Differential Network Analysis (DNA)

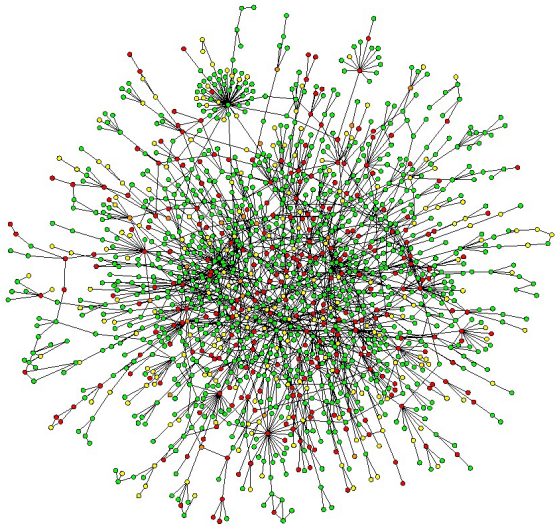
An overview

Antonio Miñarro (aminarro@ub.edu)

GRBIO July 17, 2025

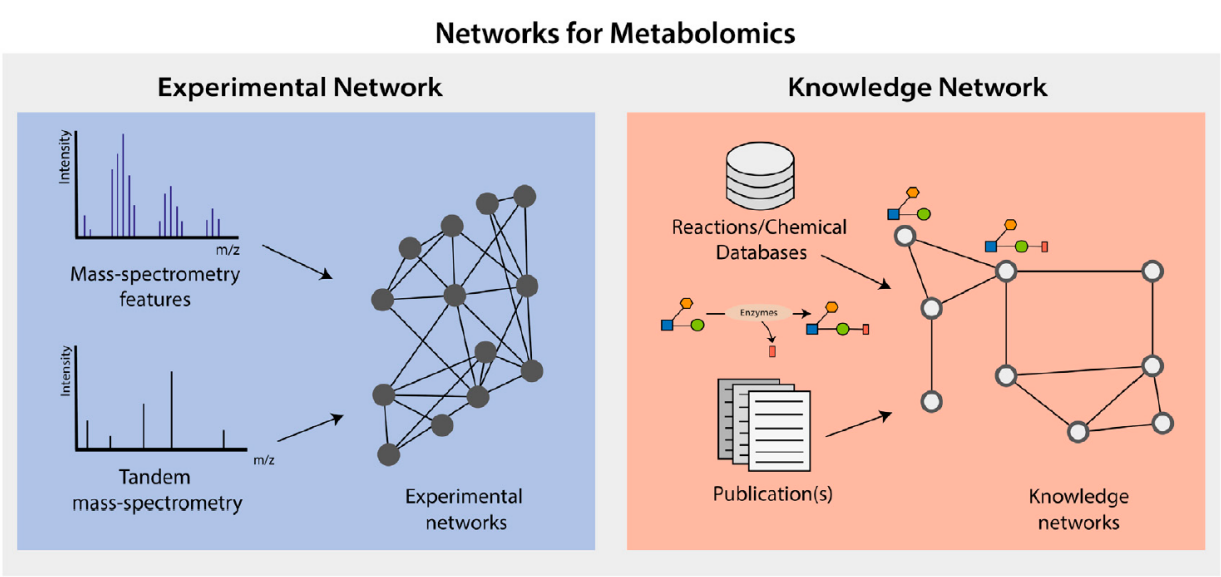
Introduction

- Networks provide additional information for the analysis of biological data beyond the traditional analysis that focuses on single variables.
- Several networks can be built from a single dataset (or a list of metabolites), where each network represents different relationships, such as statistical (correlated metabolites), biochemical (known or putative substrates and products of reactions), or chemical (structural similarities, ontological relations).

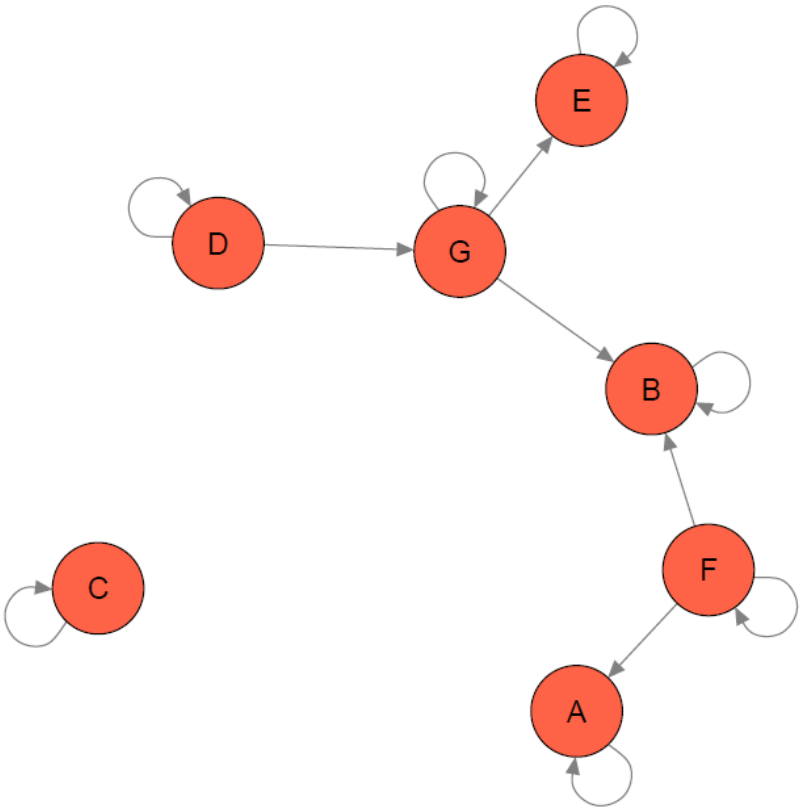


Knowledge and experimental networks

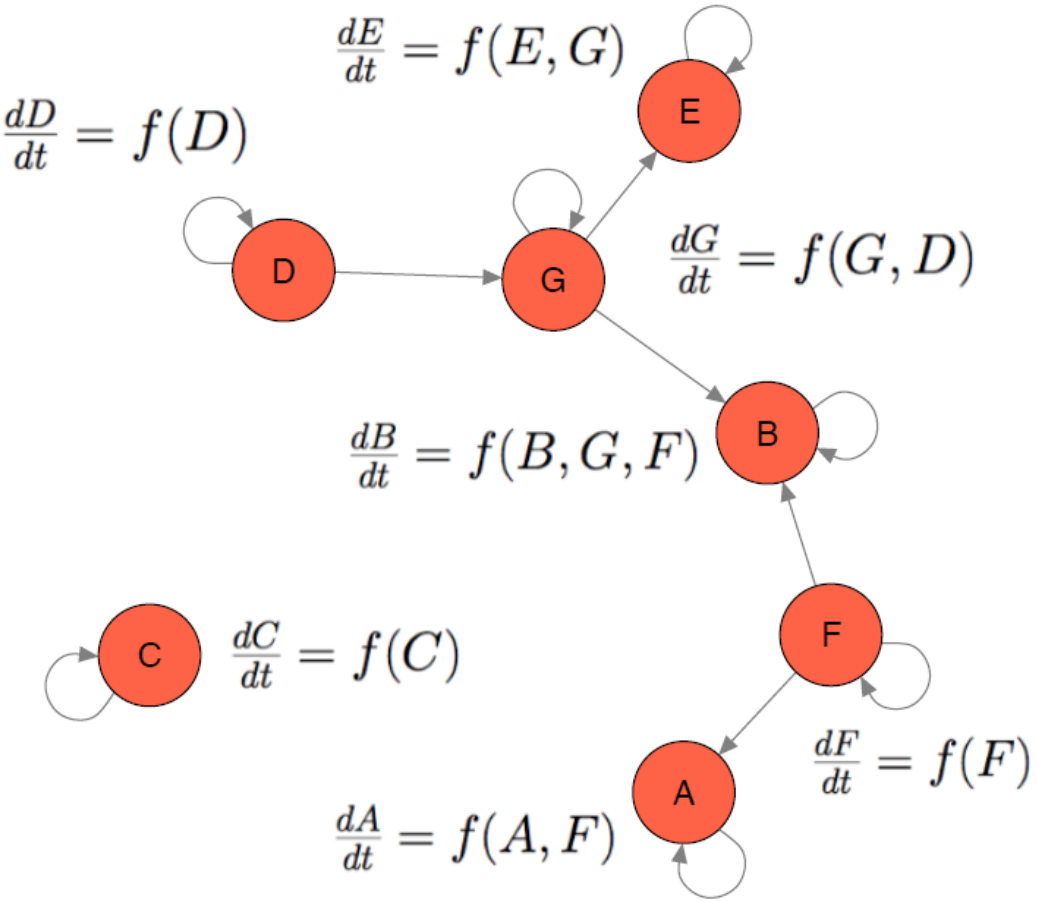
- Knowledge networks are generated from biochemical or biological knowledge and allow interpreting omics data in the context of prior biological knowledge, such as metabolic pathways and enzymatic reactions.
- Experimental networks are generated from the omics data itself, based on relationships between elements in the data (e.g., spectral similarity, or correlation).



