Estimating cross-source relationships from wide big data using component- and network-analysis

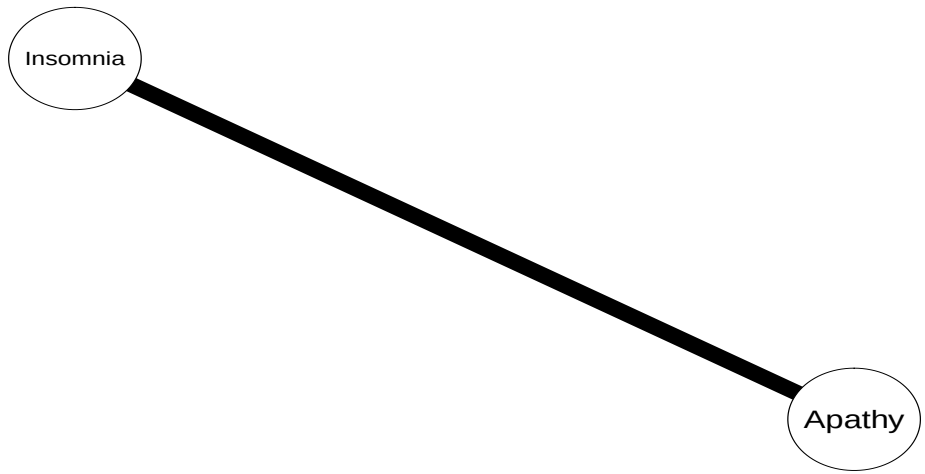
Pia Tio
Lourens Waldorp & Katrijn Van Deun

University of Amsterdam & Tilburg University
the Netherlands

June 8, 2018

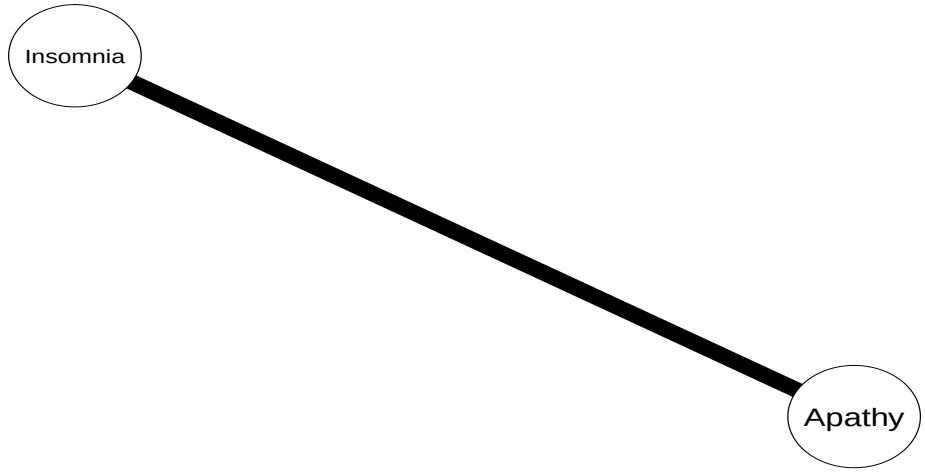
Network analysis

Idea of mutually interacting entities
- Nodes represent variables of interest
- Edges represent conditional depenency relationships



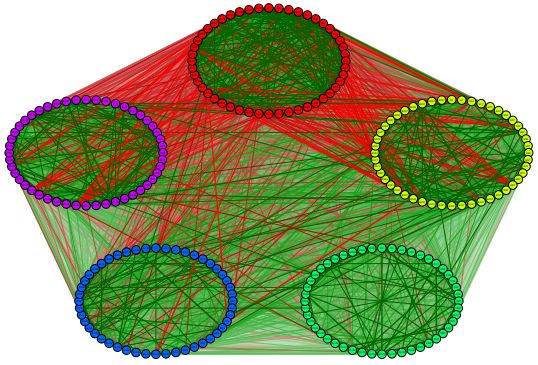
Network analysis

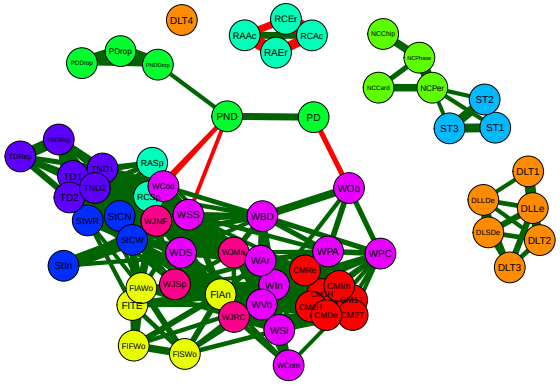
Idea of mutually interacting entities
- Nodes represent variables of interest
- Edges represent conditional depenency relationships

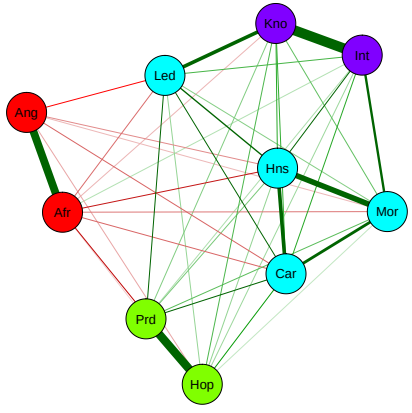


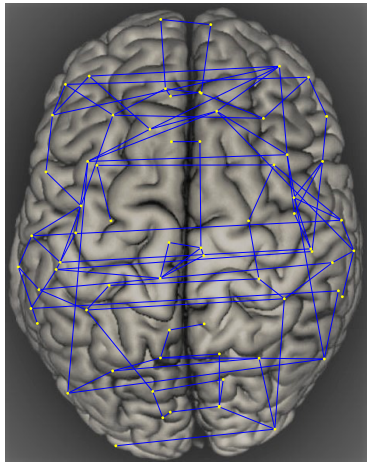
Key idea: For **Gaussian data**, inverse covariance matrix $\Sigma^{-1} =$ strength conditional dependence relationships

