Indirect Comparison of Interaction Graphs

Ulrich Mansmann, Markus Schmidberger, Ralf Strobl and Vindi Jurinovic

Abstract A strategy for testing differential conditional independence structures (CIS) between two graphs is introduced. The graphs have the same set of nodes and are estimated from data sampled under two different conditions. The test uses the entire pathplot in a Lasso regression as the information on how a node connects with the remaining nodes in the graph.
The interpretation of the paths as random processes allows defining stopping times which make the statistical properties of the test statistic accessible to analytic reasoning. A resampling approach is proposed to calculated p-values simultaneously for a hierarchical testing procedure. The hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no node in the graph shows a differential CIS). If the global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual nodes is reached.
The strategy is applied to association patterns of categories from the ICF in patients under post-acute rehabilitation. The patients are characterized by two different conditions. A comprehensive understanding of differences in the conditional independence structures between the patient groups is pivotal for evidence-based intervention design on the policy, the service and the clinical level related to their treatment.
Due to extensive computation, parallel computing offers an effective approach to implement our explorative tool and to locate nodes in a graph which show differential CIS between two conditions.

Ulrich Mansmann, Markus Schmidberger and Vindi Jurinovic
IBE, LMU Munich, Germany, e-mail: mansmann@ibe.med.uni-muenchen.de

Ralf Strobl
Institute for Health and Rehabilitation Sciences, LMU Munich, Germany

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_14,

1 Introduction

We present a statistical strategy to detect changes in the conditional independence structure (CIS) between elements under different conditions. For example, the elements could be the genes which are annotated to a certain pathway. The conditions may be defined by two different diseases and two datasets containing the corresponding gene expression information measured in tissues from the respective patients. Finally, the CIS between the genes of the pathway may be estimated by an appropriate method (Schäfer & Strimmer 2005, Meinshausen & Bühlmann 2006, Wainwright et al. 2006, Friedman et al. 2007, Banerjee et al. 2008).
The detection of nodes which show differences in the way they connect to other nodes is straightforward by visual inspection of both graphs. But, it is difficult to decide which of the detected nodes show a differential CIS between both conditions caused by systematic differences (true positives) and which are statistical artefacts caused by the algorithm or random fluctuation in the data (false positives). A similar problem exists for the nodes with equal CIS's between both conditions. It is difficult to discriminate between true or false negatives. The goal of this paper is to present a strategy to detect a set of nodes with differential CIS under a controlled error rate.
The proposed strategy to detect the set of nodes is called indirect because an explicit estimation of the CIS between nodes is avoided. A direct test calculates the test statsitic from the estimated graphs. For example it can be based on a resampling (permutation) approach which works as follows:

- Choose a metric to measure differential connectivity between two graphs. This can be done by the Structural Hamming Distance (*SHD*).
- Estimate the two graphs by a specific algorithm from the given data and determine the SHD_{obs} between both graphs.
- Permute the data units between both data sets, estimate both graphs for permutation i and calculate the specific SHD_i $(i = 1, \ldots, R)$.
- Determine a permutation p-value by $\#\{SHD_{obs} < SHD_i\}/R$.

Related ideas can be found in Balasubramanian et al. (2004) or Ruschhaupt (2008). The strategy proposed will use a global test for a set of nodes. Furthermore, a given hierarchy of clusters within the set of nodes is considered. The hierarchy has to be derived from specific domain knowledge. For each cluster C we will test the null hypothesis $H_{0,C}$: The cluster C does not contain any node with a differential CIS to other nodes of the graph.
The hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no node in the graph shows a differential CIS). If the global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual nodes is reached.
Meinshausen (2008) developed an attractive approach for hierarchical testing which will be used to solve our problem.
In computational biology, it might for example be interesting to use the Gene Ontology (Ashburner et al. 2000) when testing for the differential connectivity derived

for genes of a specific pathway or functional group. But, the Gene Ontology does not posses the hierarchical nature of the hierarchies used by Meinshausen, although the approach can be made feasible (with some more cumbersome notation) for Gene Ontology and related hierarchies derived from genomic domain knowledge (Goeman & Mansmann 2008).
Since simple hierarchies do not exist for problems in computational biology, we study for illustrative reasons an example from human functioning where the nodes are respective categories defined by the International Classification of Functioning, Disability and Health (ICF, WHO (2001)).
The paper is organized as follows: Section 2 introduces the methodological aspects of the test statistic with which we compare graphs. It states theorems to describe properties of the test statistics, and defines the sampling approach to perform the hierarchical test procedure. Section 3 presents the example and Section 4 will discuss our approach. The Appendix offers some results to the properties of the test statistics.

2 Methods

Consider the p-dimensional multivariate distributed random variable $X = (X_1, \ldots, X_p)$ which is the outcome of a Markov random field (MRF). A Markov random field is specified by an undirected graph $G = (N, E)$, with node set $N = 1, 2, \ldots, p$ and edge set $E \subset N \times N$. The structure of this graph encodes certain conditional independence assumptions among subsets of the p-dimensional random variable X, where variable X_i is associated with node $i \in N$.
For multivariate Gaussian data, the article Meinshausen & Bühlmann (2006) solved the fundamental problem of estimating the structure of the underlying graph given a set of n samples from the MRF and showed that L_1-regularization can lead to practical algorithms with strong theoretical guarantees. For multivariate binary data, Wainwright et al. (2006) provides comparable results. Both methods use L_1-regularized regression (linear and logistic), in which the neighbourhood of any given node is estimated by performing regression subject to an L_1-constraint. Neighbourhood selection estimates the CIS separately for each node in the graph and is hence equivalent to variable selection for regression models. The proposed neighbourhood selection schemes are consistent for sparse high-dimensional graphs. Consistency depends on the choice of the penalty parameter which can be derived from controlling the probability of falsely joining some distinct connectivity components of the graph.

