# Networks and Sparse Graphical Models

Ernst Wit

e.c.wit@rug.nl

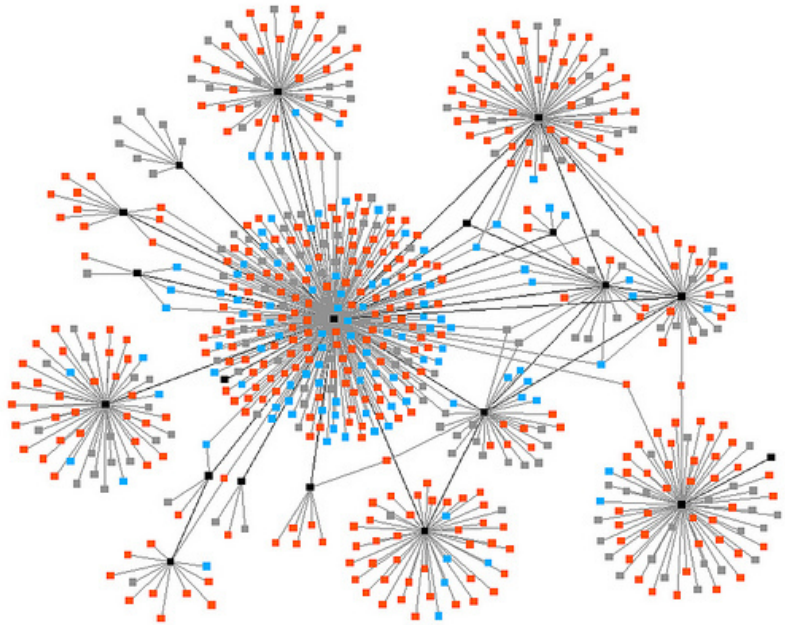
Johann Bernoulli Institute
University of Groningen

21 January, 2014

rijksuniversiteit groningen

# Motivation: Features of a Dynamic Genomic Process

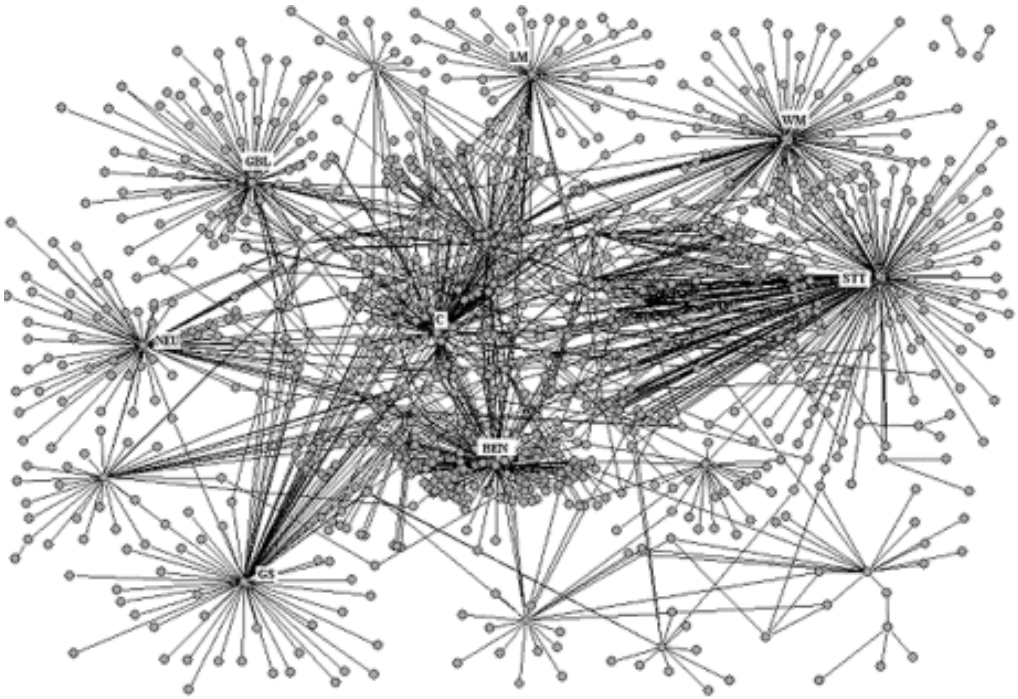
- "**p** >> **n**":
  - Number of observations smaller than number of variables.
  - Thousands of variables and hundreds of observations.
- **Structure:**
  - Highly complex and structured phenomenon.
  - Possibly with additional topographical structure (small world).
- **Sparsity:** only small number of links between nodes.



rijksuniversiteit groningen

# "Network" as Graphical Model

**In Genomics:** typically have measurements of nodes
**Examples:** RNA-seq, GWAS, proteomics

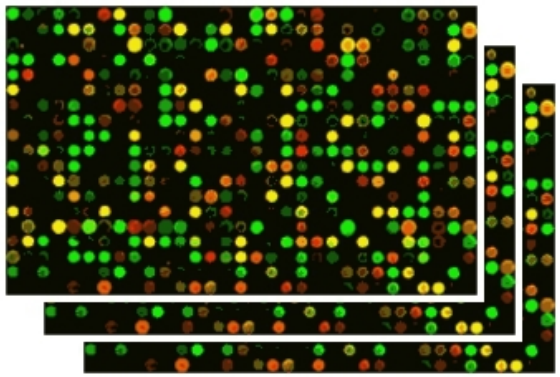



**Proposal:** interpret network as *conditional independence* relations.

rijksuniversiteit groningen

# Motivation: Dynamic Genomic Networks

- Transcription: snap-shot of gene activity in time and space.
- Microarray and RNA-seq data measure gene activity.



**Running example: T-cell time-series dataset.**

- Temporal expression of 58 genes for 10 spaced time points.
- At each time point there are 44 separate measurements.

## Definition (Aim)

Determine dynamic genomic graph $G$ on basis of $\{Y_{gt}^{(i)}\}_{gti}$.

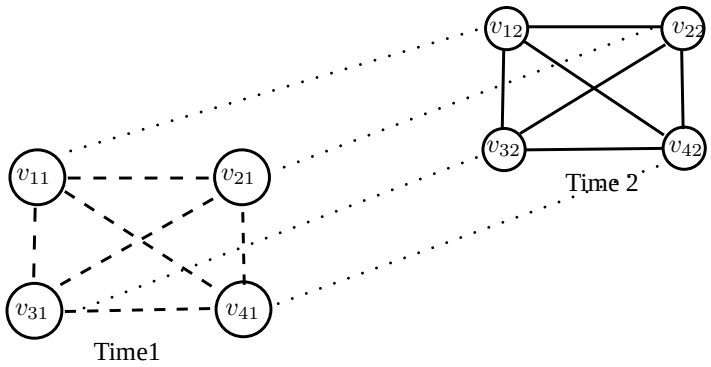
# Dynamic Genomic Graphs

- $\Gamma$ be a set of "genes".
- $T$ be a set of ordered "time points".

