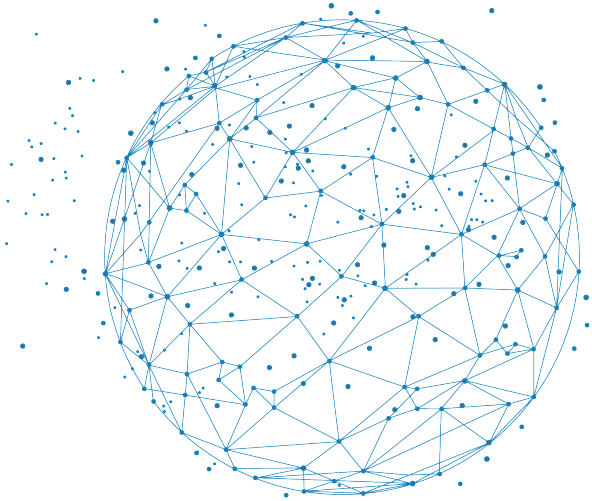


First edition 2023 in Fréjus



Theoretical part Networks

Anaïs Baudot
Morgane Térézol

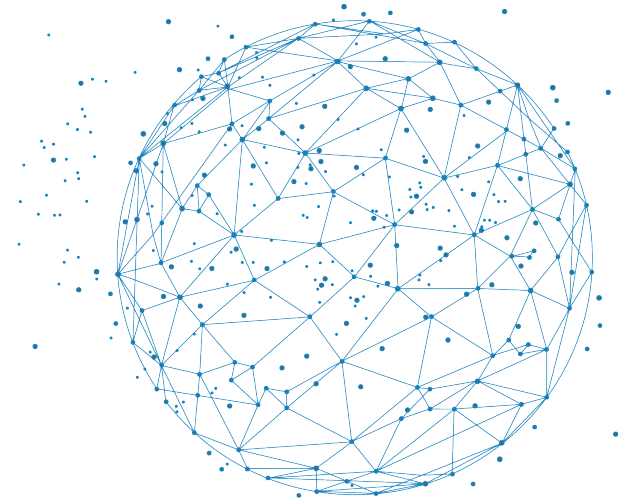
DOI version final



- Chapter 0: General introduction, key concepts
- Chapter 1: Network Construction
- Chapter 2: Network Analysis - measures
- Chapter 3: Network Analysis - algorithms
- Wrap-up

Chapter 0

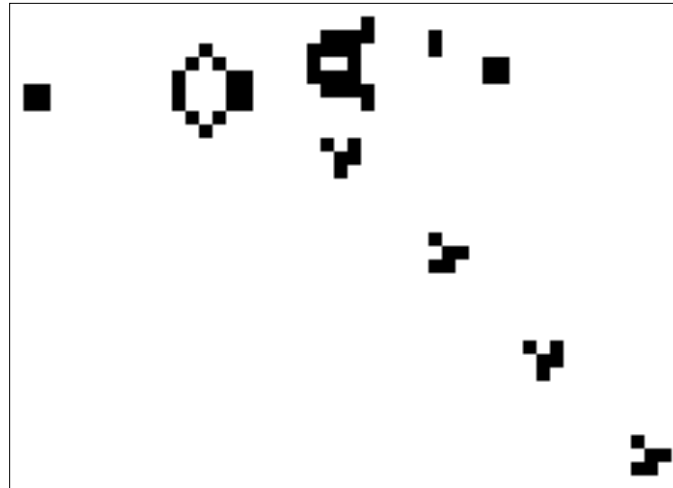
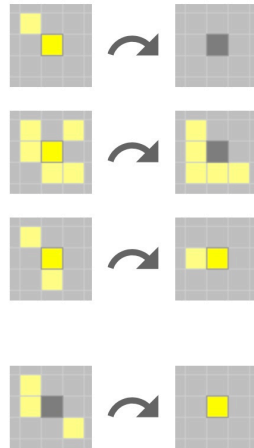
General Introduction





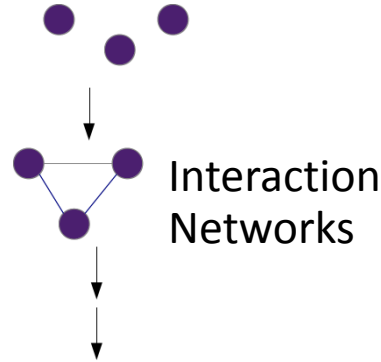
- many components
- global behavior cannot be deduced from knowledge on the components of the system
- interactions between components
- emerging properties

The sum is greater than the parts





Systems component : genes/proteins



Emerging properties : phenotypes

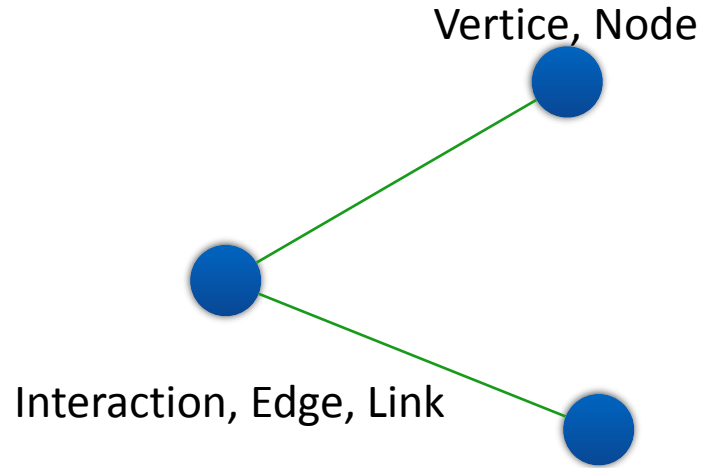


Phenotype does not emerge from isolated biological molecules but from their interactions



$G = (V, E)$, where V is the set of vertices and $E \subseteq (V \times V)$ is the set of edges

Graph, Network, Web



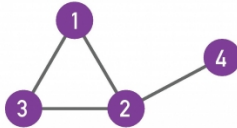


Adjacency matrix

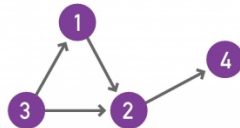
a. Adjacency matrix

$$A_{ij} = \begin{matrix} & A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & & A_{22} & A_{23} & A_{24} \\ A_{31} & & A_{32} & A_{33} & A_{34} \\ A_{41} & & A_{42} & A_{43} & A_{44} \end{matrix}$$

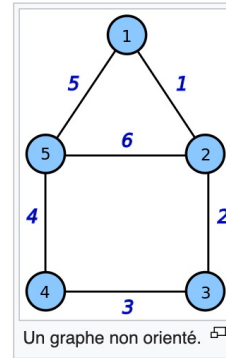
b. Undirected network



c. Directed network



Incidence matrix



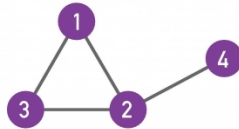


Adjacency matrix

a. Adjacency matrix

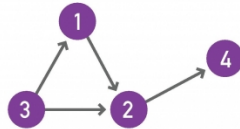
$$A_{ij} = \begin{matrix} & A_{11} & A_{12} & A_{13} & A_{14} \\ & A_{21} & A_{22} & A_{23} & A_{24} \\ & A_{31} & A_{32} & A_{33} & A_{34} \\ & A_{41} & A_{42} & A_{43} & A_{44} \end{matrix}$$

b. Undirected network



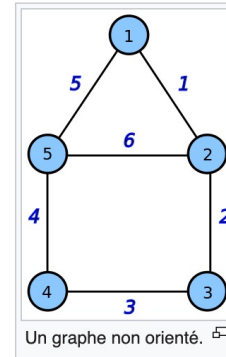
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

c. Directed network



$$A_{ij} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Incidence matrix

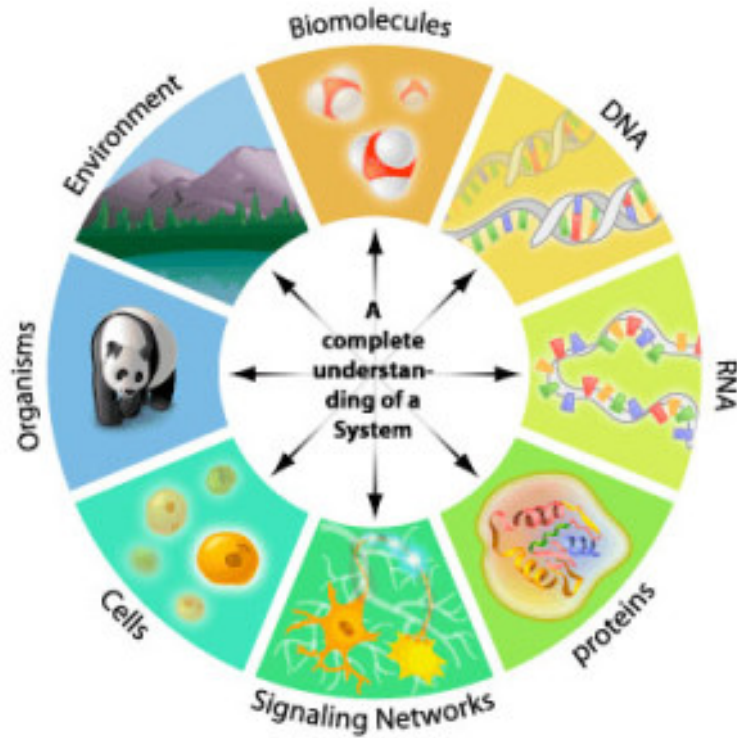


edges

nodes

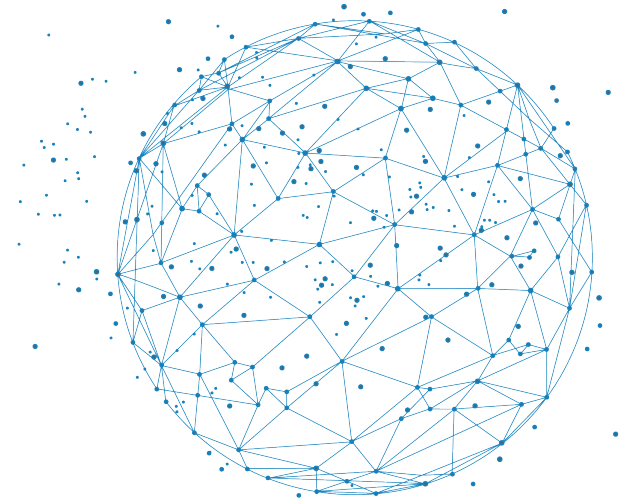
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

