Connections matter

A conversation between statistics and biology

Monica Chiogna

EuroBioC 2020, 16th December

Department of Statistical Sciences, Alma Mater Studiorum, University of Bologna

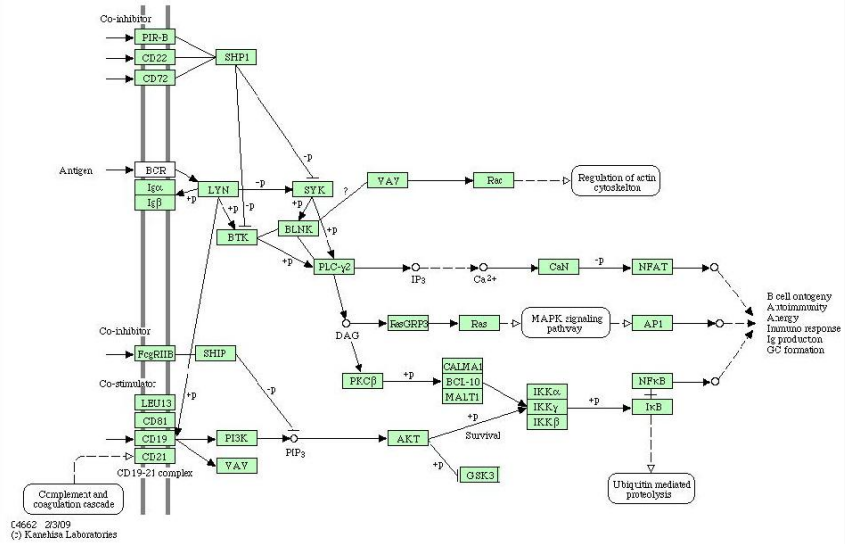
A convergence of interests

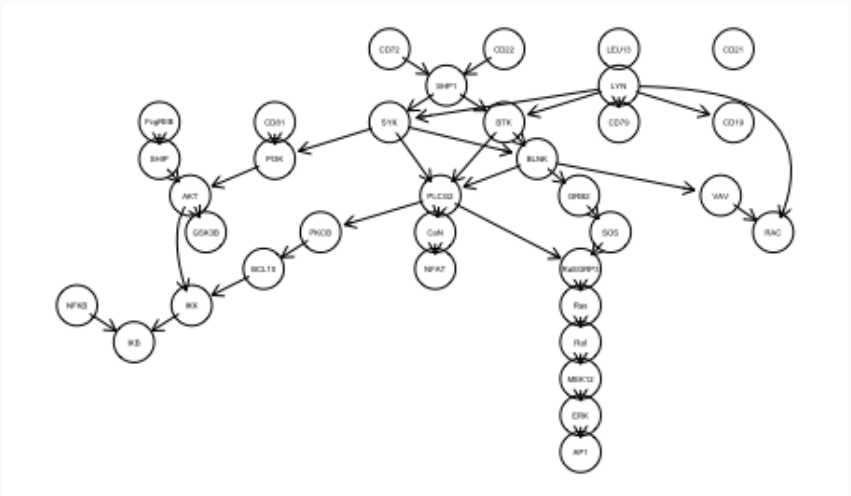
- Biology has become a data intensive science, making urgent the development of innovative and transformative tools contributing to the understanding of complex patterns in high-dimensional contexts.
- Statistics is an inherently collaborative discipline, developing in response to scientific needs.
- To enable mutual exchange of conceptual understanding, the use of a “common language” is essential.

The need for a common language

- A biologist understands gene transcription by identifying specific transcription factors, their binding sites, the role of RNA polymerase and the genes that get activated.
- For a statistician, these crucial facets are extraneous details. What matters are the probability distributions associated with the involved quantities and quantification of the forces involved in the processes.

Graphs: the connecting language





A promising route

As pathways are the best representation of biological experimentally validated knowledge of a specific process;

we make use of the topology of the pathway within the theory of graphical modelling. According to the research goal, we use directed acyclic graphs (DAGs), gaussian graphical models (GGMs), mixed graphs (ongoing).

(Note that a plurality of statistical methods are based on the use of pathways as simple lists of genes, but do not make use of the relations among them).

Outline

- Essentials on graphical models
- Connecting biology to statistics
- Aswering biological questions: doing and seeing
- And beyond

Essentials on graphical models

Conditional independence

Random variables X_1 and X_3 are *conditionally independent* given the random variable X_2, usually written $X_1 \perp\!\!\!\perp X_3 | X_2$, if

$$\mathcal{L}(X_1 | X_2, X_3) = \mathcal{L}(X_1 | X_2)$$

Intuitively: knowing X_2 renders X_3 irrelevant for explaining/predicting X_1.

Conditional independence allows to factorize the joint distribution of the variables into smaller, more tractable, units

$$\begin{aligned} X_1 \perp\!\!\!\perp X_3 | X_2 \longrightarrow p(x_1, x_2, x_3) &= \psi_1(x_1, x_2)\ \psi_2(x_2, x_3) \\ &= p(x_1, x_2)\ p(x_3 | x_2) \\ &= p(x_1 | x_2)\ p(x_2, x_3) \end{aligned}$$

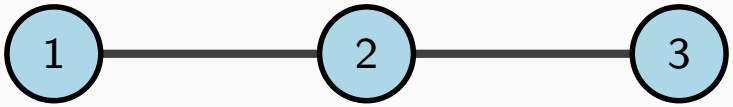
Graphical representation

Each node, v say, is associated to random variable X_v.

Edges represent connections between vertices

- undirected
- directed
- bidirected (not covered here)

Undirected graphs (UGs)



- vertices: $V = \{1, 2, 3\}$
- cliques: $C_1 = \{1, 2\}$ and $C_2 = \{2, 3\}$
- separator: $S = \{2\}$

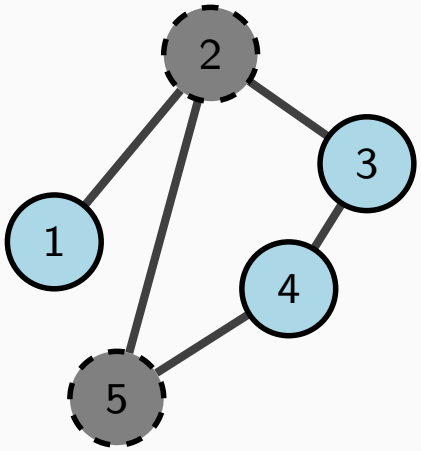
Semantic of the representation

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

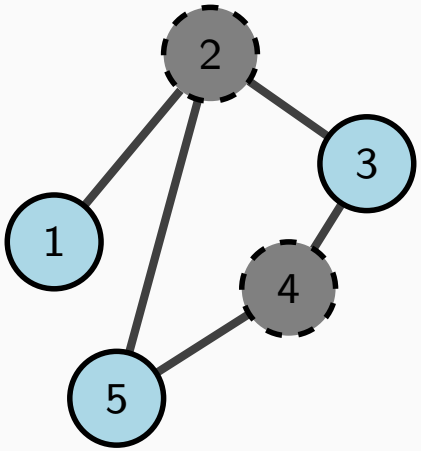
$$p(x_1, x_2, x_3) = p(x_1, x_2)\ p(x_3|x_2) = p(x_1|x_2)\ p(x_2, x_3)$$

$$p(x_V) = p(x_{C_1})p(x_{C_2}|x_S) = p(x_{C_2})p(x_{C_1}|x_S)$$

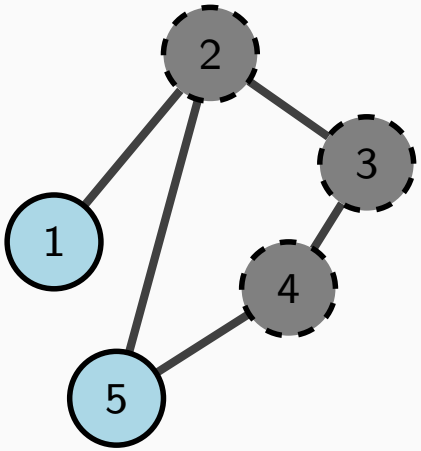
Global



Local

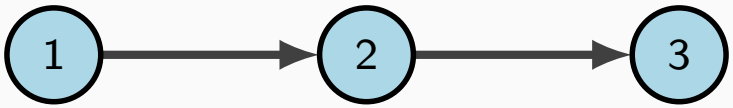


Pairwise



$$(G) \Longrightarrow (L) \Longrightarrow (P)$$

Directed acyclic graphs (DAGs)



