Network Inference and Properties

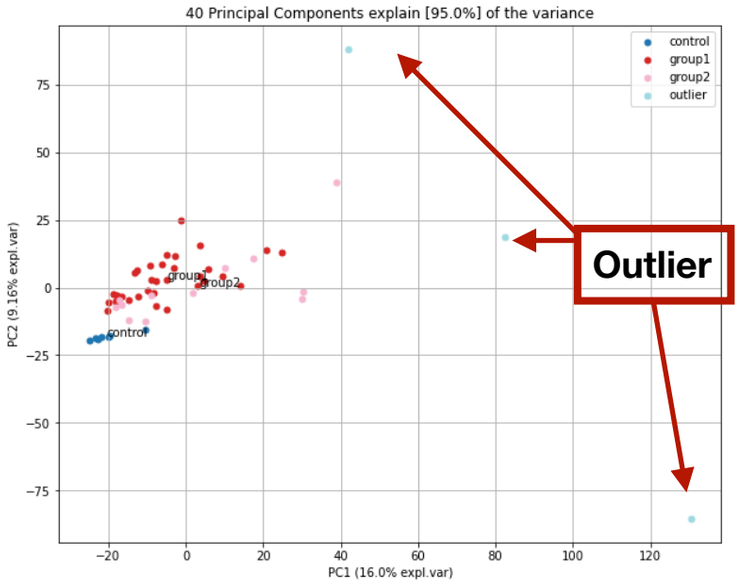
Rui Benfeitas
NBIS - National Bioinformatics Infrastructure Sweden
Science for Life Laboratory, Stockholm
Stockholm University

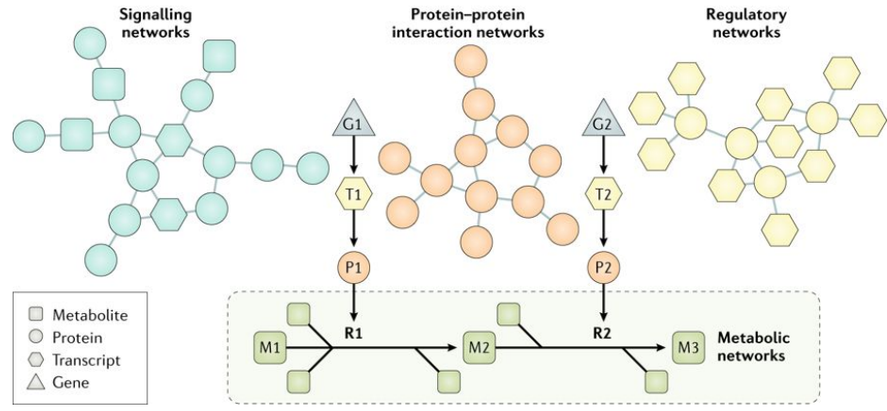
rui.benfeitas@scilifelab.se

Overview

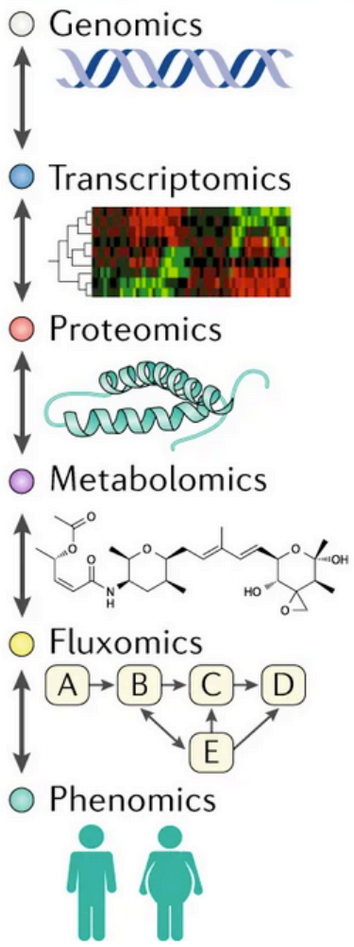
1. Introduction to network analysis
2. Terminology
3. **Network inference**
4. Key network properties
5. Community analysis

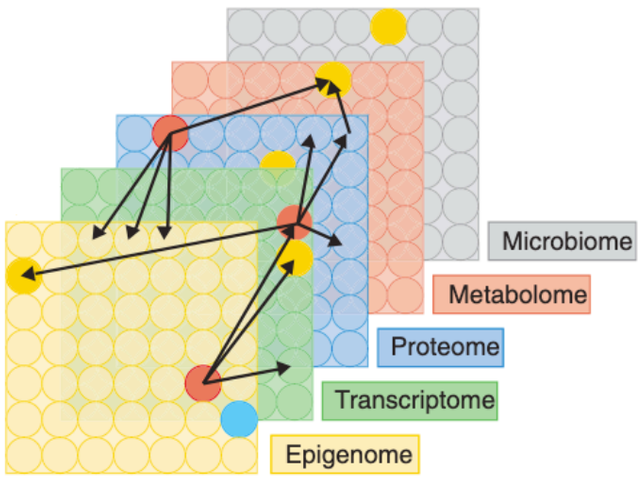
Building networks





Raw → Pre-processing → Distance calculation → Graph analysis





Hasin 2017

Piening 2018

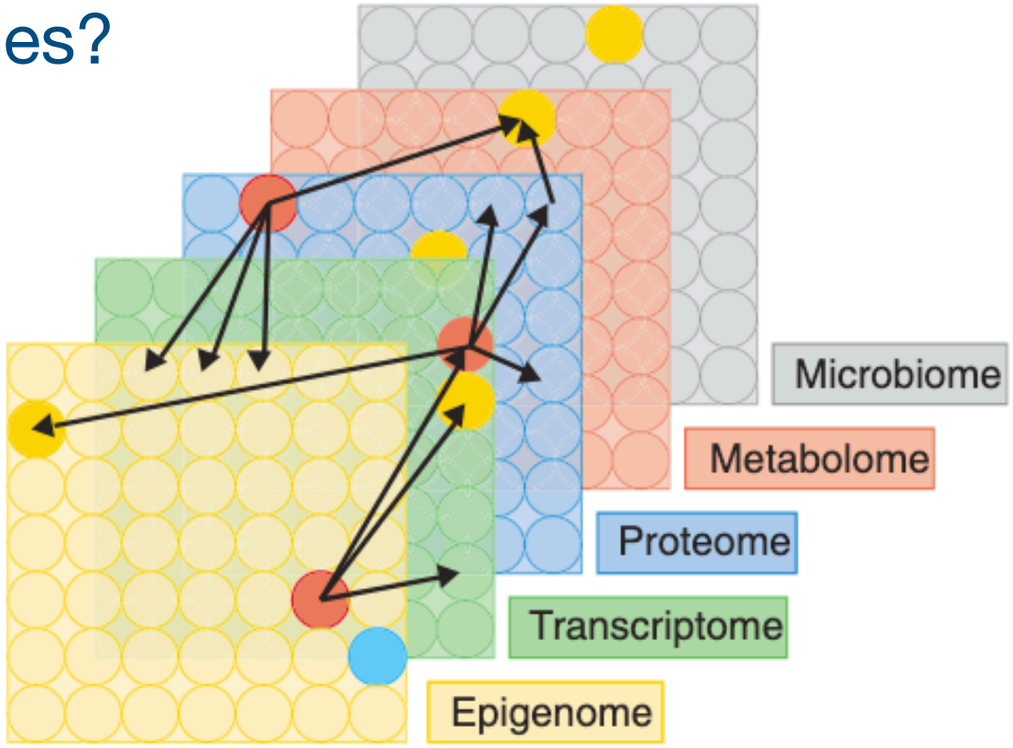
Mardinoglu 2018

Interomic vs Intraomic networks

Networks may be build for individual omics or for their integration

What is my biological question?

- Do I want to analyse vertical relationships between features?
- Biological motivation for integrating omics with different coverage (e.g. transcriptomic and proteomic)
- Do I want to extract functional properties?



Different approaches for network inference

1. Feature association
2. K-nearest neighbour graph (k-NNG) construction

No prior graph structure

3. Knowledge-based
4. Genome-scale metabolic models

Based on prior information

1. Association analysis

Balanced dataset for group sizes

GroupA (80 samples) vs GroupB (20 samples)

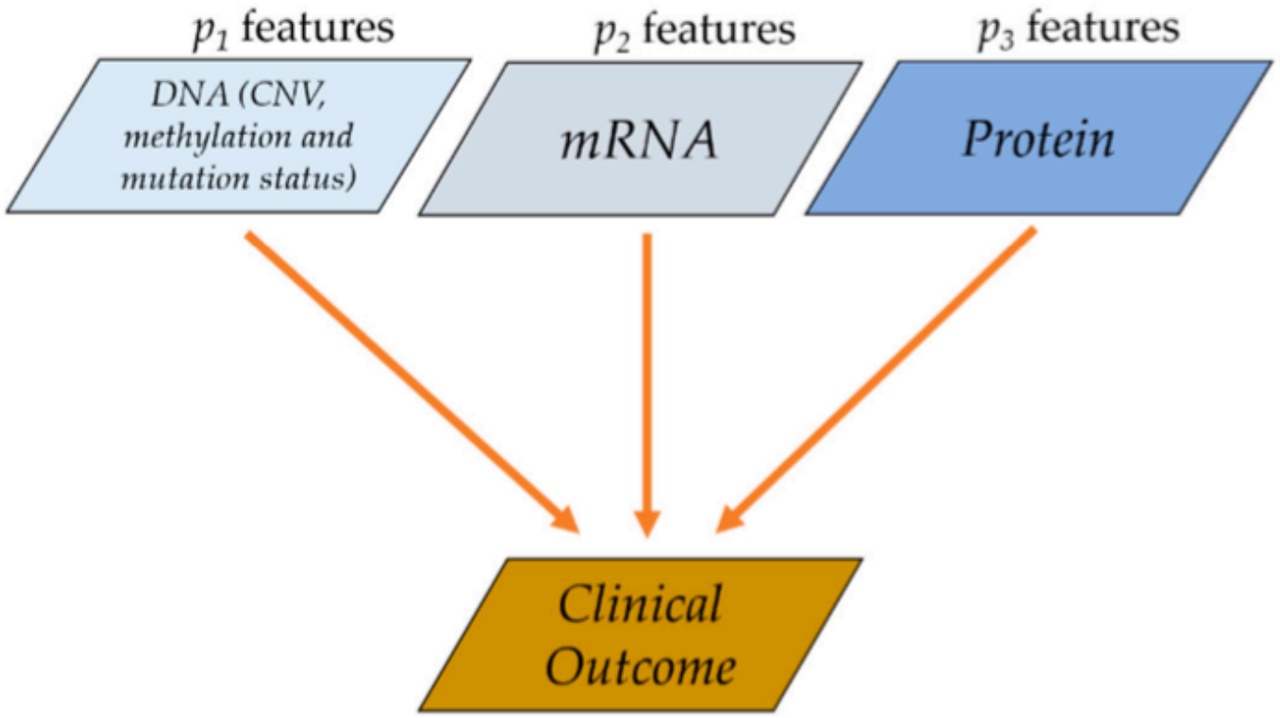
GroupA (50 samples) vs GroupB (50 samples)

Common approach: compute correlations between different features

- Spearman

- Pearson

Extend known associations



1. Association analysis

Easy to interpret

Unweighted vs weighted ($-1 \leq \rho \leq 1$)

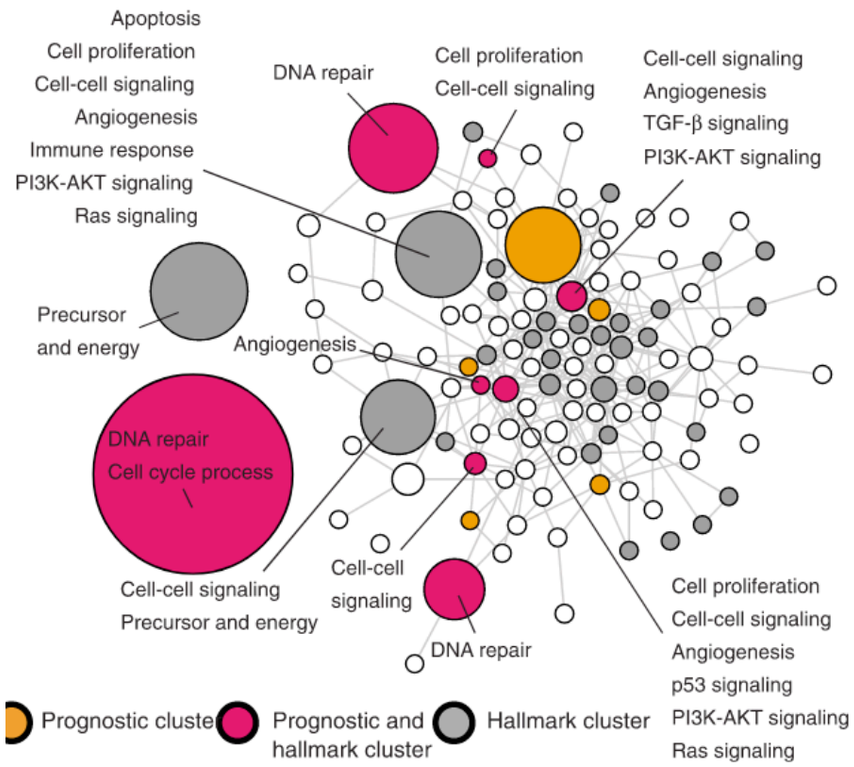
Unbalanced networks

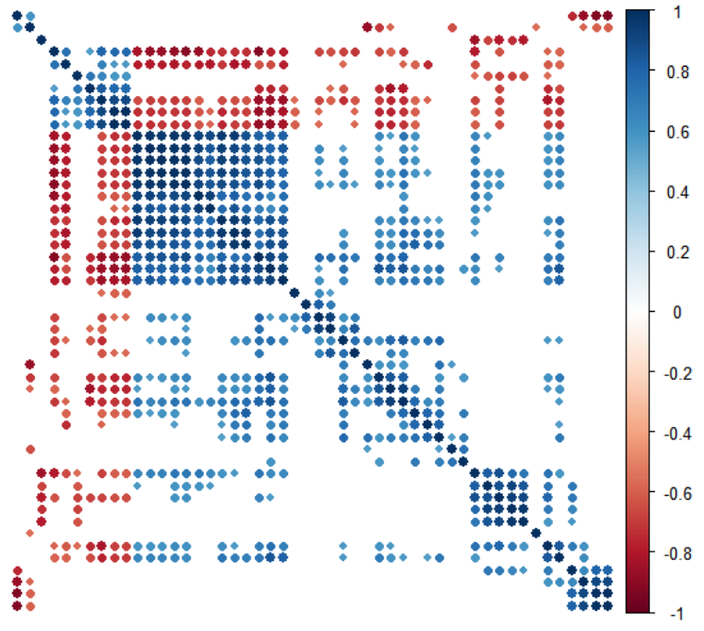
Prone to type I errors

Filtering

- FDR vs Bonferroni
- Correlation coefficient cutoff

Need adjustment to possible confounding factors





Wu 2019
Uhlen 2017

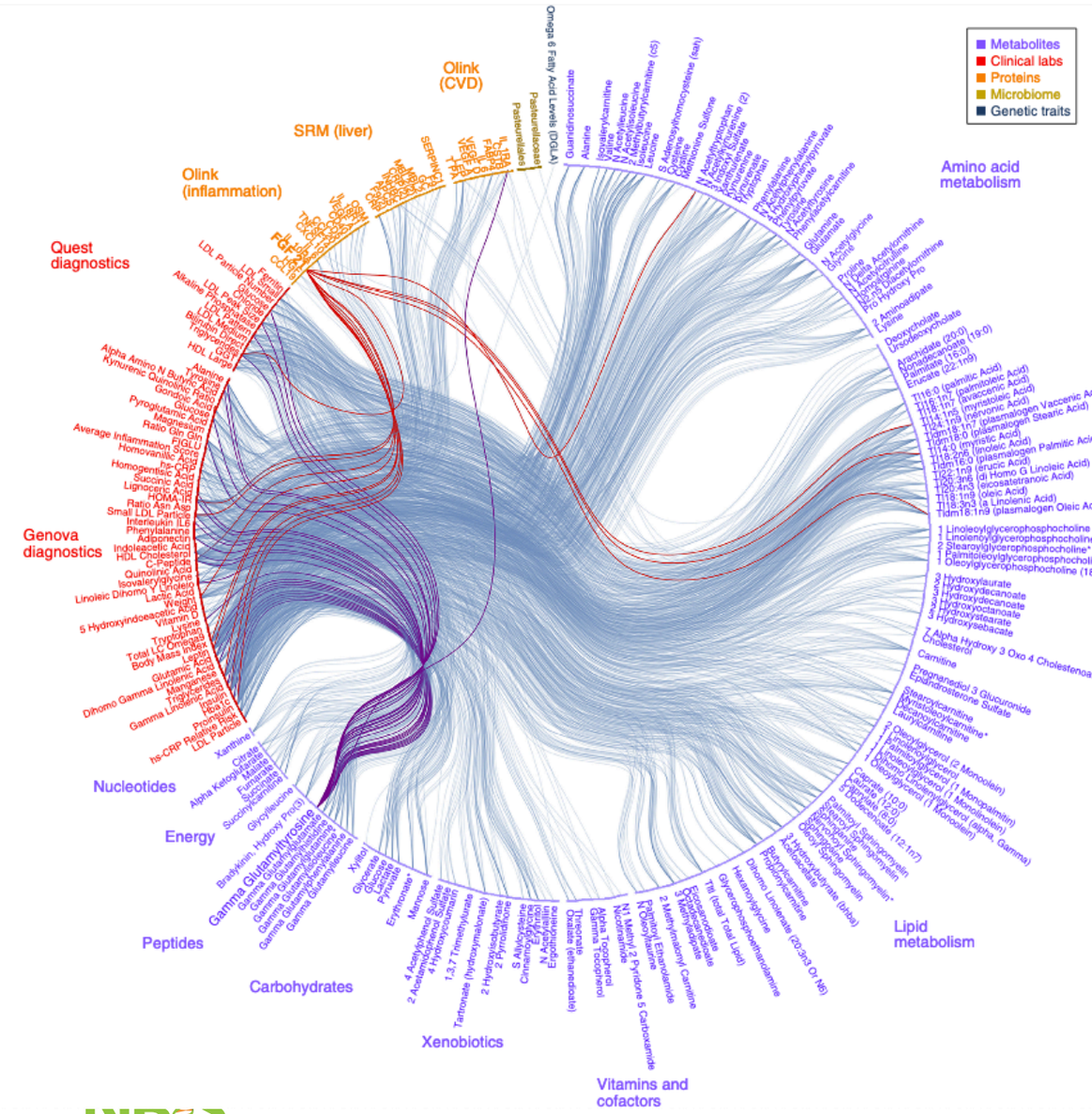
1. Association analysis

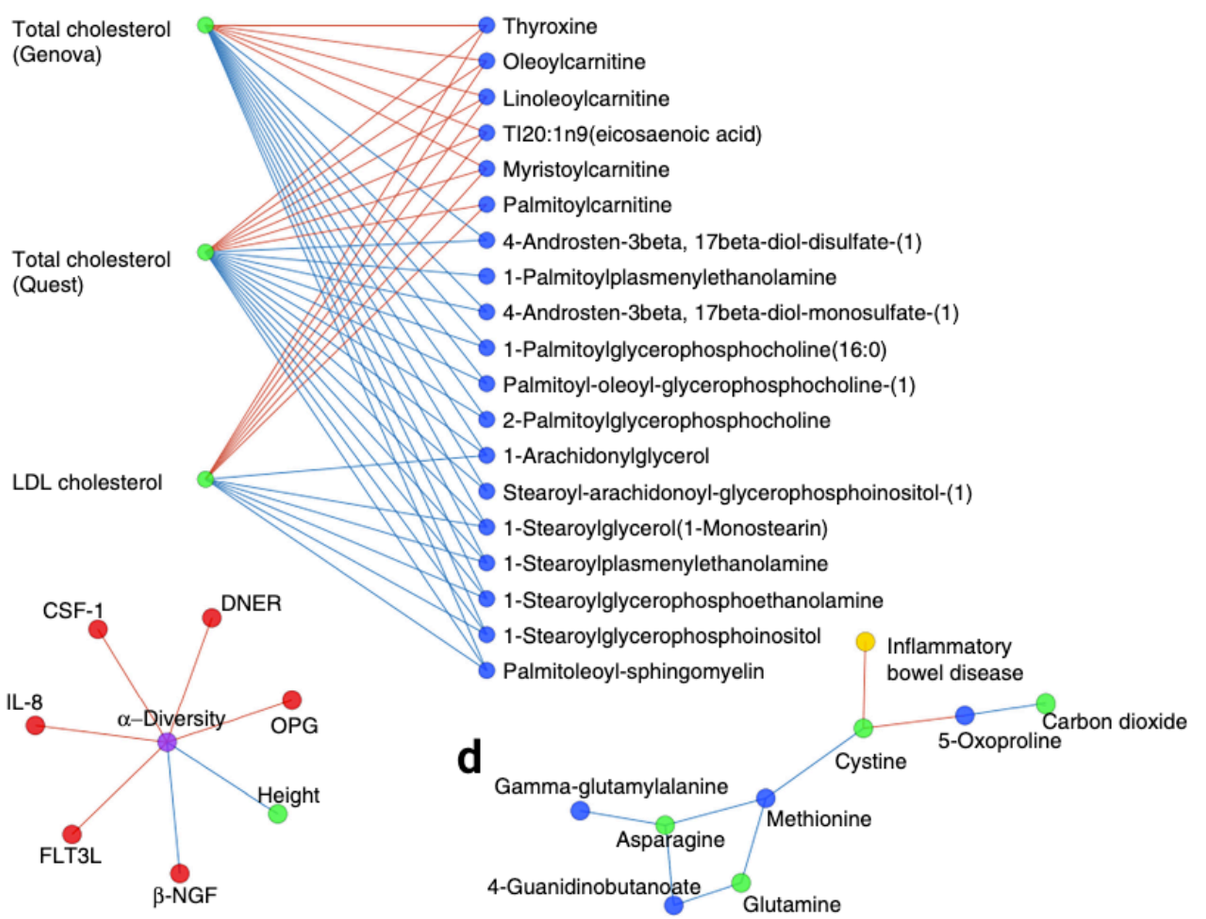
Adjusting for confounding factors: partial correlation analysis