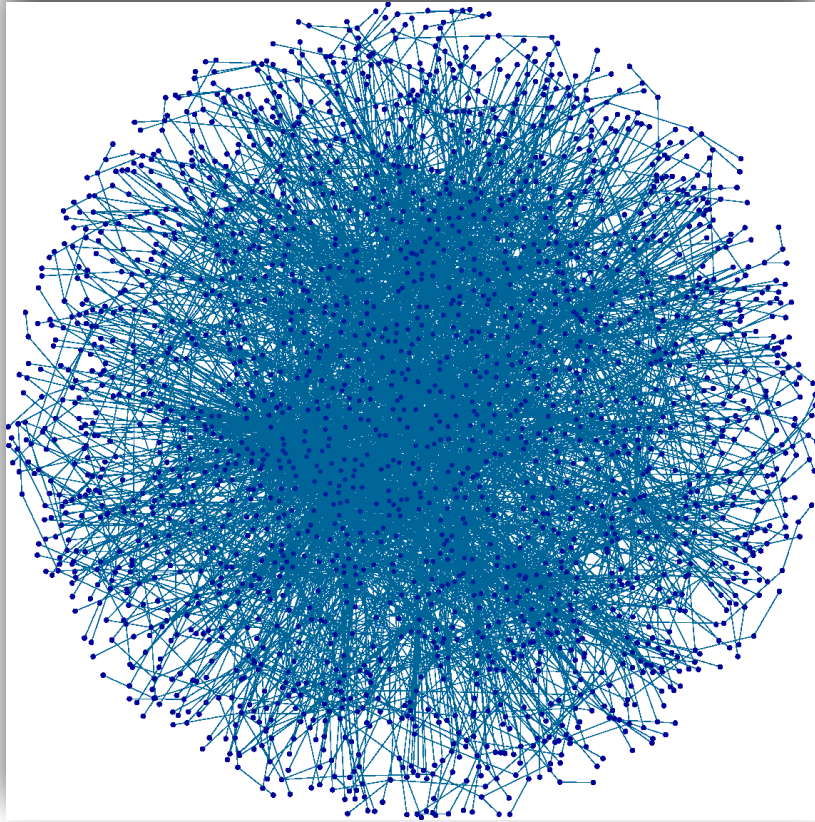
Key concept 3: Multi-scale Data



Chapter 1

Network Construction



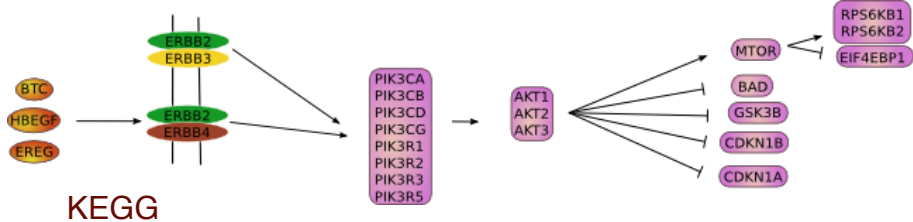


Data sources?

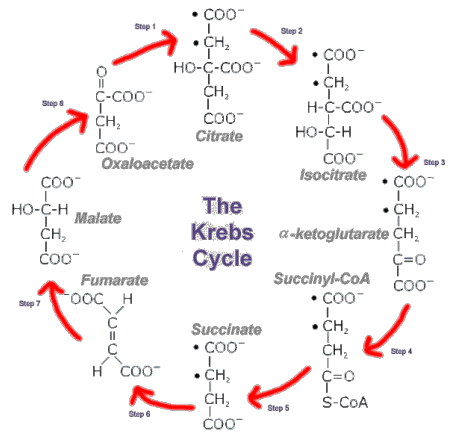
- Expert knowledge and literature
- High-throughput screenings
- Inference from omics data



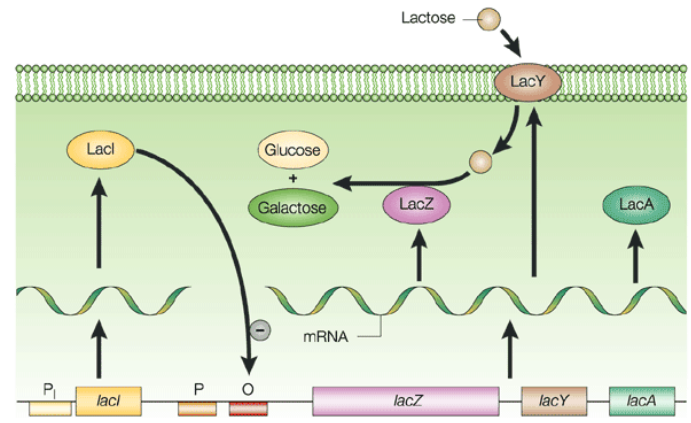
Signaling networks



Operon / Regulatory networks

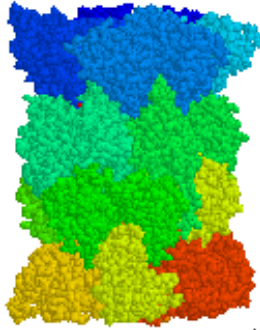


Metabolic Cycle

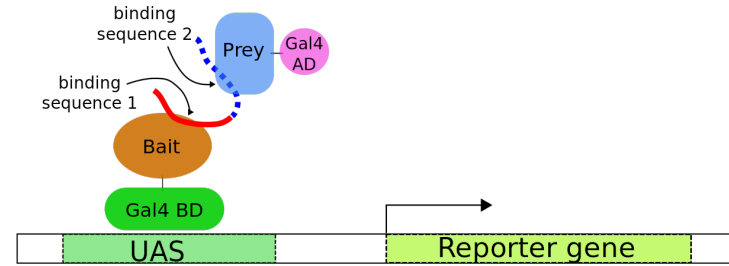


Opéron lactose

Nature Reviews | Genetics
Shuman et al.

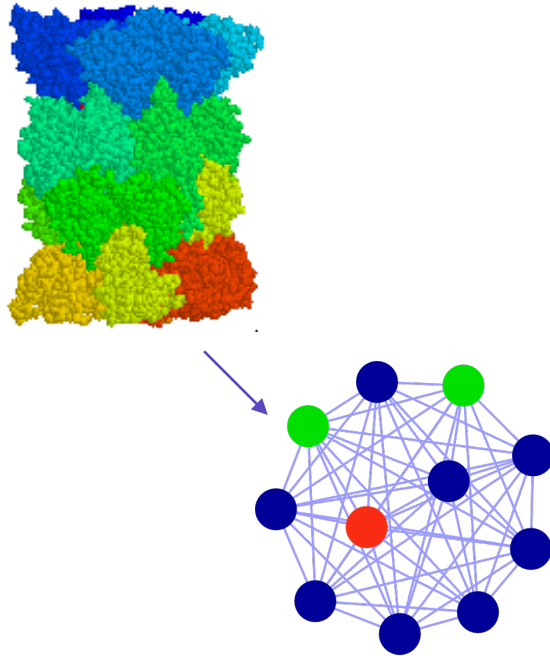


Molecular complex interactions
=> GFP-trap

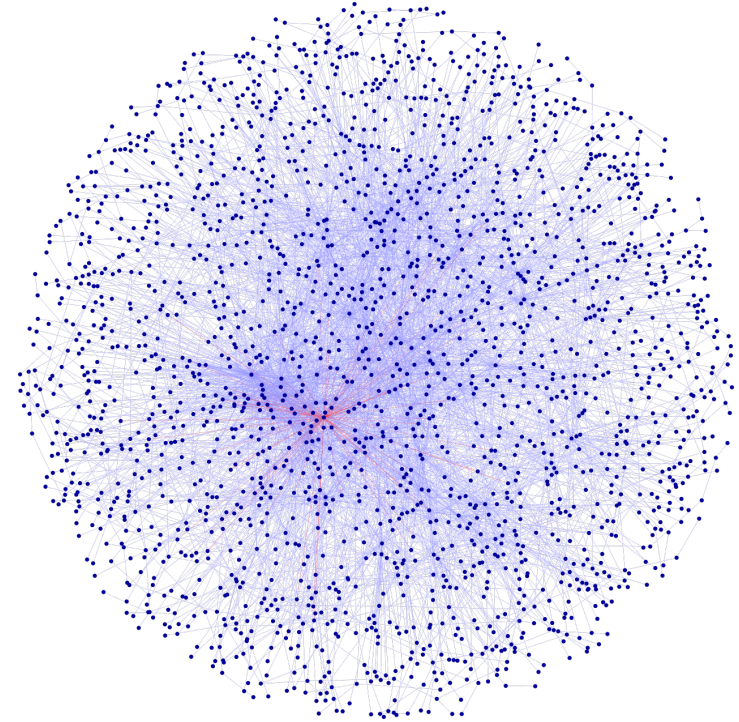


Protein-Protein interactions
=> Yeast 2-hybrids screens

Thousands of interactions in model organisms (since 2001)



Protein complexes



Protein-protein interaction networks

DIP (dip.doe-mbi.ucla.edu)

IntAct (www.ebi.ac.uk/intact)

MINT (mint.bio.uniroma2.it/mint)

BioGRID (www.thebiogrid.org)

BIND (www.blueprint.org)

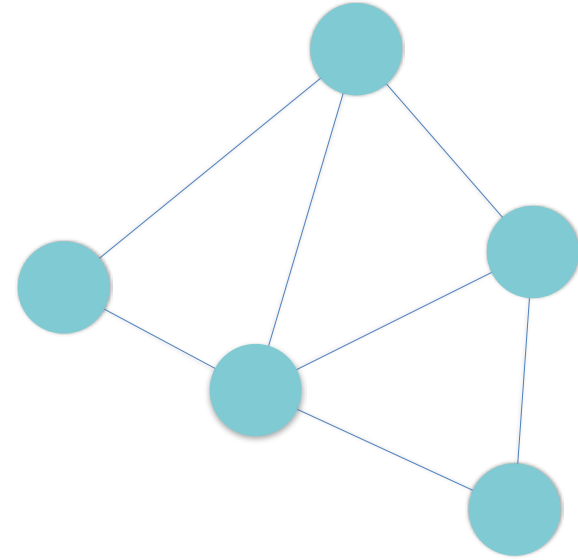
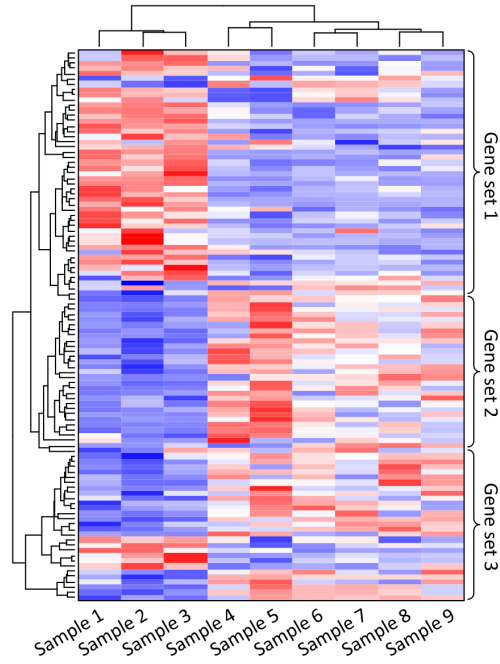
Reactome (<http://reactome.org/>)

NDEX (<https://www.ndexbio.org/>)