Any node, v say, is conditional independent of its non-descendants given its immediate parents, $pa(v)$ say. So, for example, $\{3\} \perp\!\!\!\perp \{1\}|\{2\}$

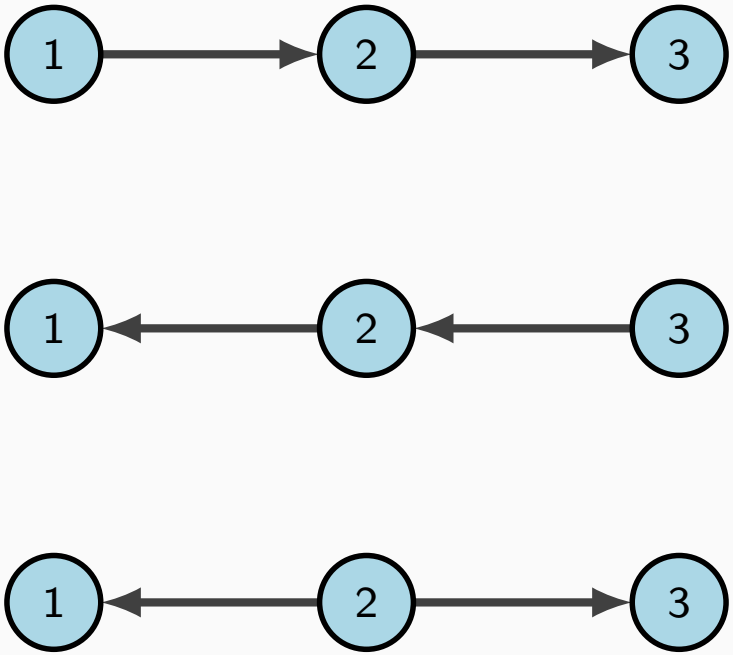
Here, the factorization involves the "families" $(x_v; x_{pa(v)})$

$$p(x_V) = \prod_{v \in V} p(x_v | x_{pa(v)})$$

The meaning of arrows

- Directed arrows between variables embody the idea of a "non-symmetrical relationship" between nodes.
- This is a pure artifact.
- Indeed, many DAGs can represent the same conditional independence property

Equivalent DAGs



The three DAGs all represent the same single property $X_1 \perp\!\!\!\perp X_3 | X_2$, and are all equally valid for this purpose.

Causal DAGs

- A totally different interpretation of this graph is in terms of a *causal* DAG
- Here, we say that X_2 is (in some sense) a "common cause" of both X_1 and X_3, which are otherwise causally unrelated.
- Under this causal interpretation, DAGs in previous slide are no longer interchangeable.

Linking statistics to biology

Preliminaries: convert pathways into graphical models

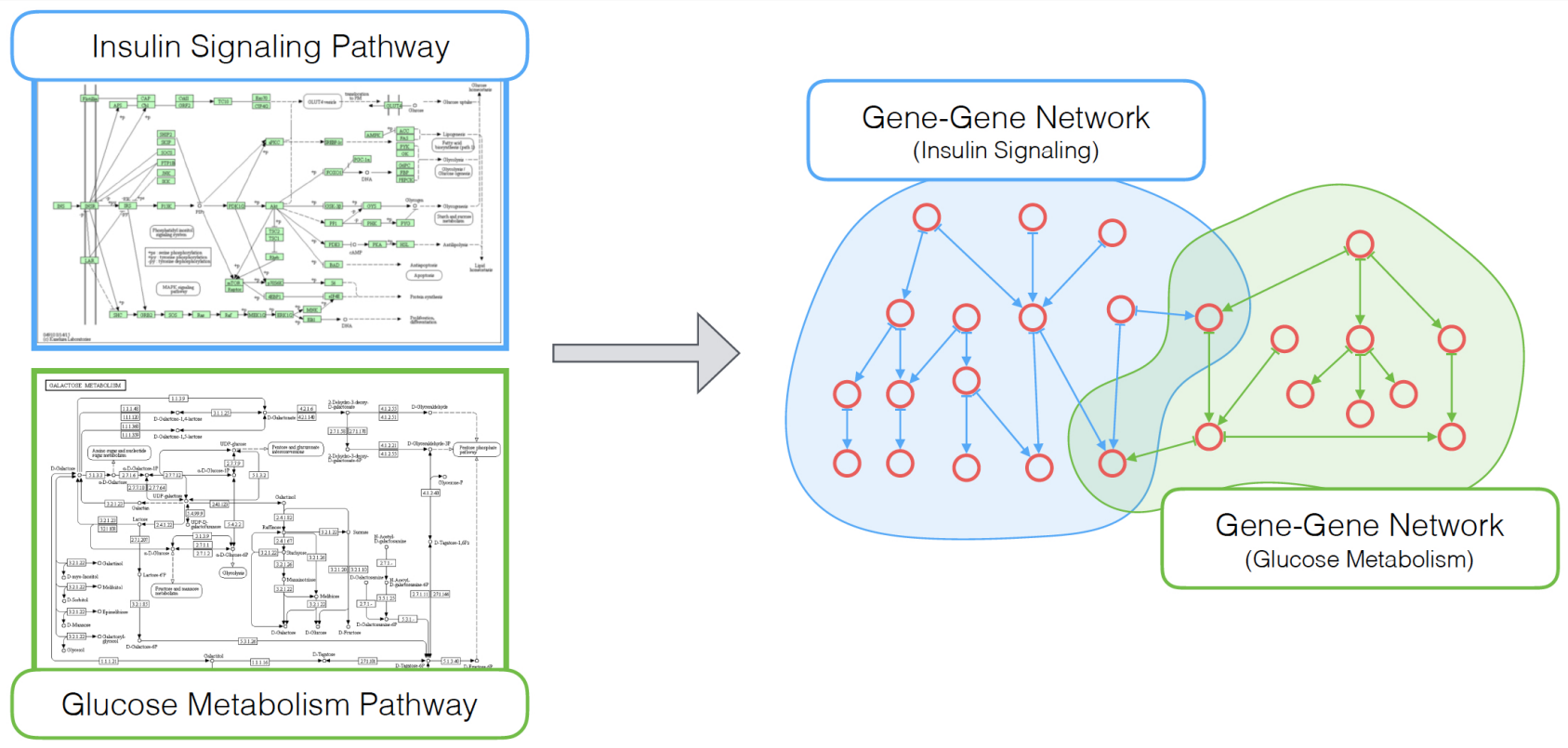


Figure 1: From biological knowledge to statistical models

Pathways conversion

- Not trivial, as annotation posits all sorts of complications, e.g., different types of relations, loops, compounds, ... (Djordjilovic et al., 2015)
- However, it is always possible to convert pathways into DAGs.
- This preliminary conversion can be conveniently performed by the `R` package `graphite` (Sales et al., 2012, 2018)

Doing and Seeing

Answers to most biological questions can be associated to the results two kinds of activities on a graphical model, i.e. *Doing* and *Seeing*

- *Doing* involves applying a disturbance to a system, typically an external intervention
- *Seeing* involves passive observation of a system in its natural state.