## Definition (Dynamic genomic graph)

A dynamic genomic graph is a pair $G = (V, E)$.

- Vertices: $V = \{v_{ij}\}_{i \in \Gamma, j \in T}$, where $\Gamma$ and $T$ are finite sets.
- Links: ordered pair of elements $E \subseteq V \times V$.




rijksuniversiteit groningen

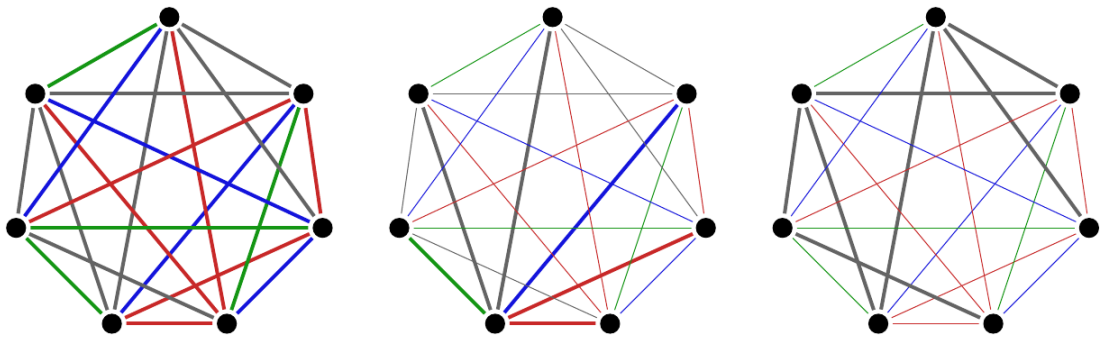
# Coloured graphs

## Definition (Coloured graph)

A coloured graph is a triplet $G_F = (V, E, F)$, where $G = (V, E)$ is a graph and $F$ is a mapping on the links, i.e.:

$$F : E \to C,$$

where $C$ is a finite set of colours.

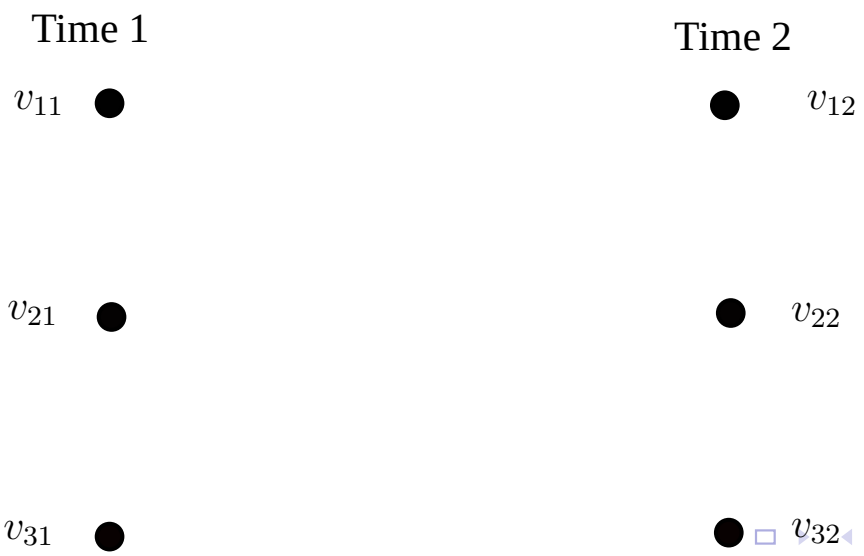


# Special kind of coloured graphs: Factorial Graphs

Denote mapping $F : E \to C$ by $E \prec F$.

In analogy with ANOVA, we define the following colouring:

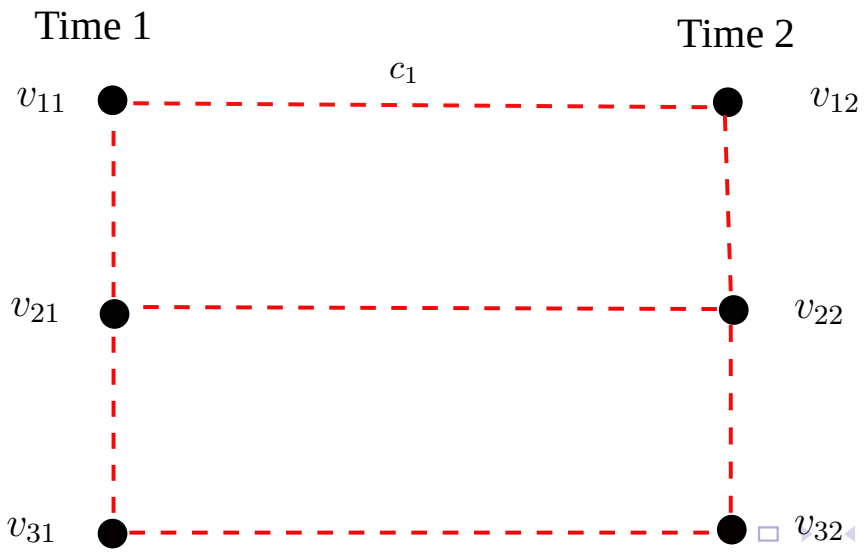
- $\boxed{E \prec 0 \Rightarrow \text{an empty graph.}}$
- $E \prec F_1$.
- $E \prec F_T$.
- $E \prec F_\Gamma$.
- $E \prec F_{\Gamma T}$.



# Special kind of coloured graphs: Factorial Graphs

Denote mapping $F : E \to C$ by $E \prec F$.
In analogy with ANOVA, we define the following colouring:

- $E \prec 0 \Rightarrow$ an empty graph.
- $\boxed{E \prec F_1\text{: same colour for all links}}$
- $E \prec F_T$.
- $E \prec F_\Gamma$.
- $E \prec F_{\Gamma T}$.

