Multi-omics data integration

appendix

Recap of Integration Strategies

Presented by the group 8

Classification of methods proposed by [Bersanelli M. et al. 2016]:

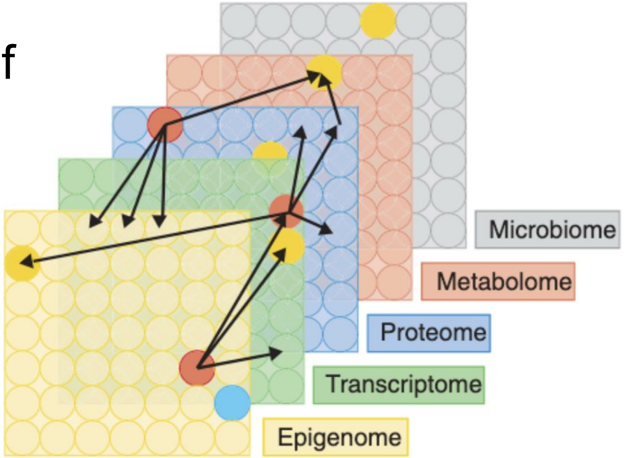
- Network-free, Non-Bayesian *(covered)*
- Network-free, Bayesian *(covered)*
- **Network-based, Non-Bayesian** (model relationships, deterministic)
- Network-based, Bayesian

Biological Networks | Intro

Which -omic layers do we want to integrate into our biological network?

Based on the integrated data we might have one of 2 types of constructed networks:

- **Intraomic** (within a single dataset)
- **Interomic** (combines the data across multiple omics datasets)



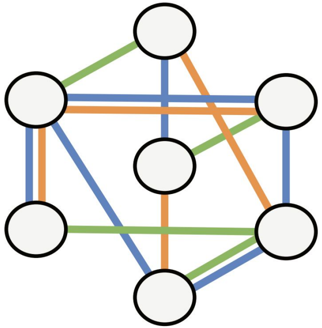
The choice what we what to integrate should be based on **the biological question**.

- Ours: gene-metabolite association studies → interomic network

[Hasin Y. et al. 2017]

Biological Networks | Intro

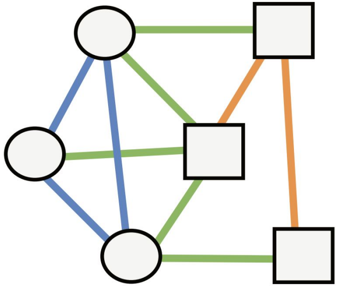
Which types of the heterogeneous network we are going to use?



Multiplex network

e.g. a molecular network where nodes are proteins, edges capture information about:

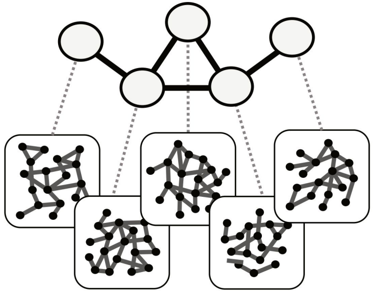
- physical interactions
- functional relationships
- sequence similarities



Typical heterogeneous network

e.g. a molecular network representing relationships among heterogeneous node types:

- genes
- transcripts
- proteins
- metabolites

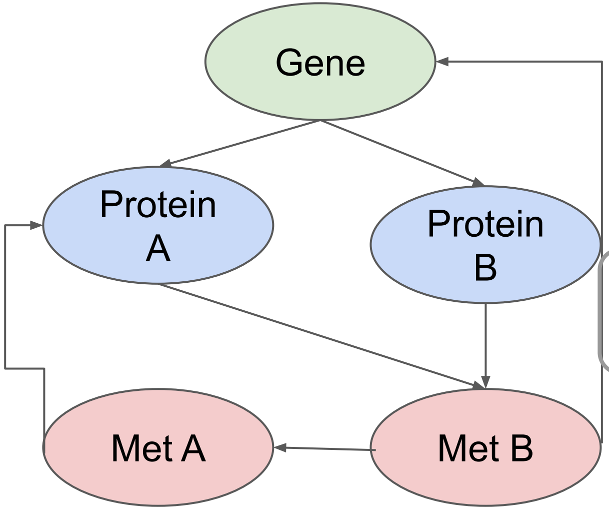


Network-of-networks

e.g. a PPI network where nodes (proteins) include structural information: amino acids are the nodes and edges link amino acids that are close enough in the protein's 3D fold