Network analysis in psychology

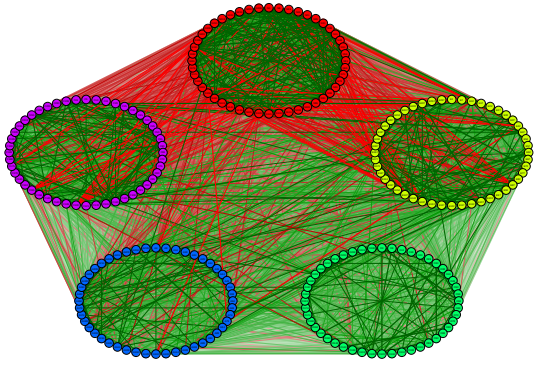


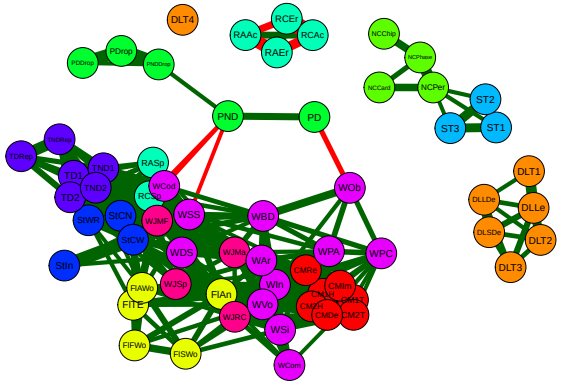




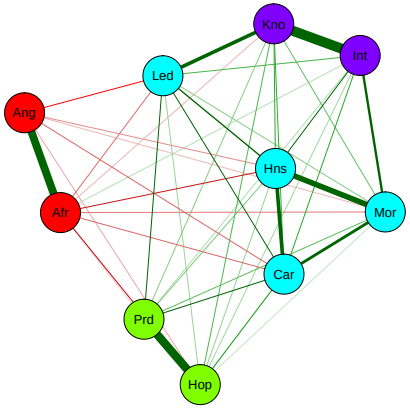


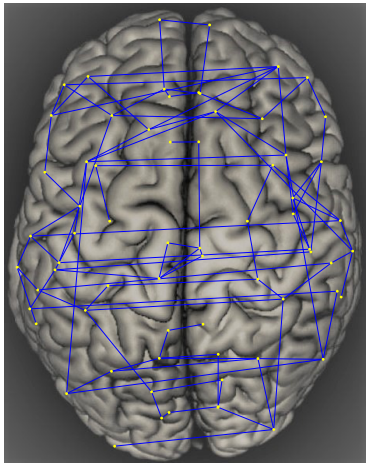
Are networks isolated island?



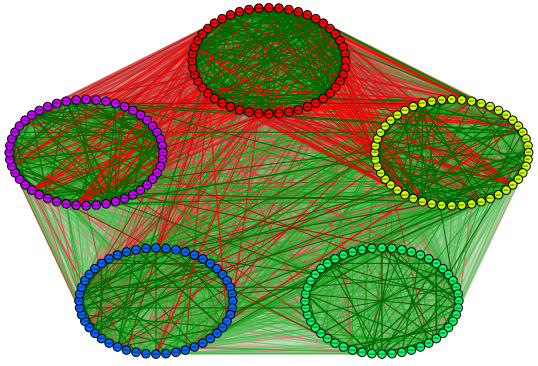


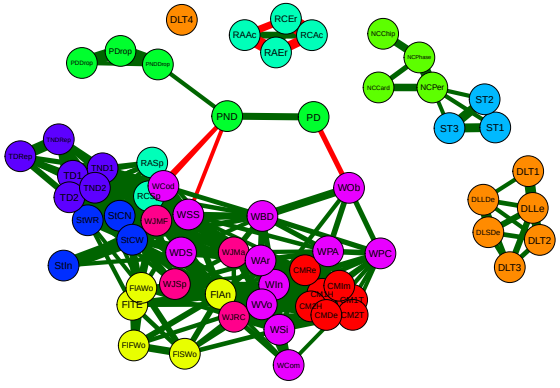
?

