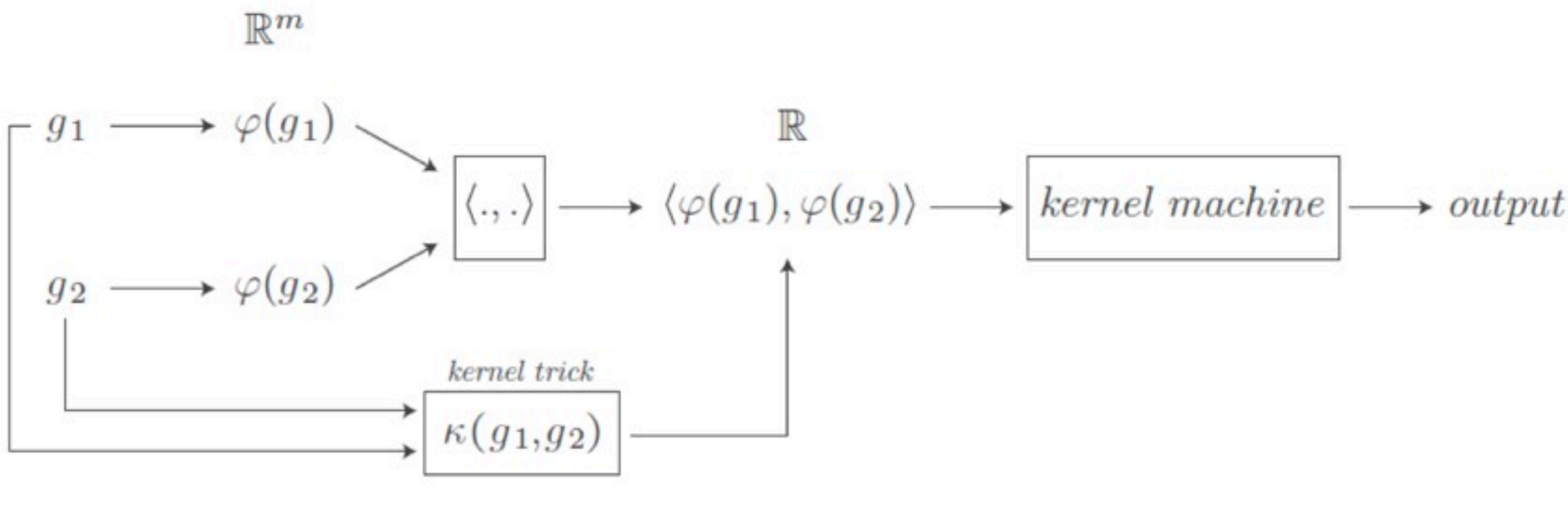
Graph dissimilarity

$$d(g,p) = \sum_{i=1}^{|E|} \sum_{j=i+1}^{|E|} \delta(e_{ij}, e'_{ij})$$

$d(g,p) = \frac{1}{2}||\mathbf{a}_g - \mathbf{a}_p||_1$ (if no missing edges)

1 [Richiardi et al., ICPR 2010]
based on [Riesen & Bunke, Int. J. Pat. Rec. Artif. Int. 2009]

Kernel trick on graphs



Leverage advances in kernel methods[1,2]

No mathematical structure other than the existence of a (valid) kernel function is necessary to use kernel machines on graphs

Many types of graph kernels applicable to brain graphs: convolution, walks/paths, ...

1 [Schölkopf & Smola, 2002] 2 [Shawe-Taylor & Cristiniani,2004]

illustration: Horst Bunke

Direct embedding and kernels

Link between direct graph embedding and graph kernels: kernelisation of a weighted GED

With $\mathbf{a}_1$,$\mathbf{a}_2$ the direct embeddings of graphs g_1,g_2, we know $d(g_1, g_2) = ||\mathbf{a}_1 - \mathbf{a}_2||_1$ is a valid weighted GED.

We can trivially obtain a (non-valid) kernel with

$$k(g_1, g_2) = e^{-d(g_1, g_2)}$$

We can also obtain a valid kernel, e.g. Von Neumann diffusion kernel[1]

$$\mathbf{B}_{ij} = max(d(g_m, g_n)) - d(g_i, g_j)$$

$$\mathbf{K} = \sum_{m}^{\infty} \lambda^m \mathbf{B}^m \ , 0 < \lambda < 1$$

1 [Kandola et al., NIPS, 2002]

Convolution graph kernels

Convolution kernel[1]: Similarity-of-graph from similarity-of-subgraph

1. Define valid kernels on substructure/subgraph

2. Combine by sum-of-products (PD functions are closed under product, PD matrices are closed under Hadamard product)

$$k(g_1, g_2) = \sum_{g1_p \in g_1, g2_p \in g_2} \prod_t k_t(g1_p, g2_p)$$

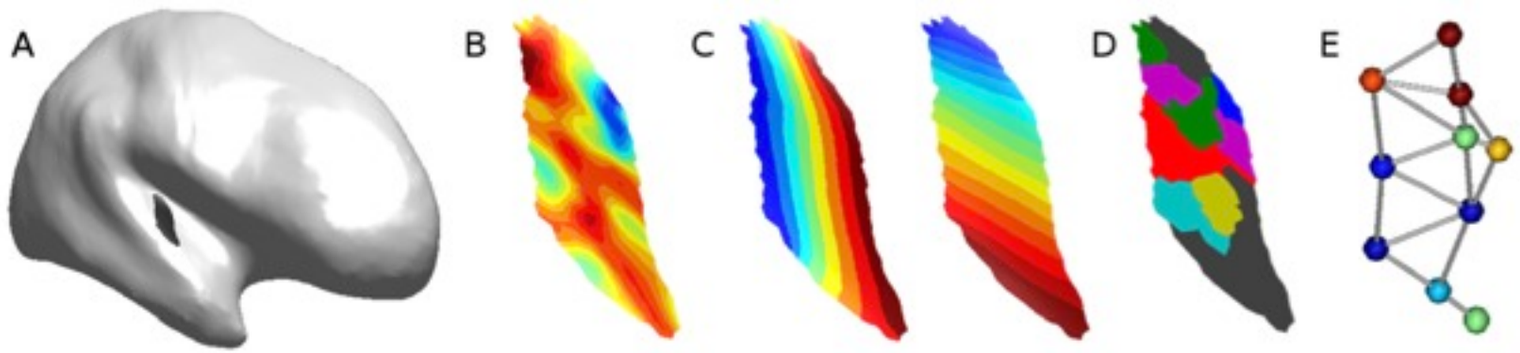
Many ways to define subgraphs

Can use modality-specific k_t

1 [Haussler, USCS TR, 1999]

Application: fMRI/auditory cortex

Multimodal graph



Vertices: auditory cortex ROIs

Vertex labels: vector: (mean activation, xpos_mean, ypos_mean)

Edge set: spatially adjacent regions (binary labels)