[Zitnik M. et al. 2024]

Gene-metabolite association studies



- Metabolic enzymes (catalyze chemical reactions)
- Kinases (phosphorylation)
- Phosphatases (dephosphorylation)

Cell regulation, signalling, metabolism

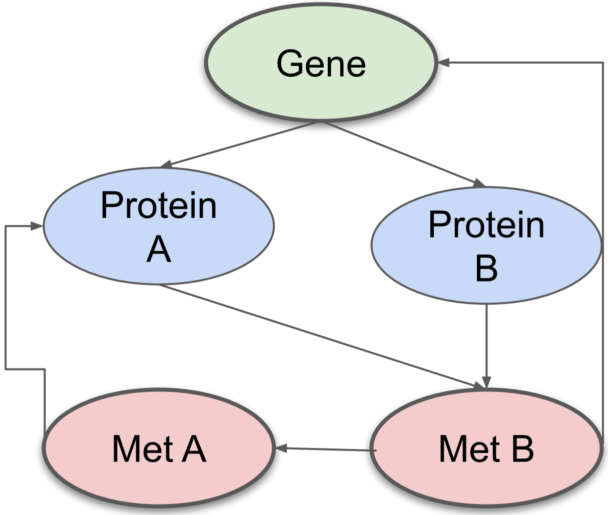
What for?

- Identifying the direct and indirect **relationships** between gene variations or expression levels and metabolites to explore **which genes** are associated with the **production/degradation/regulation** of metabolites

- How they are involved in the **metabolic pathways**

[Xu B. et al. 2024]

Gene-metabolite association network



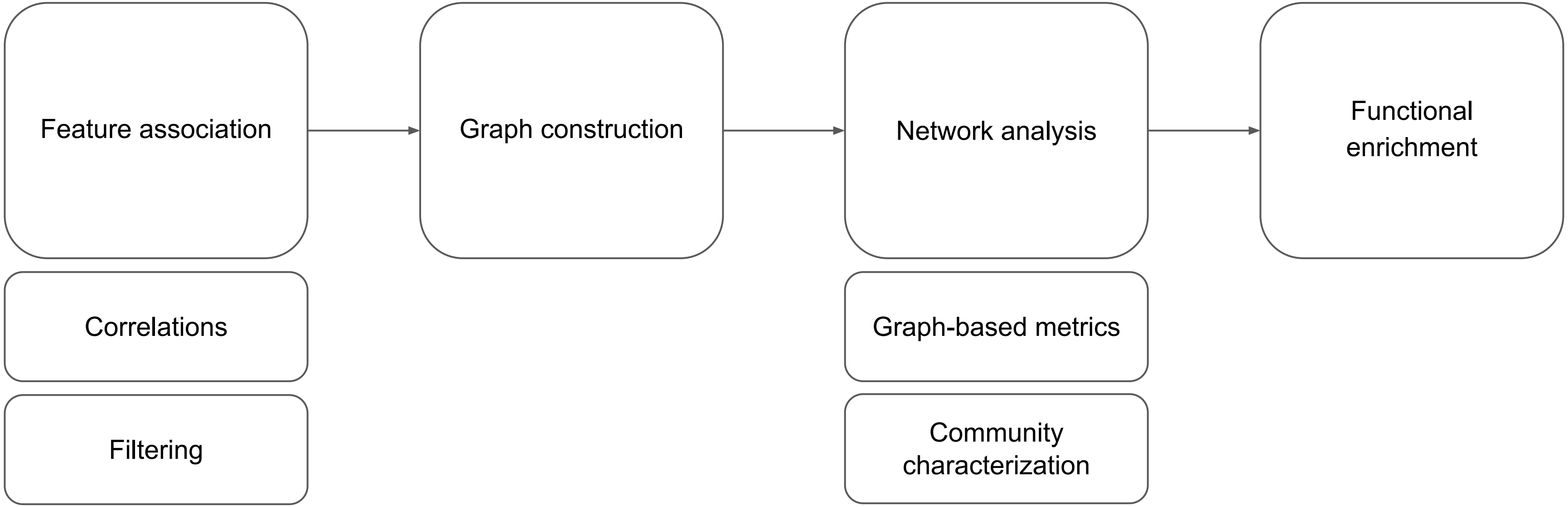
In the tutorial:

- We will focus on gene-metabolite entities
- The following relationships:
  - Gene-gene
  - Gene-metabolite
  - Metabolite-metabolite

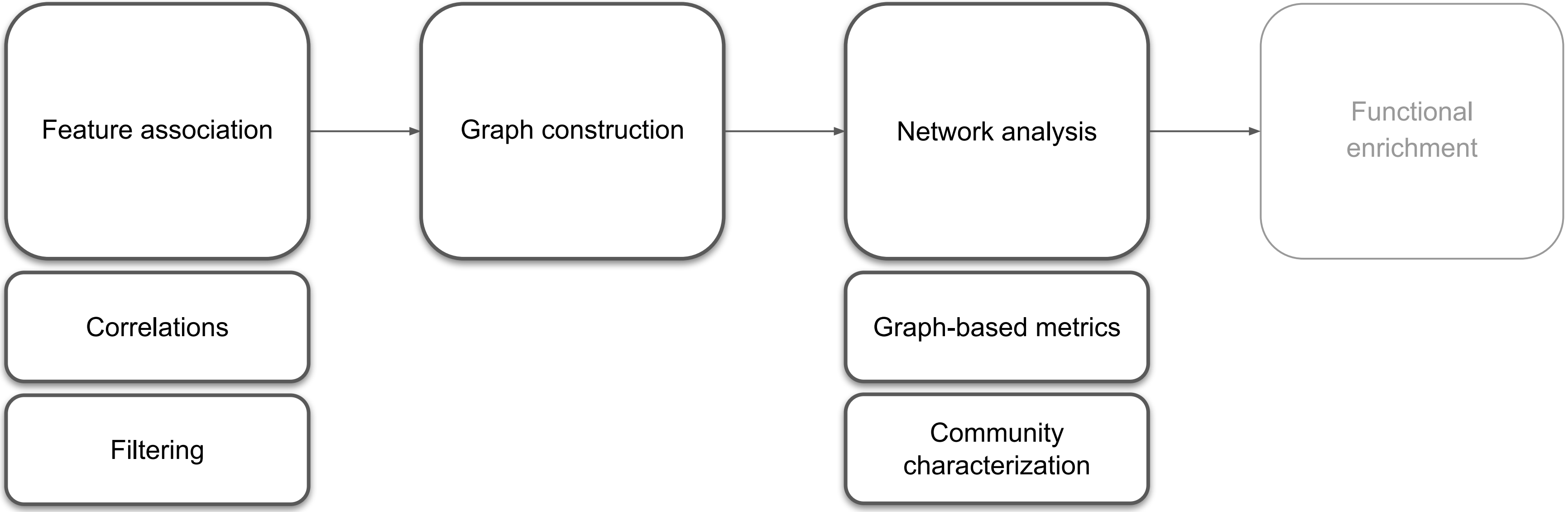
  would be considered as one type
- Indirect graph

[Xu B. et al. 2024]

