Part I: Single Networks

IBC2022 short course: Network Modeling for High-Dimensional Data

July 10th, 2022, Carel F.W. Peeters

Mathematical & Statistical Methods group — Biometris, Wageningen University & Research

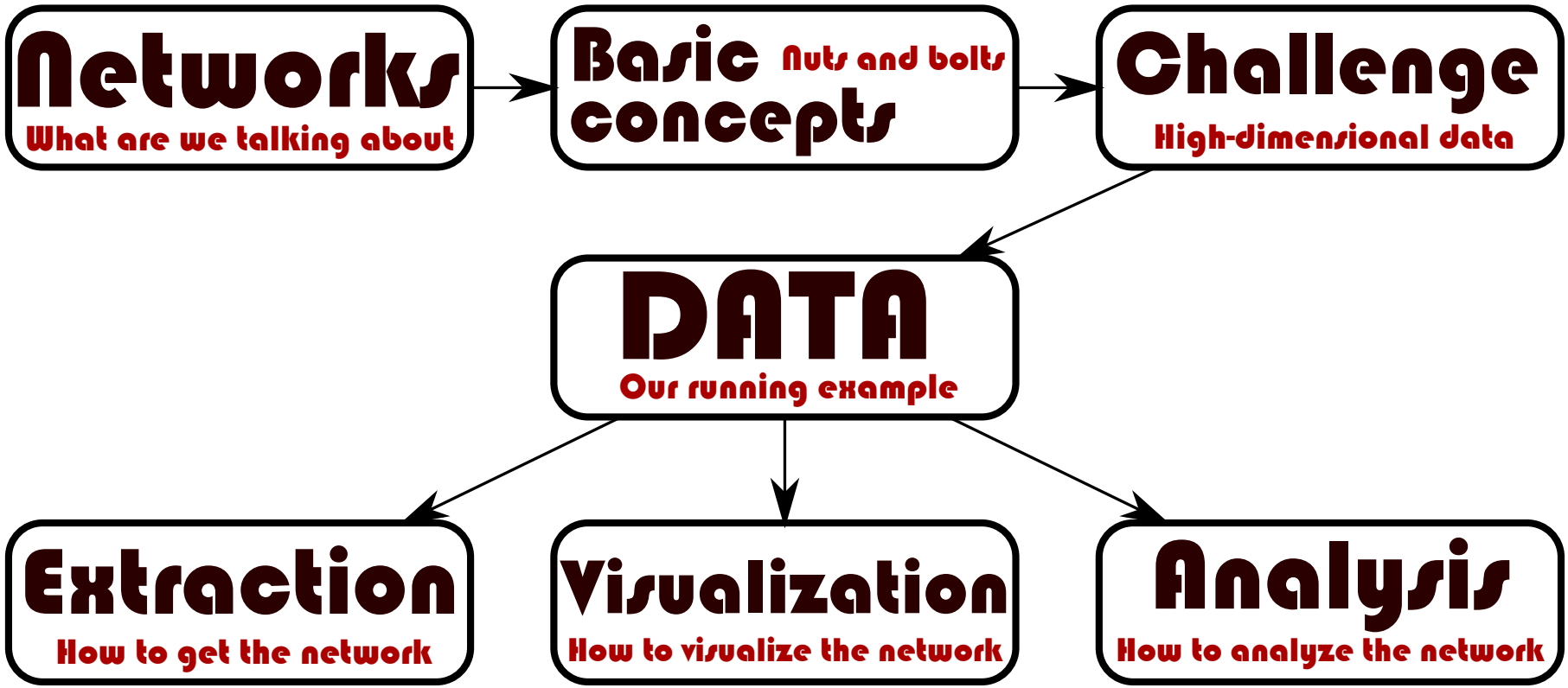
Schedule:

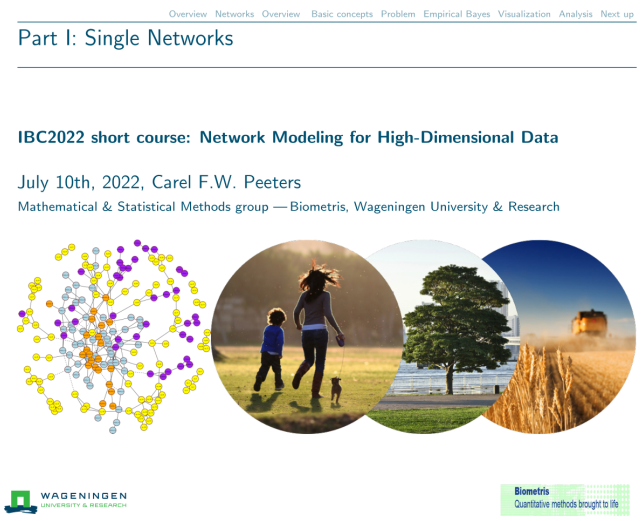
The course is divided a 4 submodules. Each submodule consists of a short lecture and a corresponding hands-on practical.

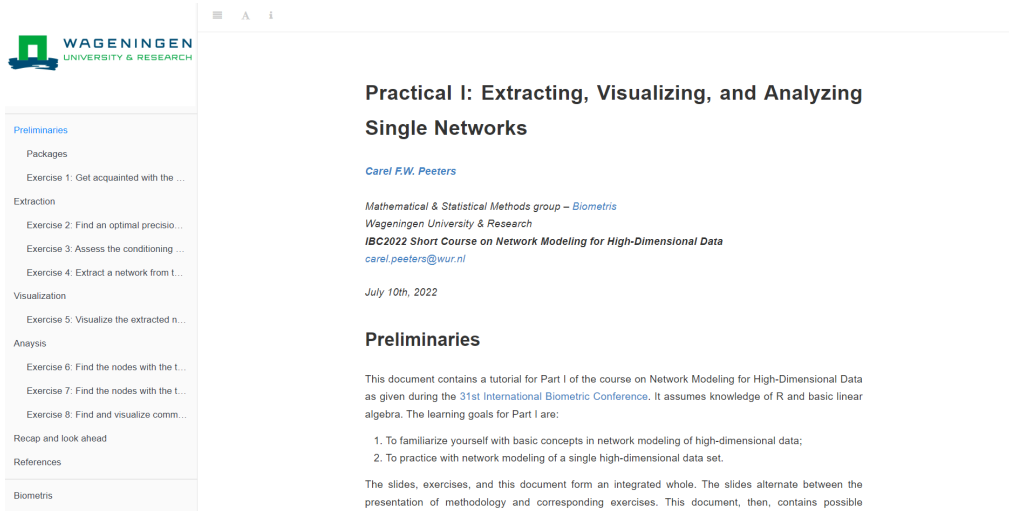
09:00 – 10:30	**Submodule 1: Extracting, visualizing and analyzing single networks** Associated literature: DOI R packages used: rags2ridges
10:30 – 11:00	Break
11:00 – 12:30	**Submodule 2: Jointly extracting, visualizing and analyzing multiple networks** Associated literature: DOI R packages used: rags2ridges
12:30 – 13:30	Lunch break
13:30 – 15:00	**Submodule 3: Extracting, visualizing and analyzing networks from time-course data** Associated literature: DOI R packages used: ragt2ridges
15:00 – 15:30	Break
15:30 – 17:00	**Submodule 4: Miscellanea and extensions** Associated literature: R packages used: porridge

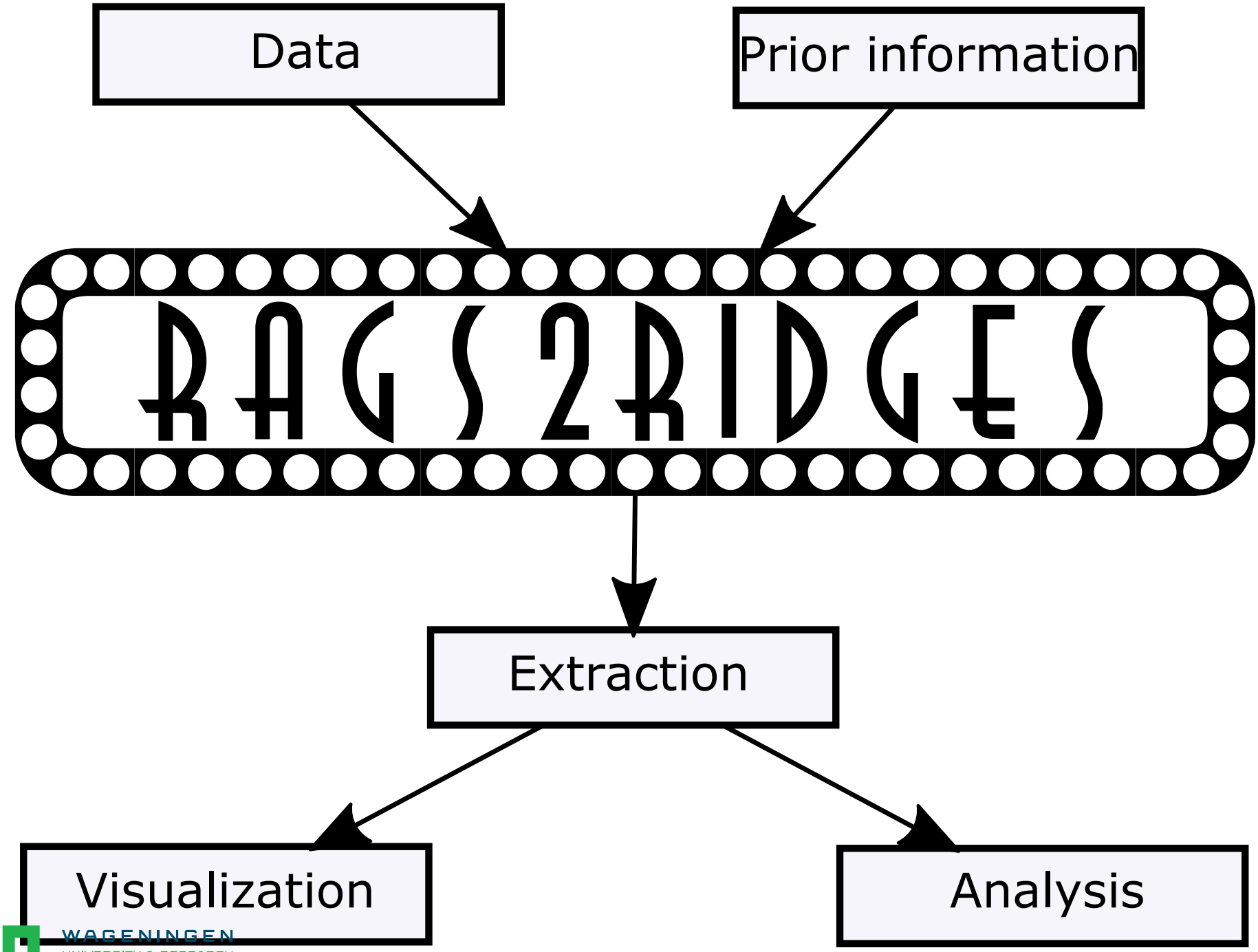
Carel F.W. Peeters
Wageningen University & Research

Wessel N. van Wieringen
Amsterdam University medical centers
VU University Amsterdam


Networks
What are we talking about
Basic concepts
Nuts and bolts
Challenge
High-dimensional data
DATA
Our running example
Extraction
How to get the network
Visualization
How to visualize the network
Analysis
How to analyze the network


Overview Networks Overview Basic concepts Problem Empirical Bayes Visualization Analysis Next up
Part I: Single Networks
IBC2022 short course: Network Modeling for High-Dimensional Data
July 10th, 2022, Carel F.W. Peeters
Mathematical & Statistical Methods group — Biometris, Wageningen University & Research
WAGENINGEN
UNIVERSITY & RESEARCH
Biometris
Quantitative methods brought to life


WAGENINGEN
UNIVERSITY & RESEARCH
Preliminaries
Packages
Exercise 1: Get acquainted with the ...
Extraction
Exercise 2: Find an optimal precisio...
Exercise 3: Assess the conditioning ...
Exercise 4: Extract a network from t...
Visualization
Exercise 5: Visualize the extracted n...
Anaysis
Exercise 6: Find the nodes with the t...
Exercise 7: Find the nodes with the t...
Exercise 8: Find and visualize comm...
Recap and look ahead
References
Biometris
Practical I: Extracting, Visualizing, and Analyzing Single Networks
Carel F.W. Peeters
Mathematical & Statistical Methods group – Biometris
Wageningen University & Research
IBC2022 Short Course on Network Modeling for High-Dimensional Data
carel.peeters@wur.nl
July 10th, 2022
Preliminaries
This document contains a tutorial for Part I of the course on Network Modeling for High-Dimensional Data as given during the 31st International Biometric Conference. It assumes knowledge of R and basic linear algebra. The learning goals for Part I are:
1. To familiarize yourself with basic concepts in network modeling of high-dimensional data;
2. To practice with network modeling of a single high-dimensional data set.
The slides, exercises, and this document form an integrated whole. The slides alternate between the presentation of methodology and corresponding exercises. This document, then, contains possible