2.1 Defining the Test Statistic

For the specific node i the corresponding path plot of the regression coefficients for the L_1-regularized regression can be interpreted as a $p-1$ dimensional random process indexed by the penalty parameter λ: $B^{(i)}(\omega, \lambda) = (\beta_\lambda^{i,j})_{j \in N \setminus \{i\}}$ where $\lambda \geq 0$.

The randomness is introduced by the random data sample. For each of the $p-1$ paths of $B^{(i)}$ it is possible to determine the stopping time $\tau^{(i,j)} = \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\}$, $j \in N \setminus \{i\}$. For a fixed set of nodes $N = 1,\ldots,p$ we observe two i.i.d. samples of sizes n and m: $D_X = \{x^{(1)},\ldots,x^{(n)}\}$ and $D_Y = \{y^{(1)},\ldots,y^{(m)}\}$ with $x^{(i)} = (x_1^{(i)},\ldots,x_p^{(i)})$ and $y^{(k)} = (y_1^{(k)},\ldots,y_p^{(k)})$. For node i in node set N the path plots derived from both data sets are compared by counting the number of common non-zero regression coefficients given penalty parameter λ_x for the path plot derived from data D_X and penalty parameter λ_y for the path plot derived from data D_Y : $\Psi_i(\lambda_x,\lambda_y)$. This function is integrated over the range of the penalty:

$$\Psi_i = \iint\limits_{[0,\infty[\times[0,\infty[} \Psi_i(\lambda_x,\lambda_y)d\lambda_x d\lambda_y \tag{1}$$

The random variable Ψ_i can also be calculated from the stopping times introduced above:

$$\Psi_i = \sum_{j\in N\setminus\{i\}} \tau_X^{(i,j)} \cdot \tau_Y^{(i,j)} \tag{2}$$

where stopping times $\tau_X(\tau_Y)$ are derived from the path plot inferred from data $D_X(D_Y)$.
It is also possible to calculate a Ψ_N for the entire graph or a Ψ_C related to the subset $C \subset N$ of nodes:

$$\Psi_N = \sum_{i\in N} \Psi_i \text{ and } \Psi_C = \sum_{i\in C\subset N} \Psi_i \tag{3}$$

We define

$$\Psi_i^{max} := \sum_{j\in N\setminus\{i\}} \max\{\tau_X^{(i,j)}, \tau_Y^{(i,j)}\}^2 \tag{4}$$

The following theorem supports the intuition that the value of Ψ_i is larger when the same *MRF* defines the distribution of the data in both conditions than when the distributions are defined by two different *MRF*s.

Theorem 1. *Let P_1 and P_2 be two equal MRFs over the same set of nodes N, D_X and D_Y two i.i.d. samples from both MRFs of size n resp. $m(n) = n \cdot \frac{1-\pi}{\pi}$. The quantity $\pi = \frac{n}{n+m}$ is the fixed percentage of the sample size of D_X on the total number of observations. Then for an arbitrary small $\varepsilon > 0$ and each node $i \in N$ there is an $n(\varepsilon)$ such that for all $n > n(\varepsilon)$ it holds*

$$P[\Psi_i^{max} \leq \Psi_i + \varepsilon] > 1 - \varepsilon. \tag{5}$$

A sketch of the proof is given in the Appendix.

2.2 A Permutation Test

Theorem 2. *Let P_1 and P_2 be two identical Markov random fields (MRFs) which generate the data under the two conditions of interest. Let Q be the mixture distribution created from P_1 and P_2 with mixture proportion π (for component P_1). For an arbitrary small $\varepsilon > 0$, a value $w > 0$, and each node $i \in N$ there is an $n(\varepsilon)$ such that for all $n > n(\varepsilon)$ it holds*

$$|P[\Psi_i^* > w] - P[\Psi_i^{\#} > w]| < \varepsilon. \tag{6}$$

The value of Ψ_i^* is calculated from the data sets D_X^* [n i.i.d. samples from P_1] and D_Y^* [$m = n\cdot(1-\pi)/\pi$ i.i.d. samples from P_2] and $\Psi_i^{\#}$ is calculated from the data sets D'_X [n i.i.d. samples from Q] and D'_Y [$m = n\cdot(1-\pi)/\pi$ i.i.d. samples from Q].

A sketch of the proof and a possible extension to a wider class of null-hypotheses is given in the Appendix.

The theorem above states for each node in N: Under the null-hypothesis (*equal MRFs*) the distribution of the test statistics can be generated from a permutation sampling of the observed data.

The permutation procedure calculates S samples simultaneously for each node i: $\Psi_i^{\#(S)}$. The permutation p-value for node i is derived as $p_i = |\{r : \Psi_i^{\#(S)} > \Psi_i^{\#}\}|/S$. It is straight forward to extend the theorem to test statistics for the set of nodes ($\Psi_N = \sum_{i\in N} \Psi_i$) or specific subsets of nodes ($\Psi_C = \sum_{i\in C\subset N} \Psi_i$). Correspondingly, it is possible to calculate permutation p-values for arbitrary sets of nodes.