Classifier design

Gaussian kernels for vertices, linear for edges

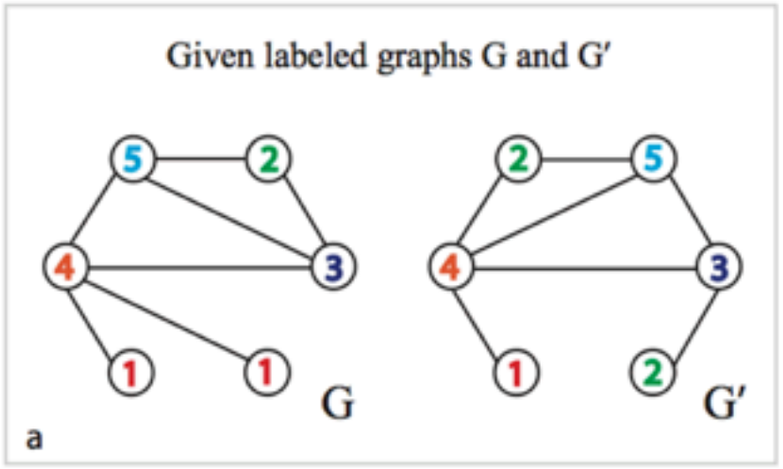
Subgraphs: paths of length two

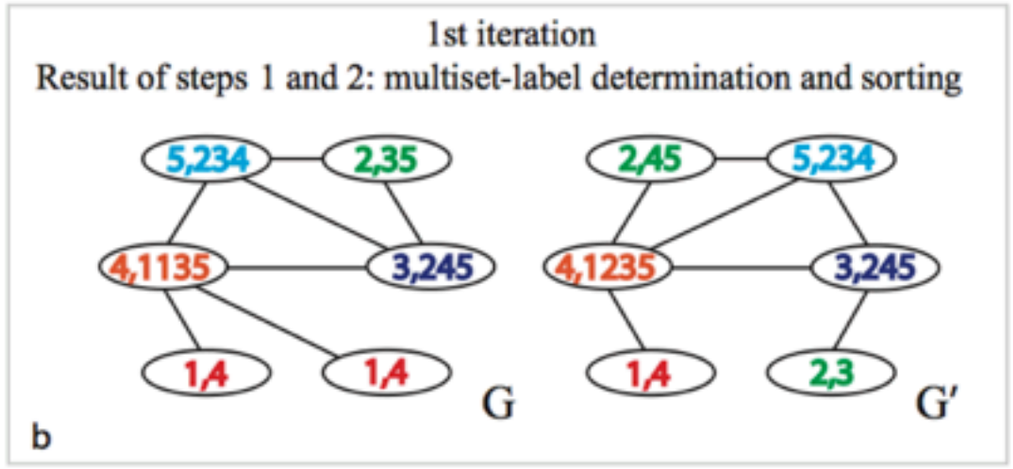
Results

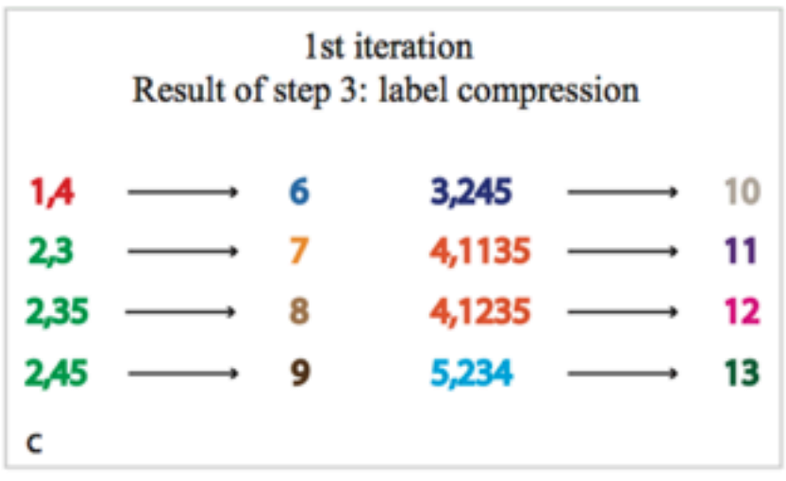
Tonotopic decoding with 5 frequencies (300-4000 Hz), N=9, subparcellation of Heschl gyri: 36-45% accuracy (chance: 20%)

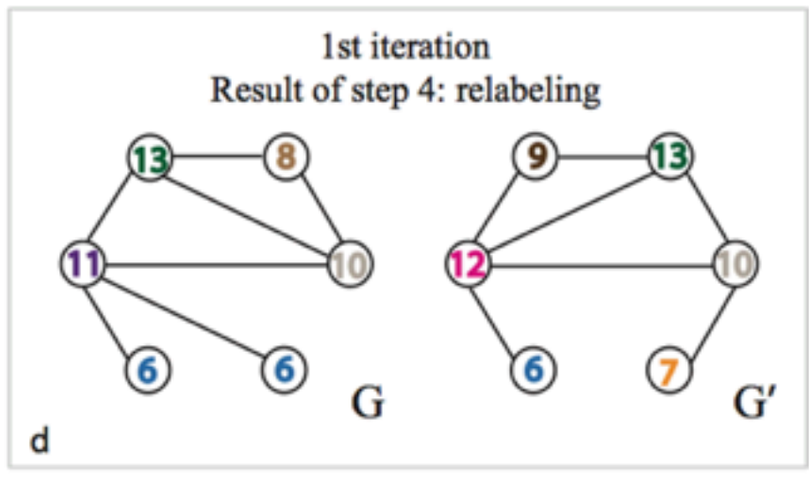
[Takerkart et al., MLMI, 2012]

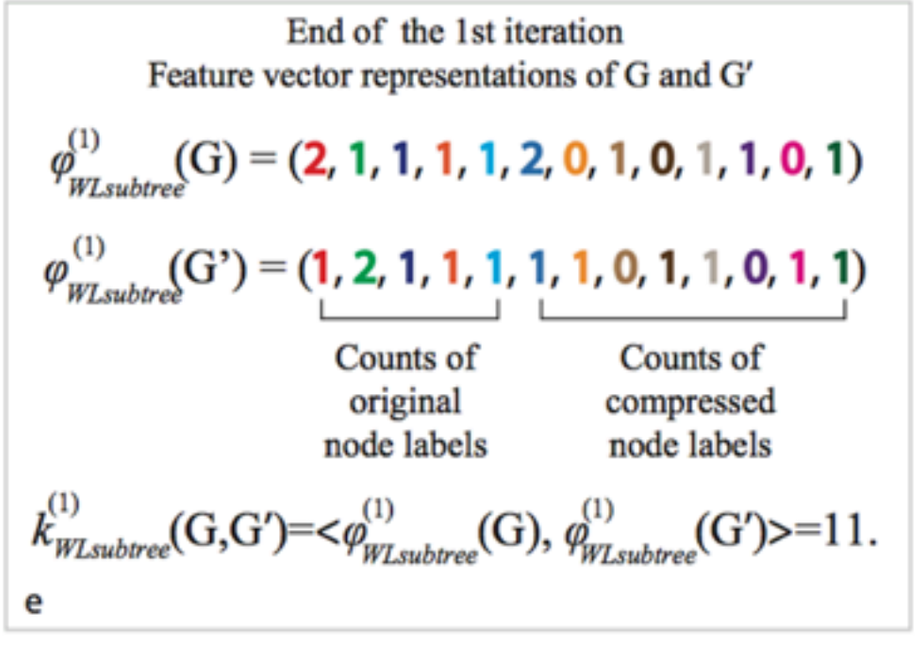
Weisfeiler-Lehman subtree kernel











[Shervashidze et al., JMLR, 2010]

Application: fMRI/decoding house vs face

fMRI brain graph

Data: Haxby, N=6, 12 runs, 9 volumes / category / run, no alignment between subjects

Vertices: voxels in ventral temporal cortex

Vertex labels: degree

Edge set: thresholded correlation (?)

Results

66% accuracy (±12%) with non-category specific mask.
Better on synthetic data.

[Vega-Pons & Avesani, PRNI, 2013]

ML summary: pros and cons

Direct embedding:

+ satisfactory prediction on several datasets
+ easy mapping of discriminative pattern
- cursed representation (O(D^2))

Dissimilarity embedding:

+ low-dimensional representation (O(N))
- setting costs is not trivial
- performs worse than direct embedding on most small-graph datasets

Graph/vertex attribute embedding:

+ low-dimensional representation (O(|V|))
+ interpretable in terms of graph properties
- many attributes are weakly discriminative

Graph kernels

+ Well suited for multimodality, custom similarity measures, domain-specific knoweldge
+ Well suited for large graphs (kernel trick - avoid explicit inner product)
- Generic graph kernels may not work well on brain graphs

Overview of approaches

Machine Learning

embeddings, kernels

matrix stats

topological properties

Stats

mass-univariate, non-parametric, relaxed/two-step

topological properties

Network science

community structures

Statistical testing on graphs

Brain graphs have challenging properties

- Non-independence of edge labels - non-IID data
- High dimensional edge space ($O(|V|^2)$)
- Structured adjacency matrix (SPD)

Choice of method depends on scale of interest

- Whole-brain: graphwise testing
- “Subnetwork of regions”: subgraphwise testing
- Two regions: edgewise testing

Graphwise: Mantel test

Test statistic[1]: strength of relationship between two matrices **X**,**Y**

$$z = \sum_{i,j \ i \neq j} \mathbf{X}_{ij} \mathbf{Y}_{ij}$$

Often use normalised version $z' = cor(vec(\mathbf{X}), vec(\mathbf{Y}))$

Test procedure: permutation of rows&cols

Can be used directly on adjacency matrix of brain graphs

$$z' = cor(vec(\mathbf{A}_1), vec(\mathbf{A}_2)) = cor(\mathbf{a}_1, \mathbf{a}_2)$$

Null hypothesis: there is no relationship between the topology of the two brain graphs

1 [Mantel, Cancer research, 1967], with principle from [Daniels, Biometrika, 1944]