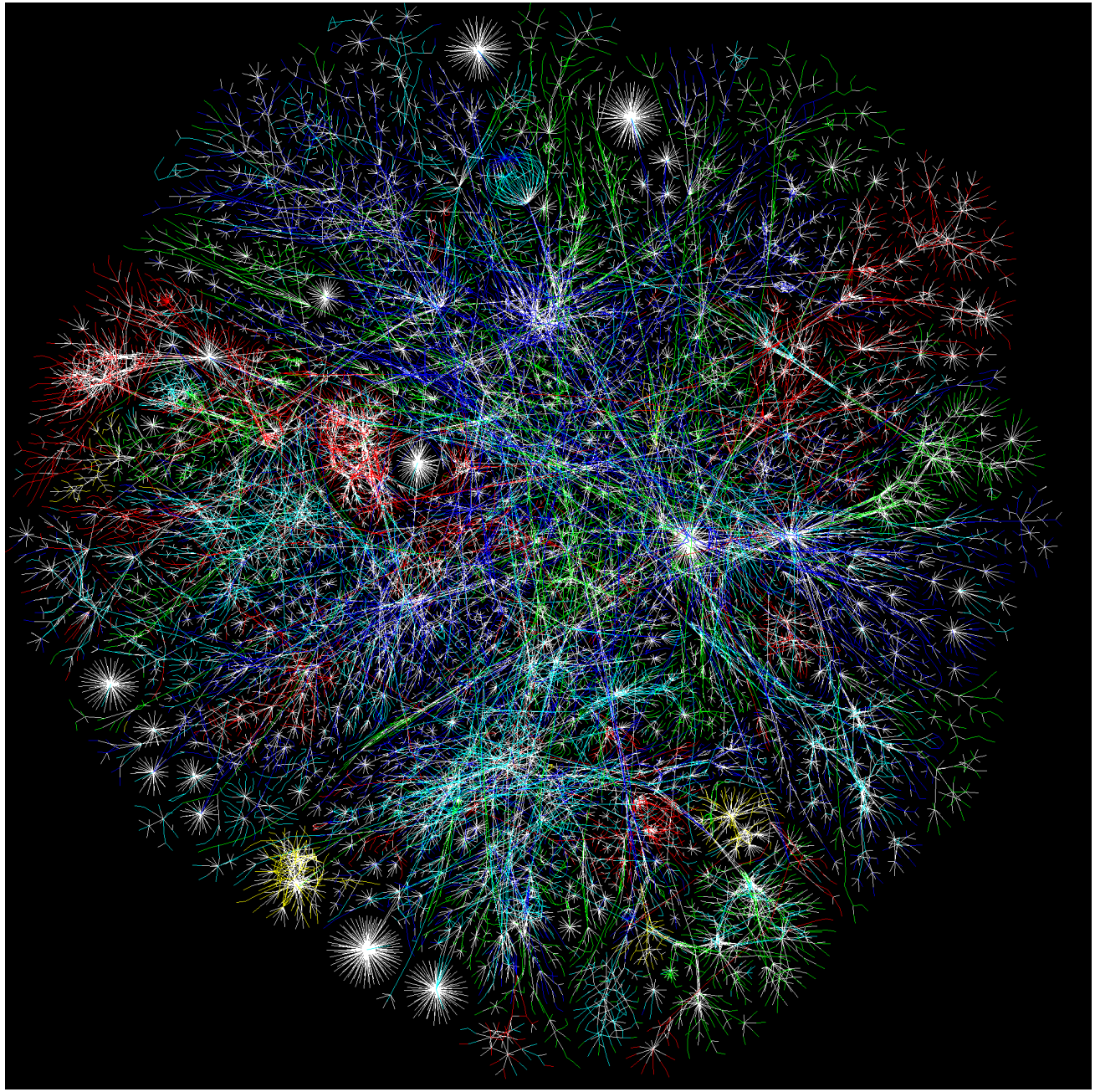

Data
Prior information
RAGS2RIDGES
Extraction
Visualization
Analysis

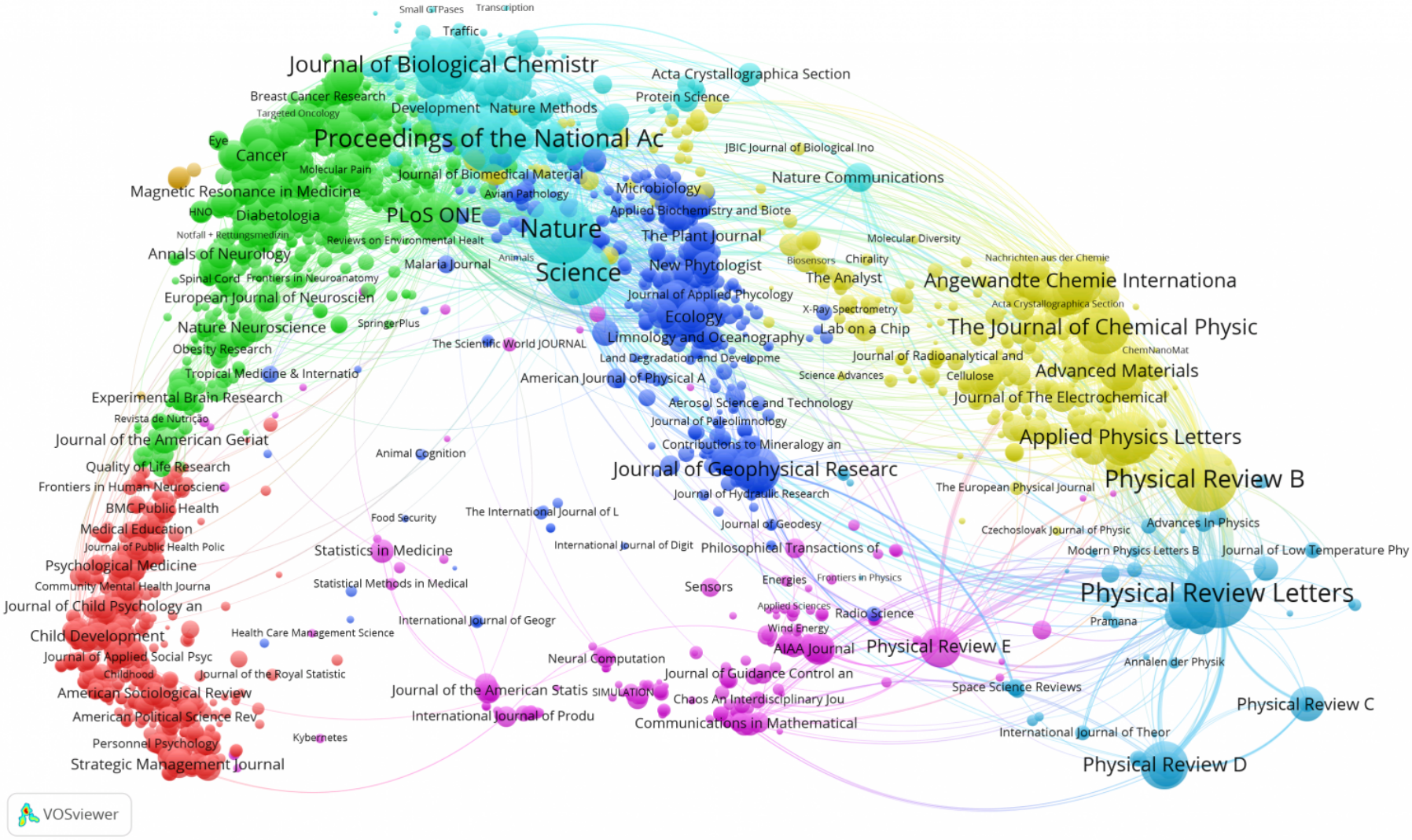

¨This has put the `IT´ in KITT¨
- The Hoff

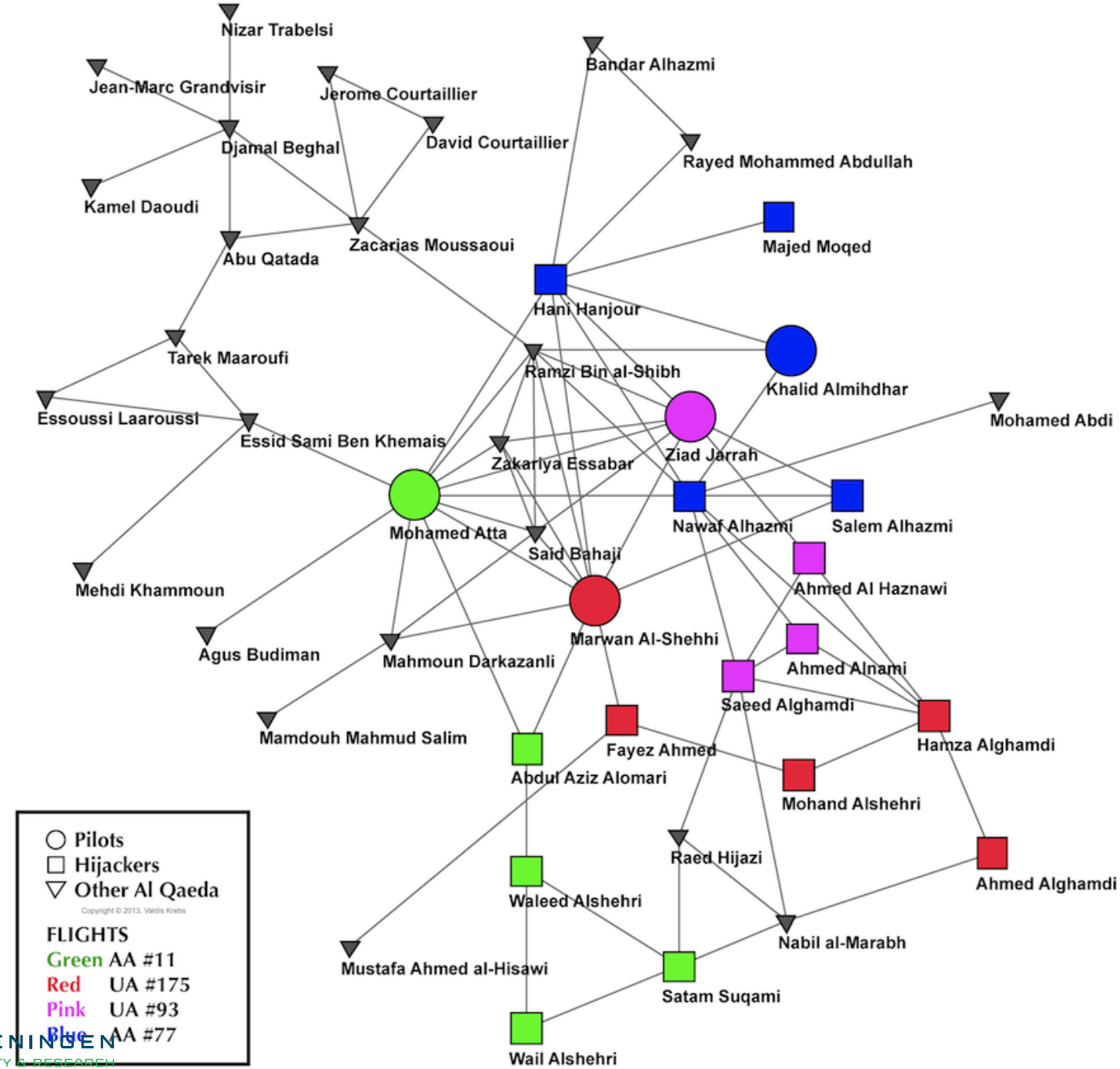
Networks

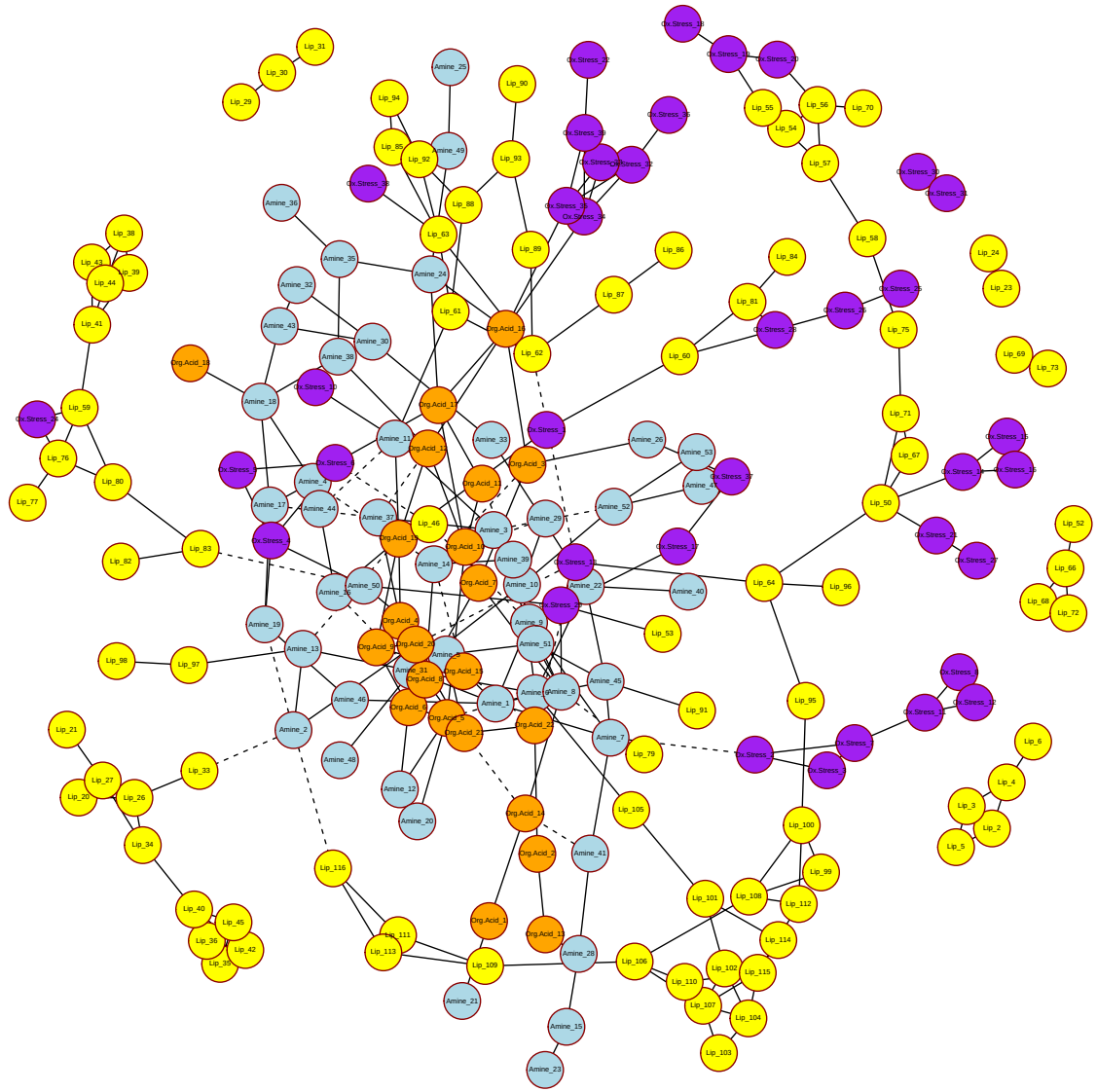
What are we talking about


NO!




Small GTPases
Transcription
Traffic
Journal of Biological Chemistr
Acta Crystallographica Section
Protein Science
Breast Cancer Research
Development
Nature Methods
Targeted Oncology
Proceedings of the National Ac
Eye
Cancer
JBIC Journal of Biological Ino
Molecular Pain
Journal of Biomedical Material
Nature Communications
Magnetic Resonance in Medicine
Avian Pathology
Microbiology
HNO
Diabetologia
PLoS ONE
Applied Biochemistry and Biote
Nature
Notfall + Rettungsmedizin
Reviews on Environmental Healt
The Plant Journal
Molecular Diversity
Annals of Neurology
Malaria Journal
Animals
Science
New Phytologist
Biosensors
Chirality
Nachrichten aus der Chemie
Spinal Cord
Frontiers in Neuroanatomy
The Analyst
Angewandte Chemie Internationa
European Journal of Neuroscien
Journal of Applied Phycology
Acta Crystallographica Section
Ecology
X-Ray Spectrometry
Nature Neuroscience
SpringerPlus
Lab on a Chip
The Journal of Chemical Physic
Limnology and Oceanography
Obesity Research
The Scientific World JOURNAL
ChemNanoMat
Land Degradation and Developme
Journal of Radioanalytical and
Tropical Medicine & Internatio
American Journal of Physical A
Science Advances
Cellulose
Advanced Materials
Experimental Brain Research
Aerosol Science and Technology
Journal of The Electrochemical
Revista de Nutrição
Journal of Paleolimnology
Journal of the American Geriat
Applied Physics Letters
Contributions to Mineralogy an
Animal Cognition
Quality of Life Research
Journal of Geophysical Researc
Physical Review B
Frontiers in Human Neuroscienc
Journal of Hydraulic Research
The European Physical Journal
BMC Public Health
Food Security
The International Journal of L
Journal of Geodesy
Czechoslovak Journal of Physic
Advances In Physics
Medical Education
International Journal of Digit
Philosophical Transactions of
Modern Physics Letters B
Journal of Low Temperature Phy
Journal of Public Health Polic
Statistics in Medicine
Psychological Medicine
Energies
Frontiers in Physics
Sensors
Physical Review Letters
Community Mental Health Journa
Statistical Methods in Medical
Journal of Child Psychology an
Applied Sciences
Radio Science
Pramana
International Journal of Geogr
Child Development
Health Care Management Science
Wind Energy
AIAA Journal
Physical Review E
Journal of Applied Social Psyc
Neural Computation
Annalen der Physik
Childhood
Journal of the Royal Statistic
Journal of Guidance Control an
American Sociological Review
Journal of the American Statis
SIMULATION
Space Science Reviews
Chaos An Interdisciplinary Jou
Physical Review C
American Political Science Rev
International Journal of Produ
Communications in Mathematical
Kybernetes
International Journal of Theor
Personnel Psychology
Physical Review D
Strategic Management Journal
VOSviewer






Basic concepts
Nuts and bolts

Network

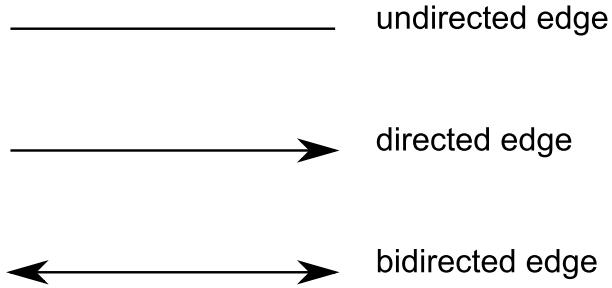
Collection of nodes that may have pairwise relationships. Represented by *graph*

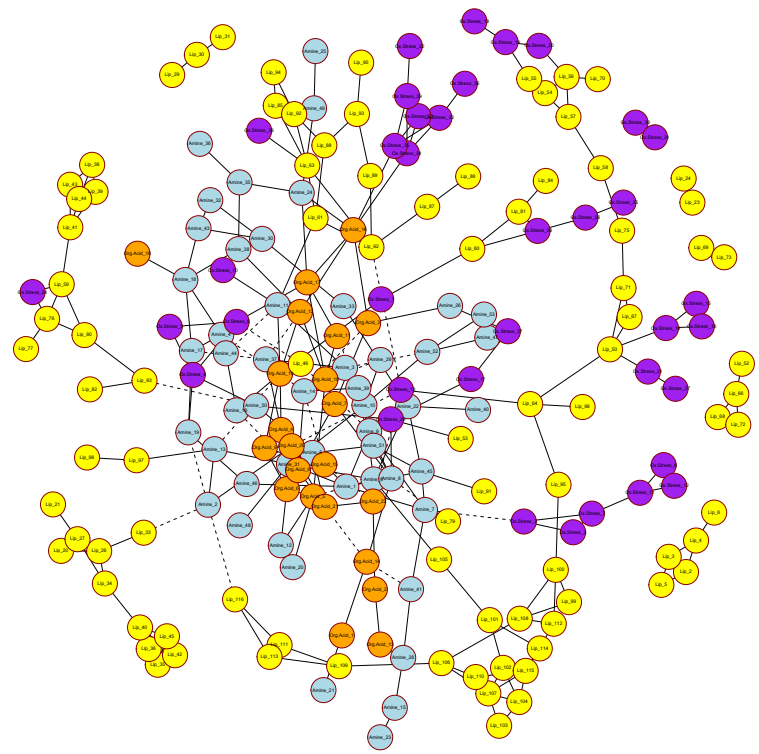
Vertices

- *Node* or *vertex* represents feature

Edges

Edge or *connection* represents some functional pairwise relation

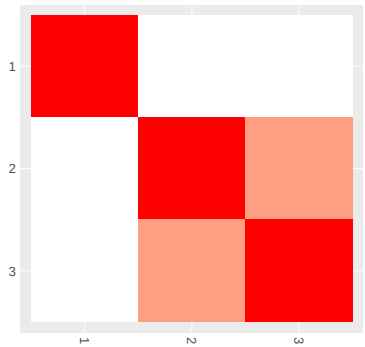


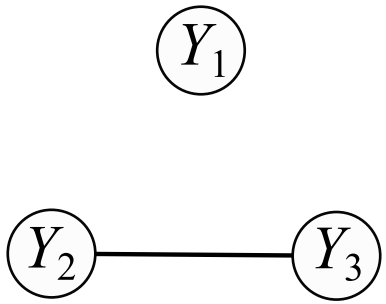


Example

Three variables: Y_1, Y_2, and Y_3

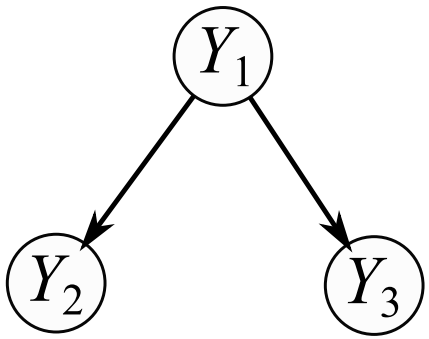
$$\text{cor}(Y_1, Y_2) = 0$$
$$\text{cor}(Y_1, Y_3) = 0$$
$$\text{cor}(Y_2, Y_3) \neq 0$$



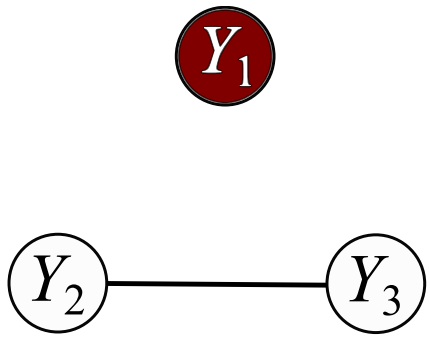


Marginal dependence

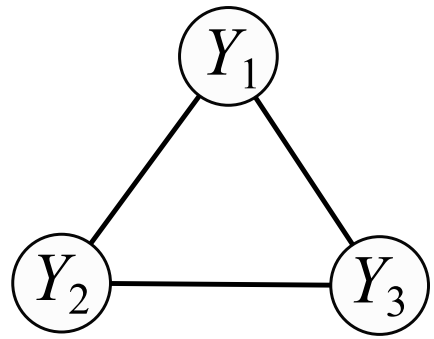
Undirected edge represents marginal dependence