Complex System

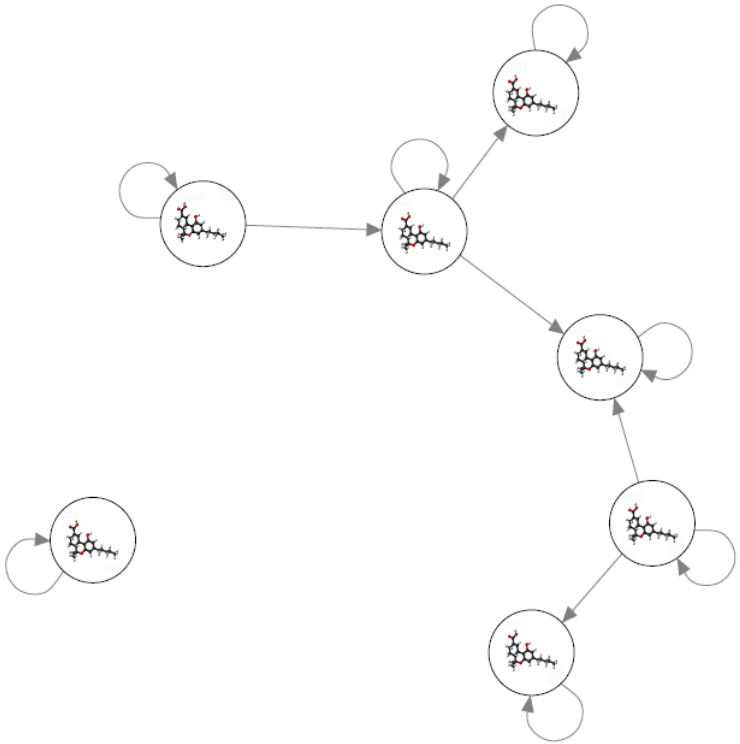


Complex System

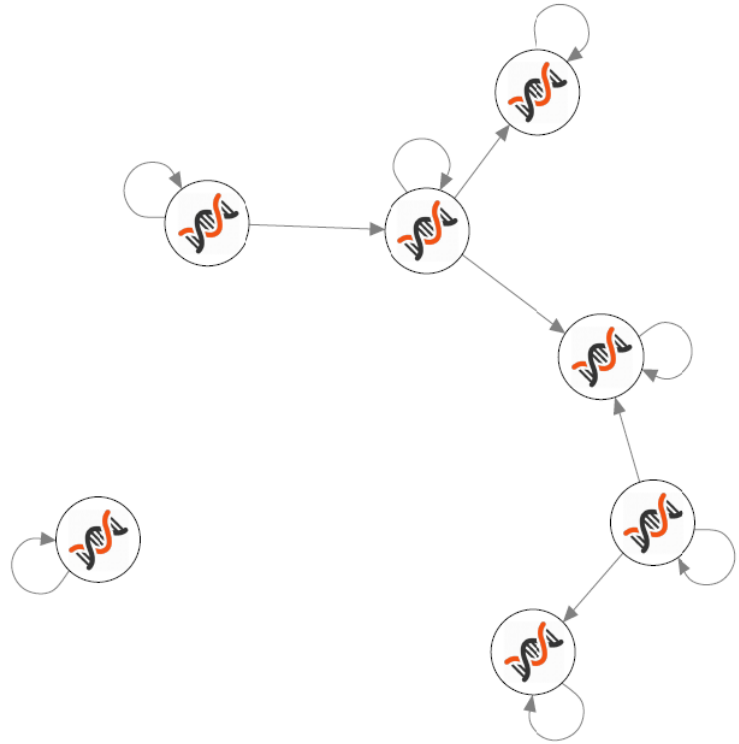


Metabolite concentration or Gene expressions

Metabolite concentration

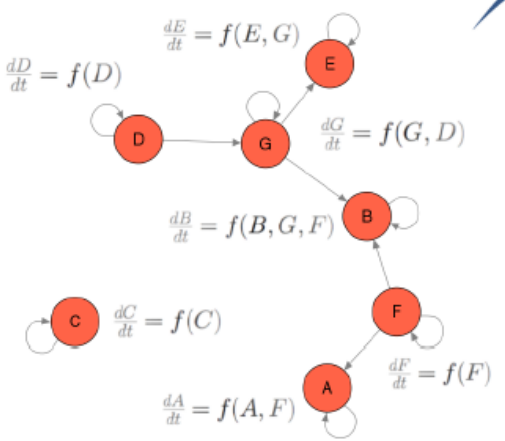


Gene expressions



Observations

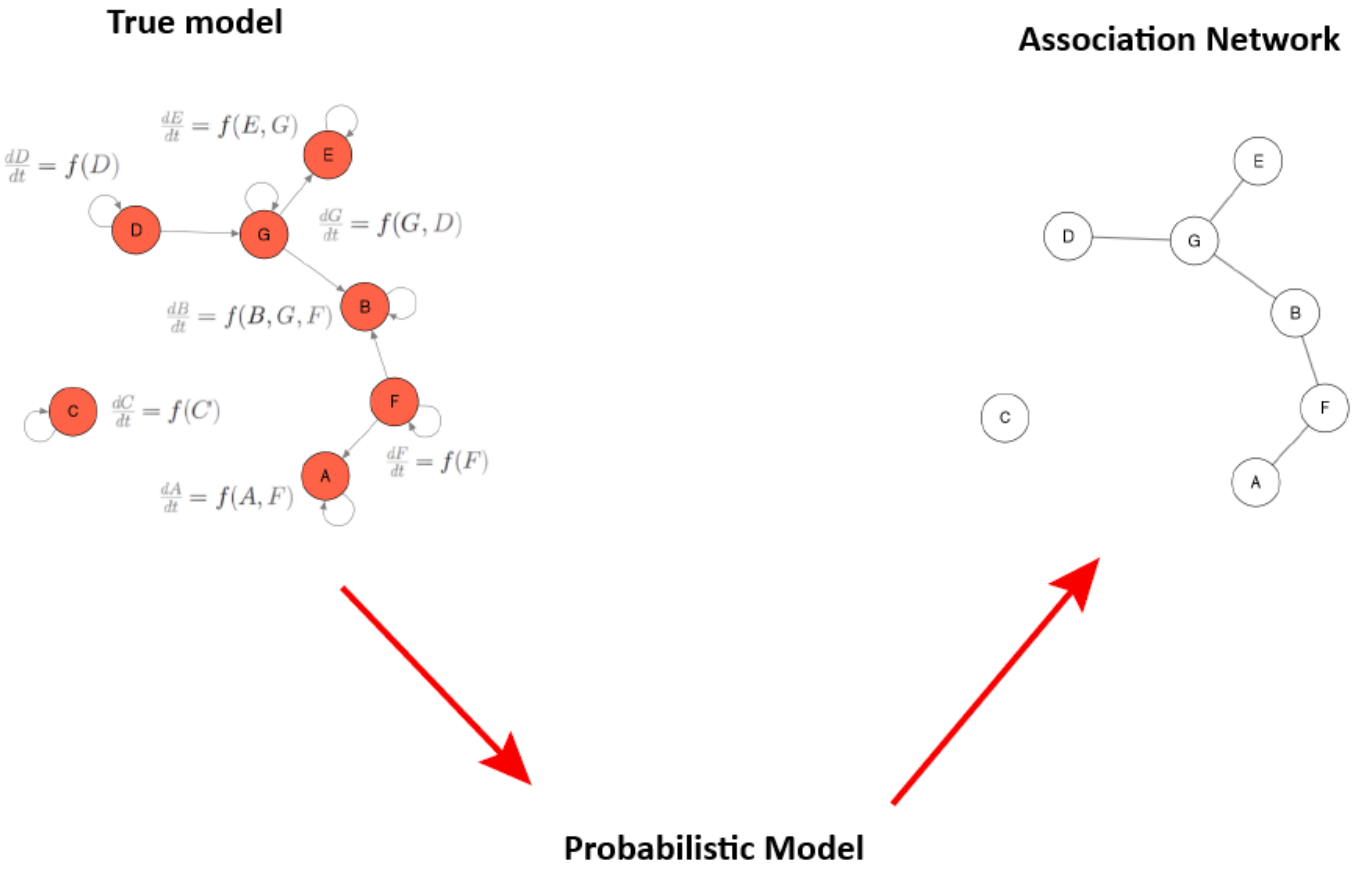
Sample Observations



	A	B	C	D	E	F	G
1	3.45	1.11	−0.91	3.18	−0.88	2.28	2.12
2	1.72	2.14	0.78	0.82	0.36	1.03	1.28
3	0.11	−3.12	1.21	1.80	0.57	0.99	1.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$N-2$	−1.29	1.65	4.62	1.74	−0.44	−2.80	3.13
$N-1$	−0.02	1.54	5.21	1.71	0.04	−0.02	3.00
N	1.18	0.82	1.05	8.77	−0.82	0.05	2.01

Recover the structure of the system

From true models to Networks



Pairwise correlations

- Let's assume a mutivariate normal distribution over all random variables.
- A simple method for inferring the network of (linear) dependencies among a set of variables is to compute all pairwise correlations and subsequently to draw the corresponding graph.
- A major drawback of correlation networks, however, is their inability to distinguish between direct and indirect associations.
- Correlation coefficients are generally high in large-scale omics data sets, suggesting indirect associations.