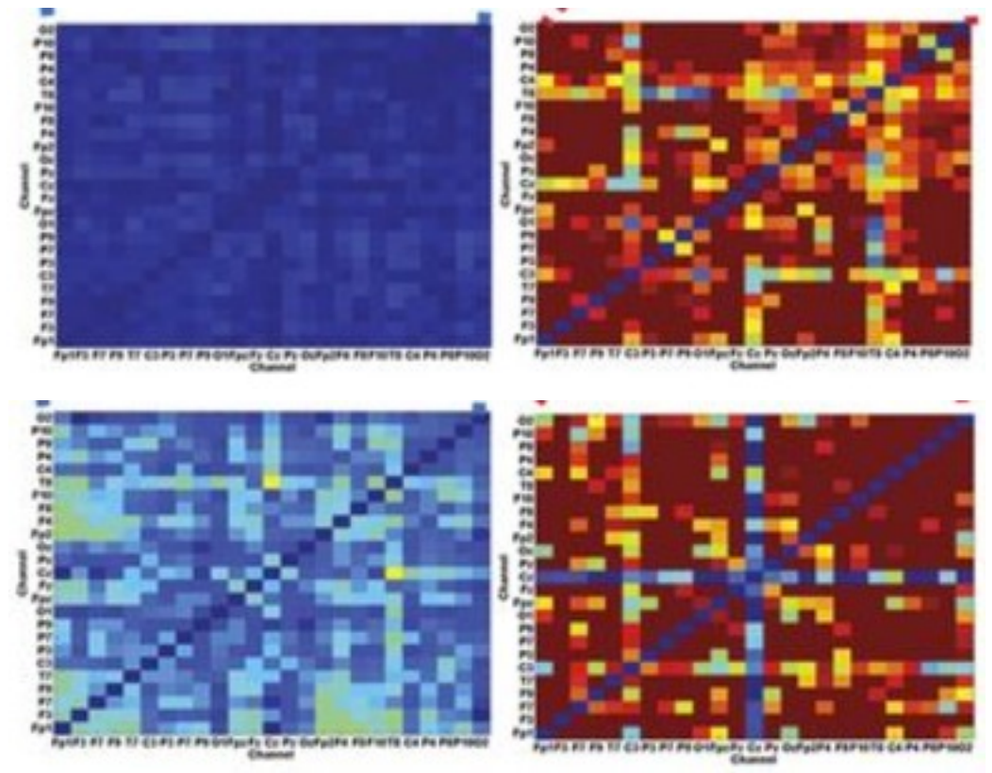
Applications: EEG/pre-term babies

Goal: compare spatial correlations between low-mode and high-mode (bursts) EEG activity in pre-term and full-term babies

Data: 10 FT, 11 PT, sleep, 5 mins selected

Vertices: 25 Channels (remontaged)

Edge labels: linear regression coefficient for each re-quantized, censored bivariate amplitude pair. Thresholded via surrogate data.



Results

Low/high difference in full-term babies, not in pre-term. Network communication is predominantly bursty in babies.

Pre-term/full-term differences in the low mode. Low-mode activity is spatially reorganised during gestation.

[Omidvarnia, Cerebral Cortex, 2013]

Edgewise: mass-univariate + MTP

The most commonly used approach in the literature is mass-univariate

If edge labels given by corr, Gaussianise: $A'_{ij} = tanh^{-1}(A_{ij})$

Test statistic: (typically) two-sample t-test

Test procedure: (typically) FDR

This has many drawbacks

High-dimensionality means we are at risk of false positives from multiple comparisons, so need MTP

Edges and their labels are not independent from the vertex they are attached to (must use an MTP for dependent tests)

Mass-univariate, may miss subthreshold covariations

Application: fMRI/brain state decoding

Goal: classify movie-watching vs resting from fMRI connectivity graph

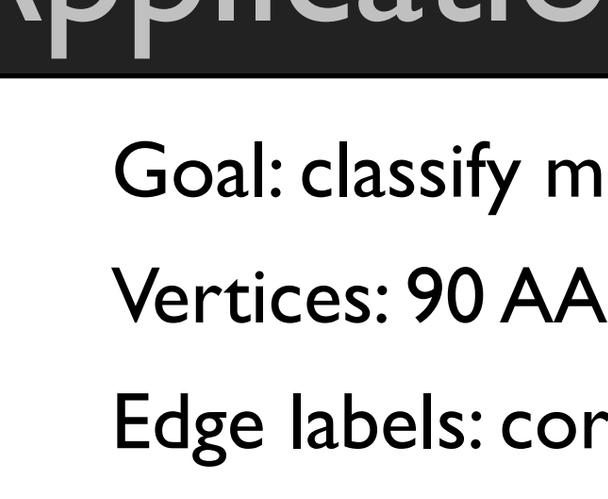
Vertices: 90 AAL regions

Edge labels: correlation of wavelet coeffs in 0.06–0.11 Hz

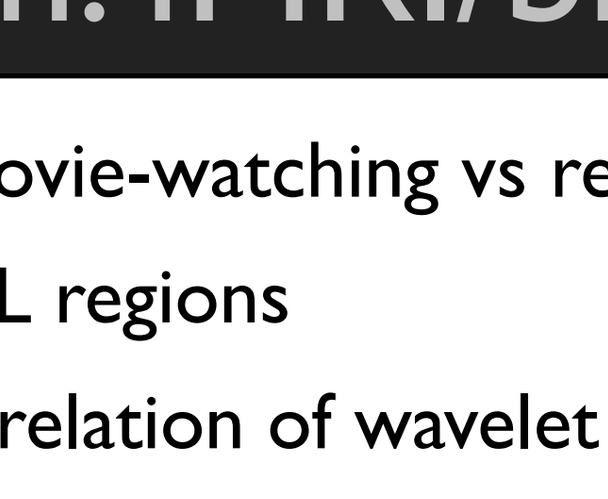
Results

23/4005 edges significant (cuneus + occipital lobe), superior temporal

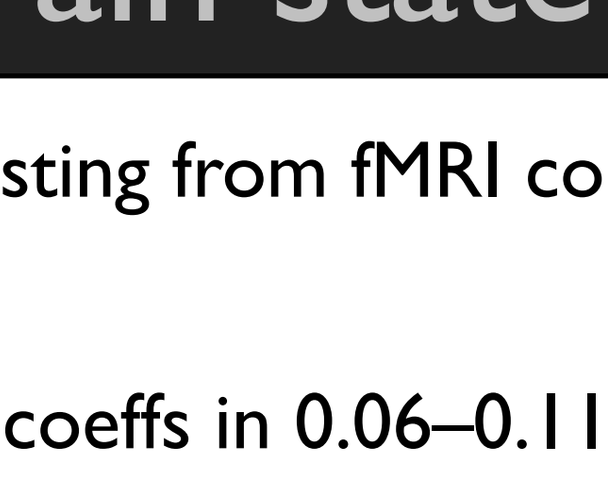
Edges found are a subset of those found with multi-band ML approach



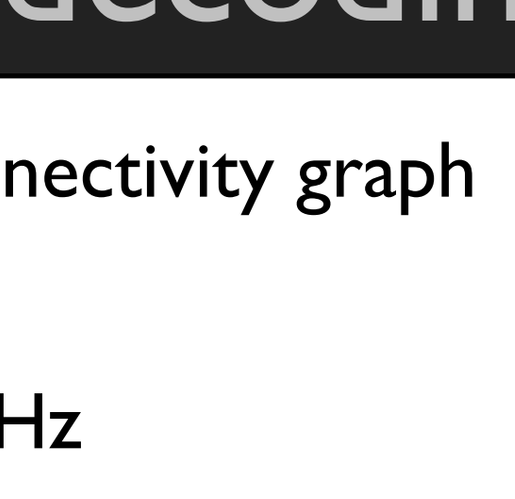
rest avg



movie avg



t test p-val



significant diffs
(5% FDR)

[Richiardi et al., NeuroImage, 2011]

Subgraphwise: two-step tests

Exploit positive dependency between tests

Same idea as Gaussian Random Field (smoothness), but applied to irregular domain of graphs

Group edges (tests) by some criterion

Zalesky's Network-based statistic[1]

Apply mass-univariate testing, threshold, compute connected components, record sizes

Permute group labels, recompute component sizes, get p-value

Other, more general variants exist with various ways of choosing subgraphs[2]

1 [Zalesky, NeuroImage, 2010] 2 [Meskaldji, PLoS one, 2011]

