

Network Inference and Properties

Rui Benfeitas

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

rui.benfeitas@scilifelab.se



SciLifeLab



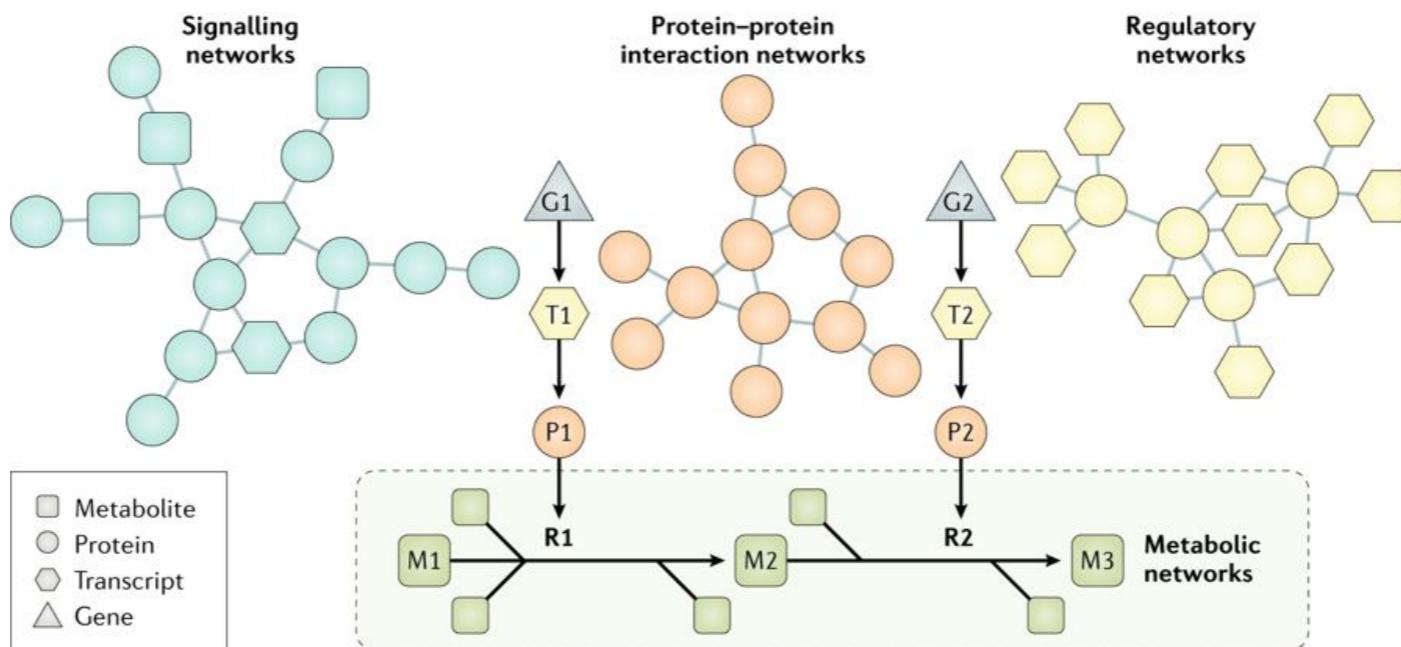
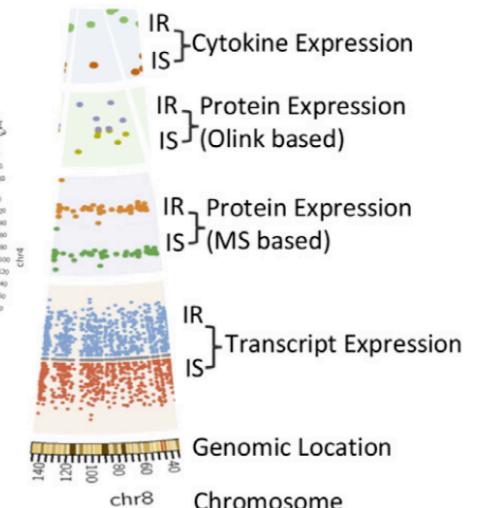
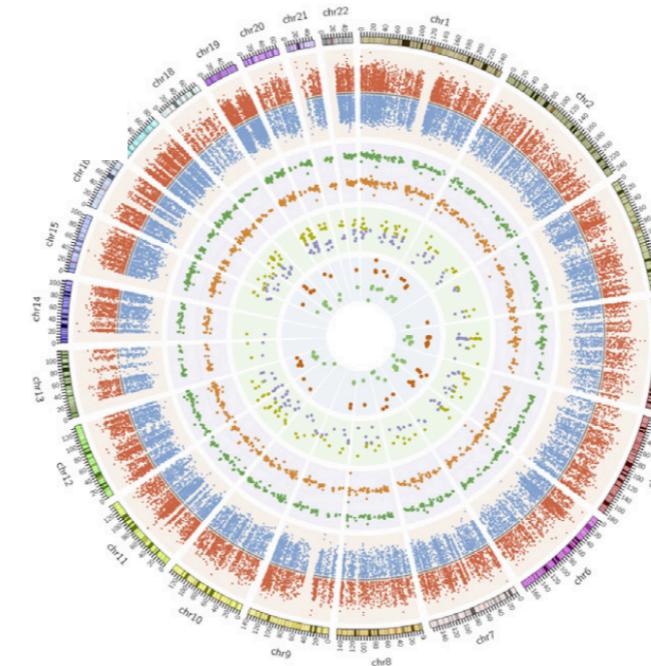
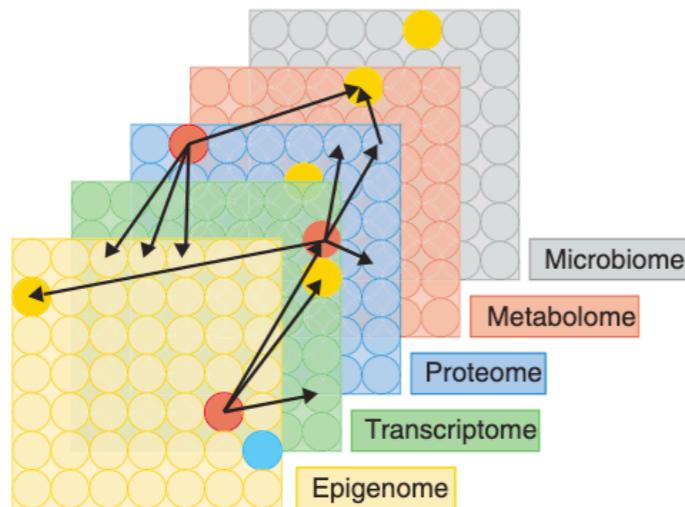
Overview

1. Introduction to network analysis
2. Terminology
- 3. Network inference**
- 4. Key network properties**
- 5. Community analysis**
- 6. Visualization**
- 7. Workshop**

Biological network inference

Building networks

How to go from raw data to a graph tractable format?



Hasin 2017
Piening 2018
Mardinoglu 2018

Important considerations

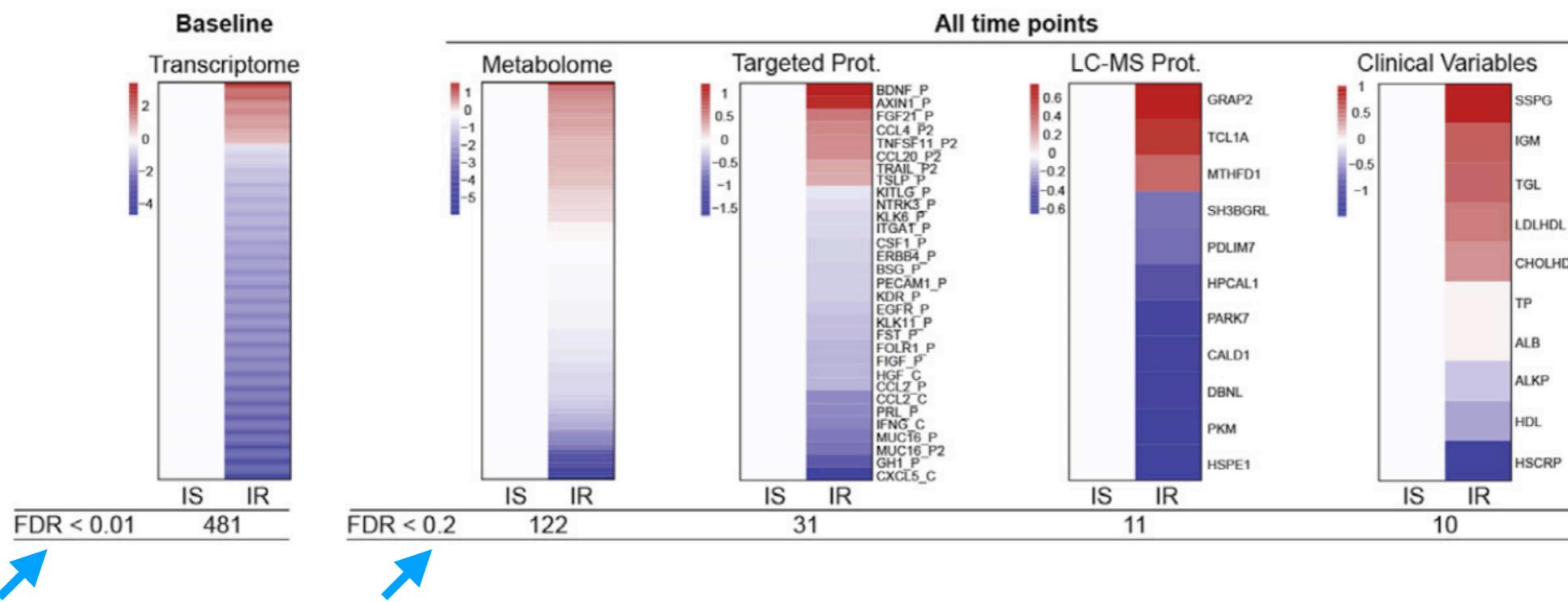
Standard statistics should not be ignored

Importance of power and sample size

Statistical power and significance in high throughput studies

Batch effect correction should be applied if needed, but avoided whenever possible

- Signal loss
- Overconfidence
- Alternative: include batch as a covariate
- Non-parametric analyses
- Limma's [removeBatchEffects\(\)](#), [NormalizerDE\(\)](#), sva's [ComBat\(\)](#)

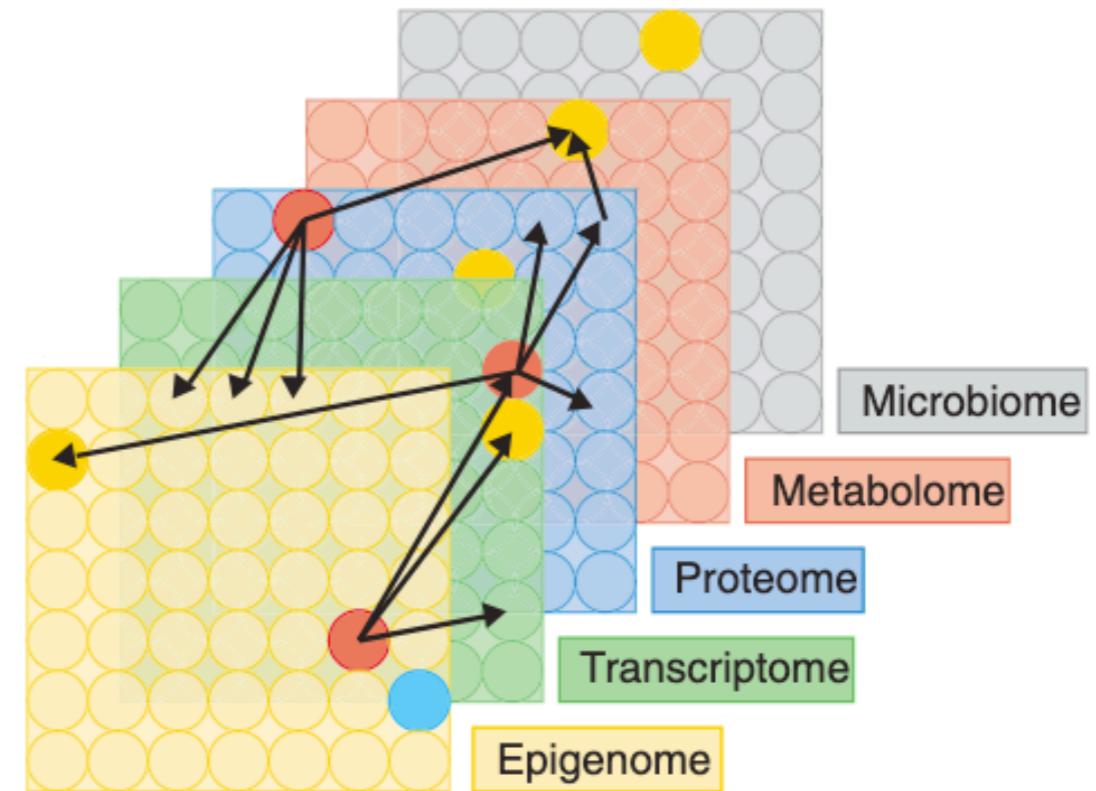


Interomic vs Intraomic networks

Networks may be build for individual omics or for their integration

Should I even integrate different omics? What is my biological question?

- Do I want to analyse vertical relationships between features?
- Why integrate omics with different coverage such as transcriptomic and proteomic data
- Do I want to extract functional properties?
- Am I predicting biomarkers?

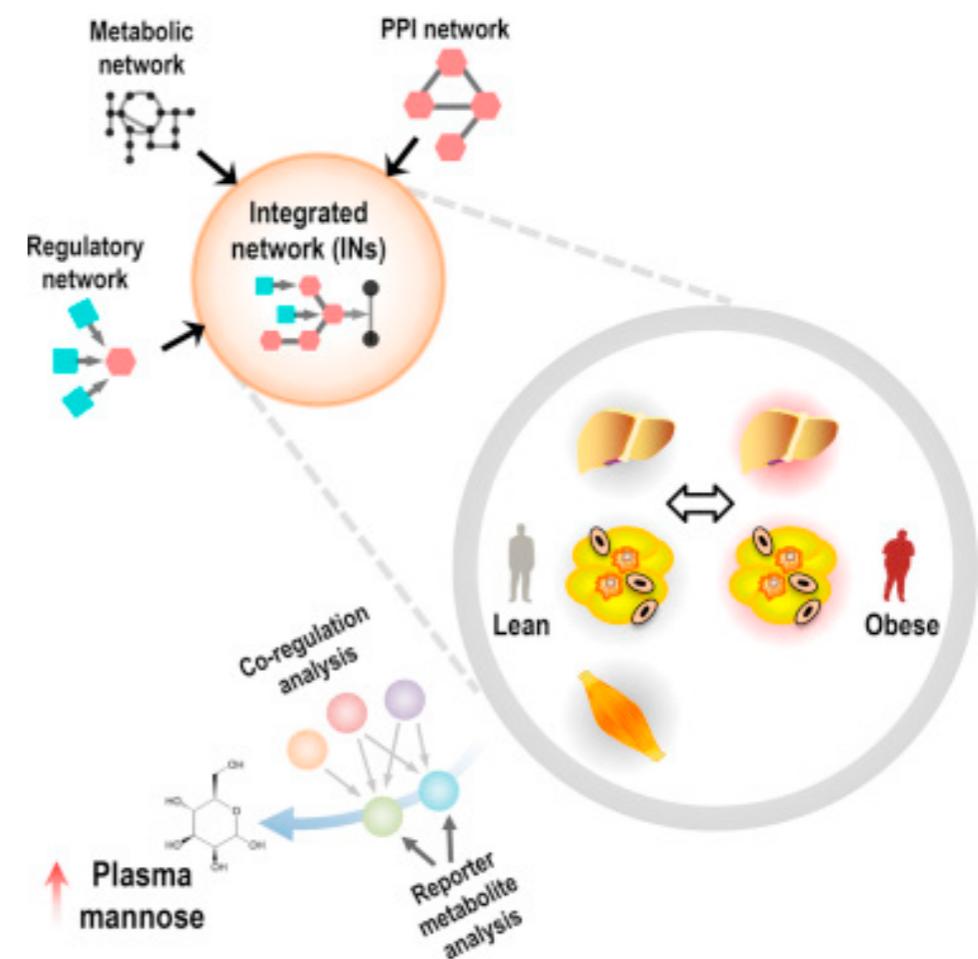


Interomic vs Intraomic networks

Multi-modal (k -partite) networks may be generated from different sources

- Transcription-factor - Gene (DNAseq)
- Gene-gene (Co-expression, PPI, GEMs)
- Gene-metabolite (GEM)
- Metabolite-metabolite (GEM)

Integrated Networks



Different approaches for network inference

- | | |
|---|---------------------------------------|
| 1. Feature association | No prior graph structure |
| 2. K-nearest neighbour graph (k-NNG) construction | |
| 3. Pathway-based | Based on available information |
| 4. Genome-scale metabolic models | |
| 5. Network deconvolution | Filter indirect effects |

1. Association analysis

Balanced dataset, standard pre-processing of each omics needed

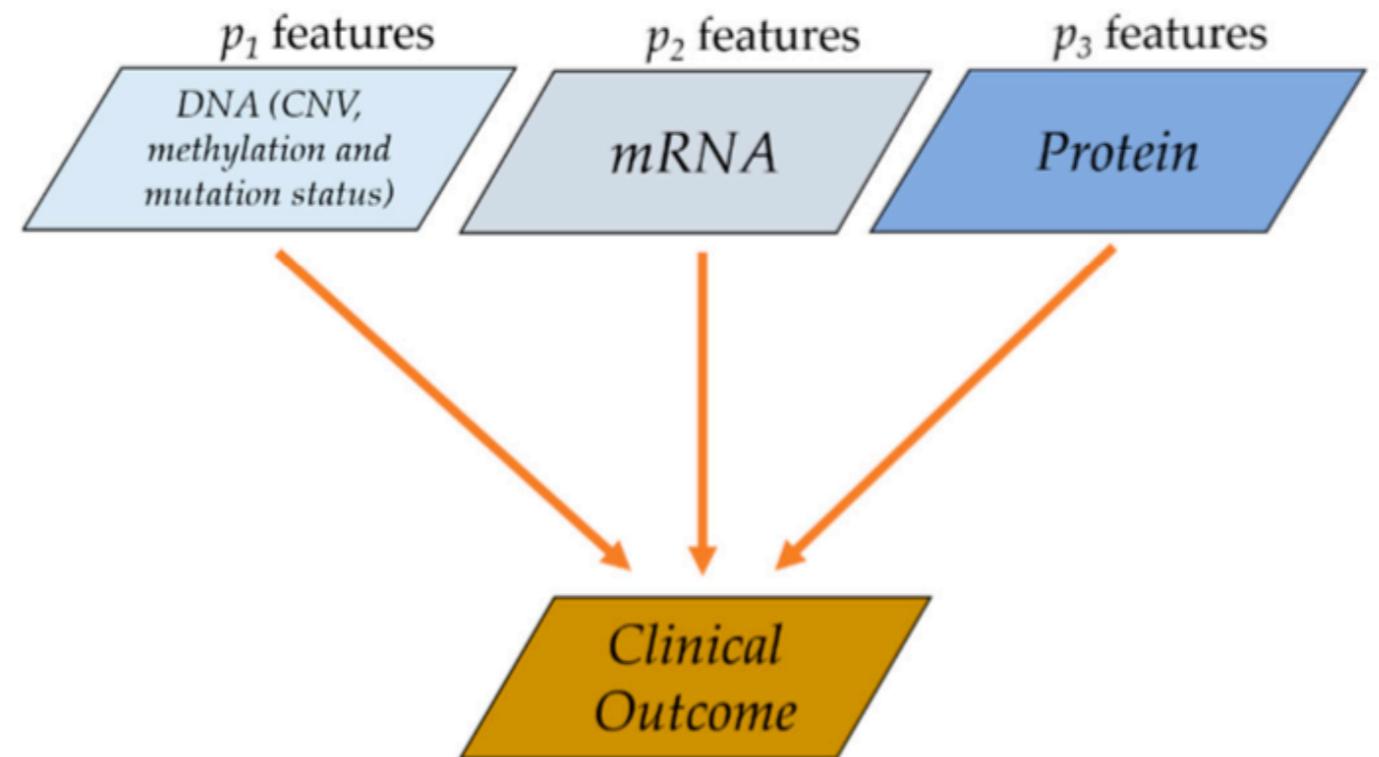
Normalization *may* be needed to make omic datasets comparable (e.g. standardization)