Conditional Independence

- Suppose that we have a collection of random variables $(X_v)_{v \in V}$ with a joint density. Let A, B and C be subsets of V and let $X_A = (X_v)_{v \in A}$ and similarly for X_B and X_C. Then the statement that X_A **and** X_B **are conditionally independent given** X_C, written

$$A \perp\!\!\!\perp B \mid C,$$

means that for each possible value of x_C of X_C, X_A and X_B are independent in the conditional distribution given $X_C = x_C$. So if we write $f()$ for a generic density or probability mass function, then one characterization of $A \perp\!\!\!\perp B \mid C$ is that

$$f(x_A, x_B \mid x_C) = f(x_A \mid x_C)\, f(x_B \mid x_C).$$

An equivalent characterization is that the joint density of (X_A, X_B, X_C) factorizes as $f(x_A, x_B, x_C) = g(x_A, x_C)\, h(x_B, x_C)$

Gaussian Graphical models (GGMs)

- Gaussian graphical models (GGMs) circumvent indirect association effects by evaluating conditional dependencies in multivariate Gaussian distributions
- To each node j, we associate a random variable X_j, for $j = 1, 2, ..., p$.
- The edges between the variables in the network indicate conditional independence: no edge between node i and j, implies X_i and X_j are independent given all other remaining variables X_k, $X_i \perp\!\!\!\perp X_j \mid X_k$.
- The combination of the network and a distribution over all random variables $X_1, X_2, ..., X_p$ is called a graphical model. When the multivariate distribution over all random variables is Gaussian (normal), we call the network together with the distribution a **Gaussian Graphical Model (GGM)**.
- It turns out that whenever the correlation ρ_{ij} between the variables X_i and X_j is $\neq 0$, then there is a path of edges in the network between nodes i and j.
- However, we cannot determine what the path is by considering only correlations.

Precision matrix

- Suppose the (joint/full) covariance matrix $\Sigma = (\sigma_{ij})$ is positive definite and therefore invertible. The precision matrix or concentration matrix is the matrix inverse of the covariance matrix,

$$\Theta = \Sigma^{-1}.$$

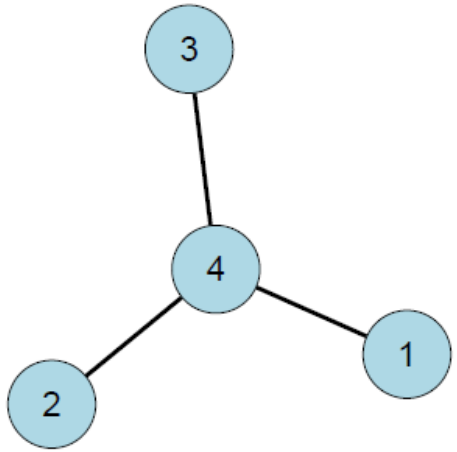
- For univariate distributions, the precision matrix degenerates into a scalar precision, defined as the reciprocal of the variance, $\theta = \frac{1}{\sigma^2}$.
- 0s in the precision matrix Θ play a central role in a GGM (Lauritzen 1996).

$$\theta_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_k$$

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Gaussian Graphical Model

$$\Sigma^{-1} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 3.45 & 0 & 0 & 3.18 \\ 0 & 2.14 & 0 & 0.82 \\ 0 & 0 & 3.21 & 1.05 \\ 3.18 & 0.82 & 1.05 & 8.77 \end{pmatrix} \end{array} \iff$$



$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Partial correlation

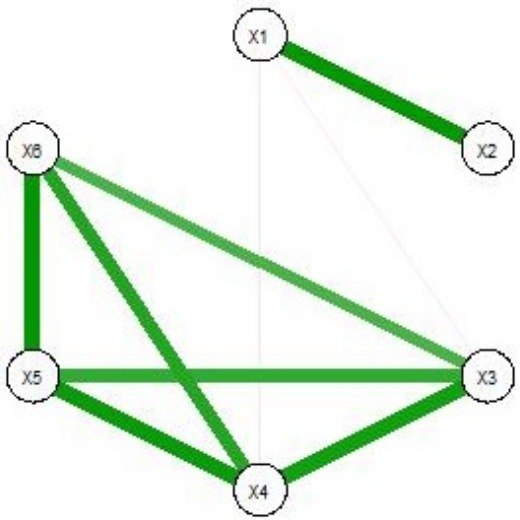
- Because correlations are easier to interpret, the partial correlations are often used in GGMs instead of partial covariances.
- The partial correlation between X and Y given a set of n controlling variables $Z = Z_1, Z_2, ..., Z_n$, written $\rho_{XY \cdot Z}$, is the correlation between the residuals e_X and e_Y resulting from the linear regression of X with Z and of Y with Z, respectively.
- The partial correlation can also be written in terms of the joint precision matrix. Consider a set of random variables, $\mathbf{V} = X_1, \ldots, X_n$ of cardinality n. We want the partial correlation between two variables X_i and X_j given all others, i.e., $\mathbf{V} \setminus \{X_i, X_j\}$. Suppose the (joint/full) covariance matrix $\Sigma = (\sigma_{ij})$ is positive definite and therefore invertible. If the precision matrix is defined as $\Theta = (\theta_{ij}) = \Sigma^{-1}$, then the partial correlation matrix $P = (\rho_{X_i X_j \cdot \mathbf{V} \setminus \{X_i, X_j\}})$ is given by

$$\rho_{X_i X_j \cdot \mathbf{V} \setminus \{X_i, X_j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$$

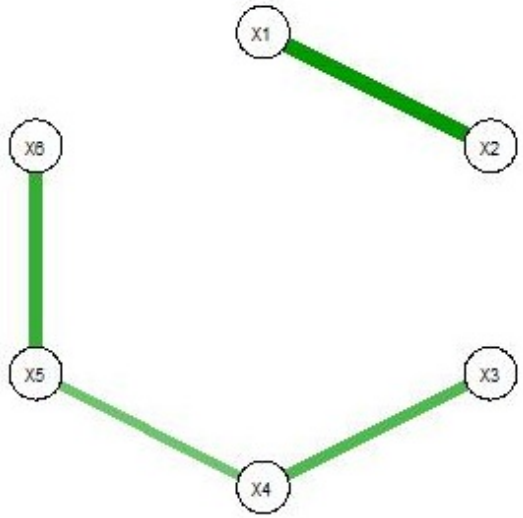
Networks based on the correlation matrix and on the partial correlation matrix

- The key idea behind GGMs is to use partial correlations as a measure of dependence of any two variables.
- A GGM is an undirected graph in which each edge represents the pairwise correlation between two variables conditioned against the correlations with all other variables (partial correlation coefficients).
- Hence, assuming that the partial covariances are 0, leads to the variables being independent when the multivariate distribution is Gaussian.

Based on Correlation matrix



Based on Partial Correlation matrix



Estimation of GGMs via regression coefficients

- In a multiple linear regression framework the regression of variable i on the remaining variables is

$$X_i = \sum_{j \neq i} X_j \beta_{ij} + e_i$$

  The indices ij in the coefficient βij make clear that it is about the relation between variables i and j in the network.

- The relation between the coefficient β_{ij} and the precision matrix Θ is (Lauritzen 1996)

$$\beta_{ij} = -\frac{\theta_{ij}}{\theta_{ii}}$$

So if $\beta_{ij} = 0$ then also $\theta_{ij} = 0$.