True mechanism



Not observing Y_1:
Spurious association



Observing Y_1:
Saturated graph

Partial correlation

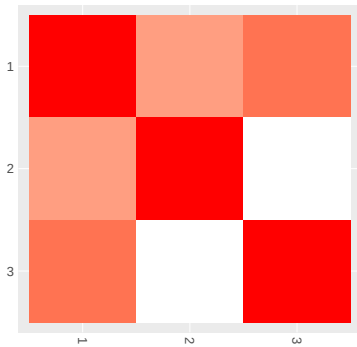
Measures degree of association between two random variables when controlling for remaining variables

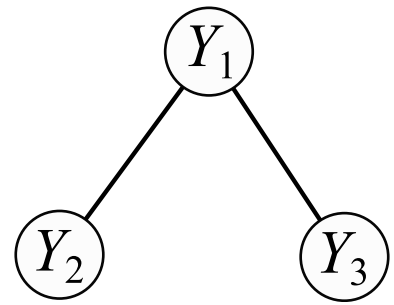
Conditioned correlation

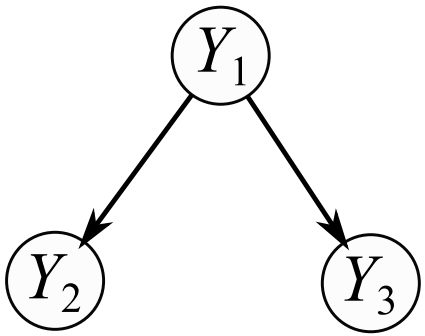
$$\text{cor}(Y_1, Y_2|Y_3)$$
$$\text{cor}(Y_1, Y_3|Y_2)$$
$$\text{cor}(Y_2, Y_3|Y_1)$$

If, e.g., $\text{cor}(Y_2, Y_3|Y_1) = 0$, we say Y_2 and Y_3 are independent *given* Y_1

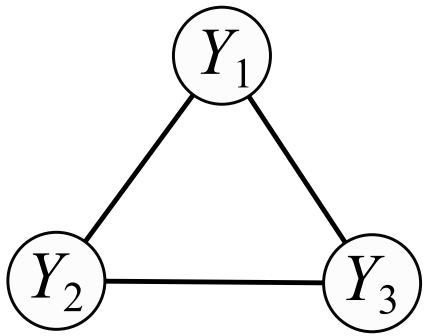
$$\text{cor}(Y_1, Y_2|Y_3) \neq 0$$
$$\text{cor}(Y_1, Y_3|Y_2) \neq 0$$
$$\text{cor}(Y_2, Y_3|Y_1) = 0$$



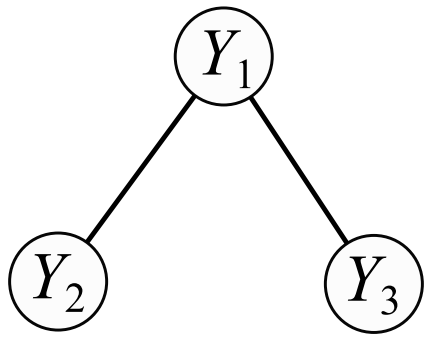




True mechanism



Correlation graph



Conditional independence graph

Graphical modeling

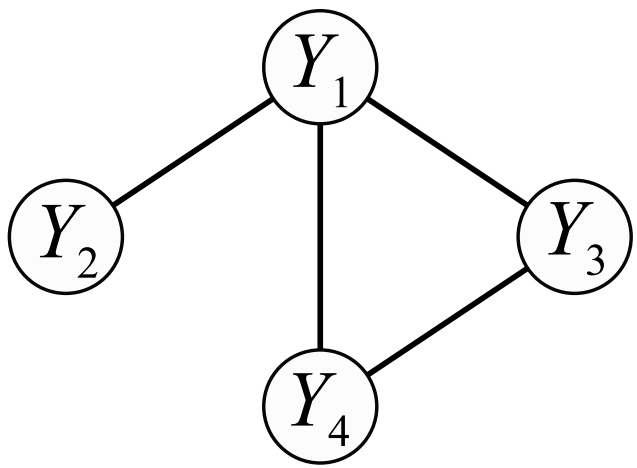
Class of models using graphs to express conditional (in)dependence relations between random variables

Gaussian setting

- Vertices: Correspond to random variables with normal distribution
- Edges: Correspond to the dependence structure
- Say $\mathbf{y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, and define $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Omega}$. Then, for $a, b \in$ vertex set V, $a \neq b$

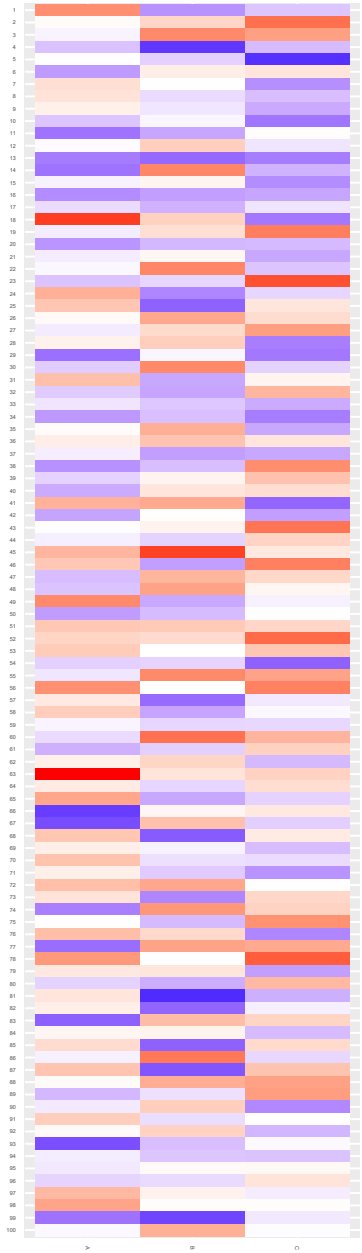
$$-\frac{\omega_{ab}}{\sqrt{\omega_{aa}\omega_{bb}}} = 0 \iff \omega_{ab} = 0 \iff a \perp\!\!\!\perp b | V \setminus \{a, b\} \iff a \neq b$$