Application: fMRI/Schizophrenia

Goal: discriminate patients with Schizophrenia

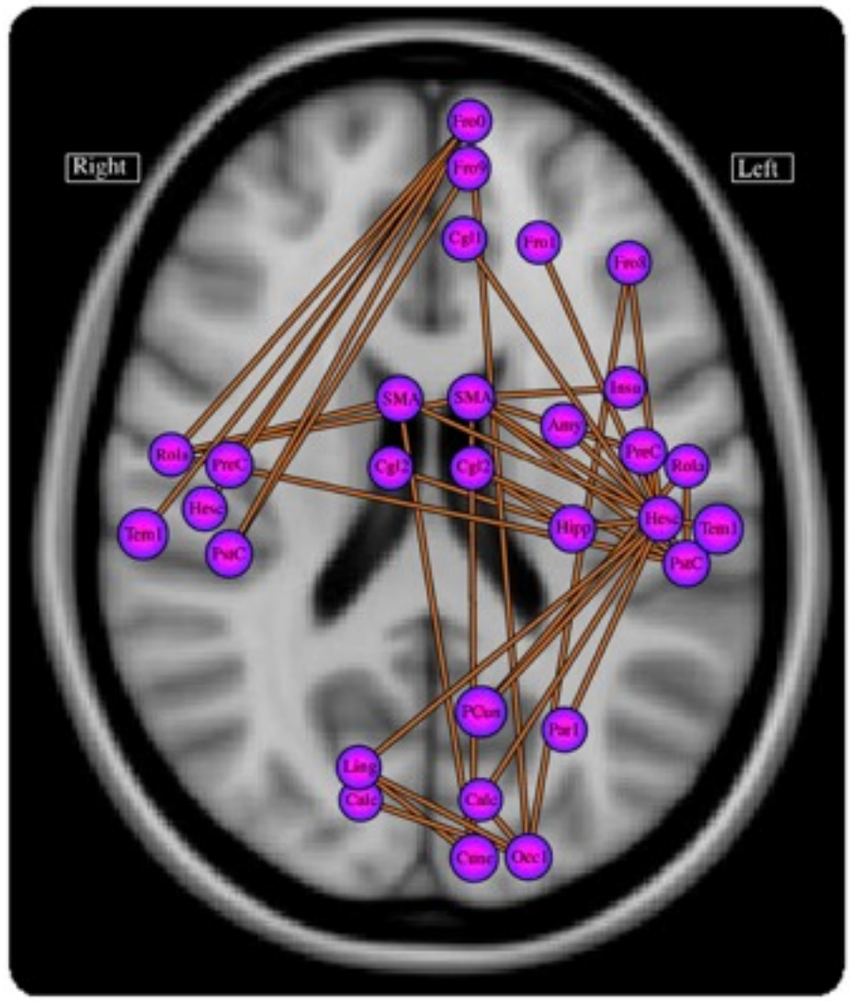
Data: 15HC, 12 SZ, 1.5T, TR=2s, rest, 17 mins

Vertices: AAL 74

Edge labels: wavelet correlation, 0.03-0.06 Hz

Results

(a) FDR ($q = 10\%$) (b) NBS ($p = 0.037$)



[Zalesky et al., NeuroImage, 2010]

Stats summary: pros and cons

Graphwise/Mantel:

+ Simple procedure, (normalised) test statistic is clear
+ Cross-modal testing
- No mapping

Subgraphwise/two-step:

+ Elegantly deal with multiple comparisons
+ Relevant scale for inference to study distributed processes
+ Mapping jointly significant edges / subgraph
- Null hypothesis may be hard to interpret

Edgewise/mass-univariates

+ Low-dimensional representation (O(|V|))
+ Interpretable in terms of graph properties
- Many attributes are weakly discriminative

Overview of approaches

Machine Learning

embeddings, kernels

matrix stats

topological properties

Stats

mass-univariate, non-parametric, relaxed/two-step

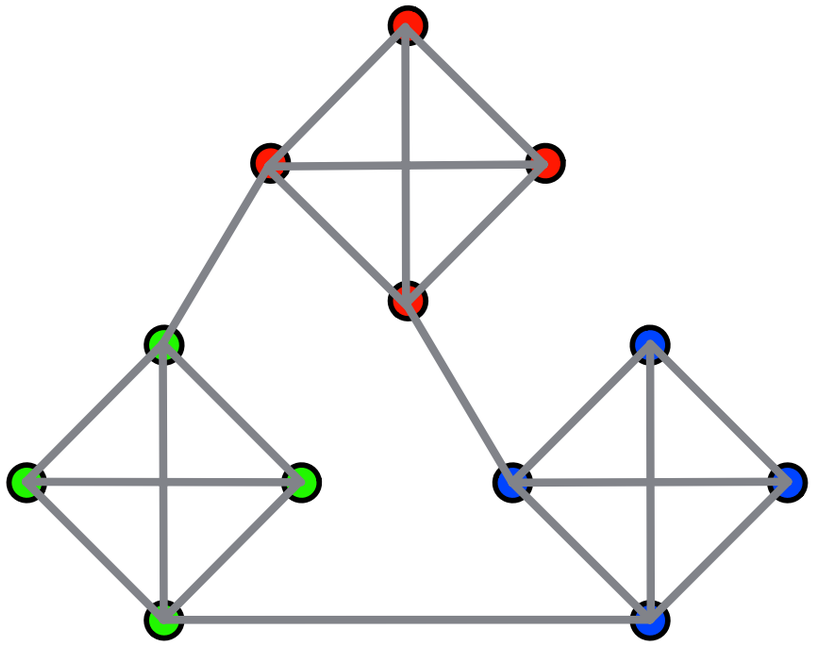
topological properties

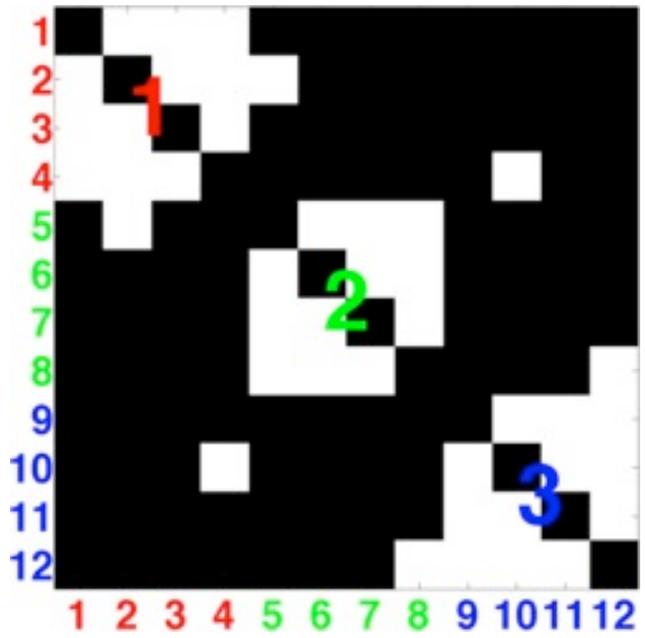
Network science

community structures

Network science techniques

Brain graphs have identifiable subgraphs ("modules", "communities") in several modalities





The partition into communities can be used to compare brain graphs between subjects or modalities at various scales

Whole-brain: graphwise community structure

"Subnetwork of regions": individual communities

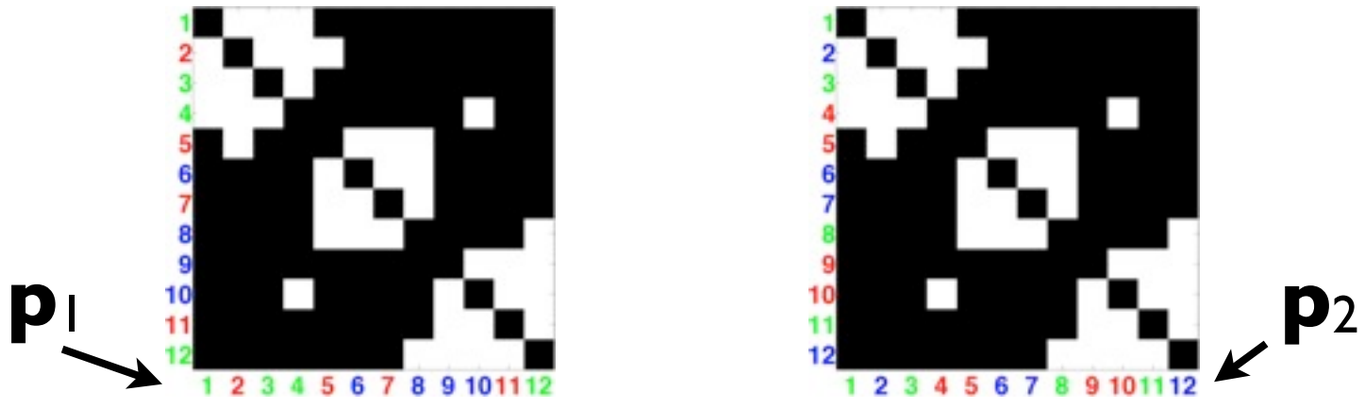
Single region: community membership (not shown)

Graphwise: NMI between partitions

Similarity between community assignments of two graphs as a proxy of their similarity