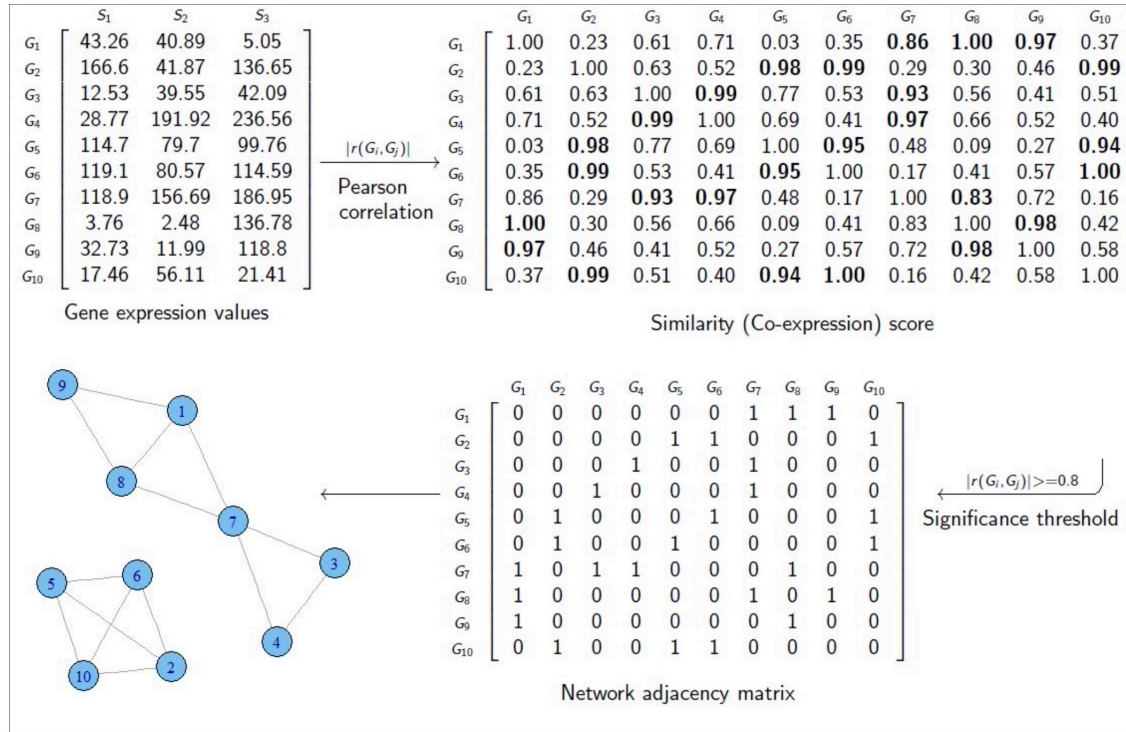
**International
Molecular
Exchange
Consortium**

Data source 3: Inference from omics data





Example: inference of co-expression networks from transcriptomics data





Greenfield A, Madar A, Ostrer H, Bonneau R (2010) DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. PLoS ONE 5(10): e13397.

Saint-Antoine, M. M. & Singh, A. Network inference in systems biology: recent developments, challenges, and applications. *Current Opinion in Biotechnology* **63**, 89–98 (2020).

Algorithm Class	Temporal Data Required?	Directionality	Advantages	Disadvantages	Examples
Correlation	No	Undirected	<ul style="list-style-type: none"> • Fast, scalable • Detection of feed-forward loops, fan-ins, and fan-outs 	<ul style="list-style-type: none"> • Possibly over-simplistic • False positives for cascades 	WGCNA [13] PGCNA [14]
Regression	No	Directed	<ul style="list-style-type: none"> • Good overall accuracy 	<ul style="list-style-type: none"> • Bad detection of feed-forward loops, fan-ins, and fan-outs 	TIGRESS [15], GENIE3 [16], bLARS [17]
Bayesian - Simple	No	Directed	<ul style="list-style-type: none"> • Performance on small networks 	<ul style="list-style-type: none"> • Performance on large networks. • Inability to detect cycles 	[19,20]
Bayesian - Dynamic	Yes	Directed	<ul style="list-style-type: none"> • Performance on small networks • Detection of cycles and self-edges 	<ul style="list-style-type: none"> • Performance on large networks. 	[21]
Information Theory	No	Undirected (at least in simplest form)	<ul style="list-style-type: none"> • Detection of feed-forward loops, fan-ins, and fan-outs • Similar to correlation methods, with better accuracy 	<ul style="list-style-type: none"> • False positives for cascades 	ARACNE [25], CLR [26], MRNET [27], PIDC [28]
Phixer	No	Directed	<ul style="list-style-type: none"> • Parsimonious output due to pruning step. 	<ul style="list-style-type: none"> • Possible loss of overall accuracy due to pruning step (this can be removed if the user chooses) 	[31]



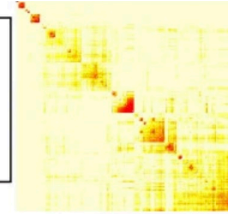
Expression data



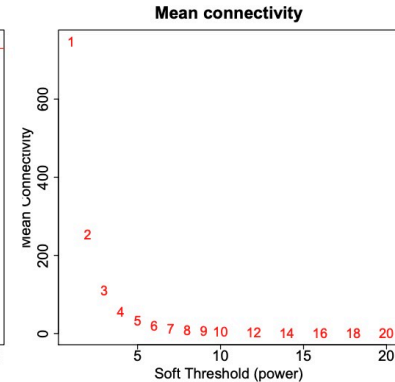
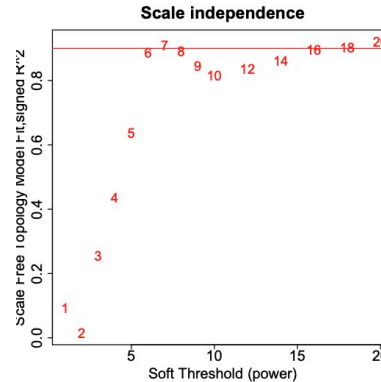
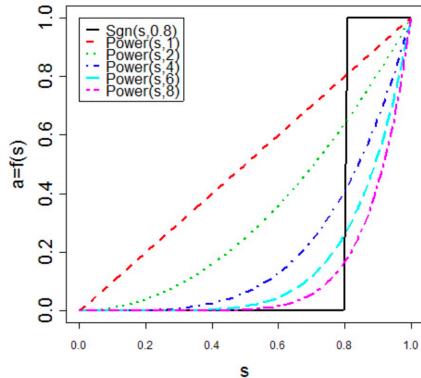
Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

Tools: correlation as a measure of co-expression



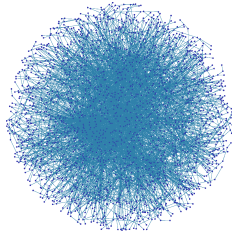
Additional step of soft-thresholding



Conclusion: Many Biological Networks



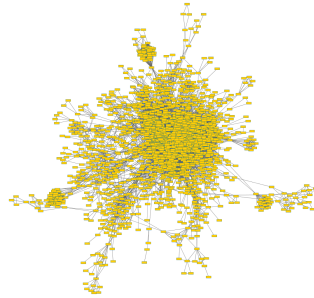
PPI



~60 000 edges

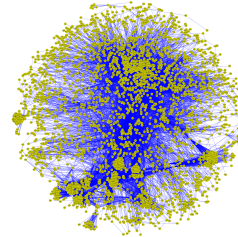
Experimental
networks

Complexes



~40 000 edges

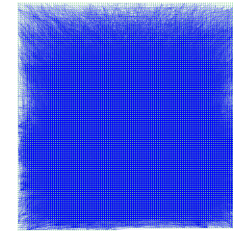
Pathways



~250 000 edges

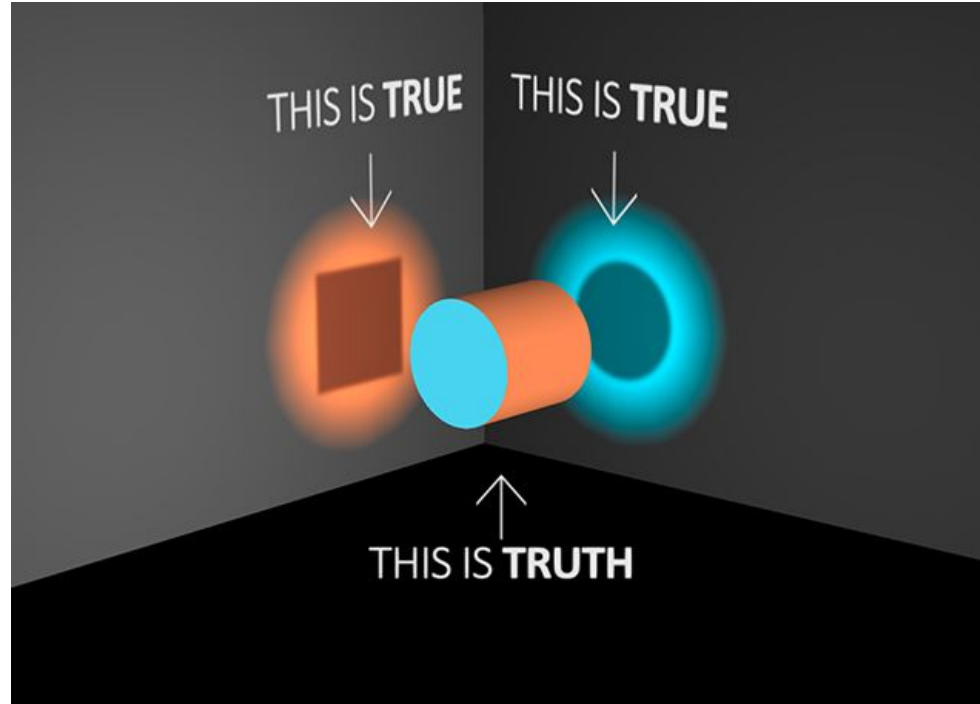
Curated
networks

**Correlation of
expression**



~1 400 000 edges

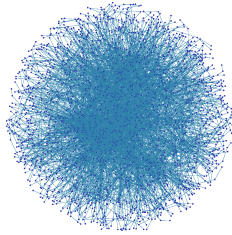
Inferred
networks



Conclusion: Many Biological Networks



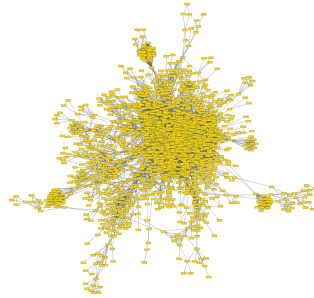
PPI



~60 000 edges

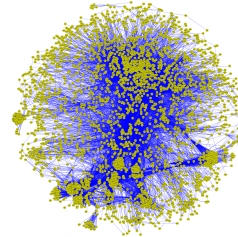
Experimental
networks

Complexes



~40 000 edges

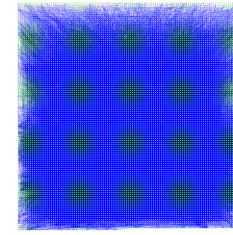
Pathways



~250 000 edges

Curated
networks

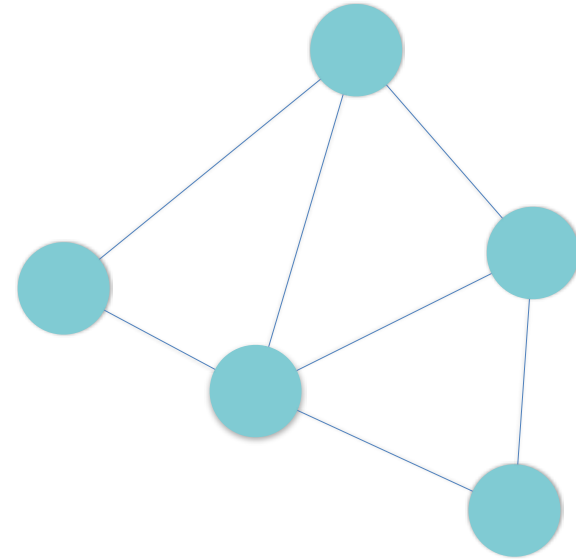
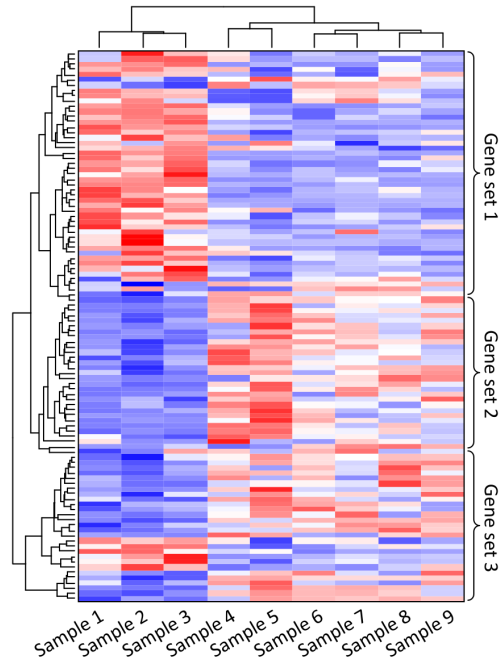
**Correlation of
expression**