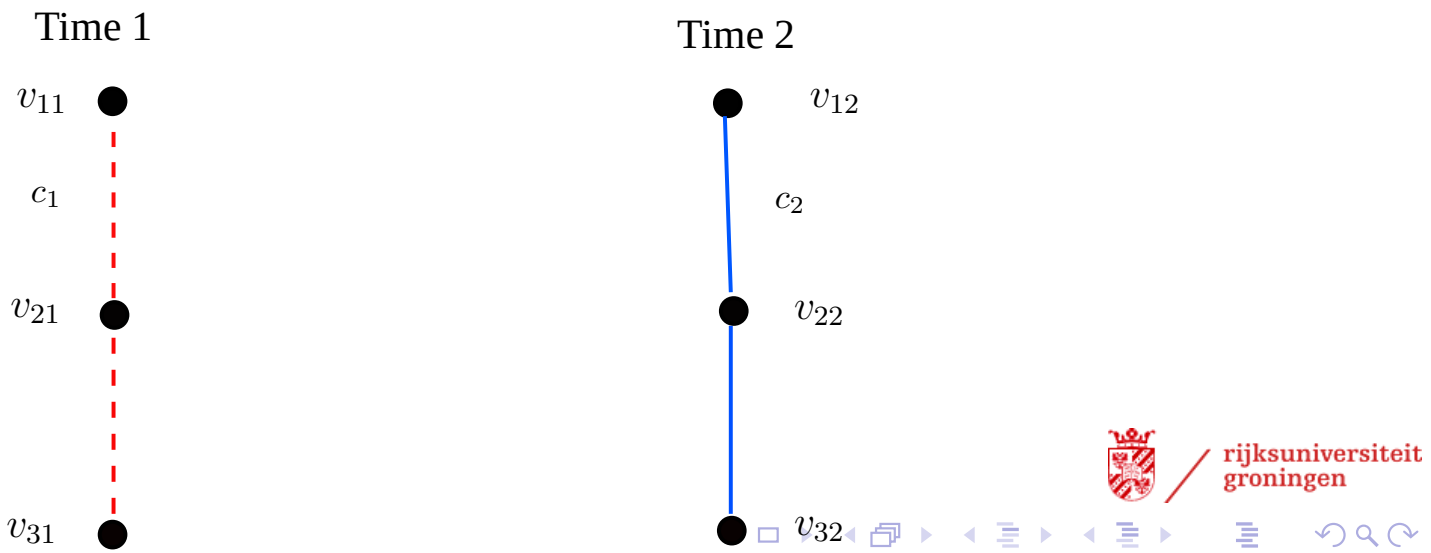



rijksuniversiteit groningen

# Special kind of coloured graphs: Factorial Graphs

Denote mapping $F : E \to C$ by $E \prec F$.
In analogy with ANOVA, we define the following colouring:

- $E \prec 0 \Rightarrow$ an empty graph.
- $E \prec F_1$.
- $\boxed{E \prec F_T\text{: same colour across all genes}}$
- $E \prec F_\Gamma$.
- $E \prec F_{\Gamma T}$.

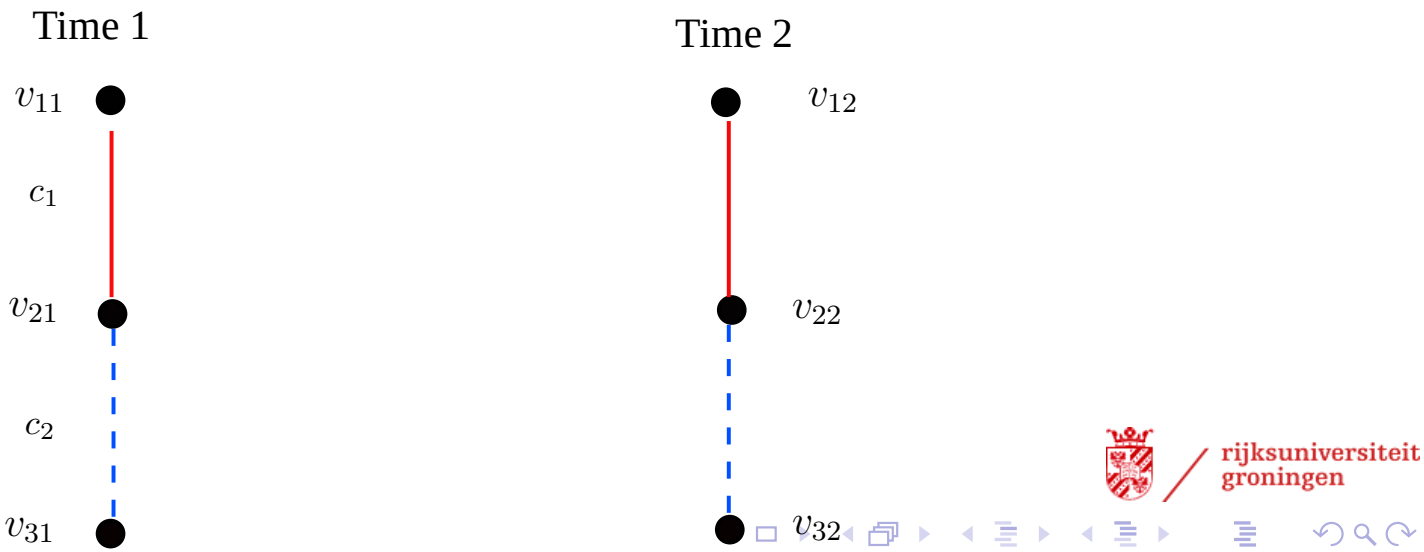



rijksuniversiteit groningen

# Special kind of coloured graphs: Factorial Graphs

Denote mapping $F : E \to C$ by $E \prec F$.
In analogy with ANOVA, we define the following colouring:

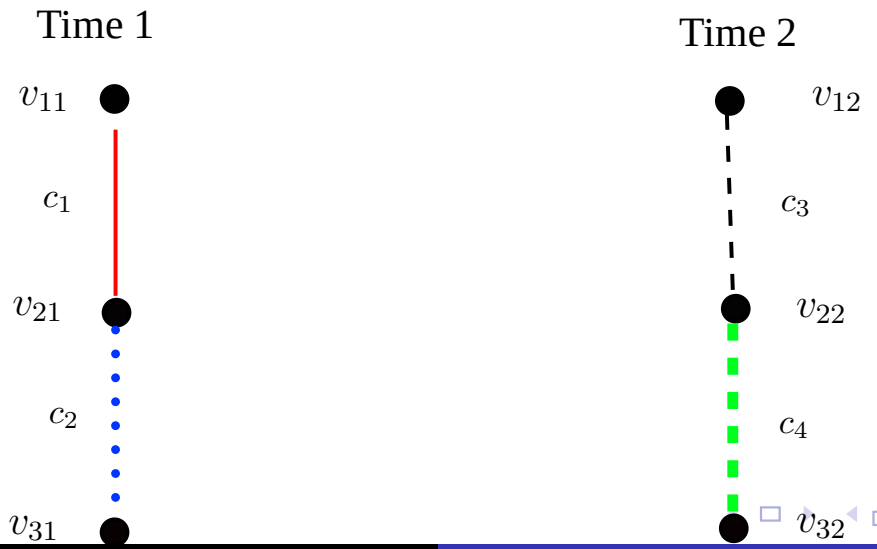
- $E \prec 0 \Rightarrow$ an empty graph.
- $E \prec F_1$.
- $E \prec F_T$.
- $\boxed{E \prec F_\Gamma\text{: same colour across all times}}$
- $E \prec F_{\Gamma T}$.