$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{bmatrix}$$

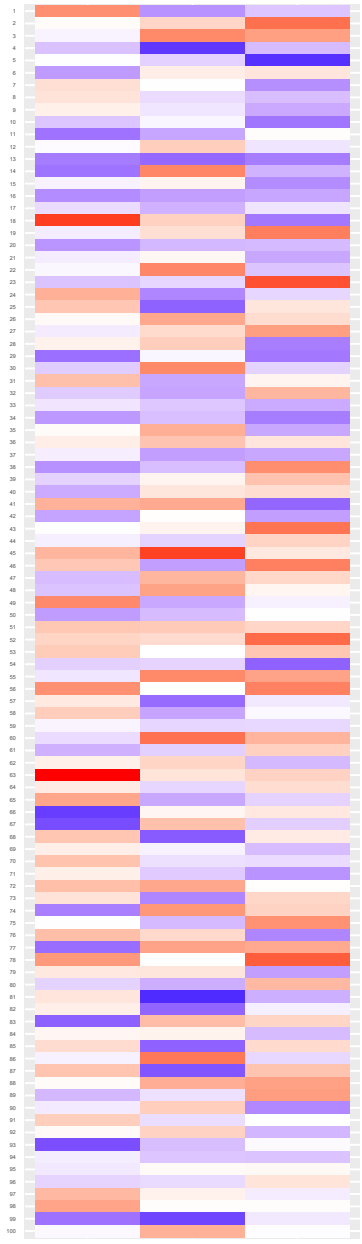



Challenge
High-dimensional data

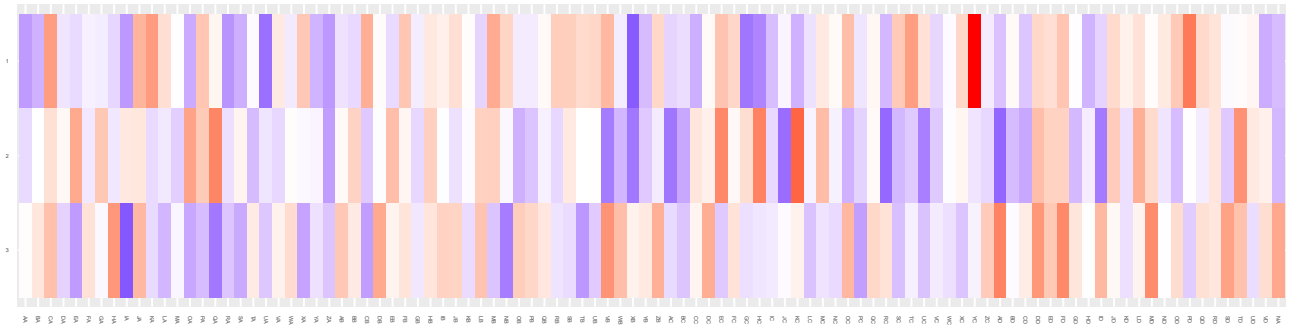
$$n > p$$



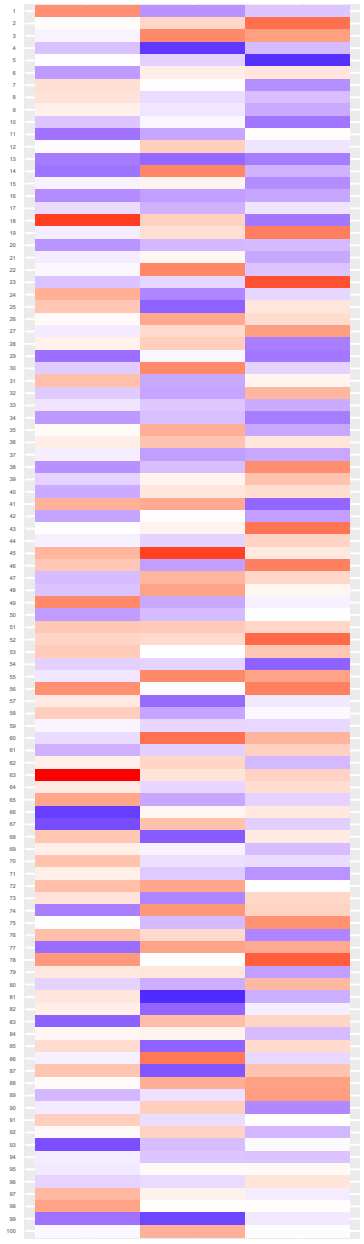
$n > p$



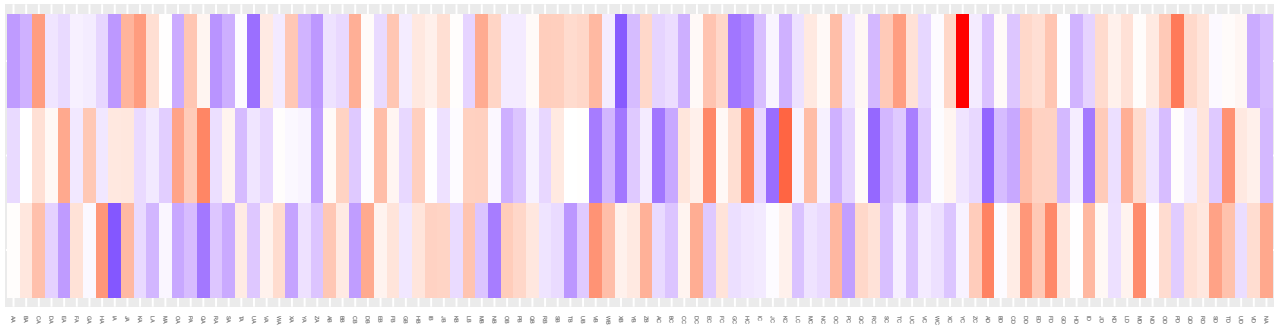
$p > n$



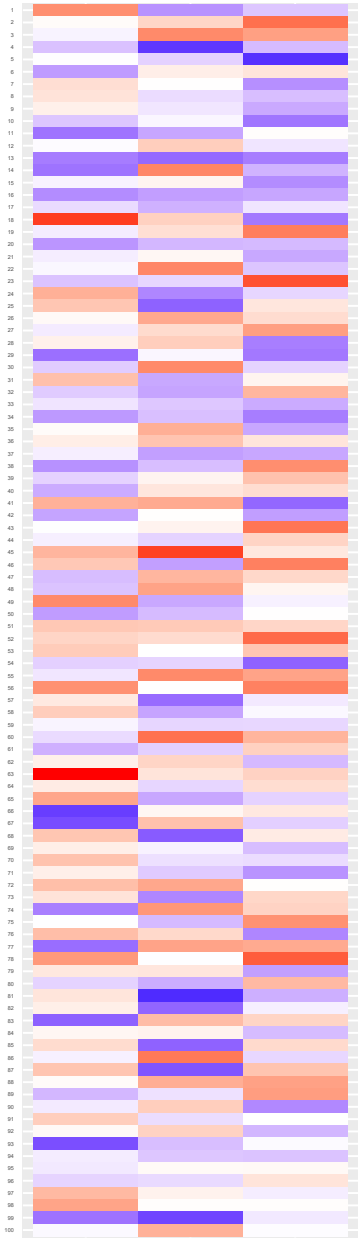
$n > p$



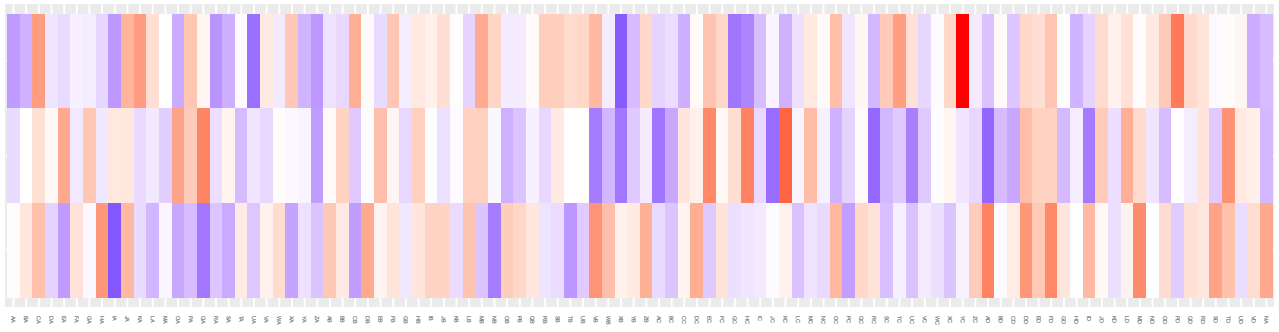
$p > n$

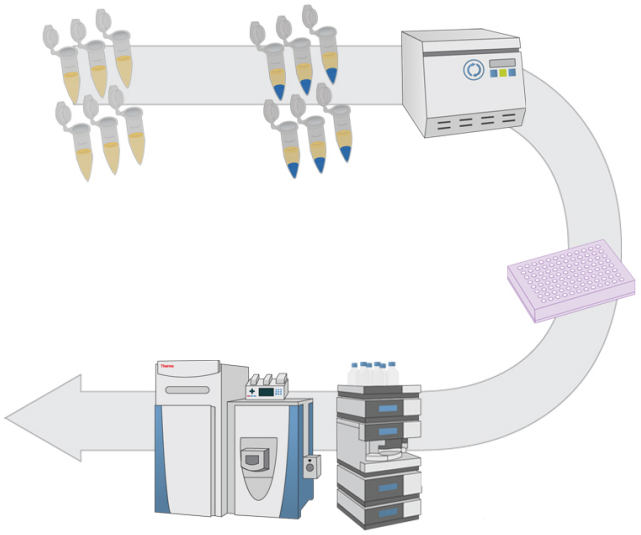


$n > p$

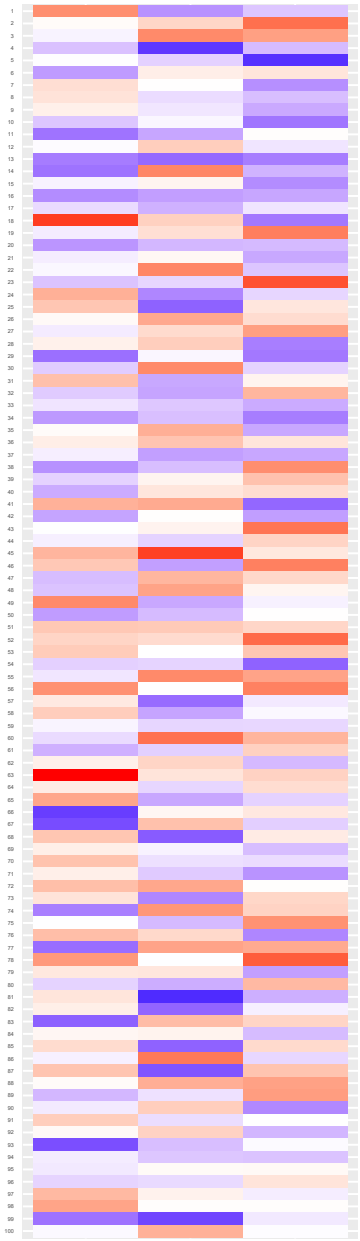


$p > n$

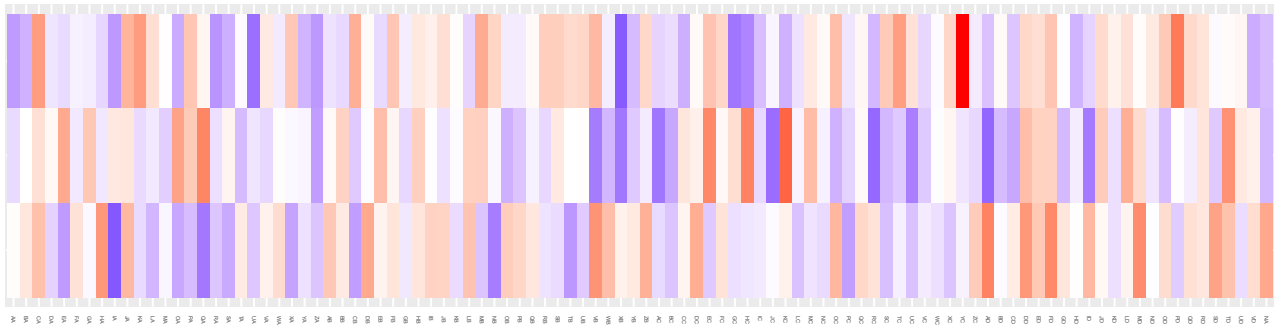


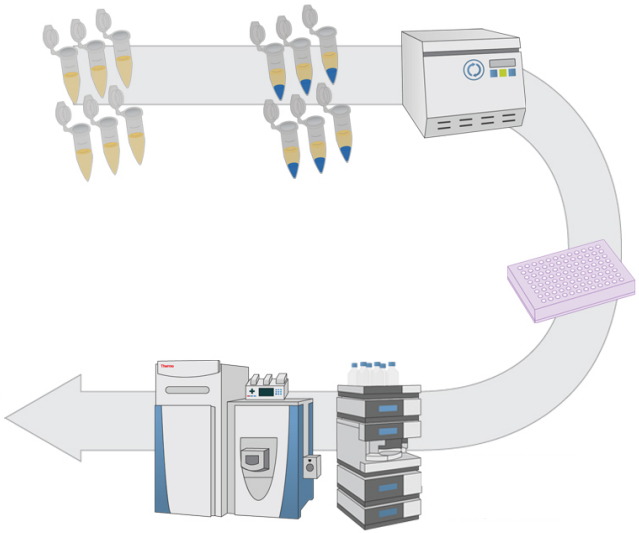


$n > p$



$p > n$



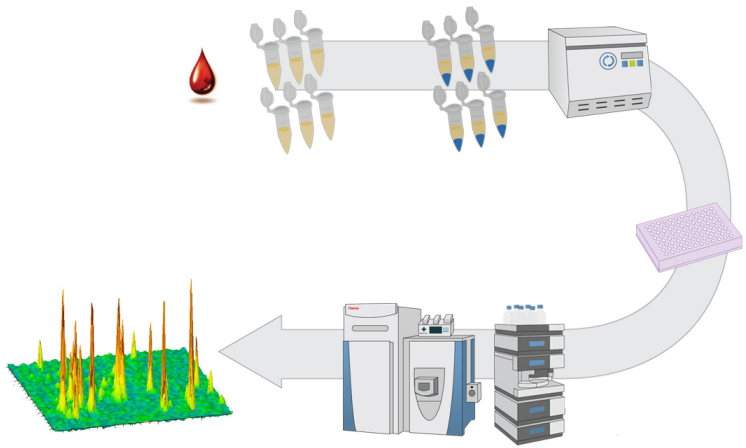


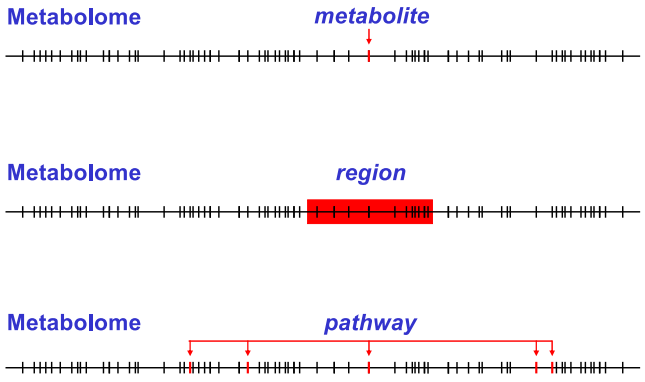
Inverse correlation when $n > p$

```
> ## Obtain some data
> p = 3
> n = 20
> X = matrix(rnorm(n*p), nrow = n, ncol = p)
> R <- cor(X)
>
> ## Obtain inverse and check
> Ri <- solve(R)
> R %*% Ri
              [,1]          [,2]          [,3]
[1,] 1.000000e+00  2.515349e-17  0.000000e+00
[2,] 3.469447e-18  1.000000e+00 -2.775558e-17
[3,] 6.938894e-18 -2.775558e-17  1.000000e+00
```

Inverse correlation when $p > n$

```
> ## Obtain some (high-dimensional) data
> p = 3
> n = 2
> X = matrix(rnorm(n*p), nrow = n, ncol = p)
> R <- cor(X)
>
> ## Try to obtain inverse
> Ri <- solve(R)
Error in solve.default(R) :
  Lapack routine dgesv: system is exactly singular
```

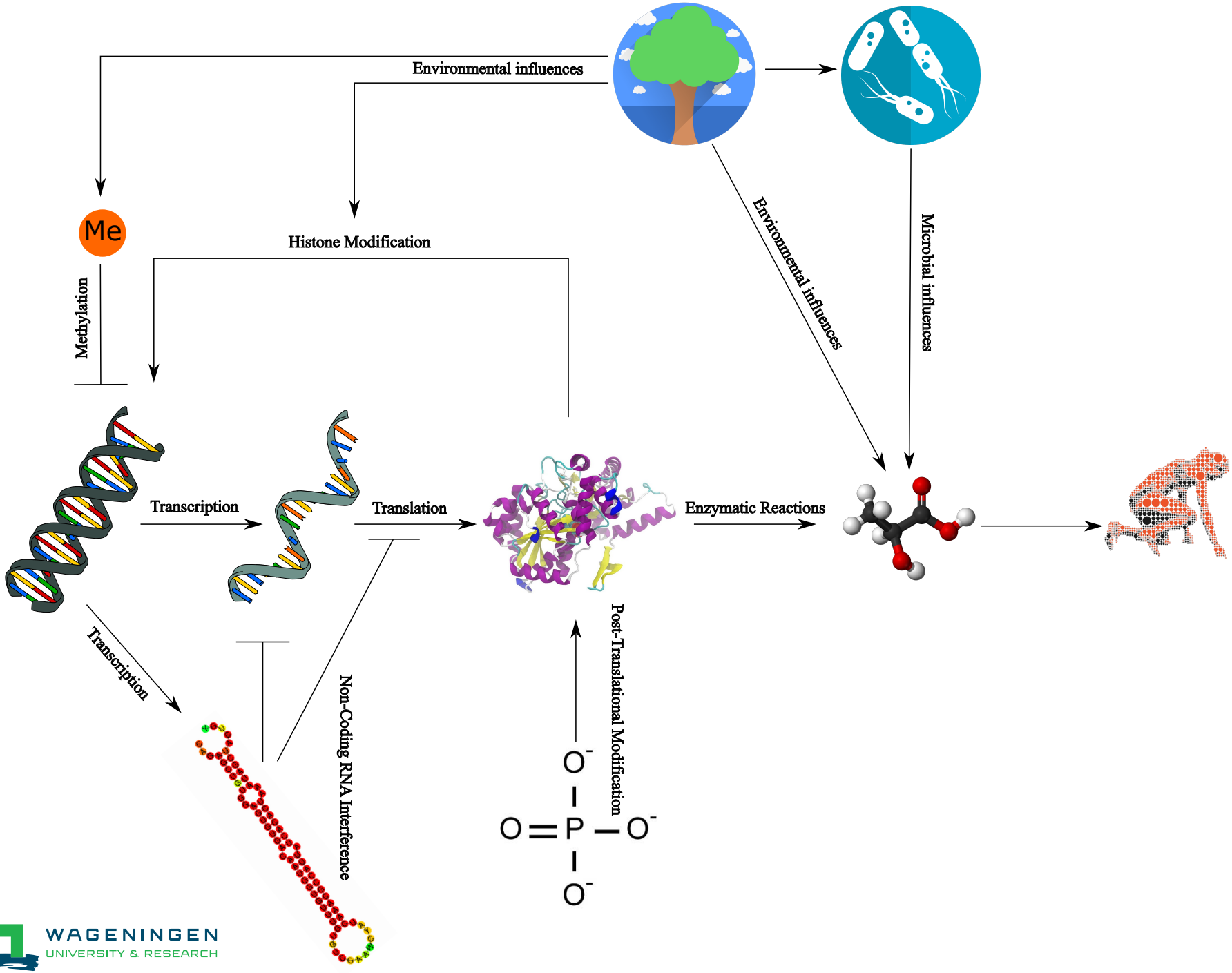




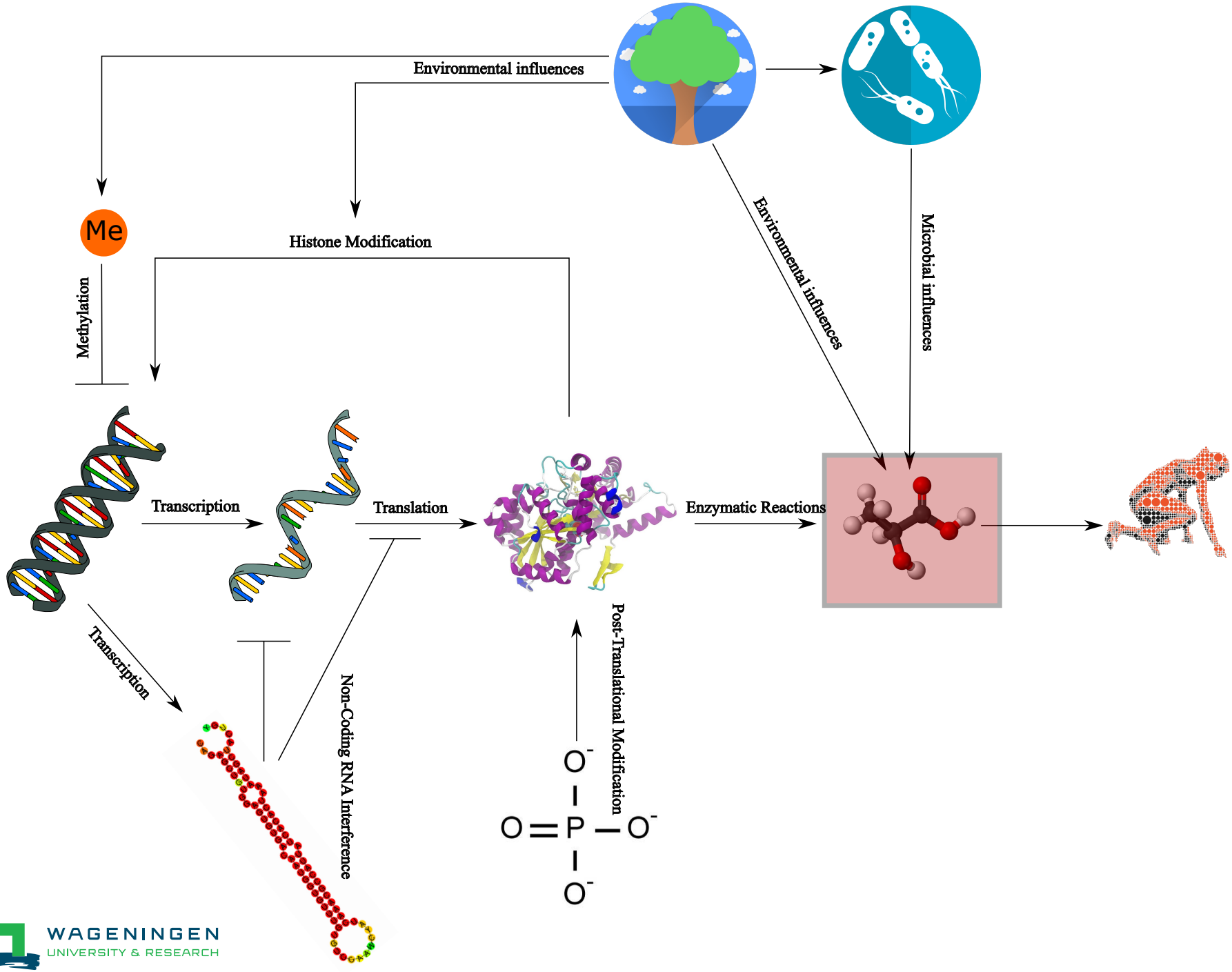
Problem

- Let $\mathbf{S} \equiv \hat{\mathbf{\Sigma}}$ denote the sample covariance matrix on n realizations $\mathbf{y}_i$
- When $p \approx n$ or $p > n$, $\mathbf{S}$ is ill-behaved or singular
- The empirical precision $\mathbf{S}^{-1} \equiv \hat{\mathbf{\Omega}}$ is then undefined

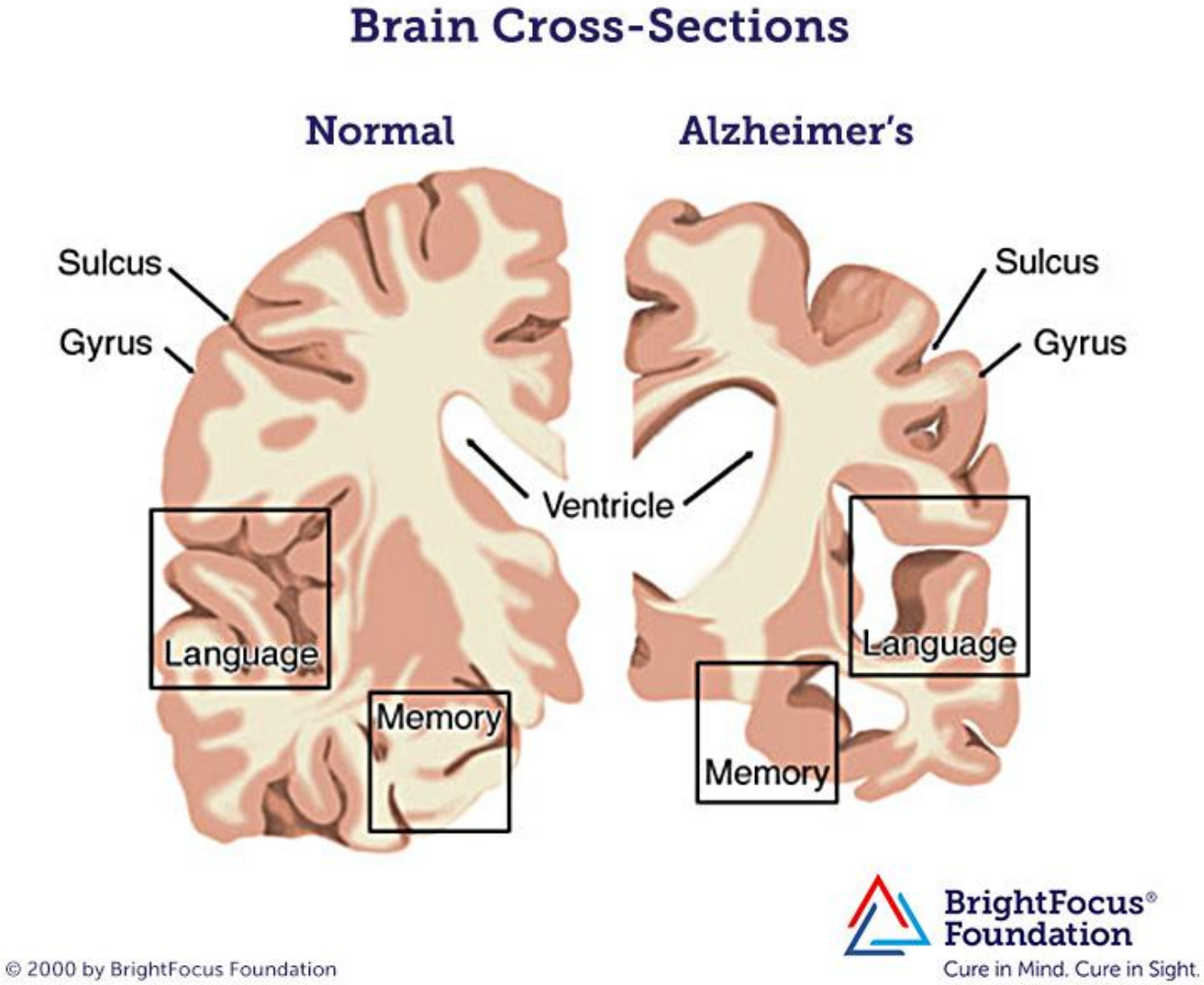

DATA
Our running example



WAGENINGEN
UNIVERSITY & RESEARCH



WAGENINGEN
UNIVERSITY & RESEARCH


Brain Cross-Sections
Normal
Alzheimer's
Sulcus
Gyrus
Ventricle
Language
Memory
Sulcus
Gyrus
Language
Memory
© 2000 by BrightFocus Foundation
BrightFocus® Foundation
Cure in Mind. Cure in Sight.