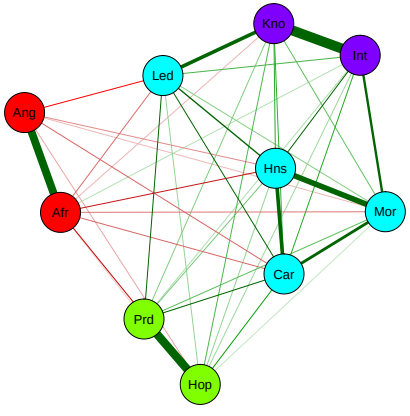


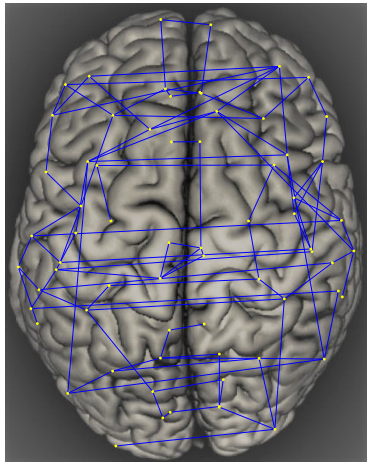


Goal: Estimating unqiue cross-source relationships using wide big data









Challenges estimating cross-source relationships

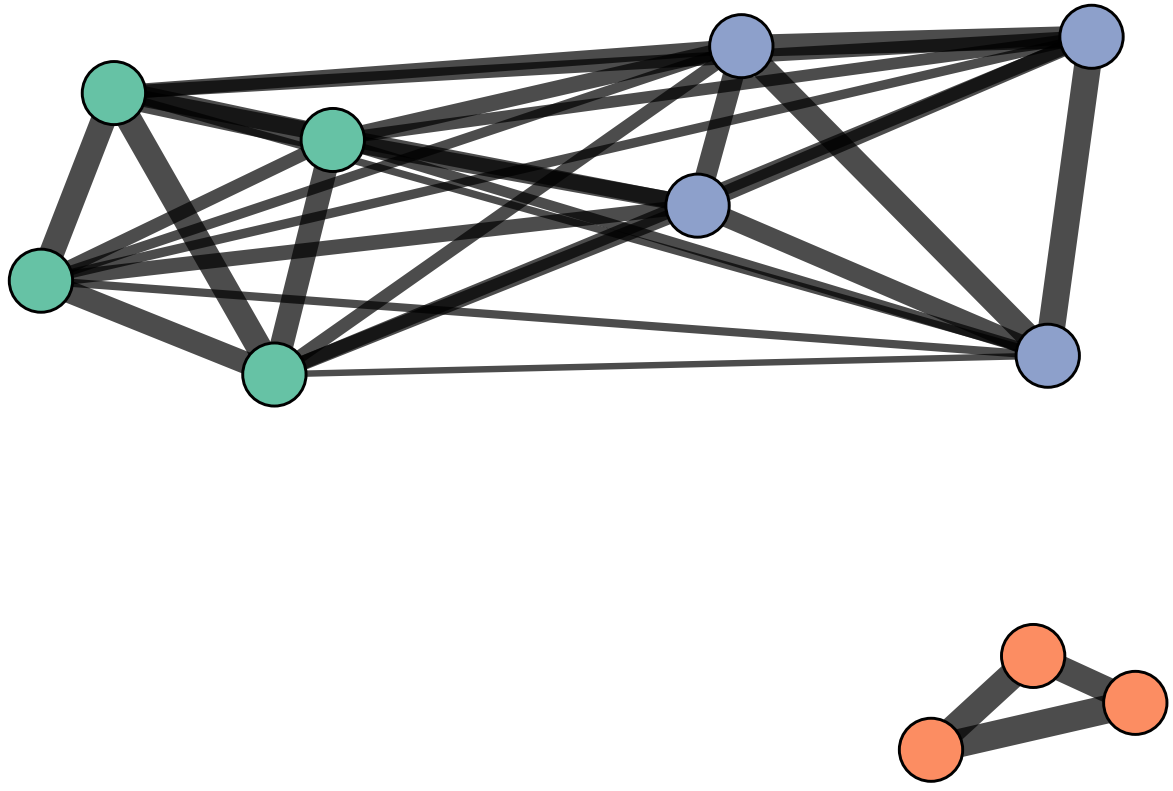
1. Different data sources may have different characteristics
2.
3.

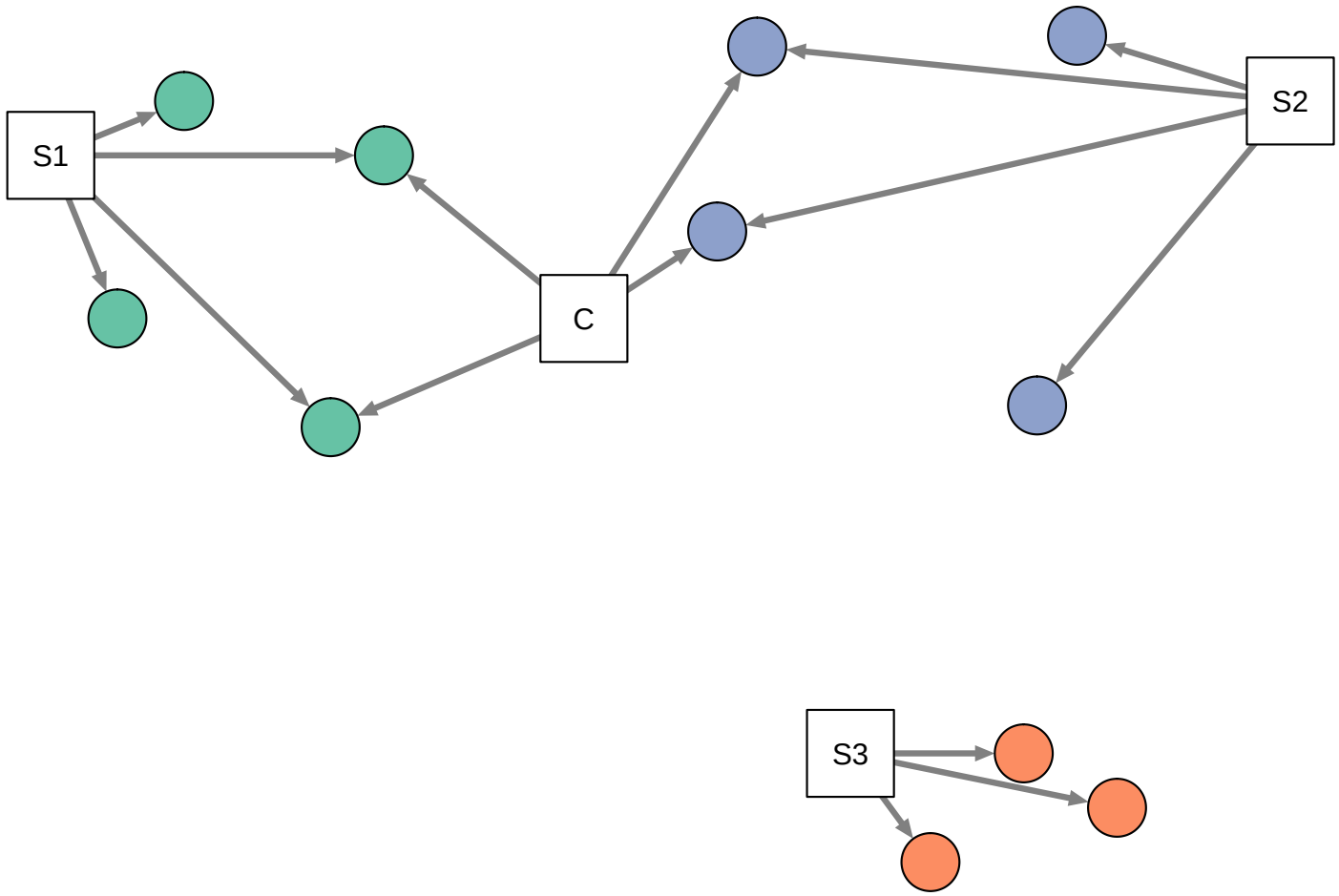
Challenges estimating cross-source relationships