- depending on the following steps

Common approach: compute correlations between different features

- Spearman
- Pearson

Extend known associations



1. Association analysis

Easy to interpret

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

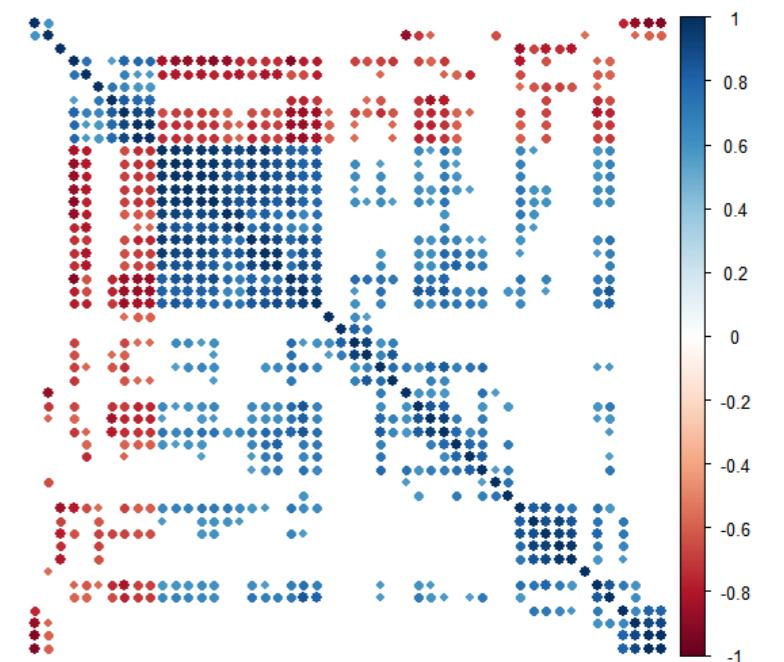
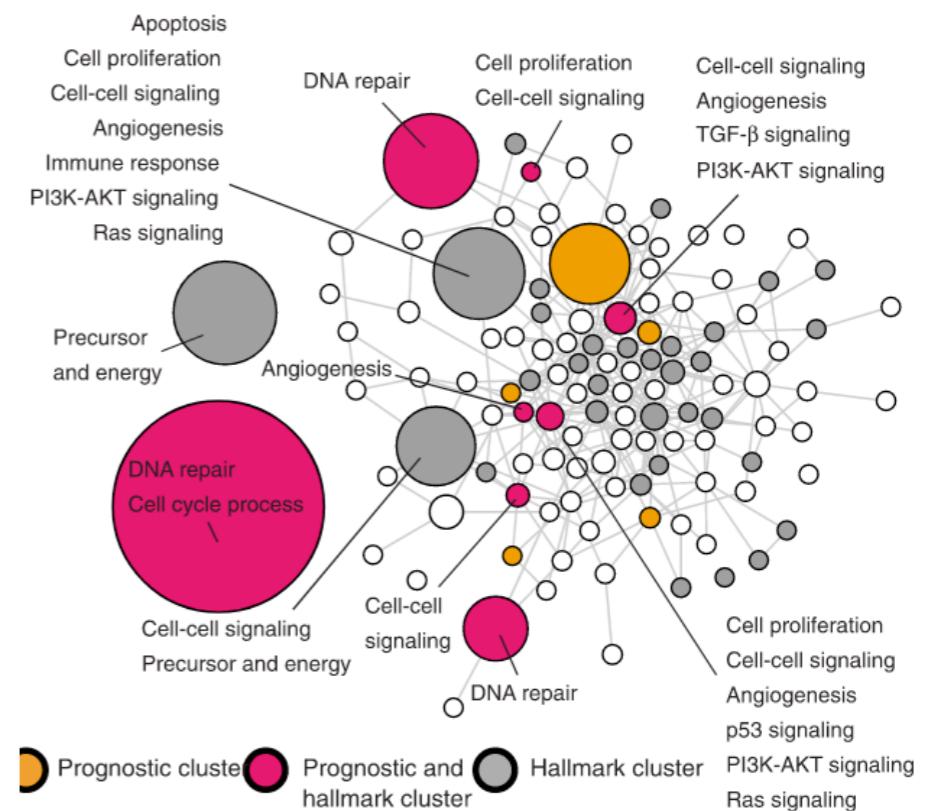
Unbalanced networks

Prone to type I errors

Being conservative in the statistic methods is essential

- FDR vs Bonferroni
- LFC cutoff

Need adjustment to possible confounding factors



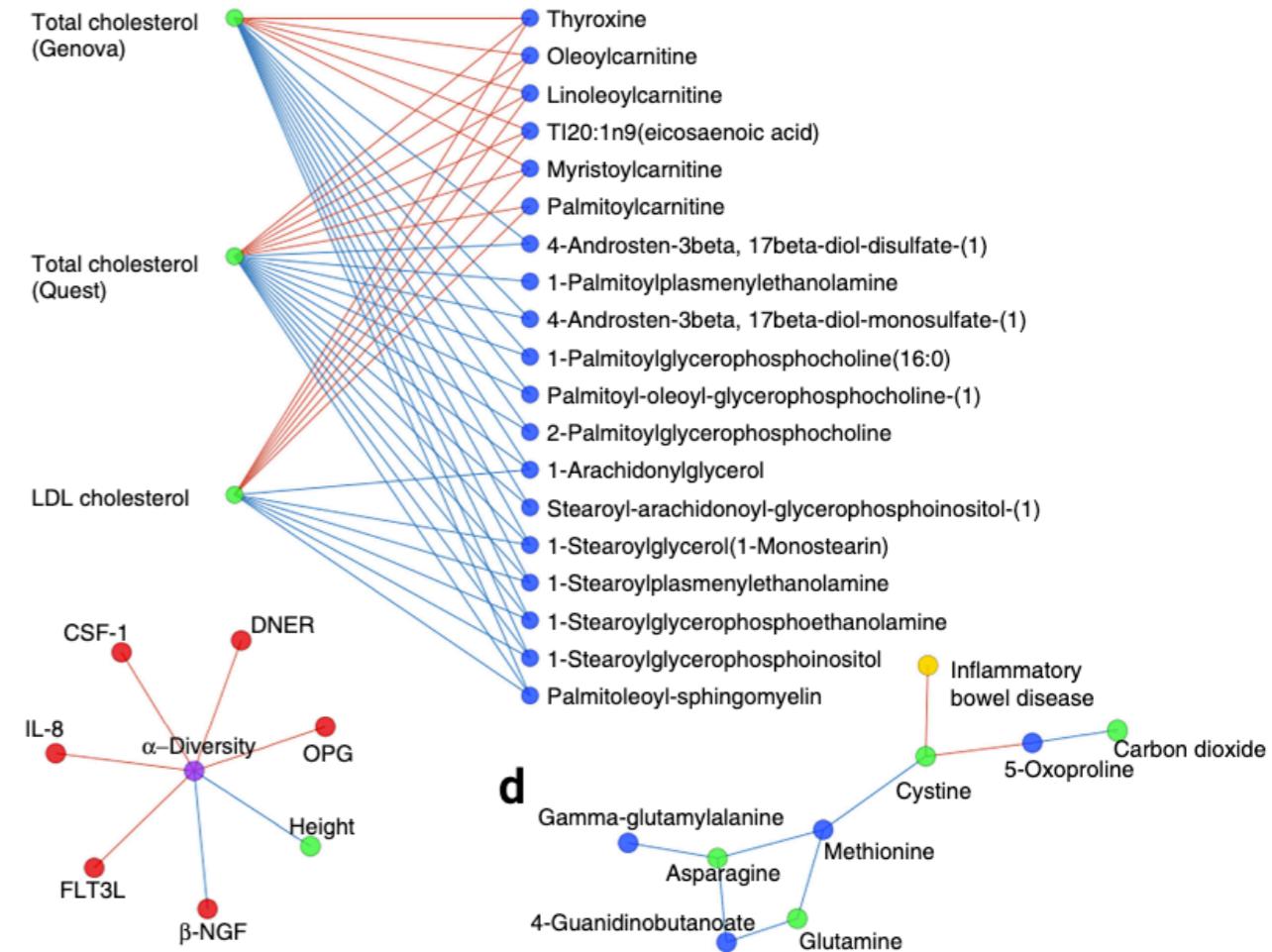
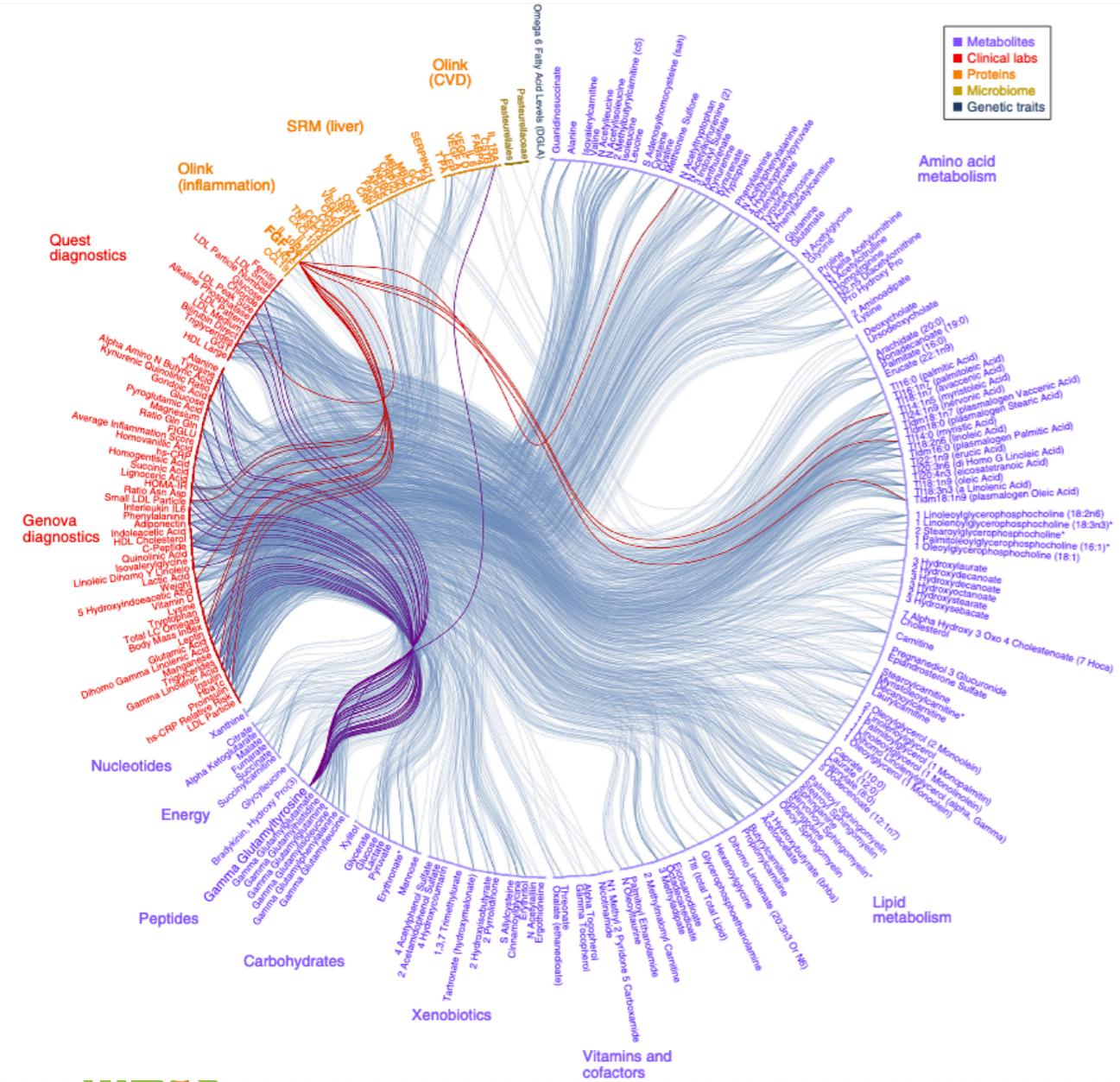
1. Association analysis

Adjusting for confounding factors: partial correlation analysis

Still considers linear relationship between variables

Below:

- gender and age are known confounding factors
- feature regression on confounding factors, followed by correlation on the residuals of each model

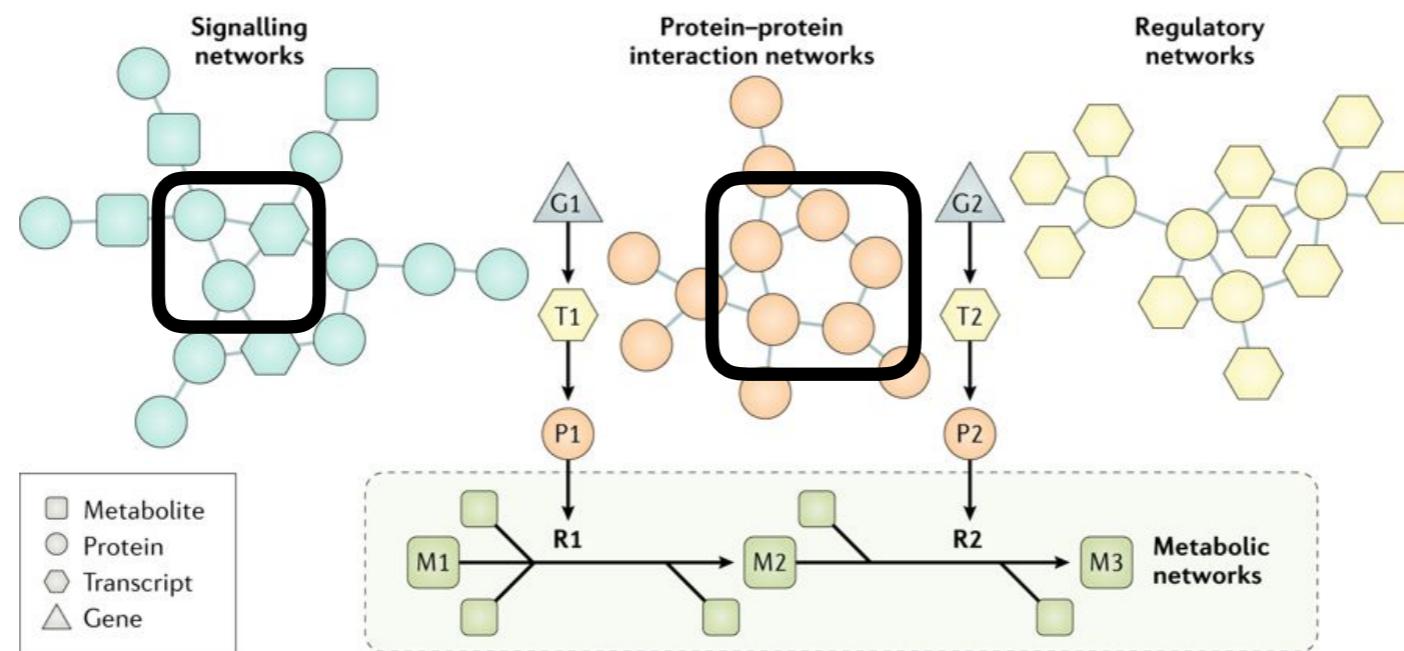


Computationally expensive representations may be simplified

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



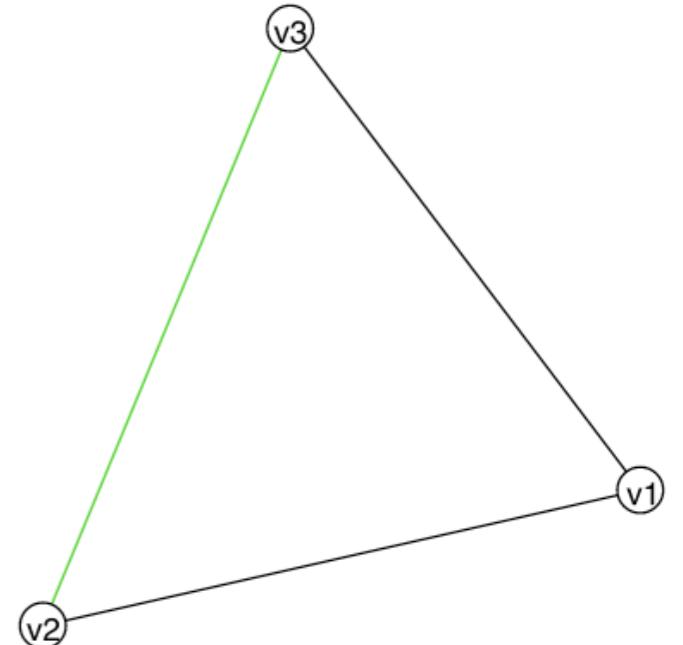
Clarke 2011
Krywinski 2013
Sham 2014
Nygaard 2016
Piening 2018
other refs as links

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

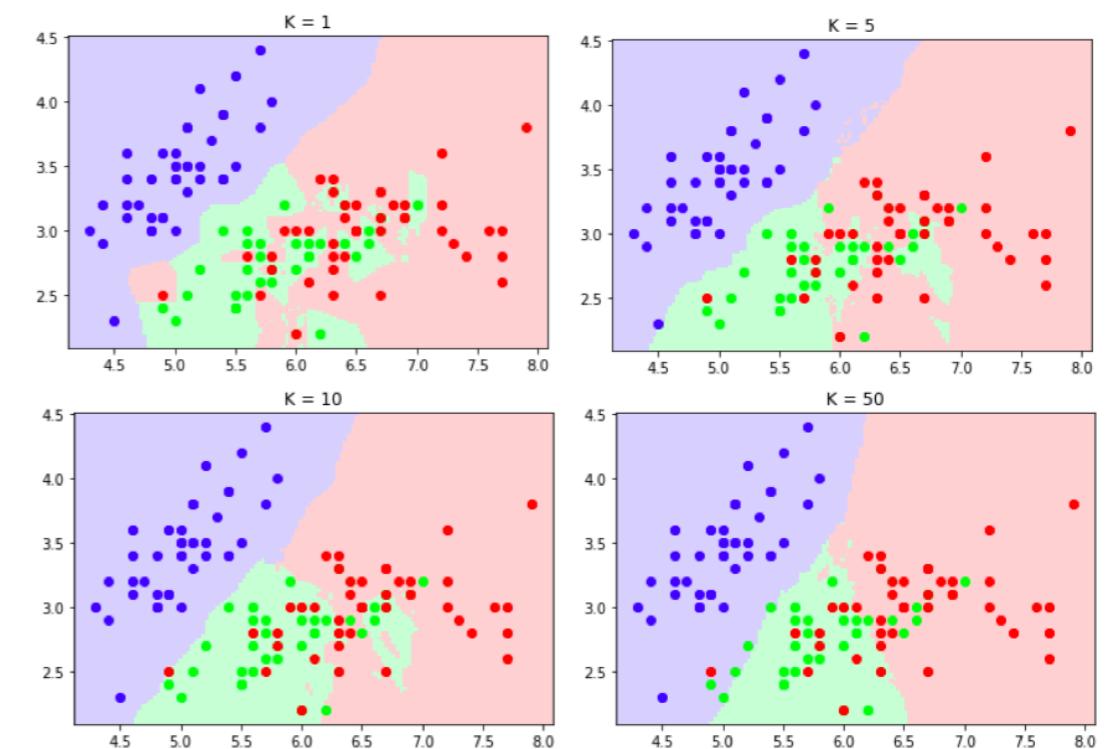
2. k -nearest neighbour graph

High k is smooth, but biased (underfitting)

Low k is accurate, but noisy (overfitting)

Optimum k may be identified:

- k -fold cross validation (sup. learning)
- $k = \sqrt{n}$ (unsupervised)
- ad-hoc



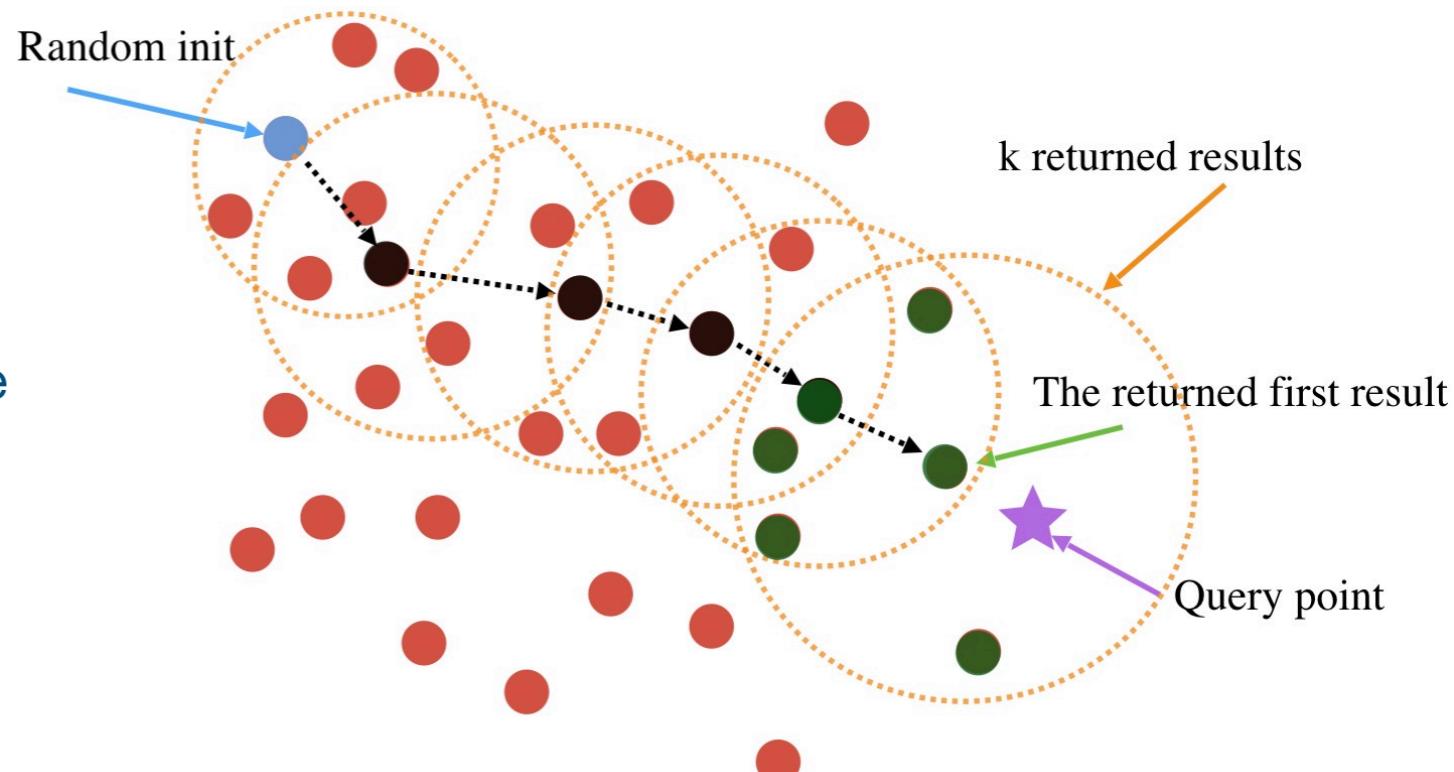
2. k -nearest neighbour graph

Brute force vs NN-Descent

The neighbour of my neighbour is likely to also be my neighbour

Given a query point B :

1. Start with random node v ;
2. For each v , find the k nearest objects;
3. Compare its neighbours' neighbours
4. Repeat until no further improvement is possible
5. Improve local neighbourhood



3. Knowledge-based graph creation

Database-derived

- PPI
- TFRN
- Metabolic Atlas
- ...