# Special kind of coloured graphs: Factorial Graphs

Denote mapping $F : E \to C$ by $E \prec F$.
In analogy with ANOVA, we define the following colouring:

- $E \prec 0 \Rightarrow$ an empty graph.
- $E \prec F_1$.
- $E \prec F_T$.
- $E \prec F_\Gamma$.
- $\boxed{E \prec F_{\Gamma T}\text{: all different colours}}$



# Natural Partitions

## Definition (Natural partition)

Let $E = \{S_i, N_i\}_{i=0}^{n_T-1}$ be subsets of links where $S_i, N_i$ are defined as follows:

$$S_i = \{\{(v_{jt}, v_{j,t+i}), (v_{j,t+i}, v_{jt})\} | j \in \Gamma, t = 1, \ldots, n_T - i\},$$

and

$$N_i = \{\{(v_{jt}, v_{k,t+i}), (v_{k,t+i}, v_{jt})\} | \forall j \neq k \in \Gamma, t = 1, \ldots, n_T - i\}.$$

The natural partitions imply subgraphs of $G$ and imply partions of $\mathbf{\Theta}$ for GGMs:

$$\mathbf{\Theta} = \left[\begin{array}{cc|cc|cc|cc} S_0 & N_0 & S_1 & N_1 & S_2 & N_2 & \ldots & \ldots \\ & S_0 & N_1 & S_1 & N_2 & S_2 & \ldots & \ldots \\ \hline & & S_0 & N_0 & S_1 & N_1 & S_2 & N_2 \\ & & & S_0 & N_1 & S_1 & N_2 & S_2 \\ \hline & & & & S_0 & N_0 & S_1 & N_1 \\ & & & & & S_0 & N_1 & S_1 \end{array}\right]$$

# Factorial Graphical Models

## Definition (Gaussian graphical models for factorially coloured graphs)

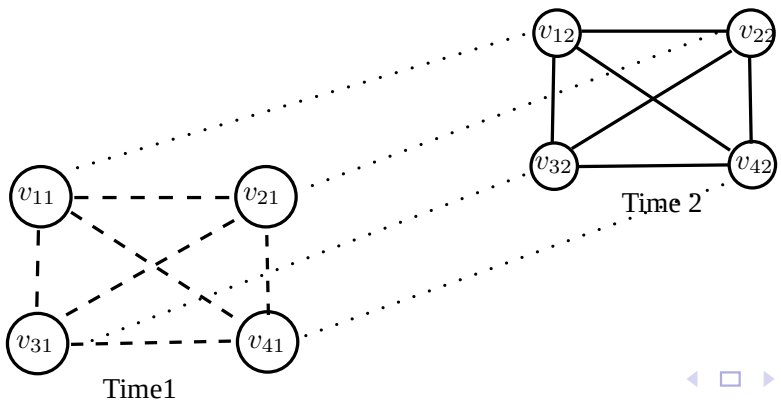
A factorial Gaussian graphical model is a graphical model defined on:

- a dynamic factorial graph $G = (V, E, F)$, where
- a factorial colouring $F$ is applied *separately* to natural partitions

$$S_i \prec F_{S_i}, \quad N_i \prec F_{N_i}, \quad i = 0, \ldots, n_T - 1$$

- which determines $\Theta$ in

$$\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Theta}^{-1}).$$



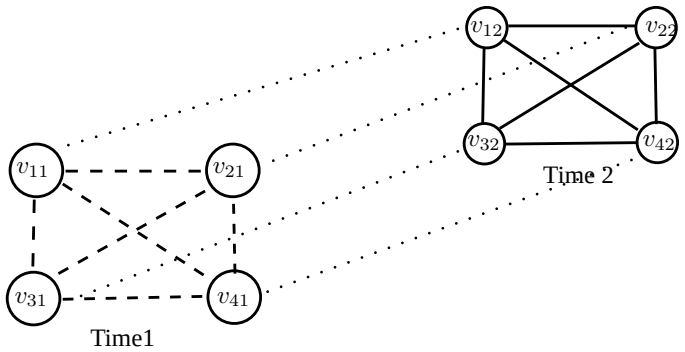

rijksuniversiteit groningen

# Example: Factorial Gaussian Graphical Model

**Model:**

$$(S_0 \prec 1), \quad N_0 \prec F_T, \quad S_1 \prec 1, \quad N_1 \prec 0.$$

**Factorial coloured graph:**




**Precision Matrix:**

$$\boldsymbol{\Theta} = \left[\begin{array}{cccc|cccc} \theta_1 & \theta_2 & \theta_2 & \theta_2 & \theta_4 & 0 & 0 & 0 \\ & \theta_1 & \theta_2 & \theta_2 & 0 & \theta_4 & 0 & 0 \\ & & \theta_1 & \theta_2 & 0 & 0 & \theta_4 & 0 \\ & & & \theta_1 & 0 & 0 & 0 & \theta_4 \\ \hline & & & & \theta_1 & \theta_3 & \theta_3 & \theta_3 \\ & & & & & \theta_1 & \theta_3 & \theta_3 \\ & & & & & & \theta_1 & \theta_3 \\ & & & & & & & \theta_1 \end{array}\right]$$

# Penalized likelihood for GGMs

- Consider an experiment: $|\Gamma|$ genes measured across $|T|$ time points.
- Assume $n$ iid samples $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$, where $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_{\Gamma T}^{(i)})$.
- Assume $\mathbf{Y}^{(i)} \sim N(\mathbf{0}, \mathbf{\Theta}^{-1})$, then

**Likelihood:**

$$l(\mathbf{\Theta}|\mathbf{y}) \propto \log(|\Theta|) - \mathrm{tr}(\mathbf{S\Theta}).$$

**AIM:** Optimization of penalized likelihood:

$$\hat{\mathbf{\Theta}} := \mathrm{argmax}_{\mathbf{\Theta}}\{l(\mathbf{\Theta}|\mathbf{y})\}$$

subject to

- $\mathbf{\Theta} \succeq 0$;
- $||\mathbf{\Theta}||_1 \leq 1/\lambda$;
- some factorial colouring $F$.