2.3 Hierarchical Testing

Now it is straightforward to combine our approach with the hierarchical testing principle of Meinshausen (2008). The principle allows using the same resampling sample to calculate p-values for each element of the hierarchy.

For the following, we assume that a hierarchy $\aleph$ is given, which is a set of clusters $C \subset \{1,\ldots,p\}$. The cardinality of a cluster C is denoted by $|C|$. The root node $\{1,\ldots,p\}$ contains all nodes of the graph and has cardinality p. The hierarchical structure implies that any two clusters $C, C' \in \aleph$ either have an empty intersection, or that one cluster is a subset of the other.

To take the multiplicity of the testing problem into account, p-values have to be adjusted. Define for cluster C the adjusted p-value as $p_{adj}^C := p_C \cdot \frac{p}{|C|}$ where p is the total number of nodes in the graph and $|C|$ is the number of nodes in the cluster of interest. The adjustment amounts to multiplying the p-value of each cluster C with the inverse of the fraction $|C|/p$ of variables it contains. The adjustment is thus resolution dependent. At coarse resolutions, the penalty for multiplicity is weak, and it increases for finer resolution levels. The p-value of the root node is thus unadjusted, whereas individual variables receive a Bonferroni-type adjustment.

The hierarchical testing procedure rejects now all hypotheses $H_{0,C}$ with $C \in \aleph$ for which (a) the adjusted p-value p^C_{adj} is below or equal to the specified level and (b) the parent node is rejected (this is always considered to be fulfilled for the root node).
Note that condition (b) is not a severe restriction. The null hypothesis $H_{0,C}$ of a cluster C is by definition always true if the null hypothesis $H_{0,pa(C)}$ is true for the parent cluster $pa(C)$. Hence it makes sense to stop testing in subtrees of clusters whose null hypothesis could not be rejected.
Using the definition of a hierarchically adjusted p-value $p^C_{h,adj} = \max_{D \in \aleph, C \subseteq D} p^D_{adj}$, the set of clusters which are rejected in the hierarchy $\aleph$ on the level α is then given by $C_{rejected} = \{C \in \aleph, p^C_{h,adj} < \alpha\}$.
Control of the family-wise error rate can now be achieved. The set of clusters that fulfill the null hypothesis $H_{0,C}$ is denoted by $C_0 = \{C \in \aleph, H_{0,C} \text{ is fulfilled}\}$. Family-wise error rate control entails that the probability of rejecting any cluster in C_0 is smaller than the pre-specified level α.

Theorem 3. *For $C_{rejected}$ and C_0 as defined above, the family-wise error rate is controlled at level α:*

$$P(C_{rejected} \cap C_0 = \emptyset) = 1 - \alpha \tag{7}$$

Proof is given by Meinshausen (2008).

2.4 Computational Issues

Calculations are done in the statistical computing software `R (V 2.9.0)` (R Development Core Team 2009). The working horse of the Lasso Regression is the computationally efficient gradient-ascent algorithm as proposed by Goeman (2009b) and implemented in Goeman (2009a). The permutation test was parallelized.
A total of 1000 samples were created to perform the comparison of the CIS between both conditions. The calculation is very computer intensive and sample calculations are independent from each other. Therefore, the functions and data are distributed to different processors. In the `R` language different technologies and approaches for parallel computing exist (Schmidberger et al. 2009). We use the snow package (`snow`) with the Rmpi package (`Rmpi`) for the communication between the processors. The code is executed at 1000 processors using the super computer HLRB2 at the Leibniz-Rechenzentrum in Munich (Germany). To guarantee a different random number stream in every `R` session, an additional `R` package rlecuyer (`rlecuyer`) is applied.
An `R`-package which offers the needed algorithms is under preparation.

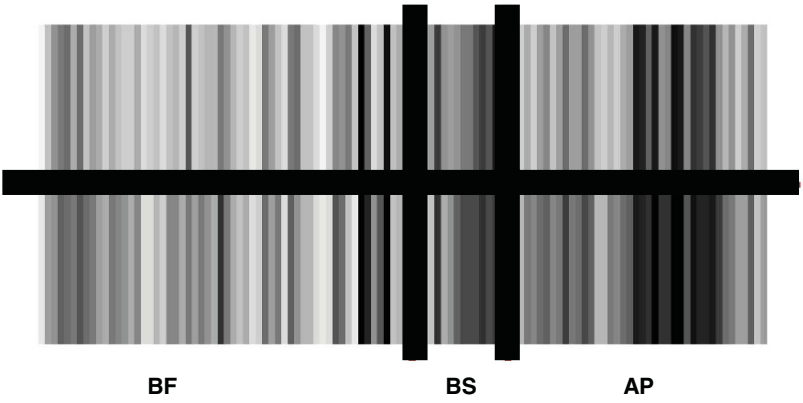

Fig. 1 Relative frequences of impairments and limitations. Patients with condition B are above the horizontal black line, patients with condition A below. BF = Body Functions, BS = Body Structures, AP = Activities & Participation

3 Example

The example studies a multivariate binary situation. It is taken from the field of rehabilitation science.