Below:

- gender and age are known confounding factors

- feature regression on confounding factors, followed by correlation on the residuals of each model



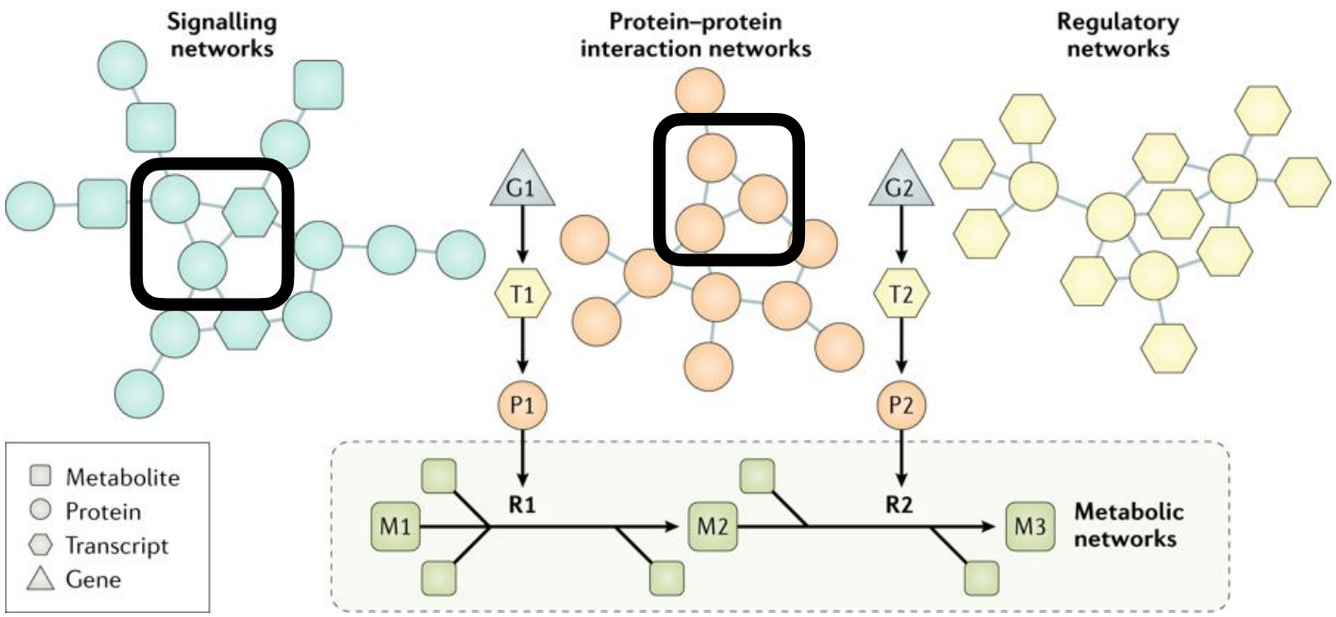


Computationally expensive representations may be simplified

Does your graph have many cliques? **Possibly noisy**

Graph contraction simplifies the graph by successively grouping cliques

Problem: reduces information and prevents studying many properties of the graph



Clarke 2011
Krzywinski 2013
Sham 2014
Nygaard 2016
Piening 2018
other refs as links

2. *k*-nearest neighbour graph

1. For each pair of features (u, v), compute a distance metric:

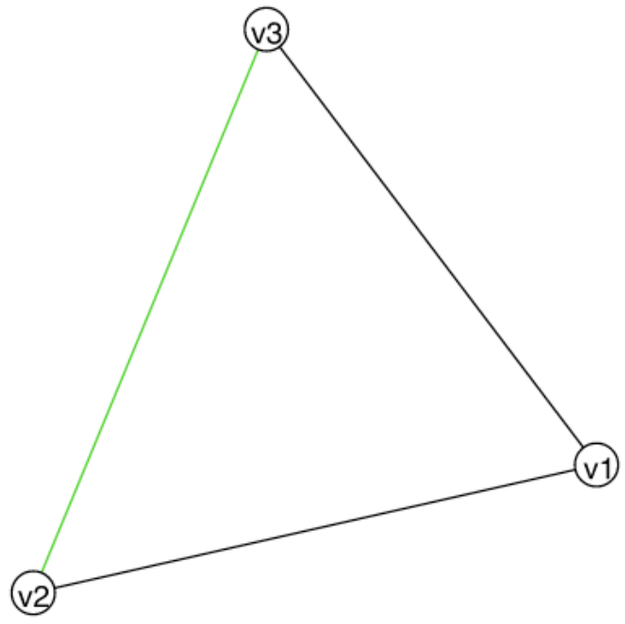
- Correlation

- Euclidean

- Jaccard

- ...

2. For each feature, select the *closest k* neighbours



Efficiency (not scalable, compute all neighbours for every node)

Generates well-structured graph

Simple as it reduces the number of features

Loses potentially important information because *k* is fixed

2. *k*-nearest neighbour graph

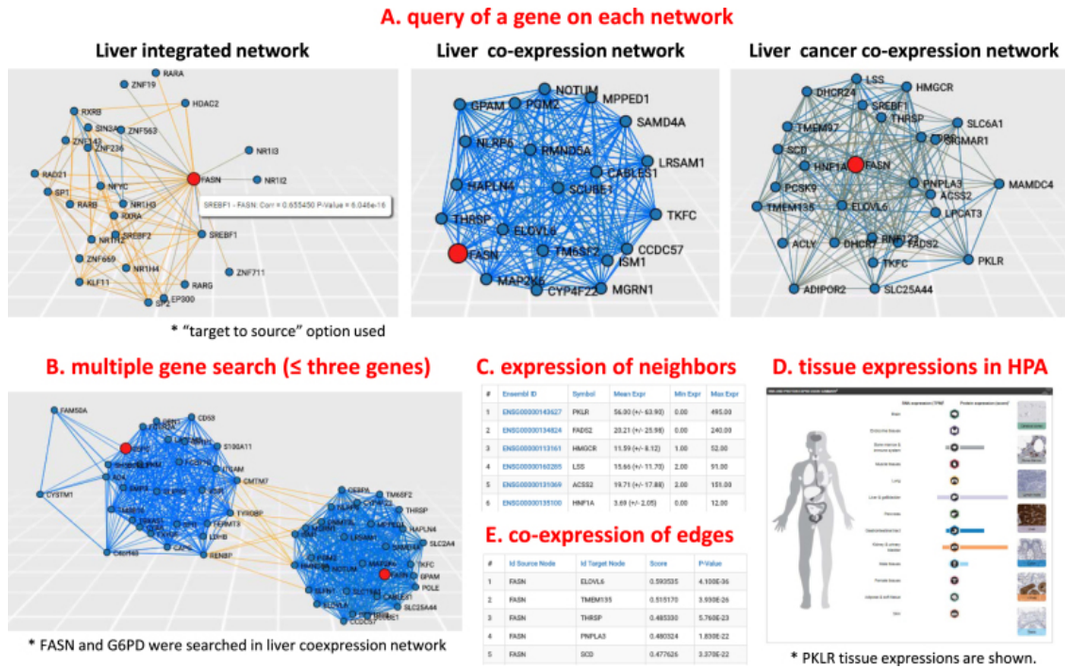
High *k* is smooth, but biased (underfitting)

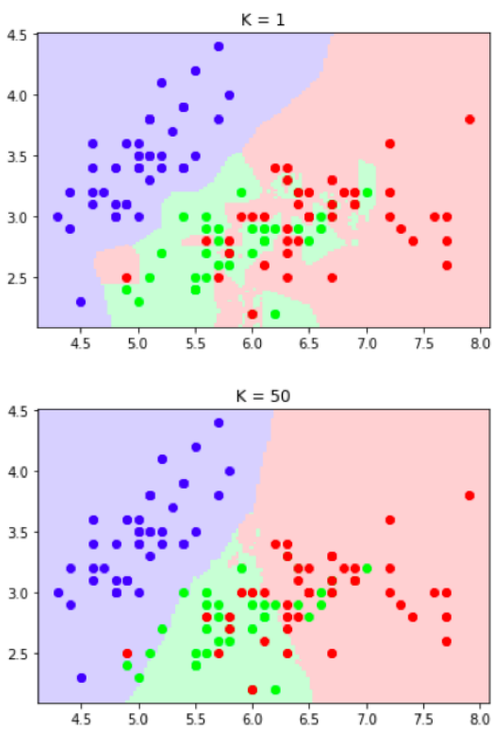
Low *k* is accurate, but noisy (overfitting)

Optimum *k* may be identified:

- cross validation

- ad-hoc





3. Knowledge-based graph creation

Database-derived

- PPI
- TFRN
- Metabolic Atlas
- …

Many reference databases

KEGG

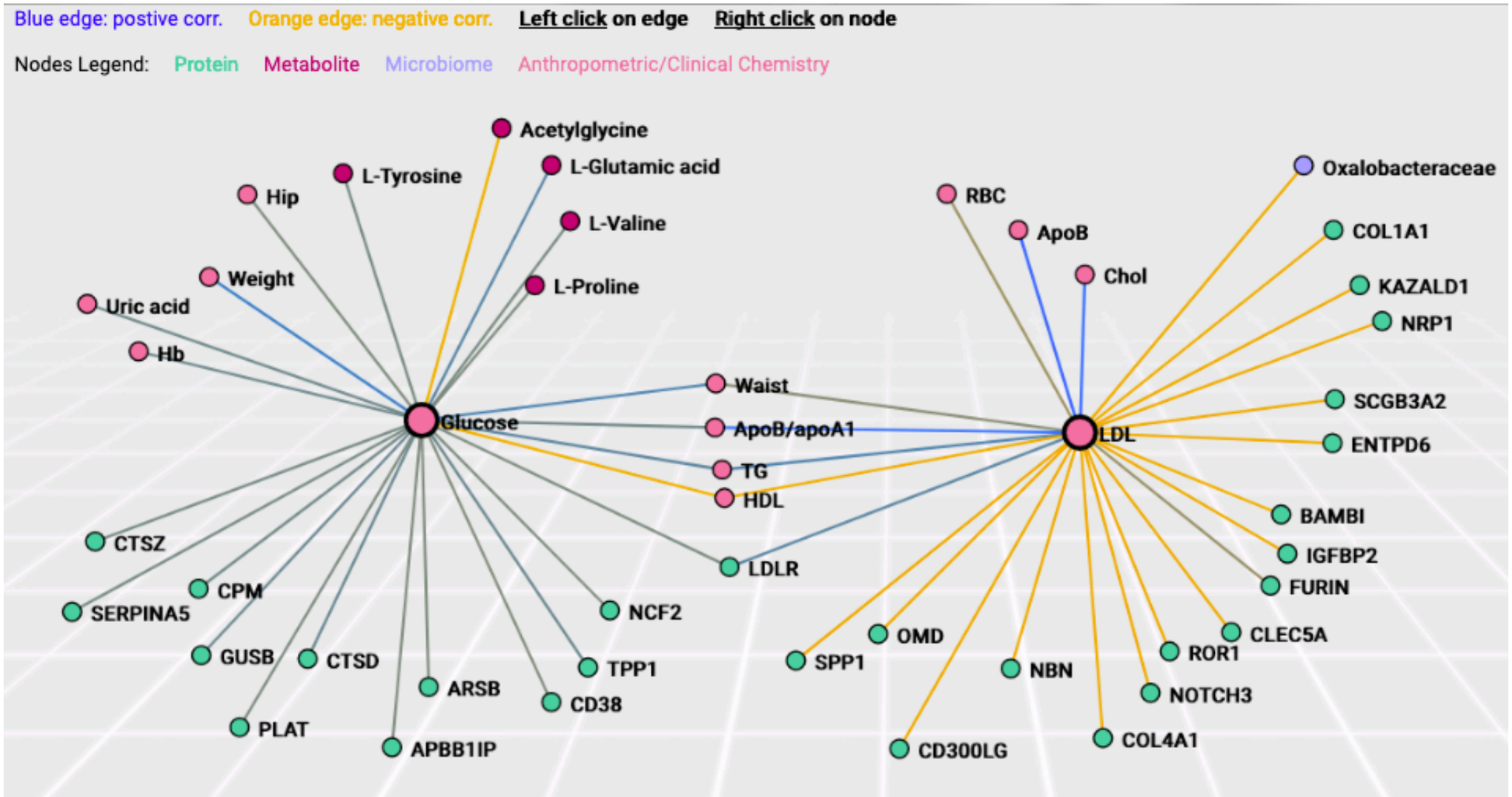
Reactome

WikiPathways

STRING-DB

3. Knowledge-based graph creation

Multi-omic biological networks



3. Knowledge-based graph creation

NAR December 2019: 1637 databases

OXFORD ACADEMIC | Journals

You are here: NAR Journal Home » Database Summary Paper Categories

NAR Database Summary Paper Category List

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

3. Knowledge-based graph creation

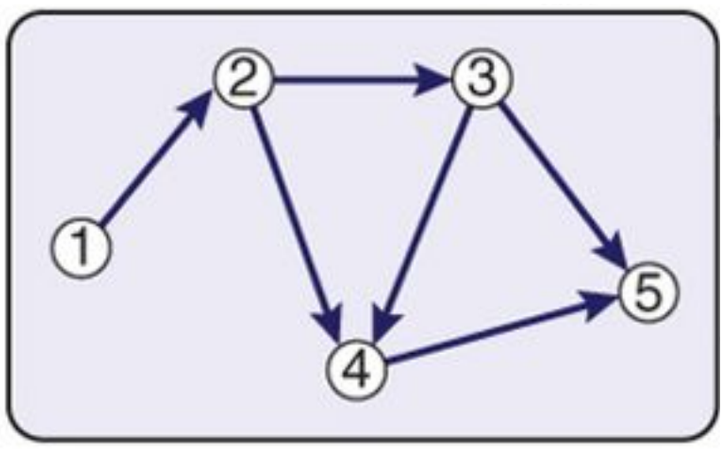
Little overlap among reference pathways

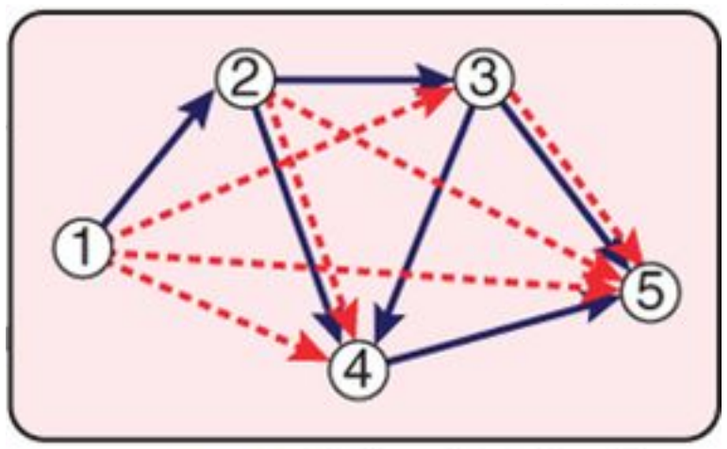
3. Knowledge-based graph creation

How to overlay your data based on known interactions?