Node-wise regression

- In nodewise regressions, we obtain estimates of the edges of the network by considering in turn each node as the dependent variable and the other nodes as predictors.
- We, therefore, make p regressions with each $p-1$ predictors (the remaining variables), and obtain in total $p(p-1)$ estimates, so two for each edge.
- We select a variable X_i, regress it on the rest, and assign an edge between X_i and another variable X_j if the coefficient β_{ij} is nonzero.
- We use linear regression since the mean of a univariate conditional Gaussian is modeled by a linear combination of all other variables.
- Whe $n > p$ the estimates are unique, but when $n < p$ the least squares solution is no longer unique.

Penalized regression

- The covariance matrix is not invertible when we have more variables than individuals because it becomes singular.
- To overcome the problem we can use a penalized form of regression such as the least absolute shrinkage and selection operator (lasso) (Tibshirani 1996).
- In an ordinary least squares regression

$$L_{\beta_i}(\mathbf{x}) = \frac{1}{2n} \sum_{k=1}^{n} \left(x_{ki} - \sum_{j \neq i} x_{kj} \beta_{ij} \right)^2$$

Penalized Regression (2)

For regression with the **lasso**, the coefficients β_{ij} for each node i are obtained by a combination of least squares and a penalty, the sum of absolute values of the coefficients weighted by a penality parameter.

$$L_{\beta_i,\lambda}(\mathbf{x}) = \frac{1}{2n}\sum_{k=1}^{n}\left(x_{ki} - \sum_{j\neq i} x_{kj}\beta_{ij}\right)^2 + \lambda\sum_{j\neq i}|\beta_{ij}|$$

- The lasso has the advantage that by estimating the coefficients β_{ij}, the coefficients are also selected because some of the estimated coefficients will be exactly 0 (Tibshirani 1996).
- Additionally, the lasso has the advantage that it decreases prediction error (i.e., the error between prediction using the estimate and the true parameter).
- The lasso penalty parameter, usually denoted by λ, needs to be determined, and so, an estimate depends on the value λ chosen. Specifically, larger values of λ result in stronger regularization. Often the value for λ is obtained by using cross-validation.

Lasso Technique Under the Generalized Linear Model Framework

- The Lasso technique can be extended to the framework of Generalized Linear Models (GLMs), which allows for modeling a wide range of data types, including binary, count, and continuous data.
- GLMs extend linear models by allowing the response variable y to have a distribution from the exponential family (e.g., binomial, Poisson, normal). The model is specified by:
  - **Link function**: $g(\mu) = \eta$, where μ is the mean of the response variable and η is the linear predictor.
  - **Linear predictor**: $\eta = X\beta$, where X is the matrix of predictors and β is the vector of coefficients.

Lasso Regularization in GLMs

Lasso regularization adds a penalty term to the GLM to perform variable selection and regularization. The objective function for Lasso in the GLM framework is:

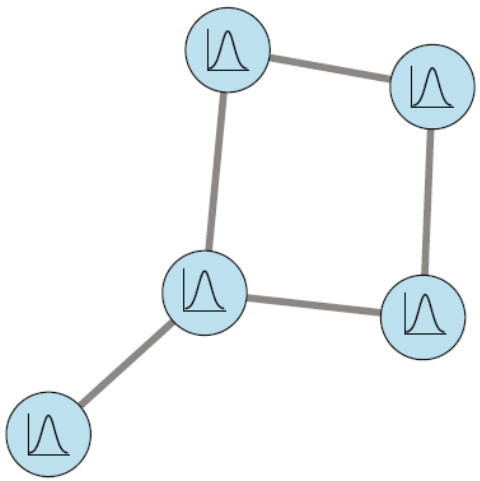
$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, \eta_i) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Here:

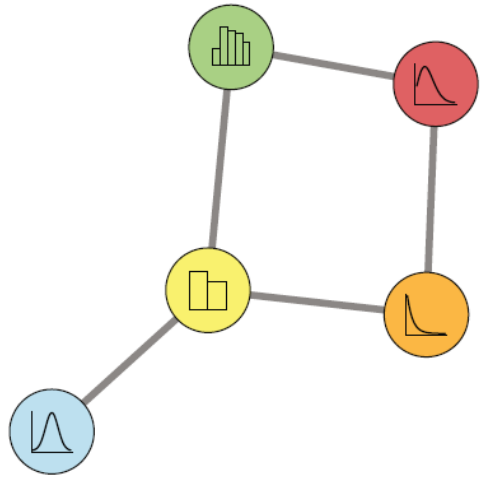
- $\mathcal{L}(y_i, \eta_i)$ is the loss function (e.g., deviance) for the i-th observation.
- $\eta_i = X_i\beta$ is the linear predictor for the i-th observation.
- λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients.
- N is the number of observations.
- p is the number of predictors.

Mixed Graphical Models MGMs

- Mixed Graphical Models **MGM** allow one to combine an arbitrary set of conditional univariate members of the exponential family in a joint distribution.
- A penalized LASSO regression is performed in the Generalized Linear Model (GLM) framework with a link-function appropriate for the node at hand.



Gaussian Graphical Model



Mixed Graphical Model

Estimating Mixed Graphical Models by node-wise regression with regularization

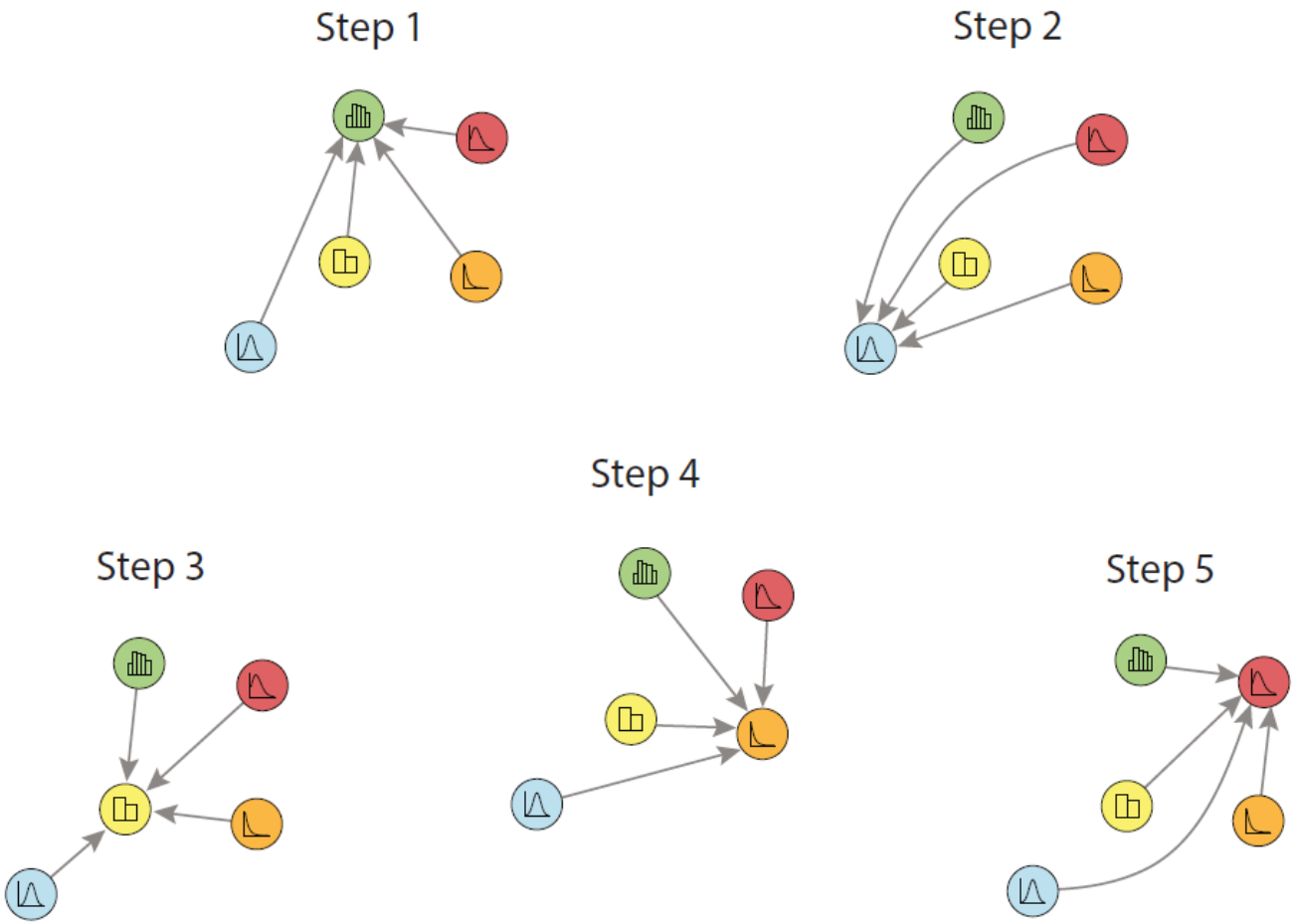
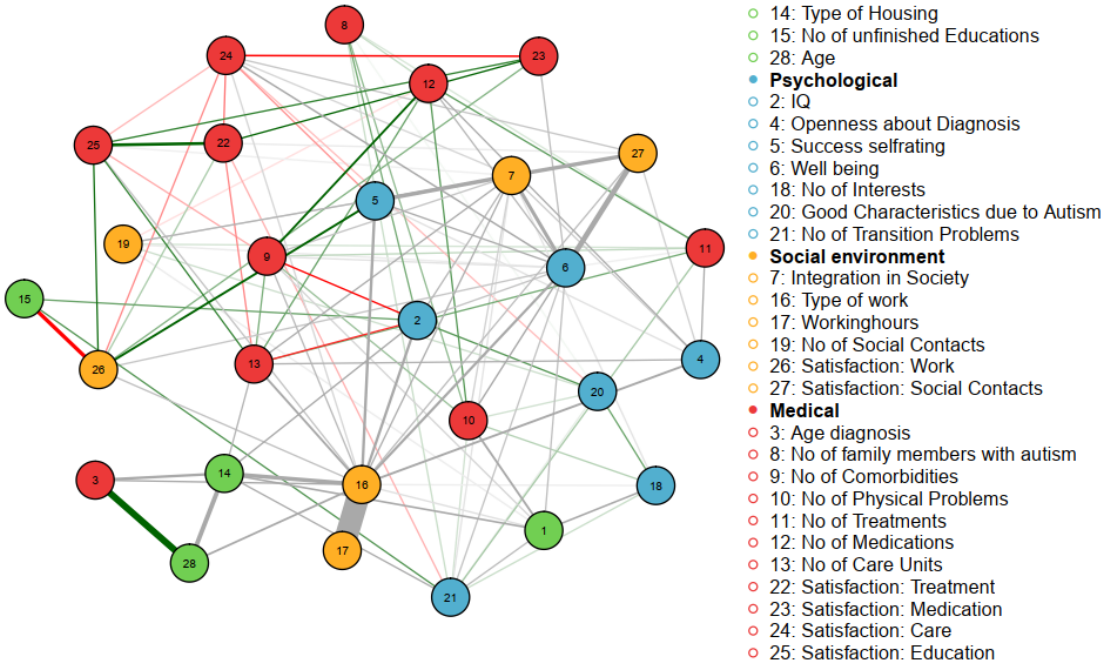