Functioning and disability are universal human experiences. Over the life span people may experience limitations in functioning in relation to health conditions including an acute disease or injury, a chronic condition, or aging. A standard language for the analysis of functioning is provided by the International Classification of Functioning, Disability and Health (ICF, WHO (2001)).

A secondary analysis of observational cross sectional data of patients from five early post-acute rehabilitation units is performed. The ICF is used to measure functioning and contextual factors. We look at the components Body Functions, Body Structures, and Activities & Participation. The presence of an impairment or limitation was binary coded for each of the 122 categories considered.

616 patients (mean age 63 years, 46% male) were included. 56% had health condition A. Figure 1 shows the profiles of the 122 ICF-categories measured during post-acute rehabilitation between the 343 patients with health condition A and the 273 patients with health condition B.

Besides the comparison of functional profiles it is important to understand stability and distinctiveness of functioning across health conditions. This can be achieved by revealing patterns of associations between distinct aspects of functioning, the ICF categories.

Based on the proposal of Wainwright et al. (2006), CIS graphs can be estimated for each condition and visually compared as done in Figure 2.

Besides a simple visual inspection for differential CIS we apply the combination between our test and the hierarchical testing procedure. The hierarchy is defined by

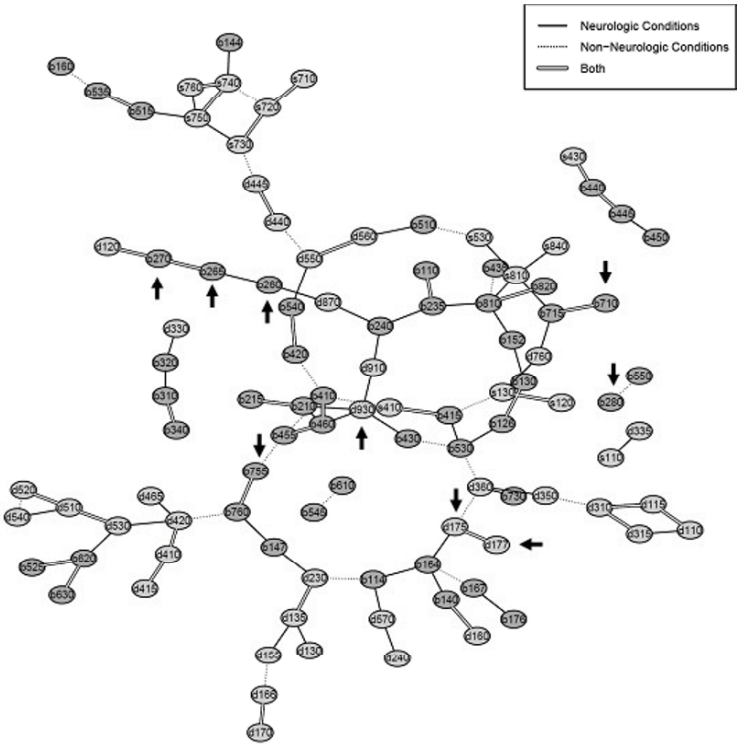

Fig. 2 Estimated CIS graphs and visual presentation of common edges as well as different edges within both graphs. The arrows point to the nodes with significant differential CIS on a 5% significance level ($\alpha = 0.05$). The three ICF components are presented in different colours: orange: Body Functions, green: Body Structures, blue: Activities & Participation

the ICF itself (WHO 2001). The root node contains all 122 ICF categories analyzed. The first level of the hierarchy is defined by the ICF-components. The second level is determined by the ICF-chapters; the first branching level in the classification gives the third level. The fourth level is given by the single categories of interest.
Table 1 gives the nodes with a significant differential CIS according to the hierarchical test procedure. The complete results are presented in the Appendix.

Our strategy was able to detect a set of nodes with differential CIS in the multivariate ICF profiles between patients under two different disease conditions as well as clusters in the hierarchy (result not shown). The identified set holds a family wise error rate of 5%, i.e. the probability of containing at least one false positive node under the null-hypothesis is 5%. The result on differential CIS is combined with two graph estimates and their naive comparison with respect to present or missing edges between both graphs by adding black arrows to the naive graph comparison. Some of the arrows point to nodes ($b260, b280, 285, b755, d175, d930$) which under condition A show a different connectivity than under condition B. Some of the arrows point

Table 1 nodes with significant differential CIS

Code	title	p.value	p.value.adj
b260	Proprioceptive function	0.00018	0.02056
b265	Touch function	1e-05	0.00124
b270	Sensory functions related to temperature and other stimuli	1e-05	0.00124
b280	Sensation of pain	0.00015	0.01722
b710	Mobility of joint functions	4e-05	0.00497
b755	Involuntary movement reaction functions	1e-05	0.0014
d175	Solving problems	1e-05	0.01244
d177	Making decisions	0.00025	0.02855
d930	Religion and spirituality	1e-05	0.0056

to nodes ($b170$, $b270$, $d177$) that show the same connectivity in both graphs. This can be understood by looking more closely to the null-hypothesis which is rejected: differential CIS can also be produced by different regression coefficients given the same connectivity. For example the odds ratio between $b710$ and $b715$ ($b270$ and $d120$, $b270$ and $b265$) is 3.91 (19.27, 29.12) in patients with condition A and 13.12 (10.89, 19.29) in patients with condition B.