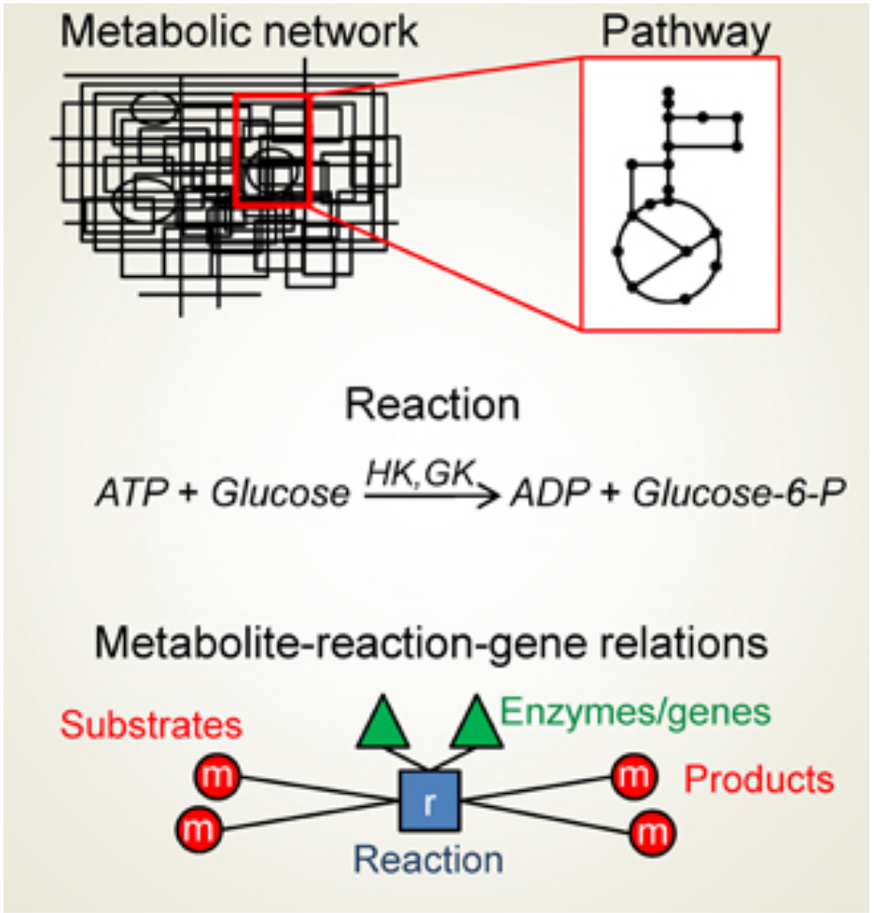
- Filter your predicted interactions based on known information? (intersection)

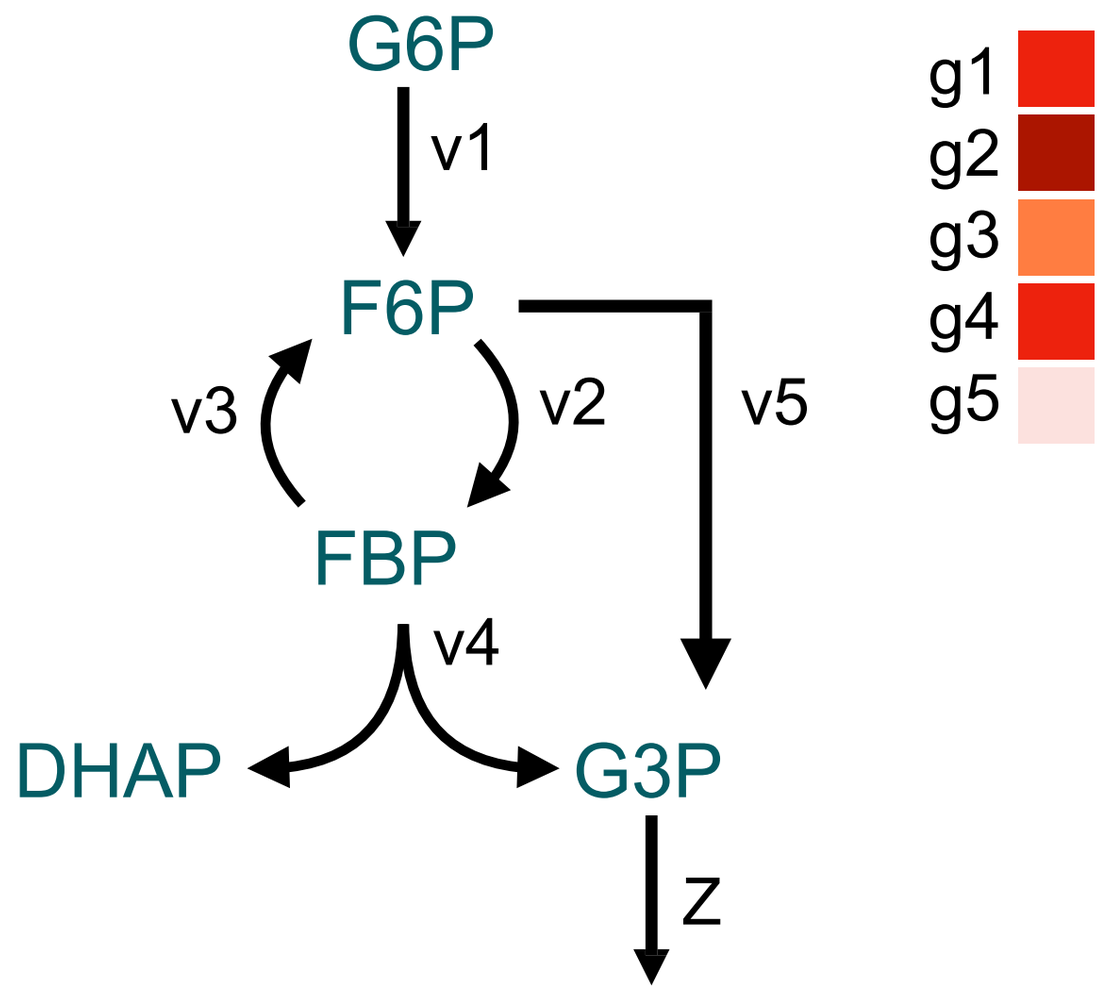
- Add interactions that are not found in the reference networks?





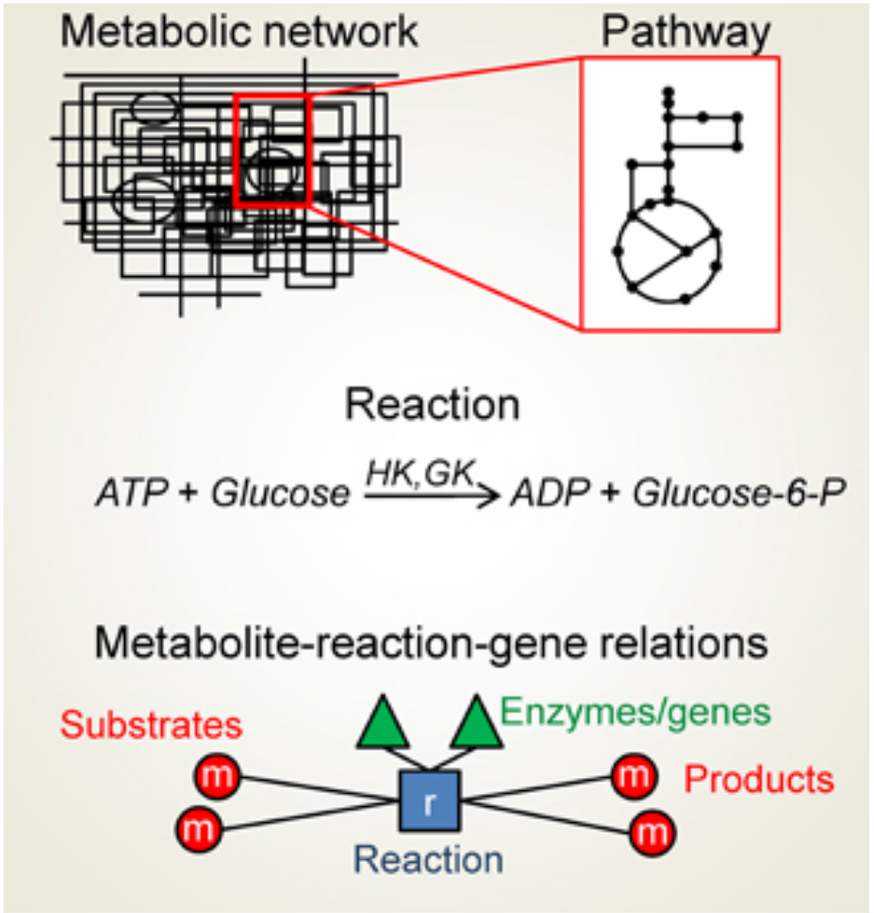
4. Genome-scale metabolic models as integrative networks

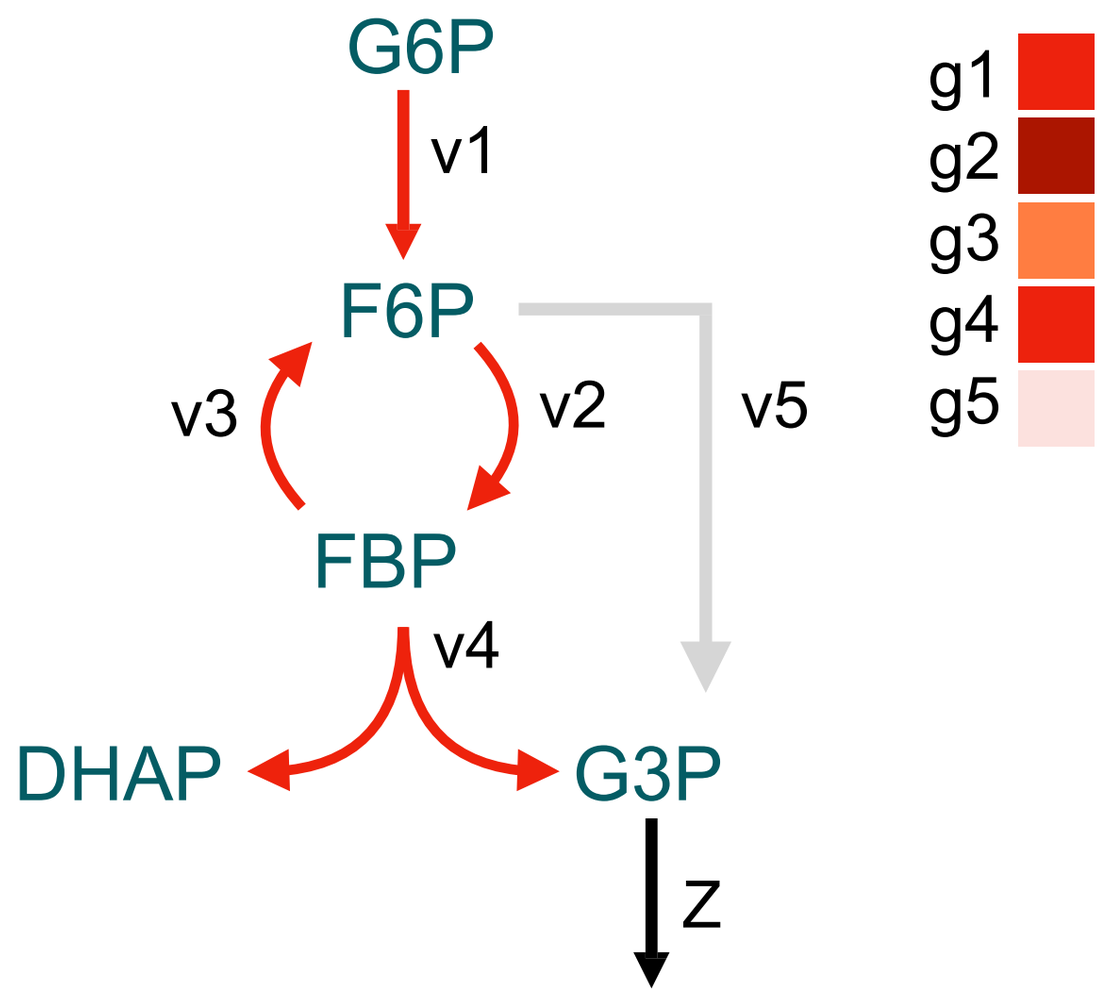




Väremo 2013
Orth et al Nat Biotechnol (2010).

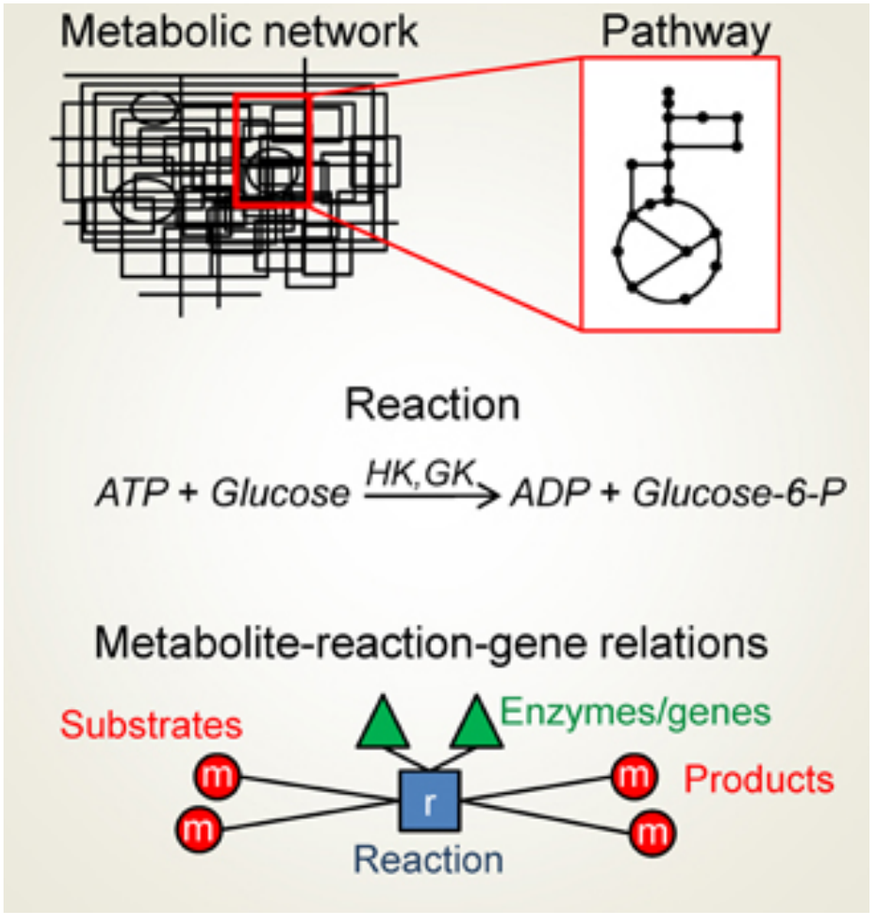
4. Genome-scale metabolic models as integrative networks

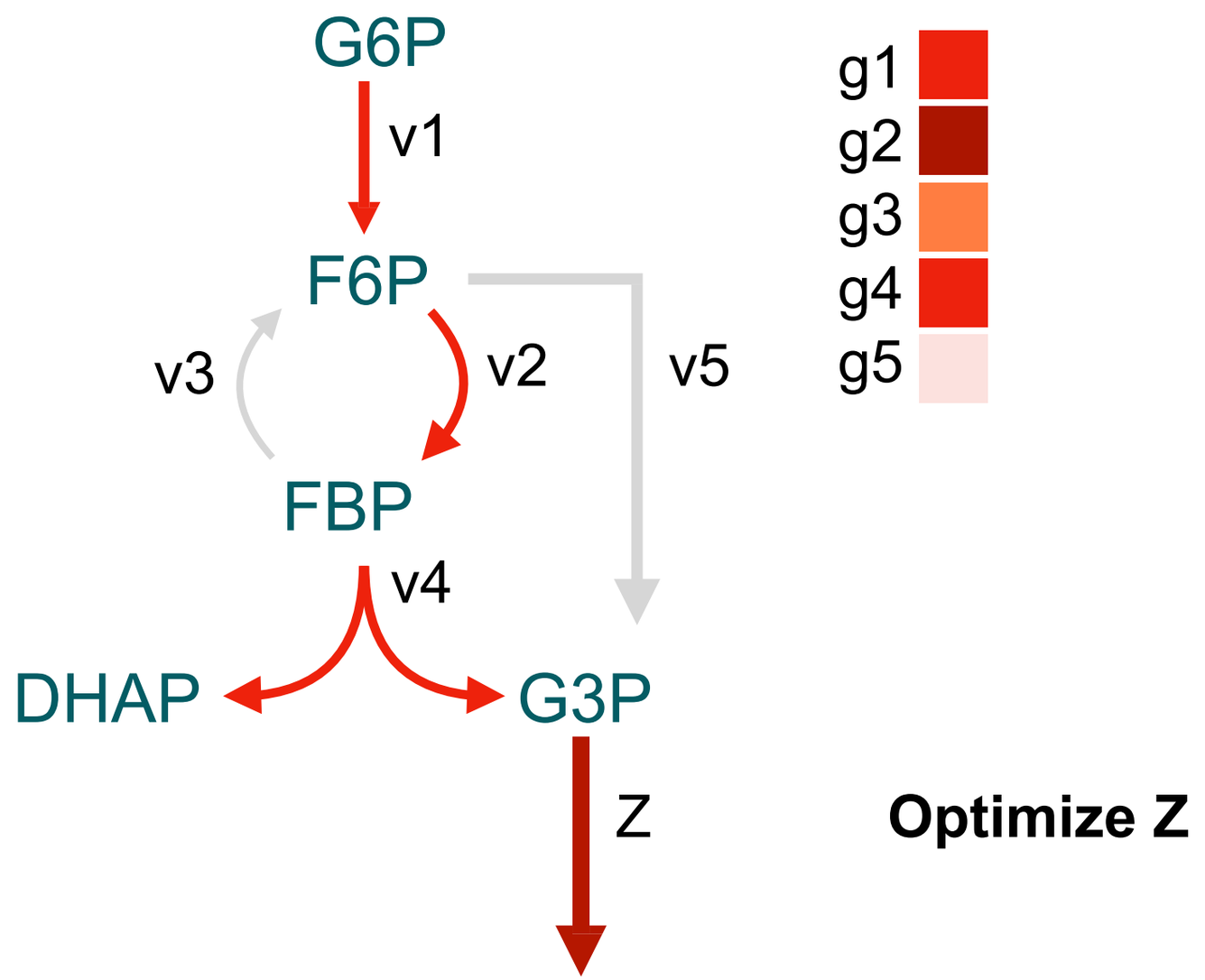




Väremo 2013
Orth et al Nat Biotechnol (2010).

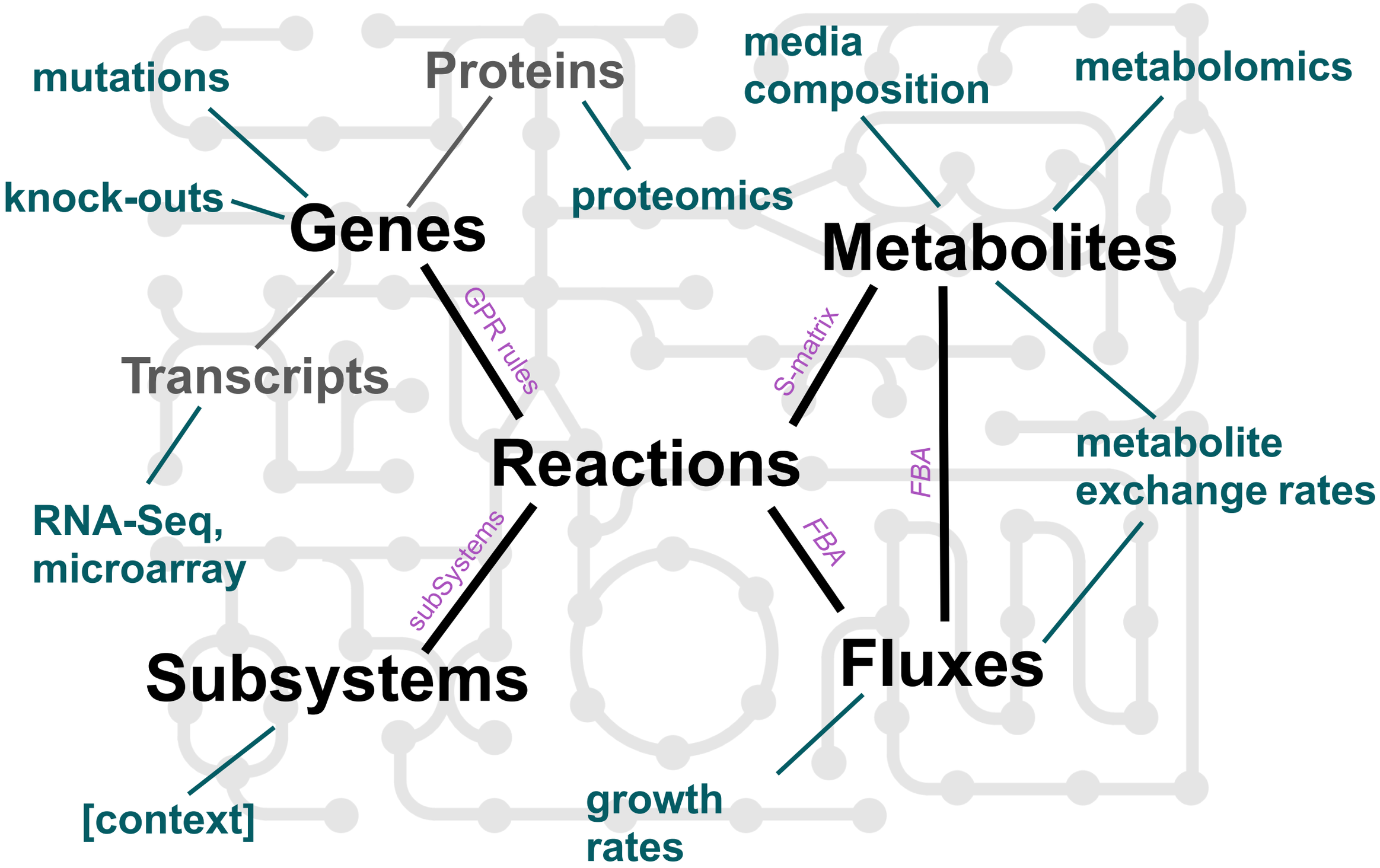
4. Genome-scale metabolic models as integrative networks