Image: Complexity Lab Utrecht (CLU)

Application: Autism data (Haslbeck and Waldorp 2020)



Application: Resting state fMRI data (Haslbeck and Waldorp 2020)

The dataset consists of BOLD measurements of 68 voxels for 240 time points, where the average sampling frequency is 2 seconds.

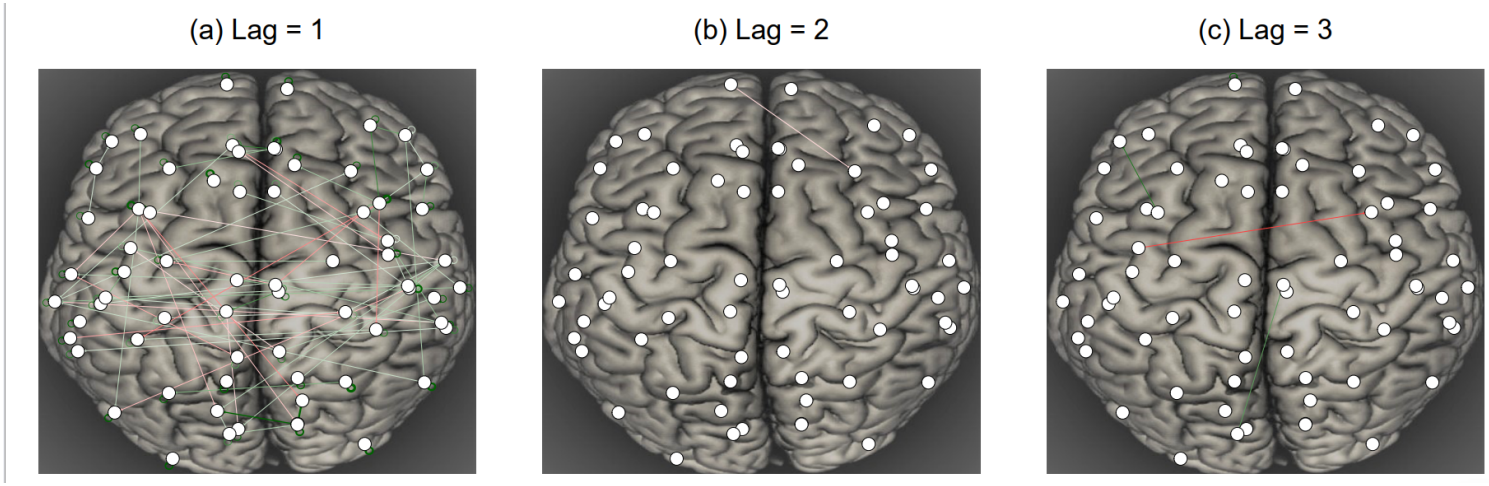


Figure 2: Visualization of the fitted mVAR model, where we depict the parameters separately for each lag. Red edges indicate positive relationships, green edges indicate negative relationships. The width of the edges is proportional to the absolute value of the edge-parameter.

For the lag of size one, many coefficients are nonzero. In contrast, for the lags of size two and three only few coefficients are nonzero. For a typical fMRI data analysis, this could mean that it is

Application (Peron et al. 2021)

CLINICAL NUTRITION

Articles Publish Topics About Contact

ORIGINAL ARTICLE · Volume 40, Issue 10, P5288-5297, October 2021 · *Open Access* Download Full Issue

Crosstalk among intestinal barrier, gut microbiota and serum metabolome after a polyphenol-rich diet in older subjects with "leaky gut": The MaPLE trial

Gregorio Peron [a,b,1] · Giorgio Gargari [c,1] · Tomás Meroño [a,b] · Antonio Miñarro [d] · Esteban Vegas Lozano [d] · Pol Castellano Escuder [a,d] · Raúl González-Domínguez [a,b] · Nicole Hidalgo-Liberona [a,b] · Cristian Del Bo' [c] · Stefano Bernardi [c] · Paul Antony Kroon [e] · Barbara Carrieri [f] · Antonio Cherubini [f] · Patrizia Riso [c] · Simone Guglielmetti [c] · Cristina Andrés-Lacueva [a,b] Show less

Affiliations & Notes Article Info

Show Outline

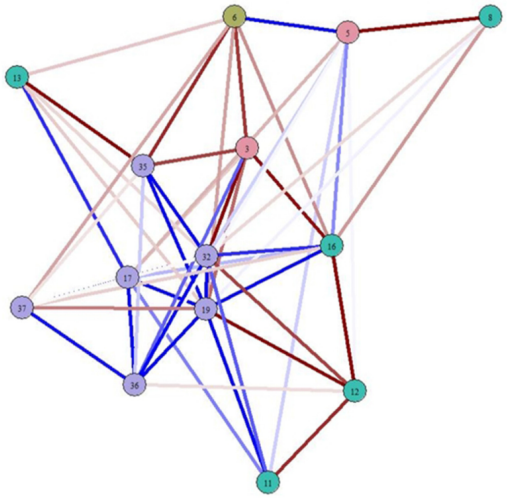
Summary

Background &aim

The MaPLE study was a randomized, controlled, crossover trial involving adults ≥60 y.o. (n = 51) living in a residential care facility during an 8-week polyphenol-rich (PR)-diet. Results from the MaPLE trial showed that the PR-diet reduced the intestinal permeability (IP) in older adults by inducing changes to gut microbiota (GM). The present work aimed at studying the changes in serum metabolome in the MaPLE trial, as a further necessary step to depict the complex crosstalk between dietary polyphenols, GM, and intestinal barrier.