~1 400 000 edges

Inferred
networks

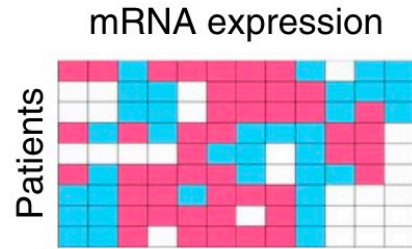
Data source 3: Inference from omics data



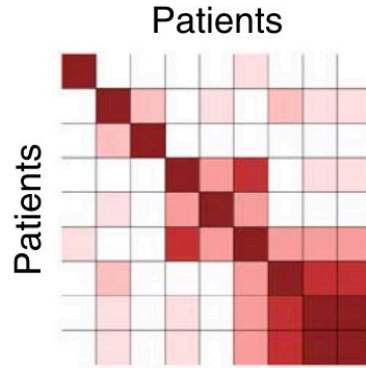
Network of features?
Network of samples?



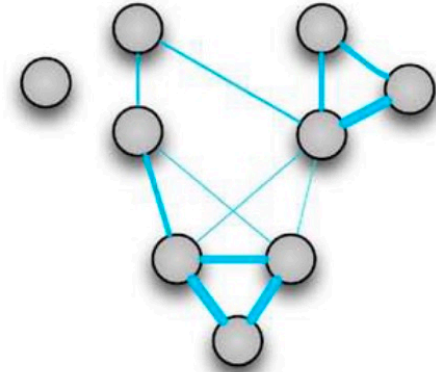
a Original data



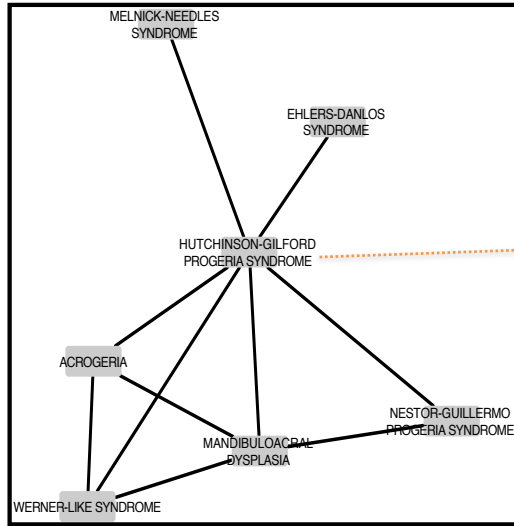
b Patient similarity matrices



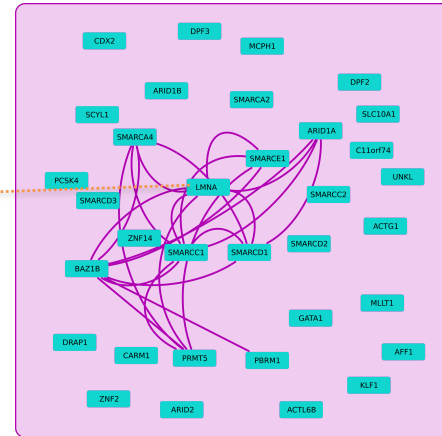
c Patient similarity networks



SNF, Wang et al. 2014

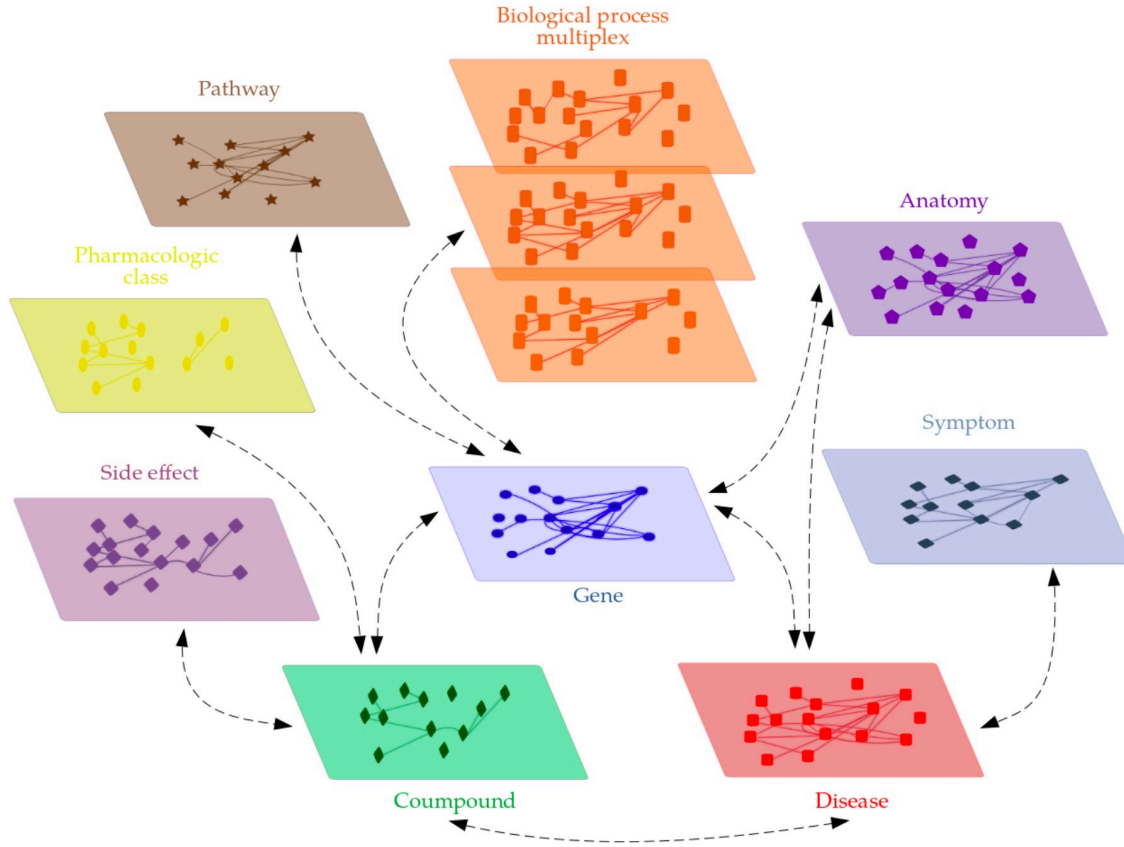


Disease-Disease Network



Molecular Network (e.g. PPI)

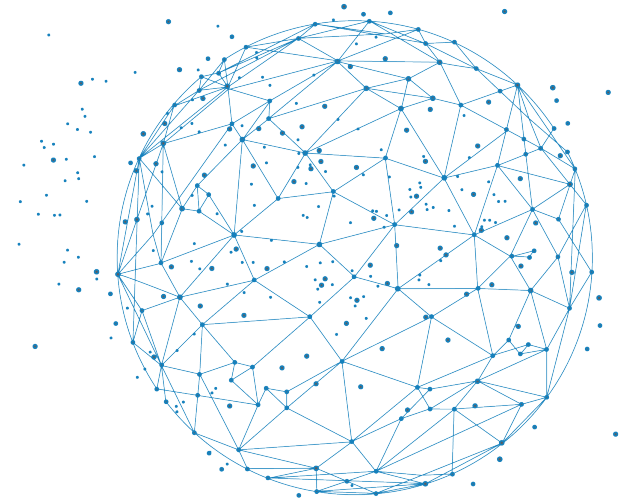
Multi-layer Networks \Leftrightarrow Knowledge graphs ?



Hetionet, adapted
from Himmelstein et
al. 2017

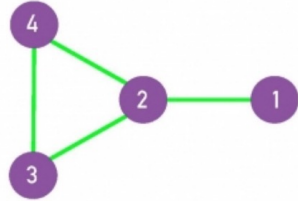
Chapter 2

Network Analysis - Measures

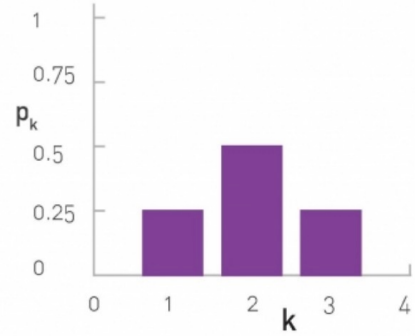




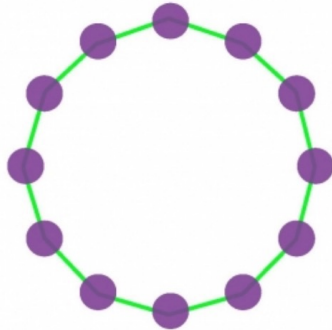
a.



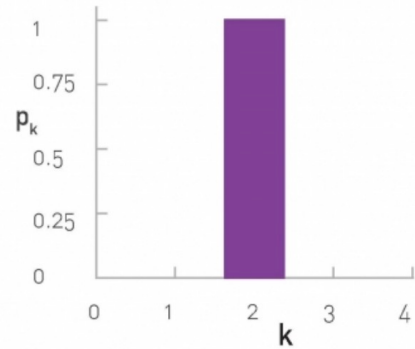
b.