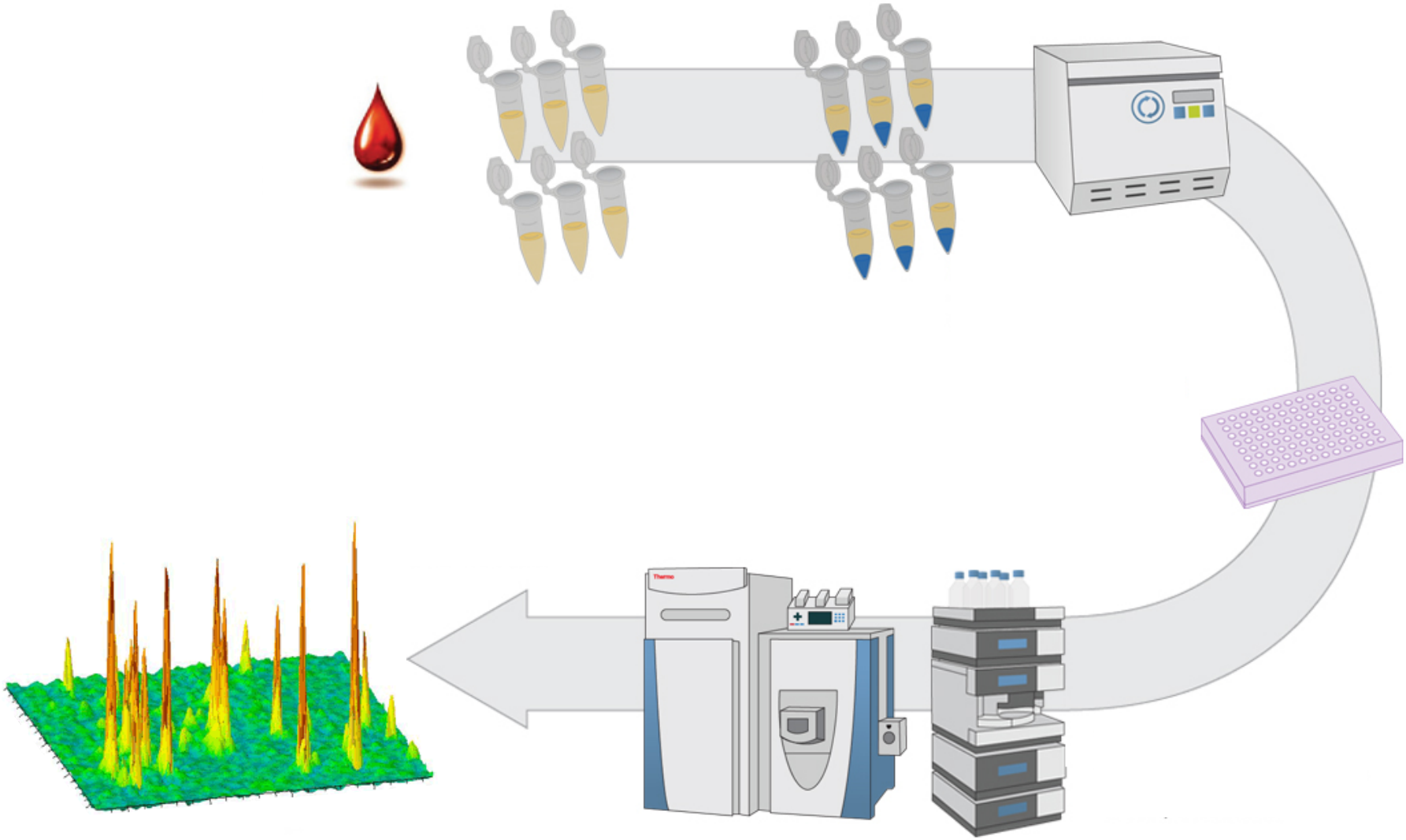
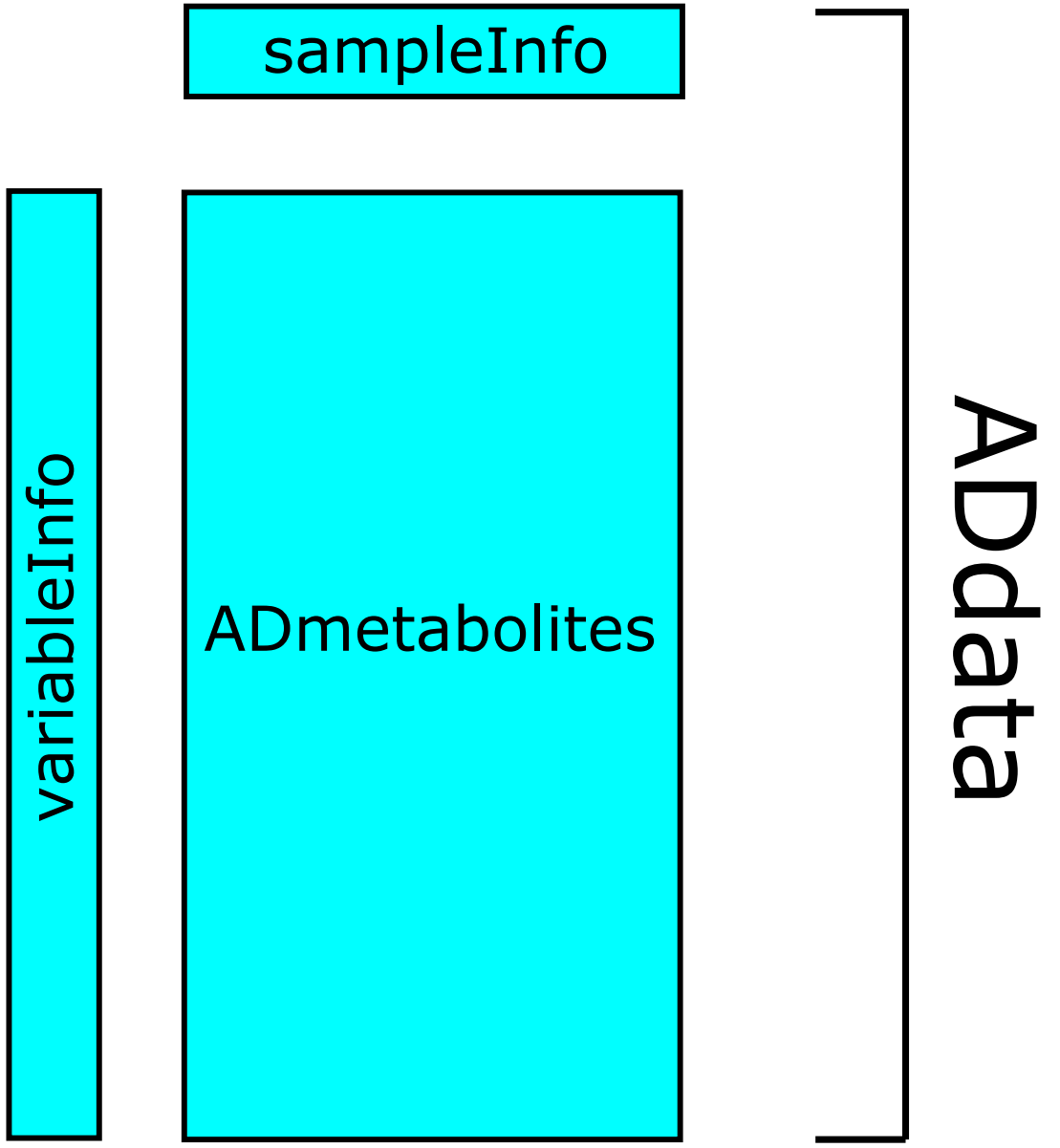


Illustration adapted from: http://planetorbitrap.com/untargeted-metabolomics#.Vzw6yfmLRaQ & http://metabolomicsplatform.com/metabolomics-overview/


sampleInfo
variableInfo
ADmetabolites
ADdata

Exercise 1: Get acquainted with the data

Get acquainted with the data

```
## Set working directory to convenience
setwd("")

## Needed package
library("rags2ridges")


## Data packaged as 3 data objects in ADdata object
data("ADdata", package = "rags2ridges")

## To probe objects one could use:
objects()
head(ADmetabolites)
head(sampleInfo)
colnames(variableInfo); table(variableInfo)
```

Variables (features)

Observations	1	2	3	4	5	.....	p
1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	.....	y_{1p}
2	y_{21}	y_{22}	y_{23}	y_{24}	y_{25}	.....	y_{2p}
3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}	.....	y_{3p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
n	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}	.....	y_{np}

- Amines
- Lipids
- Organic Acids
- Oxidative Stress

p = 230: 53 Amines, 116 Lipids, 22 Organic acids, 39 Oxidative stress compounds
n = 127: 40 AD class 1, 87 AD class 2


Extraction
How to get the network

Maximize

$$\underbrace{\ln|\mathbf{\Omega}| - \text{tr}(\mathbf{S\Omega})}_{\text{log-likelihood}} - \underbrace{\frac{\lambda}{2}\|\mathbf{\Omega} - \mathbf{T}\|_2^2}_{\ell_2\text{-penalty}}$$

- $\mathbf{T}$ denotes a p.d. symmetric target matrix
- $\lambda \in (0, \infty)$ denotes a penalty parameter

Analytic penalized ML estimator

$$\hat{\mathbf{\Omega}}(\lambda) = \left\{ \left[\lambda \mathbf{I}_p + \frac{1}{4}(\mathbf{S} - \lambda\mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda\mathbf{T}) \right\}^{-1}$$

Behavior

i. $\hat{\boldsymbol{\Omega}}(\lambda) \succ 0$, for all $\lambda \in (0, \infty)$;

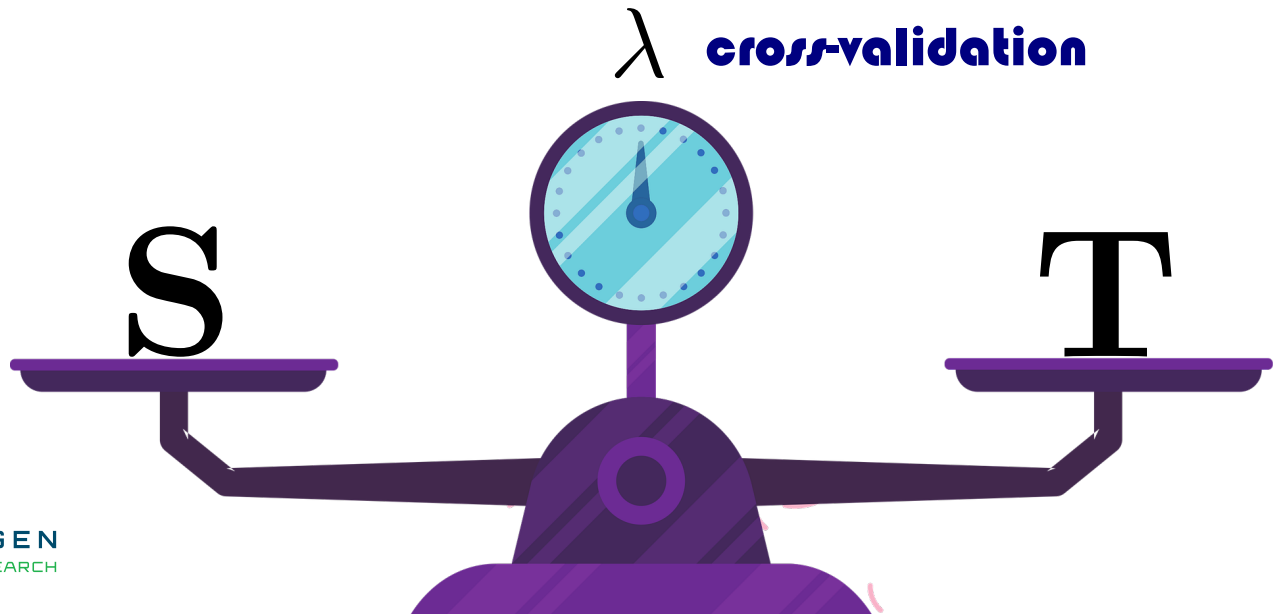
ii. $\lim_{\lambda \to 0^+} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{S}^{-1}$ if $p < n$;

iii. $\lim_{\lambda \to \infty} \hat{\boldsymbol{\Omega}}(\lambda) = \mathbf{T}$.

Consistency

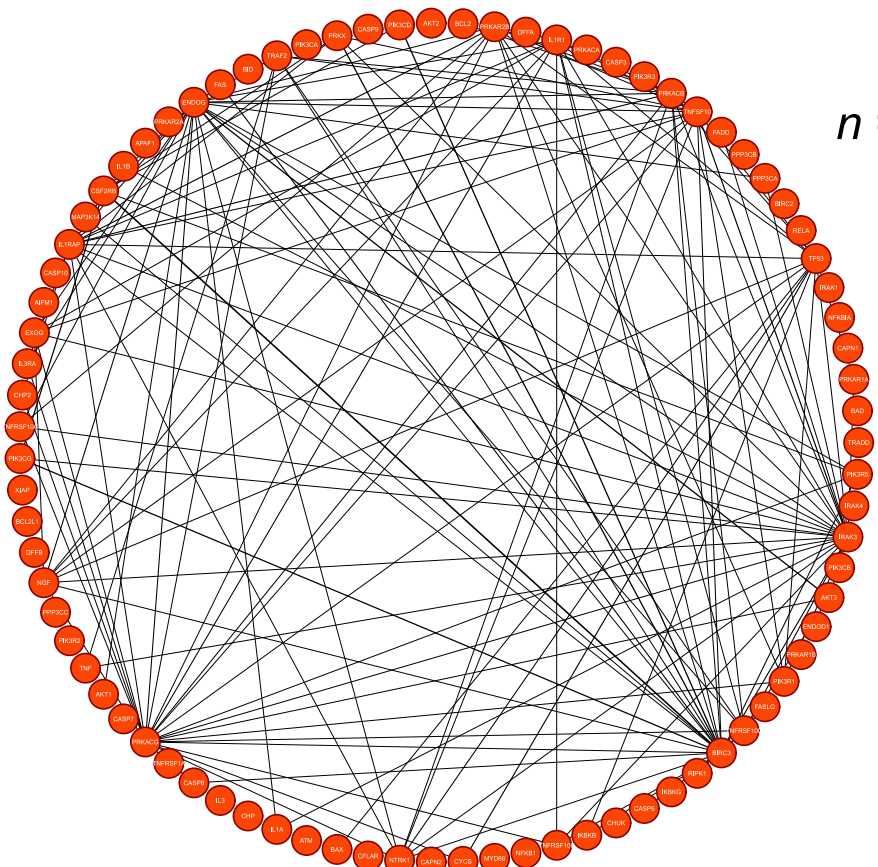
i. $\lim_{n \to \infty} \mathbb{E}\left[\hat{\boldsymbol{\Omega}}_n(\lambda_n)\right] \longrightarrow \lim_{n \to \infty} \mathbb{E}\left(\mathbf{S}_n^{-1}\right) = \boldsymbol{\Omega}$;

ii. $\lim_{n \to \infty} \mathbb{E}\left(\|\hat{\boldsymbol{\Omega}}_n(\lambda_n) - \boldsymbol{\Omega}\|_F^2\right) = 0$.

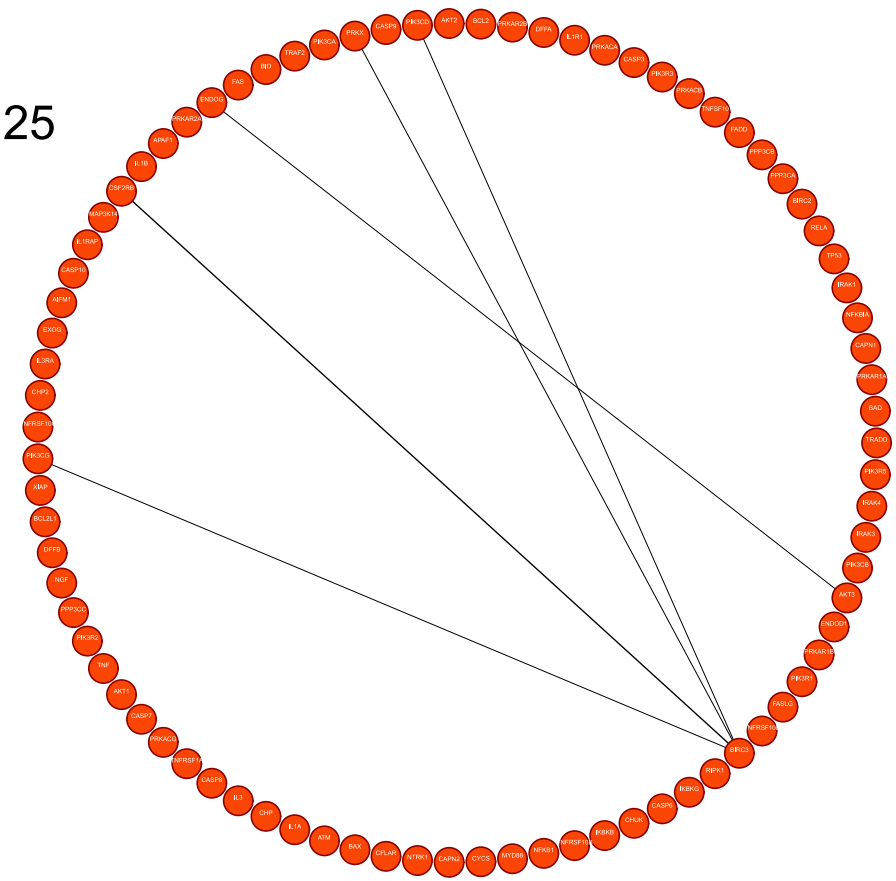


Why not maximize the ℓ_1-Penalized log-likelihood?