This is the same problem as comparing clusterings

Assignment of vertices to communities in $\mathbf{p}_i \in \mathbb{N}^{|V|}$



Measure similarity between assignment vectors, e.g.[1,2]

$$NMI(\mathbf{p}_i, \mathbf{p}_j) = \frac{2I(\mathbf{p}_i, \mathbf{p}_j)}{I(\mathbf{p}_i, \mathbf{p}_i) + I(\mathbf{p}_j, \mathbf{p}_j)}$$

from normalised table counts

Permute group labels and recompute to obtain p-value

1 [Alexander-Bloch et al., NeuroImage, 2012] 2[Ambrosen et al., PRNI, 2013]

Application: fMRI/Schizophrenia

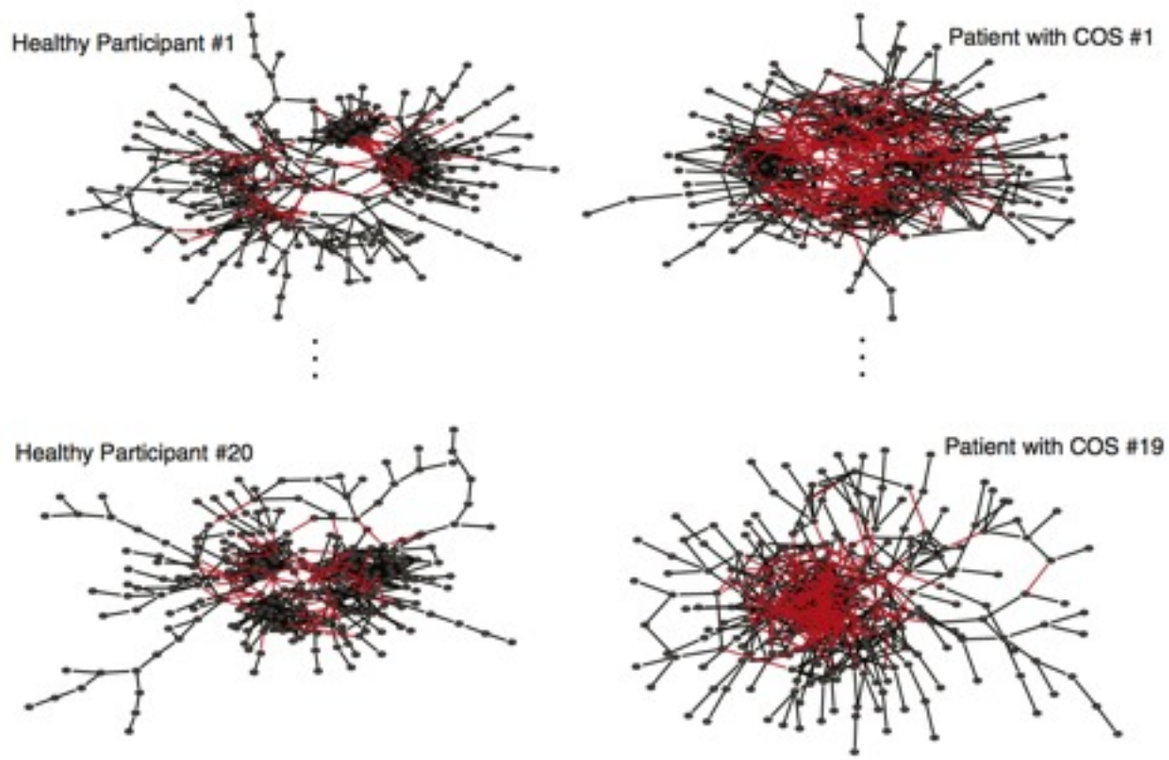
Goal: discriminate patients with schizophrenia

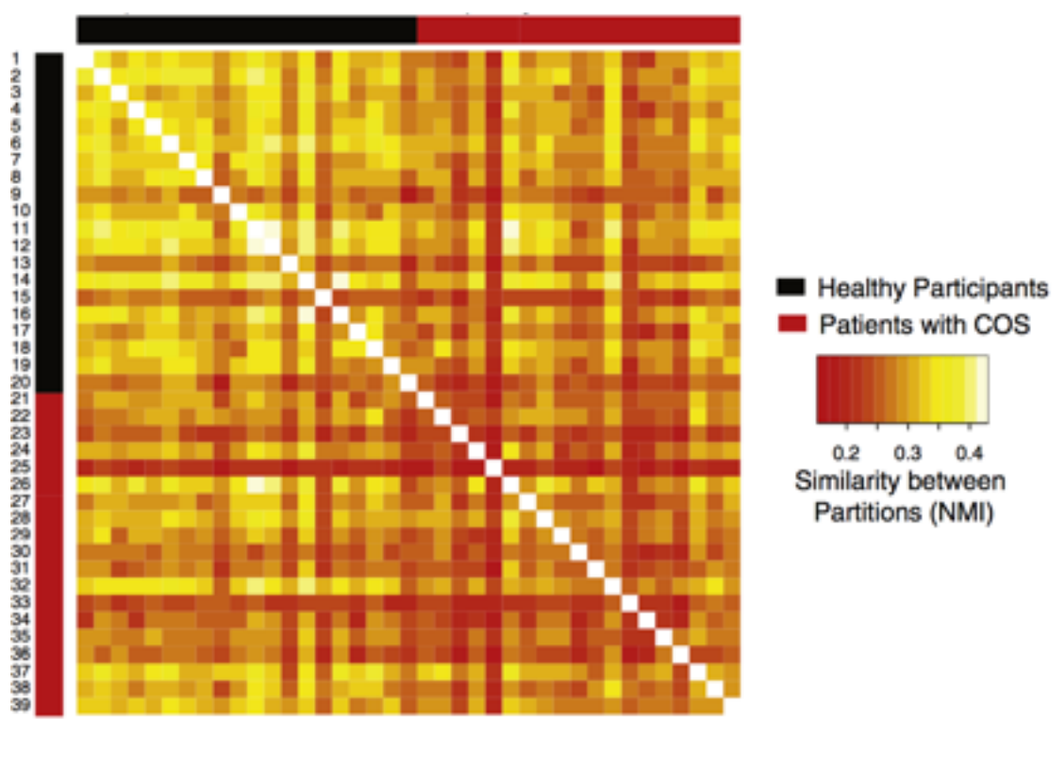
Data: 23 HC, 23 SZ, TR=2.3s, rest, 2x3 min (144 points)

Vertices: Subparcellated Harvard-Oxford, 278 regions

Edge labels: thresholded and binarised absolute wavelet correlation, 0.05-0.1Hz

Results



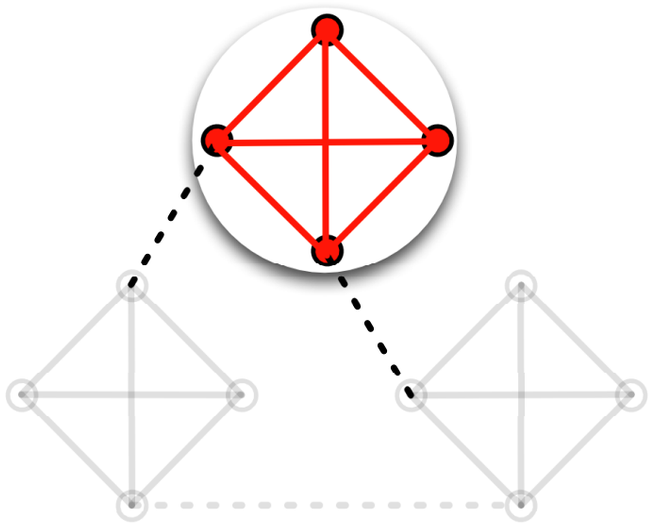


[Alexander-Bloch et al., NeuroImage, 2012]