The general workflow for the biological network analysis



The general workflow for the biological network analysis



Feature association

Correlations

Filtering

Graph construction

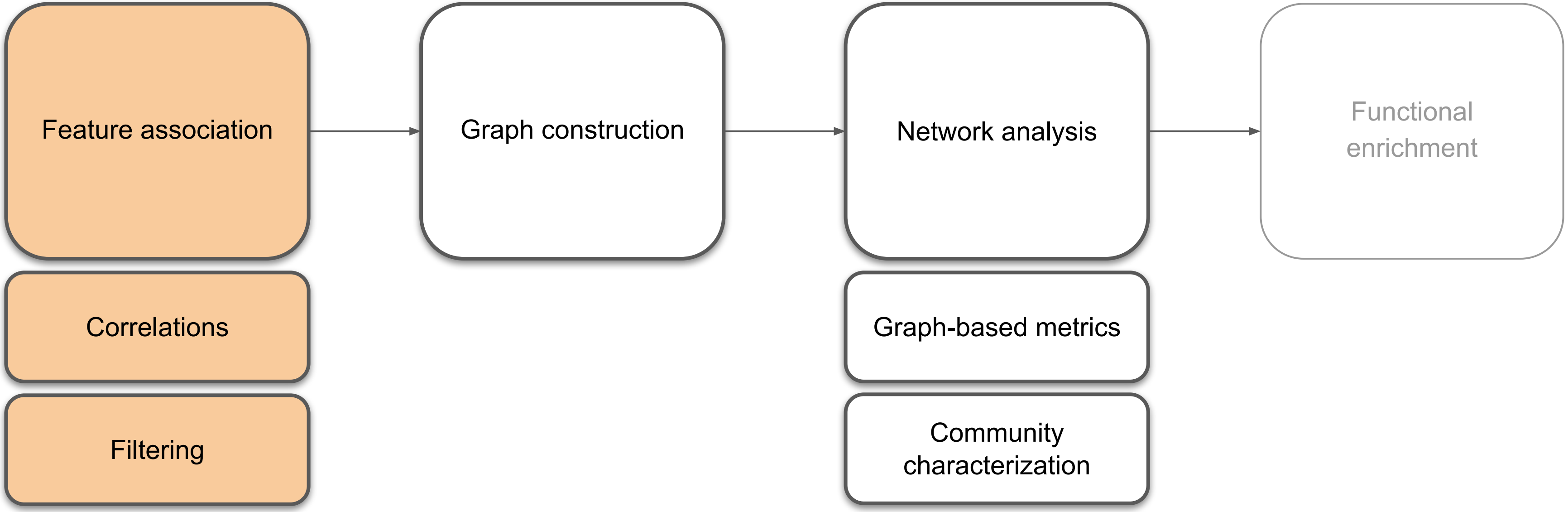
Network analysis

Graph-based metrics

Community characterization

Functional enrichment

The general workflow for the biological network analysis



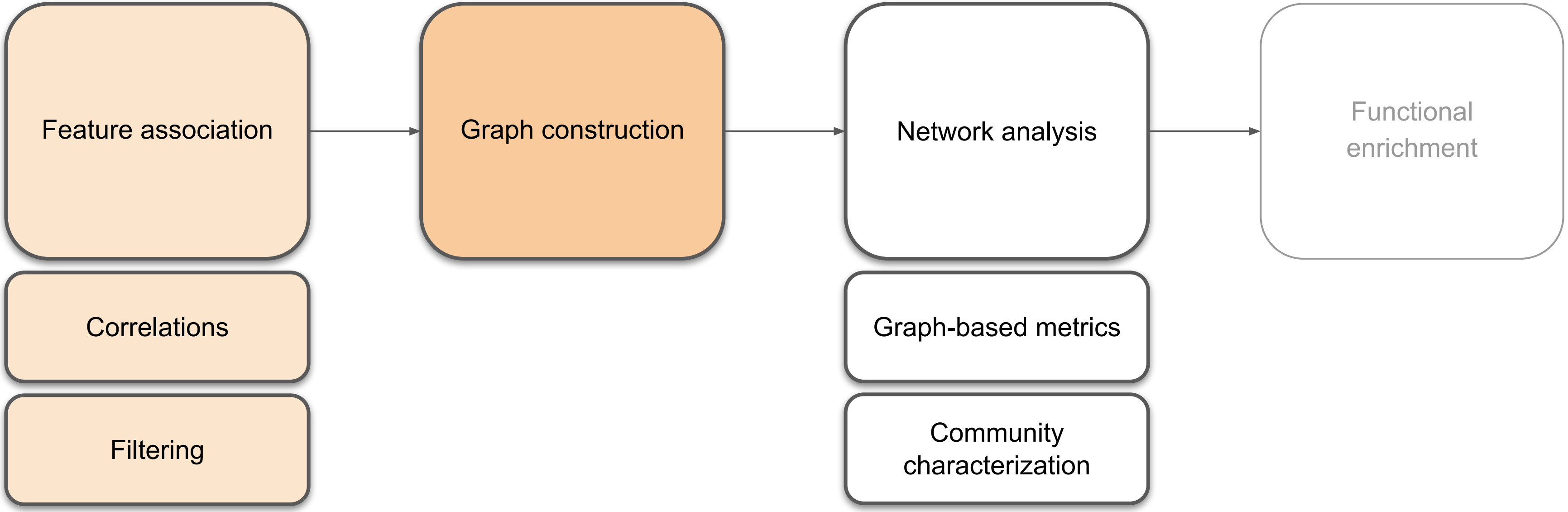
Feature association

- Standard pre-processing each of omics data, (in our case it is not needed)