$$\underbrace{\ln|\mathbf{\Omega}| - \text{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Omega})}_{\text{log-likelihood}} - \underbrace{\lambda\|\mathbf{\Omega}\|_1}_{\ell_1\text{-penalty}}$$



n = 25

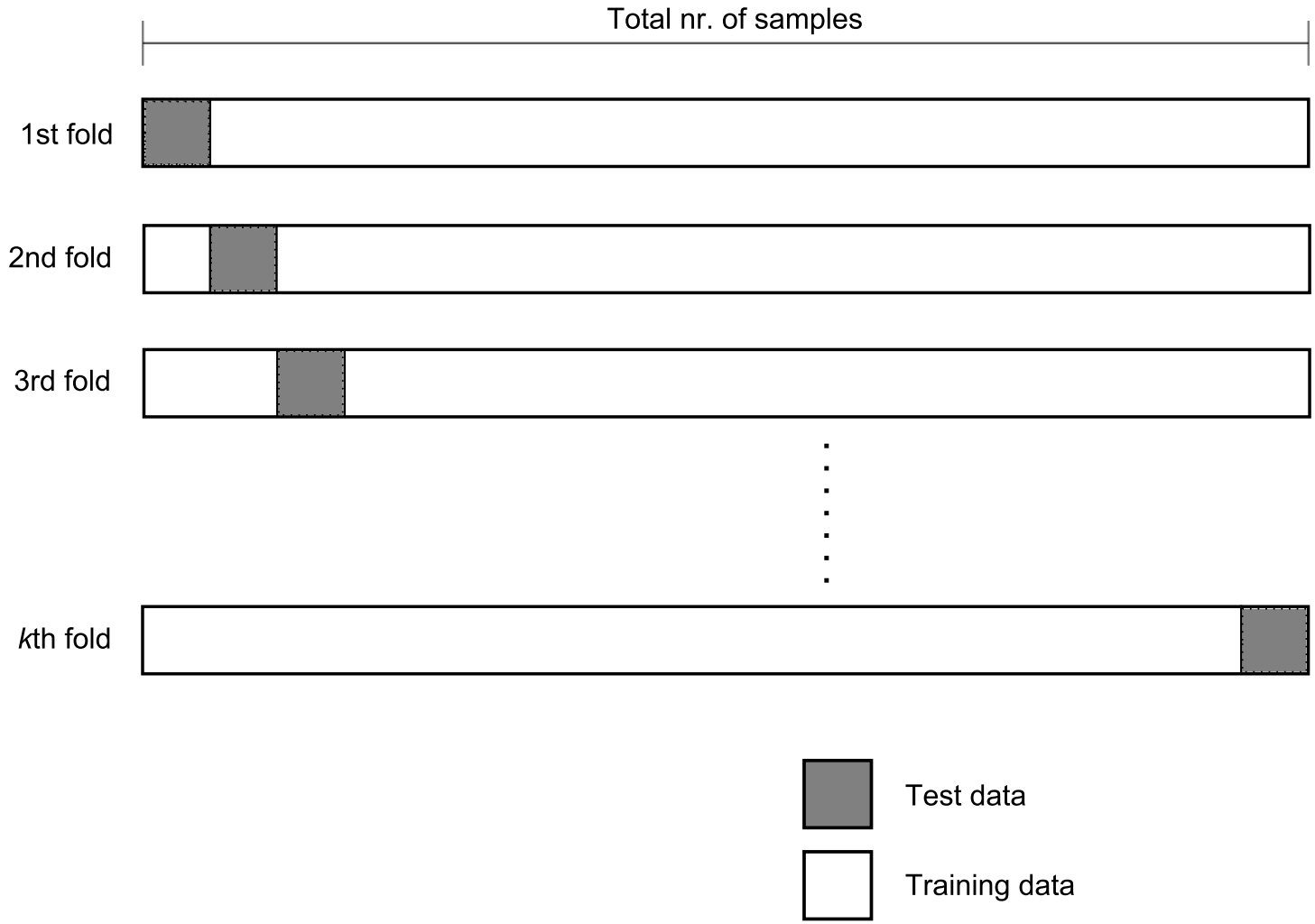


Graphical ridge

graphical lasso

K-fold cross-validation (CV)

Single iteration of *K*-fold CV



K-fold CV score

$$\varphi^K(\lambda) = \sum_{k=1}^{K} n_k \left\{ -\ln |\hat{\mathbf{\Omega}}(\lambda)_{-k}| + \text{tr}[\hat{\mathbf{\Omega}}(\lambda)_{-k}\mathbf{S}_k] \right\},$$

- n_k is the size of subset k, for $k = 1, \ldots, K$ disjoint subsets;
- $\mathbf{S}_k$ denotes the sample covariance matrix on kth test set;
- $\hat{\mathbf{\Omega}}(\lambda)_{-k}$ denotes the estimated regularized precision matrix on kth training set

Choose

$\lambda^* = \arg\min_{\lambda \in \mathbb{R}+} \varphi^K(\lambda)$

Efficiency

- Implementation uses C++ at its core
- Makes use of a root-finding (Brent) algorithm
- Utilizes rotational equivariance property when possible

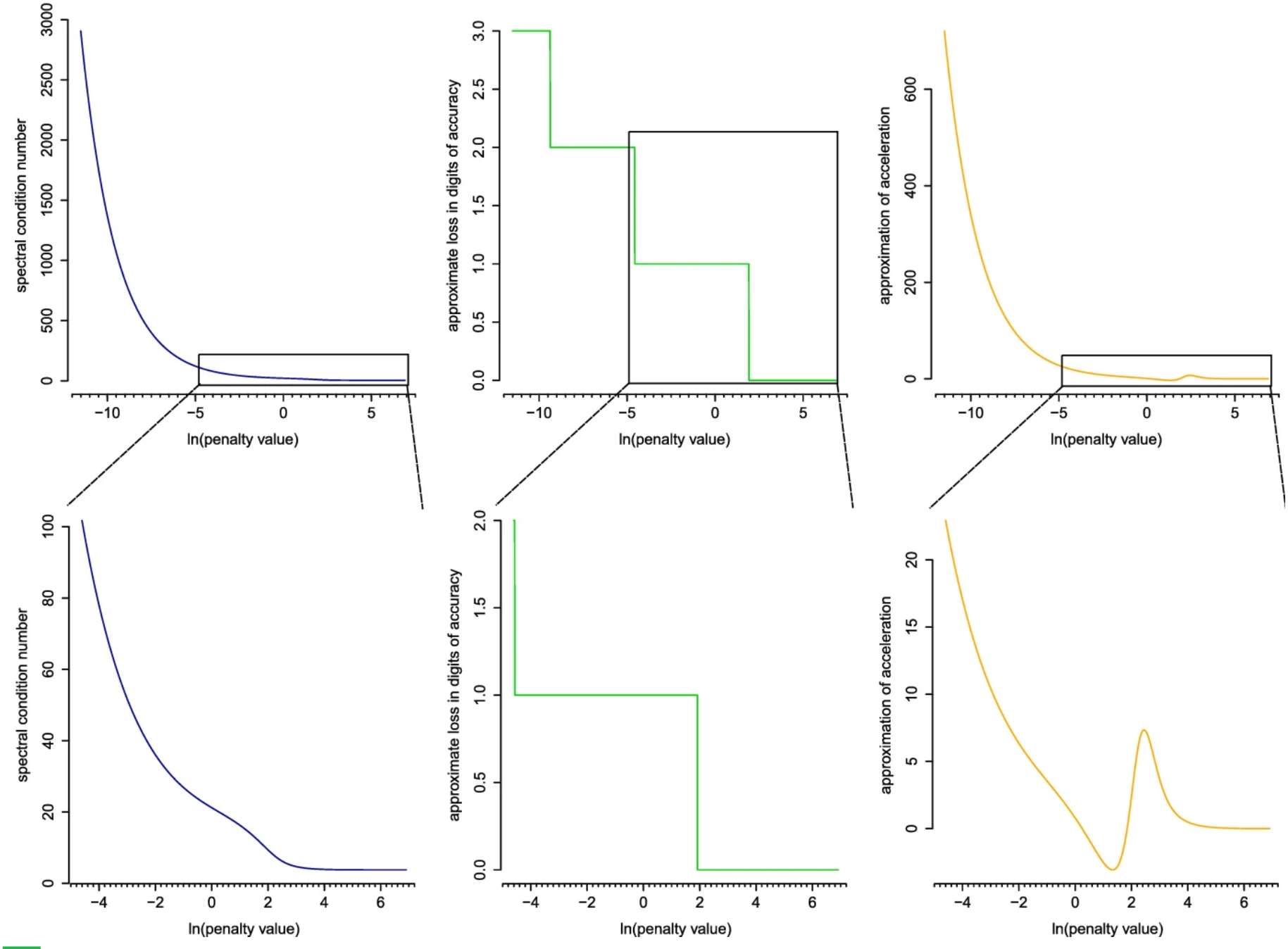
Exercise 2: Find an optimal precision matrix for the Class 2 AD data

CV function

```
optPenalty.kCVauto(
  Y,          ## data
  fold,       ## number of folds
  lambdaMin,  ## minimum penalty value
  lambdaMax,  ## maximum penalty value
  target      ## target matrix T: use default.target()
  )
```

Returns list object

- $optLambda: Optimal penalty parameter
- $optPrec: Precision estimate under optimal penalty parameter



WAGENINGEN
UNIVERSITY & RESEARCH

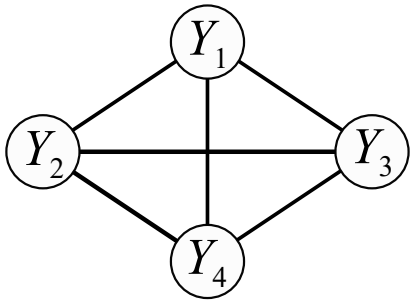
Exercise 3: Assess the conditioning of the optimal precision matrix

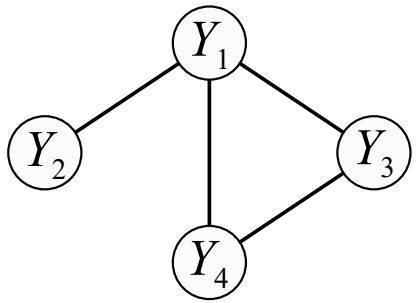
Function for condition number plot

```
CNplot(S,          ## covariance matrix data
       lambdaMin,  ## minimum penalty value
       lambdaMax,  ## maximum penalty value
       step,       ## coarseness grid
       target,     ## target matrix T
       Iaids,      ## logical for interpretational aids
       vertical,   ## logical for inclusion vertical bar
       value,      ## value if vertical = TRUE
       verbose     ## logical controlling on-screen printing
       )
```

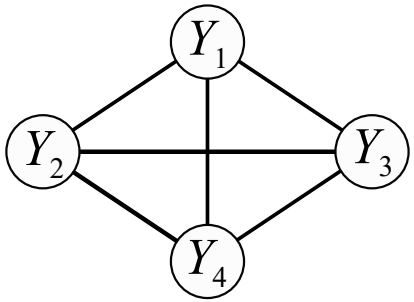
Returns list object

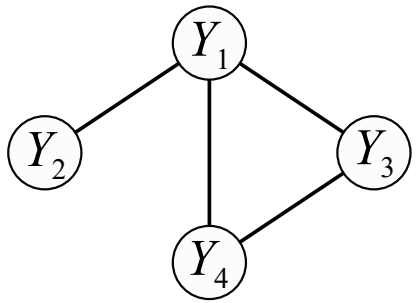
- `$optLambda`: Optimal penalty parameter
- `$optPrec`: Precision estimate under optimal penalty parameter





$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & \omega_{23} & \omega_{24} \\ \omega_{31} & \omega_{32} & \omega_{33} & \omega_{34} \\ \omega_{41} & \omega_{42} & \omega_{43} & \omega_{44} \end{bmatrix} \longrightarrow \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{bmatrix}$$





$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & \omega_{23} & \omega_{24} \\ \omega_{31} & \omega_{32} & \omega_{33} & \omega_{34} \\ \omega_{41} & \omega_{42} & \omega_{43} & \omega_{44} \end{bmatrix} \longrightarrow \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{bmatrix}$$

Scaling

$\hat{\mathbf{P}}(\lambda)$: Regularized precision estimate scaled to partial correlation form

Assume

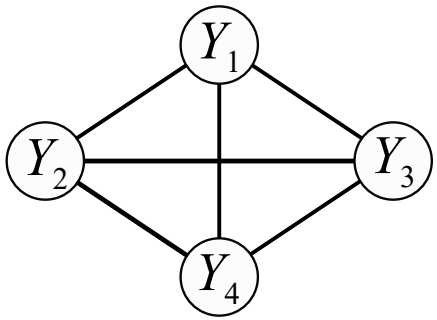
Nonredundant coefficients (indexed by $j < j'$) follow a mixture distribution:

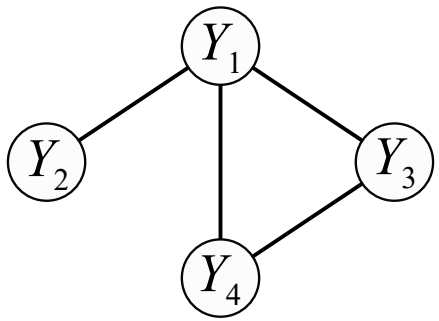
$$f\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\} = \eta_0 f_0\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}; \kappa\right\} + (1-\eta_0) f_{\mathcal{E}}\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\}$$

- $\eta_0 \in [0,1]$ is the mixture weight
- $f_0\{\cdot\}$ denotes the distribution of a null-edge
- $f_{\mathcal{E}}\{\cdot\}$ denotes the distribution of a present edge
- κ denotes degrees of freedom

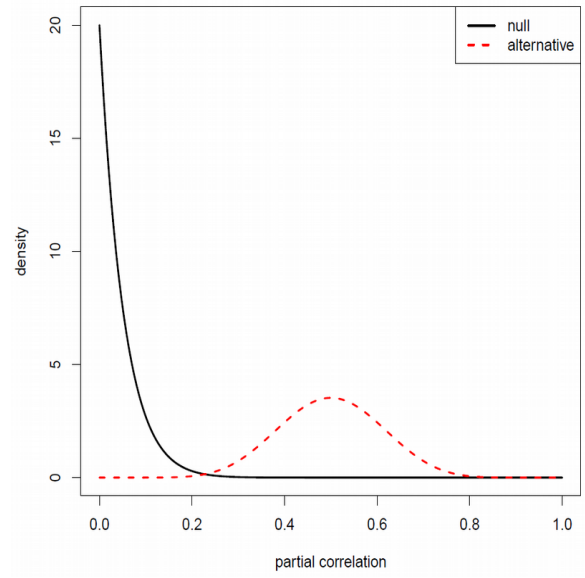
Determine

$$P\left(Y_j \neq Y_{j'} | [\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right) = \frac{\hat{\eta}_0 f_0\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}; \hat{\kappa}\right\}}{\hat{\eta}_0 f_0\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}; \hat{\kappa}\right\} + (1-\hat{\eta}_0)\hat{f}_{\mathcal{E}}\left\{[\hat{\mathbf{P}}(\lambda^*)]_{jj'}\right\}}$$





$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & \omega_{23} & \omega_{24} \\ \omega_{31} & \omega_{32} & \omega_{33} & \omega_{34} \\ \omega_{41} & \omega_{42} & \omega_{43} & \omega_{44} \end{bmatrix} \longrightarrow \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} \\ \omega_{21} & \omega_{22} & 0 & 0 \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ \omega_{41} & 0 & \omega_{43} & \omega_{44} \end{bmatrix}$$



Exercise 4: Extract network from the optimal precision matrix

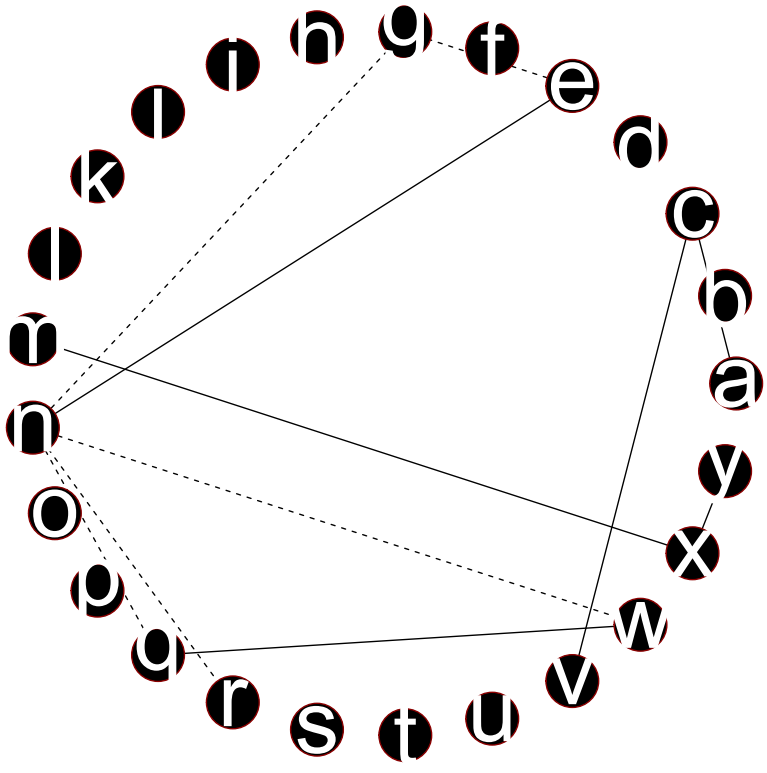
Function for support determination

```
sparsify(P,              ## estimated precision matrix
         threshold,      ## type of thresholding: we use "localFDR"
         FDRcut,         ## cut-off for 1 - lFDR
         verbose,        ## logical controlling on-screen printing
         )
```

Returns list object

- `$sparsePrecision`: Sparsified precision matrix
- `$sparseParCor`: Sparsified partial correlation matrix