# *Numquam ponenda est pluralitas sine necessitate*

William Occam (1288-1348) proposed a meta-theory of knowledge:

> *"For nothing ought to be posited without necessity."*

Can be interpreted *statistically* as a

- **Aesthetic principle:** enhances model interpretability through parsimonious representation
- **Pragmatic principle:** computability.
- **Ontological principle:** represents expectation about nature of solution.
- **Prediction principle:** bias-variance trade-off

# Inferring penalized factorial Gaussian graphical models

**LogdetPPA.** Newton-CG primal proximal point algorithm (*Wang et al., 2010*, including Kim Toh and Defeng Sun) is used to solve optimization:

$$\hat{\boldsymbol{\Theta}} := \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} -\{\log|\boldsymbol{\Theta}| - \operatorname{tr}(\boldsymbol{\Theta}\mathbf{S}) + \boldsymbol{\lambda}'\boldsymbol{\theta}^{+} + \boldsymbol{\lambda}'\boldsymbol{\theta}^{-} : A(\boldsymbol{\Theta}) = \mathbf{0},$$

$$B(\boldsymbol{\Theta}) - \boldsymbol{\theta}^{+} + \boldsymbol{\theta}^{-} = \mathbf{0}, \boldsymbol{\Theta} \succeq 0, \boldsymbol{\theta}^{+}, \boldsymbol{\theta}^{-} \geq 0\}$$

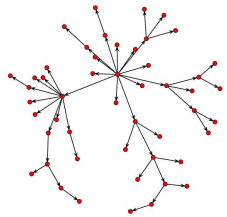
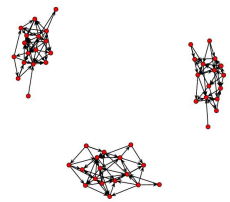
- $A(\boldsymbol{\Theta})$: linear constraints which depend on coloured graph.
- $B(\boldsymbol{\Theta})$: $\ell_1$-norm penalty on elements of precision matrix.
- $\boldsymbol{\theta}^{+}$ and $\boldsymbol{\theta}^{-}$ are additional variables (slack variables).
- $\boldsymbol{\Theta} \succeq 0$: semi-positive definite constraint.

**Solves $\hat{\Theta}$ up to** $2000 \times 2000$ .

rijksuniversiteit groningen

# Simulations: Lag 0 network identification

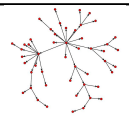
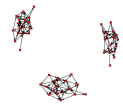
Table : $n = 50$ and $t = 10$ for various number of genes (p)

| | | $p = 50$ | $p = 100$ | $p = 50$ | $p = 100$ |
|---|---|---|---|---|---|
| SCAD | SEN | 0.987 | 0.957 | 0.992 | 0.971 |
| | SPE | 0.990 | 0.989 | 0.945 | 0.946 |
| | Distance | 6.289 | 7.484 | 9.576 | 30.192 |
| GLASSO | SEN | 0.930 | 0.946 | 0.975 | 0.923 |
| | SPE | 0.989 | 0.967 | 0.944 | 0.942 |
| | Distance | 6.821 | 13.67 | 9.641 | 30.517 |

# Simulations: Lag 1 network identification

Table : $n = 50$ and $t = 10$ for various number of genes (p)

| | |  | |  | |
|---|---|---|---|---|---|
| | | $p = 50$ | $p = 100$ | $p = 50$ | $p = 100$ |
| SCAD | SEN | 0.998 | 0.948 | 0.989 | 0.993 |
| | SPE | 0.985 | 0.899 | 0.977 | 0.959 |
| | Distance | 0.602 | 1.665 | 0.452 | 3.254 |
| EBDBN | SEN | 0.343 | 0.195 | 0.394 | 0.226 |
| | SPE | 0.615 | 0.793 | 0.599 | 0.787 |
| | Distance | 37.910 | 52.048 | 59.730 | 80.722 |
| GeneNet | SEN | 0.000 | 0.000 | 0.000 | 0.000 |
| | SPE | 0.969 | 0.971 | 0.997 | 0.999 |
| | Distance | 10.607 | 15.092 | 19.414 | 31.197 |

# T-cell data

**Aim.** Use large time-course experiment to characterize response of human T-cell line (Jurkat) to PMA and ionomycin treatment.

**T-cell time-series dataset**

- Temporal expression of 58 genes for 10 unequally spaced time points.
- At each time point there are 44 separate measurements.
- See Rangel et al. (2004) for more details.

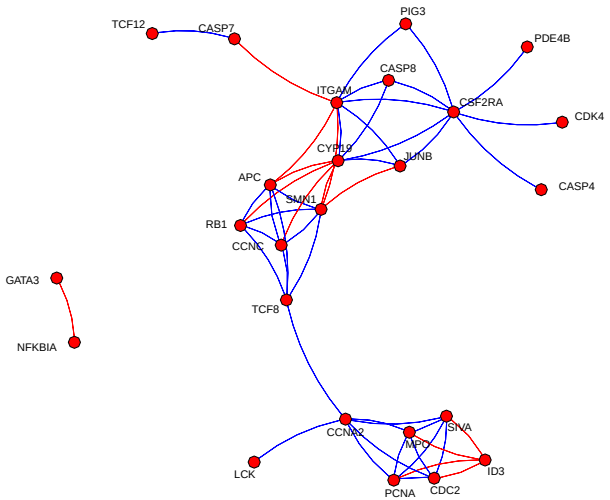
# Application to T-cell data

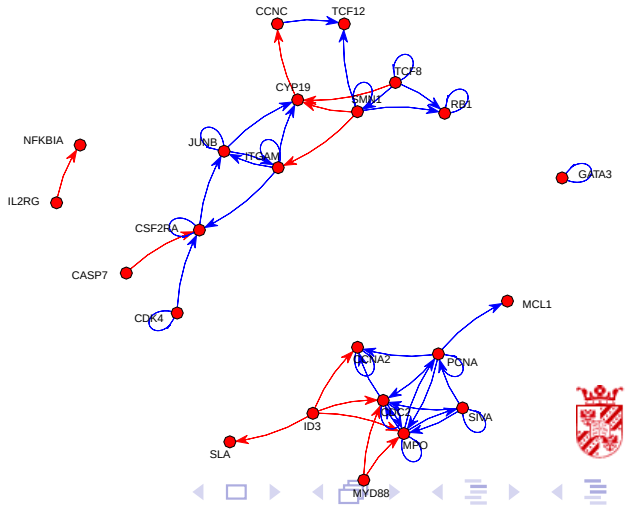
$$S_0 \prec F_1, \mathbf{N_0} \prec \mathbf{F_\Gamma}, S_1 \prec F_T, \mathbf{N_1} \prec \mathbf{F_\Gamma}$$