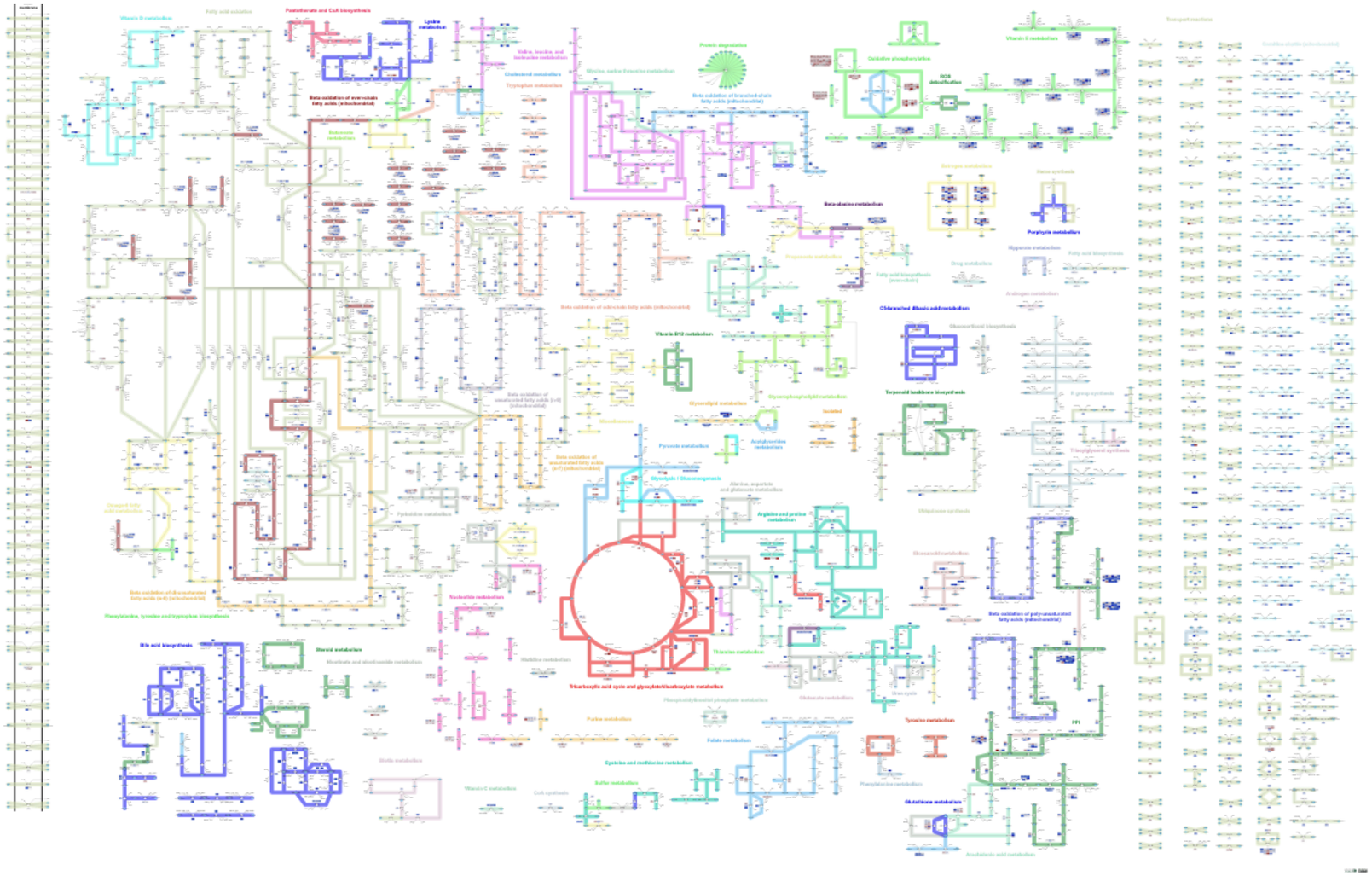
Väremo 2013
Orth et al Nat Biotechnol (2010).

4. Genome-scale metabolic models as integrative networks



4. Genome-scale metabolic models as integrative networks

metabolic ATLAS



Simulate flux distributions

Dysregulated pathways

Reporter metabolites

Essential genes

Targetable enzymes

May be combined with standard graph analysis

https://metabolicatlas.org/

4. Genome-scale metabolic models as integrative networks

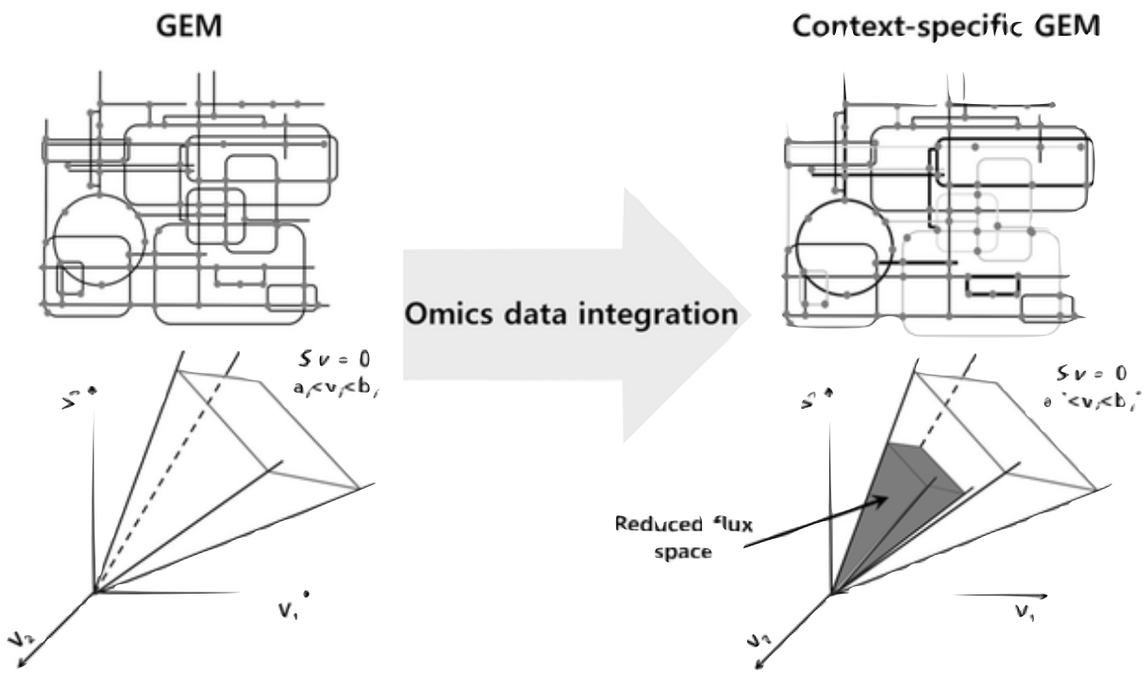
GEMs may be used to find such missing relationships, but there is a coverage issue

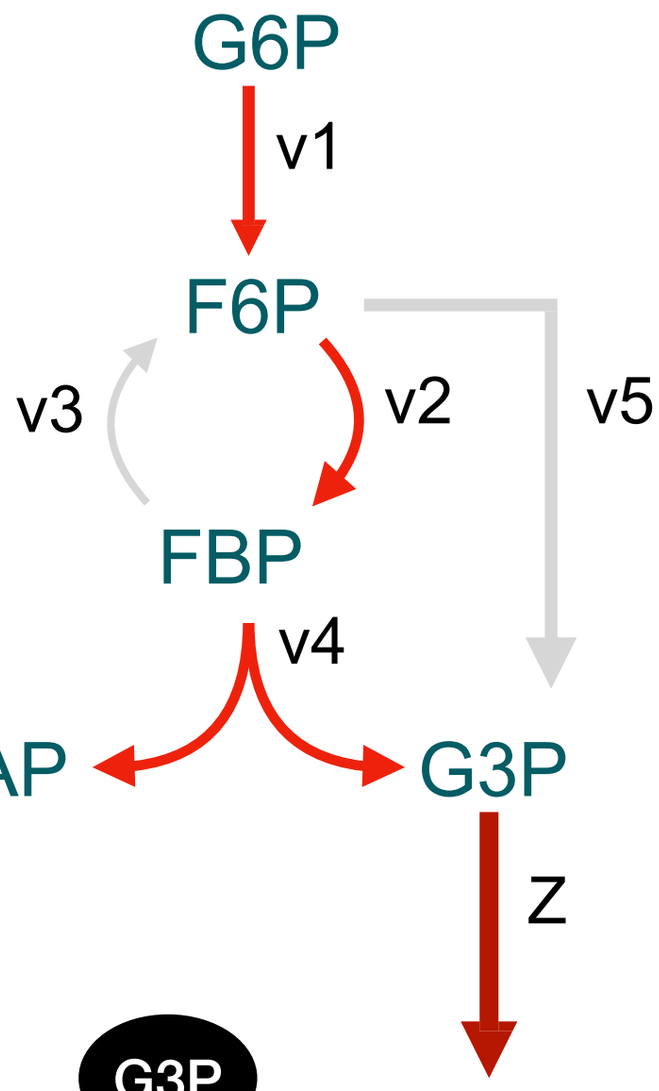
The overall strategy follows

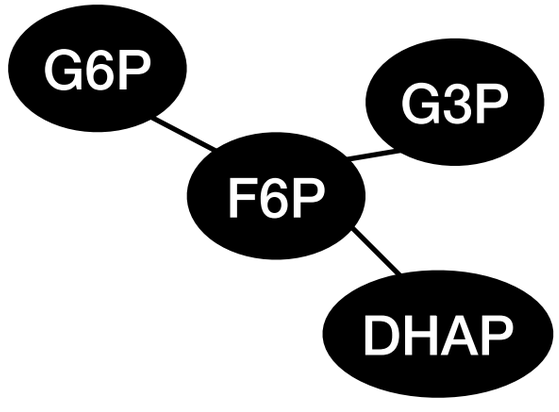
1. Integrate proteomic, transcriptomic, metabolomic, fluxomic
2. Flux distribution
3. Compute metabolite-reaction-gene relationships
4. Extract relevant relationships (met-met, gene-gene)

4b. Exclude unnecessary interactions (e.g. cofactors)

5. Downstream analysis (e.g. topology, stratification)







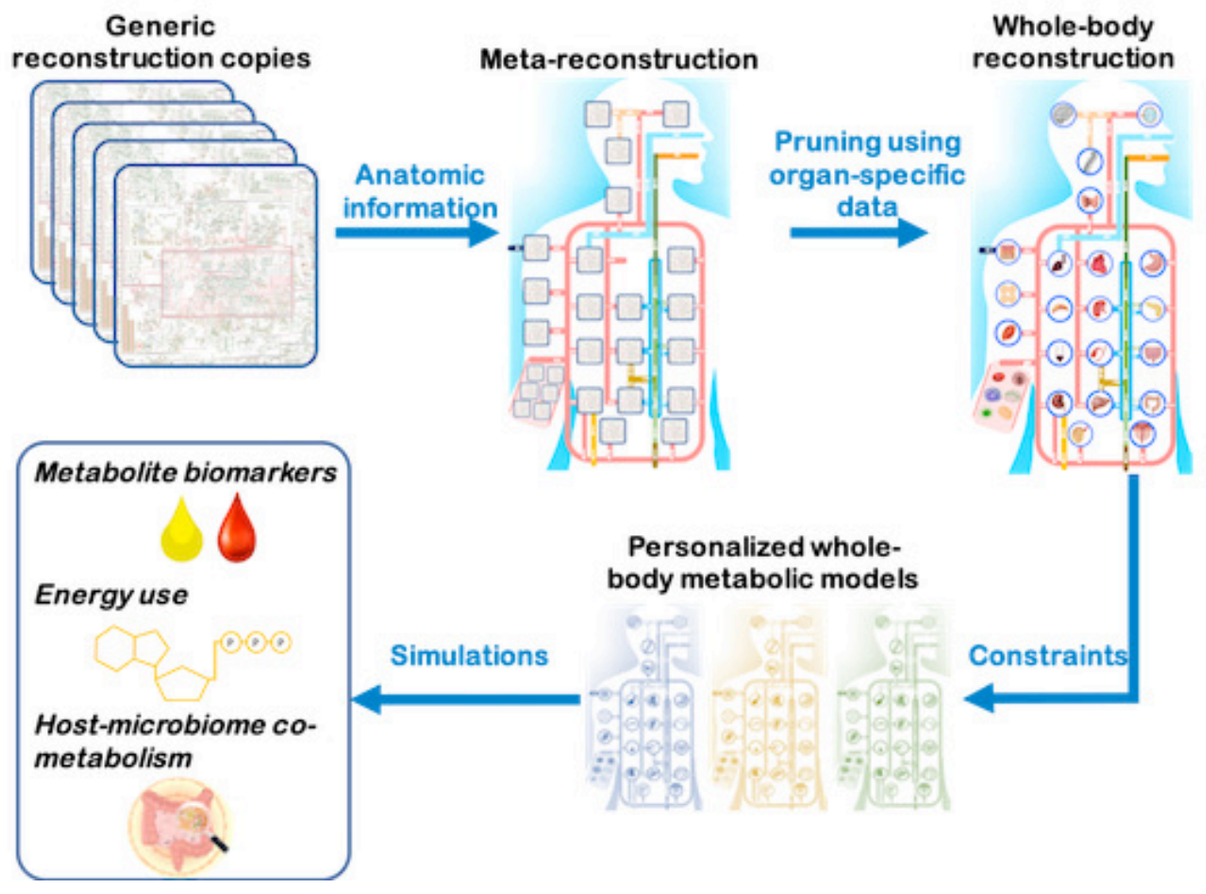
4. Genome-scale metabolic models as integrative networks

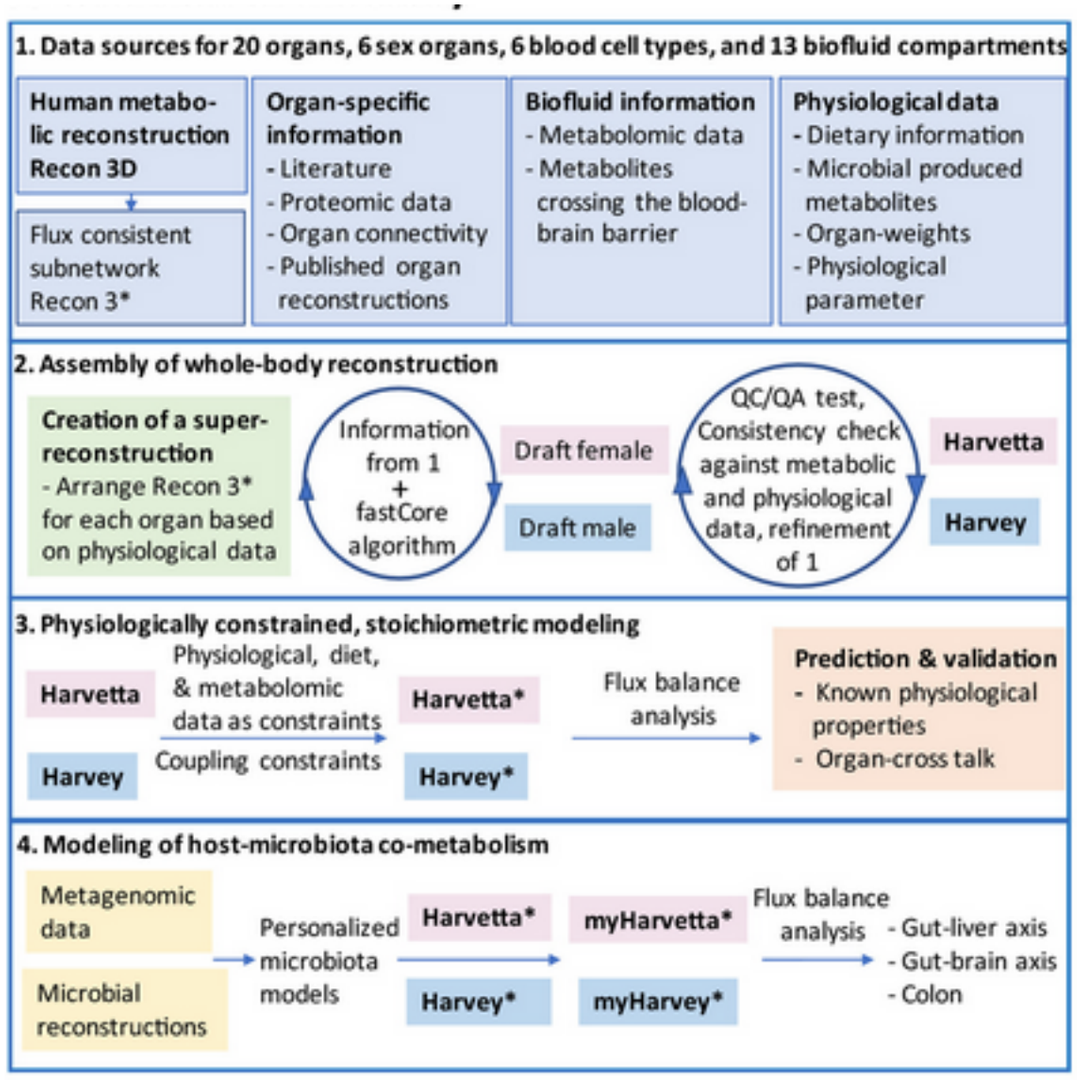
Personalised whole-body models
of host
+
gut microbiome