- Correlation between different features
  - Bonferroni correction
  - FDR

- Filtering statistically insignificant correlation values

The general workflow for the biological network analysis



Feature association

Correlations

Filtering

Graph construction

Network analysis

Graph-based metrics

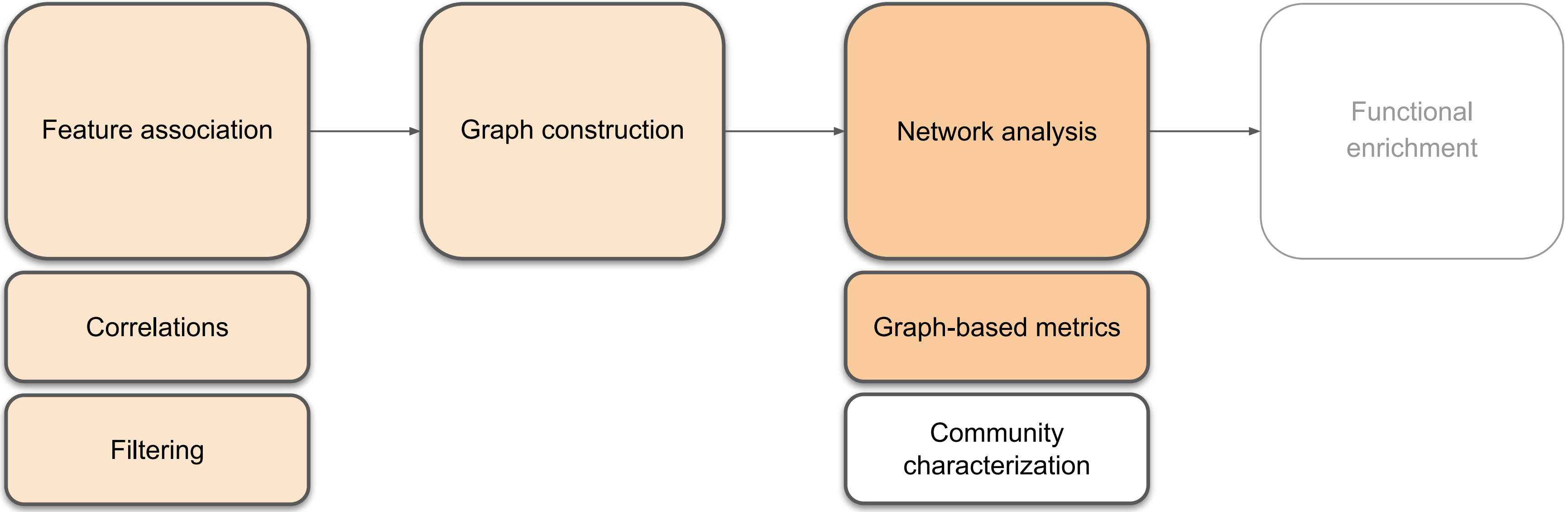
Community characterization

Functional enrichment

Graph construction

- We will focus on
  - Unweighted network based on the **filtered correlation matrix**
  - **KNN**-based graph
    - We need the **standardization** for KNN (based on Euclidean distance) to make metabolite and gene expression features comparable

The general workflow for the biological network analysis


Feature association

Correlations

Filtering

Graph construction

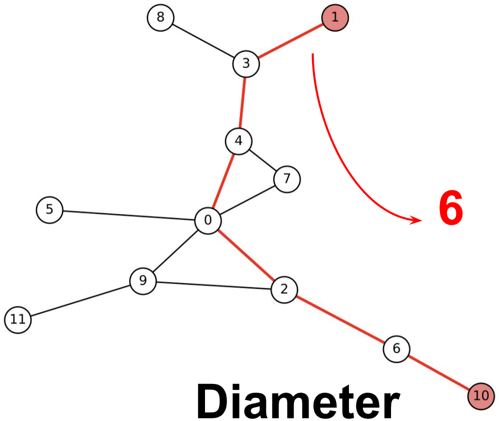
Network analysis

Graph-based metrics

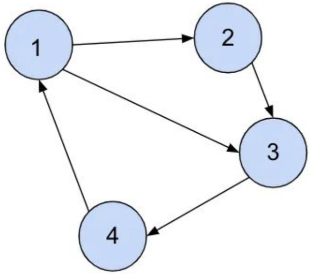
Community characterization

Functional enrichment

Network analysis | Graph-based metrics



This example illustrates a direct graph!