$$\mathbf{\Theta} = \left[\begin{array}{cc|cc|cc|cc} S_0^1 & N_0^1 & S_1^1 & N_1^1 & 0 & 0 & \dots & \dots \\ & S_0^1 & N_1^1 & S_1^1 & 0 & 0 & \dots & \dots \\ \hline & & S_0^2 & N_0^2 & S_1^2 & N_1^2 & 0 & 0 \\ & & & S_0^2 & N_1^2 & S_1^2 & 0 & 0 \\ \hline & & & & S_0^3 & N_0^3 & S_1^3 & N_1^3 \\ & & & & & S_0^3 & N_1^3 & S_1^3 \end{array}\right]$$

$N_0^1 = N_0^2 = \ldots = N_0^{10}$

$N_1^1 = N_1^2 = \ldots = N_1^{10}$




rijksuniversiteit groningen

# What have we achieved so far? And problems!

**Summary:**

- Penalized Gaussian graphical models
- Coloured graphs
- Natural partitions
- Factorially coloured Gaussian graphical models

**Problems:**

1. Factorial colouring not particularly flexible in modeling time dynamics.
2. Gaussian assumption may be too restrictive for realistic genomic data.

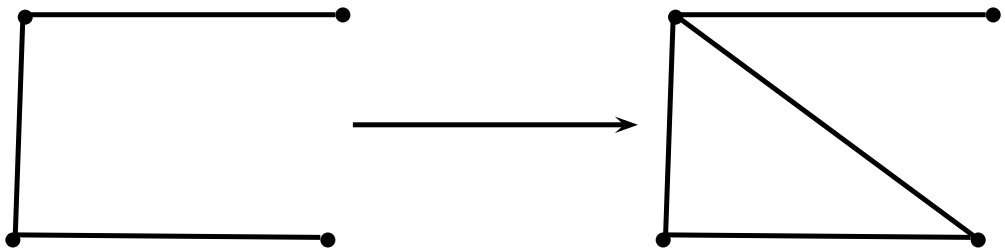
rijksuniversiteit groningen

# 1. Extension: Slowly changing graphical models

- Problem: Estimate changes in the dynamic of the network.

**Main Idea:** Penalize changes between graphs at different time points

$$||\Delta\Theta||_1 = \sum_{s=0}^{t-1}\sum_{k=0}^{t-1} ||N_s^k - N_s^{k+1}||_1.$$




- Solution: Penalized maximum likelihood subject to constraints.

rijksuniversiteit groningen

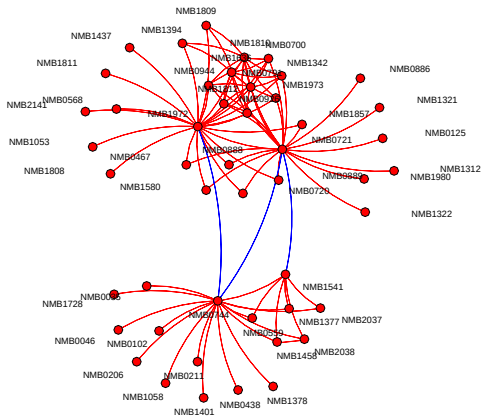
# Application to a time-course dataset

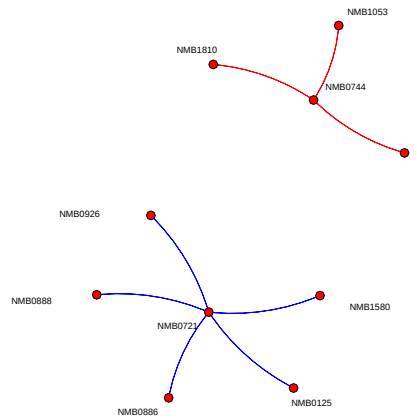
$$\mathbf{S_0} \prec F_{\Gamma T}, \mathbf{N_0} \prec 1, \mathbf{N_1} \prec 1, \mathbf{N_2} \prec 0$$

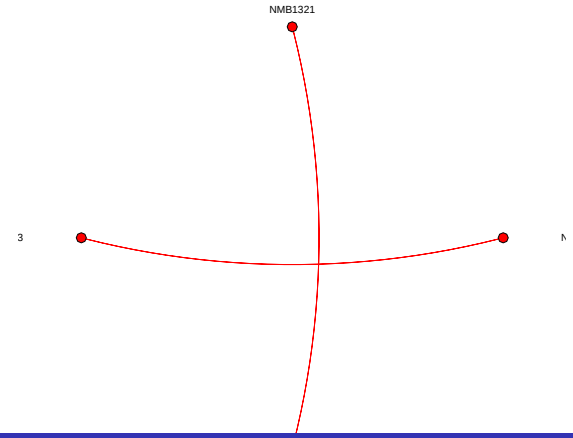
$$\Theta = \left[\begin{array}{cc|cc|cc|cc} S_0^1 & N_0^1 & S_1^1 & N_1^1 & 0 & 0 & \cdots & \cdots \\ & S_0^1 & N_1^1 & S_1^1 & 0 & 0 & \cdots & \cdots \\ \hline & & S_0^2 & N_0^2 & S_1^2 & N_1^2 & 0 & 0 \\ & & & S_0^2 & N_1^2 & S_1^2 & 0 & 0 \\ \hline & & & & S_0^3 & N_0^3 & S_1^3 & N_1^3 \\ & & & & & S_0^3 & N_1^3 & S_1^3 \end{array}\right]$$