Many reference databases

KEGG

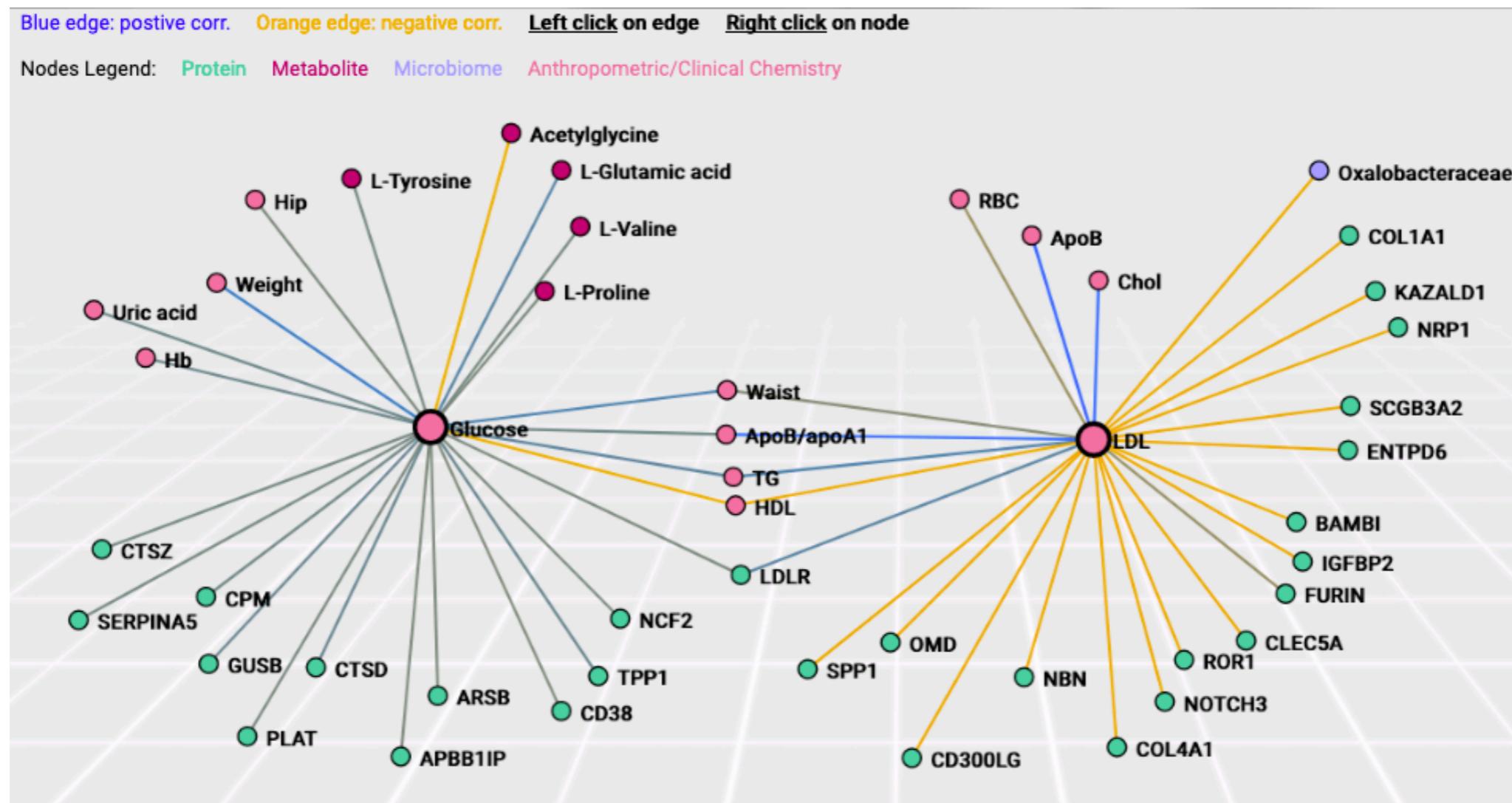
Reactome

WikiPathways

MSigDB

3. Knowledge-based graph creation

Multi-omic biological networks



3. Knowledge-based graph creation

NAR December 2018: 1613 databases



You are here: [NAR Journal Home](#) » [Database Summary Paper Categories](#)

NAR Database Summary Paper Category List

[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](#)

► [Compilation Paper](#)
► [Category List](#)
► [Alphabetical List](#)
► [Category/Paper List](#)
► [Search Summary Papers](#)

Oxford University Press is not responsible for the content of external internet sites

Collection of databases on our website

3. Knowledge-based graph creation

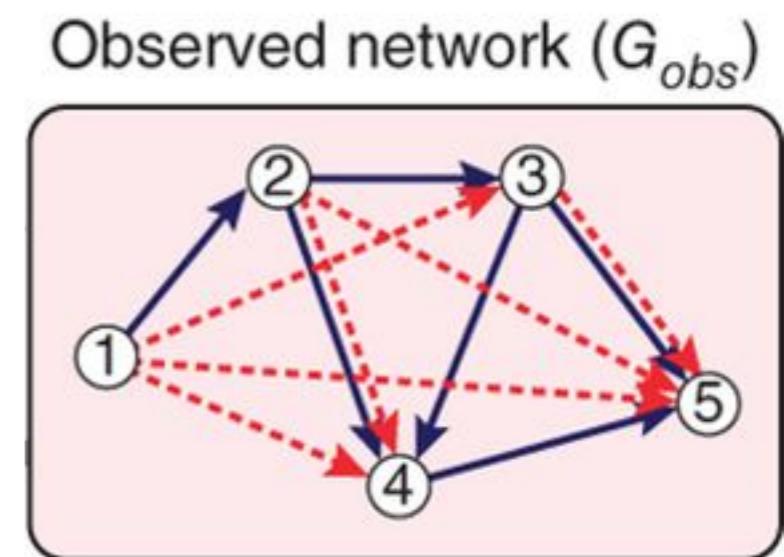
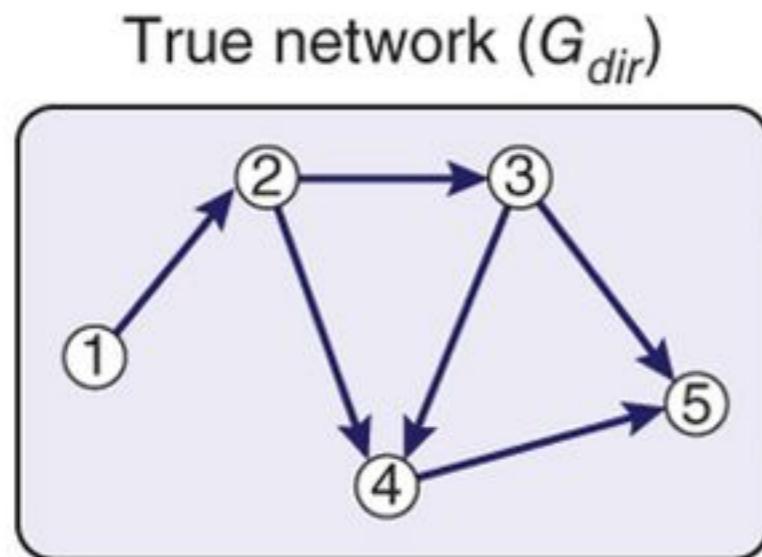
Little overlap among reference pathways



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?
- Simply consider interactions based on physical presence considering the reference networks?

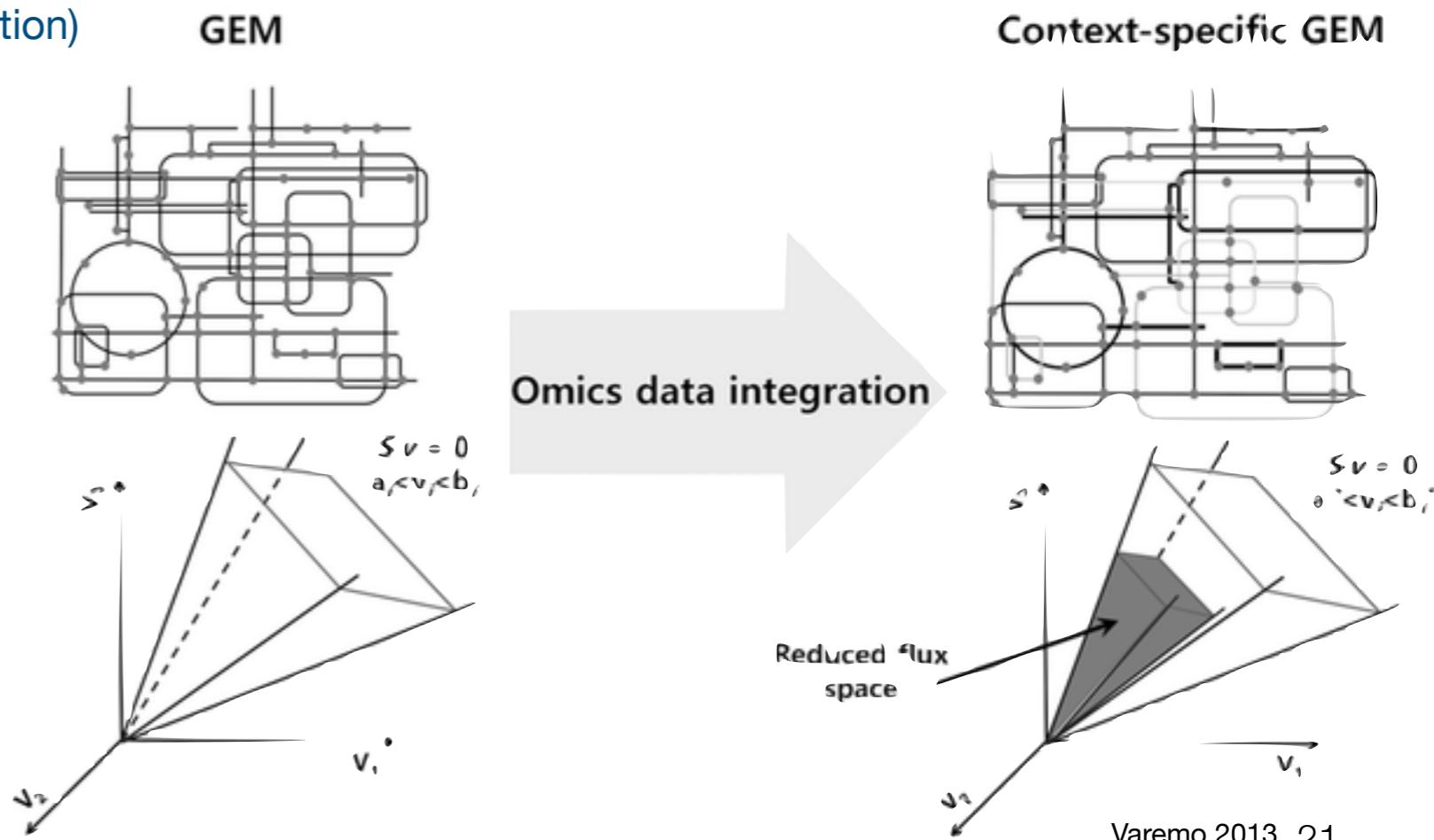


4. Using genome-scale metabolic models for graph creation

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

The overall strategy follows

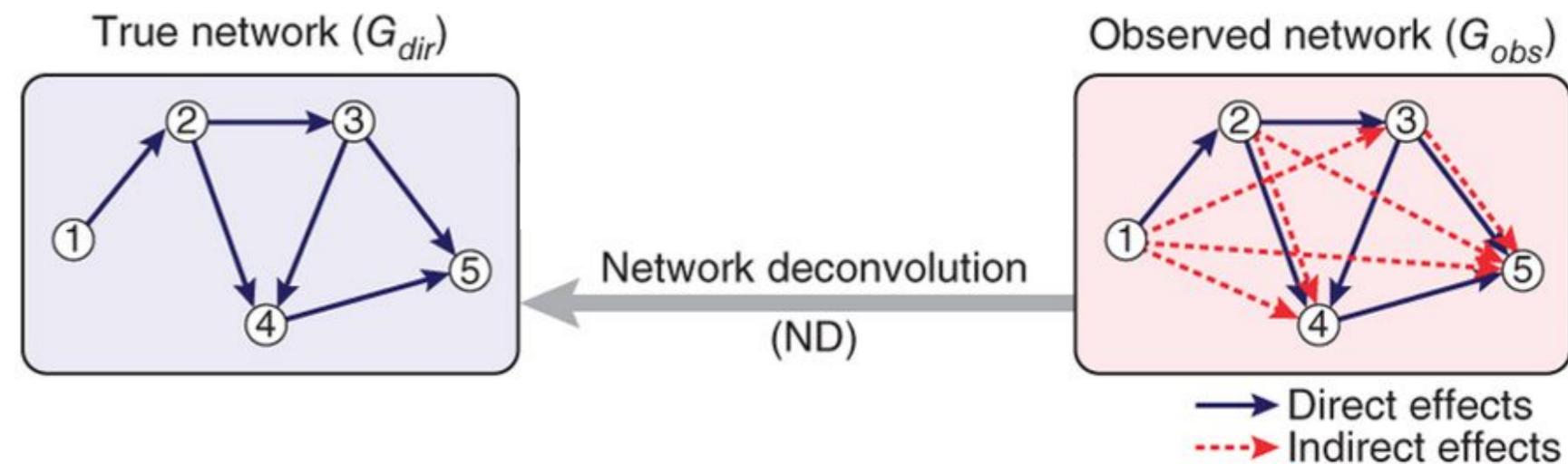
1. Integrate proteomic, transcriptomic, metabolomic, fluxomic
2. Flux distribution
3. Compute metabolite-reaction-gene relationships
4. Extract relevant relationships (met-met, gene-gene)
- 4b. Exclude unnecessary interactions (e.g. cofactors)
5. Downstream analysis (e.g. topology, stratification)



5. Network deconvolution

Biology is **noisy**, which may result in edges that are not true

- $1 \rightarrow 2$
- $2 \rightarrow 3$
- $1 \rightarrow 3$



Direct and indirect effects

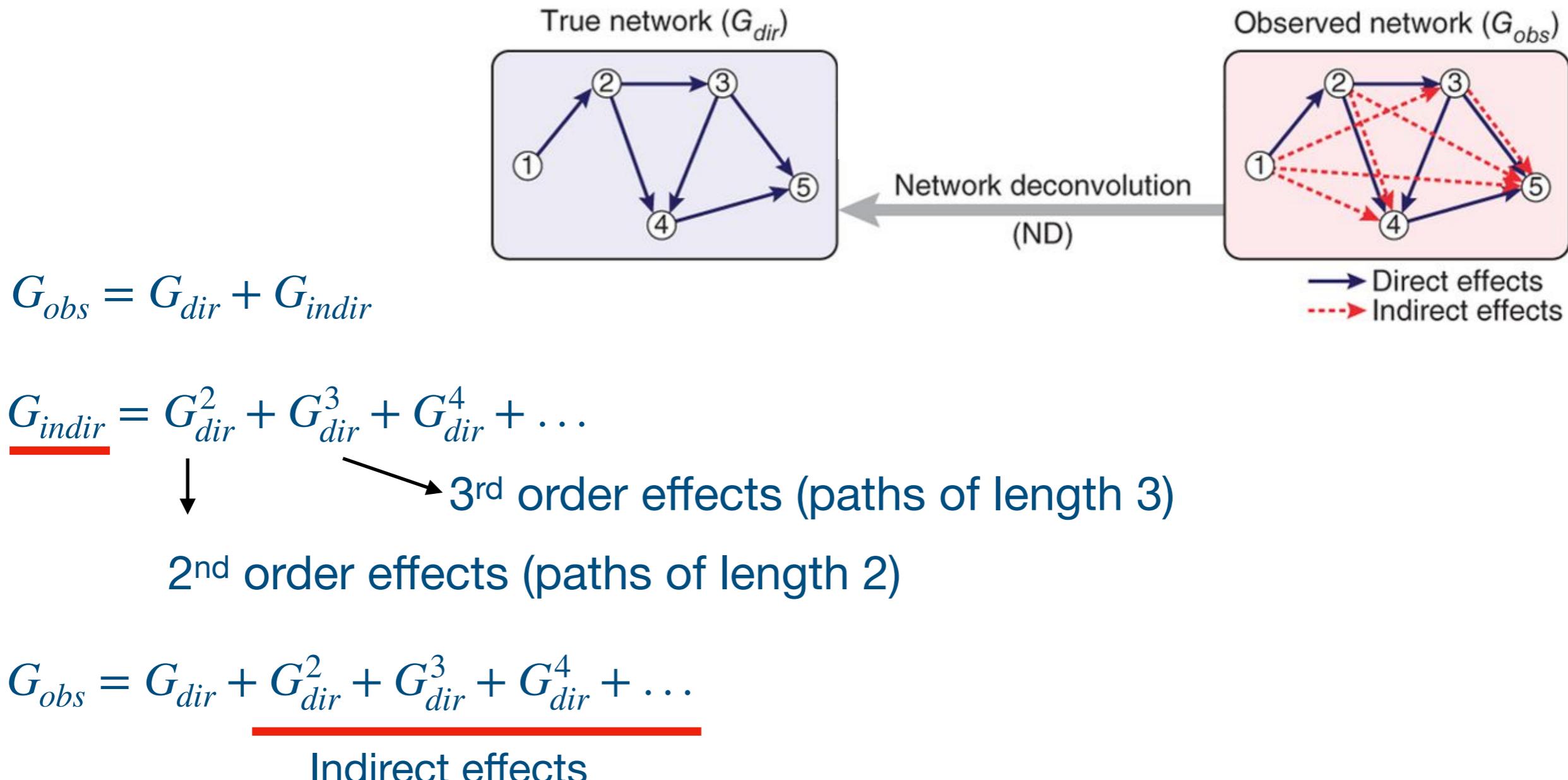
Indirect effects can be of arbitrary length:

- $2 \rightarrow 3 \rightarrow 5$
- $1 \rightarrow 2 \rightarrow 3 \rightarrow 5$
- ...