26 organs

6 blood cell types

80,000 biochemical reactions



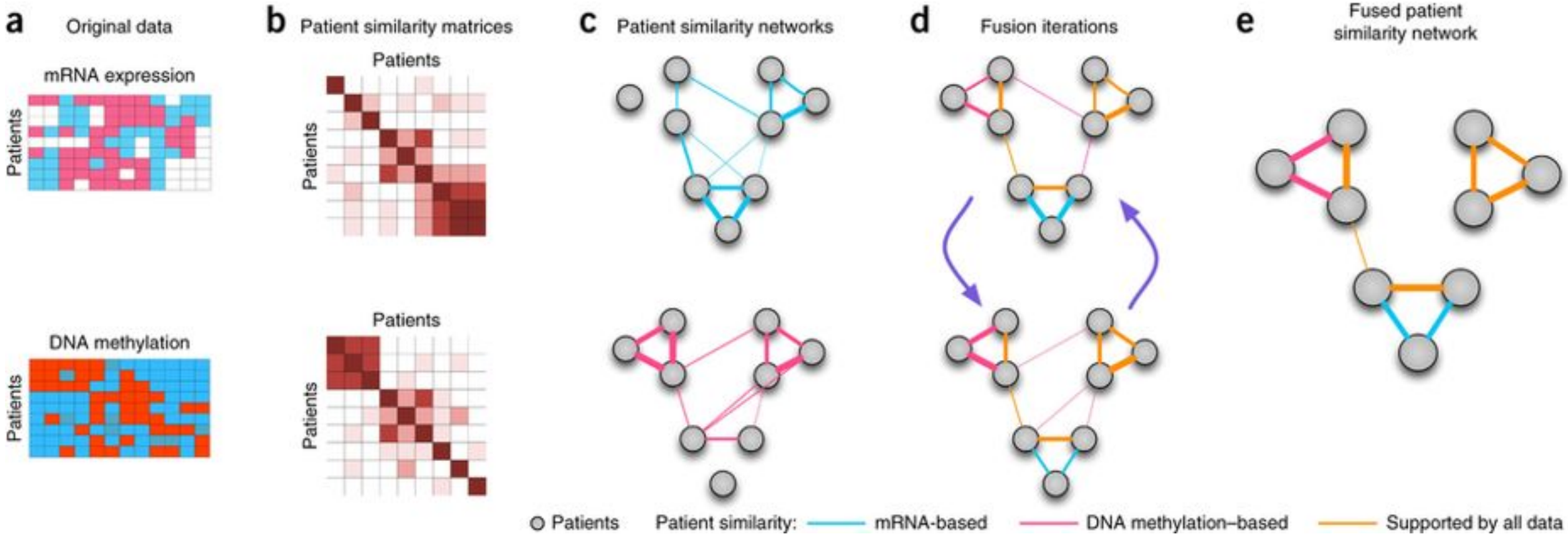


Similarity network fusion

Sample-sample clustering based on multi-omic data improves clustering

Single-omics present complementary (non-redundant) information

Enables further comparisons between clusters

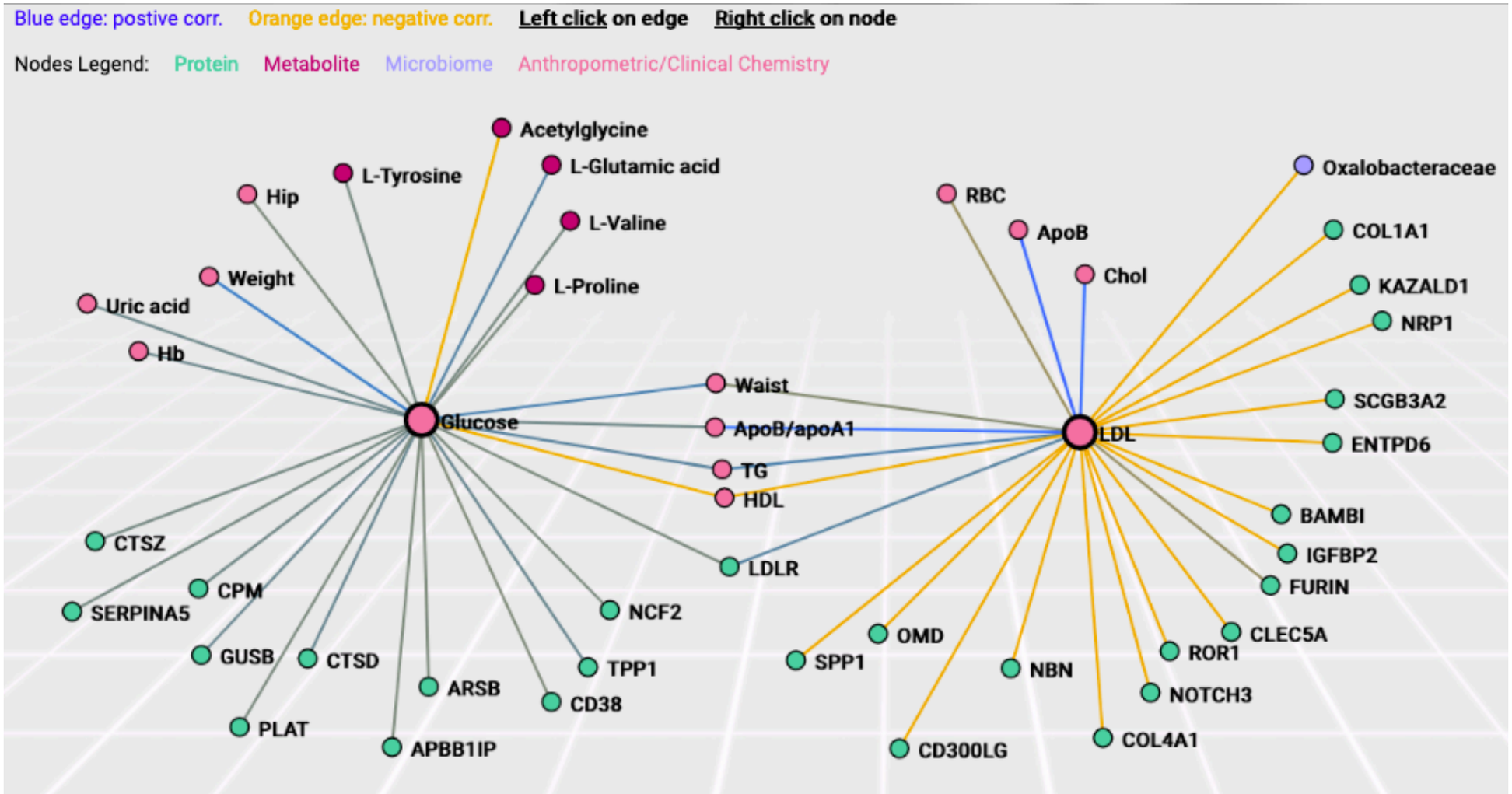


Overview

1. Introduction to network analysis
2. Terminology
3. Network inference
4. **Key network properties**
5. Community analysis

Motivation

You have built an association network (e.g. PPI, multi-omic). How to identify pivotal features, their organization, and biological characteristics?



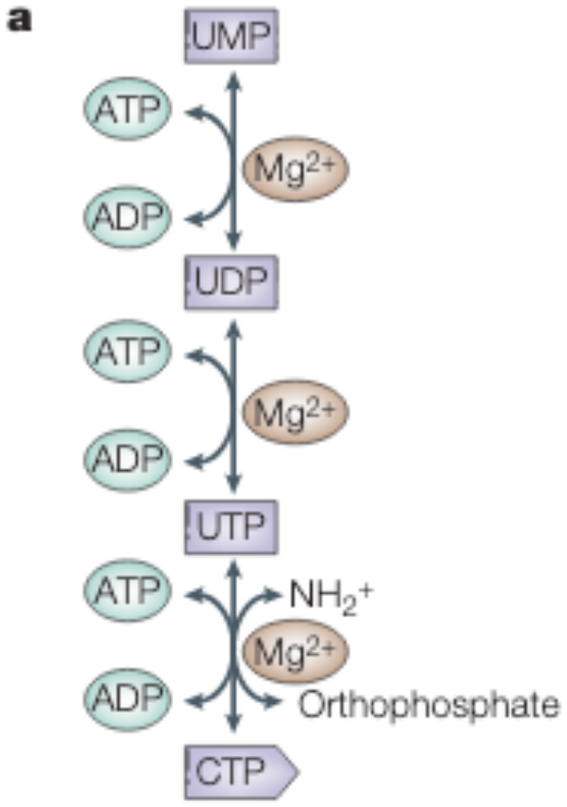
Key network properties to discuss

1. Network representations
2. Network density
3. Paths
4. Centrality
5. Clustering coefficient
6. Degree and connectivity distributions

1. Network representations

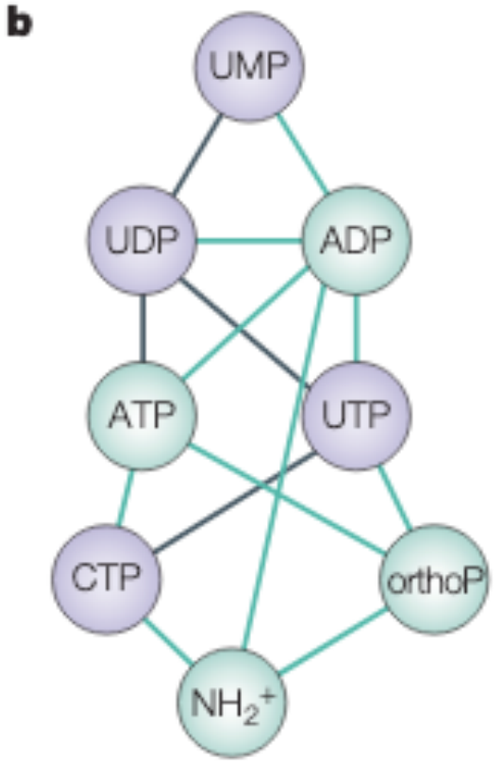
Representations of a metabolic network: pyrimidine metabolism

Metabolism



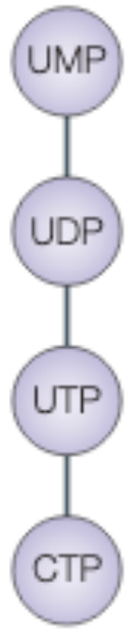
(directed graph)

Graph representation: metabolites and co-factors



(undirected graph)

metabolite-metabolite association



(undirected graph)

Other representations: Protein-Protein, Protein-Metabolite

2. Network density

A **dense graph** is a graph where the number of edges approximates the maximum possible number of edges for the given node number.

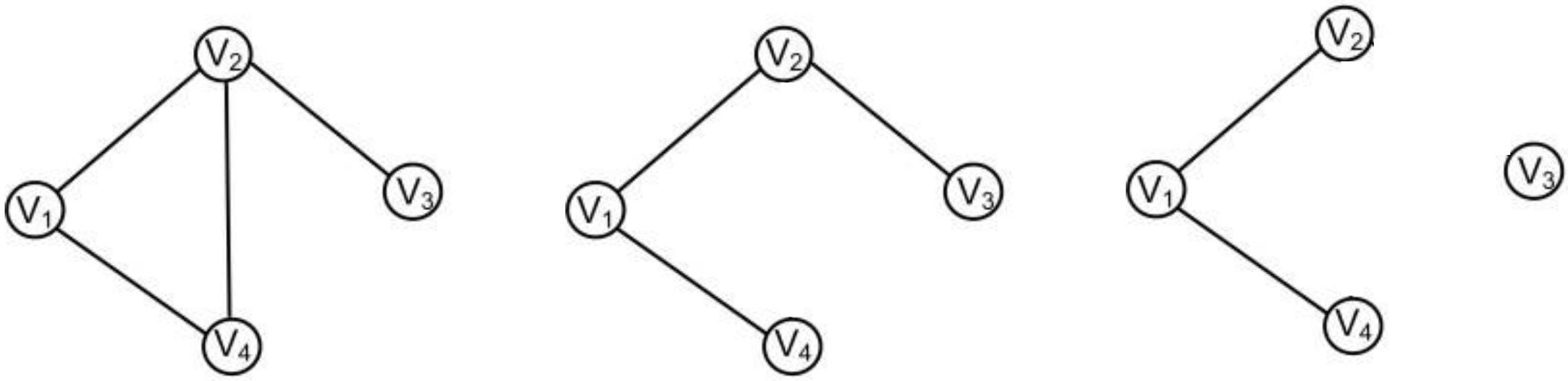
We can thus compute the network **density** (or **global connectivity**) as

$$\textbf{Undirected graphs: } D = \frac{2 * E}{V \cdot (V - 1)}$$

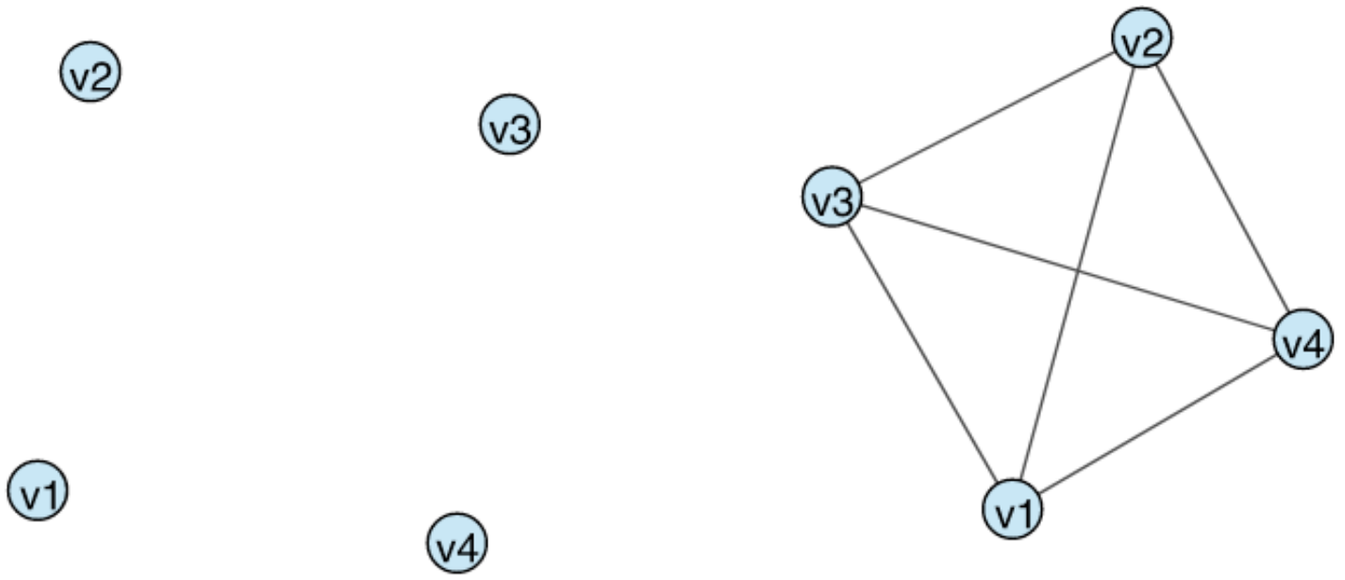
E : number of edges

V : number of vertices

$$\text{Possible edges} = \frac{V \cdot (V - 1)}{2}$$

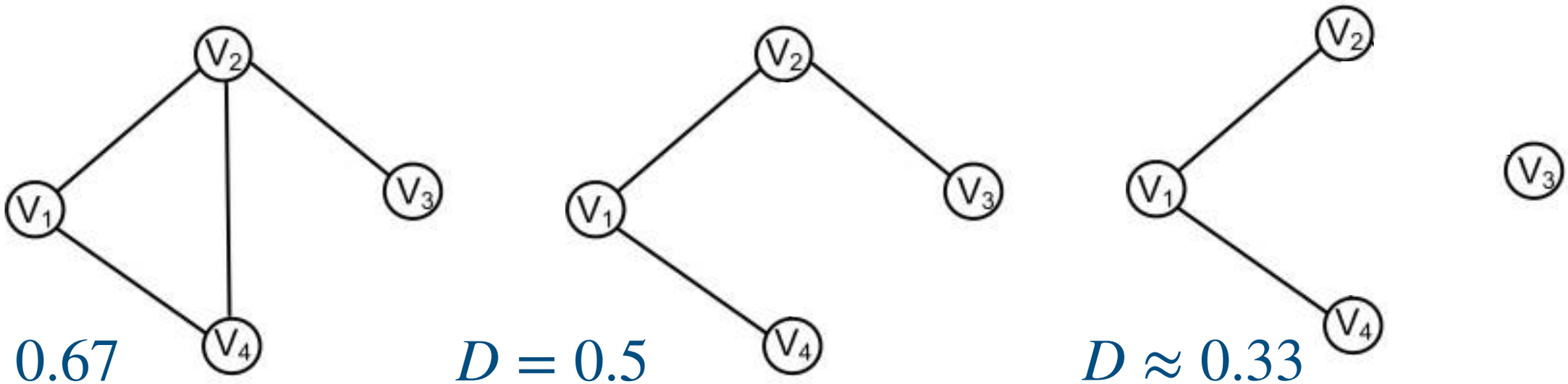


2. Network density



$$0 \leq D \leq 1$$

Higher density indicates higher associations in the network, which implies lower resilience to changes.



$D \approx 0.67$ $D = 0.5$ $D \approx 0.33$

2. Biological network density

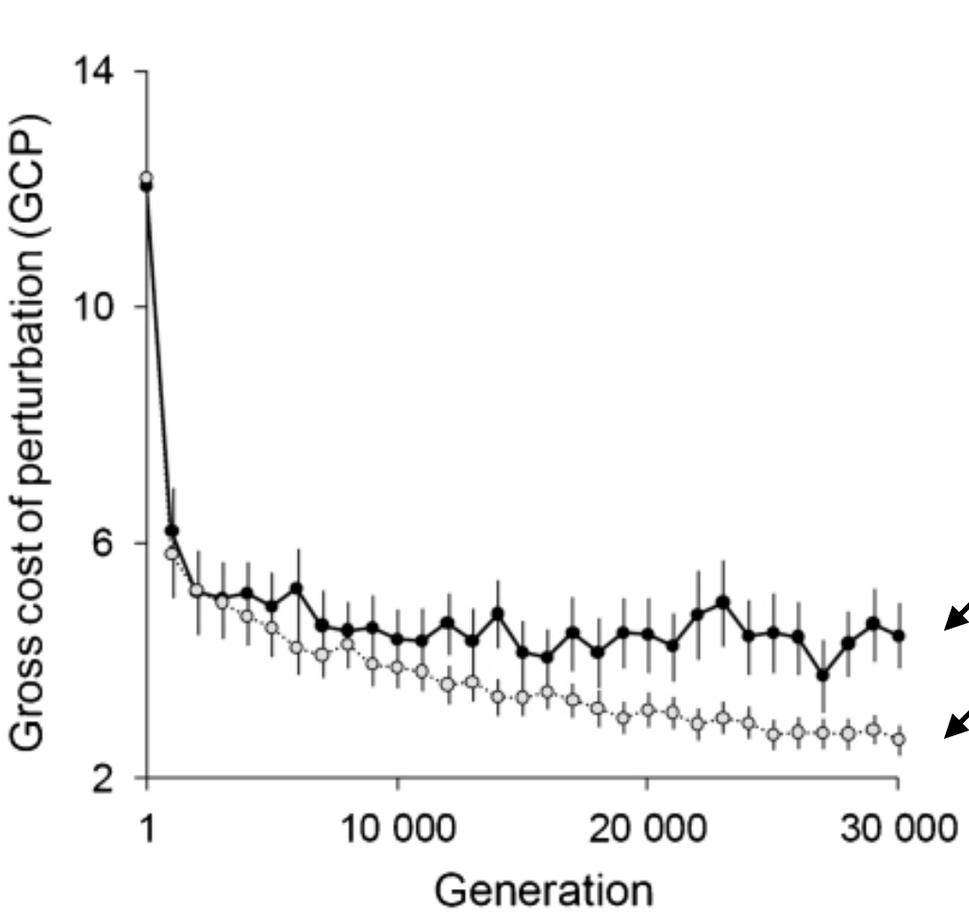
Evolutionary analysis of biological networks indicates general sparsity

Network structure must balance robustness to mutation, stochasticity and environmental queues

Sparse networks show higher robustness when accounting for costs and benefits of complexity

Table I Biological networks are sparsely connected

Organism	Interactions	Genes	D
Drosophila melanogaster	29	14	0.148
D. melanogaster	45	25	0.072
Sea urchin	82	44	0.0065
Saccharomyces cerevisiae	1052	678	0.0023
S. cerevisiae	3969	2341	0.0007
S. cerevisiae	106	56	*0.0338*
Escherichia coli[a]	578	423	0.0032
Arabidopsis thaliana[b]	18 625	6760	0.0004



dense network (density 0.9)

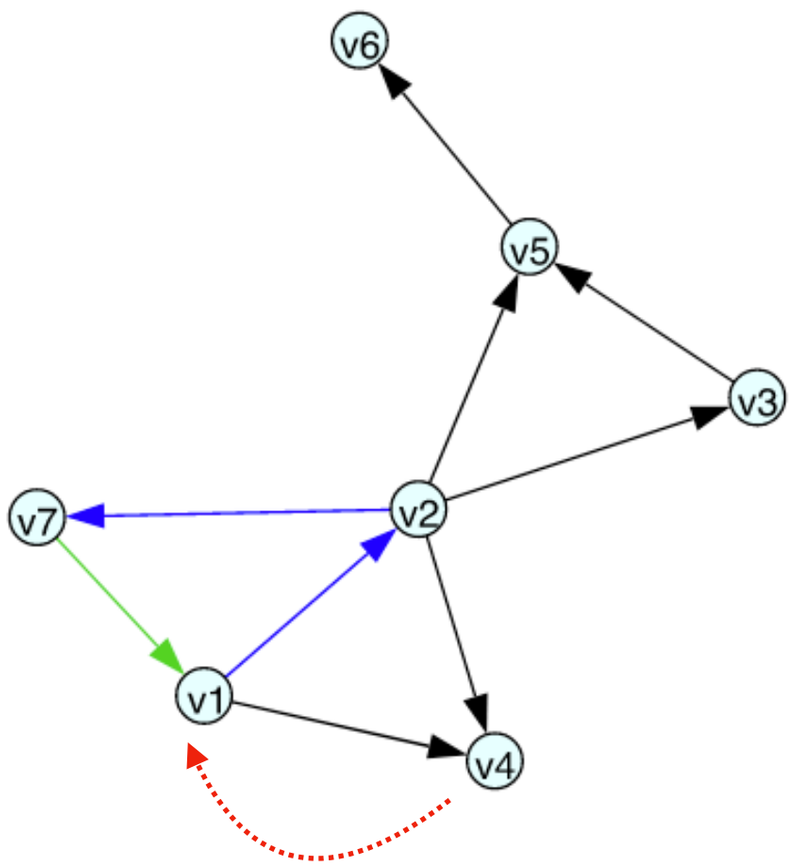
sparse network (density 0.1)