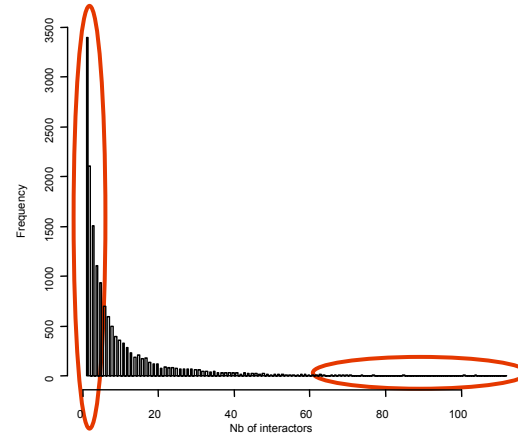
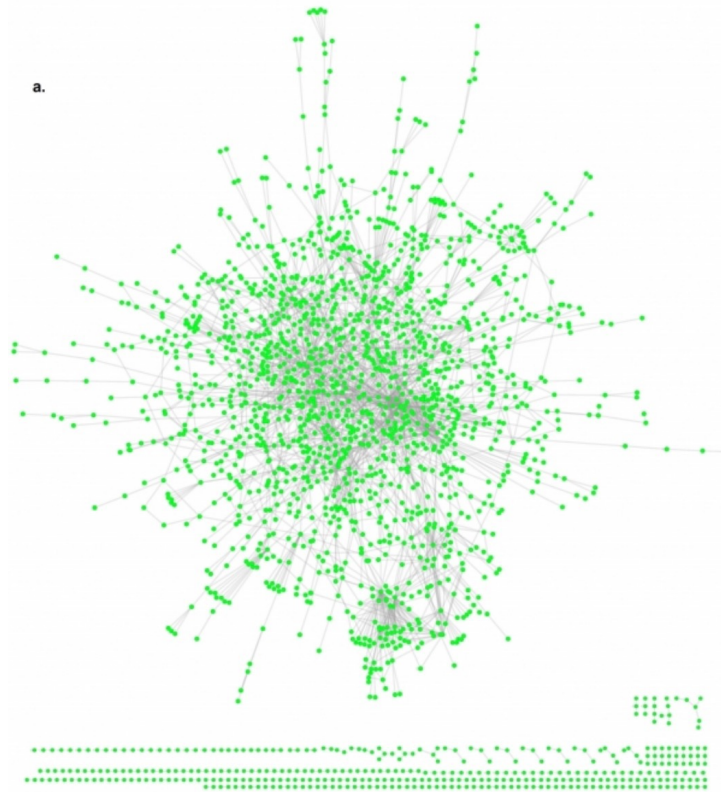


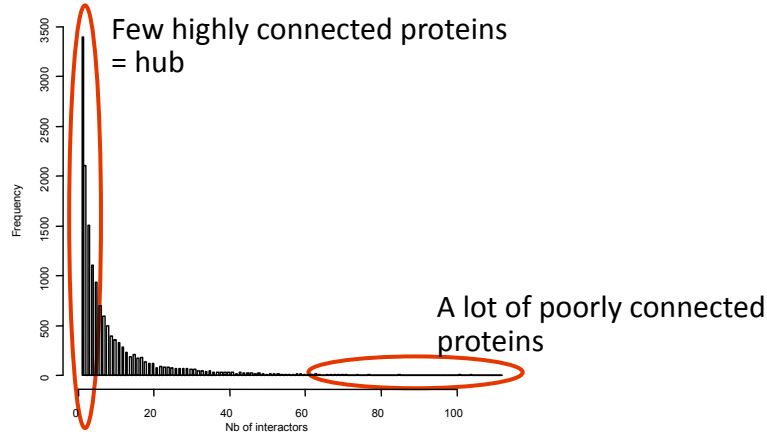
c.



d.

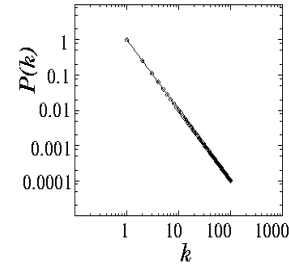






Scale-free networks

Power-law distribution



Biological interpretation?

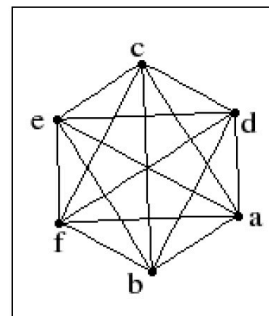
- Growth with preferential attachment (“rich get richer”) => create “hubs”
- Robust to random attack, sensitive to targeted attacks

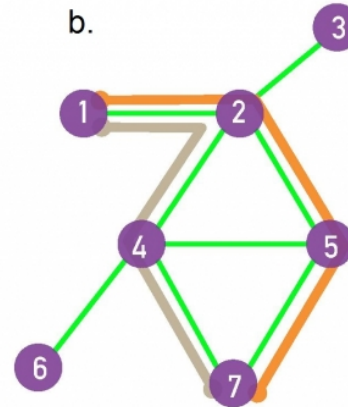
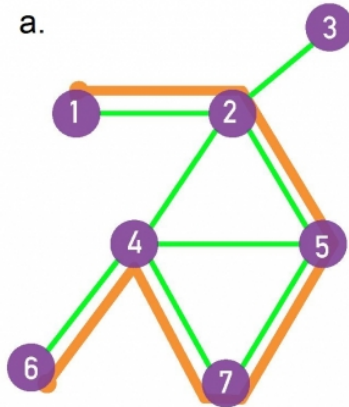


$$D(G) = \frac{|E|}{|E|_{K_n}}$$

- ▶ $D_G = 0 \Leftrightarrow |E| = 0$
- ▶ $D_G = 1 \Leftrightarrow G = K_n$

Graphe complet K_n

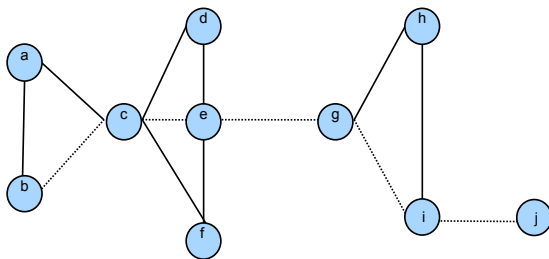




a) Walking from nodes 1 and 6 through the orange path

b) Smallest distance between 1 and 7

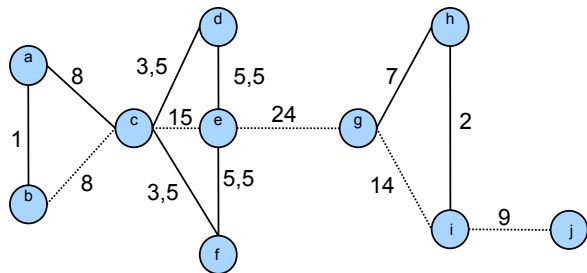
Network diameter?

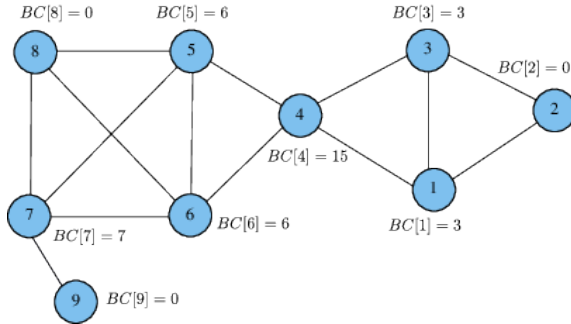


Number of shortest paths
running through an edge

=

"bottleneck"

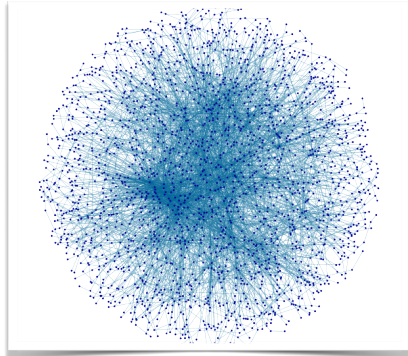




Number of shortest paths running through a node

=

“bottleneck”



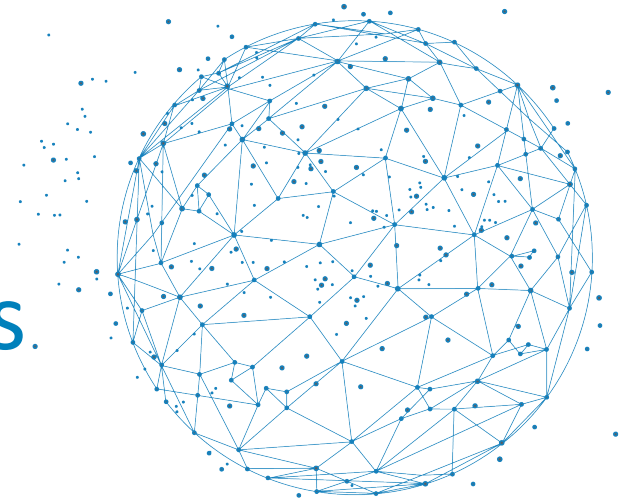
Biological interpretation ?

Correlation with gene essentiality,
gene involvement in diseases,
importance in flux transmission ...

Chapter 3

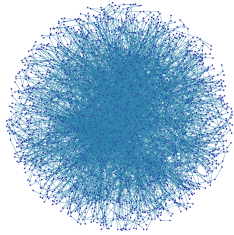
Network Analysis - Algorithms.

- Random walks / diffusion
- Clustering/partitionning
- Active modules
- Embedding