Decreasing effect with increasing path length

Indirect effects derived from algebraic manipulation for $l = 2, 3, \dots$

5. Network deconvolution



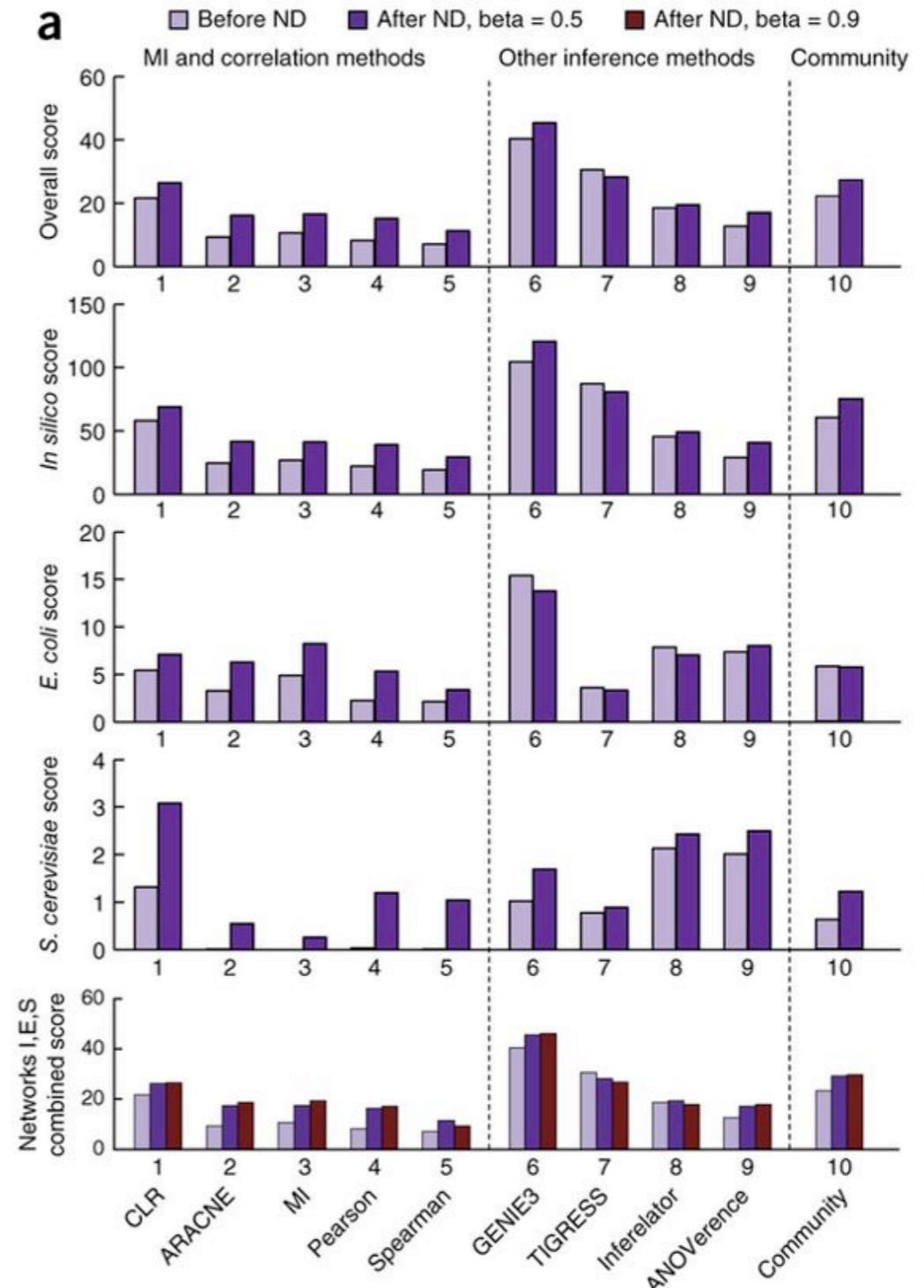
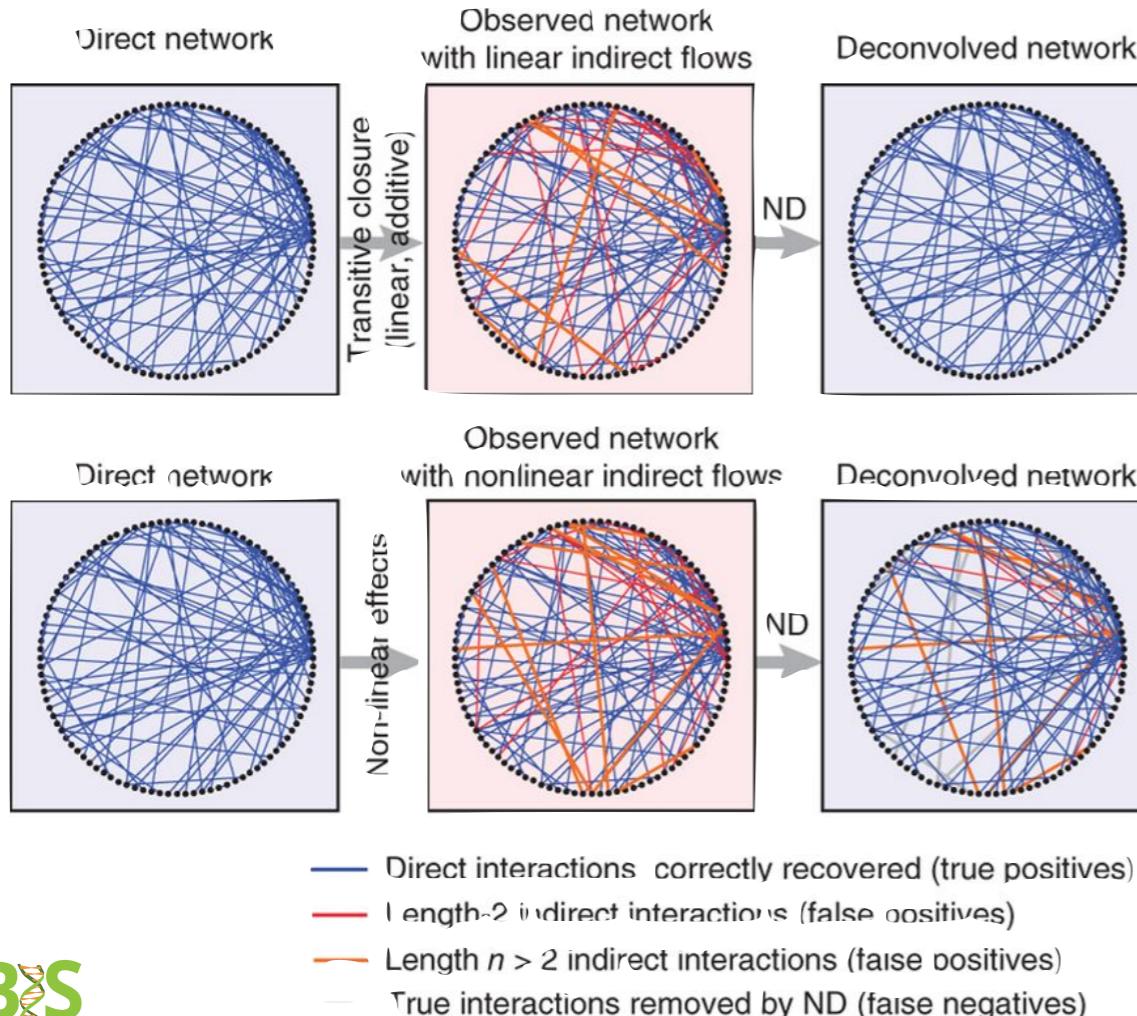
5. Network deconvolution

Nonlinear indirect effects are not always captured

May remove some direct interactions

Does not take into consideration edge weight

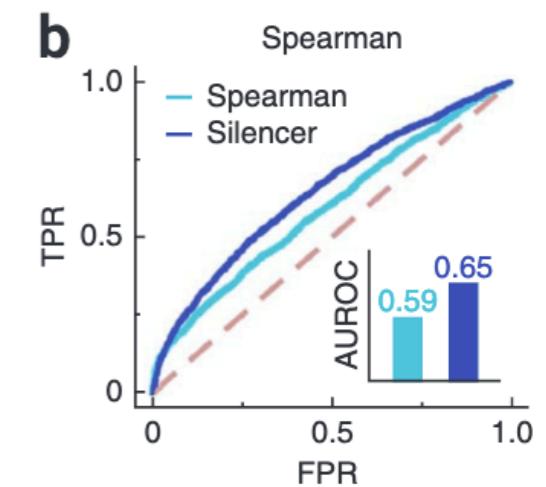
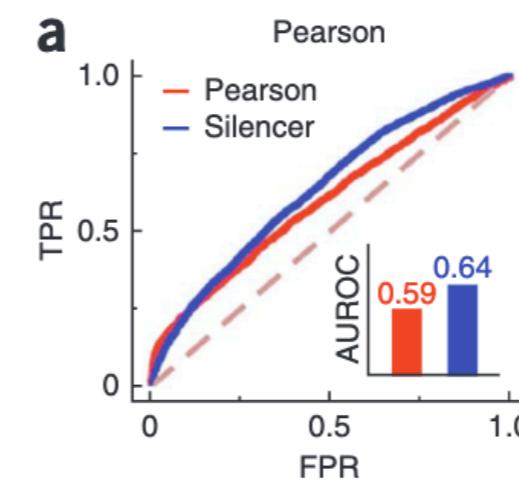
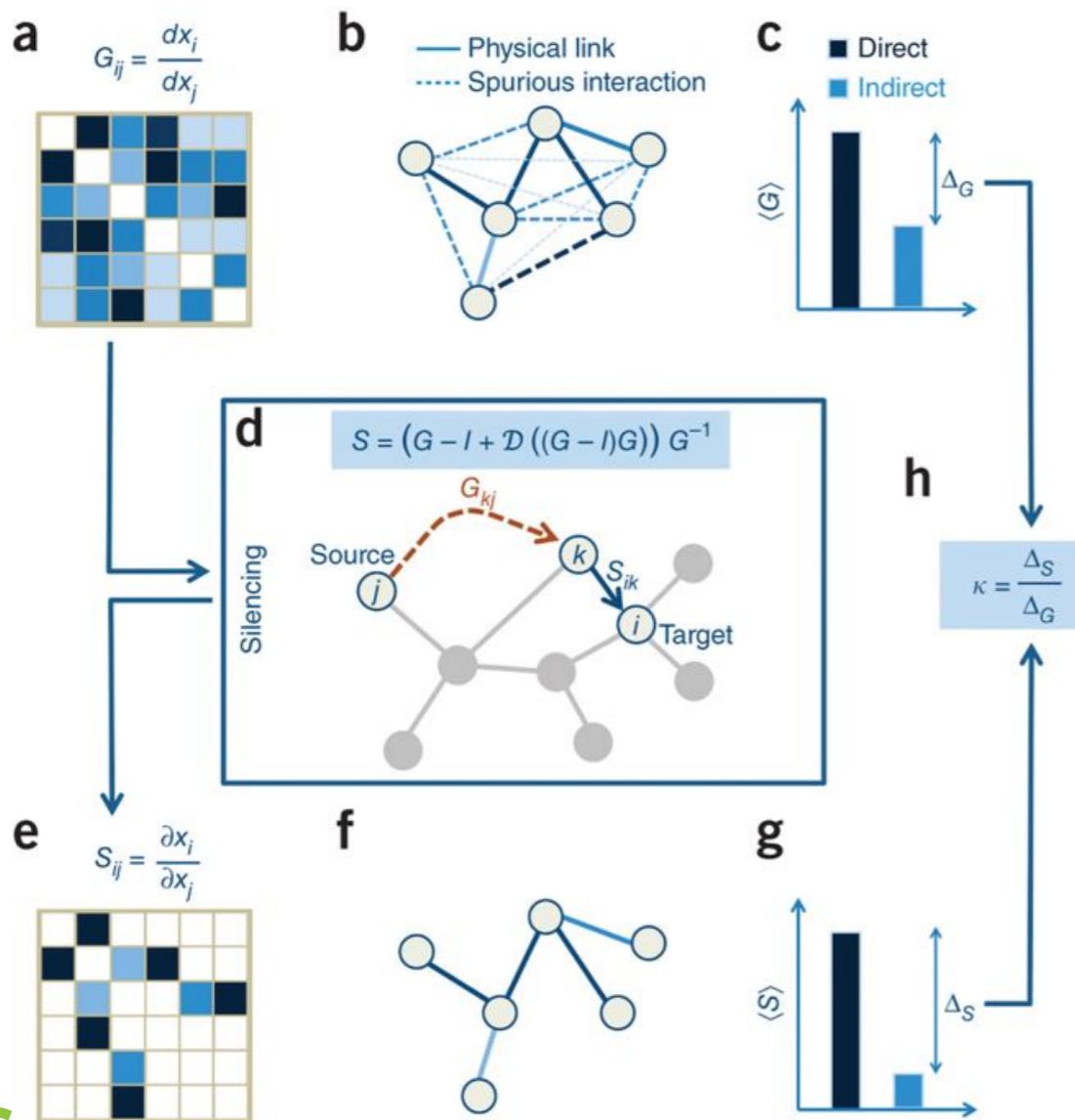
Still improve predictions (true edges: experimental)



5. Network deconvolution

Similar approach involving perturbation of individual nodes to identify indirect interactions

Test case in *E. Coli* co-expression network (4511 nodes)

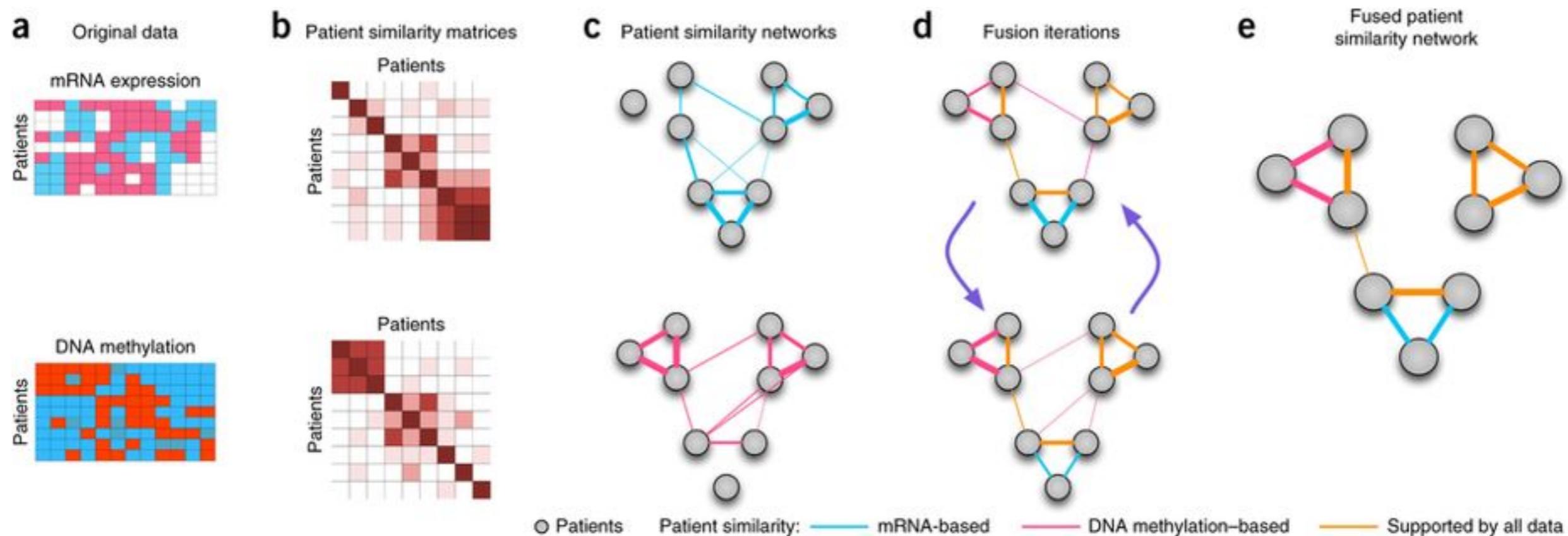


Bonus: similarity network fusion

Sample-sample clustering based on multi-omic data improves (?) clustering

Single-omics present complementary (non-redundant) information

Enables further comparisons between clusters

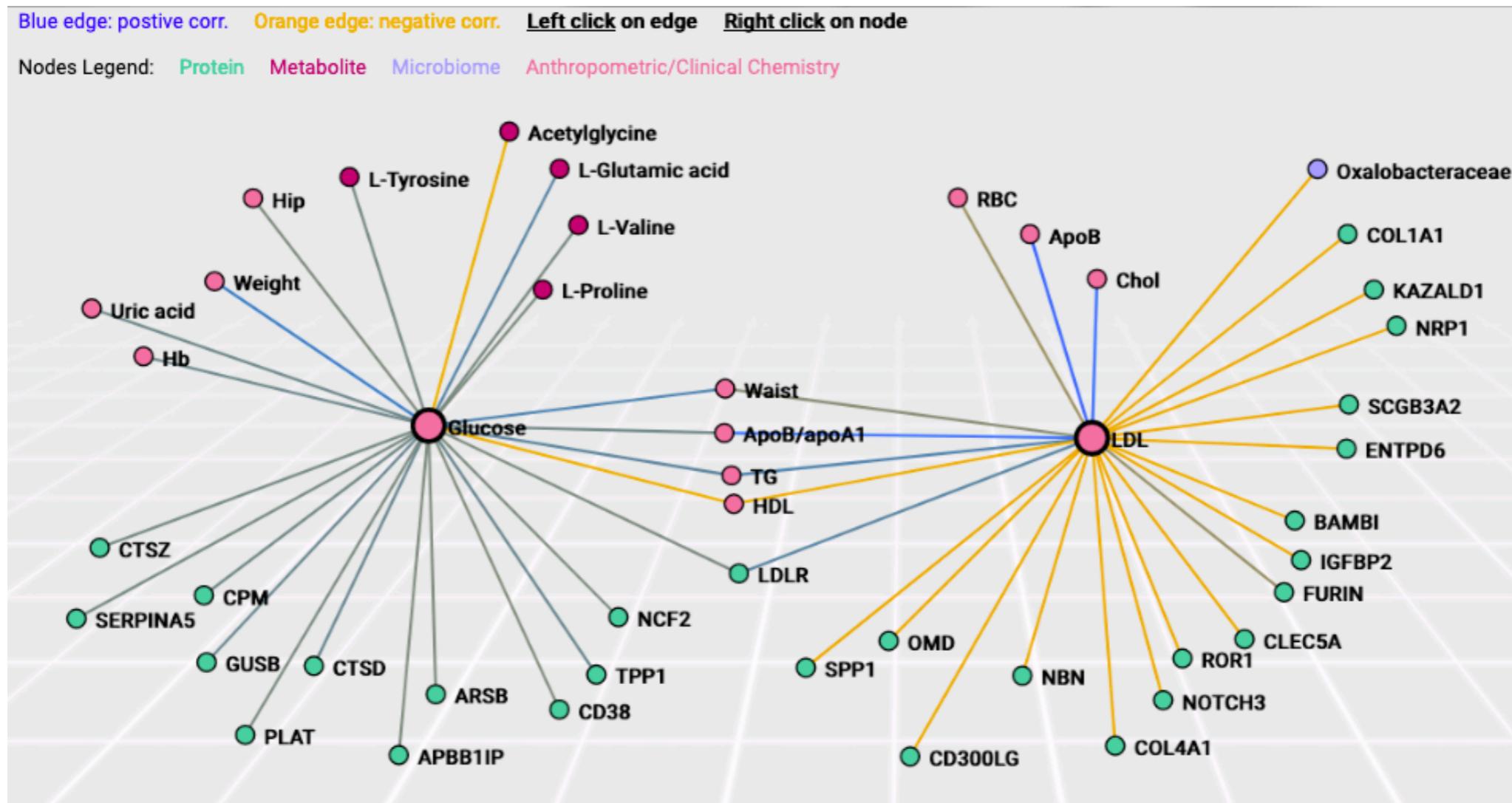


Key network properties

1. Introduction
2. Terminology
3. Network construction
- 4. Key properties**
5. Community analysis
6. Visualization
7. Workshop

Motivation

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



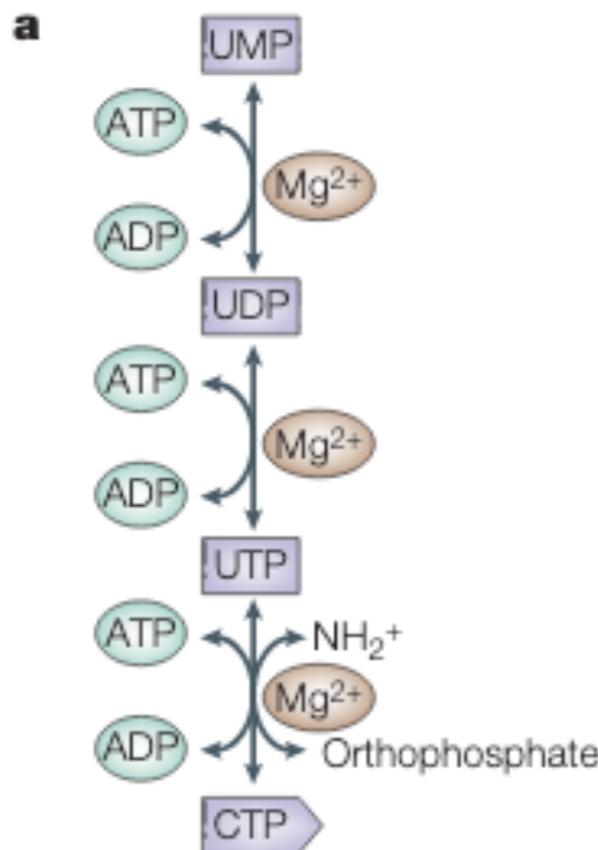
Key network properties to discuss

- 1. Network representations**
- 2. Network density**
- 3. Shortest path, mean path length and walks**
- 4. Connectivity**
- 5. Centrality**
- 6. Clustering coefficient**
- 7. Degree and connectivity distributions**