Subgraphwise: significance of communities

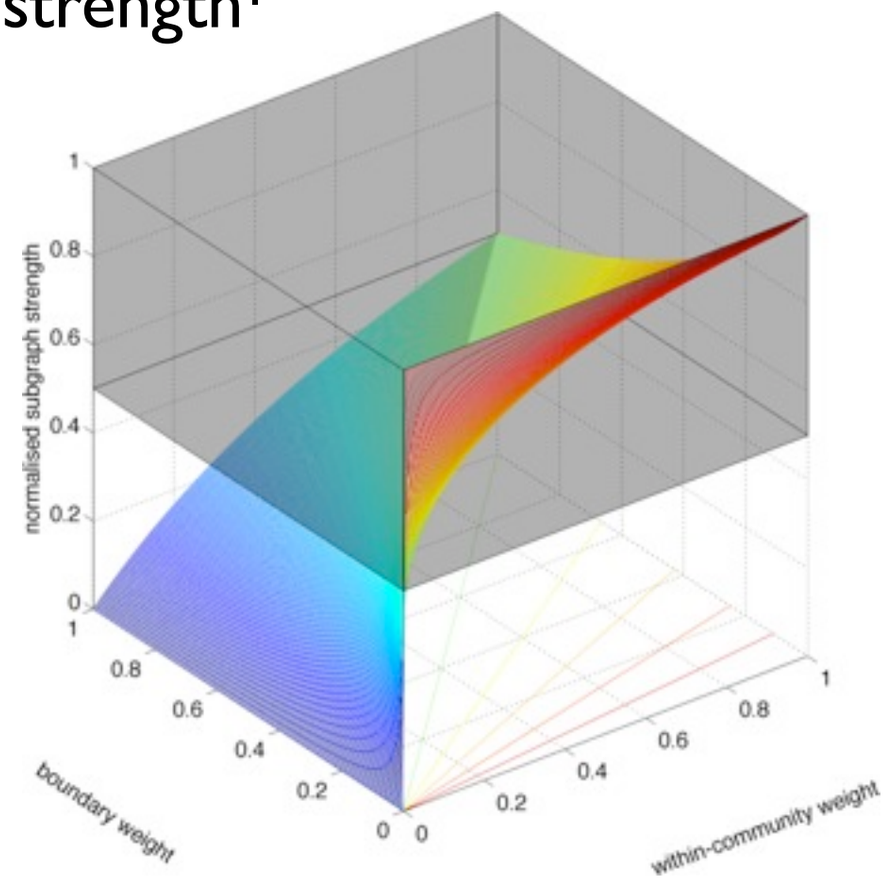
Are communities significant in both graphs?

Test statistic: normalised community strength[1]



$$S_c = \frac{W}{W + B}$$

$$S_c = \frac{\sum_{i \in V_c, j \in V_c} A_{ij}}{\sum_{i \in V_c, i \sim j} A_{ij}}$$



Test procedure: permutations of the partition vector. Null hypothesis: any other group of $|V_c|$ vertices can have as high a value of S_c.

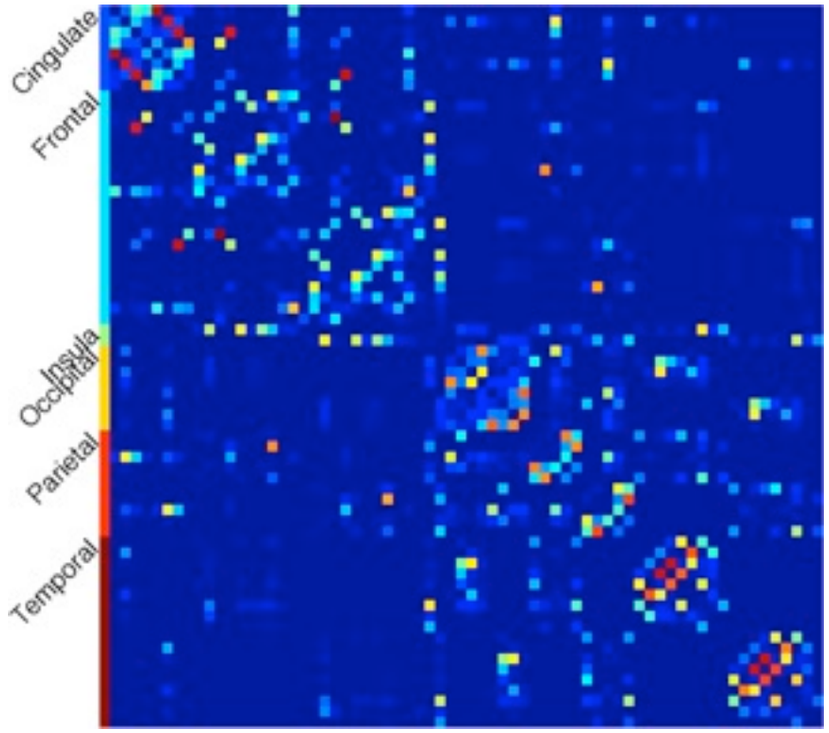
This can be used across modalities.

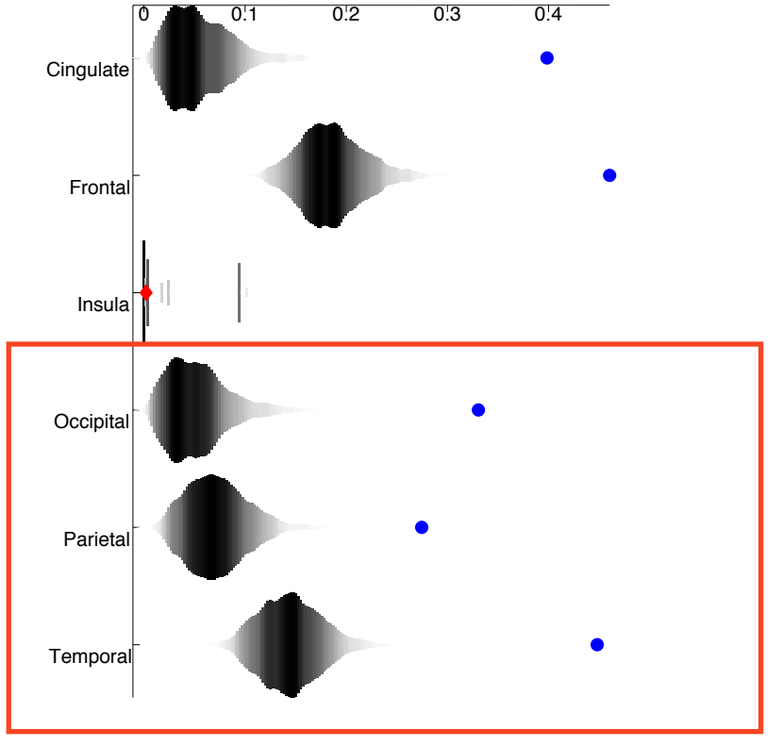
1 [Richiardi et al., PRNI, 2013]