$N_0^1$ | $|N_0^1| - |N_0^2|$ | $|N_0^2| - |N_0^3|$

N0 at time 1 | 1 | 7

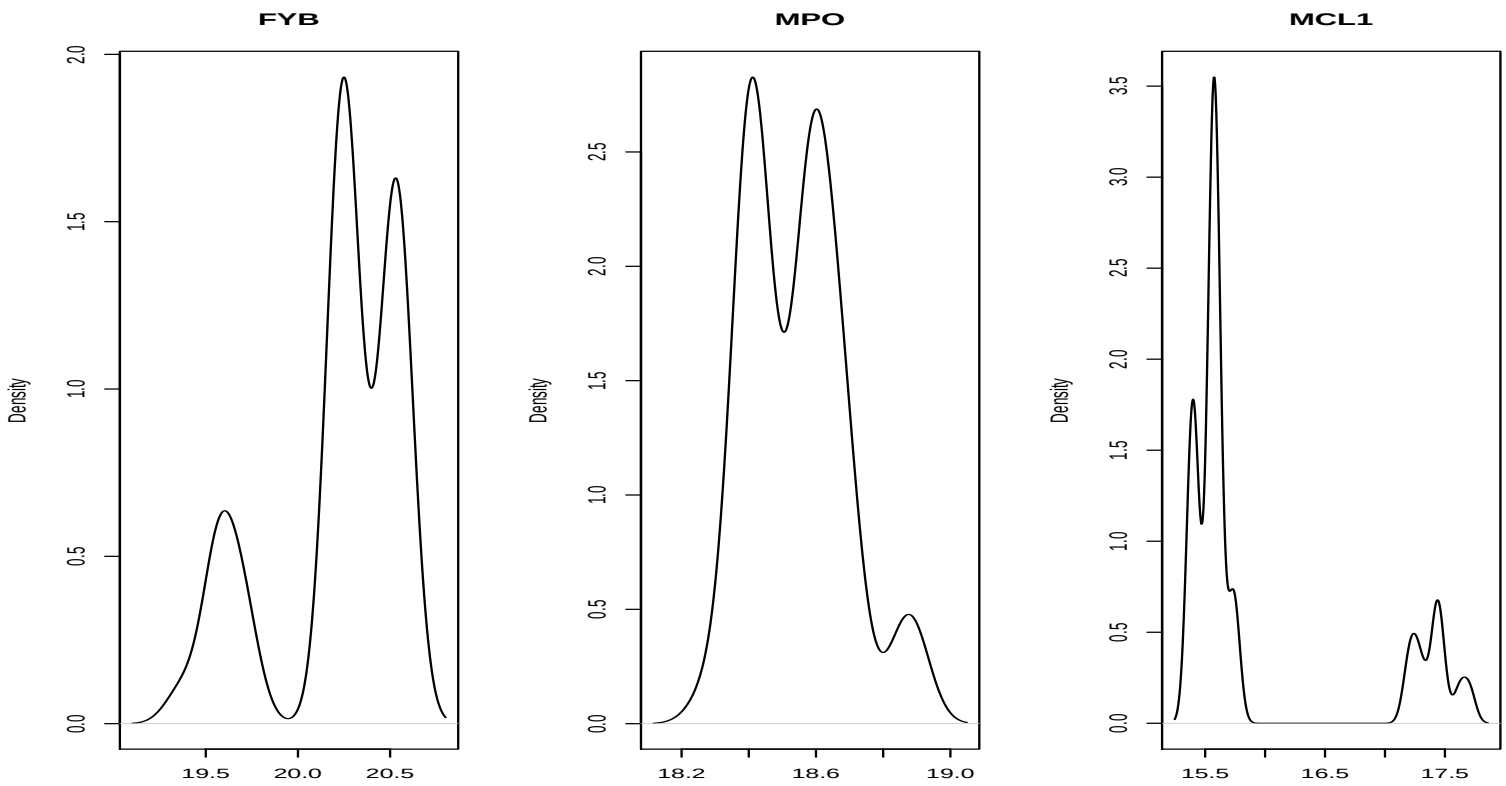


# 2. Extension: Non-Normality

- Problem: Non-Normality of the data (e.g. T-cell).



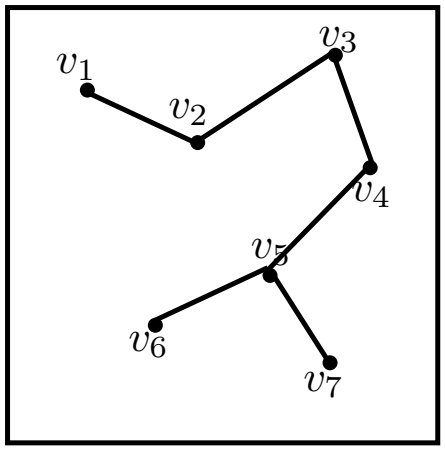

- Solution: Copula Gaussian graphical models

rijksuniversiteit groningen

# Copula Gaussian graphical models

**IDEA:**

- Graph exists on a hidden Gaussian variable $\mathbf{Z} \sim N(0, \Theta)$,
- $\mathbf{Z}$ gives rise to observed non-Gaussian data $\mathbf{Y}$.




Latent Variable: $\mathbf{Z} \sim N(\mu, \mathbf{\Theta}^{-1})$

$$F_1$$
$$F_2$$
$$\vdots$$
$$F_7$$

Marginal Distribution: $\mathbf{F}$

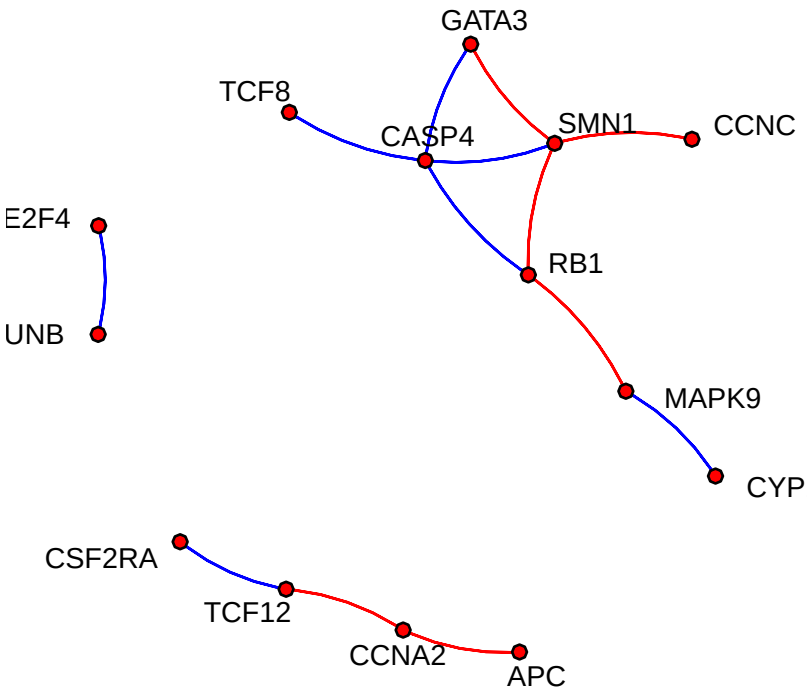
$$Y_1 = F_1^{-1}(\phi(Z_1))$$
$$Y_2 = F_2^{-1}(\phi(Z_2))$$
$$\vdots$$
$$Y_7 = F_p^{-1}(\phi(Z_7))$$