Visualization
How to visualize the network


m
x
y
w
r
q
n
g
e
a
c
v

Exercise 5: Visualize the extracted network

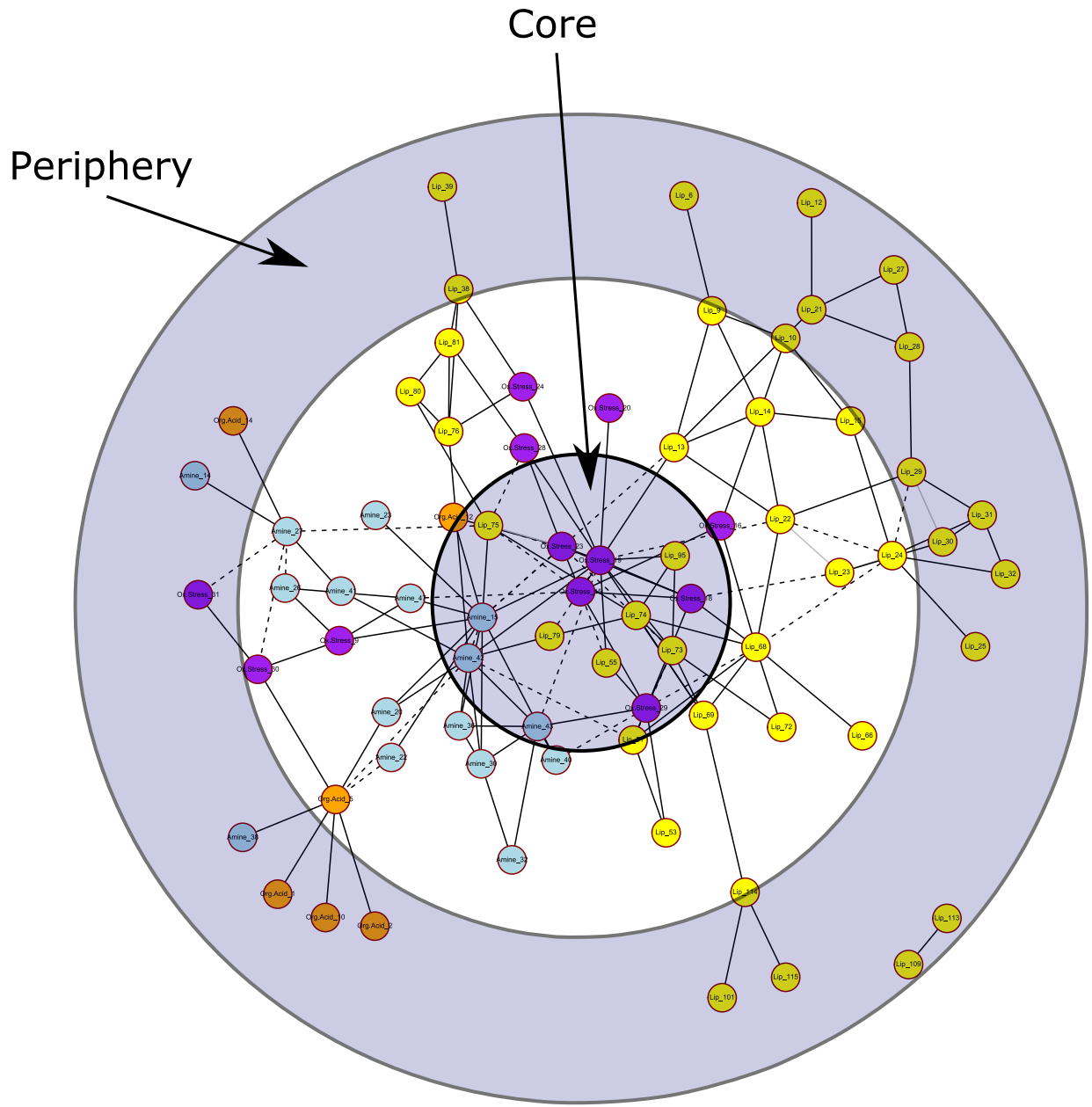
Function for visualization

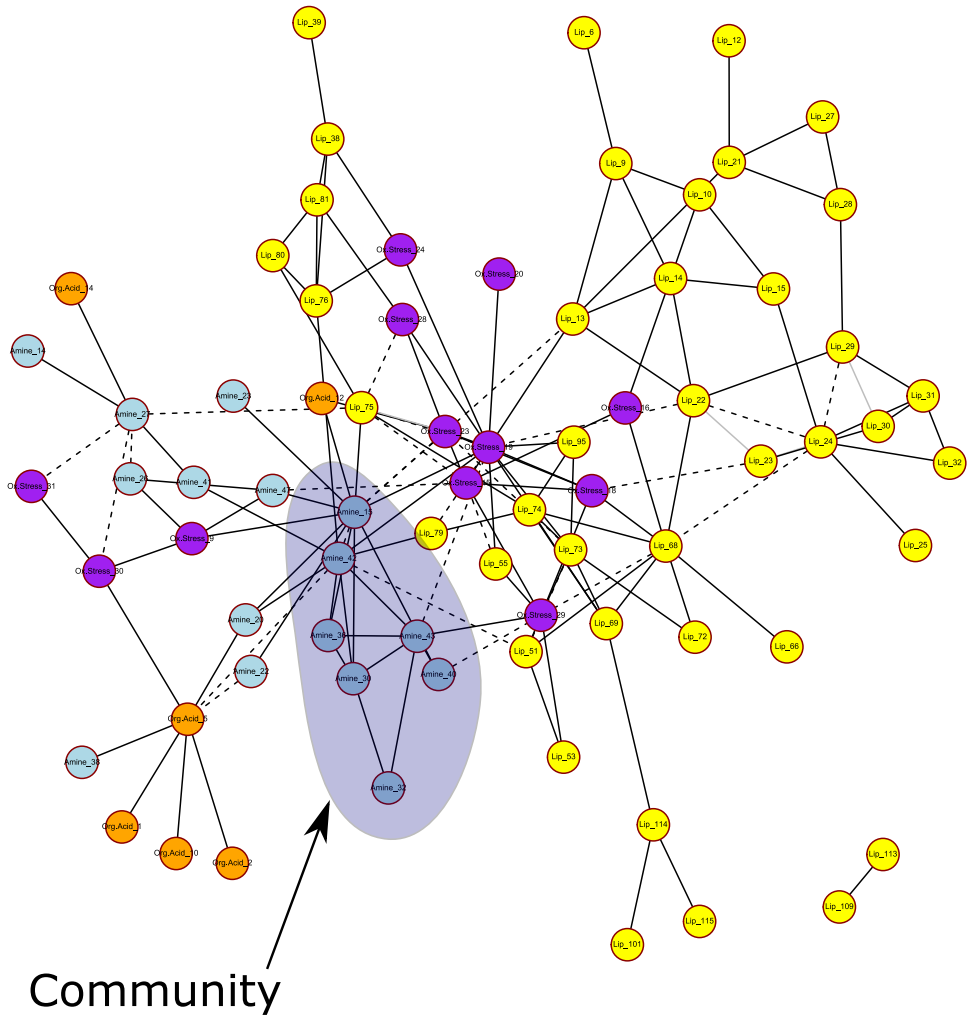
```
Ugraph(M,       ## matrix to be visualized
       type,    ## "plain", "fancy", "weighted"
       lay,     ## specifies layout
       coords, ## gives control over node coordinates
       Vcolor, ## vertex color
       Vsize,   ## vertex size
       Vcex,    ## size of labels within vertices
       ...
       )
```

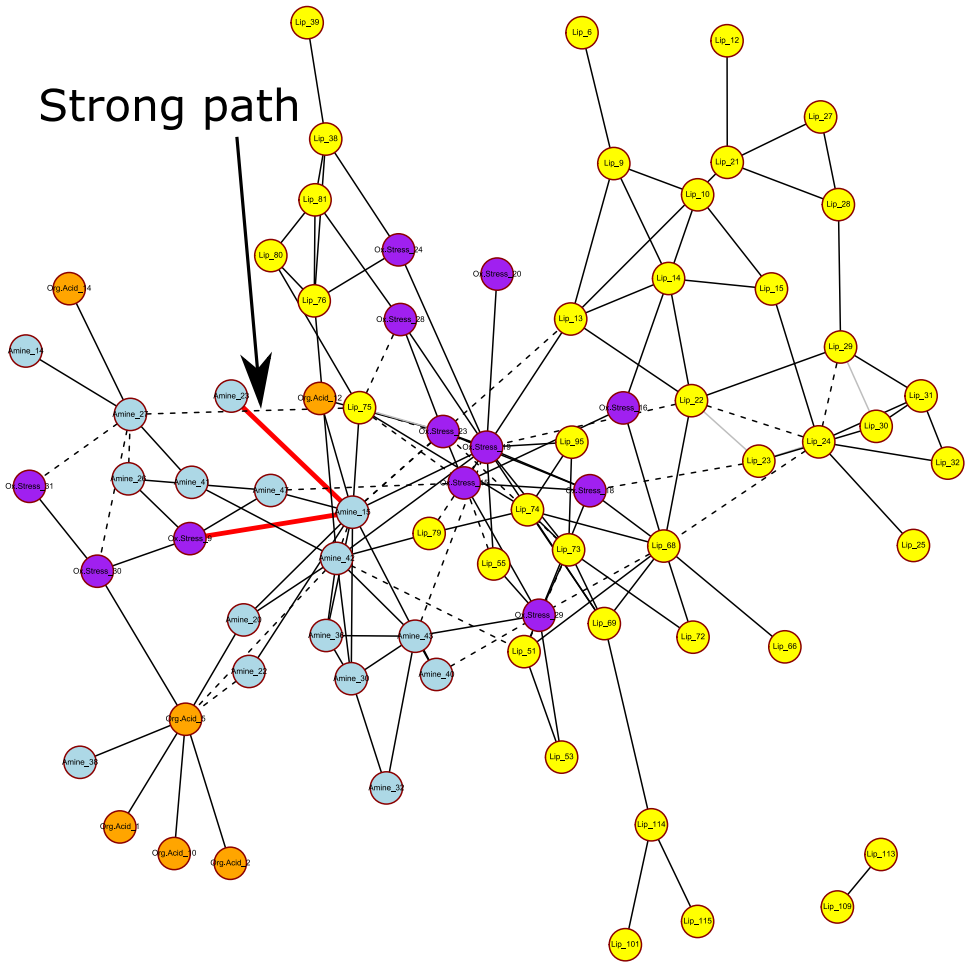
Returns matrix object

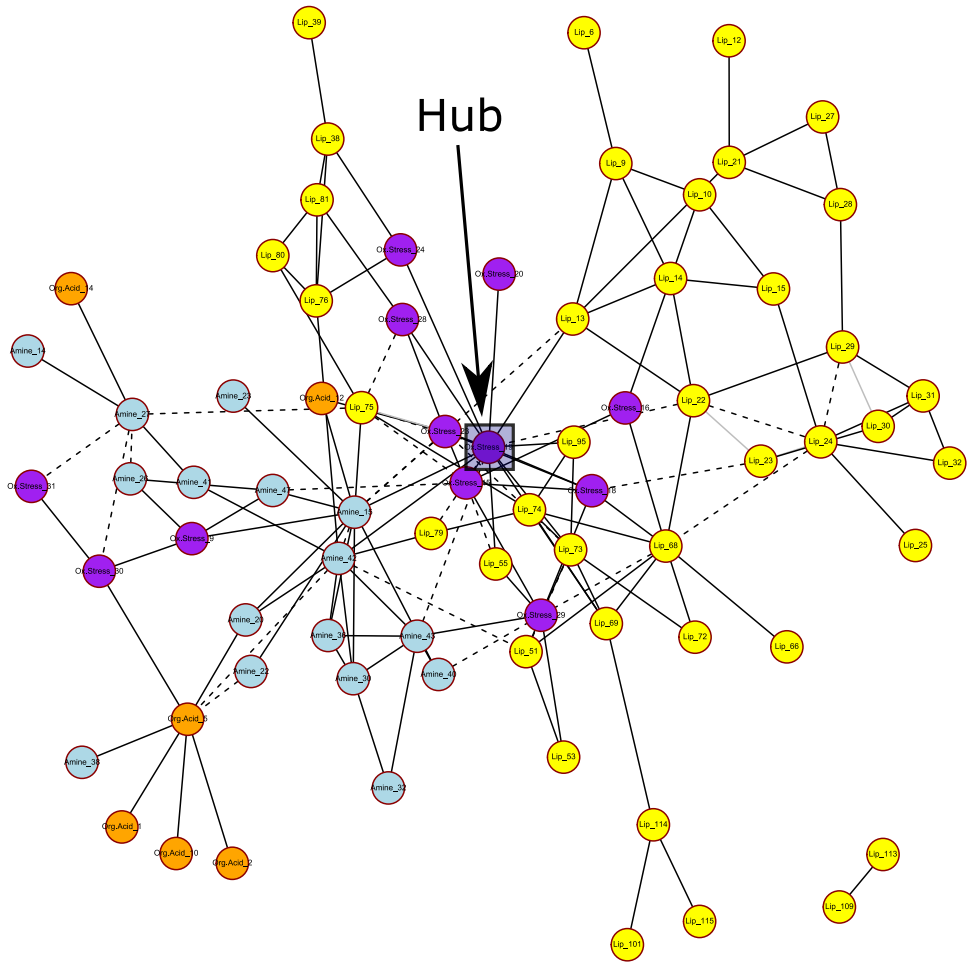
Containing the coordinates of the vertices in the given graph/network