(Spirtes et al, 2000; Pearl, 2009, Dawid, 2009, among others)

Doing

Gene silencing

- Employed for studying gene function, and for the development of therapeutics for diseases such as cancer, infectious diseases and neurodegenerative disorders
- Experiments are expensive and time consuming and the design of silencing experiments, optimal with respect to specific targeting, might require complex adaptive procedures

What we can do

- Simulating potential effects of silencing without physically performing the experiment (Djordjilovic et al., 2020)

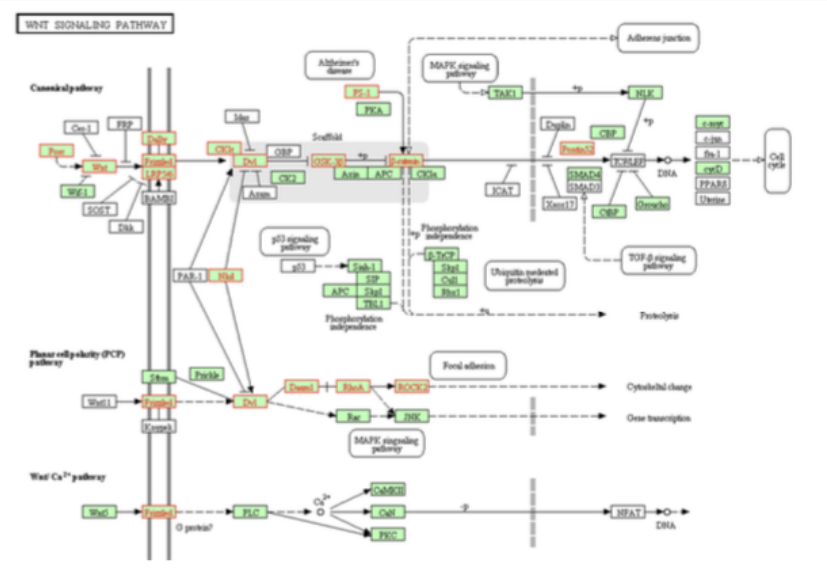
Pipeline

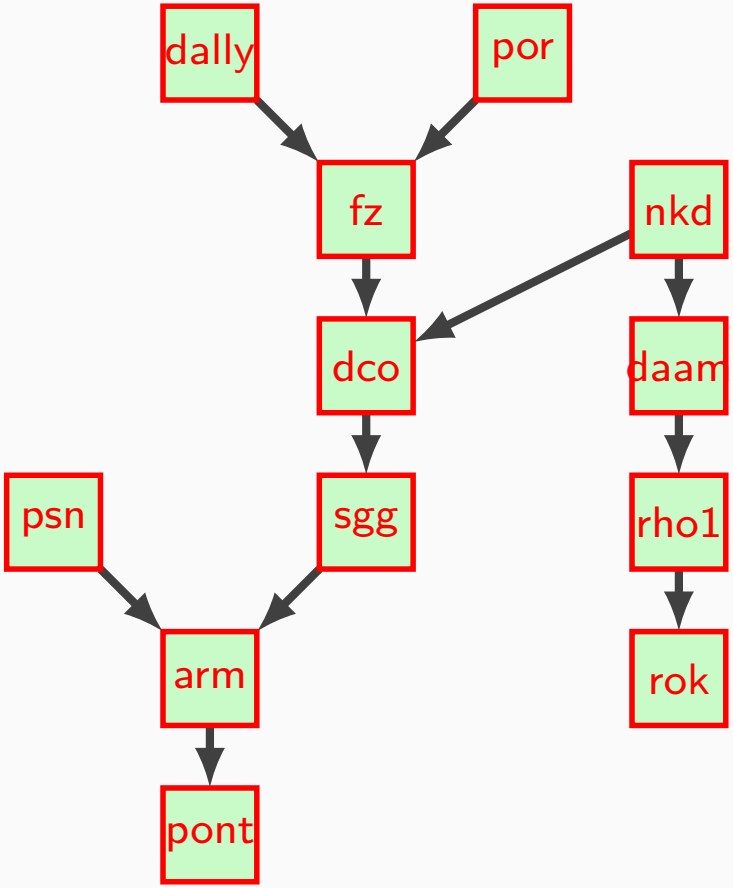
Step 1. Retrieval of DAG and estimation of the statistical model

Step 2. Statistical silencing, aka, intervention analysis

Step 1: retrieval of DAG

- Available data: *Drosophila melanogaster*, 12 genes participating in the WNT pathway. Interest is silencing the naked cuticle gene (nkd)

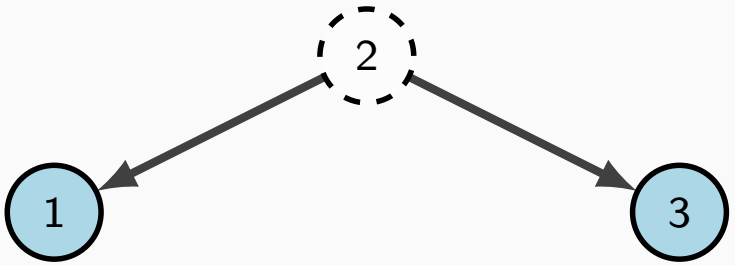




Refinement

- DAGs obtained by simple pathway conversion do not always capture *observed* statistically significant associations.
- For example, latent (not observed) factors might induce associations not depicted in the pathway.

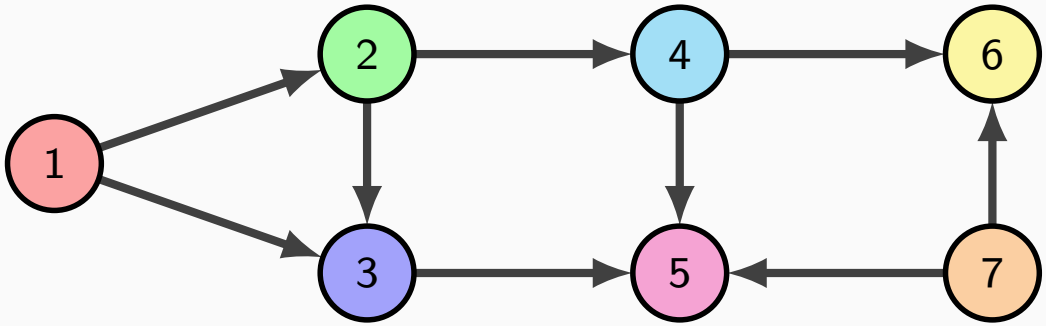
True relation



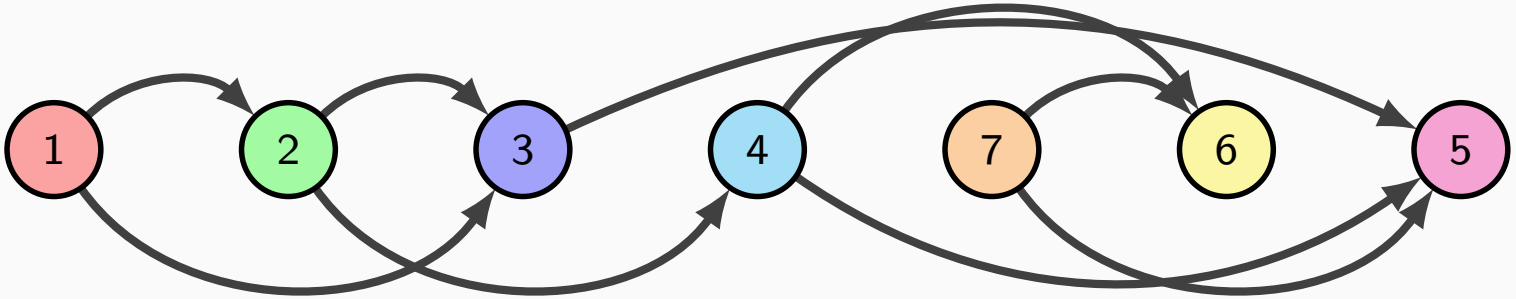
Observed association

Step 1: guided structure learning

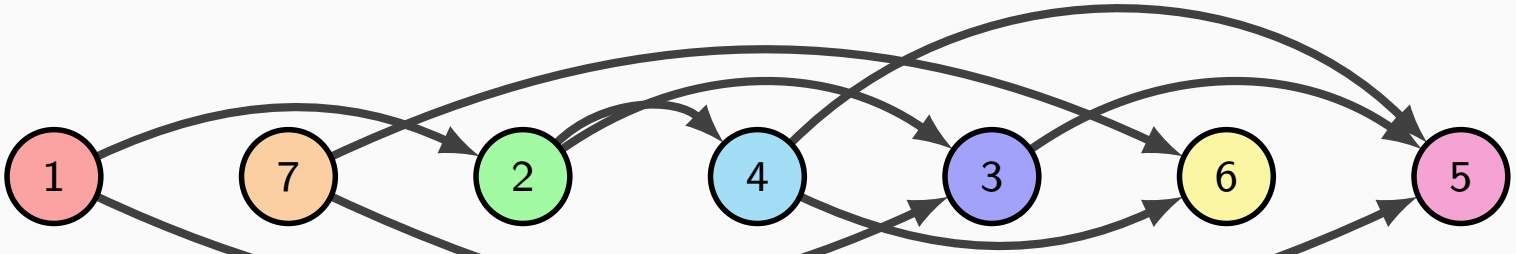
- We estimate the graph, guided by a topological ordering of the variables retrieved from the pathway DAG (Djordjilovic et al., 2017).