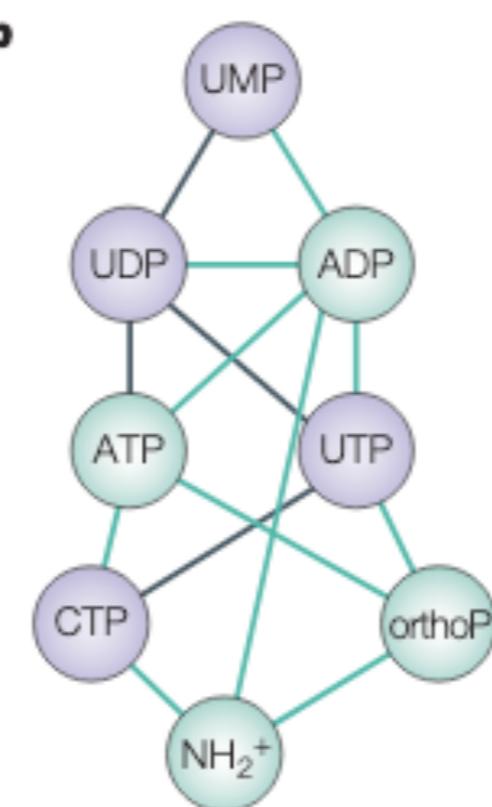
1. Network representations

Representations of a metabolic network: pyrimidine metabolism

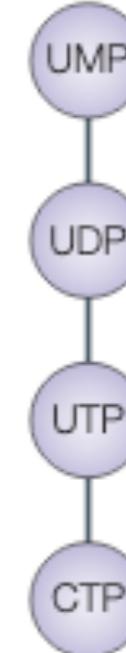
Metabolism



Graph representation:
metabolites and co-factors



metabolite-metabolite
association



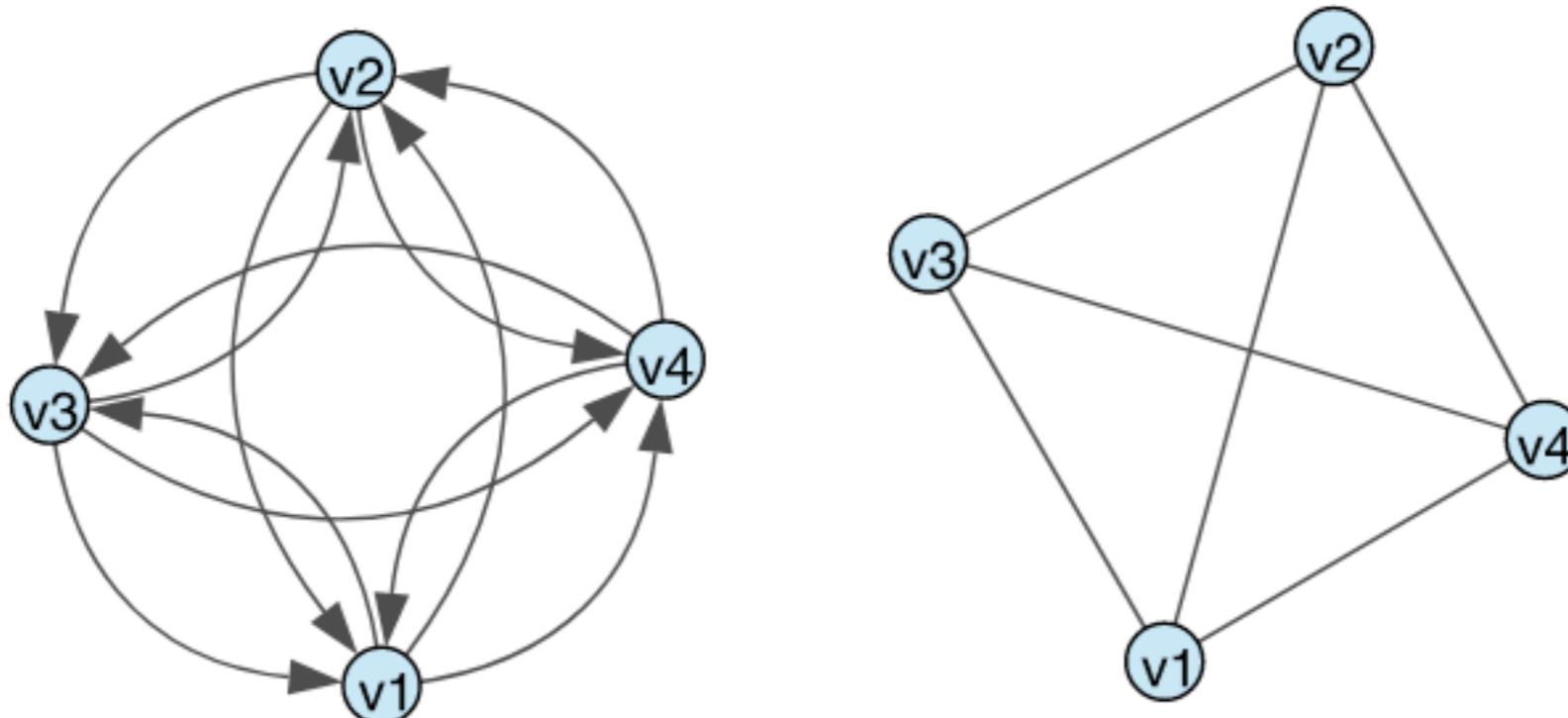
Other representations: Protein-Protein, Protein-Metabolite, ...

2. Network density

For a graph with **V nodes**, the total number of possible **edges** is given by

Directed graphs: $|V| \cdot (|V| - 1)$

Undirected graphs: $\frac{|V| \cdot (|V| - 1)}{2}$



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.

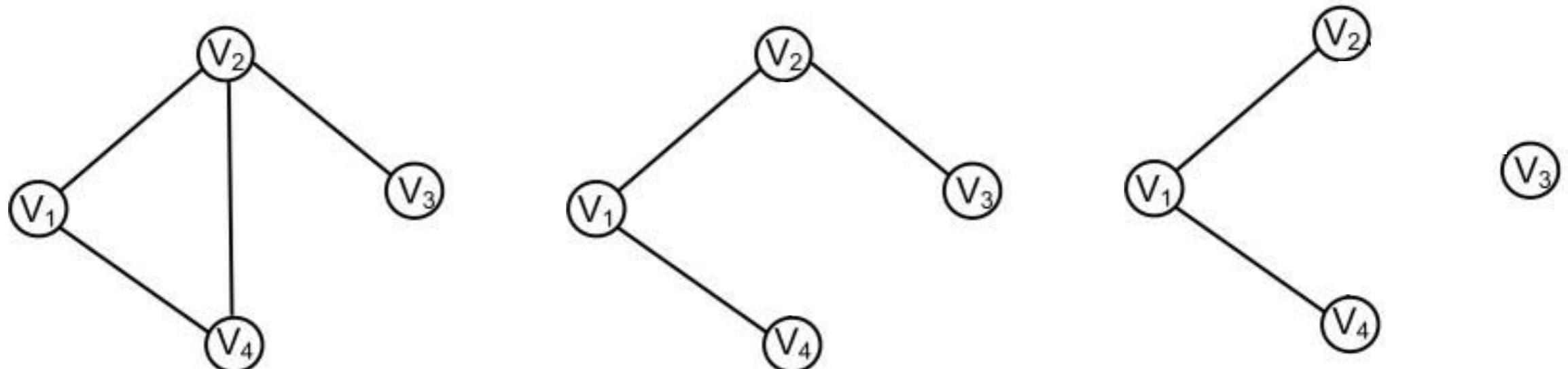
We can thus compute the network **density** (or **global connectivity**) as

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

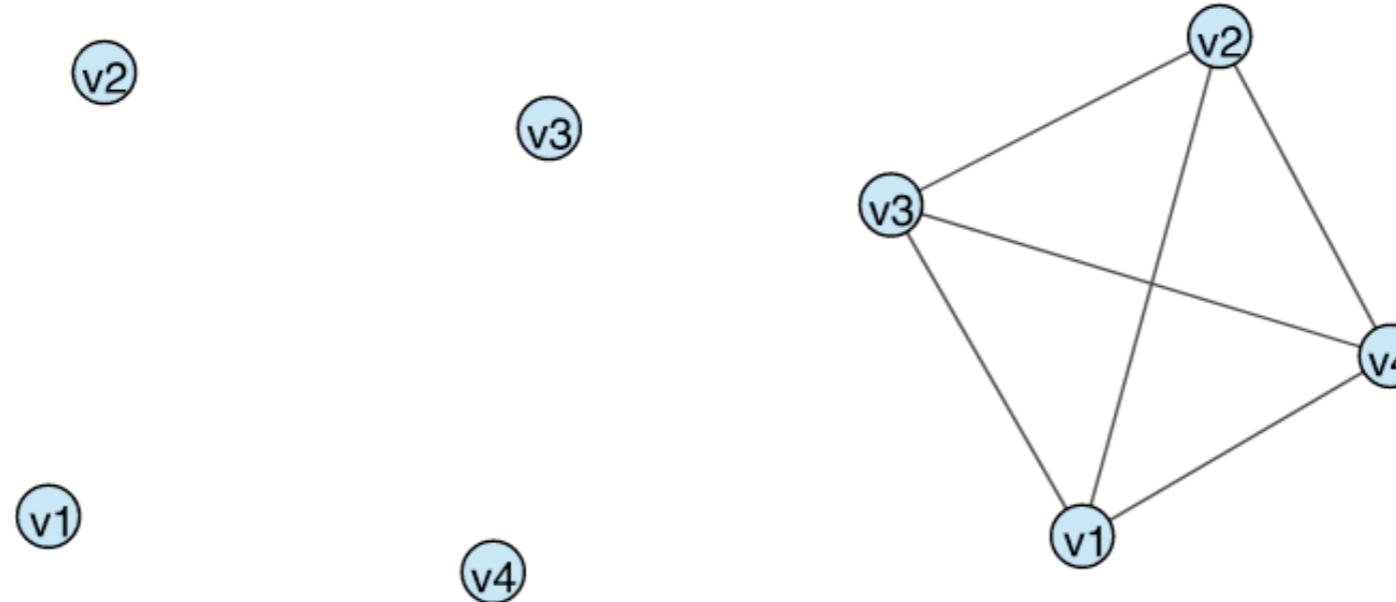
$$\text{Directed graphs: } D = \frac{2 \cdot E}{V \cdot (V - 1)}$$

E : number of edges

V : number of vertices

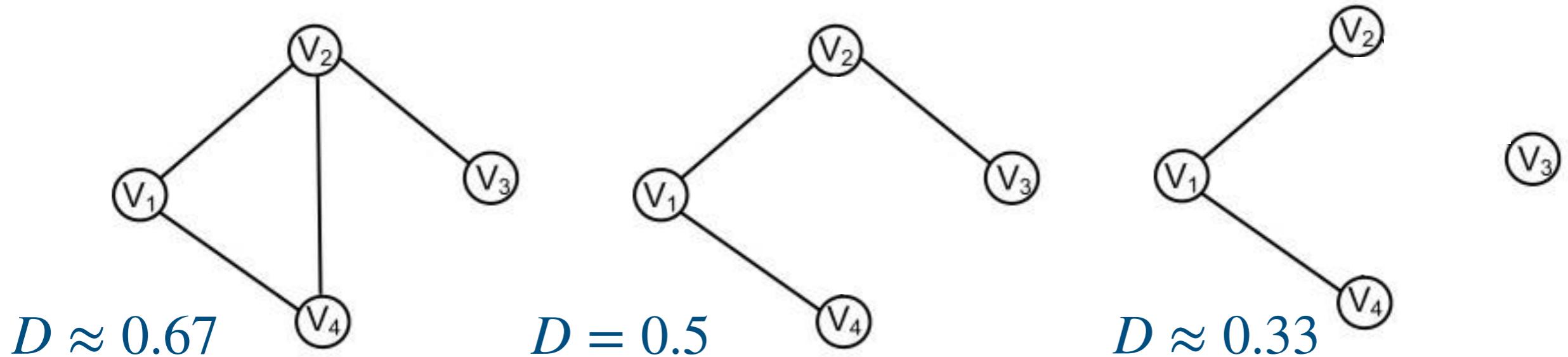


2. Network density



$$0 \leq D \leq 1$$

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



2. Biological network density

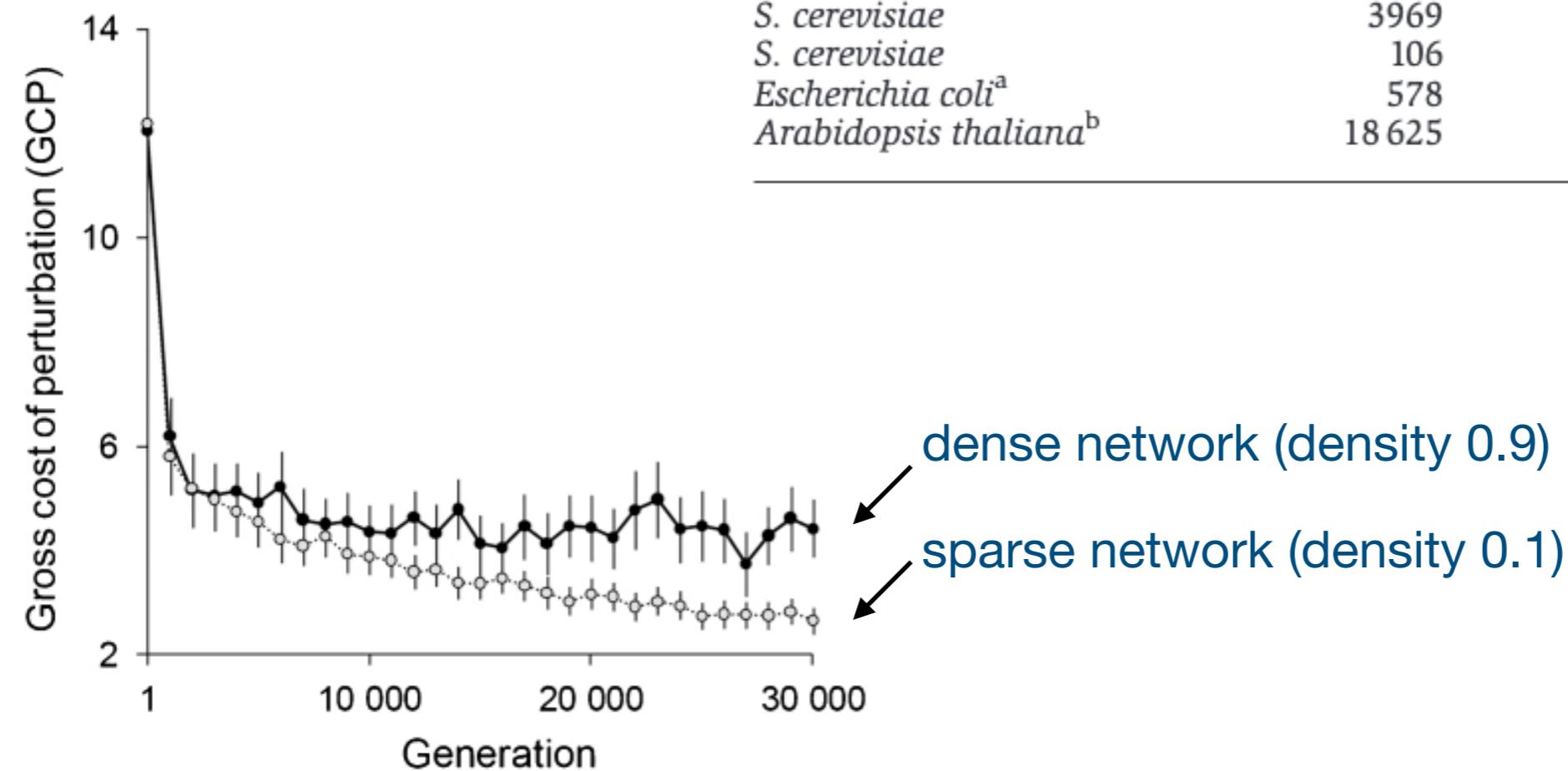
Evolutionary analysis of biological networks indicates general sparsity

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

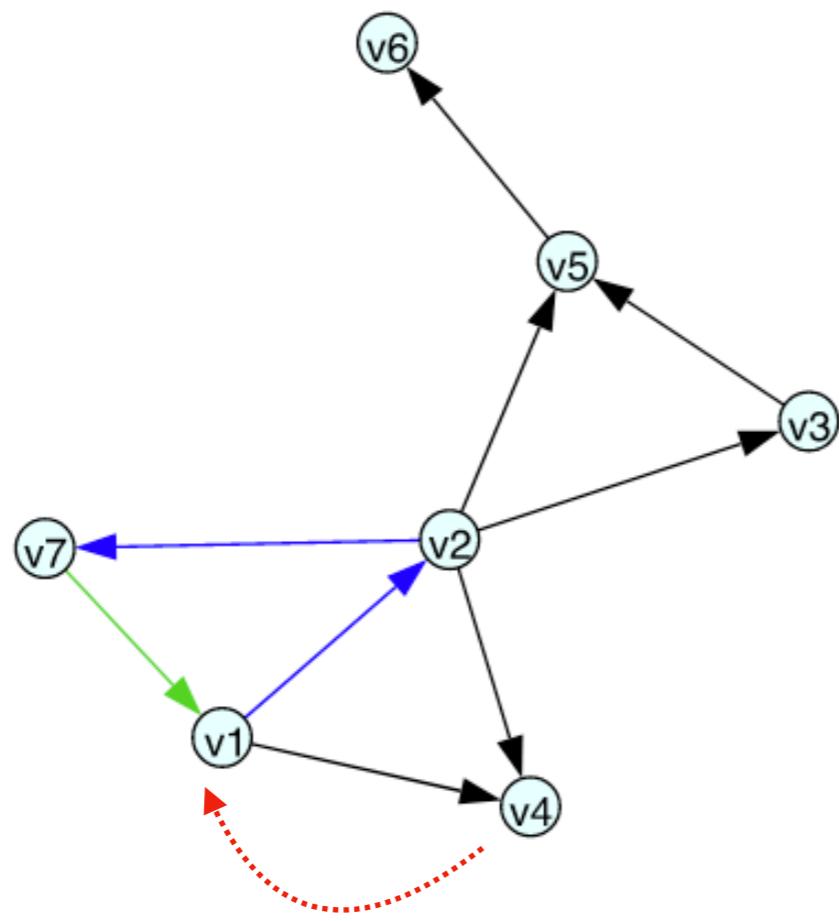
Organism	Interactions	Genes	D	K
<i>Drosophila melanogaster</i>	29	14	0.148	2.07
<i>D. melanogaster</i>	45	25	0.072	1.8
Sea urchin	82	44	0.0065	1.86
<i>Saccharomyces cerevisiae</i>	1052	678	0.0023	1.55
<i>S. cerevisiae</i>	3969	2341	0.0007	1.7
<i>S. cerevisiae</i>	106	56	0.0338	1.9
<i>Escherichia coli</i> ^a	578	423	0.0032	1.37
<i>Arabidopsis thaliana</i> ^b	18 625	6760	0.0004	2.75



3. Paths and walks

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

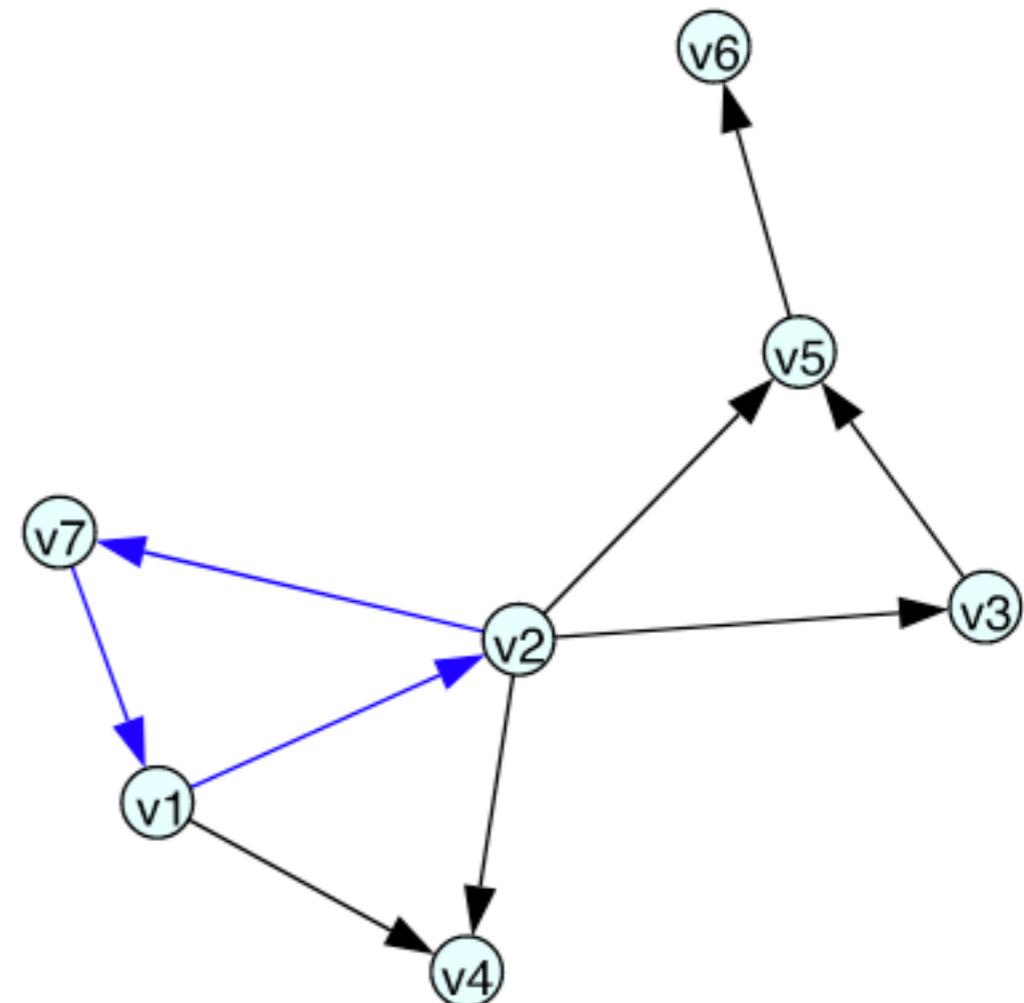
3. Paths and walks

Cycles and acyclic graphs

The **average path** gives a measure of network navigability

Algorithms are selected based on graph features (negative weights? Edge number)