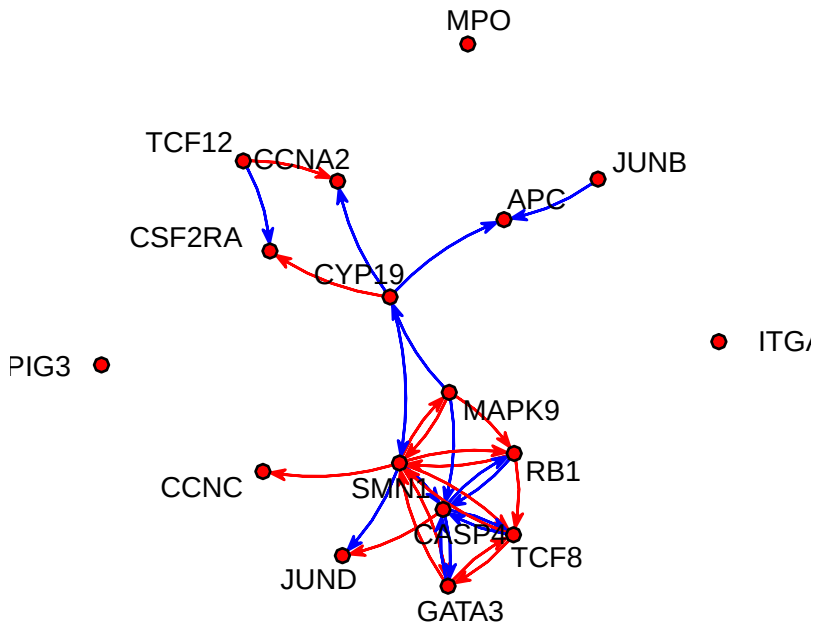
Observed: $\mathbf{Y}$

- We consider the $F_i$ as nuisance parameters.
- For continuous variables: 1-to-1 relationship between $Z$ and $Y$.
  For discrete variables, relationship is more complicated!

# Application of Gaussian copula graphical models to T-cell

$$S_0 \prec F_1, \mathbf{N_0} \prec \mathbf{F_\Gamma}, S_1 \prec F_T, \mathbf{N_1} \prec \mathbf{F_\Gamma}$$




$N_0^1 = N_0^2 = \ldots = N_0^{10}$

$N_1^1 = N_1^2 = \ldots = N_1^{10}$

# Dynamical mammary gland application

Mammary gland gene expression data:

- Microarray experiment
- using mammary tissue from female mice
- across 4 different developmental stages
- for 8,600 genes.
- 3 replicates on each of 18 time points.

30 genes have been identified as activators for developmental stages (Wit and McClure, 2004).

**Objective:**

Study interactions between these crucial mice genes.

# Inference

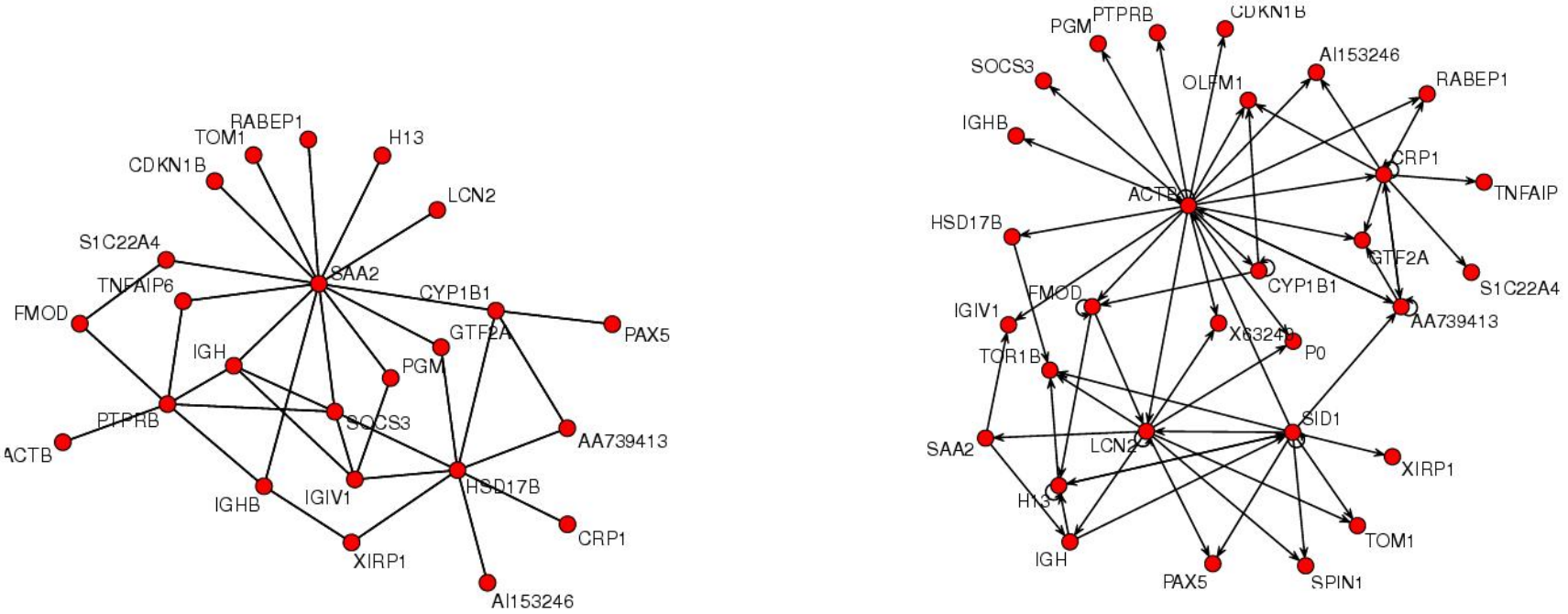



Figure : Undirected $N_0$ (left) and directed $N_1$ (right) time series chain graphical model networks inferred from mammary gland time course gene expression data.

Abegaz and Wit. "Sparse time series chain graphical models for reconstructing genetic networks". Biostatistics. 2013

rijksuniversiteit groningen

# Conclusions

- Graphical models are a convenient formulation of many genomic networks.
- Static boolean networks:
  - Sparse GLM inference via dglars.
  - Software: R package `dglars`.
- Dynamic continuous networks:
  - Chain graphical models infer sparse time dynamics.
  - Software: R package `SparseTSCGM`.