Leclerc 2008

3. Paths

Distance between nodes is measured in path length

In directed graphs, the shortest path between $(a, b) \neq (b, a)$

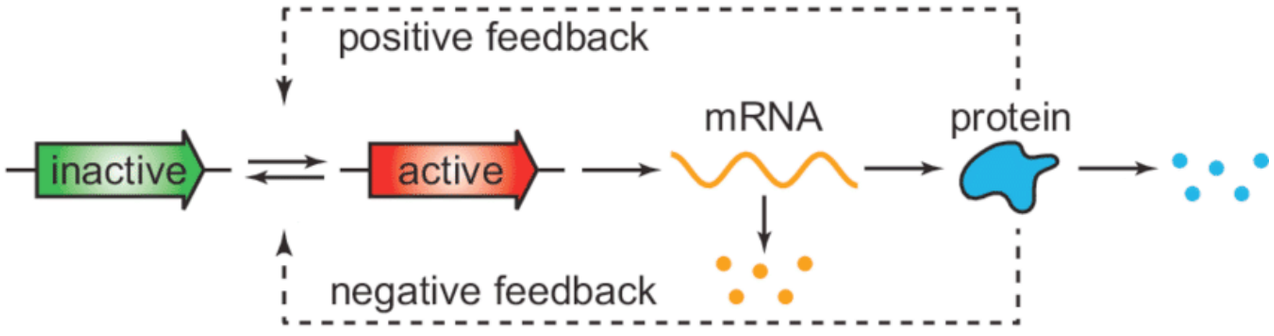


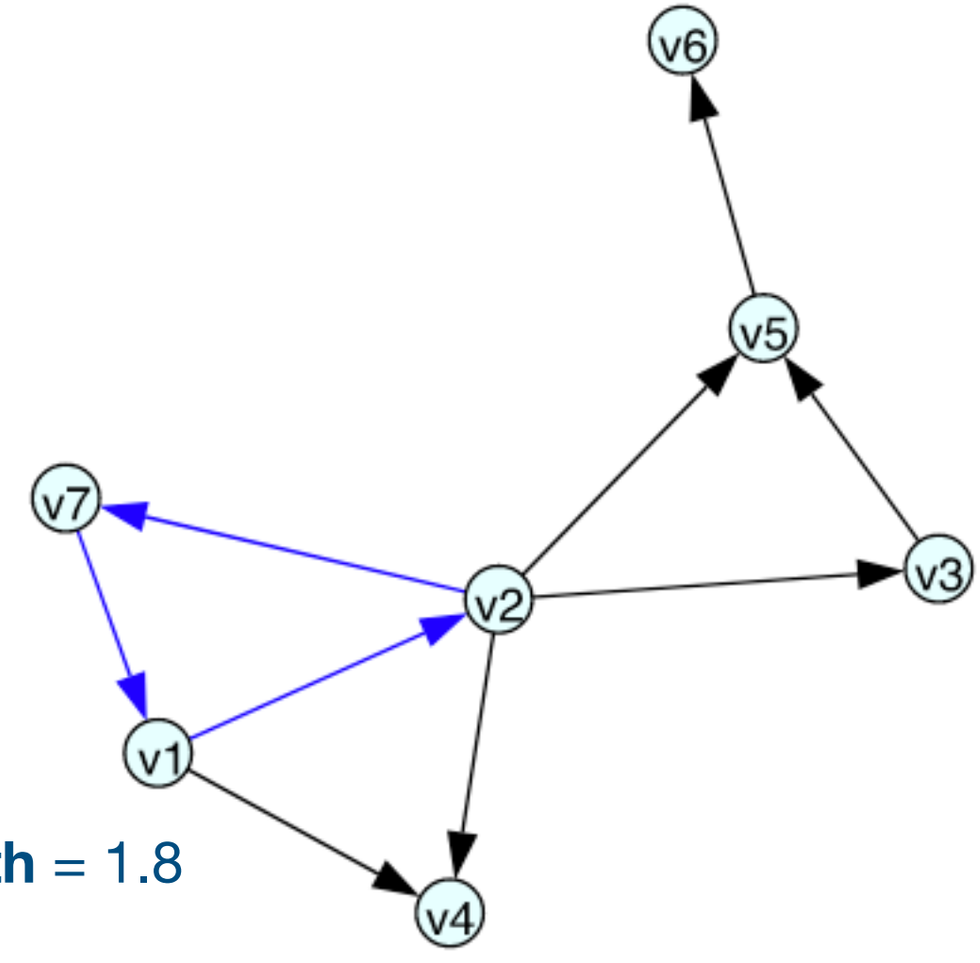
	v1	v2	v4	v3	v5	v7	v6
v1	0.0	1.0	1.0	2.0	2.0	2.0	3.0
v2	2.0	0.0	1.0	1.0	1.0	1.0	2.0
v4	inf	inf	0.0	inf	inf	inf	inf
v3	inf	inf	inf	0.0	1.0	inf	2.0
v5	inf	inf	inf	inf	0.0	inf	1.0
v7	1.0	2.0	2.0	3.0	3.0	0.0	4.0
v6	inf	inf	inf	inf	inf	inf	0.0

3. Paths

Cycles and acyclic graphs

The **average path** gives a measure of network navigability (~feature relationships)



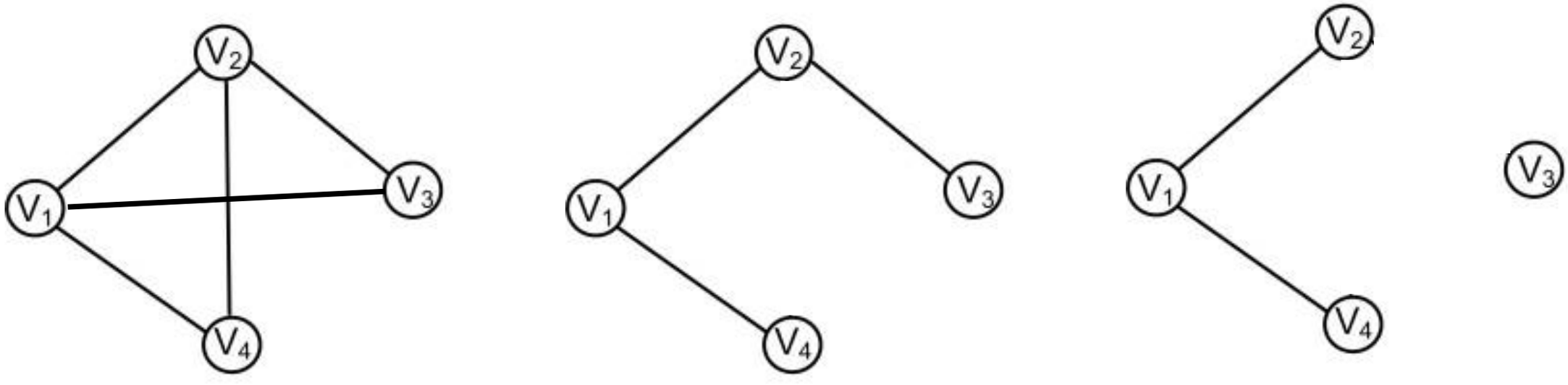


Average path length = 1.8

4. Connectivity

Node connectivity $\kappa(G)$: minimum number of **nodes** whose removal renders the network disconnected

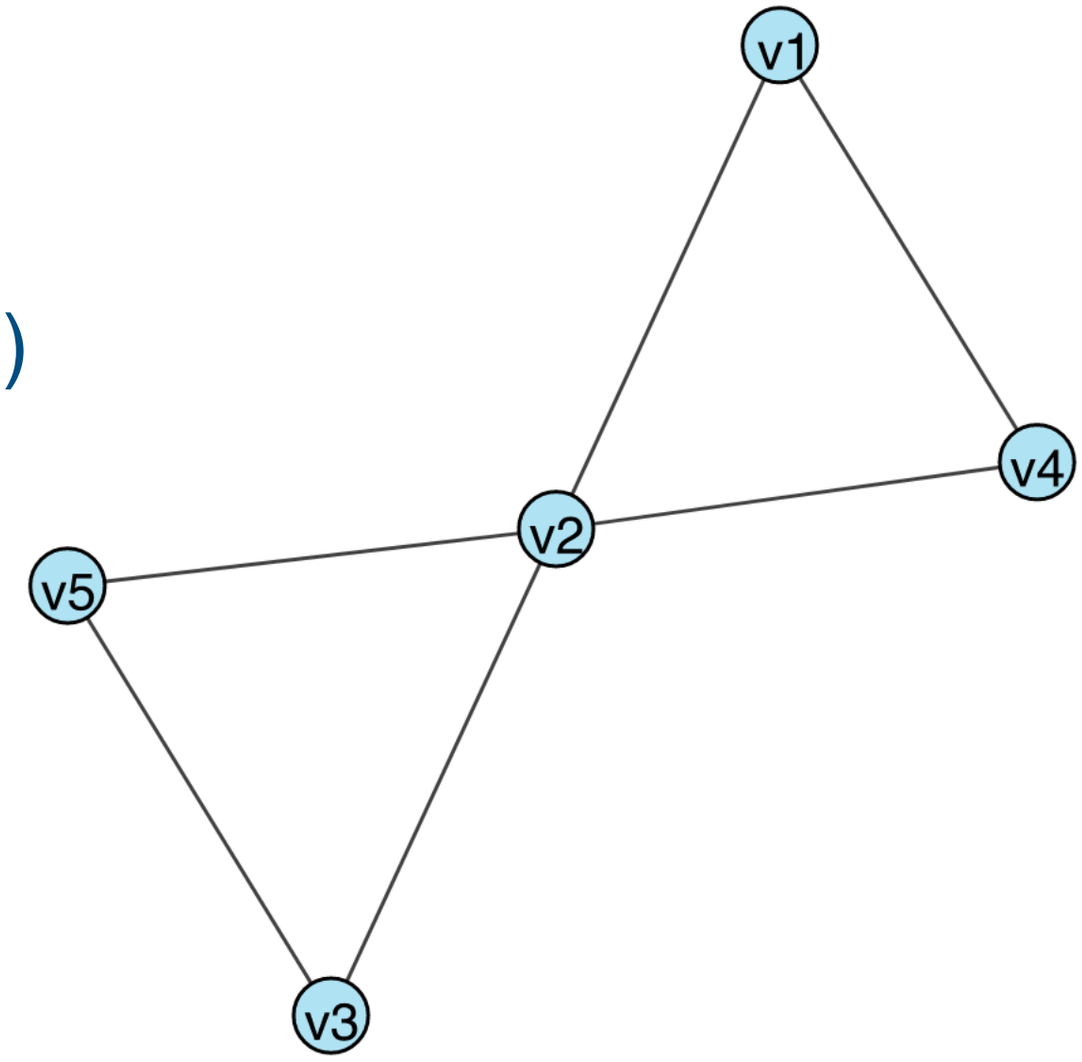
Edge connectivity $\lambda(G)$: minimum number of **edges** whose removal renders the network disconnected



4. Connectivity

$\kappa(G) = 1$; **cut:** v_2

$\lambda(G) = 2$; **bridge:** ((v_2 , v_1) & (v_2 , v_4))



Local connectivity may also be computed for any given pair of vertices

(e.g. v3,v1: v2 and associated edges)

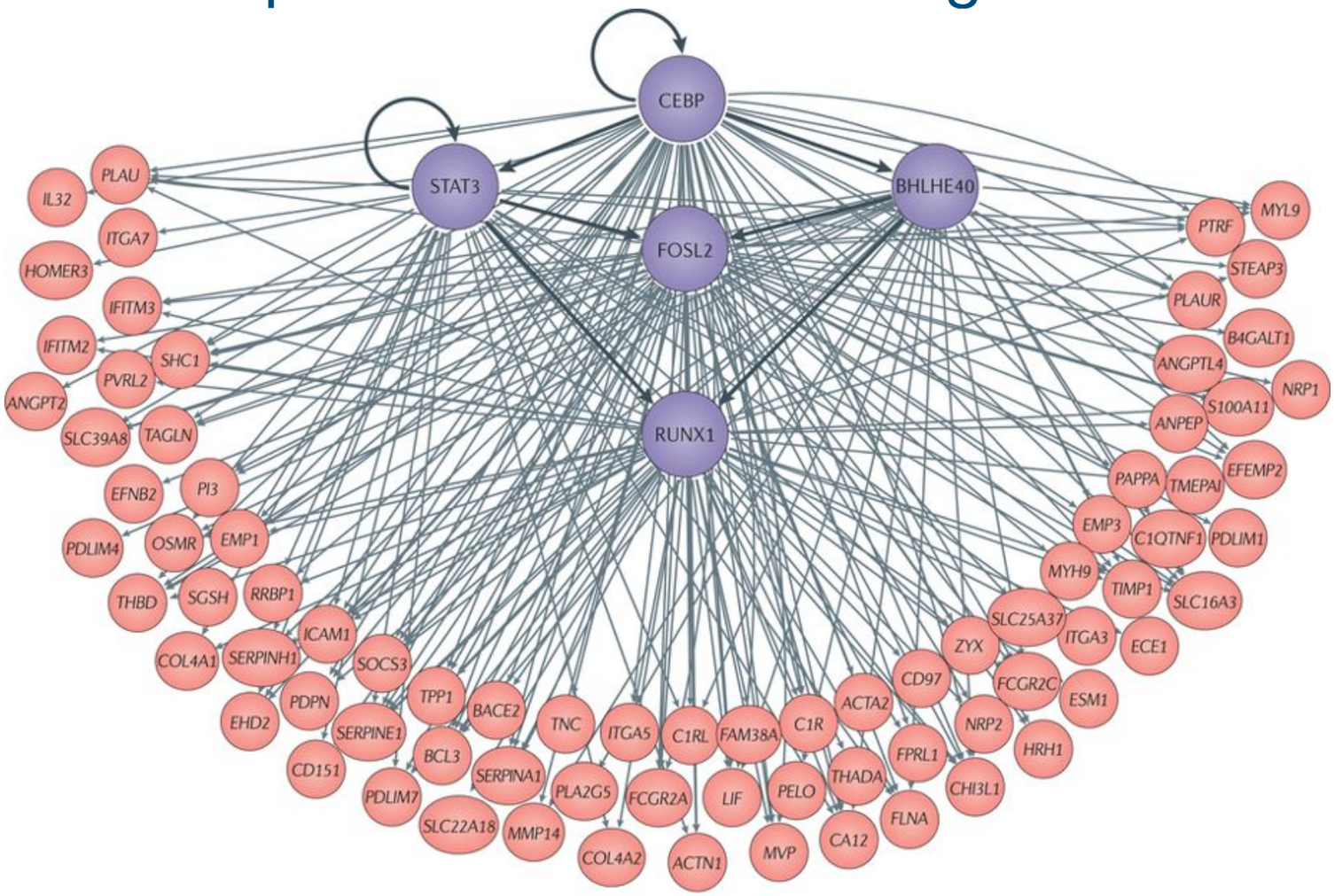
5. Centrality

Indicate the most central nodes in a network

Why look at the central nodes?

Hubs

Example: Transcription Factor Master Regulators



Nature Reviews | Cancer

5. Centrality

Indicate the most central nodes in a network

Central nodes **possibly** most important in the network

There are many different measures of centrality:

- **Degree**
- **Eccentricity**
- **Betweenness**
- *Closeness*
- *Eigenvector*
- *PageRank*
- Katz
- Percolation
- Cross-clique

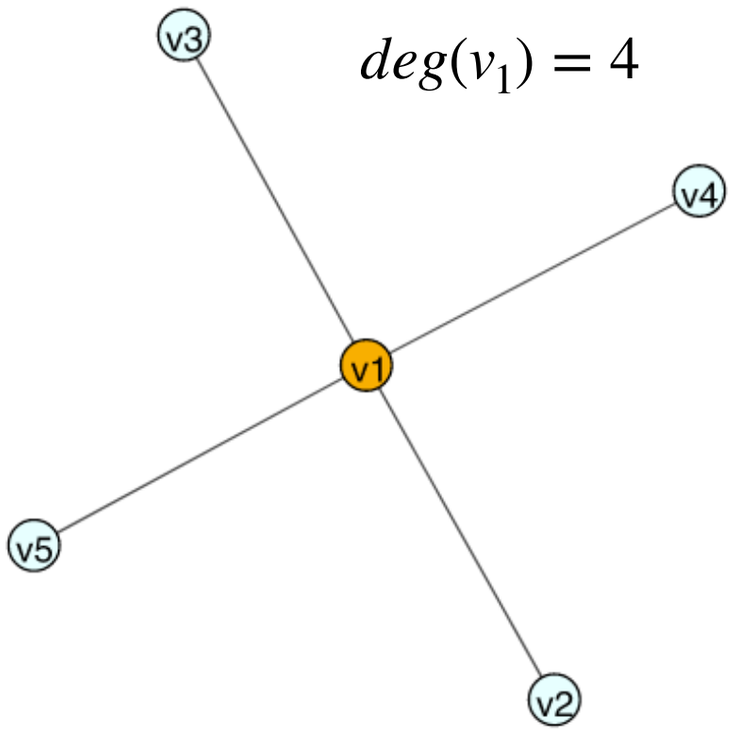
…

4. Centrality: degree centrality

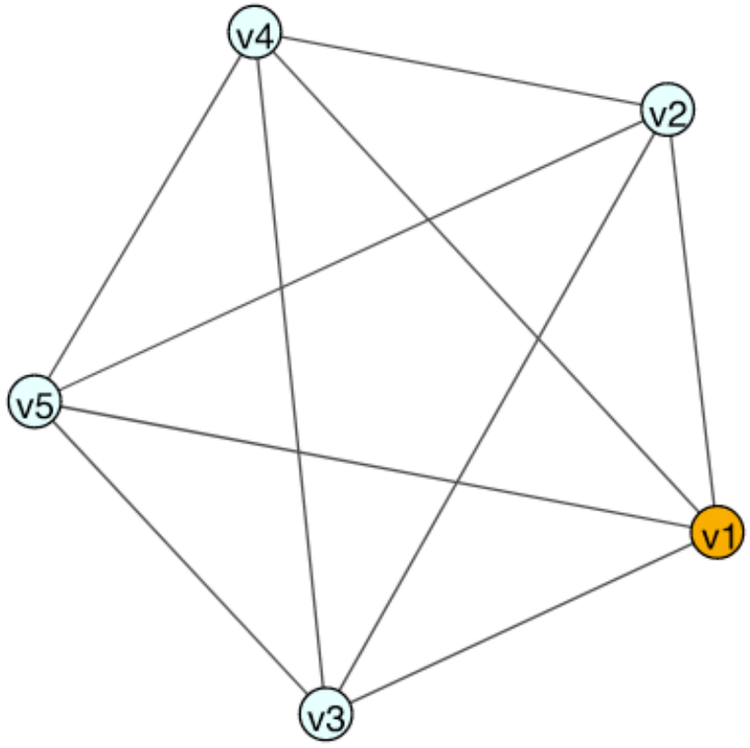
Degree indicates the number of connections with a node

$$d(v) = |N(i)|$$

where *N(i)* is the number of 1st neighbours of a node.