1. Different data sources may have different characteristics
2. Dimension of data source: variables $>$ observations
3.

Challenges estimating cross-source relationships

1. Different data sources may have different characteristics
2. Dimension of data source: variables $>$ observations
3. Which variables are involved in cross-source relationships




S1
C
S2
S3

Component-model as pre-processing step

Component model

- Reduce dimensionality of data X by summarising p variables in r components ($r < p$)
- Components maximise the amount of variance explained

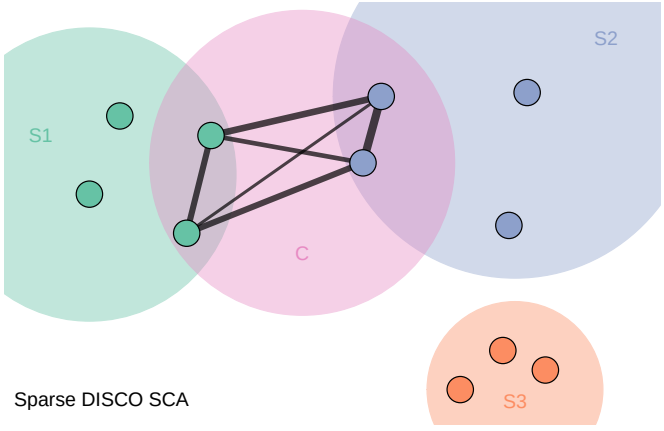
Sparse DIStinctive and COmmon Simultaneous Component Analysis (sparse DISCO SCA)

- Components can be source-specific (1 source) or common (1+ sources)

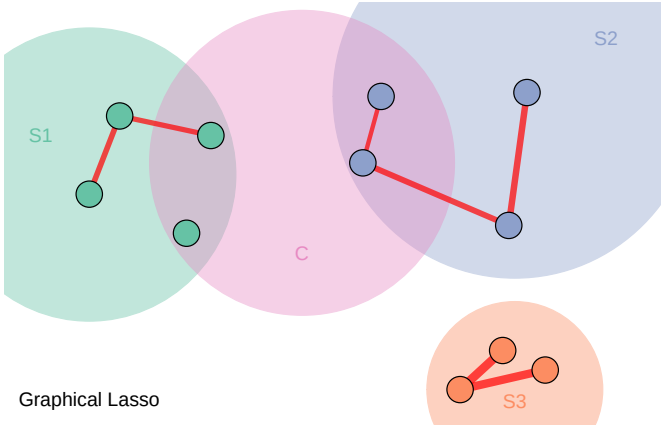
Component-model as pre-processing step

	Common	Source 1-specific	Source 2-specific
Source 1	0	0	0
	$x_{2,1}$	$x_{2,2}$	0
	$x_{3,1}$	$x_{3,2}$	0
	0	0	0
	$x_{5,1}$	$x_{5,2}$	0
	0	0	0
	0	$x_{7,2}$	0
Source 2	$x_{8,1}$	0	0
	$x_{9,1}$	0	$x_{9,3}$
	0	0	0
	$x_{11,1}$	0	0
	0	0	$x_{12,3}$
	$x_{13,1}$	0	$x_{13,3}$
	$x_{14,1}$	0	$x_{14,3}$

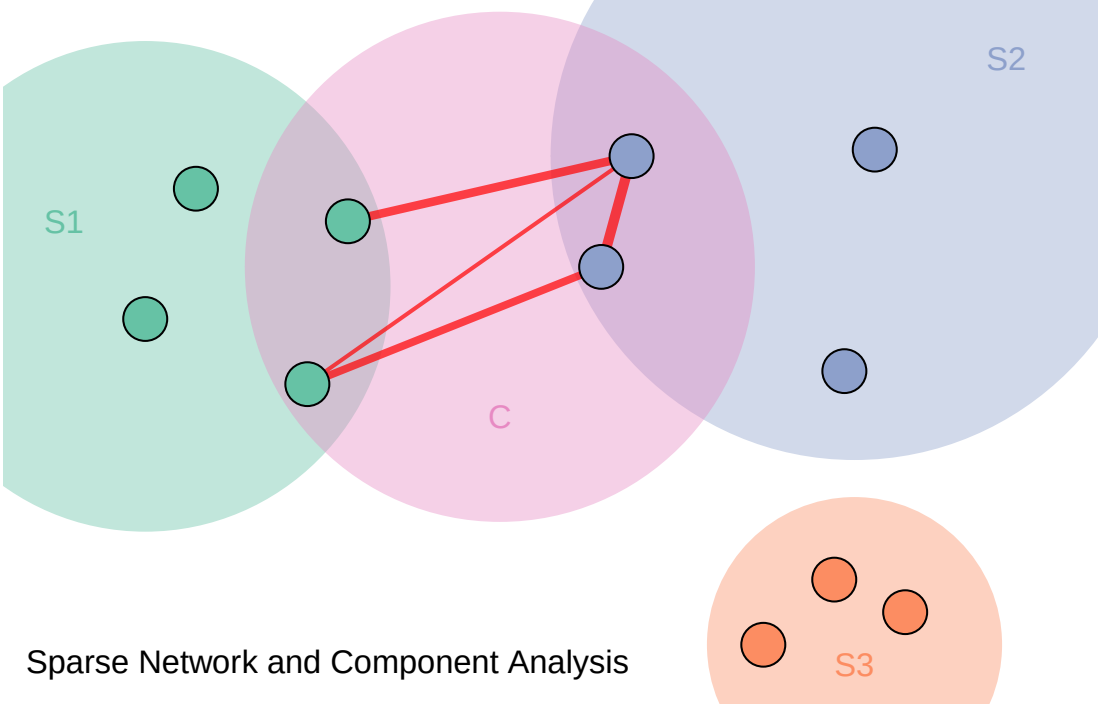
Sparse Network And Component analysis (SNAC)