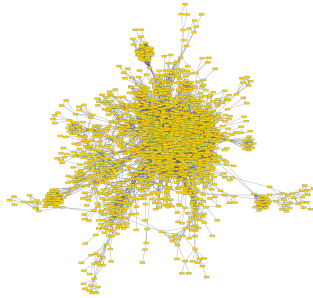


PPI



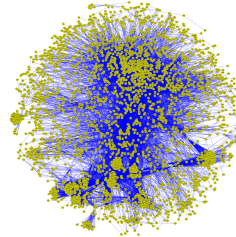
~60 000 edges

Complexes



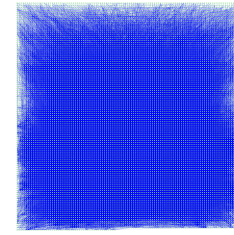
~40 000 edges

Pathways



~250 000 edges

Correlation of expression

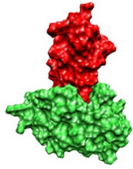


~1 400 000 edges

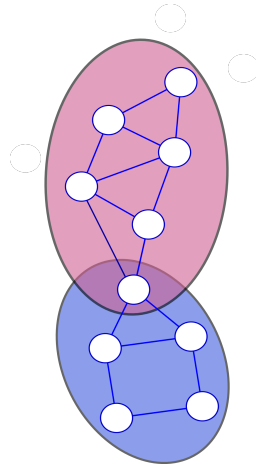
Experimental
networks

Curated
networks

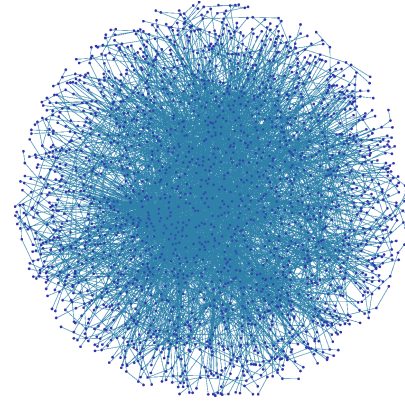
Inferred
networks



Binary interaction



modules



Interaction Networks



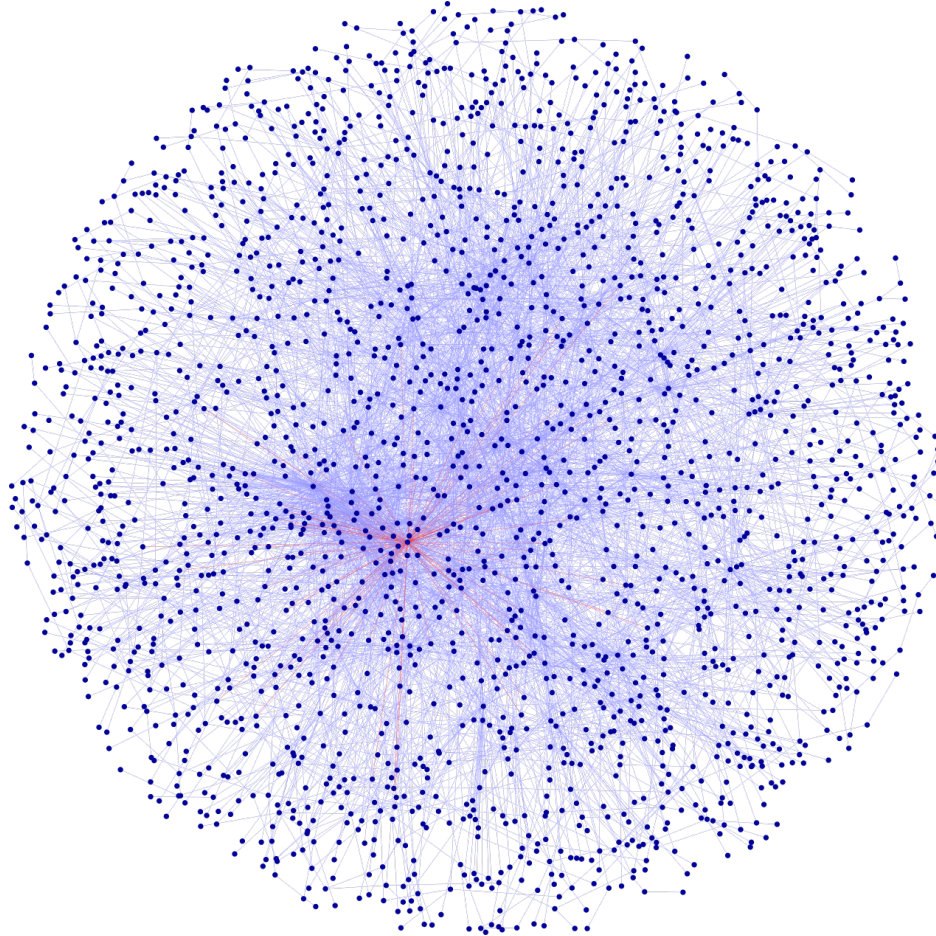


- Communities around node(s) of interest
- Communities obtained from the partitionning of the full network
- Communities obtained from the integration of quantitative data on network nodes, aka active modules

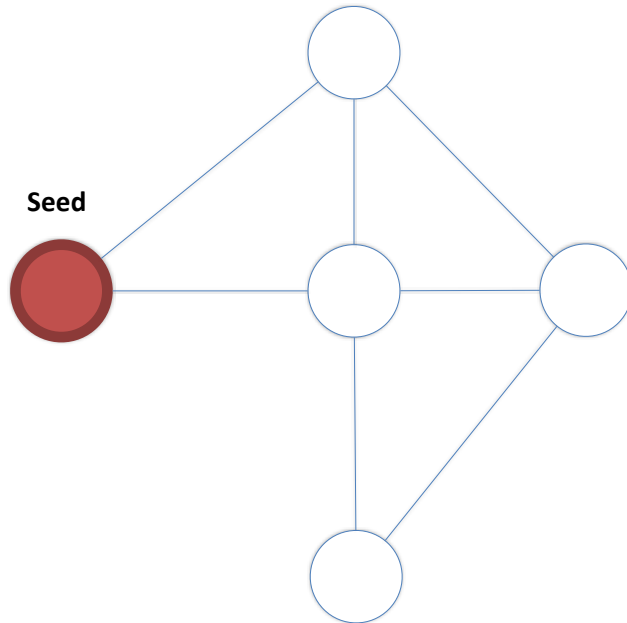
Topology alone

Topology +
quantitative data

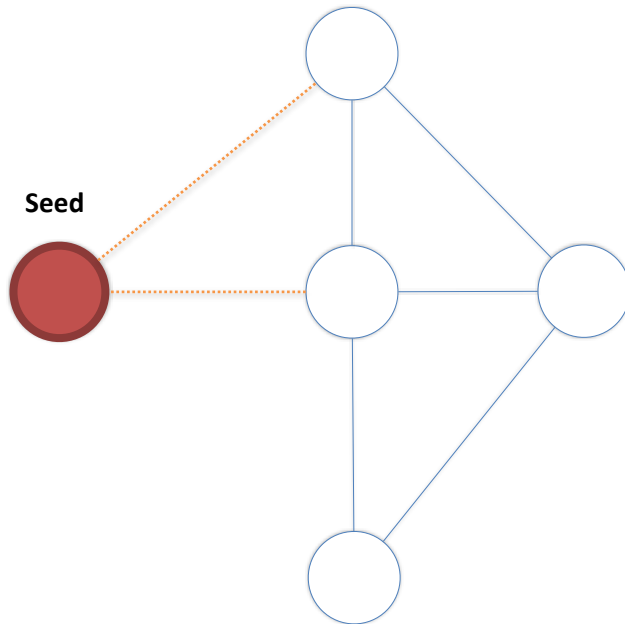
Random Walk with Restart (RWR)

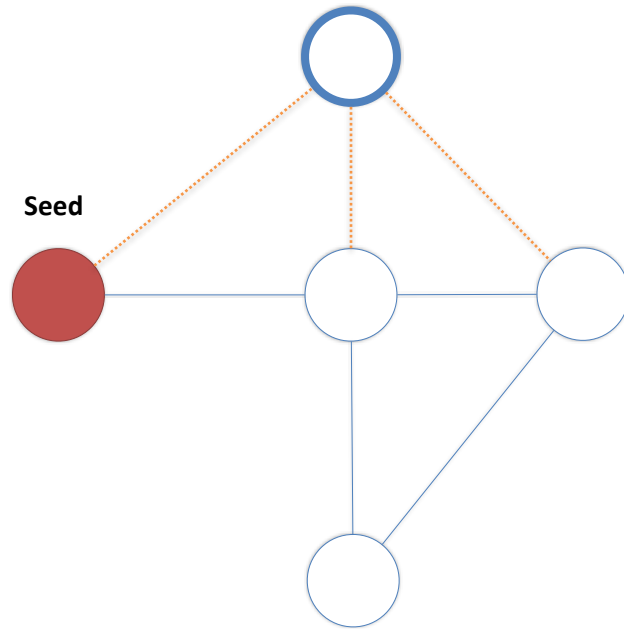


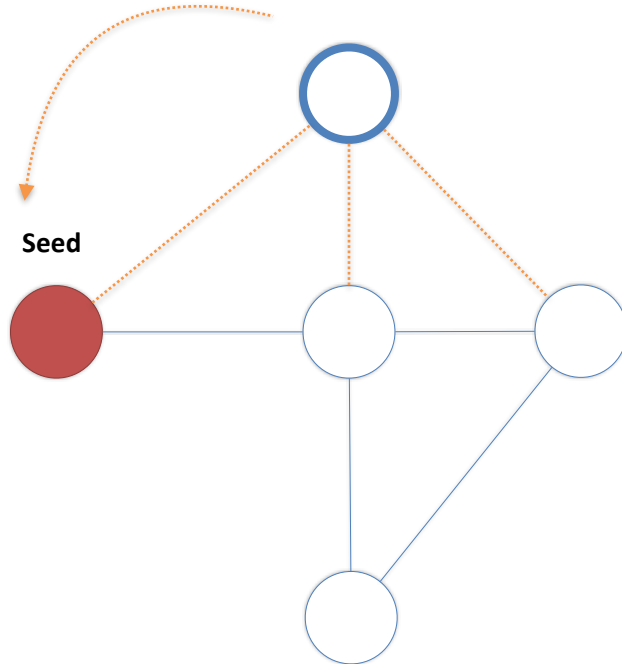
Random Walk with Restart (RWR)



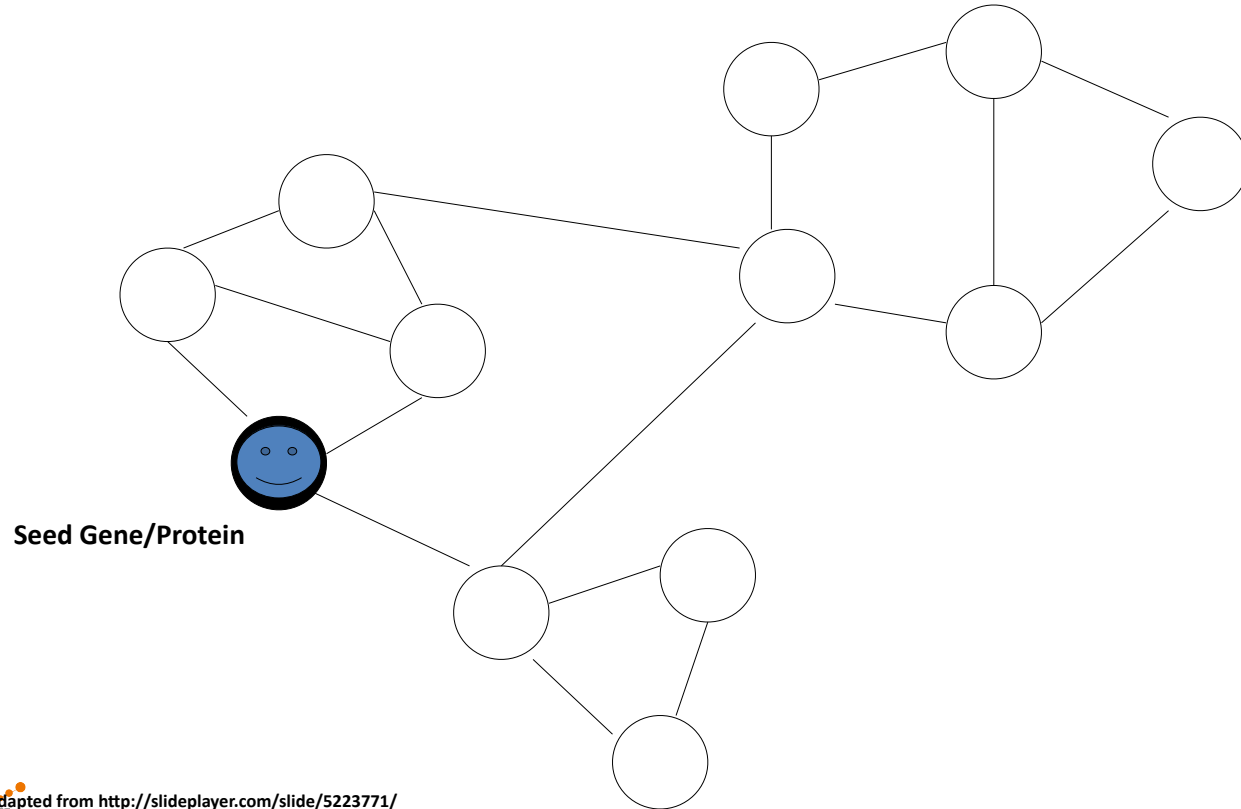
Random Walk with Restart (RWR)