Shortest path:
P(1,2) = 1
P(1,3) = 1
P(1,4) = (1,3) + (3,4) = 2
P(2,1) = (2,3) + (3,4) + (4,1) = 3
P(2,3) = 1
.....

Average path length:
(1 + 1 + 2 + 3 + 1 + ...) / (number of all shortest path)

Diameter

- the largest shortest distance between any two vertices in the graph

- larger graph diameter → more sparse and disconnected the network

Average path length

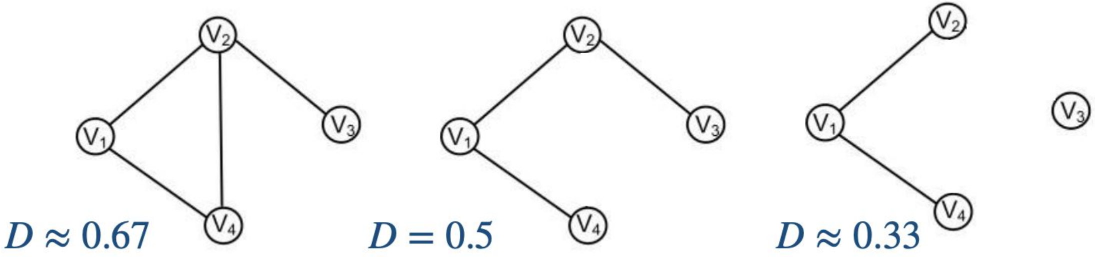
- the average distance between any two nodes in a graph

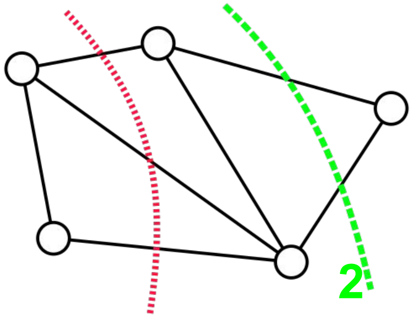
- can indicate the speed of information preparation in the graph

https://en.wikipedia.org/wiki/Diameter_%28graph_theory%29
https://medium.com/@kristinakhvatova/an-introduction-to-network-analysis-part-1-fa2a92455e34

Network analysis | Graph-based metrics

D = 2*E/V*(V-1)
E - number of edges
V - number of nodes





Density

- the ratio of the actual number of its edges and the largest possible number of edges it could have

- higher density → higher associations in the network → lower resilience to changes

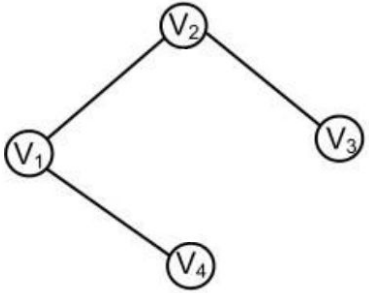
Minimum cut

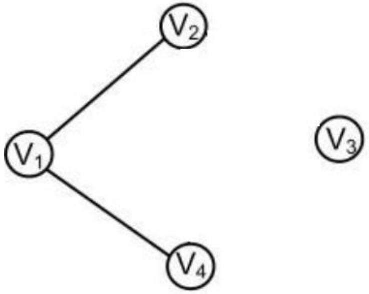
- the minimum number of edges which we need to remove to disconnect the graph

- indicates the weakest link or bottleneck in a network

https://github.com/NBISweden/workshop_omics_integration/blob/OMICSINT_H24/session_topology/lab/lectures/lecture_short.pdf
https://en.wikipedia.org/wiki/Minimum_cut