Application (Peron et al. 2021)

PR-diet



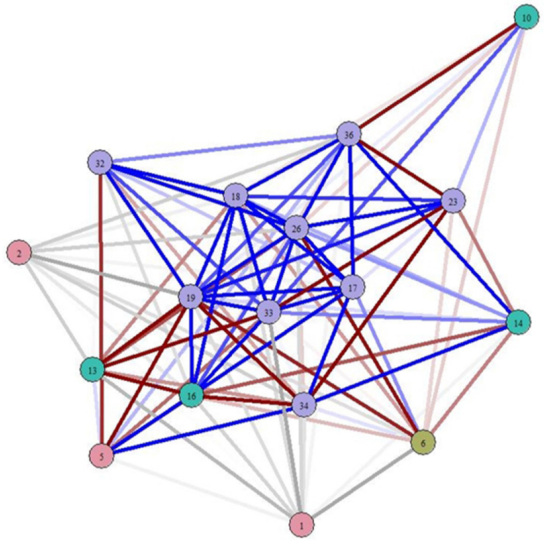
Covariables
- 3 Age
- 5 BL-Zonulin

Response
- 6 Δ Zonulin

Metabolomic
- 8 Δ Deoxycarnitine
- 11 Δ 7-MX
- 12 Δ TB
- 13 Δ HA
- 16 Δ CAT-S

Metagenomic
- 17 Δ *Lachnospiraceae*
- 19 Δ *Coriobacteriaceae*
- 32 Δ *Enterobacteriaceae*
- 35 Δ *Porphyromonadaceae*
- 36 Δ *Verrucomicrobiaceae*
- 37 Δ *Clostridiaceae*

Control diet



Covariables
- 1 Sequence
- 2 Sex
- 5 BL-Zonulin

Response
- 6 Δ Zonulin

Metabolomic
- 10 Δ 3-MX
- 13 Δ HA
- 14 Δ HPPA-S
- 16 Δ CAT-S

Metagenomic
- 17 Δ *Lachnospiraceae*
- 18 Δ *Ruminococcaceae*
- 19 Δ *Coriobacteriaceae*
- 23 Δ *Rikenellaceae*
- 26 Δ *Streptococcaceae*
- 32 Δ *Enterobacteriaceae*
- 33 Δ *Clostridiales_other*
- 34 Δ *Barnesiellaceae*
- 36 Δ *Verrucomicrobiaceae*

Changes in a network

- Let $G = (V, E)$ be a network with nodes $V = \{1, 2, ..., m\}$ and edge set $E \subseteq V \times V$. Changes in G can be due to changes in:
  - its nodes, V,
  - its edges, E,
  - both.
- Changes in the node set are common in social and communication networks, where both V and E can change as the network grows over time.
- We focus on the setting where the node set V is fixed and the goal is to identify changes in network edges, E.

Measures of difference between networks

- We restrict the discussion to comparing two networks, G^1 and G^2 with the same node set V and edges sets E^1 and E^2, or, equivalently, adjacency matrices A^1 and A^2.
- E^1 and E^2 may have been directly observed, obtained from experiments, or learned from observations on the nodes via graphical modeling approaches.
- We focus primarily on networks inferred using graphical modeling methods. For instance, in the case of GGMs, $A^s, s \in \{1, 2\}$ may correspond to estimated partial correlation matrices $\hat{P}^1, \hat{P}^2$

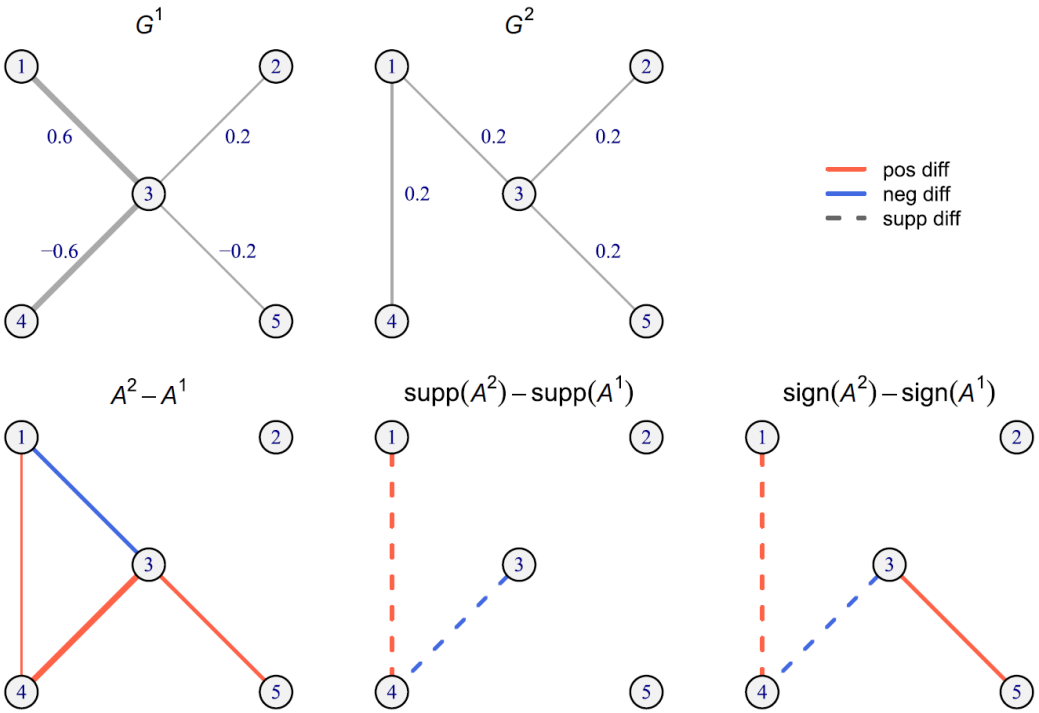
Global differences between networks

- The question here is if $A^1 = A^2$. Different norms or distance measures can be used to assess whether A^1 and A^2 are the same.
- For instance $||A^1 - A^2||$ for some matrix norm. In the case of GGMs $||\hat{P}^1 - \hat{P}^2||$.
- The topology of the space of networks offers additional measures of differences between A^1 and A^2, including (potentially vector-valued) summary measures of the two networks:
  - size and/or number of clusters
  - the average connectivity
  - the degree distribution

Local differences between networks

- In many applications, local differences between the two networks, including differences in individual edges, neighborhoods, or subnetworks, can also be of interest. This is especially the case in biological applications, where network-based biomarkers can be used to interrogate mechanisms of disease initiation and progression.
- Identifying local differences between networks can also be of interest following an affirmative global test of difference between the two networks.
- Local differences between two networks can be assessed qualitatively or quantitatively:
  - $\hat{P}^1_{jk} \neq \hat{P}^2_{jk}$
  - $j - k \in G^1$ but $j - k \notin G^2$

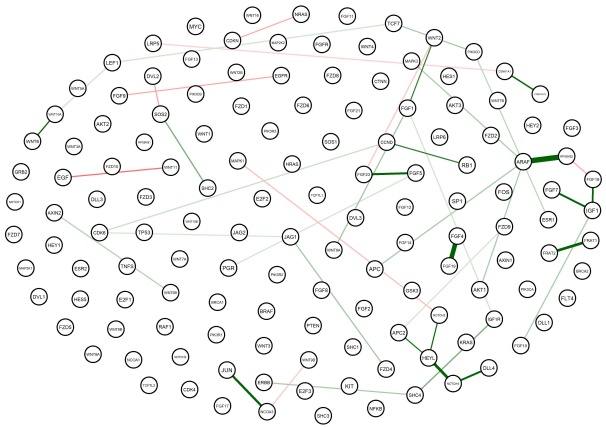
Examples of quantitative and qualitative differences in networks

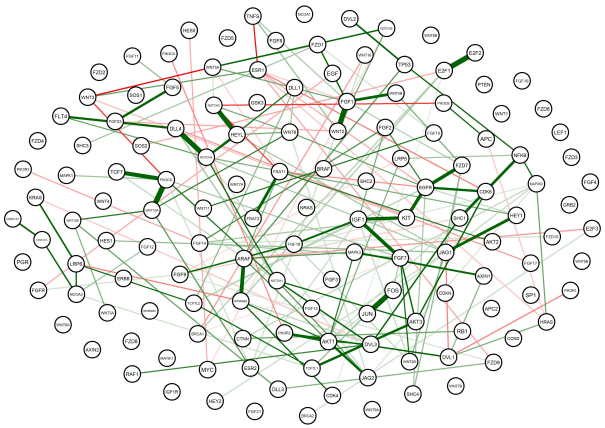


TCGA breast cancer gene expression dataset

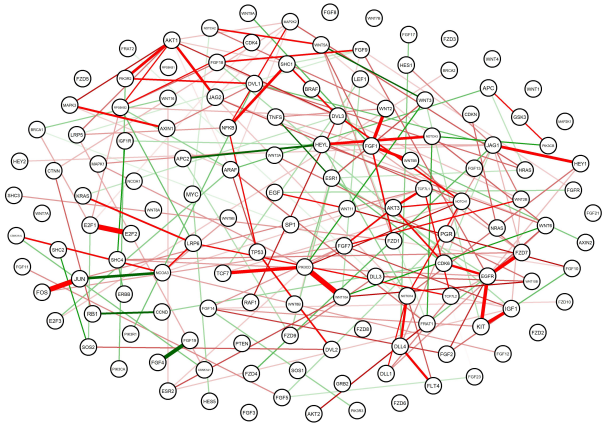
The dataset is obtained from the TCGA database. It includes gene expression measurements for 231 luminal A cancers and 95 basal-like cancers.