4 Discussion

The estimation of complex graphs from observed data is a challenging task and different strategies were developed for the case of sparse graphical structures (Schäfer & Strimmer 2005, Meinshausen & Bühlmann 2006, Wainwright et al. 2006, Friedman et al. 2007, Kalisch & Bühlmann 2007, Banerjee et al. 2008). Especially, the estimation of the conditional independence structure (CIS) in a multivariate observation is of high interest. Different data sets of the same multidimensional random variable from different conditions may be available. They may produce different CIS-graph estimates. A natural question is if the observed data give convincing evidence for a systematic difference between the CIS-graphs behind the distributions of the observed data. The meaning of convincing evidence has to be operationalized in statistical terms. This is achieved by using the family wise error rate.

It is our intention to compare the conditional independence structure (CIS) between two multivariate Gaussian or binary distributions (Ising model). Multivariate Gaussian or binary distributions define Markov random fields (MRF) and imply a graph with the nodes defined by the single components of the multivariate random variable and the edges between the nodes by the CIS. Besides different set of edges, further differences in the CISs between two distributions are created by the strength of the conditional dependencies between specific nodes. We present a test statistic which addresses both CIS-aspects.

The measure to compare two graphs uses the idea that the connectivity of a node with

the remaining nodes in a graph relates to a variable selection problem in a regression setting. In this context, the usefulness of L_1-regularized regression was demonstrated by several authors (Wainwright et al. 2006, Meinshausen & Bühlmann 2006, Friedman et al. 2007). The differential connectivity of node i in both graphs is defined as follows: (1) for each node $j \neq i$ we determine the minimal penalty parameter which shrinks the corresponding regression coefficient to zero $(\tau_X^{(i,j)}, \tau_Y^{(i,j)})$ in both conditions; (2) for each $j \neq i$ we calculate $\tau_X^{(i,j)} \cdot \tau_Y^{(i,j)}$; (3) we determine the test statistic Ψ_i as the sum of the products over all $j \neq i$. The test statistics for the entire graph or a subset C of nodes is $\Psi_N = \sum_{i \in N} \Psi_i$ and $\Psi_C = \sum_{i \in C \subset N} \Psi_i$ respectively. The test statistic is motivated by the intuition that equal *MRF*s in both conditions will produce a large value of Ψ_i, Ψ_N or Ψ_C.

It is shown in the Appendix that the $\tau^{(i,j)}$s can be calculated in principle and that formal statements about their properties can be derived.

The null-hypothesis of equal *MRF*s for both conditions is tested by a permutation approach. It allows formulating global tests on sets of nodes as well as tests for single nodes. This enables searching for a set of nodes with differential CIS in a hierarchical testing procedure. The motivation for hierarchical testing can be summarized as follows:

- Any differential connectivity at all? The CIS of a group of nodes can be tested between both graphs whether all nodes have the same CIS under each condition.
- Differential CIS in sub-clusters? If it is established that a cluster of nodes does indeed contain nodes with differential CIS, it is desirable to attribute it to one or several sub-clusters

If possible, the differential CIS in a cluster of variables is attributed to its sub-clusters. In each sub-cluster, it is again examined whether the collective effect can be attributed to even smaller sub-clusters of nodes. The procedure retains the smallest possible clusters which exhibit a significant differential CIS or helps to detect single nodes with differential CIS.

Our approach avoids estimating graphs explicitly. We did not put the direct and the indirect approach side by side. Therefore, no detailed analysis of their pros and cons is available. The indirect approach does not fix the value of a regularization parameter which has to be done when the estimate of an explicit graph is needed. The direct approach needs explicit graphs since differential CIS may be measured by the Structural Hamming Distance (*SHD*) (Kalisch & Bühlmann 2007) or other suitable methods. The hierarchical test procedure can be applied for the direct as well as the indirect approach.

One potential advantage of the proposed test statistics is its generalizability to more than two graphs. The comparison of several graphs based on the bivariate *SHD* measure is cumbersome and not illuminative. For node i and data from conditions U, V, W, X, Y, and Z (for example) it is possible to modify Ψ_i by $\Psi_i = \sum_{j \in N \setminus \{i\}} \tau_U^{(i,j)} \cdots \tau_Z^{(i,j)}$. This is the subject of further research.

In this paper we study the null-hypothesis of two equal Markov random fields. More general forms of the null-hypothesis ($H_{0,general}$) may be of interest: *The MRFs are*

different but the graphs related to both distributions have a Structural Hamming Distance of 0. A CIS graph G defines a family of probability measures (multivariate Gaussian or multivariate binary Ising Model) p_G by the corresponding CIS. The null-hypothesis $H_{0,general}$ states that the distributions which generate the data under both conditions belong to p_G. Since the mixture of two distributions from p_G is in general not in p_G it follows that the permutation approach used so far will not work anymore. We assume that $H_{0,general}$ can be tested by replacing the permutation by a more complicated resampling procedure. It may be based on a sampling scheme which creates new versions of data D_X and D_Y under the restriction that the graphs behind the distribution of X and Y have the same set of edges, $SHD(G_X, G_Y) = 0$. The proof of this assumption and the development of an efficient algorithm are topics of ongoing research.