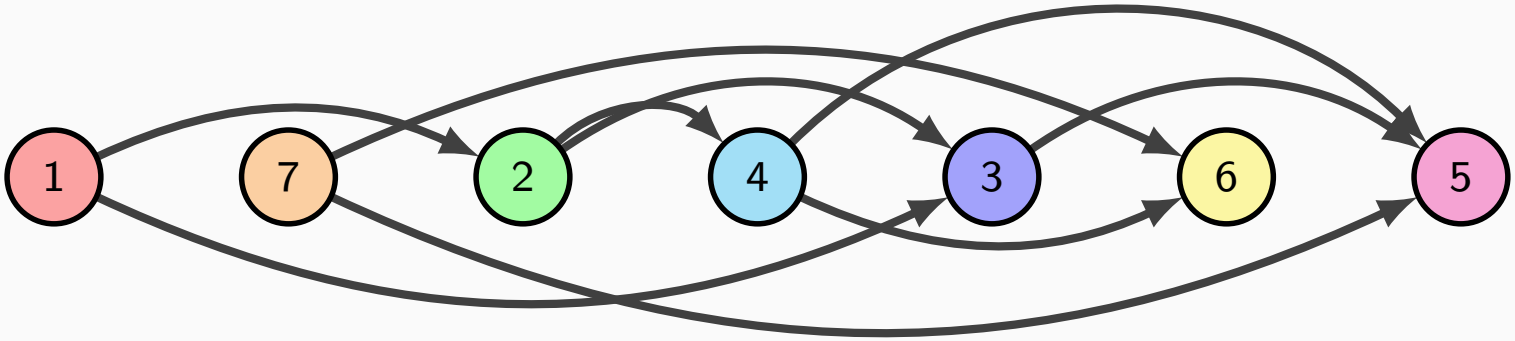
Sort 1



Sort 2



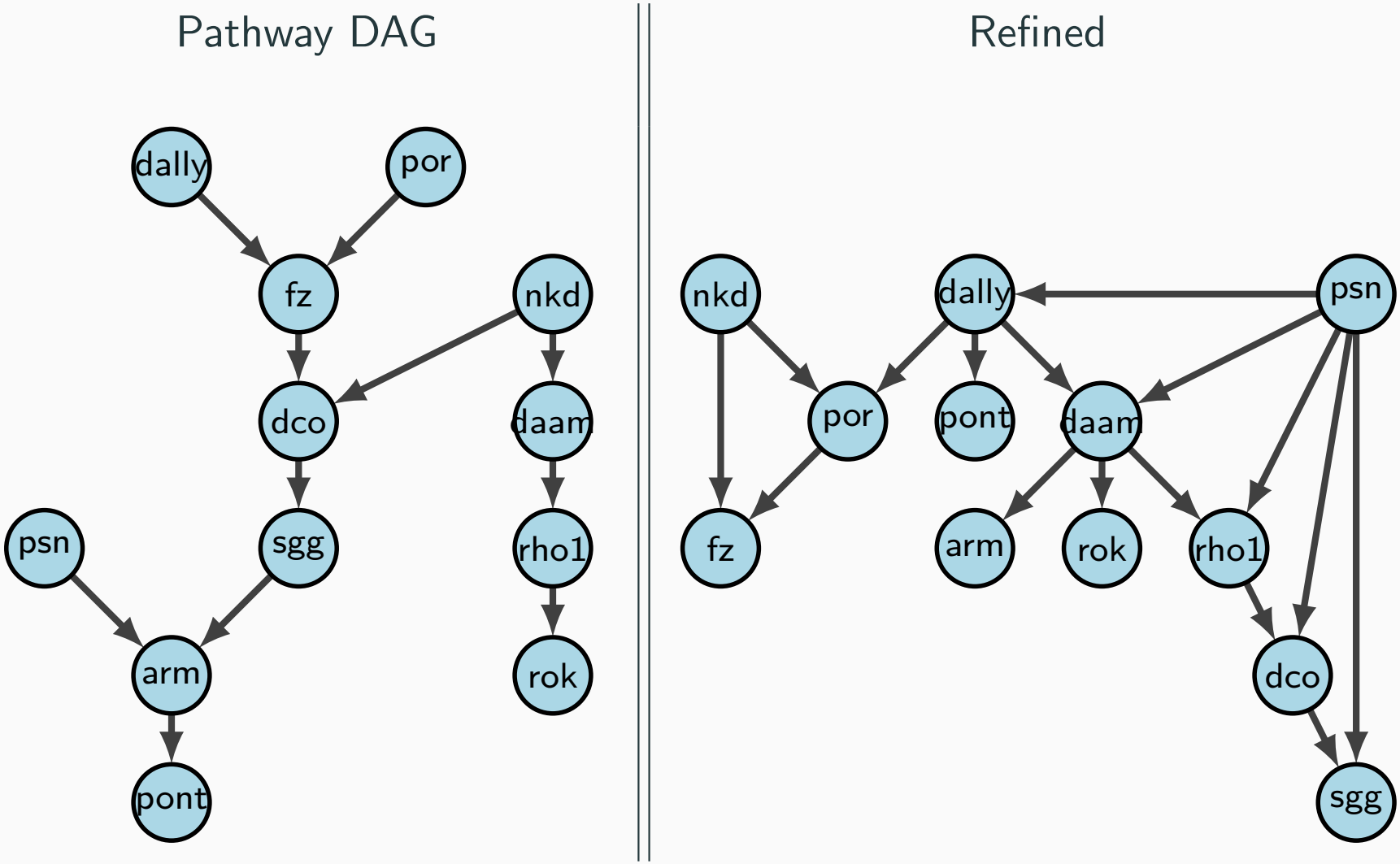
Step 1: guided structure learning with K2-type algorithms



K2 adds incrementally parents to nodes by maximizing a criterion.

Node	Potential parents
1	$\emptyset$
7	1
2	1, 7
4	1, 7, 2
...	...

Example: *Drosophila melanogaster*



Step 2: silencing

- Silencing effect of v on u

$$\delta_u = \mathsf{E}(X_u \mid\mid X_v = \alpha + 1) - \mathsf{E}(X_u \mid\mid X_v = \alpha)$$

- Coupled with a bootstrap strategy to account for uncertainty related to estimation of both the graphical structure and the causal effect.
- Coupled with shrinkage to tackle low sample sizes.

Biological validation

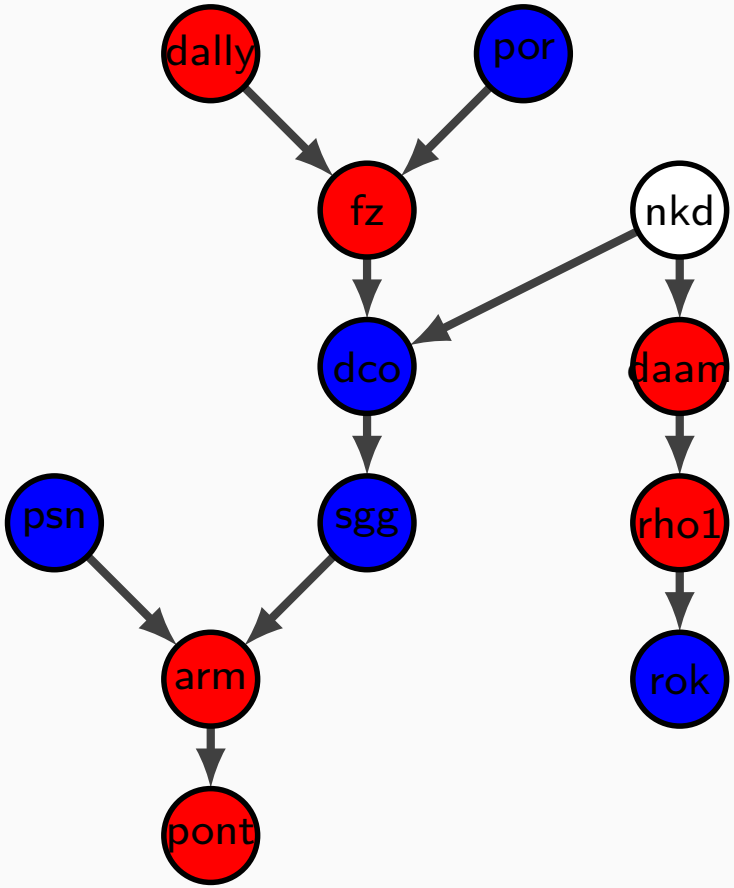
Experiment
- Expression of nkd is artificially inhibited