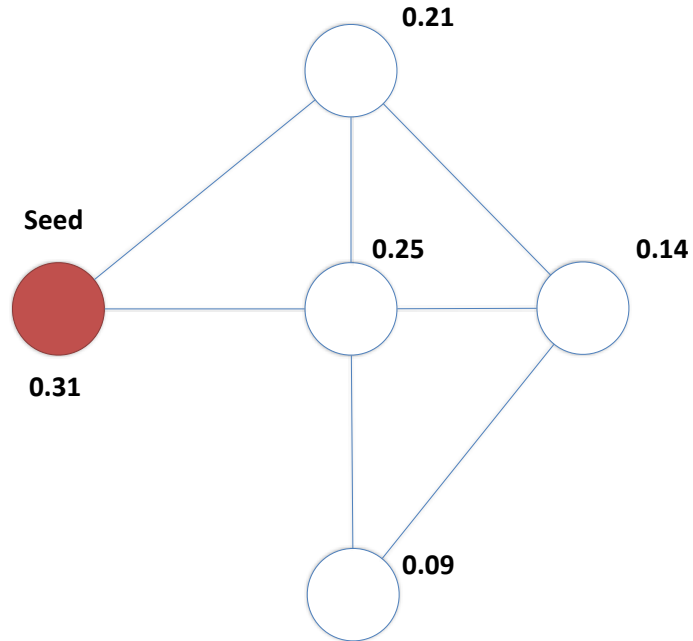




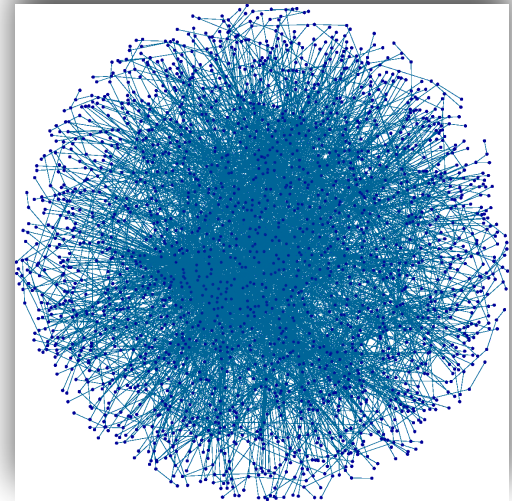


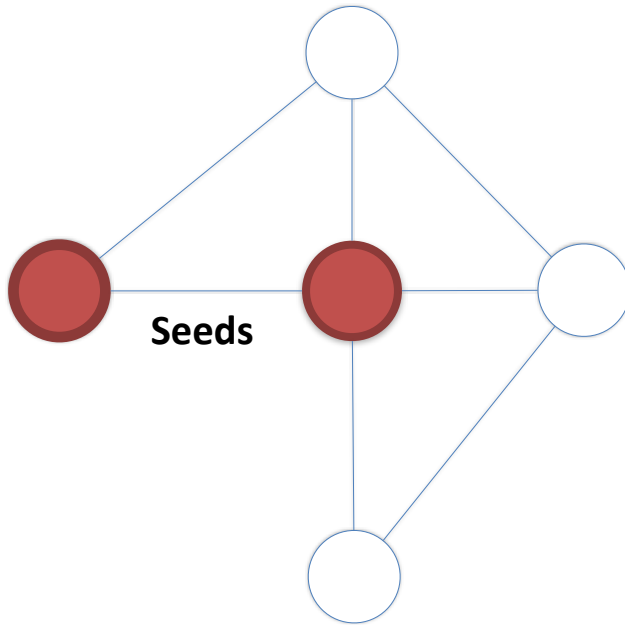
Random Walk with Restart (RWR)



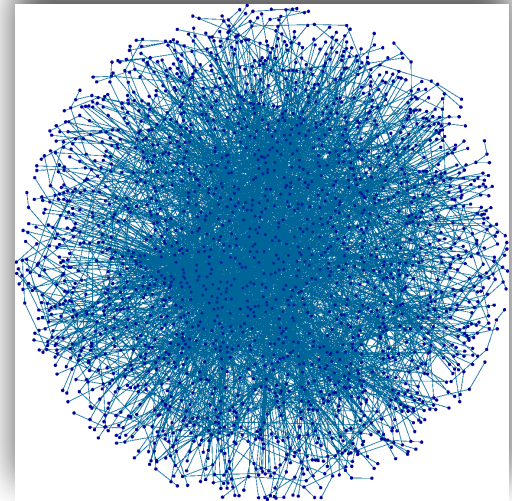


- Local and global exploration
- Proximity/pertinence score *wrt* the seed
 - Rank network nodes
 - Extract subnetworks/modules

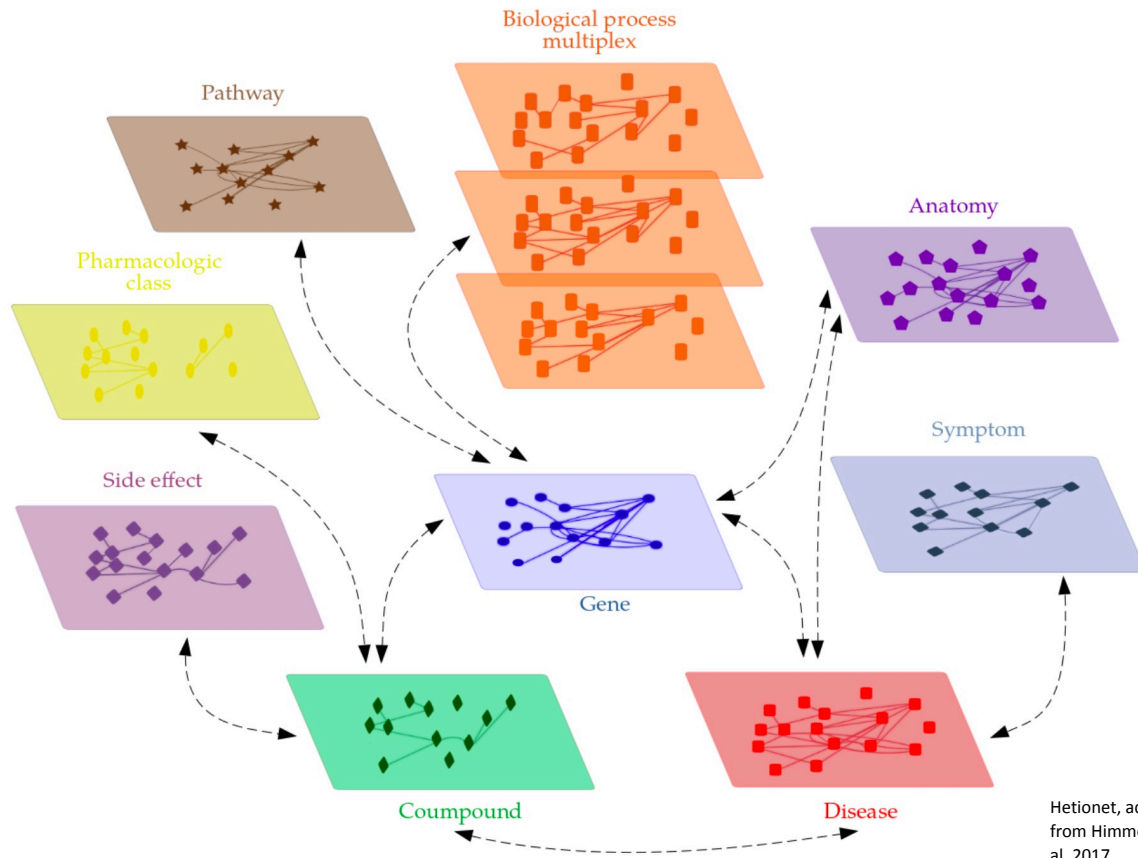




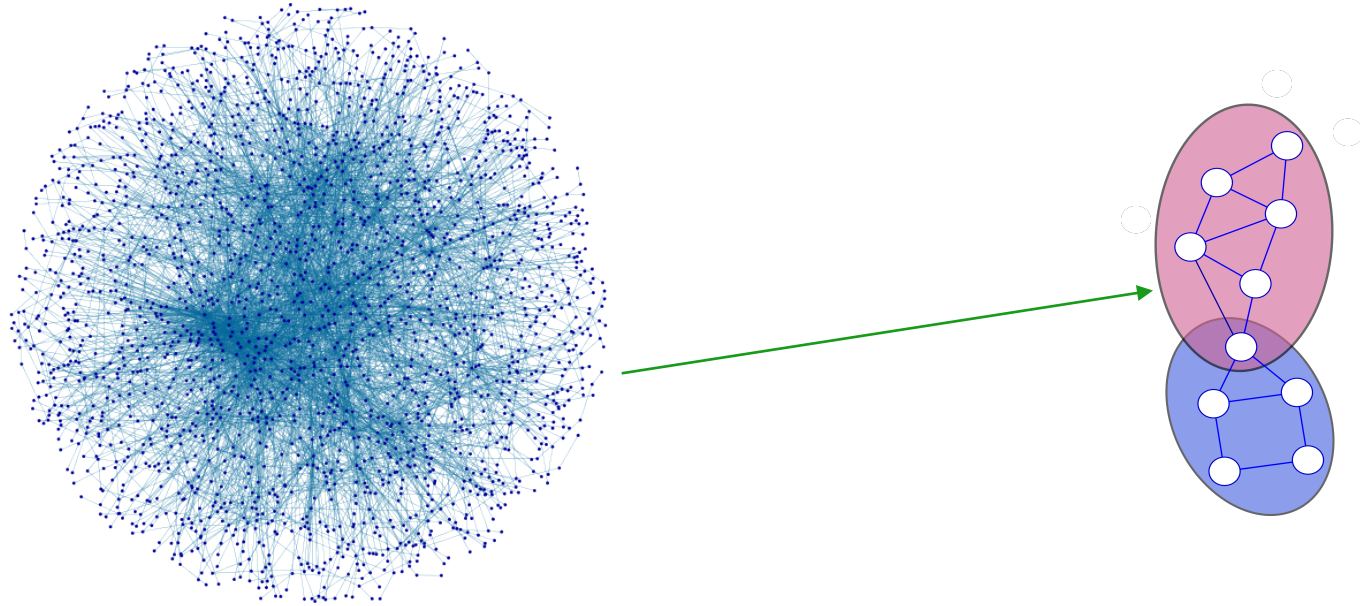
- Local and global exploration
- Proximity/pertinence score *wrt* the seed
 - Rank network nodes
 - Extract subnetworks/modules



Random Walk with Restart on Multilayer networks



Hetionet, adapted from Himmelstein et al. 2017





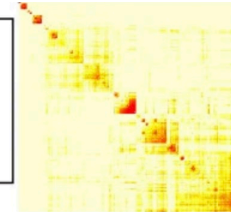
Expression data



Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

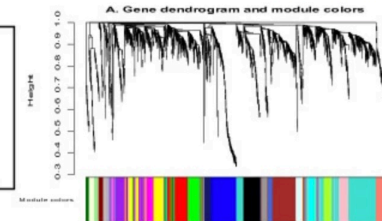
Tools: correlation as a measure of co-expression