Sparse DISCO SCA



Graphical Lasso



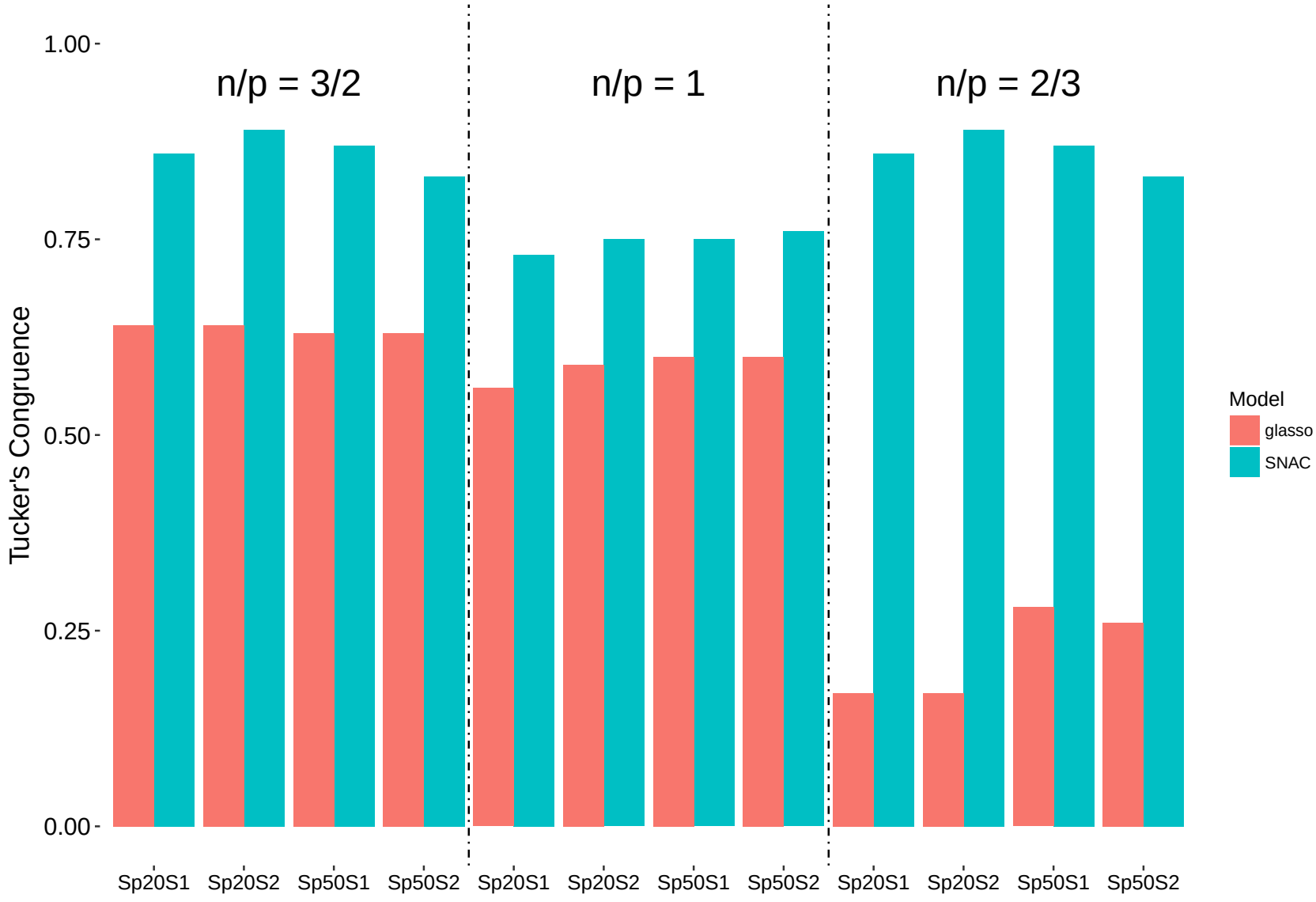
Sparse Network and Component Analysis

Does SNAC accurately recover cross-source relationships?

Simulation study
2 data sources: 1 common & 2 set-specific components
- Ratio observations/variables: $\mathbf{3/2}$, $\mathbf{1}$, $\mathbf{2/3}$
- Sparseness of component: $\mathbf{20\%}$ or $\mathbf{50\%}$
- Importance of the common component: **equal** or **less important**

Tucker's congruence
Similarity between population and estimated inverse covariance matrix

Does SNAC accurately recover cross-source relationships?