$deg(v_1) = 4$



$deg(v_1) = 4$

4. Centrality: degree centrality

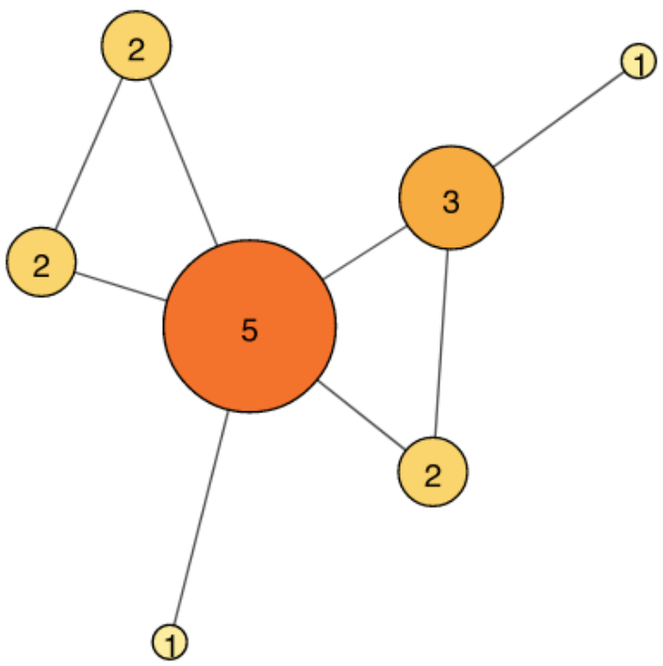
Undirected networks vs directed networks

In-degree vs **Out-degree**

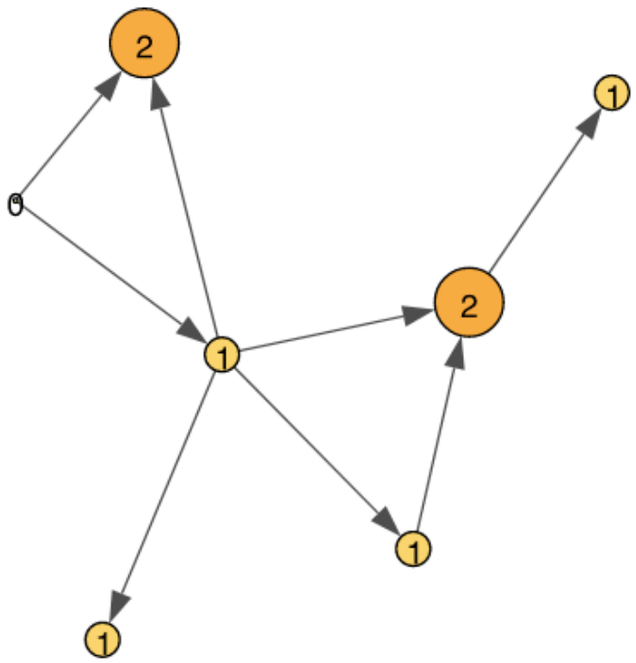
$$C_D(v_i) = \sum_{j=1}^{N} e_{ij}$$

Numbers indicate degree:

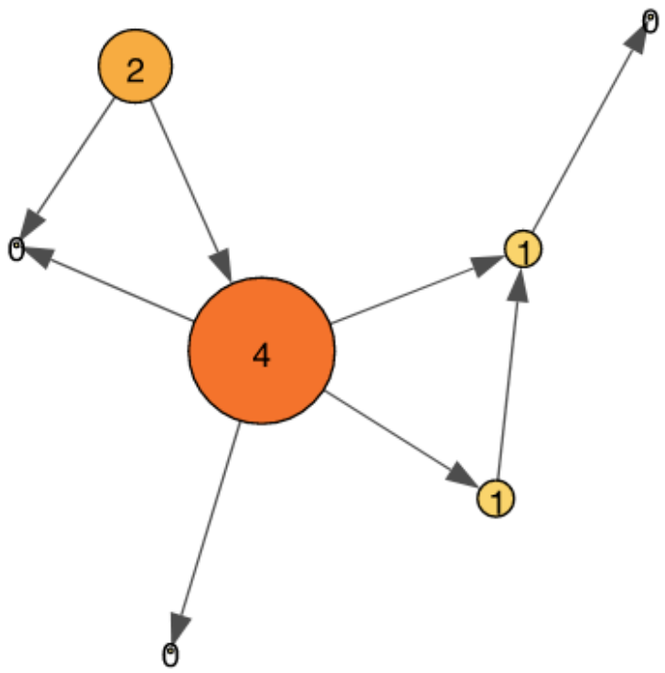
Undirected



In-degree



Out-degree



4. Centrality: degree centrality

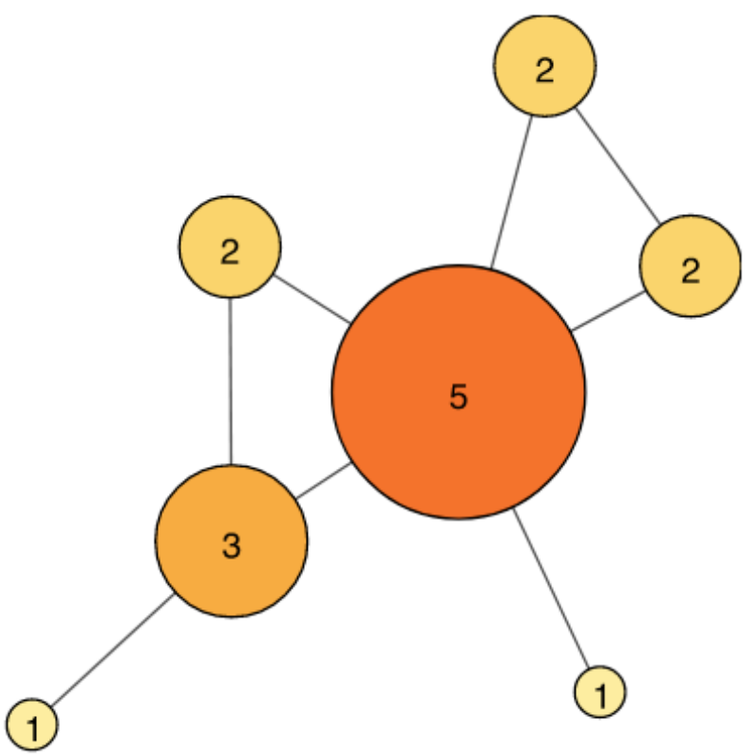
Degree centrality

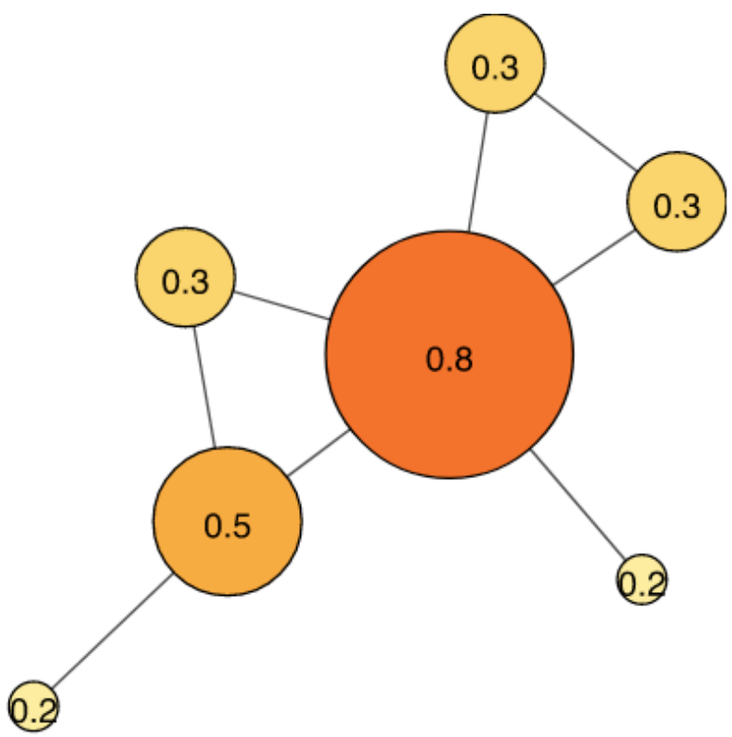
$$C_D(v_i) = \sum_{j=1}^{N} e_{ij}$$

Normalized degree centrality

$$C_D(v_i) = \frac{\sum_{j=1}^{N} e_{ij}}{N-1}$$

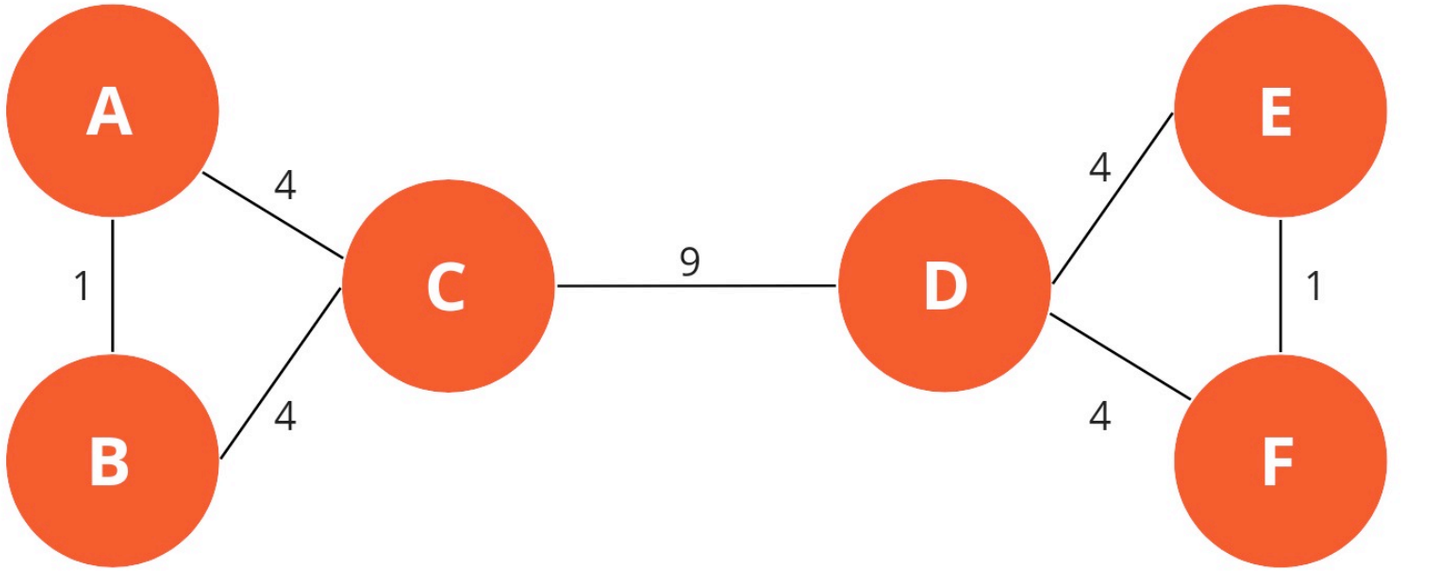
Centrality normalization allows for comparison between networks of different sizes





4. Centrality: betweenness centrality

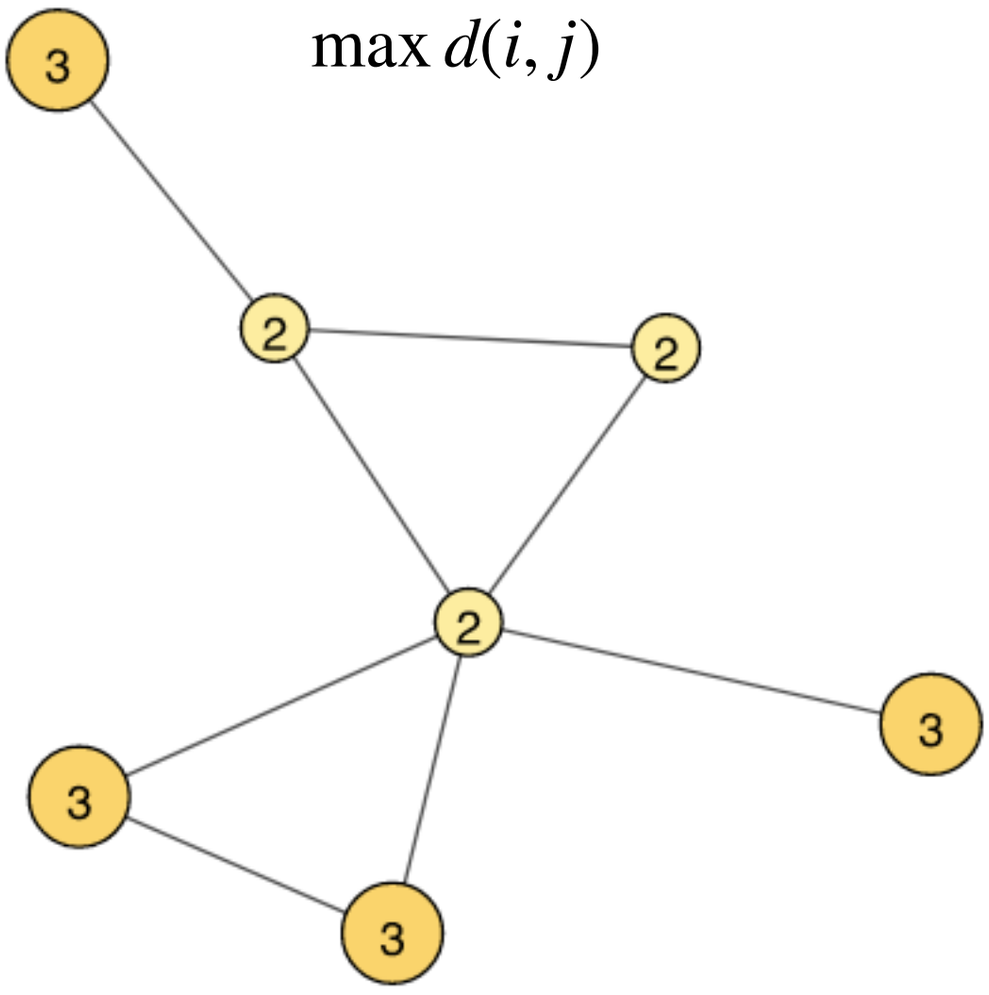
Betweenness considers the number of shortest paths passing through each edge



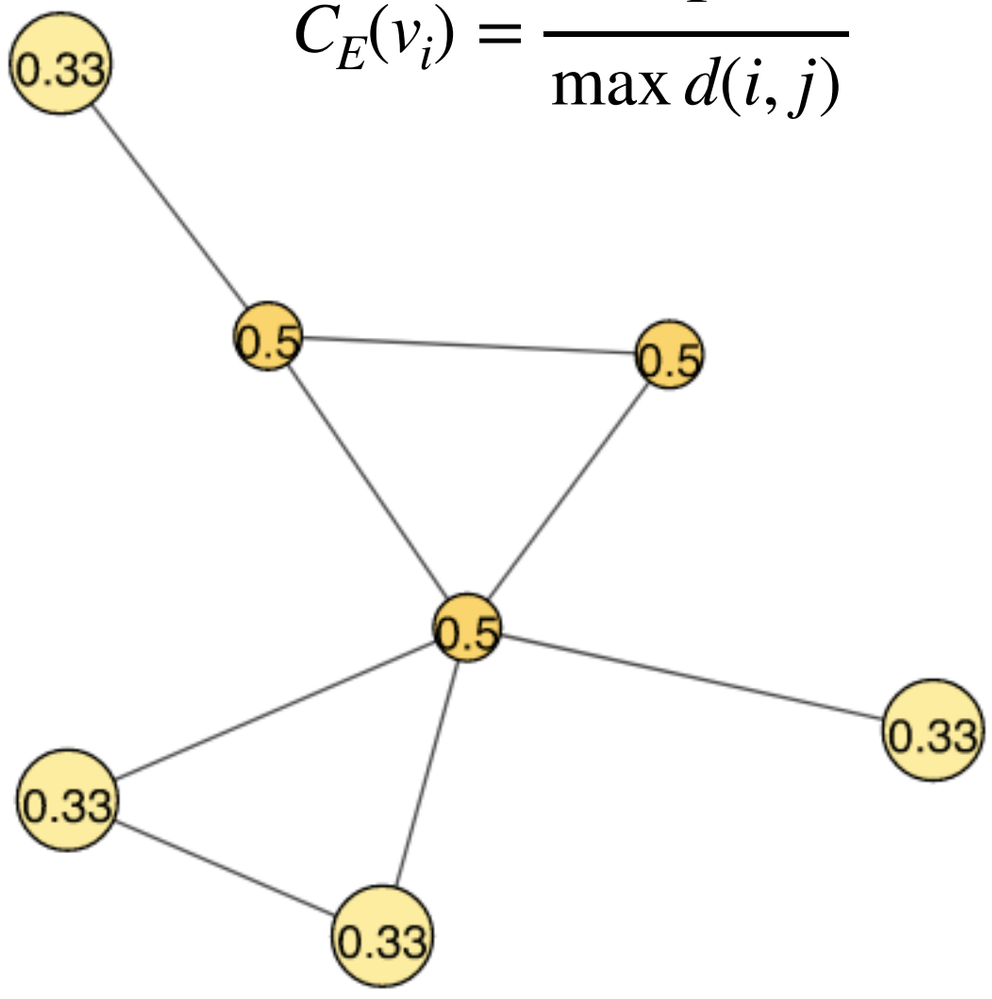
4. Centrality: eccentricity centrality

Eccentricity considers a node's maximum shortest path to all other nodes

$$\max d(i,j)$$



$$C_E(v_i) = \frac{1}{\max d(i,j)}$$



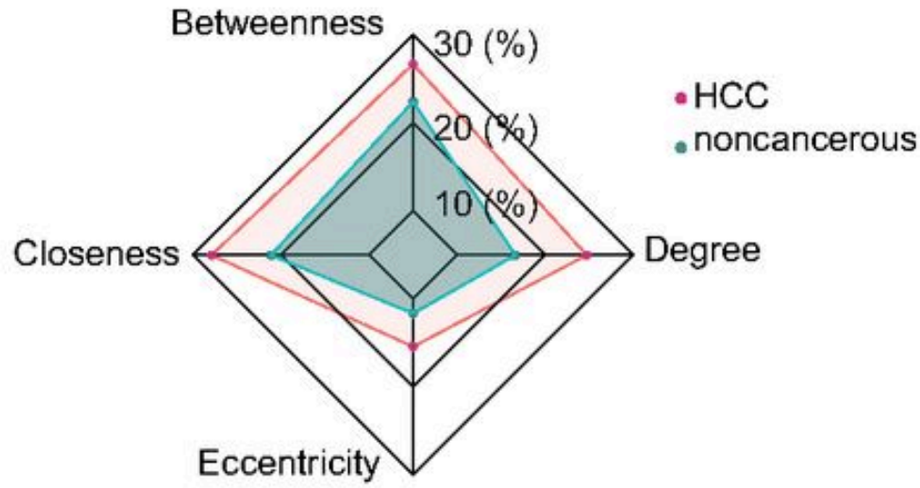
5. Centrality: limitations & influence

Node centrality is used as proxy for **importance**

Should be:

1. Complement with experimental observations
2. Compute multiple metrics and summarise joint observations
3. Compute node **influence,** modifications of centrality

- Accessibility
- Dynamic influence
- Impact
- Expected force



Measure **information transmission** rather than *connectiveness*

6. Clustering coefficient

How likely is it that two connected nodes are part of a highly connected group of nodes?

If node v_1 is connected with v_2 and v_3, it is very likely that v_2 and v_3 are also connected.