Data
- 12 measured gene expression levels
- 1 biological pathway, WNT (KEGG database)
- 14 knock-down and 14 control samples

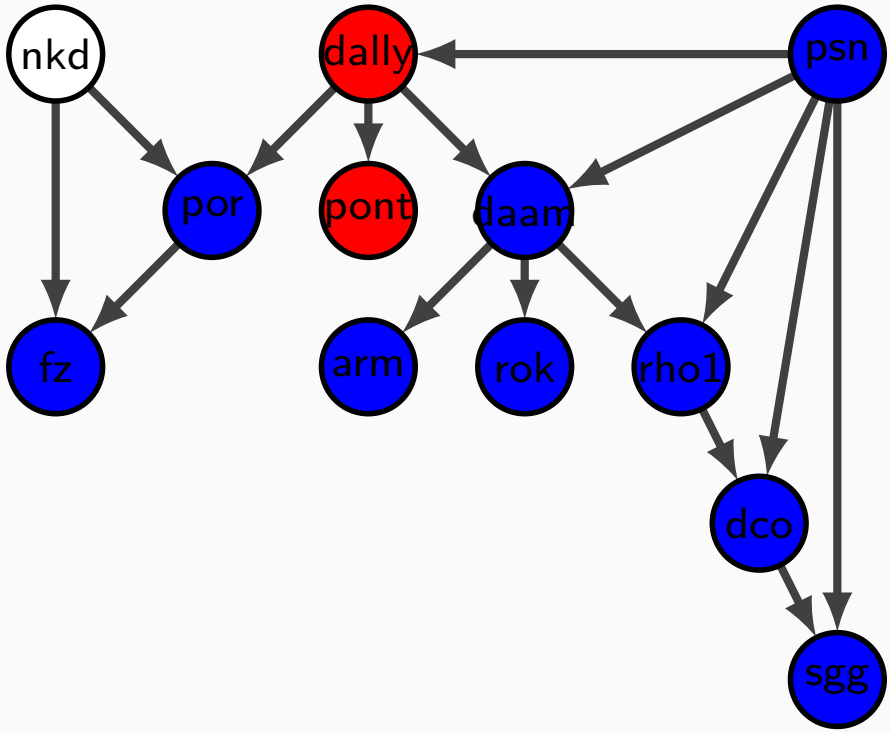
Expected results Predictions obtained via intervention calculus expected to be coherent with observed mean values after silencing

Success of predictions

● : success ● : failure



Pathway DAG



Refined

Seeing

Searching for the source of difference

- In biological networks, diseases can be modelled as perturbations that affect certain targets, which, once perturbed, propagate the perturbation through network connections.
- In practice, we often collect and compare observations from healthy individuals and observations from patients after the disease related perturbation has already taken place.
- On the basis of this comparison, it is of interest to identify the site of original perturbation, i.e., the *source of difference*, and distinguish it from the elements of the network that were affected through the process of network propagation.

Definition of the source set (Salviato et al., 2019)

- (1): healthy; (2) diseased
- *Source set*: the smallest set $D \subseteq V$ such that
  1. $\mathcal{L}^{(1)}(X_D) \neq \mathcal{L}^{(2)}(X_D)$
  2. $\mathcal{L}^{(1)}(X_{V \setminus D} | X_D) = \mathcal{L}^{(2)}(X_{V \setminus D} | X_D)$
- Intuitively: D can be seen as the minimal subset of variables explaining the difference between the two conditions. Variables outside of D are either irrelevant or redundant.

What we can do

- Estimate $D_G \supseteq D$.
- For more details, please follow Elisa Salviato talk: *SourceSet: a graphical model approach to identify primary genes in perturbed biological pathways*

Pipeline

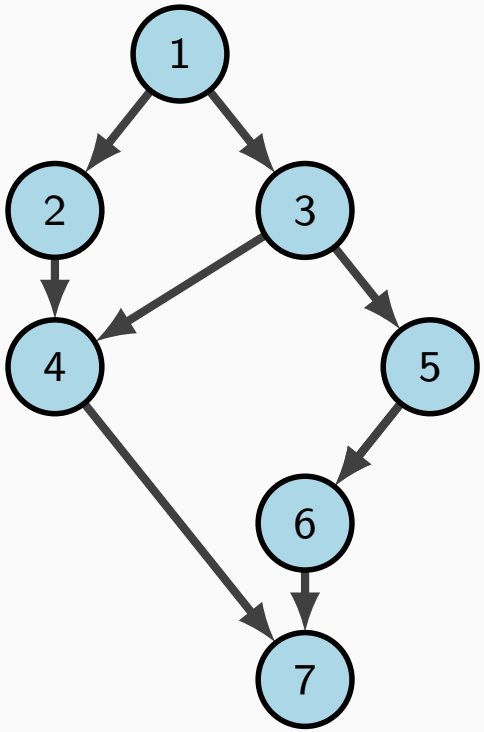
Step 1. Retrieval of DAG and conversion to UG

Step 2. Decomposition of the statistical model and estimation of the source set

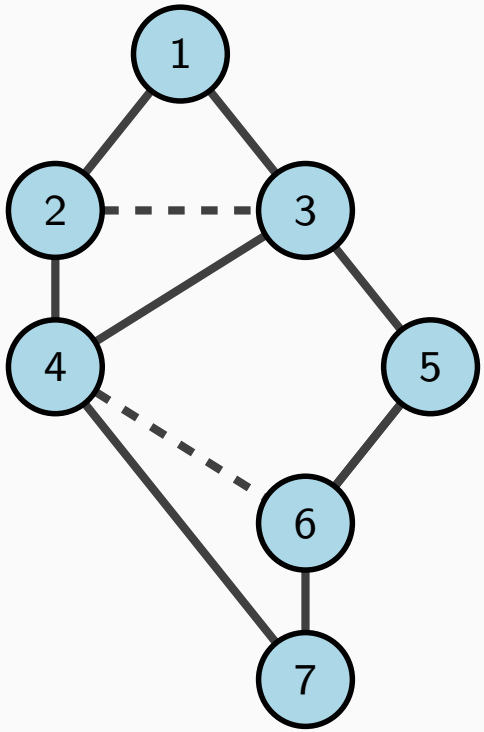
Step 1: retrieval of DAG and conversion to UG

An undirected perspective is in this context more convenient:

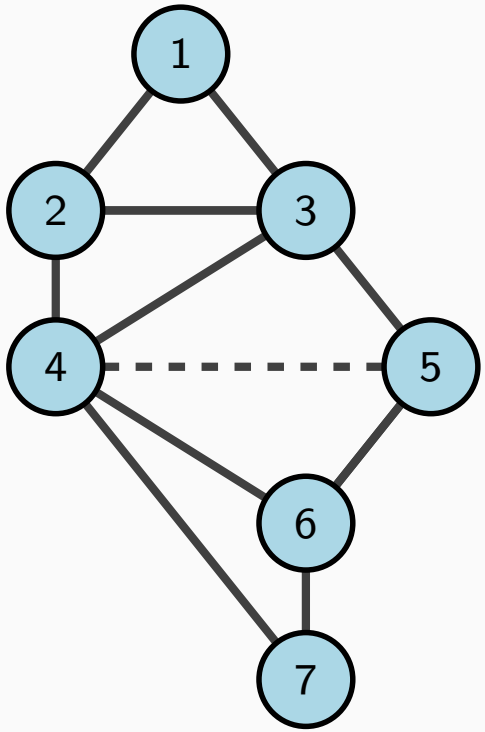
- treats all variables on equal footing
- it is not influenced by loops, feedbacks, etc. . .
- technically, it "enlarges" the model