The strategy presented in our paper depends on the implicit assumption that the data is created by *MRF*s (multivariate Gaussian or multivariate binary Ising Model). We used this assumption implicitly in our example. Tools for model validation and model assessment in a setting comparable to the data presented are under development (Gneiting 2008). It is a second line of our research to implement efficient validation strategies to assess model assumptions. A mixture of binary Ising models is in general not a binary Ising model. A mixture of binary Ising models with the same conditional independence graph does not need to have the same graph anymore because of confounding. We tried to reduce confounding by using fixed covariates in the Lasso regression. The algorithm provided by Goeman (2008) allows controlling for confounding by the incorporation of fixed covariates.

The strategy presented offers an explorative tool to detect nodes in a graph with the potential of a relevant impact on the regulatory process between interacting units in a complex process. The findings introduce a practical algorithm with theoretical guarantees. We see our result as the first step on the way to a meta-analysis of graphs. A meta-analysis of graphs is only useful if the graphs available for aggregation are *homogeneous*. The definition of the homogeneity of graphs $G_1, \ldots, G_K$ by a pairwise Structural Hamming Distance of 0 is not sufficient to describe homogeneity in a correct way. The assessment of homogeneity of graphs needs procedures like the one presented.

Acknowledgements This work is supported by the LMU*innovativ* project *Analysis and Modelling of Complex Systems in Biology and Medicine (Cluster B, Expression Analyses)*.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology:a tool for the unification of biology., *Nature Genetics* **25**: 25–29.

Balasubramanian, R., LaFramboise, T., Scholtens, D. & Gentleman, R. (2004). A graph-theoretic approach to testing associations between disparate sources of functional genomics data., *Bioin-*

formatics **20**(18): 3353–3362.
Banerjee, O., Ghaoui, L. E. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, *Journal of Machine Learning Research* pp. 485–516.
Friedman, J., Hastie, T. & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* .
Gneiting, T. (2008). Editorial: Probabilistic forecasting, *Journal of the Royal Statistical Society: Series A* **17**: 319–321.
Goeman, J. (2008). *penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*. `http://www.msbi.nl/goeman`. R package version 0.9-22.
Goeman, J. (2009a). L1 and l2 penalized regression models. `http://www.msbi.nl/goeman`. R package version 0.9-24.
Goeman, J. (2009b). L1 penalized estimation in the cox proportional hazards model., *Biometrical Journal* .
Goeman, J. J. & Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology., *Bioinformatics* **24**(4): 537–544.
Kalisch, M. & Bühlmann, P. (2007). Estimating high dimensional acyclic graphs with the pc-algorithm, *Journal of Machine Learning Research* **8**: 613–636.
Meinshausen, N. (2008). Hierarchical testing of variable importance, *Biometrika* **95**: 265–276.
Meinshausen, N. & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso, *The Annals of Statistics* **34**: 1436–1462.
R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.
Ruschhaupt, M. (2008). *Erzeugung von positiv definiten Matrizen mit Nebenbedingungen zur Validierung von Netzwerkalgorithmen für Microarray-Daten*, PhD thesis, LMU München, Fakultät für Mathematik, Informatik und Statistik, München, Germany. `http://edoc.ub.uni-muenchen.de/view/subjects/fak16.html`.
Schäfer, J. & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics., *Statistical Applications in Genetics and Molecular Biology* **4**: Article32.
Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L. & Mansmann, U. (2009). State of the art in parallel computing with R, *Journal of Statistical Software* **31**(1). `http://www.jstatsoft.org/v31/i01/`.
Wainwright, M. J., Ravikumar, P. & Lafferty, J. D. (2006). High dimensional graphical model selection using l1-regularized logistic regression, *Proceedings of Advances in neural information processing systems* **9**: 1465–1472.
WHO (2001). International classification of functioning, disability and health (icf), *Nature Genetics* .

Appendix

Properties of the Stopping Times

We look at the following penalized quadratic form

$$QF(\beta,\lambda) = (\beta-\beta^*)^t \cdot \Sigma \cdot (\beta-\beta^*) - \lambda \cdot \sum_{i=1}^{d} |\beta_i| \tag{8}$$

where β is a d-dimensional parameter vector, β^* a d-dimensional fixed vector, Σ a $d \times d$ matrix, and $\lambda \geq 0$. For given λ, QF is maximized by $\beta^{\#}(\lambda)$

$$\beta_j^{\#}(\lambda) = 2 \cdot \beta_j^* + \lambda \cdot \sum_k (-1)^{j+k} \mathrm{Det}(\Sigma_{jk}) / \mathrm{Det}(\Sigma) \tag{9}$$

for $j = 1, \ldots, d$. Σ_{jk} is a quadratic matrix derived from matrix Σ by removing line j and column k.
The penalized log-likelihood for a linear regression problem $y_i \sim \beta \cdot x_i$ is given by QF where $\Sigma = X^t \cdot X$ and $\beta^* = y \cdot X^t \cdot (X^t \cdot X)^{-1}$. The penalized log-likelihood for a logistic regression problem $\mathrm{logit}(p_i) = \beta \cdot x_i$ is approximated by QF where β^* is the ML estimate and Σ is the corresponding Fisher Information Matrix.
From (1.8) it is possible to calculate the minimal penalty parameter λ which shrinks the coefficient to 0. In terms of the notation introduced in the methods section if follows

$$\tau^{(i,j)} = \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\} = 2 \cdot \beta_j^* \cdot \frac{\mathrm{Det}(\Sigma)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk})} \tag{10}$$