The data only includes expression measurement of genes that overlap with the breast cancer pathway (hsa05224) from the KEGG database. It includes an expression matrix (samples $\times$ genes) and a vector denoting group membership.



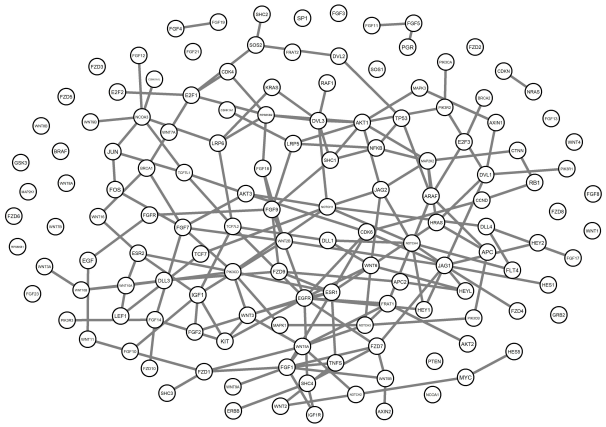


TCGA breast cancer gene expression dataset (2)



Adj_B - Adj_L: 194 different edges

TCGA breast cancer gene expression dataset (3)

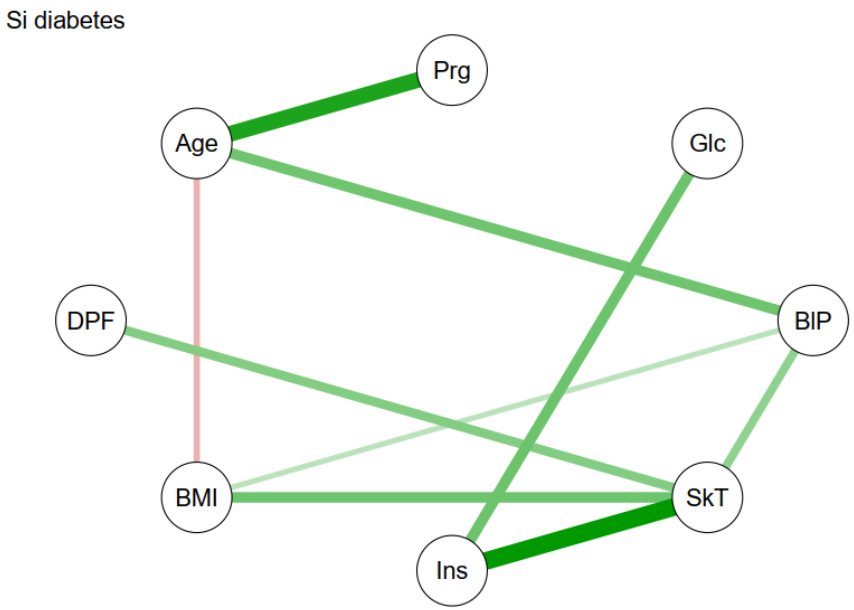


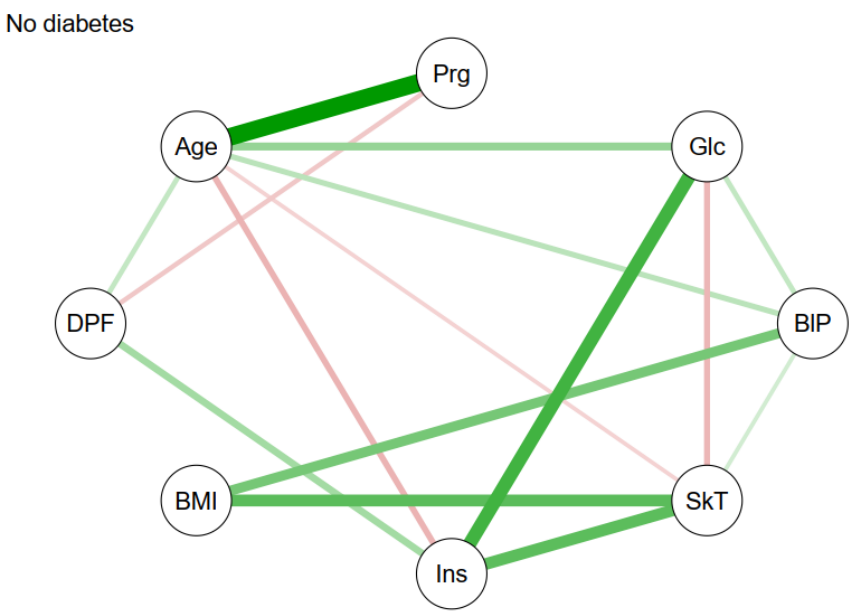
supp(Adj_B) - supp(Adj_L): 179 different edges

Network structure comparison via permutation testing

1. Estimate the network structures from the observed data
2. Select and calculate the value of a test statistic
3. Repeatedly
    a. Pool the data sets and resample
    b. Estimate the network structures
    c. Calculate the test statistic
4. Evaluate the significance

TFG 2024 Núria Serra Pons





TFG 2024 Núria Serra Pons (2)

Based in (Borkulo et al. 2023)

A^G denotes a symmetric $p \times p$ matrix with edge weights of a graphical model G

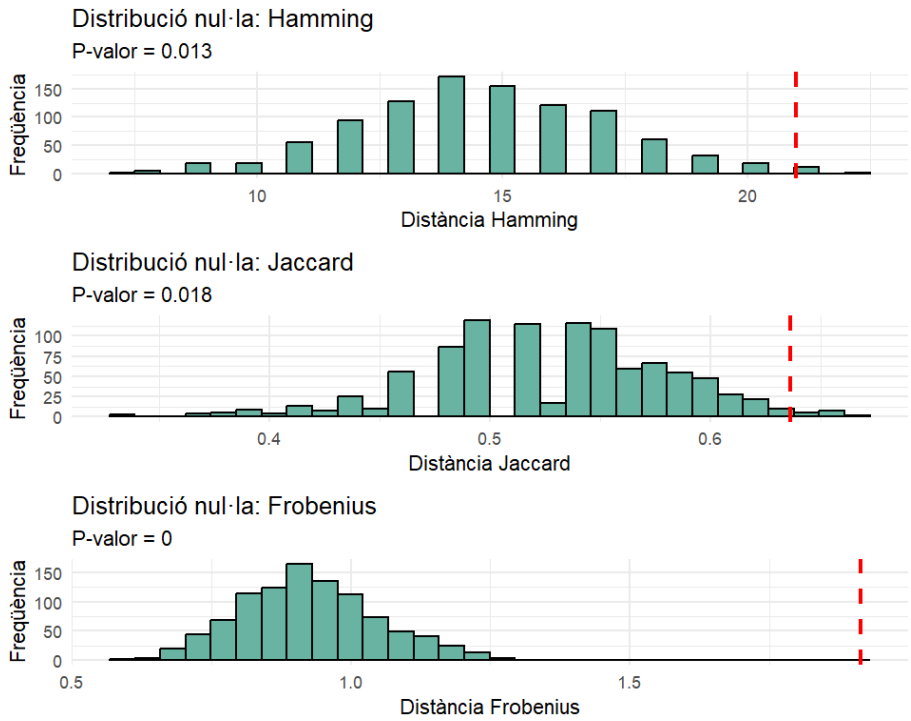
Test	Hypothesis	Test statistic	p-value
INS[1]	$H_0 : A^1 = A^2$	$M(A^1, A^2) = \max_{ij} \lvert a^1_{ij} - a^2_{ij} \rvert$	0.2208
IGS[2]	$H_0 : \sum_{i=1}^{p} \sum_{j>i} \left\lvert a^1_{ij} \right\rvert = \sum_{i=1}^{p} \sum_{j>i} \left\lvert a^2_{ij} \right\rvert$	$S(A^1, A^2) = \left\lvert \sum_{i=1}^{p} \sum_{j>p} (\lvert a^1_{ij} \rvert - \lvert a^2_{ij} \rvert) \right\rvert$	0.0249

1 **Invariant Network Structure (INS)**: all edges in the networks as a whole could be compared in an omnibus test. An efficient way to use resampling for testing differences between all edge weights, is using the maximum statistic.