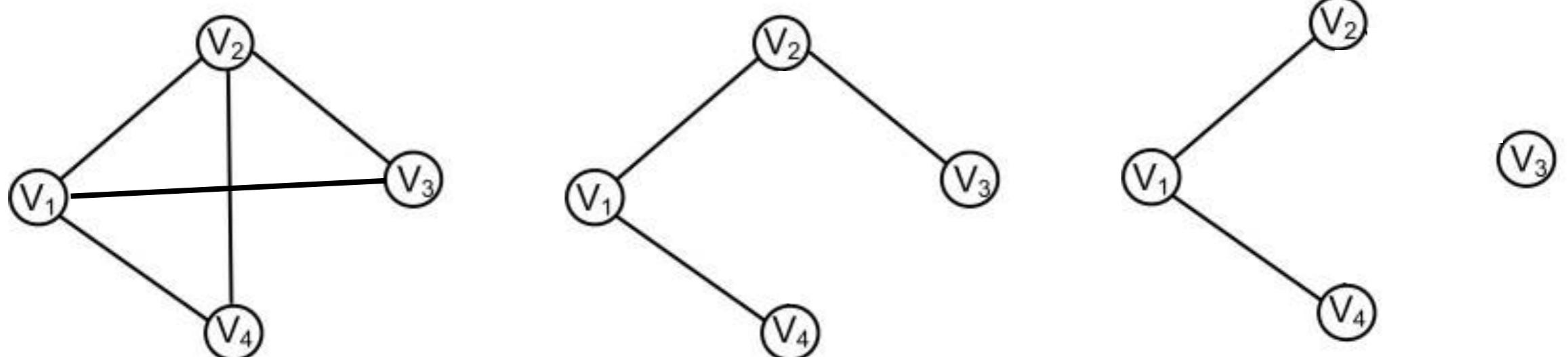
- Breath First Search
- Dijkstra
- Bellman-Ford
- Floyd-Warshall
- Johnson's



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

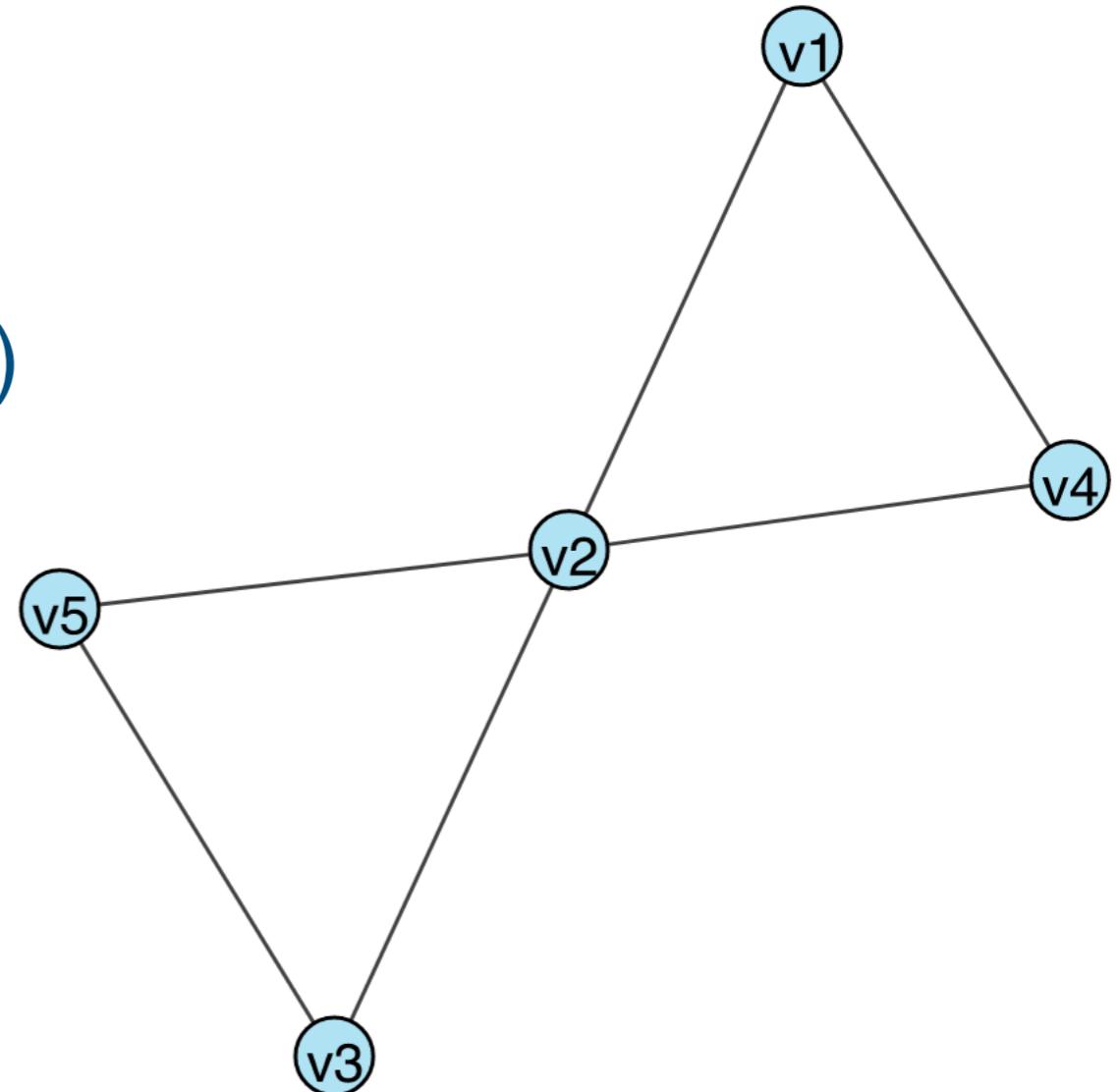


4. Connectivity

$\kappa(G) = 1$; **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. v_3, v_1 : v_2 and associated edges)

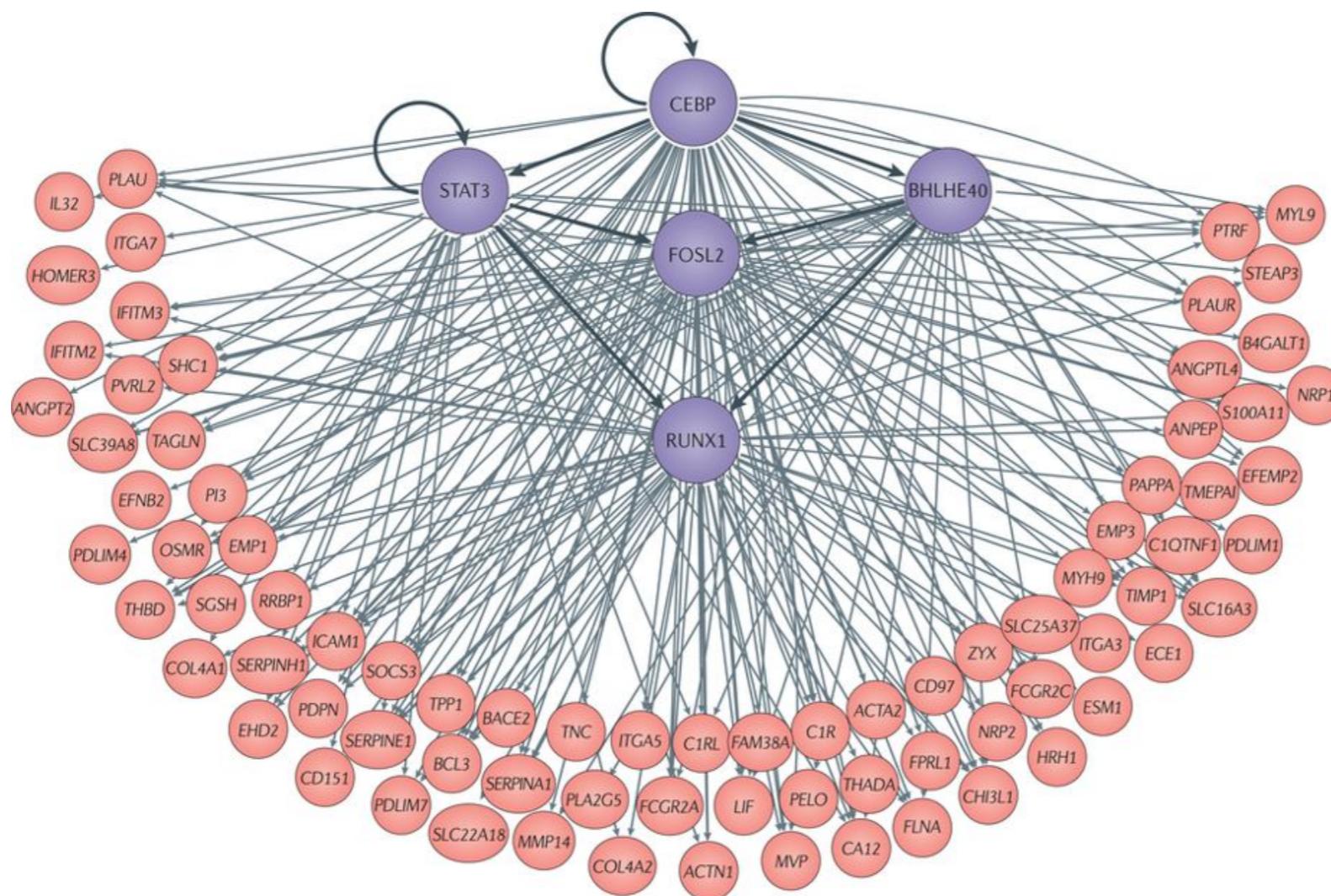


5. Centrality

Indicate the most central nodes in a network

Central nodes **possibly** most important in the network, act as **hubs**

Example: Transcription Factor Master Regulators



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
- Closeness
- Betweenness
- Eigenvector
- Katz
- PageRank
- Percolation
- Cross-clique

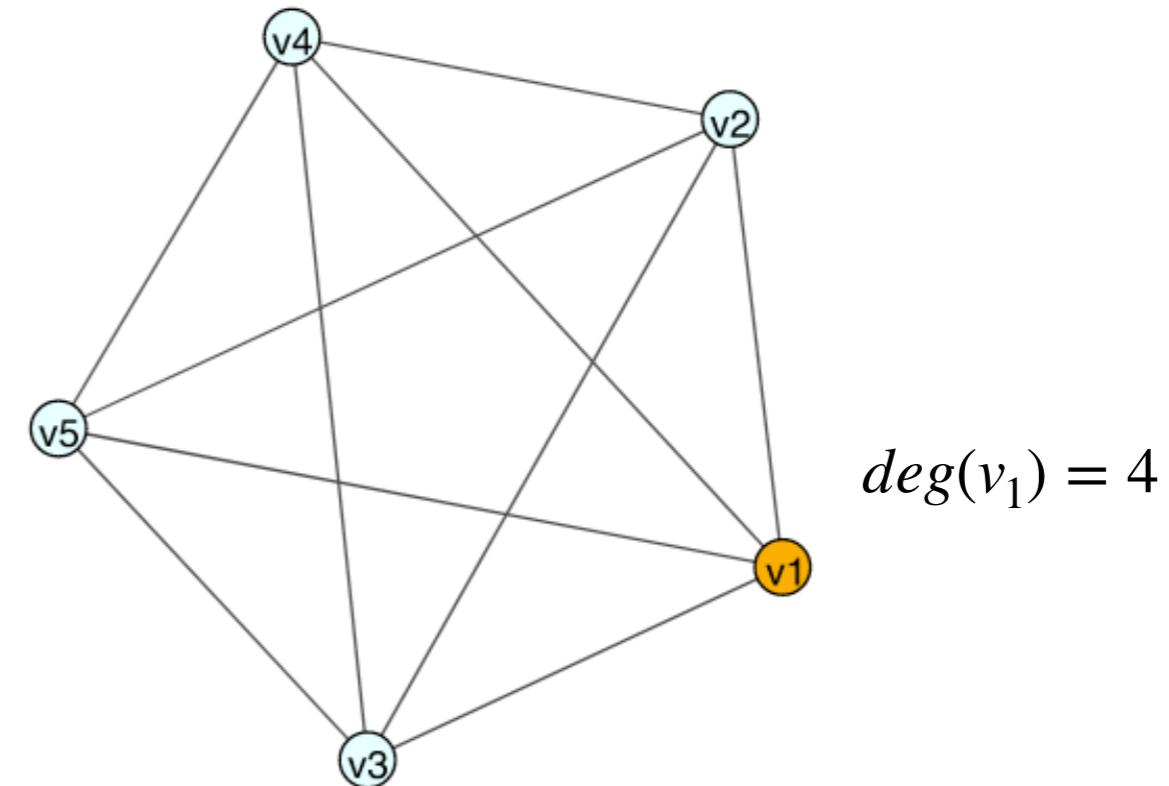
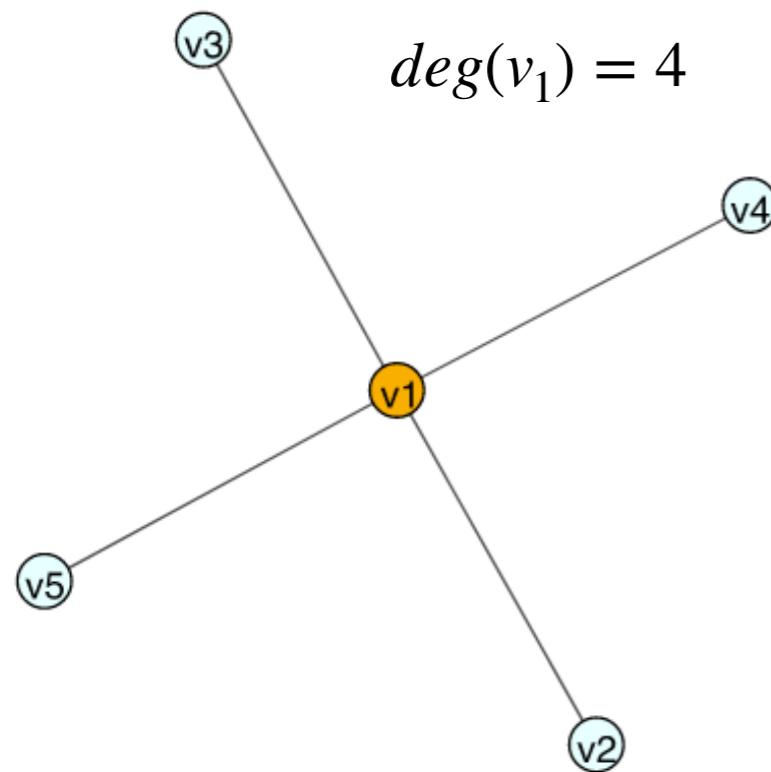
...

5. 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.



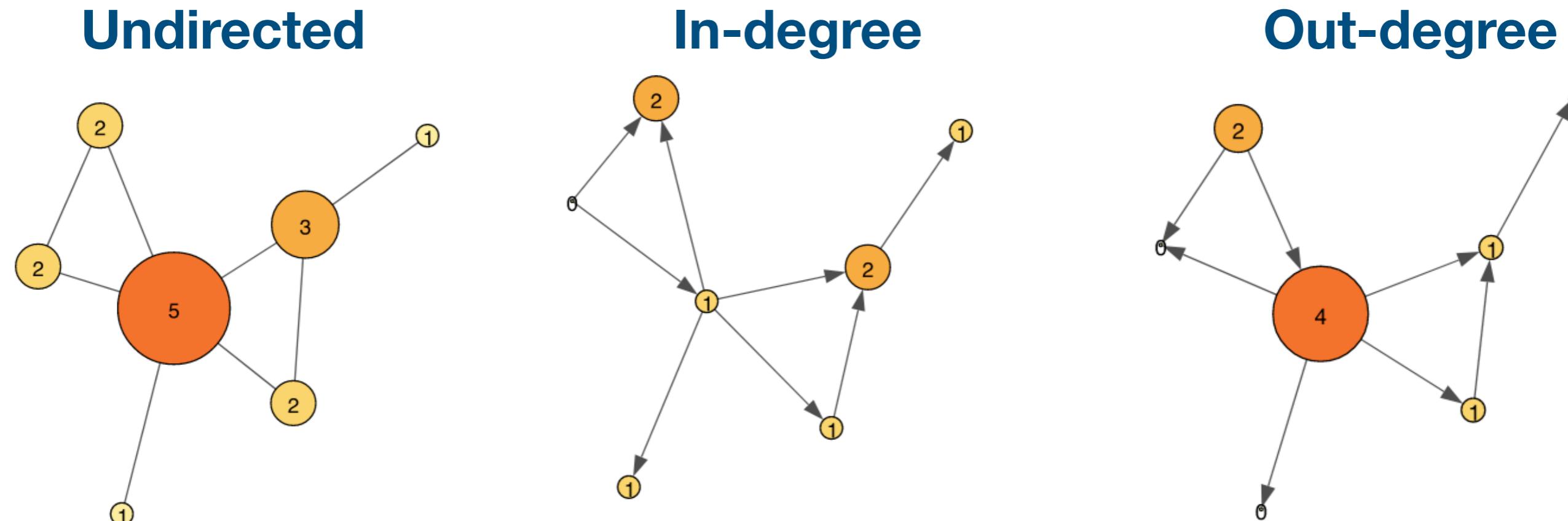
5. Centrality: degree centrality

Undirected networks vs directed networks

In-degree vs Out-degree

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

Numbers indicate degree:



5. Centrality: degree centrality

Degree centrality

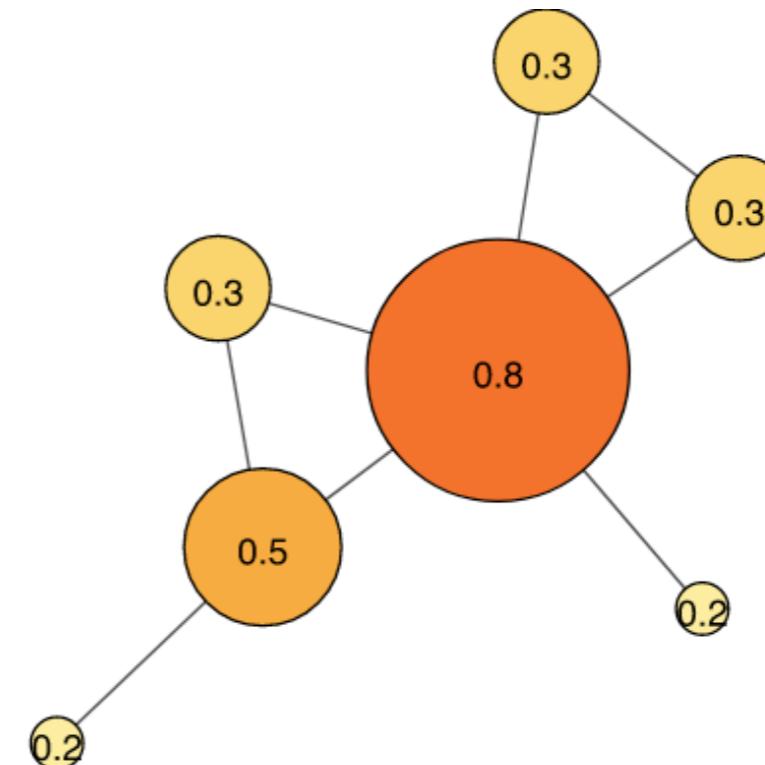
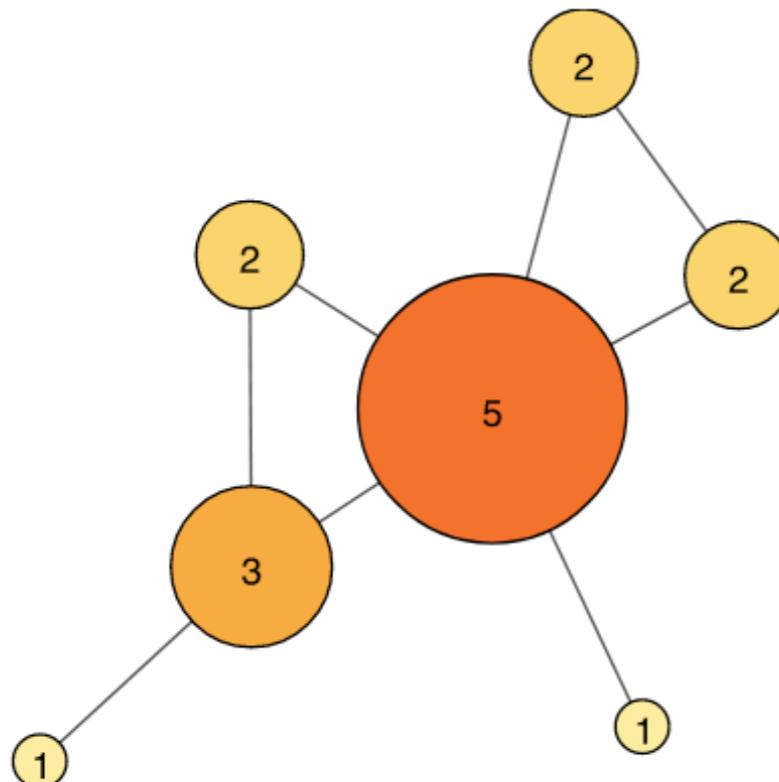
$$C_D(v_i) = \sum_{j=1}^N e_{ij}$$