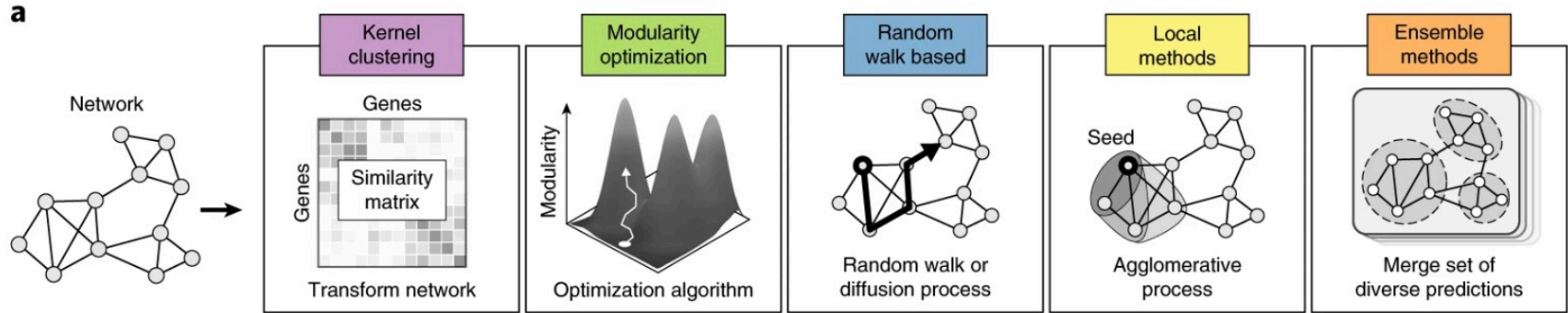
Identify modules

Rationale: module (pathway) based analysis

Tools: hierarchical clustering, Dynamic Tree Cut



A wide number of community identification algorithms

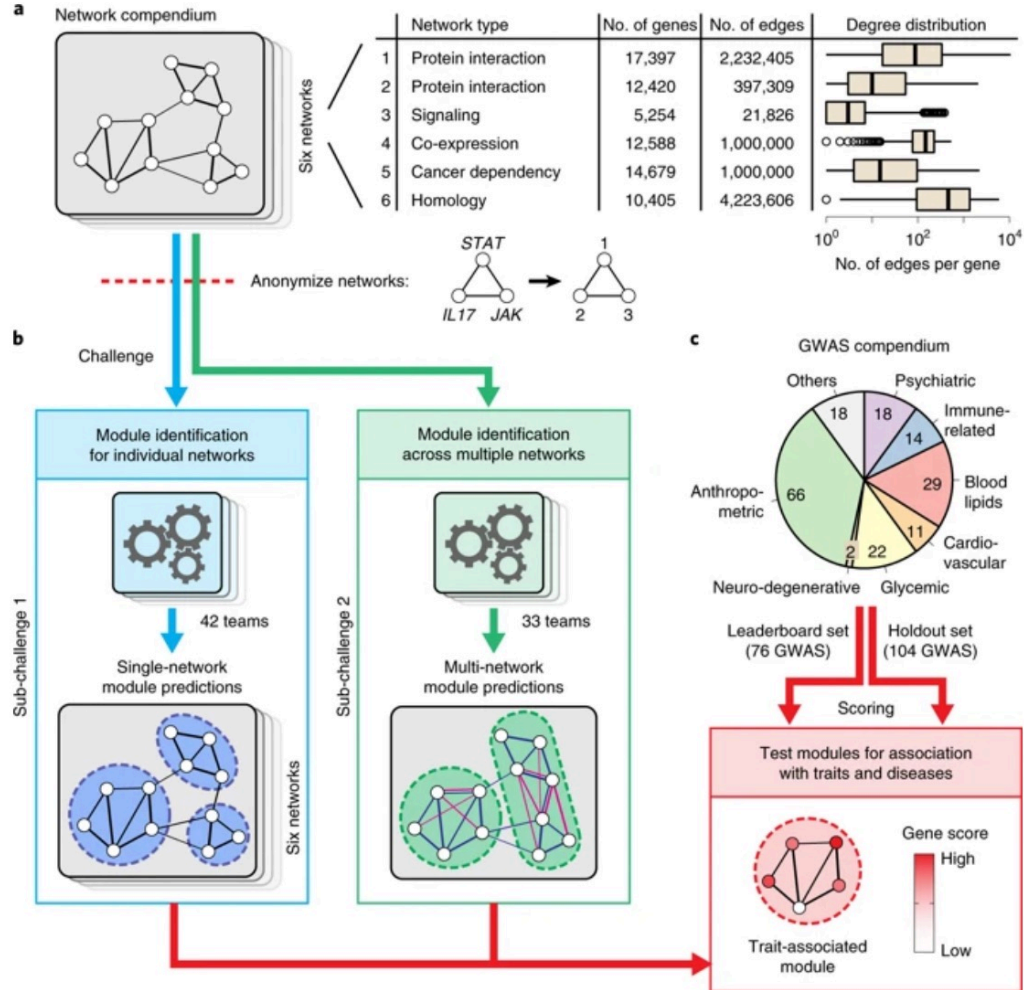


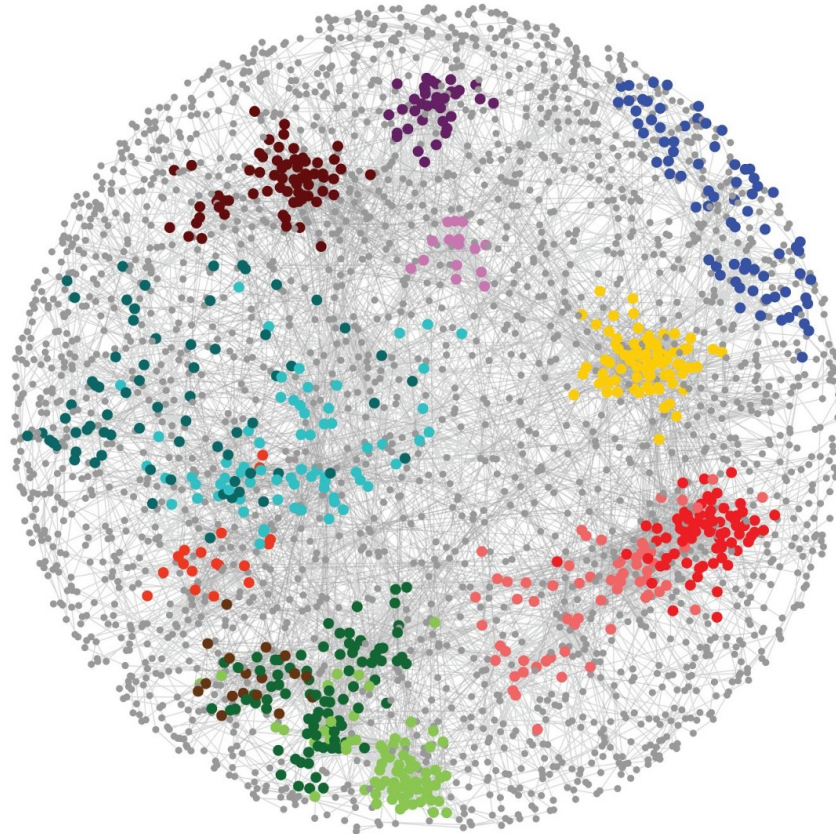
Disease Module Identification
DREAM Challenge
Discover disease pathways in genomic networks

DREAM CHALLENGES
powered by Sage BioNetworks

Unil | Université de Lausanne
SIB | Swiss Institute of Bioinformatics
Joint Research Center for Computational Genomics
RWTH AACHEN UNIVERSITY
IBM
Sage BIOMETRICS
Cell

**We need “benchmarks”
But we don’t have “Gold Standard”**







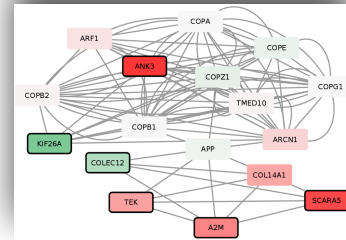
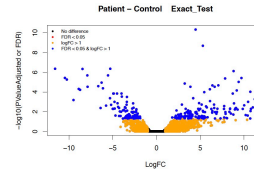
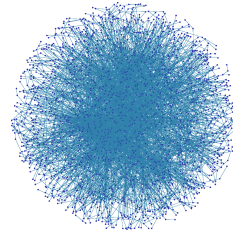
Biological Network

+

RNA-seq transcriptomics data




Find subnetworks of interest
aka “active modules”

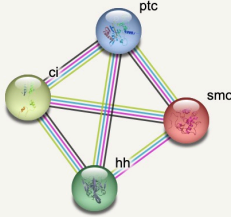


Algorithms: Greedy searches (PinnacleZ), Simulated Annealing (jActiveModules), Genetic Algorithms (COSINE) (Ideker et al. 2002, Chuang et al. 2007, Ma et al. 2011, Ozisik et al. 2017...)





[Search](#) [Download](#) [Help](#) [My Data](#)



Viewers >
Legend >
Settings >
Analysis ▾
Exports >
Clusters >
More +
Less -

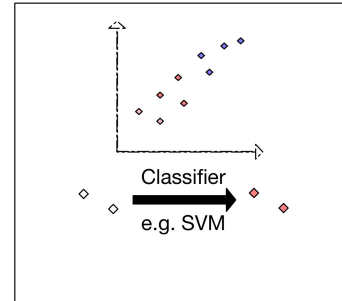
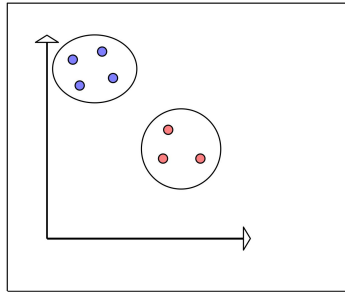
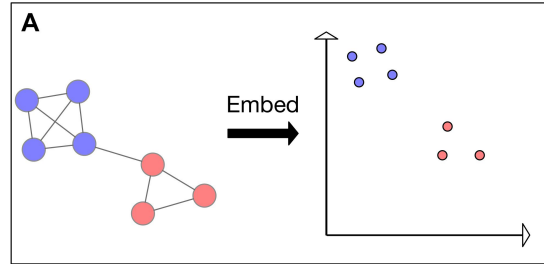
Network Stats

number of nodes: 4	expected number of edges: 0
number of edges: 6	PPI enrichment p-value: 7.22e-07
average node degree: 3	<i>your network has significantly more interactions than expected (what does that mean?)</i>
avg. local clustering coefficient: 1	

Functional enrichments in your network [explain columns](#)

>	Biological Process (Gene Ontology)			
GO-term	description	count in network	strength	false discovery rate
GO:0048099	Anterior/posterior lineage restriction, imaginal disc	3 of 4	3.42	4.81e-07
GO:0035217	Labial disc development	2 of 4	3.24	0.00020
GO:0007224	Smoothened signaling pathway	4 of 12	3.06	7.18e-09
GO:0001746	Bolwigs organ morphogenesis	3 of 9	3.06	1.73e-06
GO:0060914	Heart formation	2 of 6	3.06	0.00032

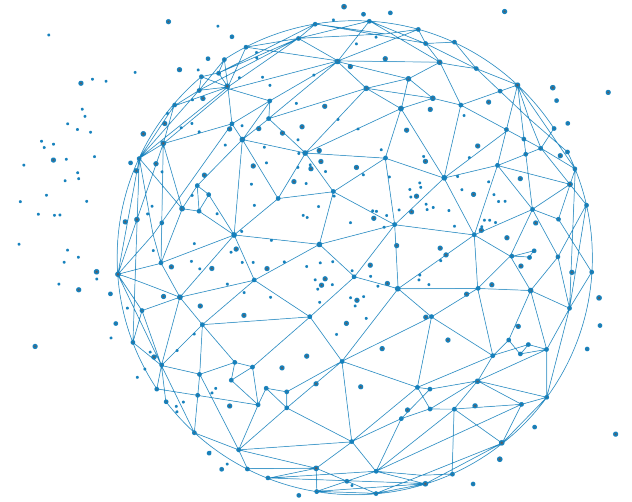
(more ...)



Embedding

Chapter 4

Wrap-up





- Intuitive
- Interpretable
- Easy to connect multi-scale/multimodal information
 - mixed integration
 - ad hoc construction
 - Robust to missing data
 - privacy?
- Scaffold to integrate different data types



Multilayer networks

- Integrate different types of nodes, different types of edges
- From databases, high-throughput screens, literature
- From different omics

Quantitative Omics + Networks

- Interaction enrichment
- Active module identification