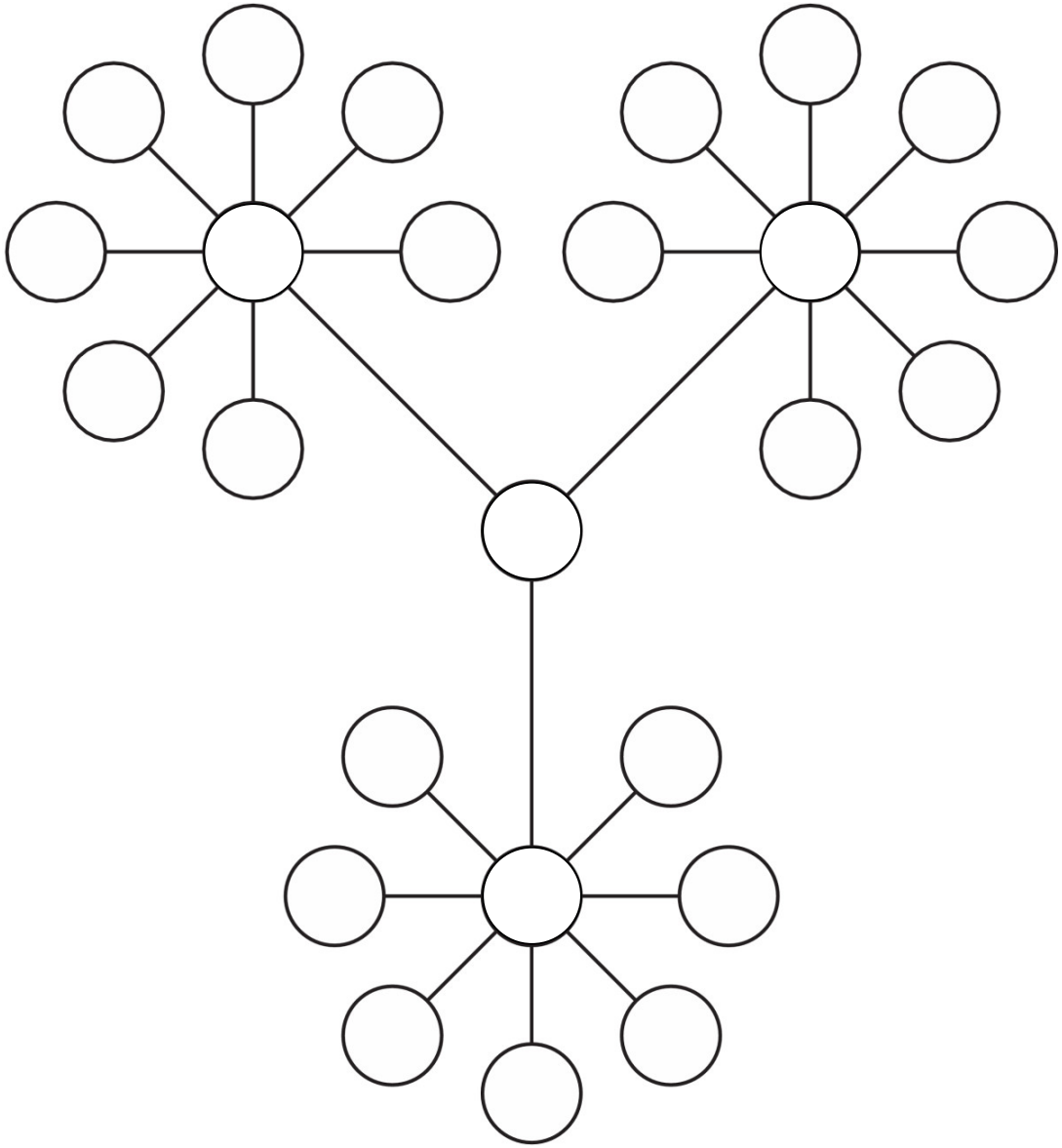

Analysis
How to analyze the network

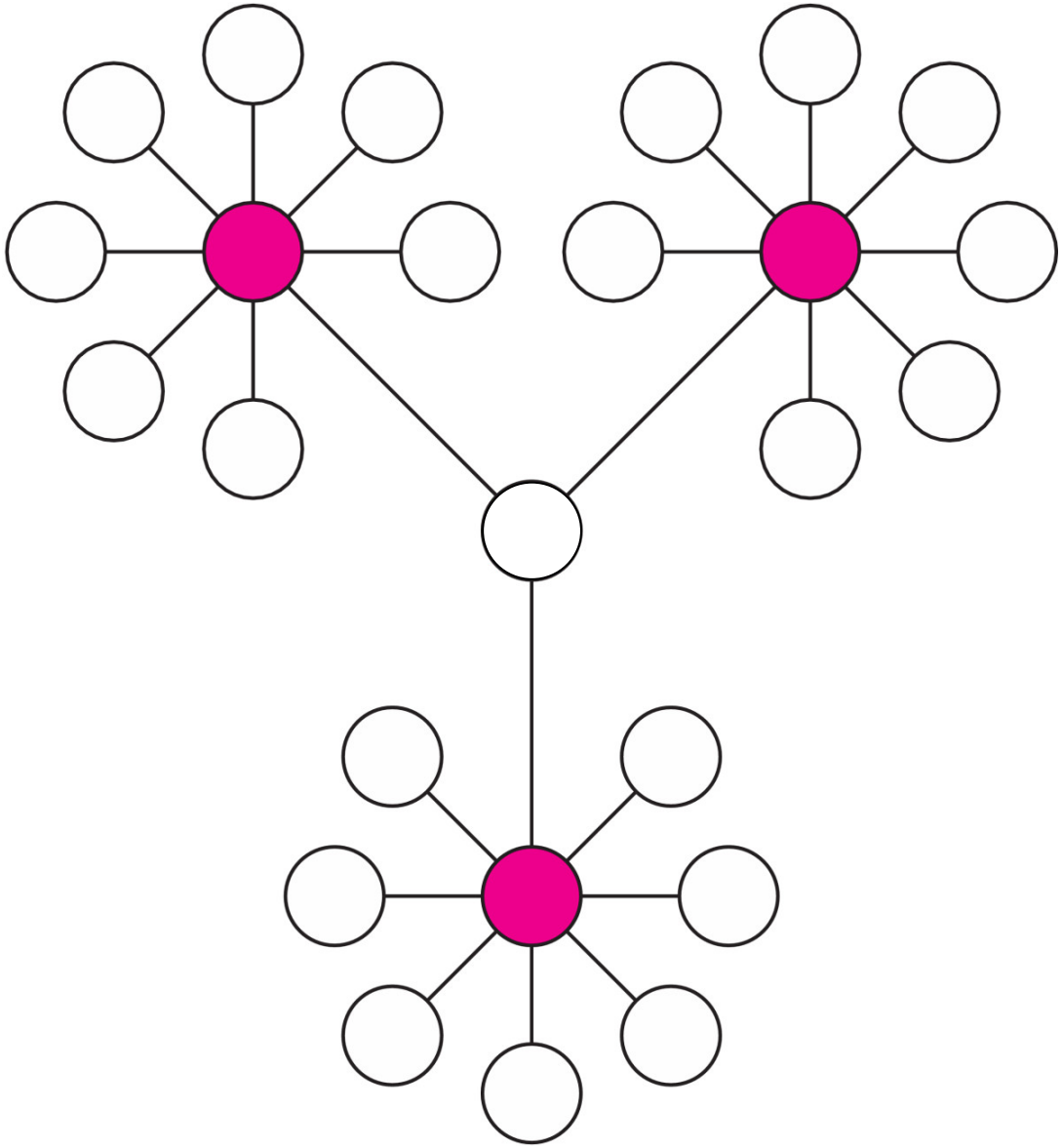

Core
Periphery

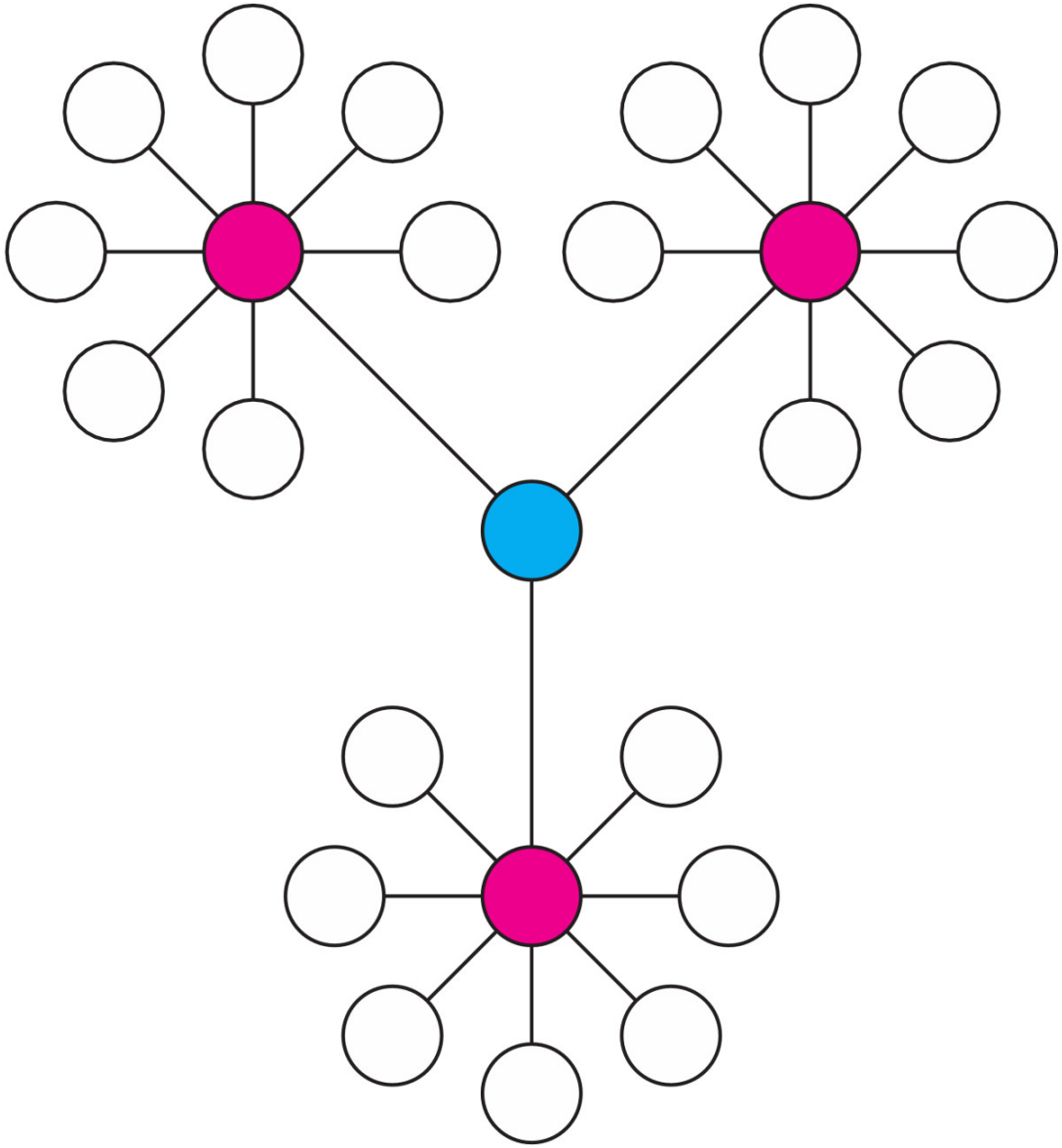

Community


Strong path


Hub







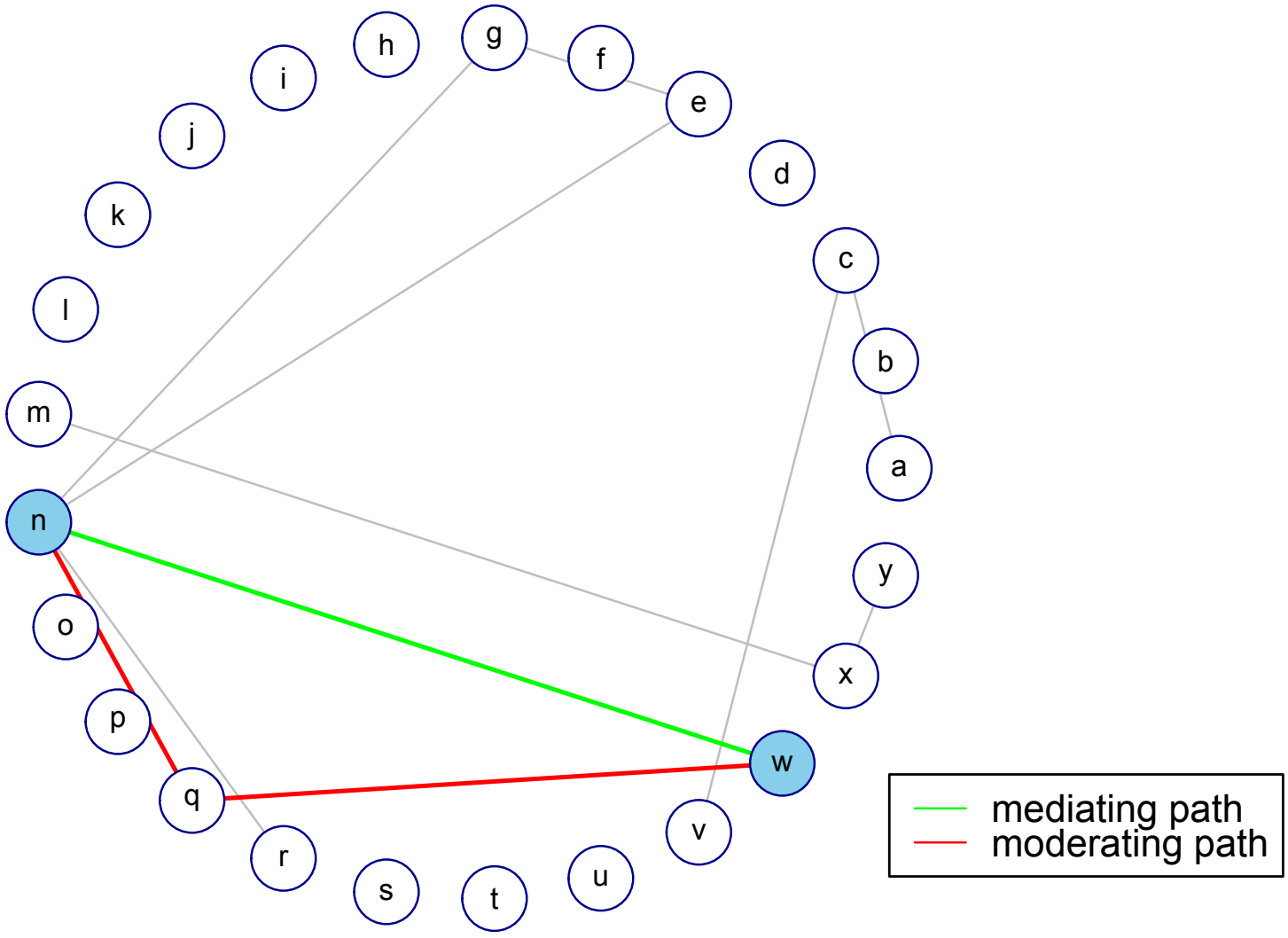
Exercise 6: Find the nodes with the top degree and betweenness centralities

Function for simple network statistics

```
GGMnetworkStats(sparseP ## sparse matrix
                )
```

Returns list object

- $degree: Degree centrality
- $betweenness: Betweenness centrality
- ...


h
g
f
i
e
j
d
k
c
l
b
m
a
n
y
o
x
p
w
q
v
r
s
t
u
mediating path
moderating path

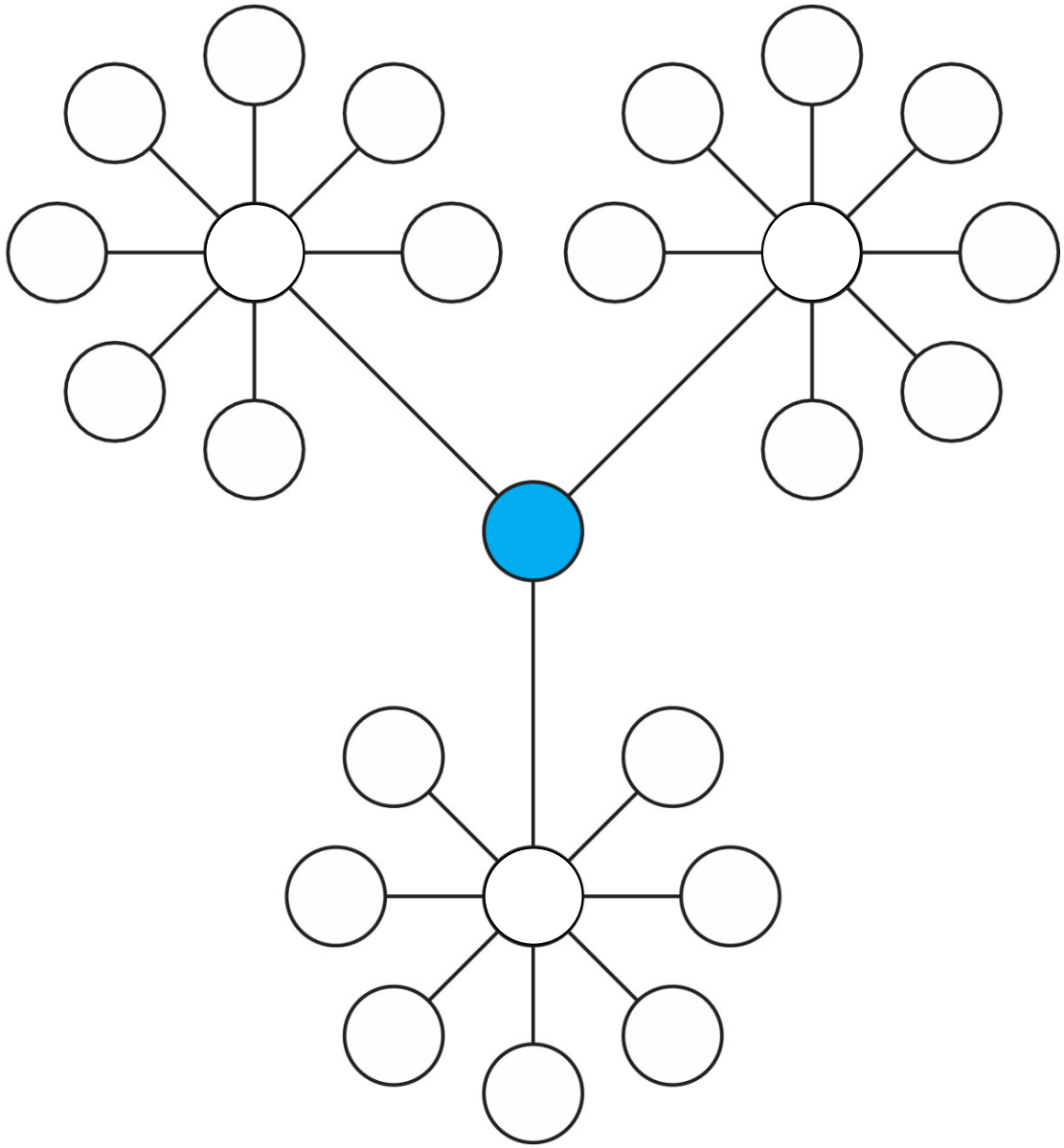
Exercise 7: Find the 2 strongest paths between Amines 1 and 2

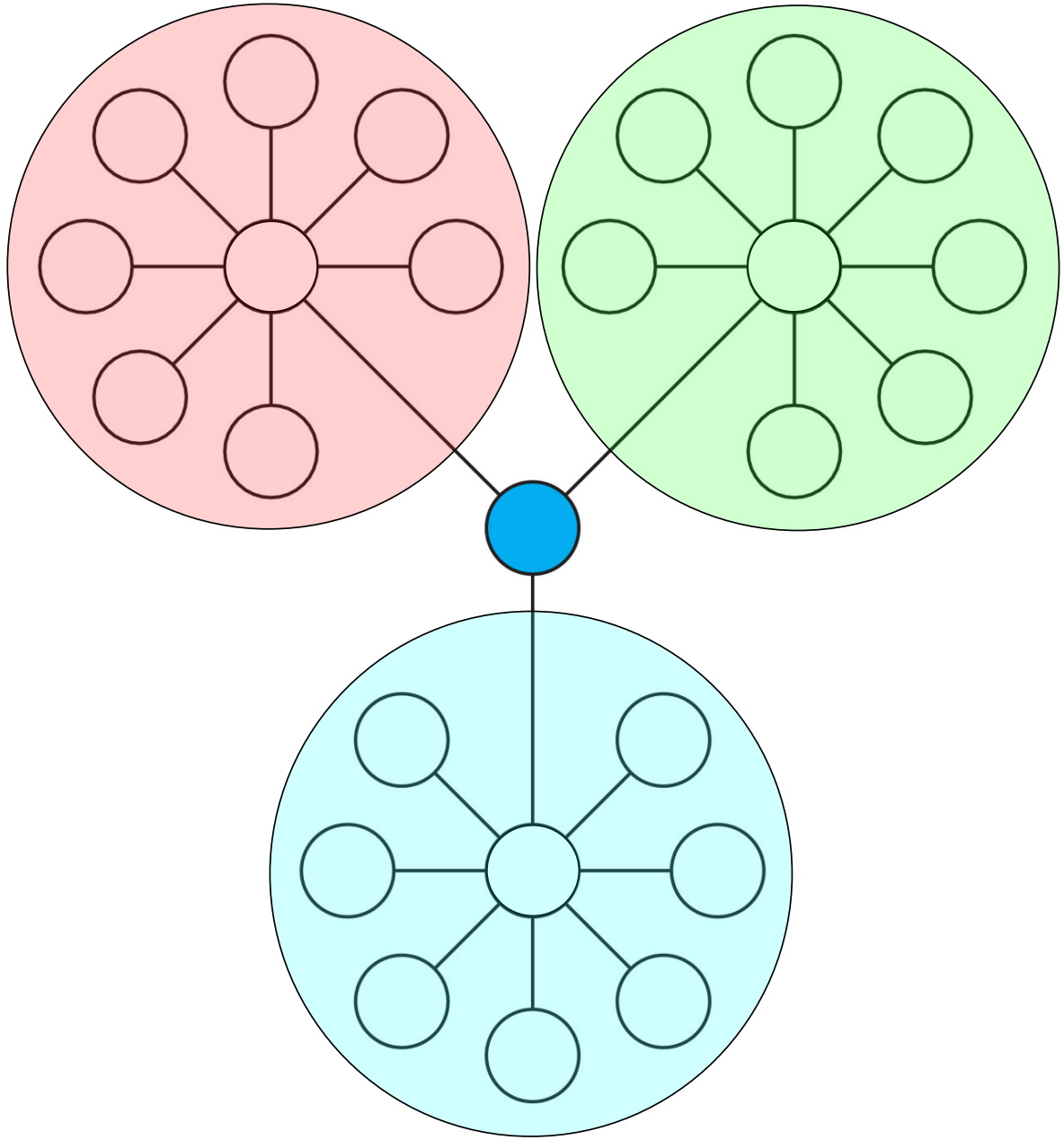
Function for path statistics

```
GGMpathStats(P0,      ## sparse precision matrix
             node1,   ## endpoint 1
             node2,   ## endpoint 2
             graph,   ## logical, should graph be produced
             ...      ## arguments to Ugraph
             )
```

Returns list object

- $pathStats: Matrix specifying paths
- ...





Exercise 8: Find and visualize communities

Function for community detection

```
Communities(P,      ## sparse matrix
            graph,  ## logical, graph is produced when TRUE
            ...     ## arguments to Ugraph
            )
```

Returns list object

- `$membership`: Community membership for each feature
- `$modularityscore`: Modularity score

Part II: Jointly extracting, visualizing, and analyzing multiple networks