The matrix Σ is derived from observed data and varies around the true value Σ^0. The variability of $\tau^{(i,j)}$ can be controlled by the following property of determinants:

$$\begin{aligned} \mathrm{Det}(\Sigma^0 + \delta \cdot \Lambda) &= \mathrm{Det}\left(\left(\Sigma^0\right)^{-1}\left(I + \delta \cdot \Lambda \cdot \left(\Sigma^0\right)^{-1}\right)\right) \\ &= \mathrm{Det}\left(\left(\Sigma^0\right)^{-1}\right) \cdot \left(I + \delta \cdot \mathrm{trace}\left(\Lambda \cdot \left(\Sigma^0\right)^{-1}\right)\right) \end{aligned}$$

It holds that

$$\begin{aligned} \tau^{(i,j)} &= \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\} \\ &= 2 \cdot \beta_j^0 \cdot \frac{\mathrm{Det}(\Sigma^0)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk}^0)} + \frac{\mathrm{Det}(\Sigma^0)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk}^0)} \cdot \varepsilon_{ij} \end{aligned}$$

where β_j^0 is the true regression coefficient. The random variables ε_{ij} concentrate on a small neighbourhood of 0 (depending of the sample size).

Statistical Properties of Ψ_i

Sketch of proof of Theorem 1.1: Without loss of generality we choose $r = 1$. We denote a variable or parameter which belongs to *MRF r* and the regression of node j on node i by $..{}_r^{(i,j)}$. We also introduce for the *MRF* $P_r (r = 1, 2)$ the notation $\Omega_r^{(i,j)} =$

$\frac{\mathrm{Det}(\Sigma^0_{jk,r})}{\sum_k (-1)^{j+k+1}\mathrm{Det}(\Sigma^0_{jk,r})}$ where $\Sigma^0_{jk,r}$ is Σ^0 which belongs to *MRF* $r(r=1,2)$ without row j and column k.
The inequality claimed in the theorem follows from

$$\Psi_i^{max} = \sum_{j \in n \setminus \{i\}} \max\{\tau_X^{(i,j)}, \tau_Y^{(i,j)}\}^2 \tag{11}$$

using the approximative representation of $\tau^{(i,j)}$ by the true parameter values and a controlled error term $\tilde{\varepsilon}$. It follows

$$\Psi_i^{max} \leq \sum_{j \in n \setminus \{i\}} \max\{\beta_1^{(i,j)} \cdot \Omega_1^{(i,j)}, \beta_2^{(i,j)} \cdot \Omega_2^{(i,j)}\}^2 + \tilde{\varepsilon} \tag{12}$$

under the null-hypothesis $\beta_1^{(i,j)} = \beta_2^{(i,j)} = \beta^{(i,j)}$ and $\Omega_1^{(i,j)} = c \cdot \Omega_2^{(i,j)} = \Omega^{(i,j)}$ where $c = c(n,\pi)$. As a consequence,

$$\Psi_i^{max} \leq \sum_{j \in n \setminus \{i\}} \{\beta^{(i,j)} \cdot \Omega^{(i,j)}\}^2 \cdot c^{-1} + \tilde{\varepsilon} \tag{13}$$

using again the approximative representation of $\tau^{(i,j)}$ it follows

$$\Psi_i^{max} = \sum_{j \in n \setminus \{i\}} \tau_X^{(i,j)} \cdot \tau_Y^{(i,j)} + 2 \cdot \tilde{\varepsilon} \tag{14}$$

where $P[|\tilde{\varepsilon}| < \varepsilon] > 1 - \varepsilon$ for $n > n(\varepsilon)$. This proves the theorem.

Sketch of proof of Theorem 1.2: The proof follows from Theorem 1 by the argument that $P_1 = P_2$ and therefore $P[\Psi_i^* \leq \Psi_i^\#] > 1 - \varepsilon$ as well as $P[\Psi_i^\# \leq \Psi_i^*] > 1 - \varepsilon$ for an arbitrary small $\varepsilon > 0$ and $n > n(\varepsilon)$. This implies $P[|\Psi_i^\# - \Psi_i^*| < \varepsilon] > 1 - \varepsilon$ and

$$\begin{aligned} P[\Psi_i^* > w] &= P[\Psi_i^* - \Psi_i^\# + \Psi_i^\# > w] = P[\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*] \\ &= P[\{\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*\} \cap \{|\Psi_i^\# - \Psi_i^*| < \delta\}] + \\ &\quad P[\{\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*\} \cap \{|\Psi_i^\# - \Psi_i^*| > \delta\}]. \end{aligned}$$

For appropriately small δ and sufficiently large n it holds

$$P[\Psi_i^* > w] \geq P[\Psi_i^\# > w + \delta] \cdot (1 - \delta) \tag{15}$$

$$P[\Psi_i^* > w] \leq P[\Psi_i^\# > w - \delta] + \delta \tag{16}$$

which proves the theorem.