DAG

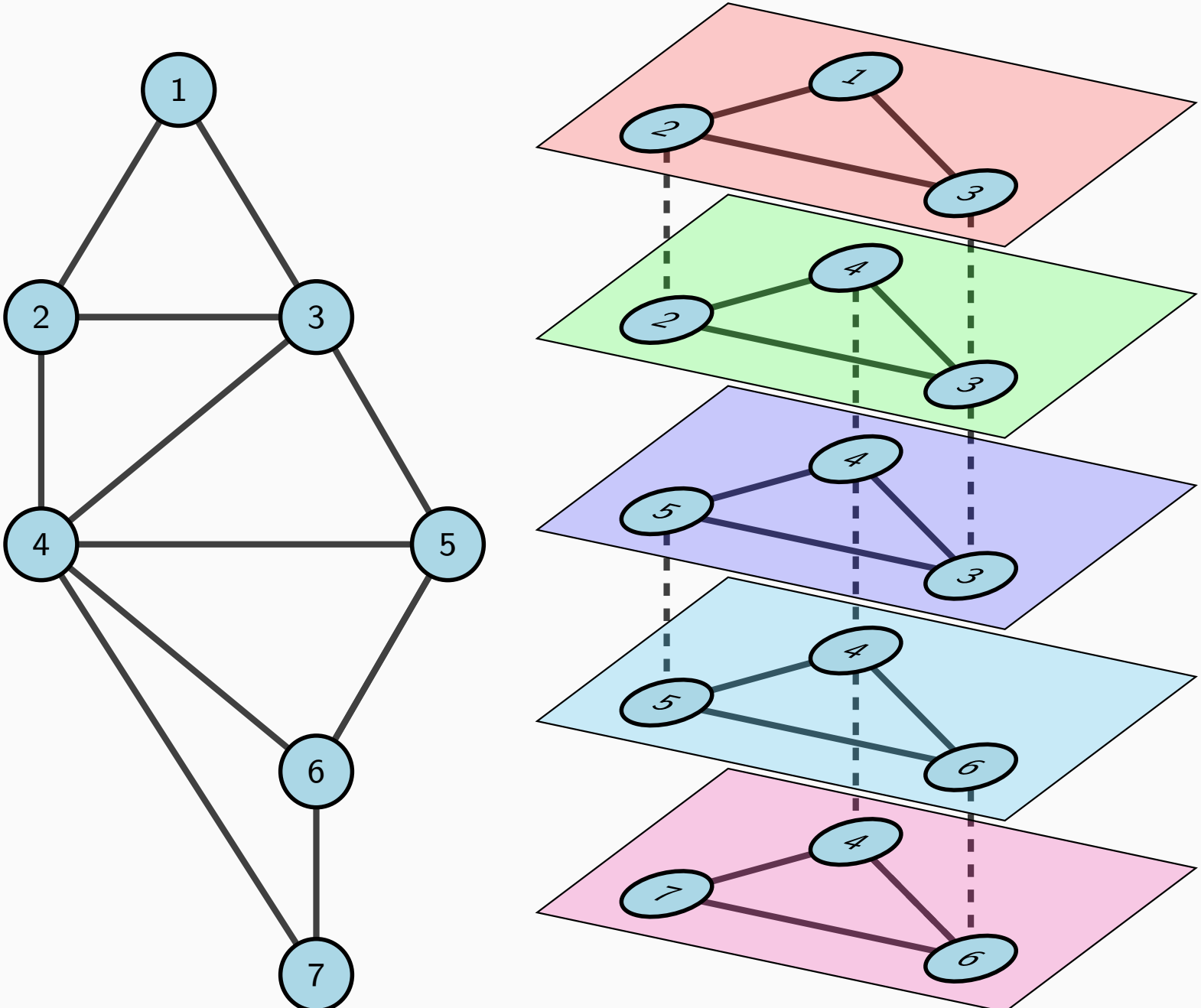


Moralization

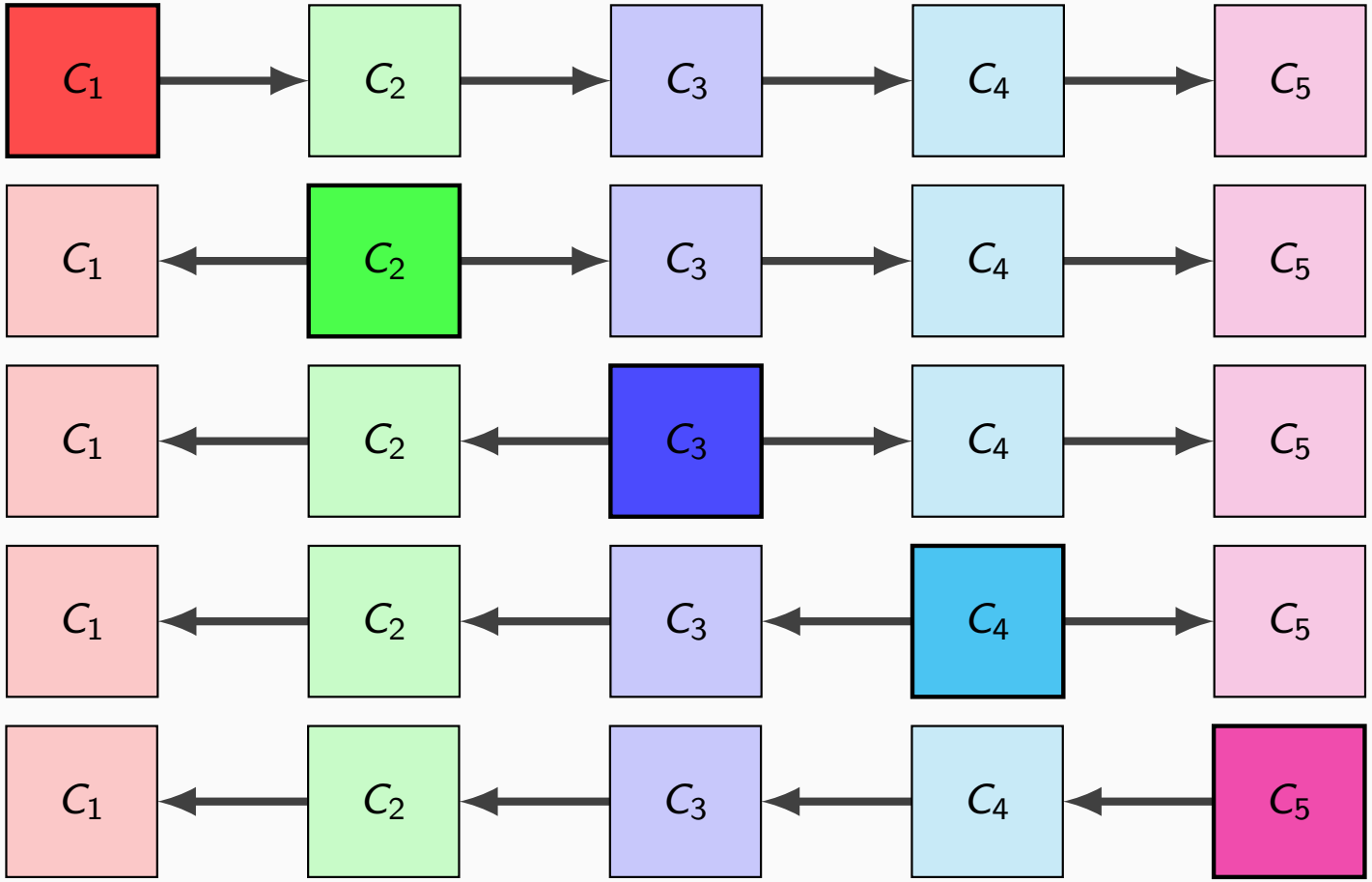


Triangulation

Step 2: Decomposition of model



Step 2: Multiple factorizations



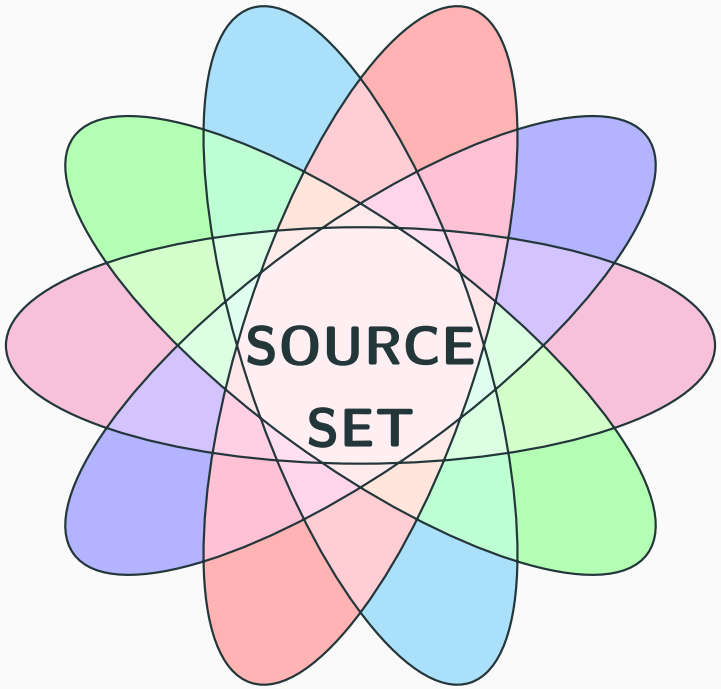
$$p(x_V) = \prod_{\{C\}} p(x_C \mid x_{pa(C)})$$