Application: multimodal correspondance

structural connectivity

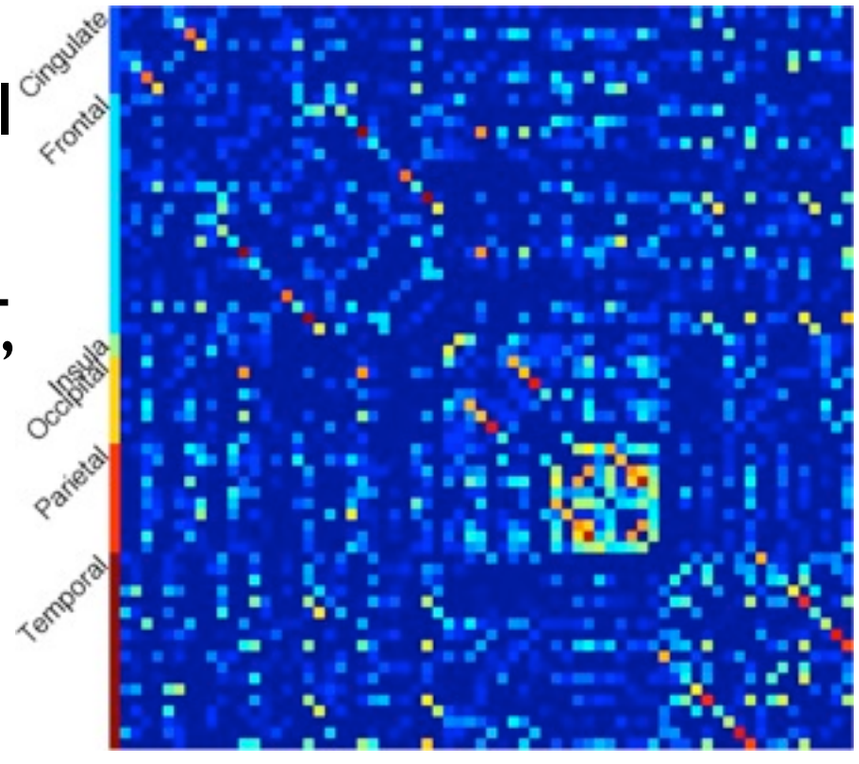
DWI, 1.5T, 30 directions

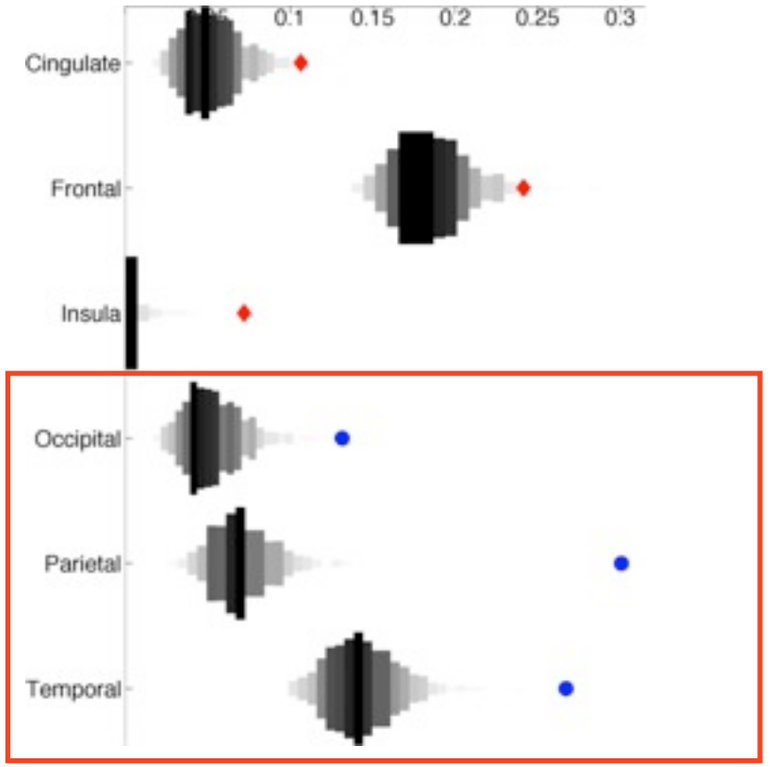




"morphological connectivity"

structural, 1.5T, 1mm voxels





[Richiardi et al., PRNI, 2013]

Network science summary: pros and cons

Graphwise/NMI:

+ Empirically works well (also on DTI[1], not shown)

+ Amenable to cross-modality testing

- Many parameters upstream: community detection algorithm, null model, etc.

Subgraphwise/community significance:

+ Interpretable quantity (weak-sense community)

+ Usable for cross-modality testing

- Sensitivity / specificity tradeoff yields false positives

1 [Ambrosen et al., PRNI, 2013]

A few links: ML - stats

Machine Learning

embeddings, kernels

matrix stats

Stats

mass-univariate, non-parametric, relaxed/two-step

Network science

community structures

Linear kernel yields the Mantel statistic

Given the direct embedding $\mathbf{a}_m$ of a graph m,

Normalise $\mathbf{a}'_m = \frac{\mathbf{a}_m - \boldsymbol{\mu}}{||\mathbf{a}_m||}$

Now the normalised Mantel test statistic $z' = \langle \mathbf{a}'_n, \mathbf{a}'_m \rangle$ is a valid kernel (linear kernel)

Dual formulation of linear SVM $f(\mathbf{a}'_m) = \sum_n \alpha_n y_n \langle \mathbf{a}'_n, \mathbf{a}'_m \rangle + \hat{b}$

In high-dim case $\forall n\ , \alpha_n \neq 0$, thus SVM is a linear combination of correlations between direct graph embeddings of all graphs in the training set. Thus both approaches intrinsically use the same measure of similarity.

	data	**class labels**
Mantel	2 graphs	unknown
SVM	all training set	available

A few links: ML - network science

Machine Learning

embeddings, kernels

topological properties

Stats

mass-univariate, non-parametric, relaxed/two-step

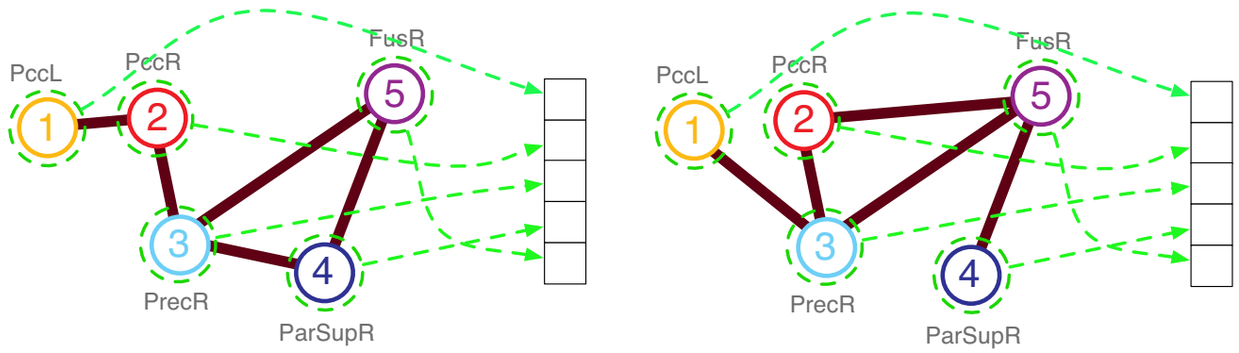
Network science

community structures

Machine learning on topological properties

We can view topological properties as "deep" feature extractors

Represent each graph and/or vertex by a vector of graph and / or vertex properties[1,2,3]



subject 1 subject 2

Intermediate step between simple embeddings and graph kernels

No complete invariants (degeneracy): use several properties[4,5]

Performance can be relatively high, especially for large graphs

1 [Cecchi et al., NIPS, 2009] 3 [Bassett et al., NeuroImage, 2012]
2 [Richiardi et al., PRNI, 2011] 4[Li et al., MLG, 2011] 5 [Bonchev et al, J Comput Chemistry 1981]

Application: fMRI/prediction from preparation

Goal: predict color/motion judgement errors, and which task the subject is preparing for, from preparation phase

Data: 10 HC, 72 x 3 conditions, TR=2s

Vertices: 70 regions from searchlight on beta map

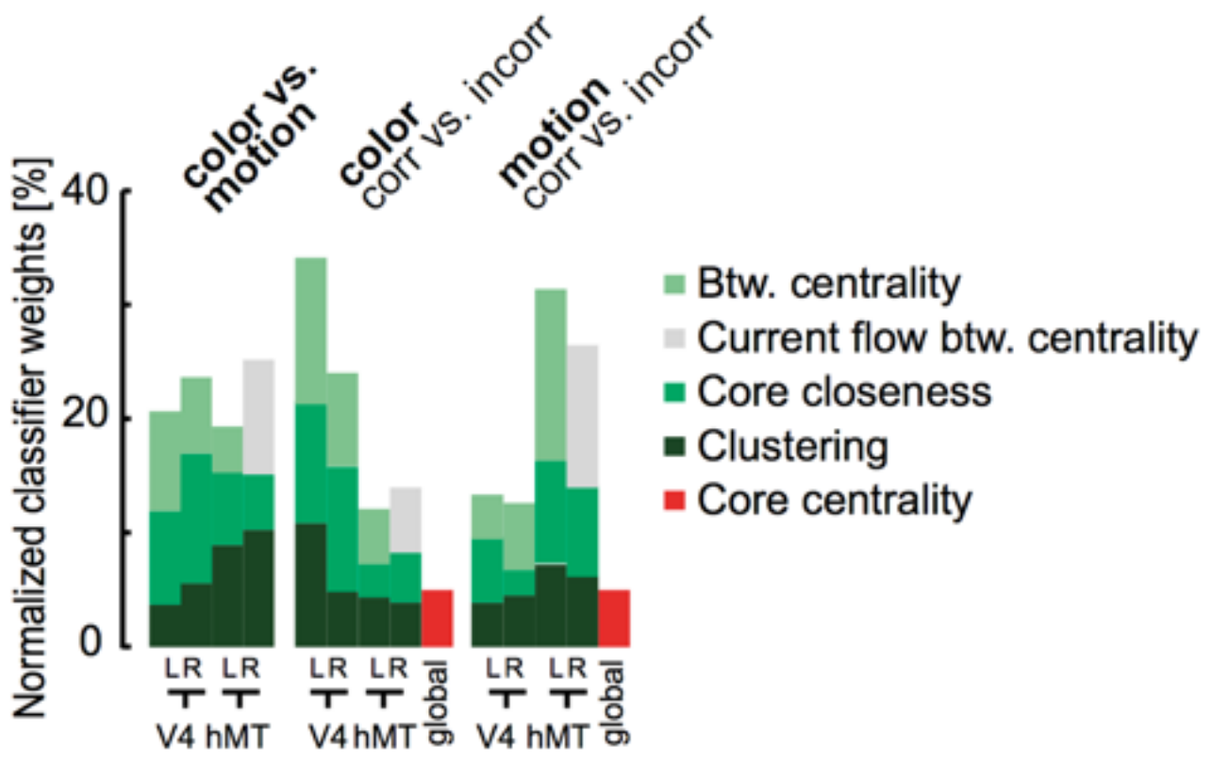
Edge labels: concatenated trials, wavelet 0.06-0.12 Hz, thresholding