Results of the Hierarchical Test Procedure on Differential Conditional Independence Structure for each Node (ICF Category)

Code	title	p.value	p.value.adj
b110	Consciousness functions	0.26916	1
b114	Orientation functions	0.19031	1
b126	Temperament and personality functions	0.07132	1
b130	Energy and drive functions	0.40246	1
b134	Sleep functions	0.1266	1
b140	Attention functions	0.81069	1
b144	Memory functions	0.13688	1
b147	Psychomotor functions	0.03534	1
b152	Emotional functions	0.57016	1
b156	Perceptual functions	0.67479	1
b160	Thought functions	0.0098	1
b164	Higher-level cognitive functions	0.00572	0.64073
b167	Mental functions of language	0.0526	1
b176	Mental function of sequencing complex movements	0.0067	0.75053
b180	Experience of self and time functions	0.80941	1
b210	Seeing functions	0.21501	1
b215	Functions of structures adjoining the eye	0.85915	1
b230	Hearing functions	0.46113	1
b235	Vestibular functions	0.95262	1
b240	Sensations associated with hearing and vestibular function	0.19814	1
b260	Proprioceptive function	0.00018	0.02056
b265	Touch function	1e-05	0.00124
b270	Sensory functions related to temperature and other stimuli	1e-05	0.00124
b280	Sensation of pain	0.00015	0.01722
b310	Voice functions	<0.00001	0.18667
b340	Alternative vocalization functions	0.09032	1
b410	Heart functions	0.79414	1
b415	Blood vessel functions	0.01307	1
b420	Blood pressure functions	0.00783	0.87703
b430	Haematological system functions	0.15039	1
b435	Immunological system functions	0.0033	0.36913
b440	Respiration functions	0.1212	1
b445	Respiratory muscle functions	0.10322	1
b450	Additional respiratory functions	0.70879	1
b455	Exercise tolerance functions	0.81957	1
b460	Sensations associated with cardiovascular and respiratory functions	0.47558	1
b510	Ingestion functions	0.01285	1
b515	Digestive functions	0.79141	1
b525	Defecation functions	0.34564	1
b530	Weight maintenance functions	0.12755	1

Code	title	p.value	p.value.adj
b535	Sensations associated with the digestive system	0.21833	1
b540	General metabolic functions	0.17652	1
b545	Water, mineral and electrolyte balance functions	0.50622	1
b550	Thermoregulatory functions	0.98723	1
b610	Urinary excretory functionsn	0.20772	1
b620	Urination functions	0.85082	1
b630	Sensations associated with urinary functions	0.13142	1
b710	Mobility of joint functions	4e-05	0.00497
b715	Stability of joint functions	0.00113	0.12604
b730	Muscle power functions	0.16681	1
b735	Muscle tone functions	0.0137	1
b755	Involuntary movement reaction functions	1e-05	0.0014
b760	Control of voluntary movement functions	0.02091	1
b770	Gait pattern functions	0.22097	1
b780	Sensations related to muscles and movement functions	0.92971	1
b810	Protective functions of the skin	0.56761	1
b820	Repair functions of the skin	0.93102	1
s110	Structure of brain	0.0736	1
s120	Spinal cord and related structures	0.05937	1
s130	Structure of meninges	0.13523	1
s410	Structure of cardiovascular system	0.72956	1
s430	Structure of respiratory system	0.12588	1
s530	Structure of stomach	0.01038	1
s710	Structure of head and neck region	0.22969	1
s720	Structure of shoulder region	0.46041	1
s730	Structure of upper extremity	0.38268	1
s740	Structure of pelvic region	0.1352	1
s750	Structure of lower extremity	0.5652	1
s760	Structure of trunk	0.47323	1
s810	Structure of areas of skin	0.4446	1
s840	Structure of hair	0.15527	1
d110	Watching	0.03047	1
d115	Listening	0.0462	1
d120	Other purposeful sensing	0.27109	1
d130	Copying	0.02501	1
d135	Rehearsing	0.08179	1
d155	Acquiring skills	0.40423	1
d160	Focusing attention	0.52833	1
d166	Reading	0.06237	1
d170	Writing	0.82457	1
d175	Solving problems	<0.00001	0.01244
d177	Making decisions	0.00025	0.02855
d230	Carrying out daily routine	0.96039	1
d240	Handling stress and other psychological demands	0.49906	1

Code	title	p.value	p.value.adj
d310	Communicating with - receiving - spoken messages	0.42238	1
d315	Communicating with - receiving - nonverbal messages	0.83199	1
d330	Speaking	0.73189	1
d335	Producing nonverbal messages	0.01302	1
d350	Conversation	0.42381	1
d360	Using communication devices and techniques	<0.00001	1
d410	Changing basic body position	3e-05	1
d415	Maintaining a body position	0.26954	1
d420	Transferring oneself	0.94277	1
d430	Lifting and carrying objects	0.31586	1
d440	Fine hand use	0.85019	1
d445	Hand and arm use	0.22326	1
d450	Walking	0.32237	1
d460	Moving around in different locations	0.00711	0.79633
d465	Moving around using equipment	0.04467	1
d510	Washing oneself	0.44569	1
d520	Caring for body parts	0.17469	1
d530	Toileting	0.29863	1
d540	Dressing	0.98399	1
d550	Eating	0.48303	1
d560	Drinking	0.0371	1
d570	Looking after ones health	0.00608	0.68055
d760	Family relationships	0.03614	1
d870	Economic self-sufficiency	0.04971	1
d910	Community life	0.07671	1
d930	Religion and spirituality	<0.00001	0.0056
d940	Human rights	0.12703	1