Normalized
degree centrality

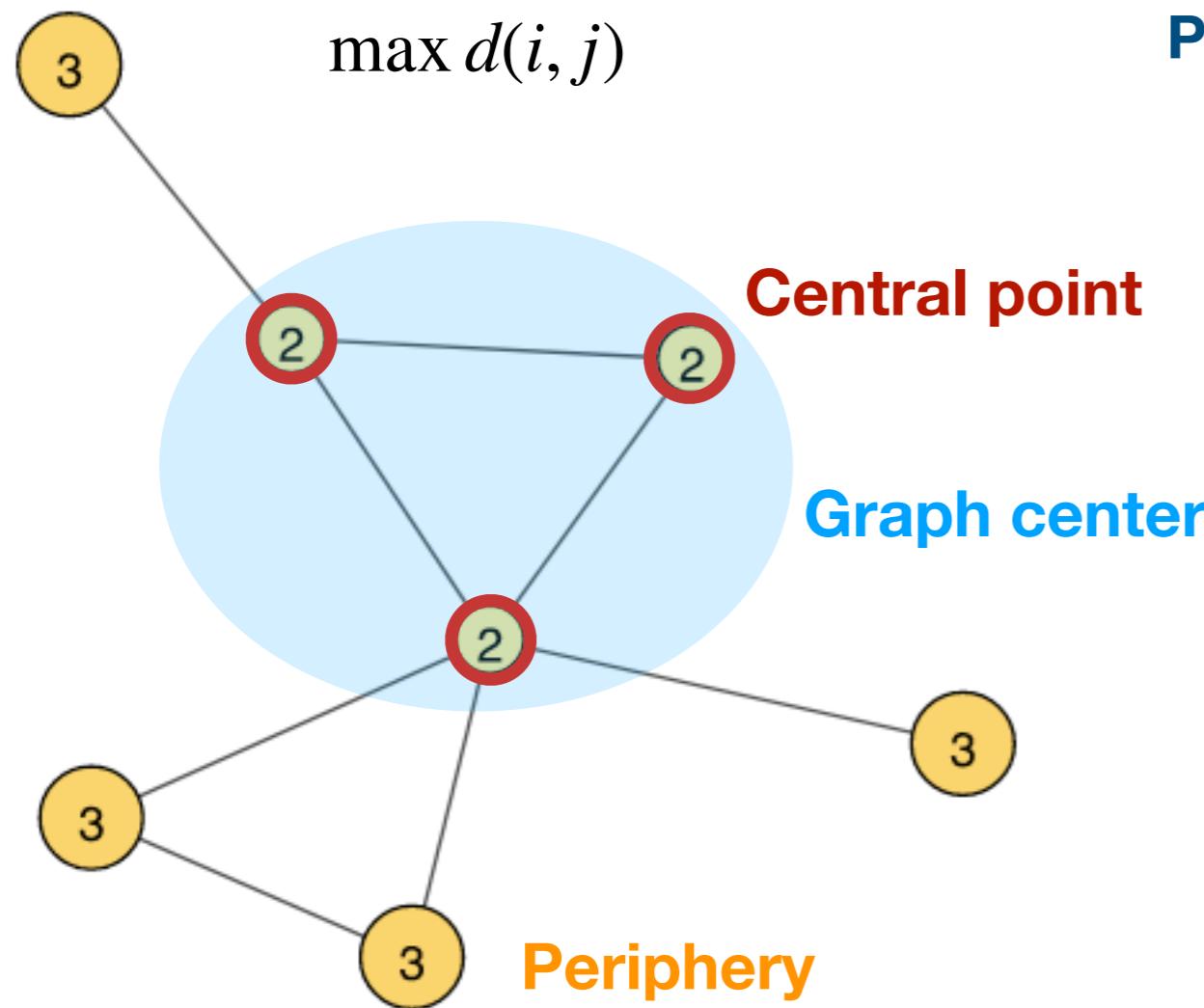
$$C_D(v_i) = \frac{\sum_{j=1}^N e_{ij}}{N - 1}$$

Normalized degree centrality accounts for the total possible number of connections

Centrality normalization allows for comparison between networks of different sizes



5. Centrality: eccentricity centrality



Network Diameter is the maximum distance

Radius is the minimum distance

Central point is that which has $d(i, j) = \text{radius}$

Graph center: set of nodes with minimum ecc

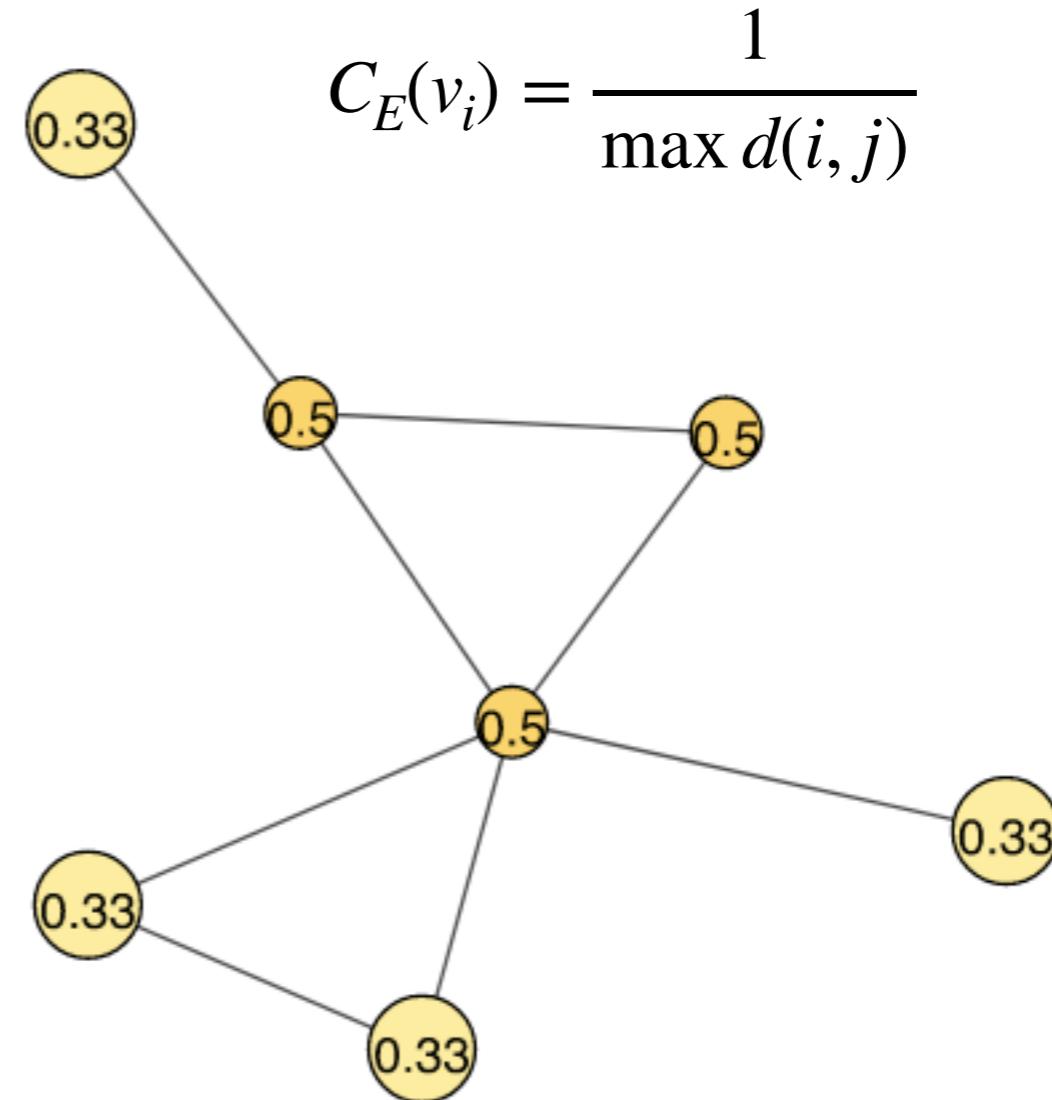
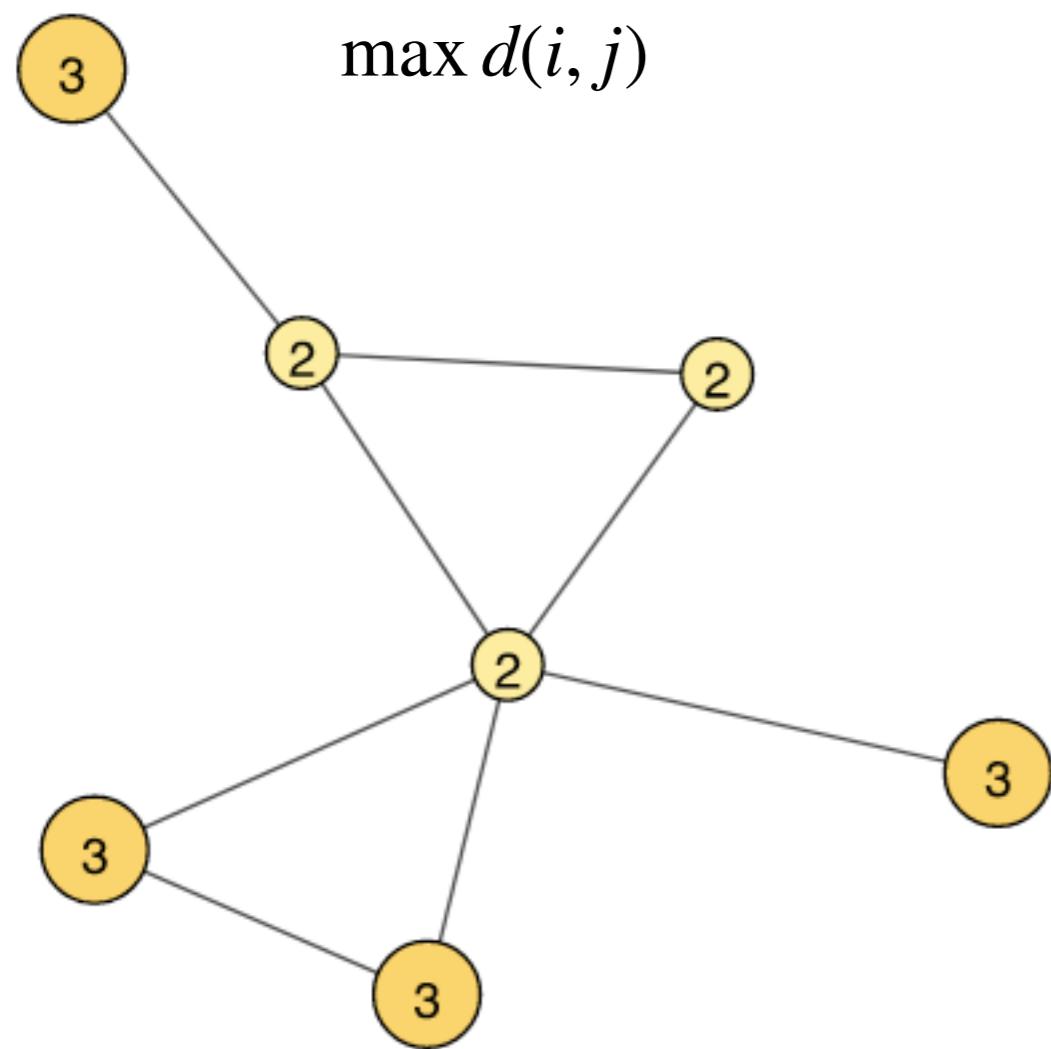
Periphery: set of nodes with $d(i, j) = \text{diameter}$

Diameter = 3

Radius = 2

5. Centrality: eccentricity centrality

Eccentricity considers a node's max path to all other nodes

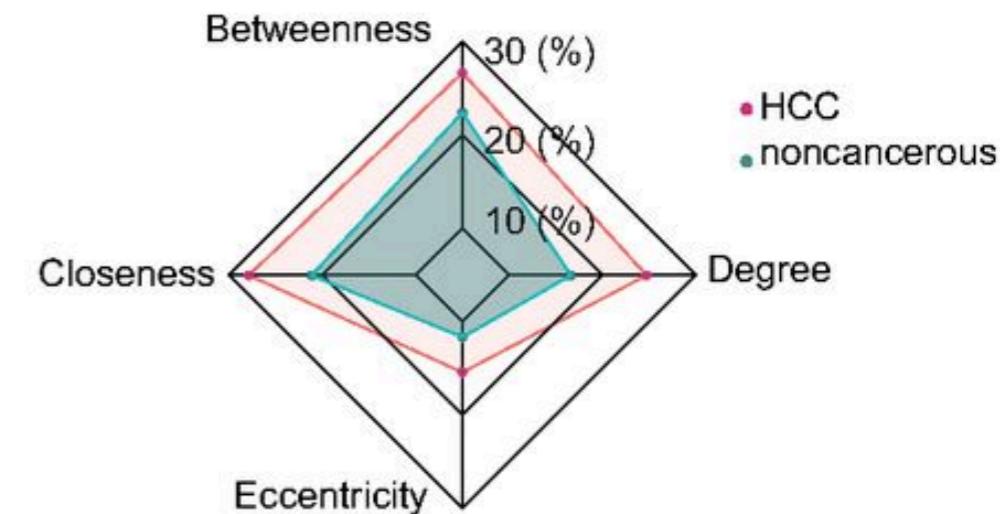


5. Centrality: limitations & influence

Node centrality does not necessarily imply **importance**

How to tackle this?

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.

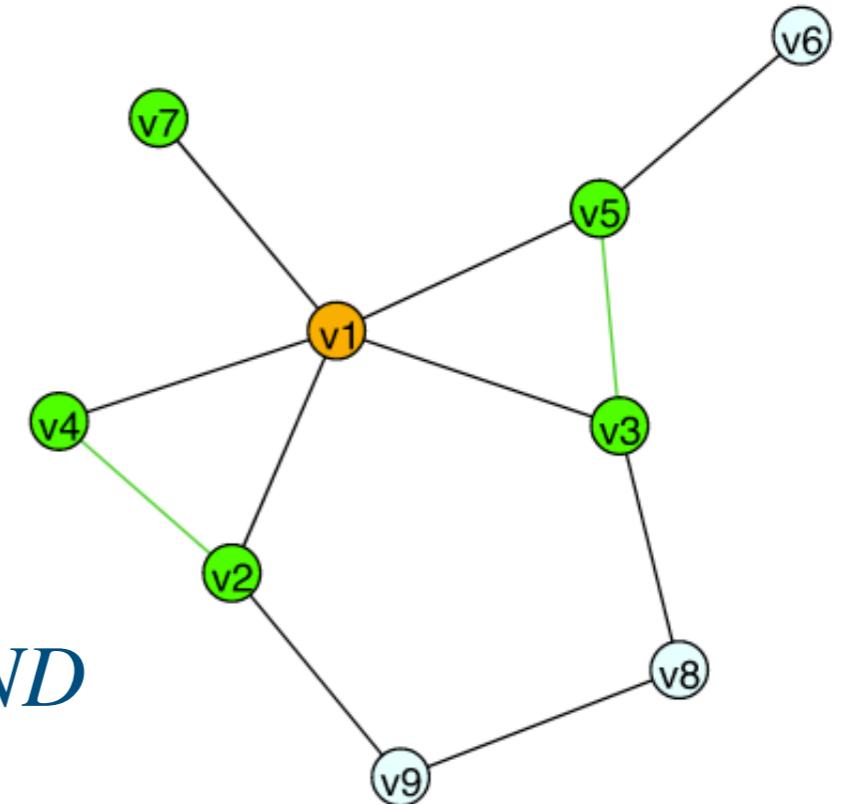
Takes into account degree of a node and the degree of its 1st neighbours

For node v_1

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

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

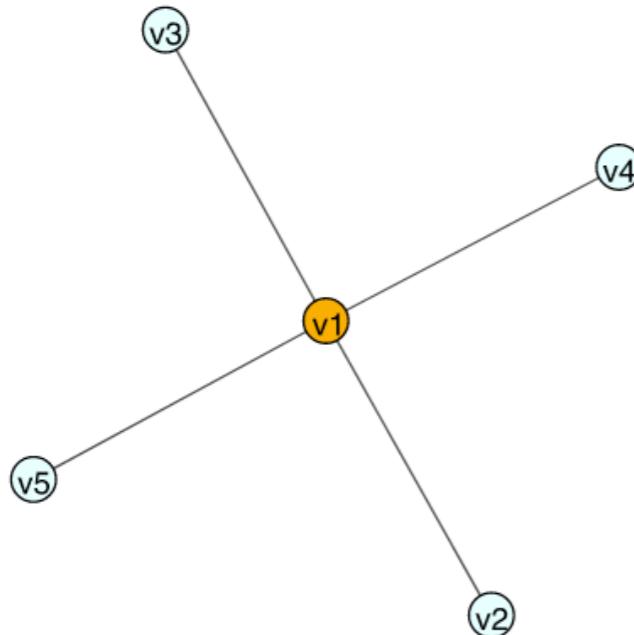
$$C(v_1) = \frac{2 \cdot 2}{5 \cdot 4} = 0.2 \quad C(v_7) = \frac{2 \cdot 0}{1 \cdot 0} = 0 \text{ or } ND$$



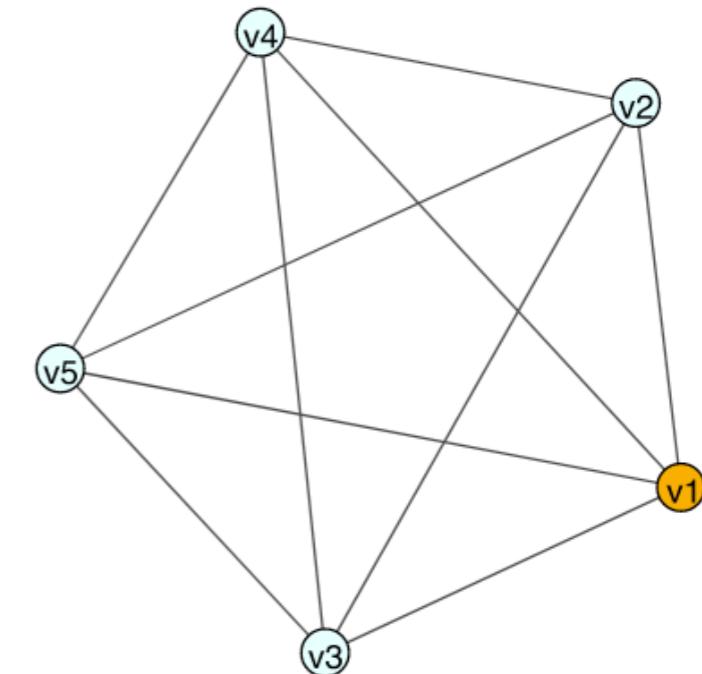
6. Clustering coefficient

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

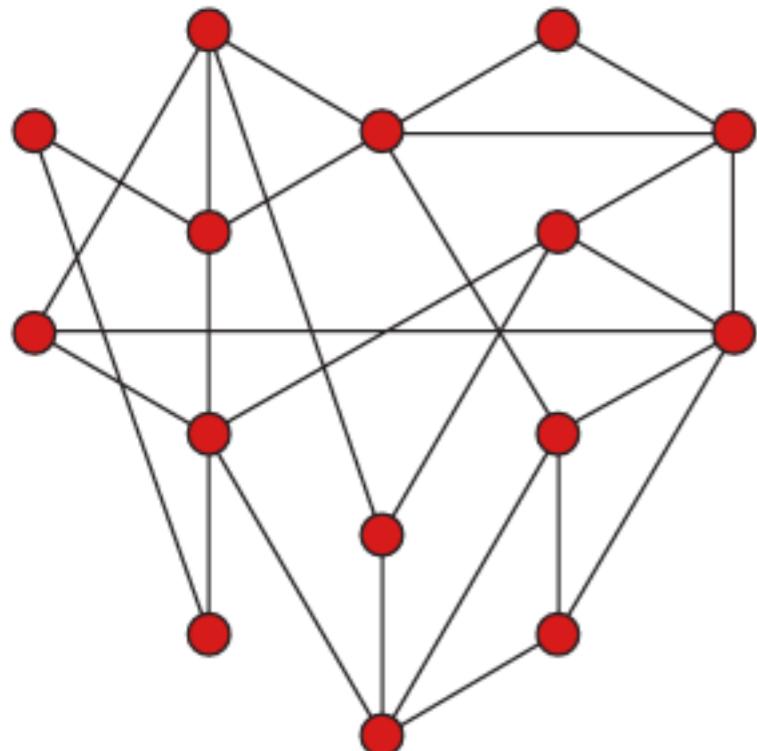
7. Degree and clustering coefficient distribution