Empirical example: Alzheimer's disease

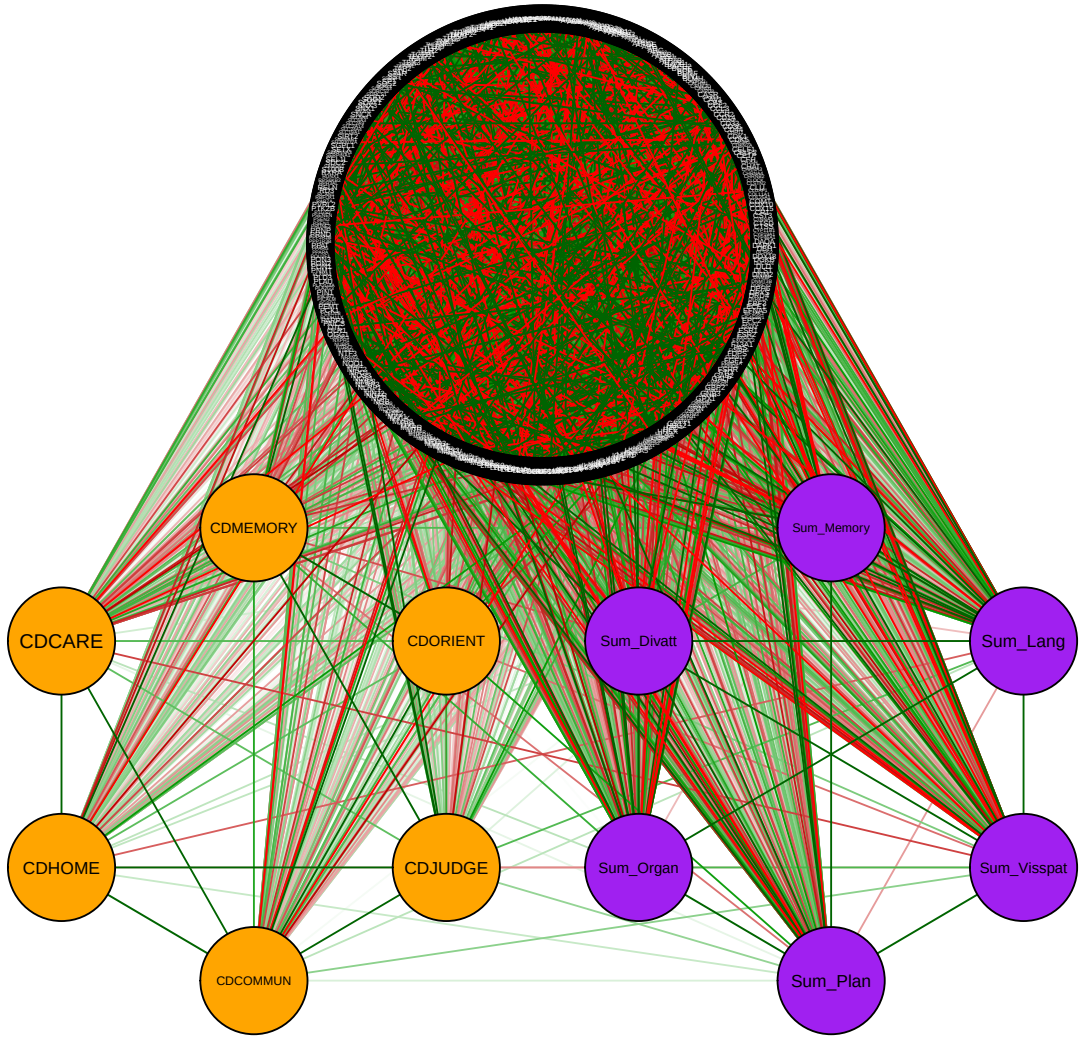
Decrease in cognitive functioning may be related to certain genes
Alzheimer's Disease Neuroimaging Initiative (ADNI) [1]

Subset of 175 participants
- 388 genes [2]
- 6 measures of everyday cognitive functioning (self-report)
- 6 measures of dementia (clinical assessment)

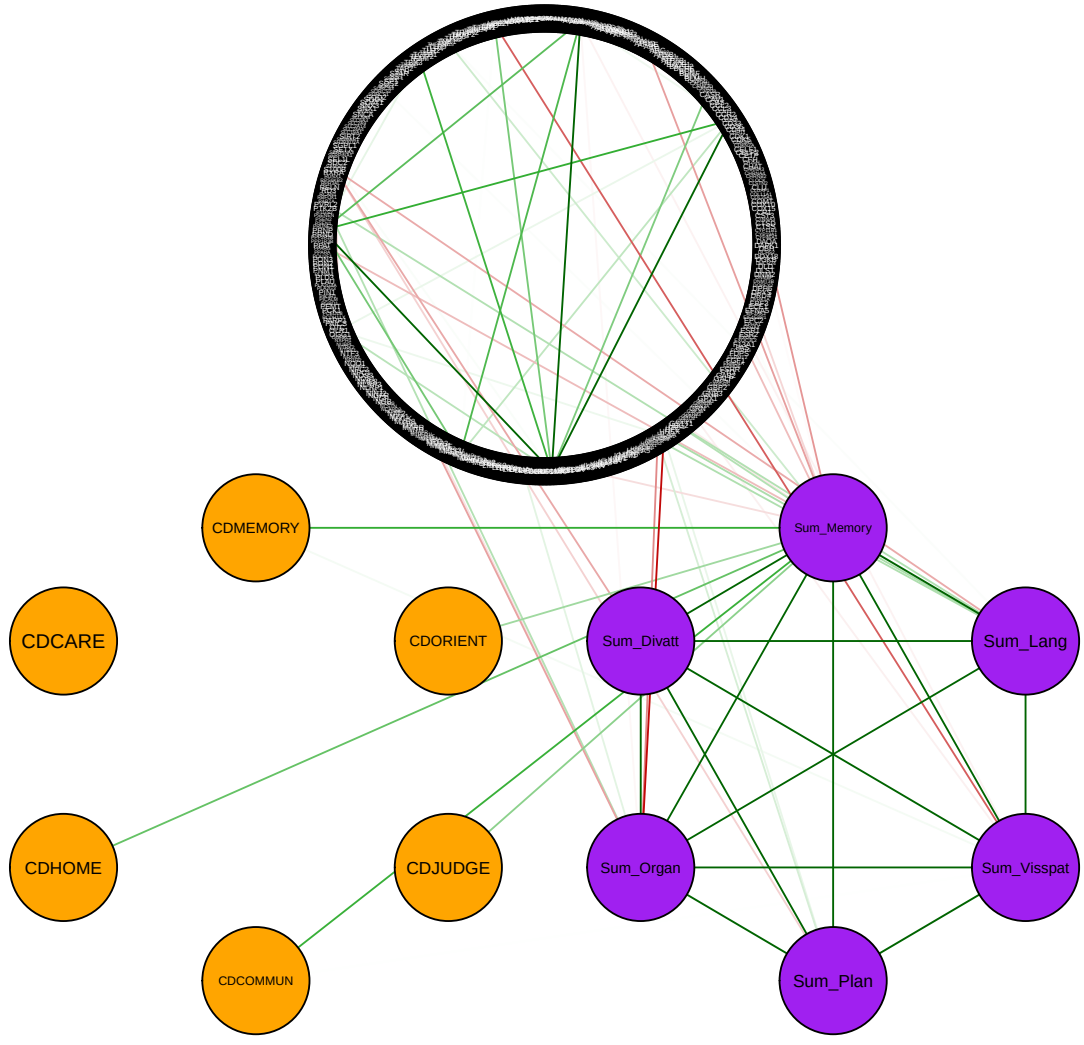
[1] adni.loni.usc.edu

[2] gene selection based on Hu, Xin, Hu, Zhnag, & Want (2017)