Network analysis | Graph-based metrics



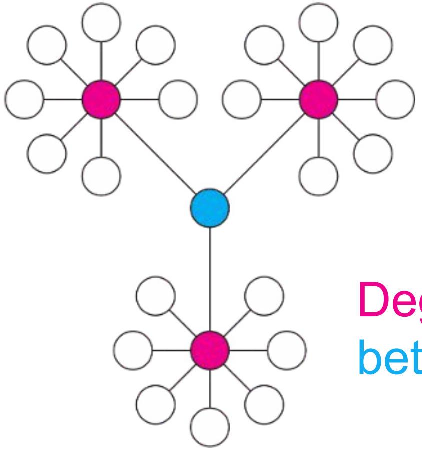


Connected graph

- there is at least 1 path connecting all nodes in a network

Disconnected graph

- some of the nodes are unreachable



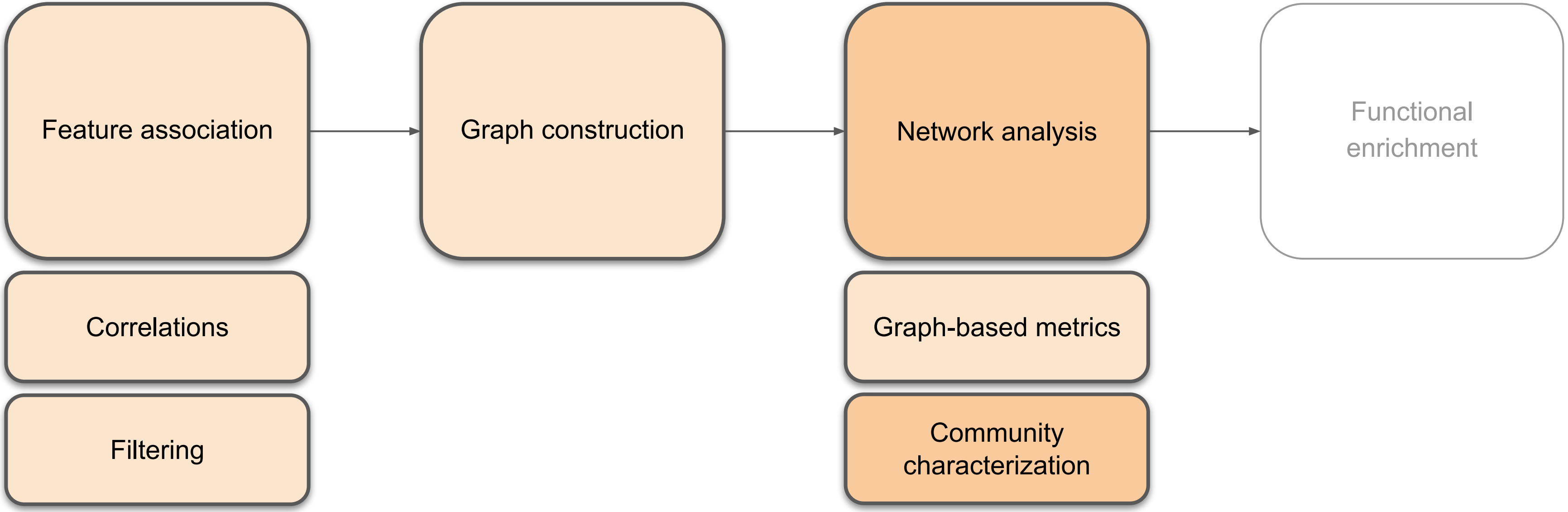
Degree vs betweenness

Centrality metrics (degree, closeness, betweenness, etc.)

- identification of the most important nodes

https://omicsforum.ca/t/what-are-the-differences-between-degree-and-betweenness-centrality/163

The general workflow for the biological network analysis



Feature association

Correlations

Filtering

Graph construction

Network analysis

Graph-based metrics

Community characterization

Functional enrichment

Network analysis | Community characterization

Louvain algorithm

Conceptually the same idea as for hierarchical clustering

0. Initialization:
   a. Assign to each node in the network its own community

1. Move nodes to the neighbouring community and save the changes if it increases the modularity (repeat until the value of modularity is not changed)