$P(k)$ gives the probability that a selected node has exactly k edges

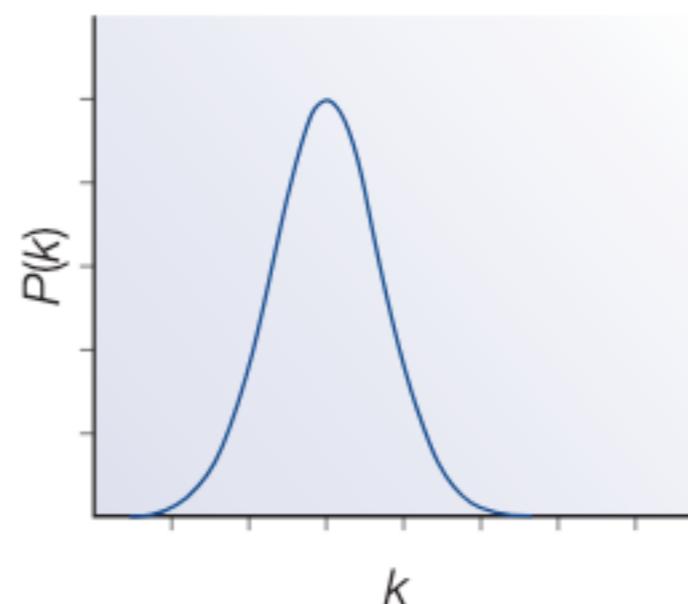
Obtained by counting number of nodes $N(k)$, for $k = 1, 2, \dots$

Allows distinguishing different kinds of networks

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



Poisson degree distribution
shows no highly connected nodes

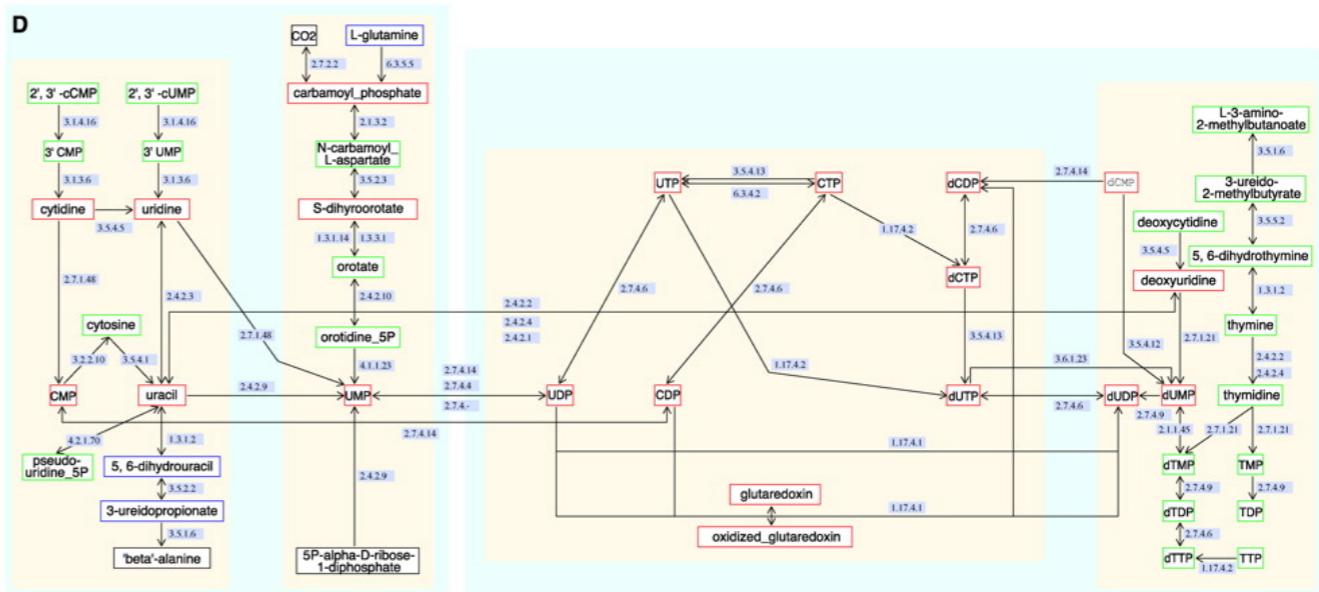
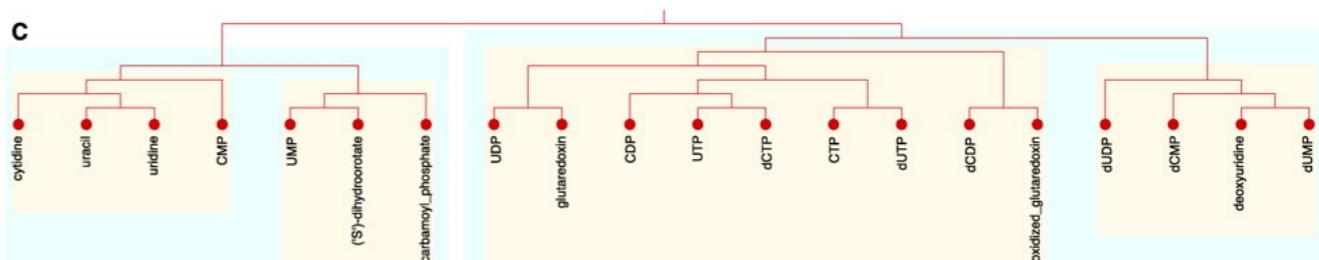
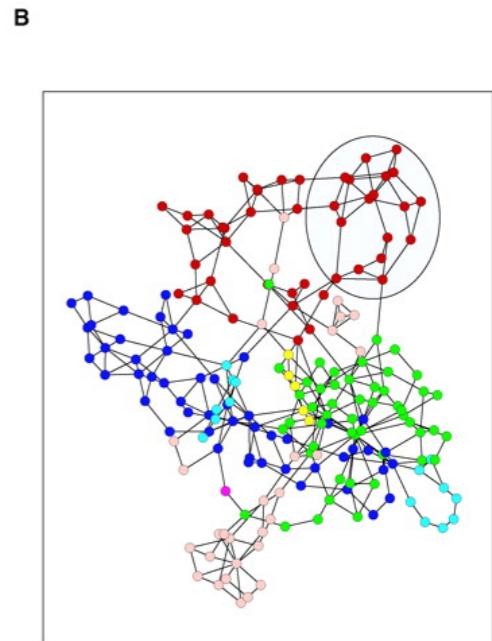


Most nodes have near $\langle k \rangle$

Metabolic networks show hierarchical topology

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

Their combination shows a hierarchical structure



7. Degree distribution

Biological networks do not follow topology features of random networks.

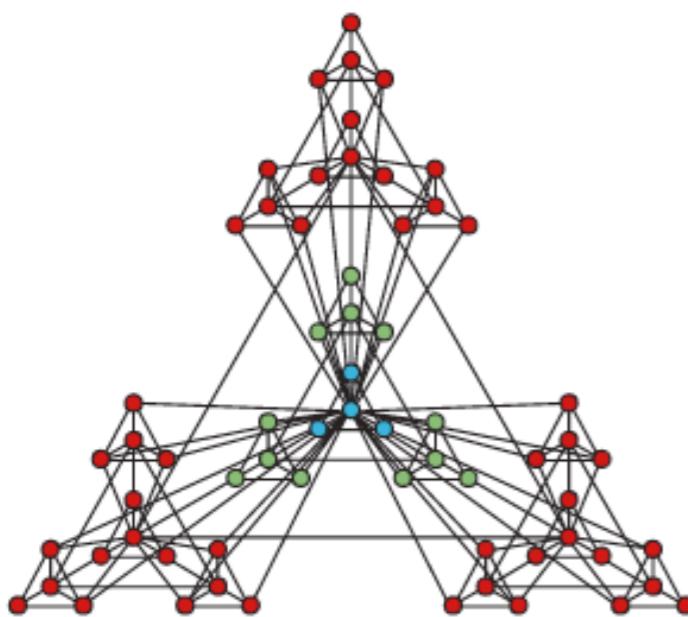
Analysis of metabolic networks of 43 organisms shows common patterns

Degree distribution *follows* the power-law $P(k) \propto k^{-\gamma}$, and is termed **scale-free**

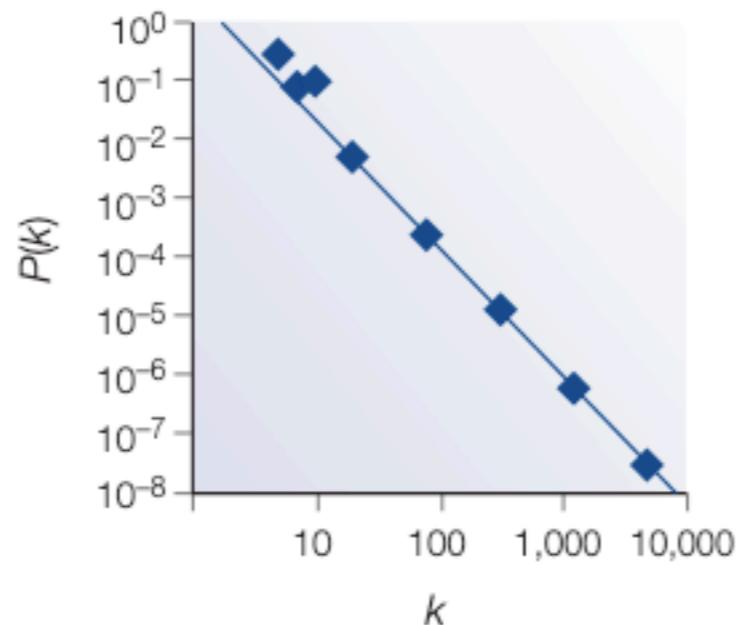
Degree exponent $2 < \gamma < 3$, where smaller γ indicates larger importance

Scale-free and Hierarchical 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



In practice:
 $\gamma < 3$:
 $P(k) \propto k^{-\gamma}$ and $P(k) \propto N$
 $\gamma > 3$:
network behaves like random

7. Degree and clustering coefficient distribution

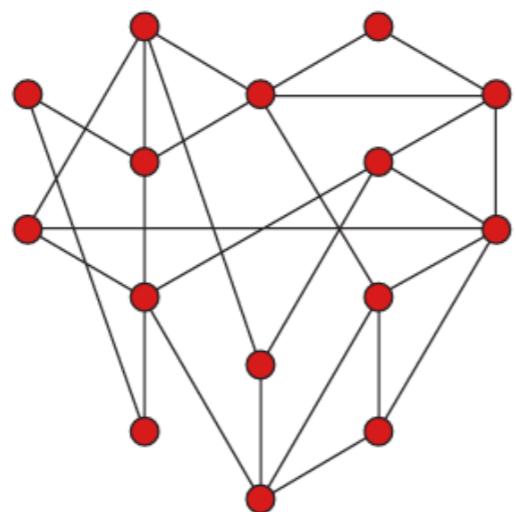
$C(k)$ shows no relationship with k in random networks: no modular organisation

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

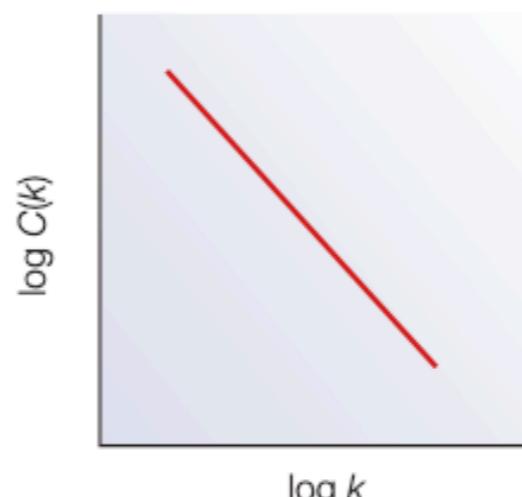
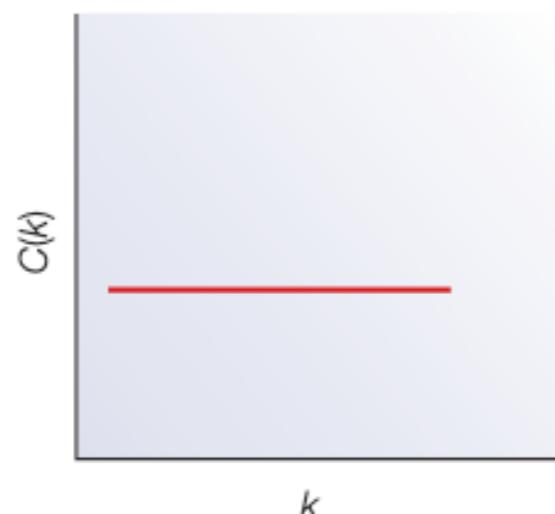
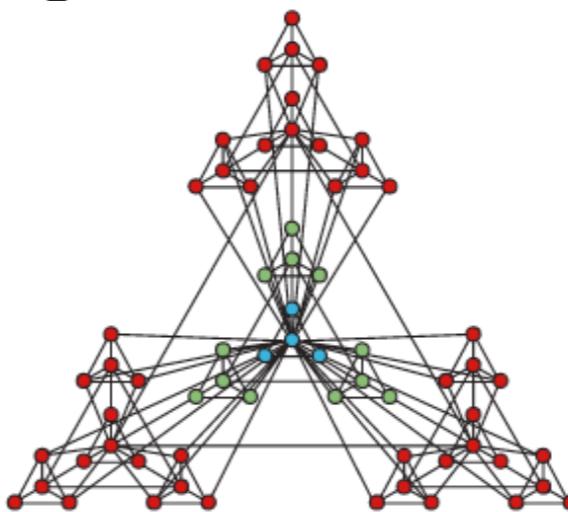
Sparingly connected nodes are part of highly modular areas

Communication between highly clustered neighbourhoods maintained by a few hubs

Random network



Hierarchical network



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

$$l(G) \approx \log N \text{ for a network of size } N$$

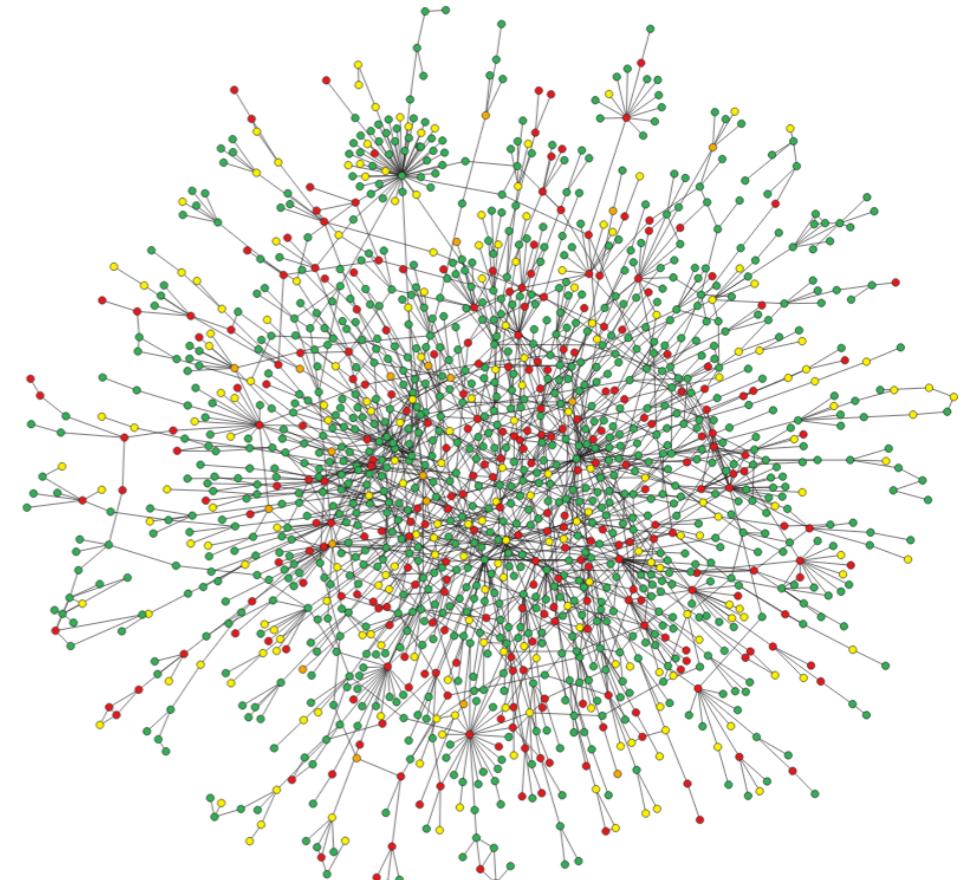
Scale-free networks show **ultra-small world**:

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

In practice, this indicates that perturbations may quickly spread throughout the network

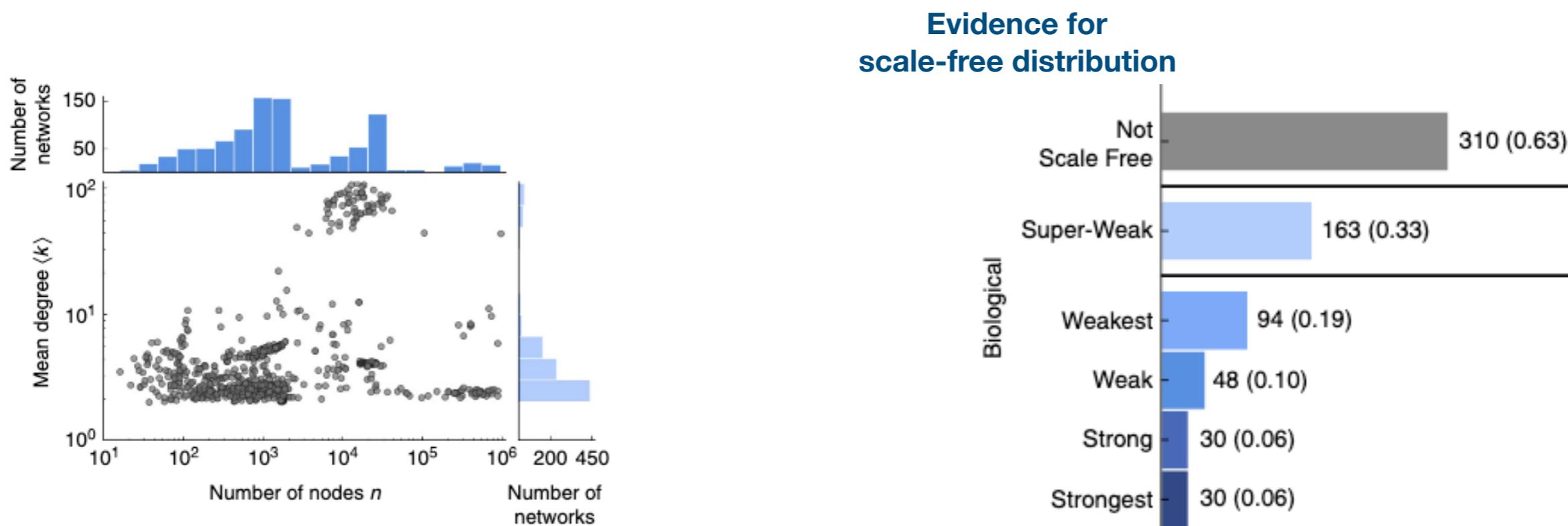
Highly central hubs tend **not** to be connected in biological networks: they are **disassortative**

(social networks: **assortative**)



7. Not all networks show scale-free behavior

Analysis of many types of networks shows that scale-free distributions are rare



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
- [Scale free networks are rare](#)
- [Modularity and community structure in networks](#)

Additional references displayed as hyperlinks in each figure.