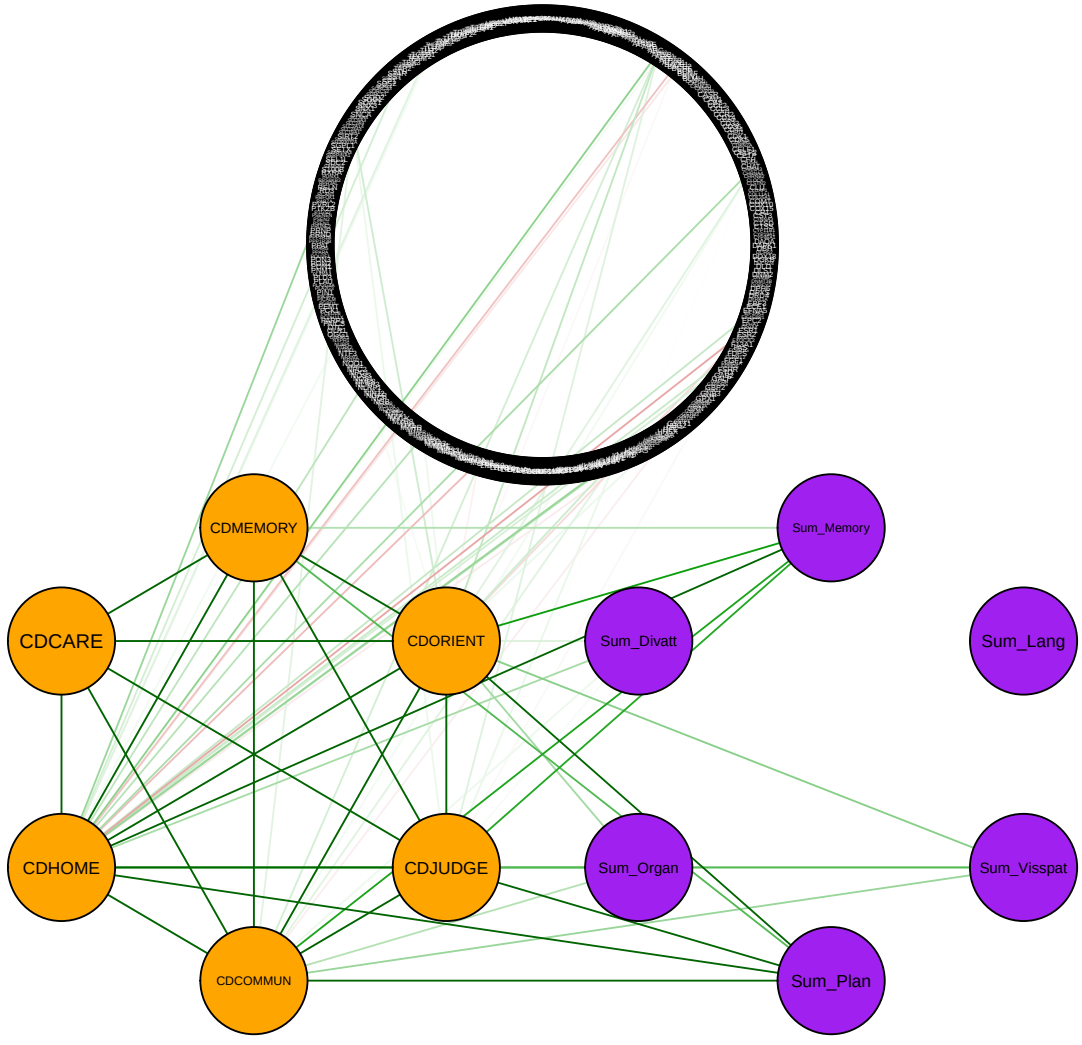
Network analysis



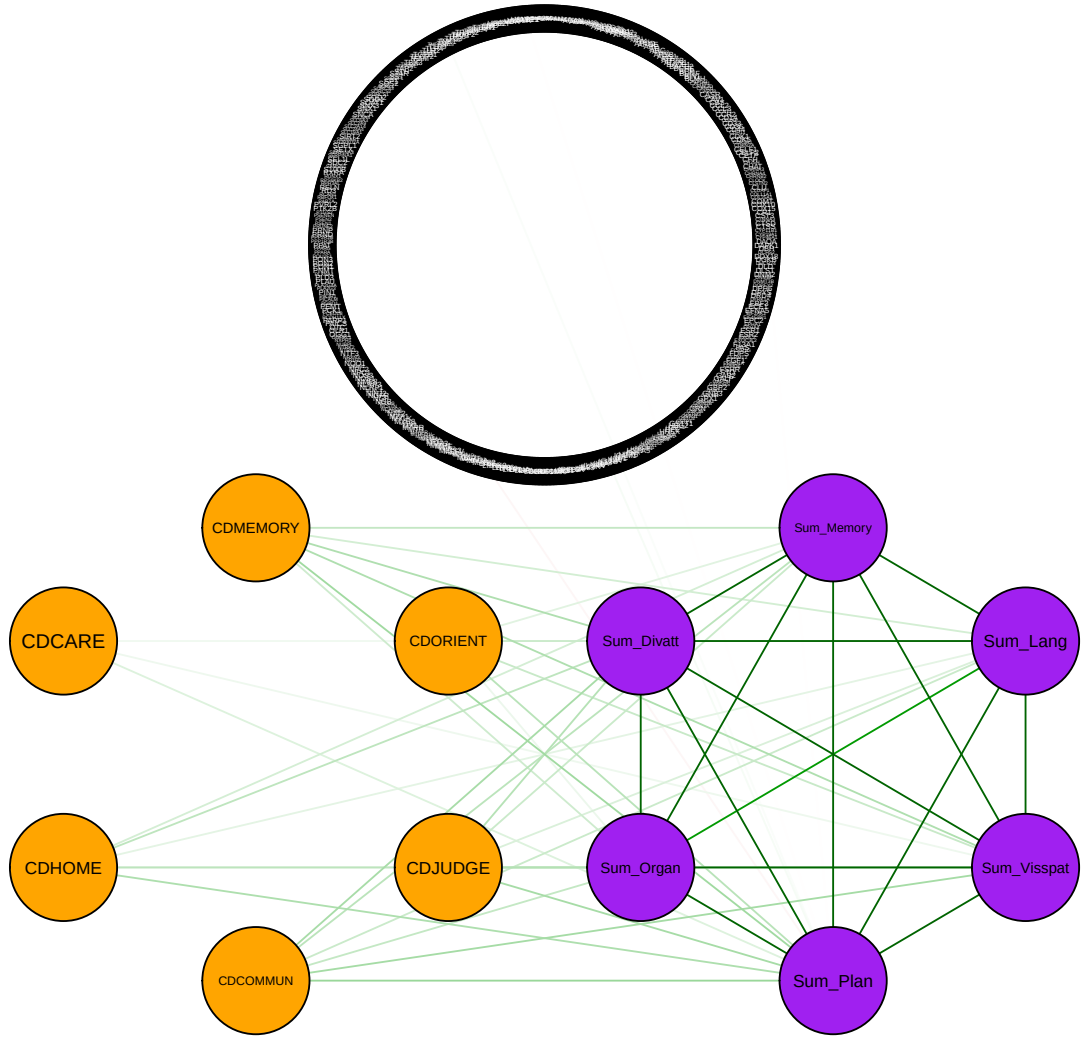
Network analysis with regularisation

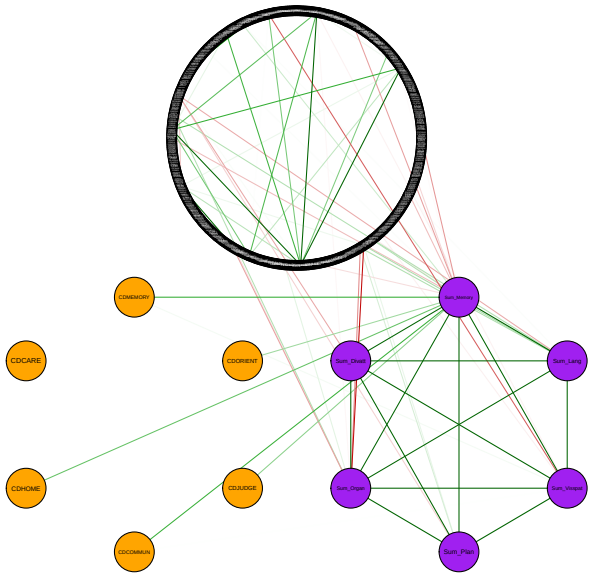


SNAC: Component 1

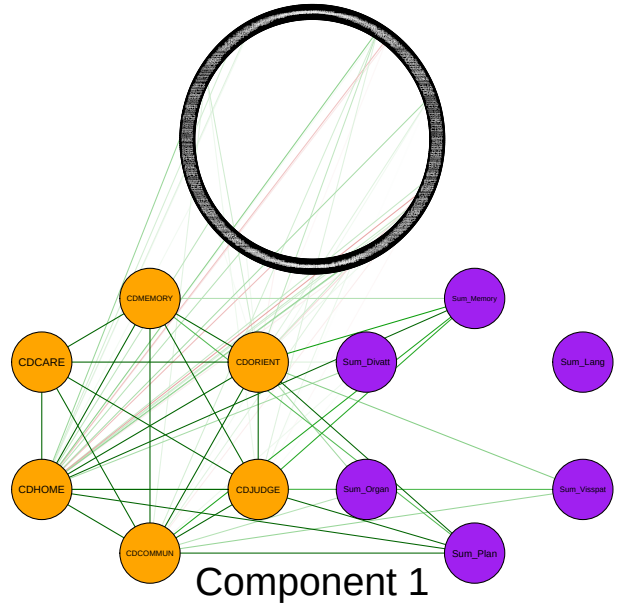


SNAC: Component 2

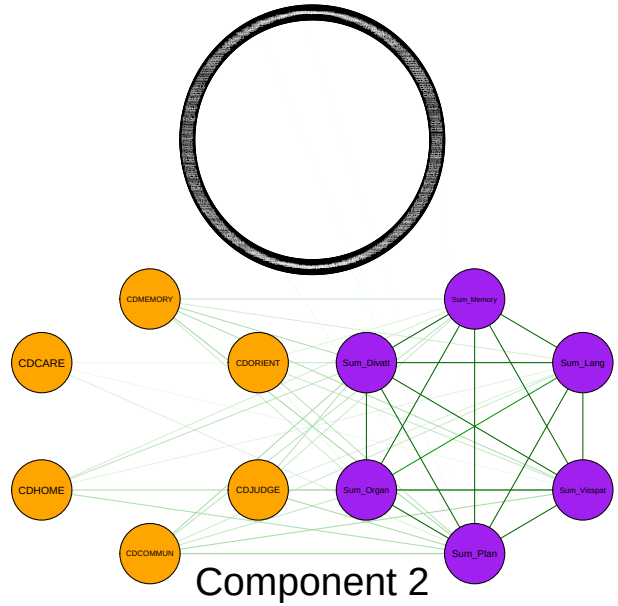




Network analysis with regularisation



Component 1



Component 2

Take home message

Sparse Network And Component (SNAC) analysis

- a **hybrid component-network analysis**
- a promising tool for modelling **unique relationships between different data sources** especially when **p** $>$ **n**
- provides insight in **how various disciplines are connected** to one another.

Note: Gaussian data only (for now)

Thank you for your attention

Questions/comments?
Please contact me at piatio@gmail.com