Level 1

2. Aggregate the graph info by reducing the communities into a single node

transition

3. Do the same procedure as in the step 1 for the aggregated version of the graph

Level 2

Stopping - the gain in modularity between 2 levels is less than threshold

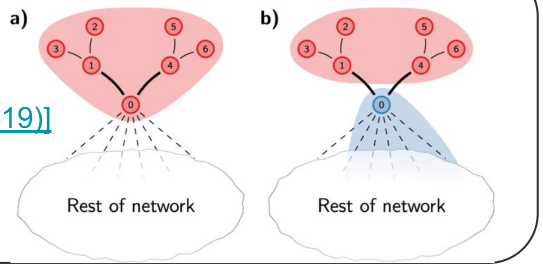
Modularity

- how well our network partitioned into the modules

- takes into account the difference between the **actual number** of edges between nodes in the community and the **expected number** of edges between them

Problem:

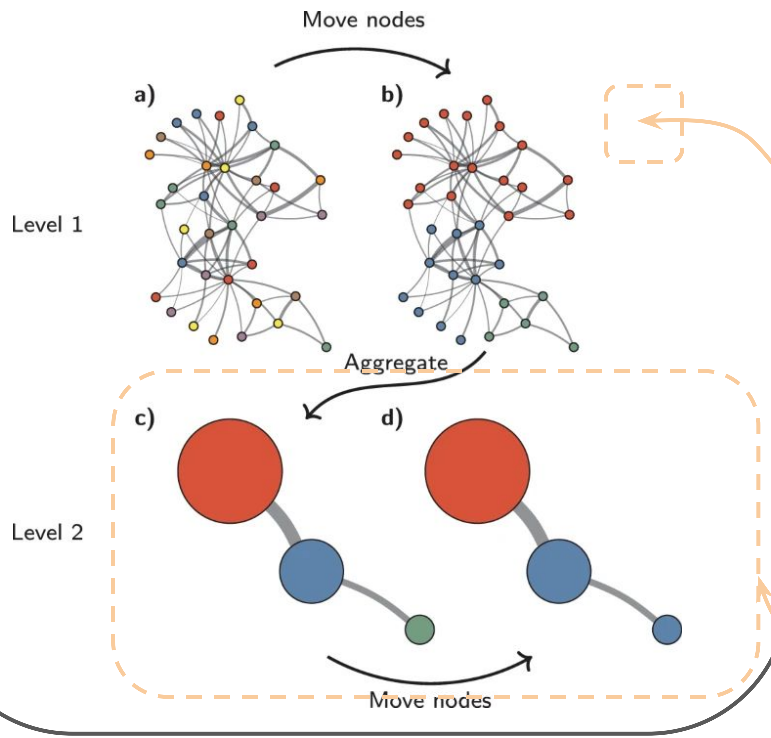
[Traag V. et al. (2019)]



https://en.wikipedia.org/wiki/Louvain_method

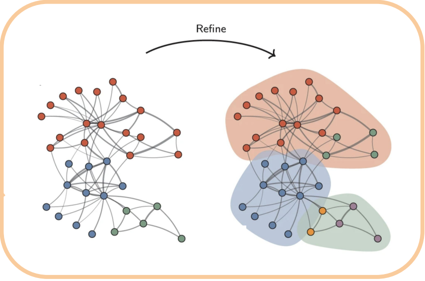
Network analysis | Community characterization

Louvain algorithm

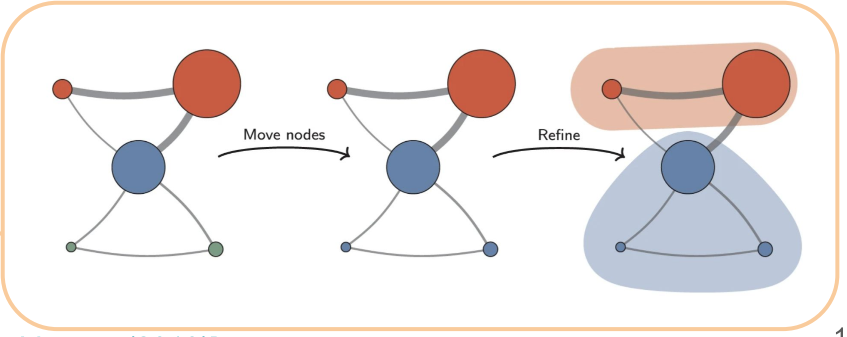


Leiden algorithm

Refinements are added



Largest possible community!



[Traag V. et al. (2019)]

Colab Notebook