Embedding: 10 vertex properties + 11 graph properties (711 dimensions)

Results

Can discriminate task and errors well above chance

Change of graph topology in V4 (color-sensitive) and hMT (motion-sensitive) is predictive of errors



[Ekmann et al., PNAS, 2012]

A few links: stats - network science

Machine Learning

embeddings, kernels

topological properties

Stats

mass-univariate, non-parametric, relaxed/two-step

Network science

community structures

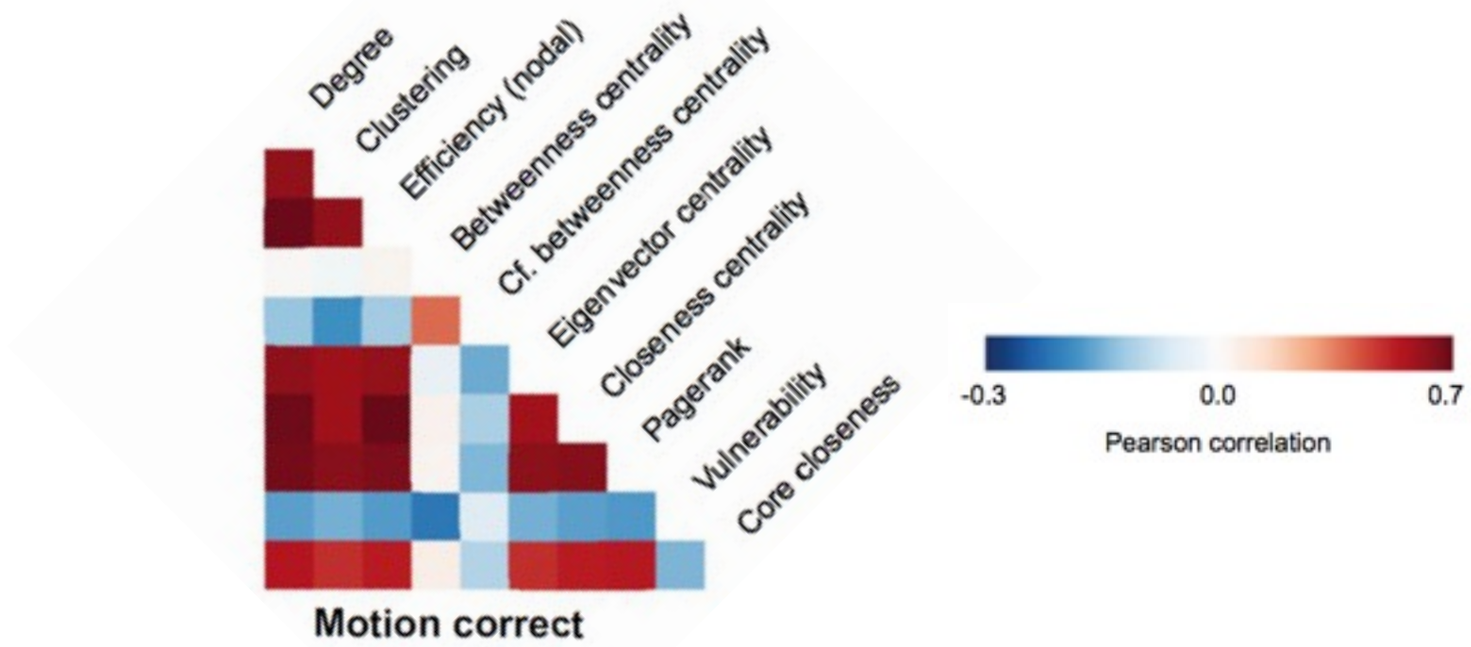
Statistical testing with topological properties

Hypothesis testing on graph/vertex properties is the most common approach to graph comparison in the neuroimaging literature[1]

This allows freedom in the choice of spatial scale

Multiple comparison problem less severe than edge stats

But...many graph properties are correlated[2,3,4]



1 see e.g.[Achard & Bullmore, PLoS CompBiol, 2007]
2 see e.g.[Lynnal et al., J. Neurosci. , 2010], 3 [Alexander-Bloch et al., Front. Syst. Neurosci., 2010]
4 [Ekmann et al., PNAS, 2012]

Application: MEG/cognitive load

Goal: study graph topology under varying cognitive load

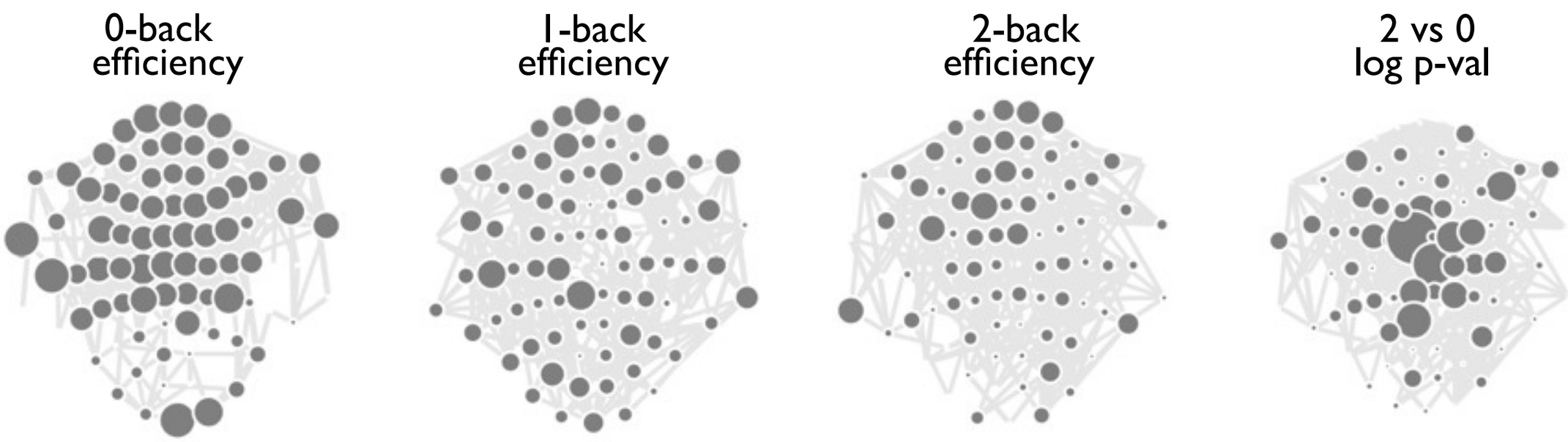
Data: 16 HC, visual memory task (0-2 back), 6 x 14 x task, MEG 1kHz sampling + 0.03-330 Hz BPF

Vertices: 87 sensors

Edge labels: trial-averaged phase synchronisation, thresholded

Results

Local efficiency decreases (less local clustering, more segregation) with increasing load in beta band



[Kitzlbicher et al., J. Neurosci, 2011]

Conclusions

Representing "connectivity" as a graph enables the application of the same inference methods across modalities, scales, and experimental paradigms

The choice of method depends on

Spatial scale of interest - whole-brain / subnetwork / region

Multimodality - Do we need to compare graphs across modalities?

Need for prediction - for clinical/marker applications, we probably want to favour predictive modelling (single-subject)

Interpretability - can we make sense of the nature of differences between graphs?

Visualisation - can we easily plot inference results?

Code[1] is available for most of these methods...

[1] jonas.richiardi@stanford.edu

Thanks

FINDlab, Stanford University
A. Altmann, M. Greicius, B. Ng

CS, Uni. Bern
H. Bunke, K. Riesen

MIPLab, UniGE/EPFL
D. Van De Ville, N. Leonardi

TU München
D. Mateus, G. Castrillon

GIPSA-LAb, INPG
S. Achard

CSIRlab, Med. Uni. Vienna
G. Langs

LabNIC, UniGE
P. Vuilleumier, M. Gschwind

Modelling and Inference on Brain networks for Diagnosis, MC IOF #299500

Subliminal ad: if you like machine learning on brain data come to Tübingen in June 2014 http://prni.org/

References

A few overview papers for graph comparison approaches

J. Richiardi, S. Achard, H. Bunke, D. Van De Ville, *Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience*, IEEE Signal Processing Magazine, May 2013, pp. 58-70

J. Richiardi & B. Ng, *Recent advances in supervised learning for brain graph classification*, Proc. GlobalSIP 2013 (in press)

G. Varoquaux, R.C. Craddock, *Learning and comparing functional connectomes across subjects*, NeuroImage (80), 2013, pp.405-415