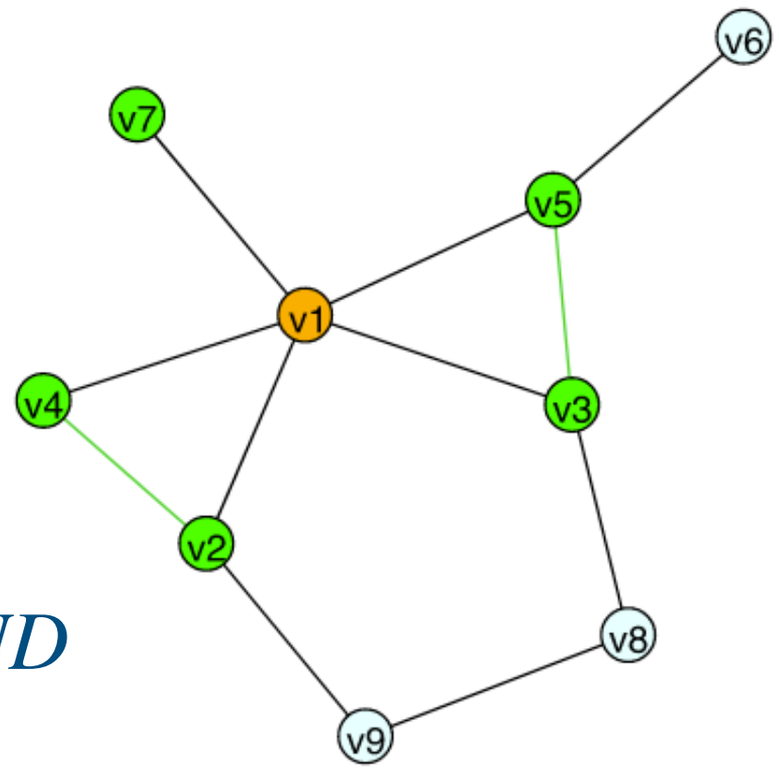
Takes into account degree of a node and the degree of its 1st neighbours

For node v_1

- $deg(v_1) = k = 5$
- n connections between 1st neighbours of $v1 = 2$

$$C_i = \frac{2 \cdot n}{k \cdot (k-1)}$$

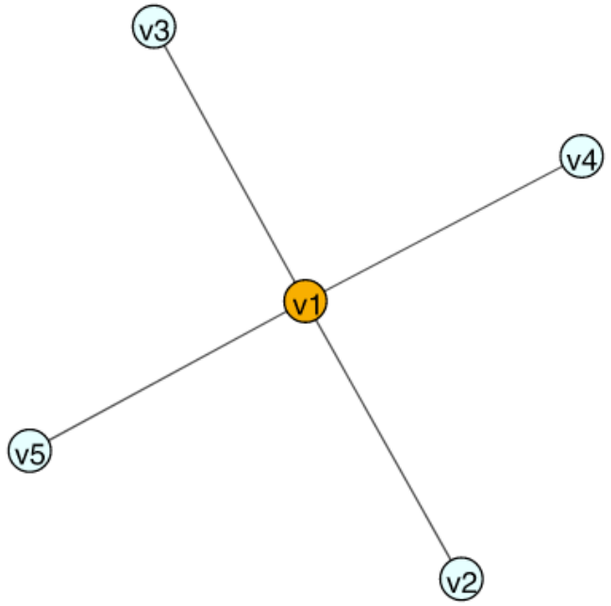
$$C(v_1) = \frac{2 \cdot 2}{5 \cdot 4} = 0.2 \qquad C(v_7) = \frac{2 \cdot 0}{1 \cdot 0} = 0 \text{ or } ND$$



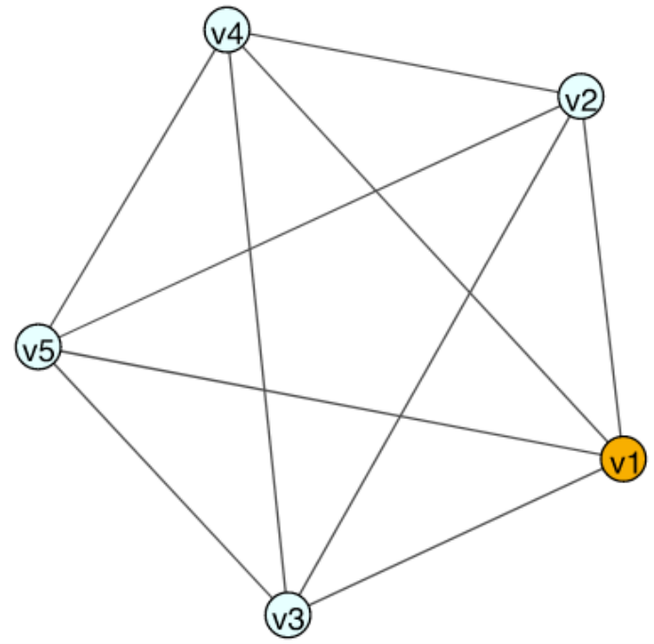
6. Clustering coefficient

$C_i = \frac{2 \cdot n}{k \cdot (k-1)}$ gives the **fraction of possible interconnections** for neighbours of node i

where $\frac{k \cdot (k-1)}{2}$ is the maximum number of triangles through a node



$$0 \leq C_i \leq 1$$

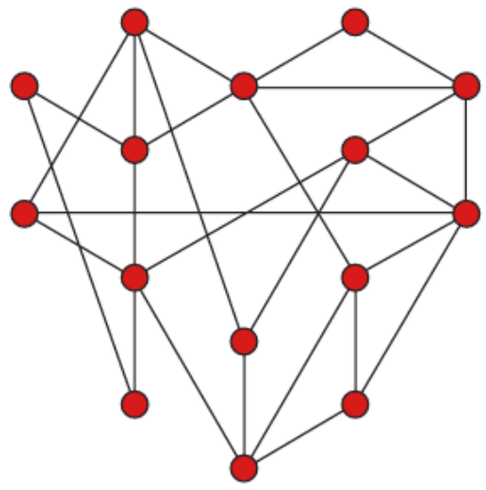


The global clustering coefficient $C(G)$ is simply the average of its clustering coefficients

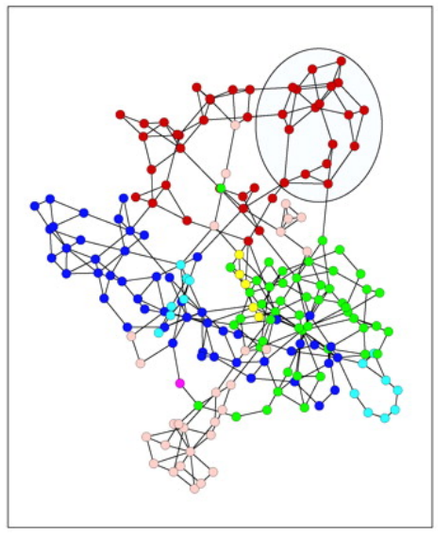
What distinguishes biological networks from random?

Do metabolic networks display different network properties from random networks?

Random network



Metabolic network

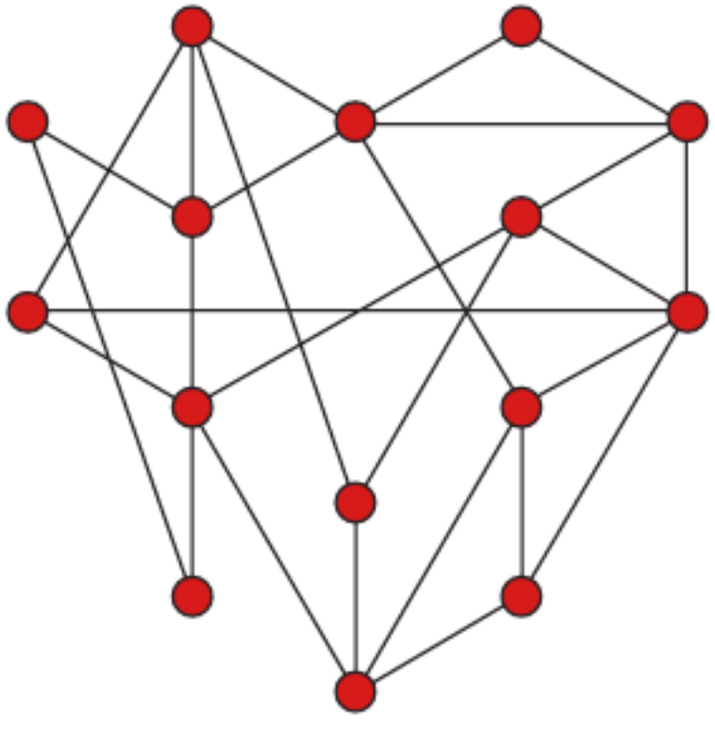


Barabasi 2004
Jeong 2000
Ravasz 2002

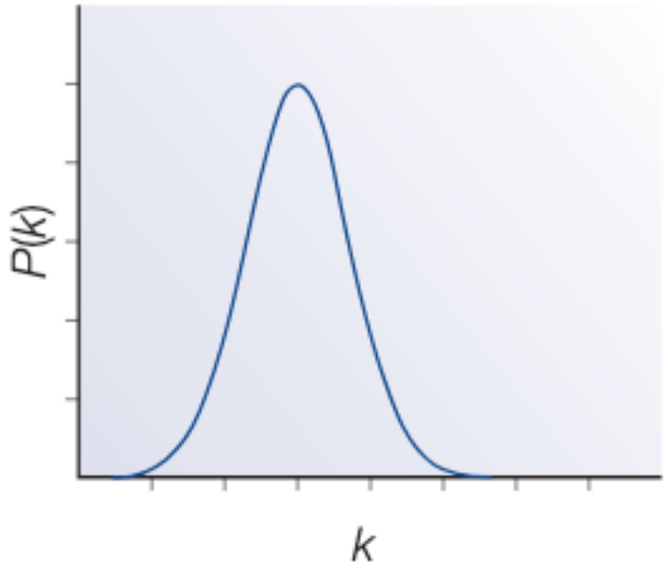
7. Degree and clustering coefficient distribution

Degree distributions allow us to compare network organization

Random network
(e.g. Erdös-Rényi model)



Poisson degree distribution
shows no highly connected nodes

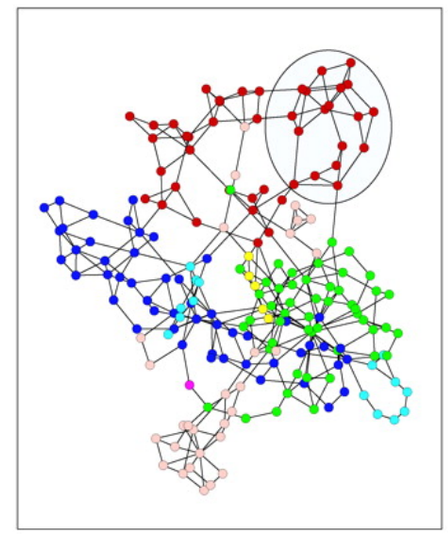


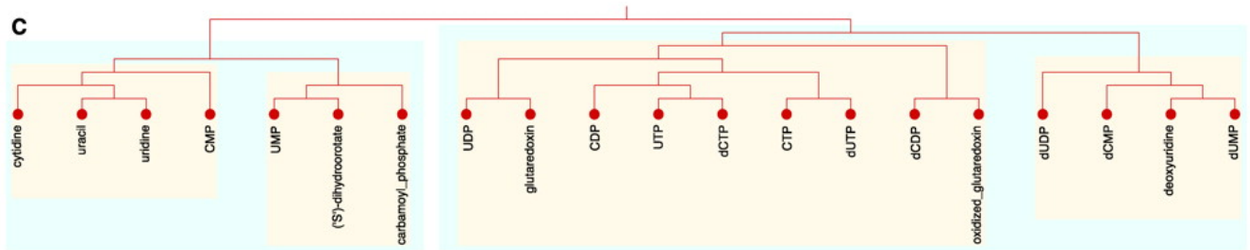
Most nodes have near <*k*>

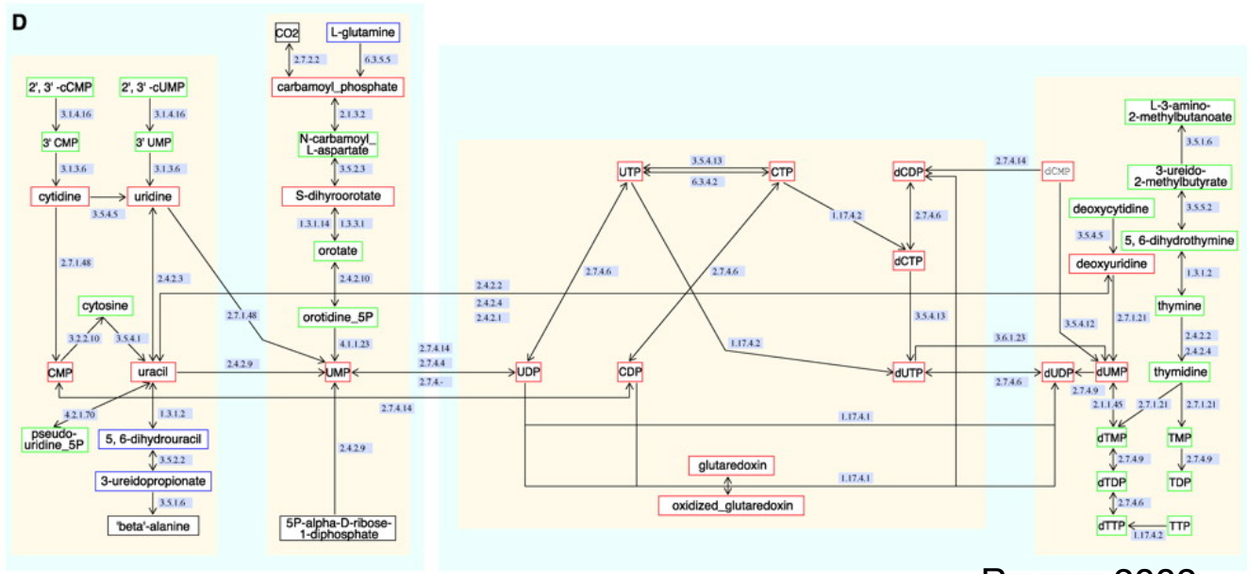
Metabolic networks show hierarchical topology

Metabolic networks of 43 organisms are organised into **small, tightly connected modules**

Their combination shows a hierarchical structure







Ravasz 2002

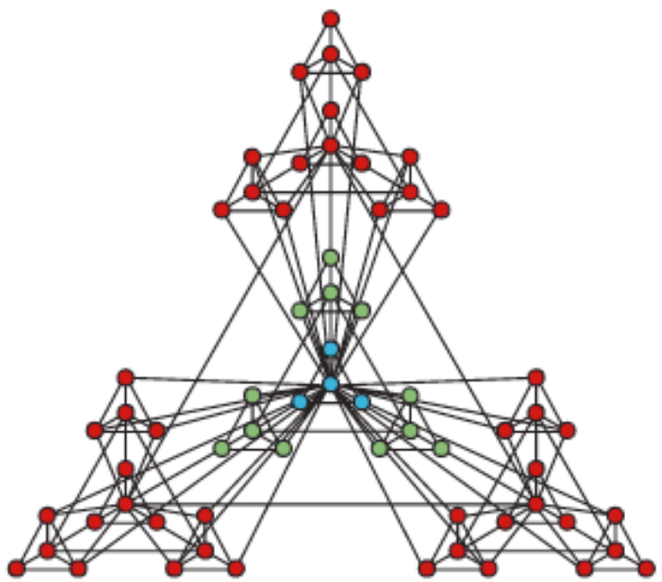
7. Degree distribution

Biological networks do not follow topology features of random networks.

Analysis of metabolic networks of 43 organisms shows common patterns

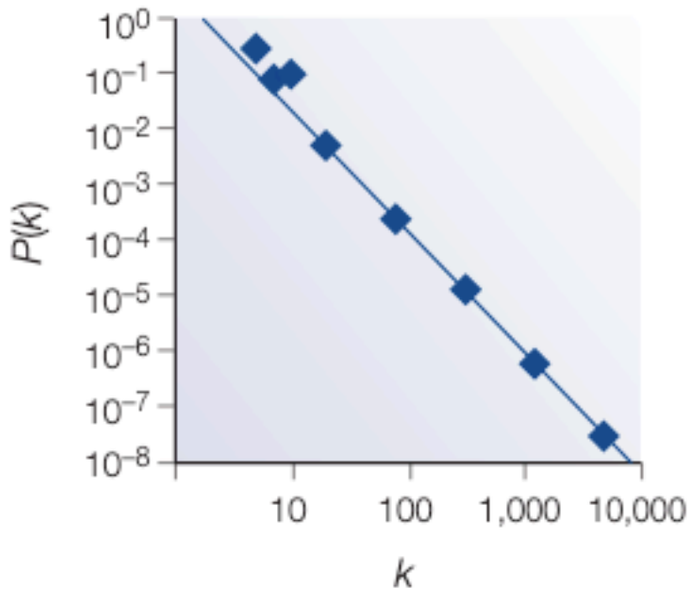
Biological networks tend to display high robustness to node failure: removal of <80% nodes still retains paths between any two nodes

Hierarchical network



Degree distribution

shows many with low degrees
a few highly connected nodes



Barabasi 2004
Jeong 2000

7. Degree and clustering coefficient distribution

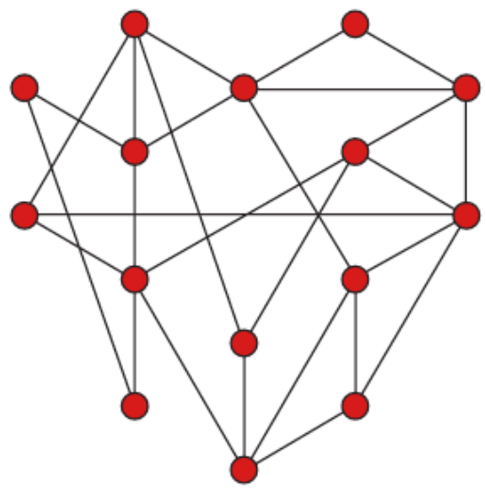
$C(k)$ shows no relationship with *k* in random networks: no modular organisation

$C(k) = k^{-1}$ in hierarchical networks

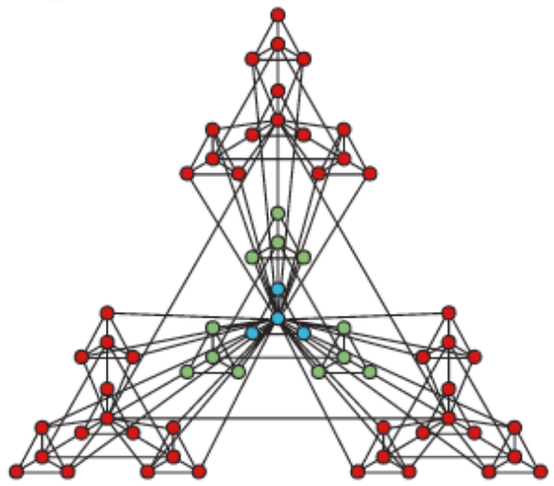
Sparsely connected nodes are part of highly modular areas

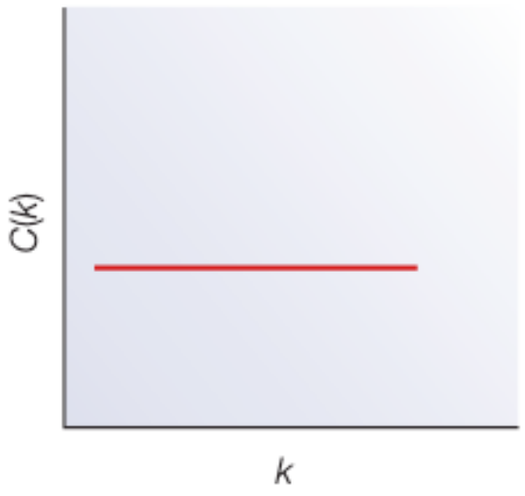
Communication between highly clustered neighbourhoods maintained by a few hubs

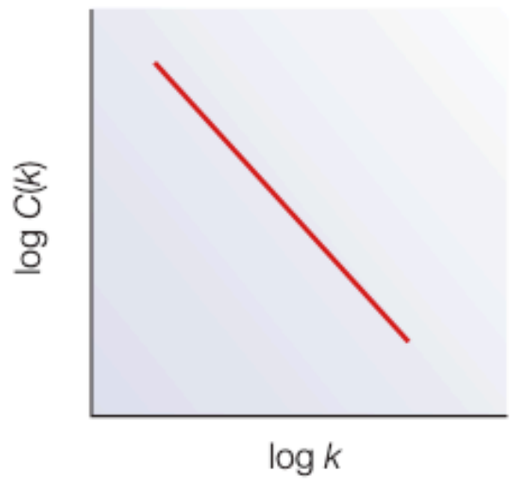
Random network



Hierarchical network







Barabasi 2004
Jeong 2000

7. Small world

Any two nodes can be connected in a small number of steps.

This is a property seen in **random networks** where the mean path length

$$l(G) \approx logN$$ for a network of size *N*

Many biological networks show **ultra-small world** properties:

$$l(G) \approx log(logN)$$

Highly central hubs tend **not** to be connected in biological networks: they are **disassortative**

(social networks: **assortative**)

Barabasi 2004

Additional reading

- Analysis of Biological Networks - General introduction into biological networks, network notation, and analysis, including graph theory.
- Using graph theory to analyze biological networks - overview of the usage of graph theory in biological network analysis
- Survival of the sparsest: robust gene networks are parsimonious - analysis of network complexity and robustness.
- Network biology: understanding the cell's functional organization - Overview of key concepts in biological network structure
- Graph Theory and Networks in Biology - extended perspective on how graph analysis is applied in biology
- Modularity and community structure in networks

Additional references displayed as hyperlinks in each figure.