Step 2: source set

- For each factorization:
  - test for each clique C

$$H_0 : \mathcal{L}^{(1)}(X_C \mid X_{pa(C)}) = \mathcal{L}^{(2)}(X_C \mid X_{pa(C)})$$

  - collect all cliques for which H_0 is rejected
- The source set is the intersection of such sets



Biological validation: STAT3 silencing

Experiment

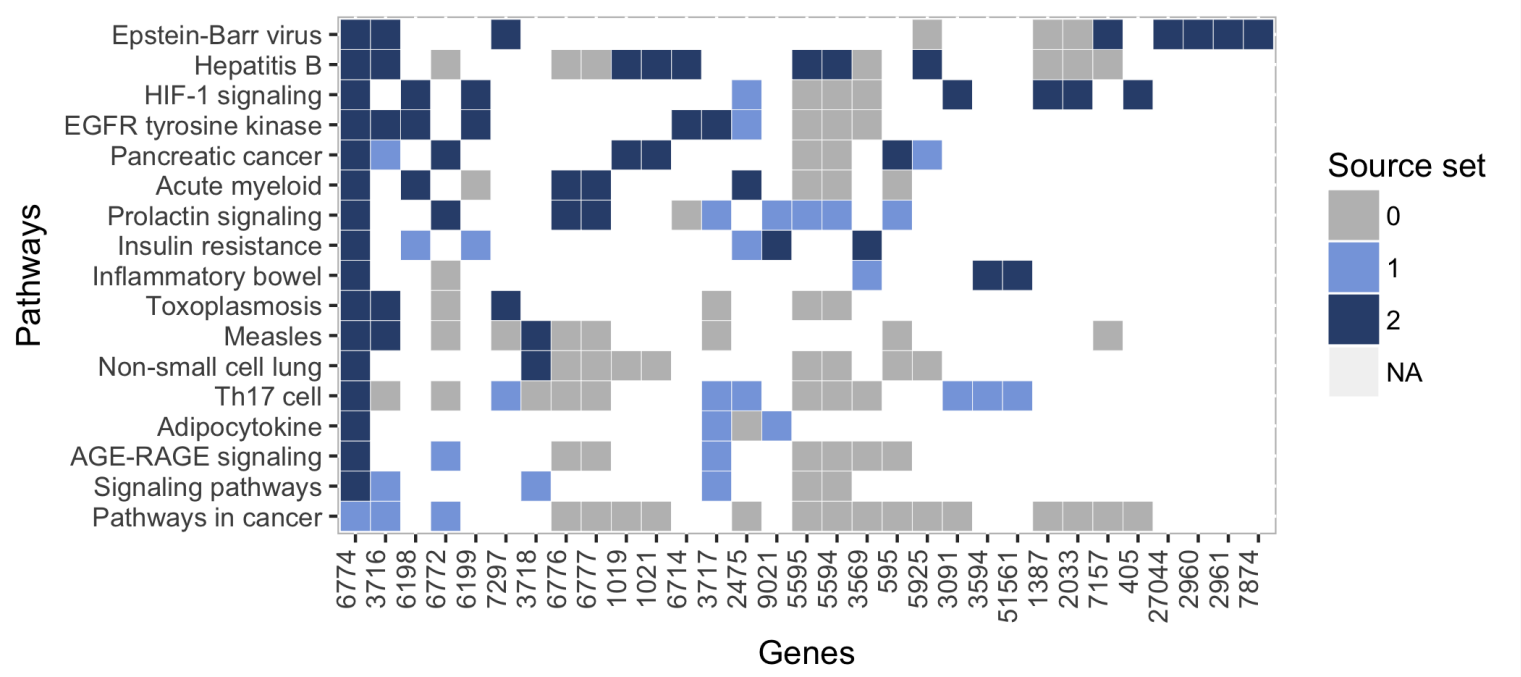
- High-Grade Glioma
- Expression of STAT3 is artificially inhibited
- The exact source of perturbation is known

Data

- 11 knock-down and 11 control samples
- 19 292 measured gene expression levels
- 17 biological pathways (KEGG database) contain STAT3 gene

Expected results STAT3 gene is expected to be included in the source set

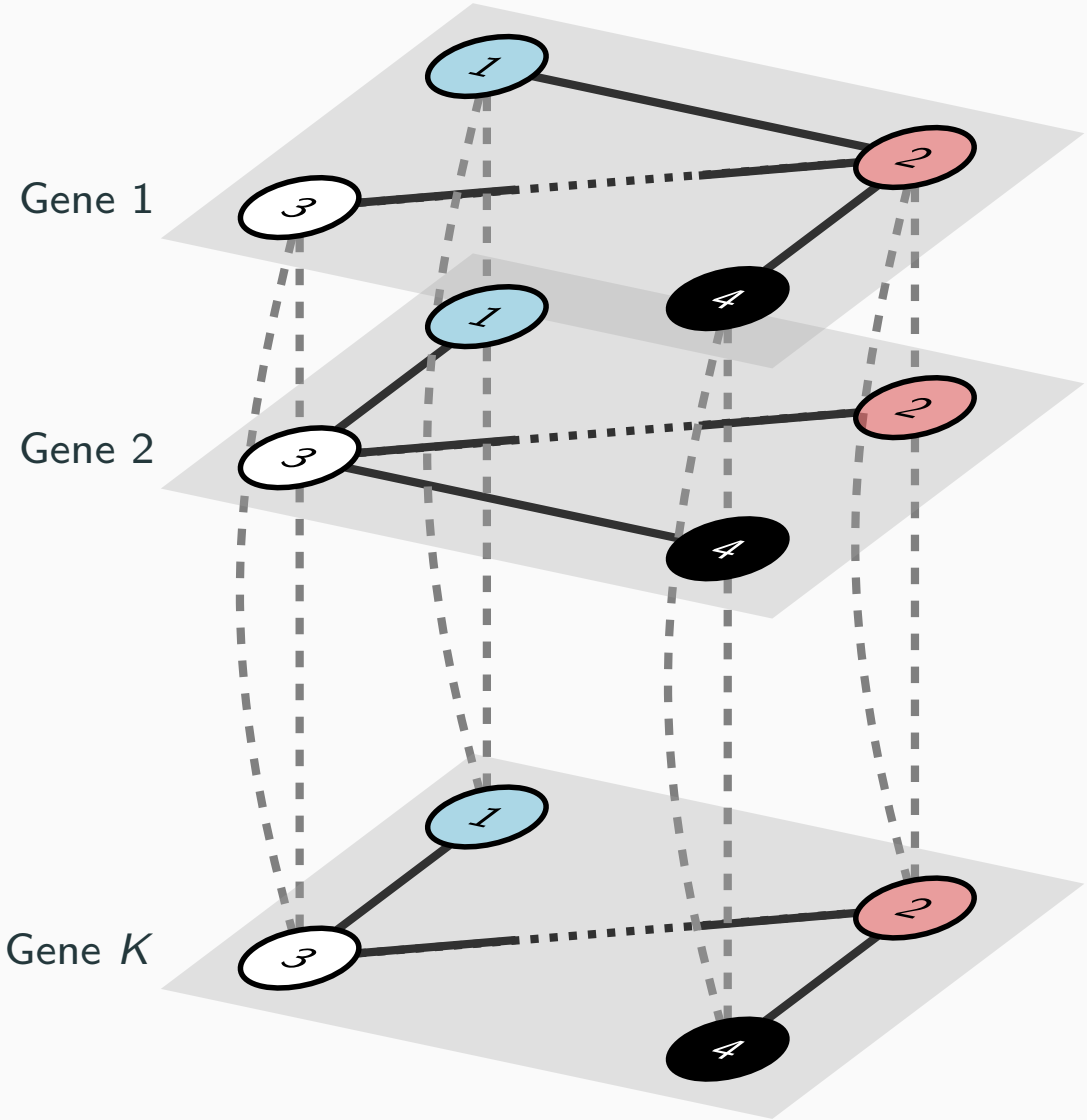
STAT3 results



- STAT3 is in the source set of 16 out of 17 pathways;
- In 4 out of 16 pathways, it is the only element of the source set.