2 **Invariant Global Strength (IGS)**: the overall level of connectivity is the same across subpopulations

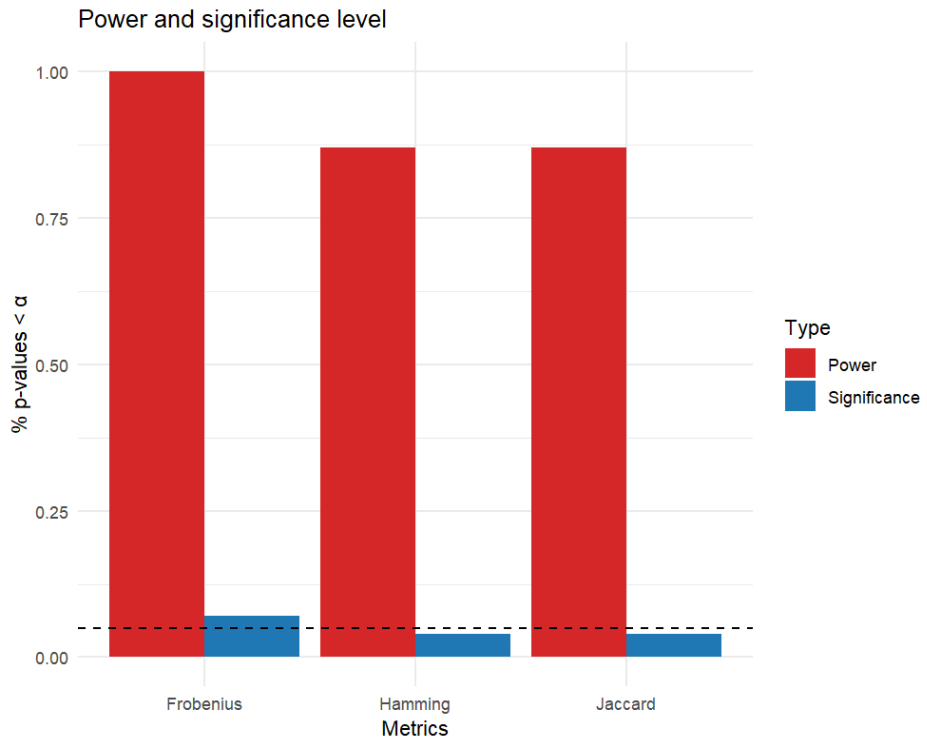
Simulation studies (1)

Two different network structures are defined via adjacency matrices, which are converted into precision matrices to simulate multivariate normal data. These datasets are then used to estimate partial correlation networks, which are compared using distance metrics and statistical testing via bootstrap and permutation methods.



Simulation studies (2)

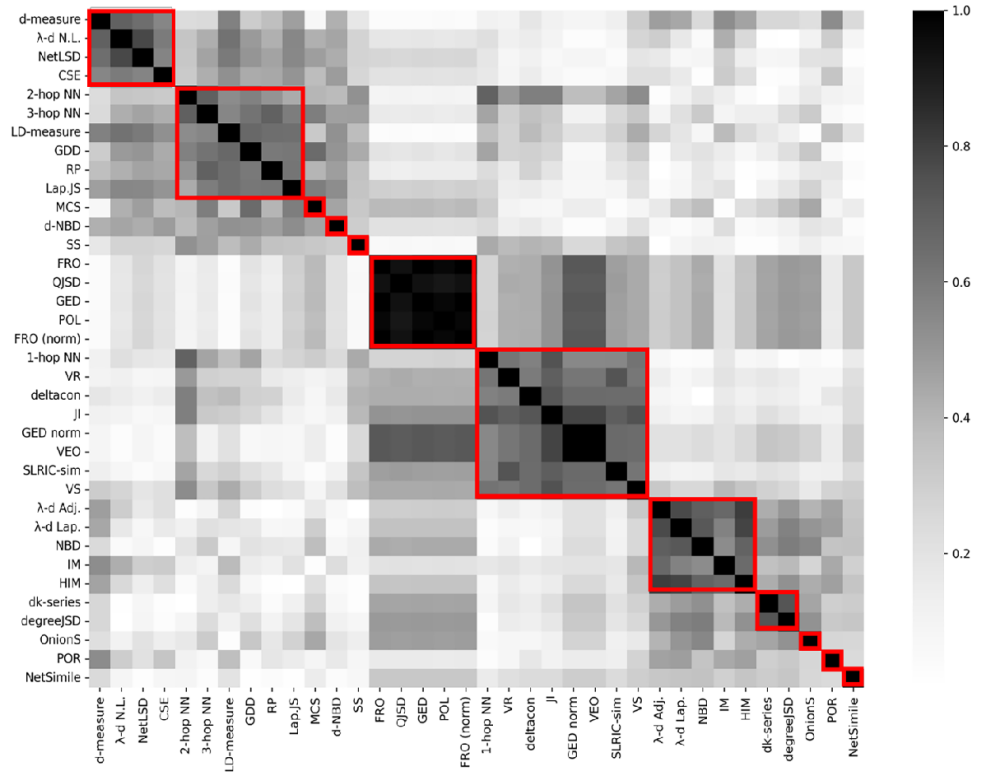
Type	Hamming	Jaccard	Frobenius
Level of significance	0.04	0.04	0.07
Power	0.87	0.87	1



Many more metrics

We've worked on three metrics but there are many more (Shvydun 2023)

1	Jaccard index (JI)
2	Graph edit distance (GED)
3	Vertex/edge overlap (VEO)
4	k-hop nodes neighborhood (k-hop NN)
5	Maximum common subgraph distance (MCS)
6	Frobenius distance (FRO)
7	Vector similarity algorithm (VS)
8	DELTACON
9	Polynomial dissimilarity (POL)
10	Graph diffusion distance (GDD)
11	Resistance perturbation (RP)
12	Vertex ranking (VR)
13	Degree Jenson-Shannon divergence (degreeJSD)
14	Portrait divergence (POR)
15	Communicability sequence entropy (CSE)
16	NetSimile measure
17	Onion spectrum (OnionS)
18	dk-series
19	λ-distances
20	Ipsen–Mikhailov (IM) distance
21	Hamming–Ipsen–Mikhailov (HIM) distance
22	NetLSD measure
23	Non-backtracking spectral distance (NBD)
24	Distributional non-backtracking spectral distance (d-NBD)
25	Quantum Jensen-Shannon divergence (QJSD)
26	Signature similarity (SimHash)
27	D-measure
28	Layer difference (LD)
29	SLRIC similarity



Software used

All calculations have been made with the software R version 4.5.0 (2025-04-11 ucrt) (R Core Team 2024).

- R packages used:
  - mgm (Haslbeck and Waldorp 2020)
  - qgraph (Epskamp et al. 2012)

References

Amara, Adam, Clément Frainay, Fabien Jourdan, Thomas Naake, Steffen Neumann, Elva María Novoa-del-Toro, Reza M Salek, Liesa Salzer, Sarah Scharfenberg, and Michael Witting. 2022. "Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation." *Frontiers in Molecular Biosciences* Volume 9 - 2022. https://doi.org/10.3389/fmolb.2022.841373.

Borkulo, Christina D. van, R. van Bork, L. Boschloo, J. J. Kossakowski, P. Tio, R. A. Schoevers, D. Borsboom, and L. J. Waldorp. 2023. "Comparing Network Structures on Three Aspects: A Permutation Test." *Psychological Methods* 28 (6): 1273–85. https://doi.org/10.1037/met0000476.

Epskamp, Sacha, Angélique O. J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, and Denny Borsboom. 2012. "qgraph: Network Visualizations of Relationships in Psychometric Data." *Journal of Statistical Software* 48 (4): 1–18.

Haslbeck, Jonas M. B., and Lourens J. Waldorp. 2020. "mgm: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data." *Journal of Statistical Software* 93 (8): 1–46. https://doi.org/10.18637/jss.v093.i08.

Lauritzen, Steffen L. 1996. *Graphical Models*. Oxford, England: Oxford University Press.

Peron, Gregorio, Giorgio Gargari, Tomás Meroño, Antonio Miñarro, Esteban Vegas Lozano, Pol Castellano Escuder, Raúl González-Domínguez, et al. 2021. "Crosstalk Among Intestinal Barrier, Gut Microbiota and Serum Metabolome After a Polyphenol-Rich Diet in Older Subjects with 'Leaky Gut': The MaPLE Trial." *Clinical Nutrition* 40 (10): 5288–97. https://doi.org/10.1016/j.clnu.2021.08.027.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Shvydun, S. 2023. "Models of Similarity in Complex Networks." *PeerJ Computer Science* 9: e1371. https://doi.org/10.7717/peerj-cs.1371.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88. http://www.jstor.org/stable/2346178.

Thank You!