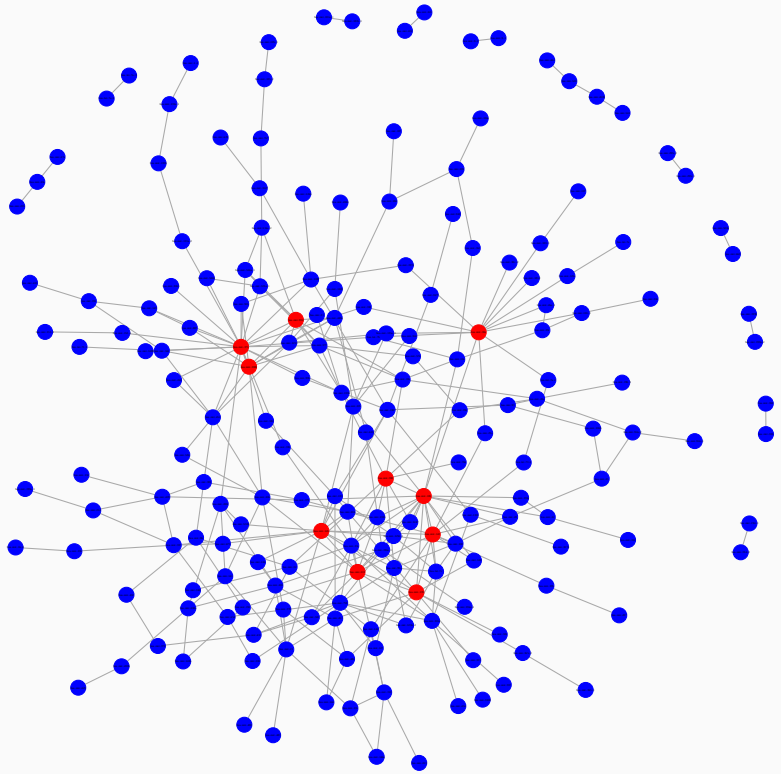
And beyond

Heterogeneous graphical models



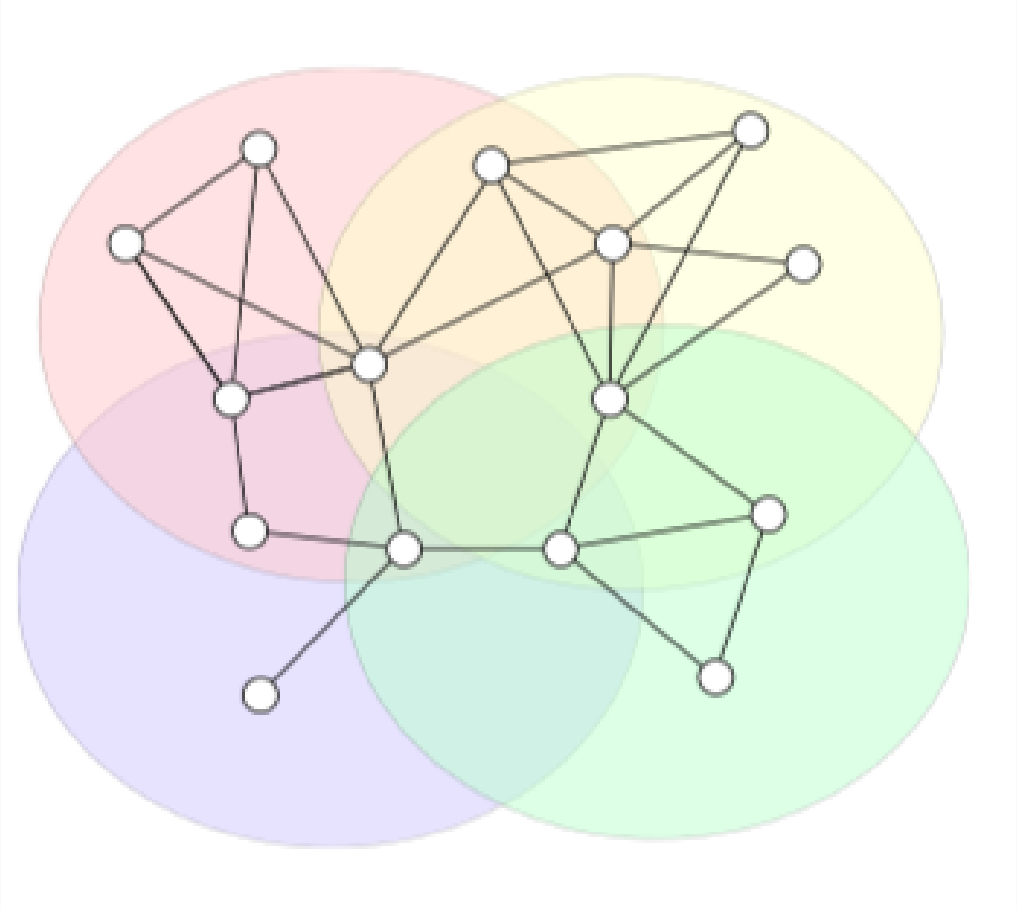
Bussoli, I. (2020). Heterogeneous Graphical Models with Applications to Omics Data. PhD Thesis. University of Padova, Italy

Structure learning



Hue Nguyen, K. and Chiogna, M. (2020). Structure learning of undirected graphical models for count data. Submitted

Graphical meta analysis



Massa, M.S. and Chiogna, M. (2013). Effectiveness of combinations of Gaussian graphical models for model building. *Journal of Statistical Computation and Simulation*

Joint work with

Enrica Calura, Chiara Romualdi, Gabriele Sales

Department of Biology, University of Padova, Italy

Davide Risso, Nguyen Thi Kim Hue, Ilaria Bussoli

Department of Statistical Sciences, University of Padova, Italy

Vera Djordjilovic

Department of Economics, Ca' Foscari University, Italy

Paolo Martini

Department of Biomedicine, University of Brescia, Italy

Sofia Massa

Nuffield Department of Population Health, University of Oxford, UK

Elisa Salviato

IFOM, the FIRC Institute for Molecular Oncology, Milan, Italy

Koen Van den Berge

UC Berkeley, California and Ghent University, Belgium

Software

GitHub

topologyGSA
SourceSet
simPATHy

graphite
clipper

learn2count
Mosclip

Some references

- Sales G, Calura E, Cavalieri D, Romualdi C (2012). graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*.
- Hue Nguyen, T.K., van den Berge, K., Chiogna, M., Risso, D. (2020) Structure learning for zero-inflated counts, with an application to single-cell RNA sequencing data. *Submitted*.
- Djordjilović, V., Chiogna, M., Romualdi, C. (2020) Simulating gene silencing through intervention analysis, *Journal of the Royal Statistical Society Series C - Applied Statistics*.
- Salviato, E., Djordjilovic, V., Chiogna, M., Romualdi, C. (2019) SourceSet: A graphical model approach to identify primary genes in perturbed biological pathways. *PLOS Computational Biology*.
- Martini,P., Chiogna, M., Calura, E., Romualdi, C. (2019) MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic Acids Research*.
- Djordjilovic, V., Chiogna, M., Vomlel, J. (2017) An empirical comparison of popular structure learning algorithms with a view to gene network inference, *International Journal of Approximate Reasoning*.
- Djordjilovic, V., Chiogna, M., Massa, M.S., Romualdi, C. (2015) Graphical modeling for gene set analysis: A critical appraisal. *Biometrical Journal*.

Thanks for connecting!