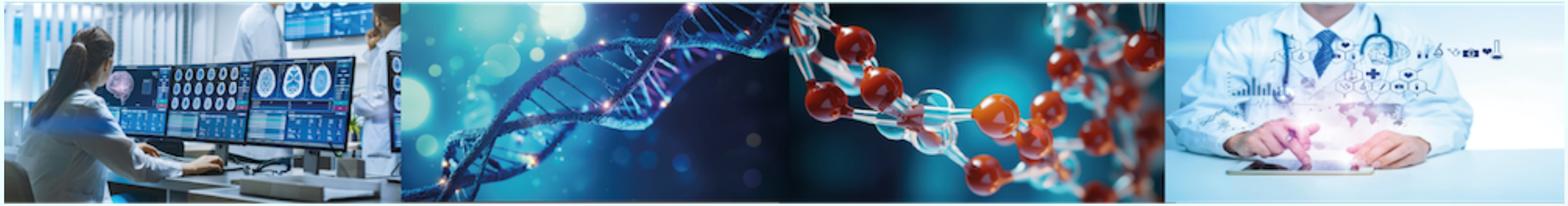


Programming for Biomedical Informatics

Lecture 11 “Biological Networks”

GitHub Website- https://github.com/biomedical-informatics/pbi

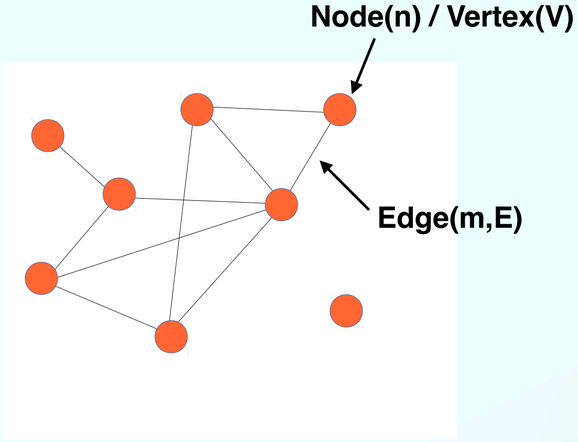
Course Website- https://groups.inf.ed.ac.uk/teaching/pbi/

Ian Simpson
ian.simpson@ed.ac.uk

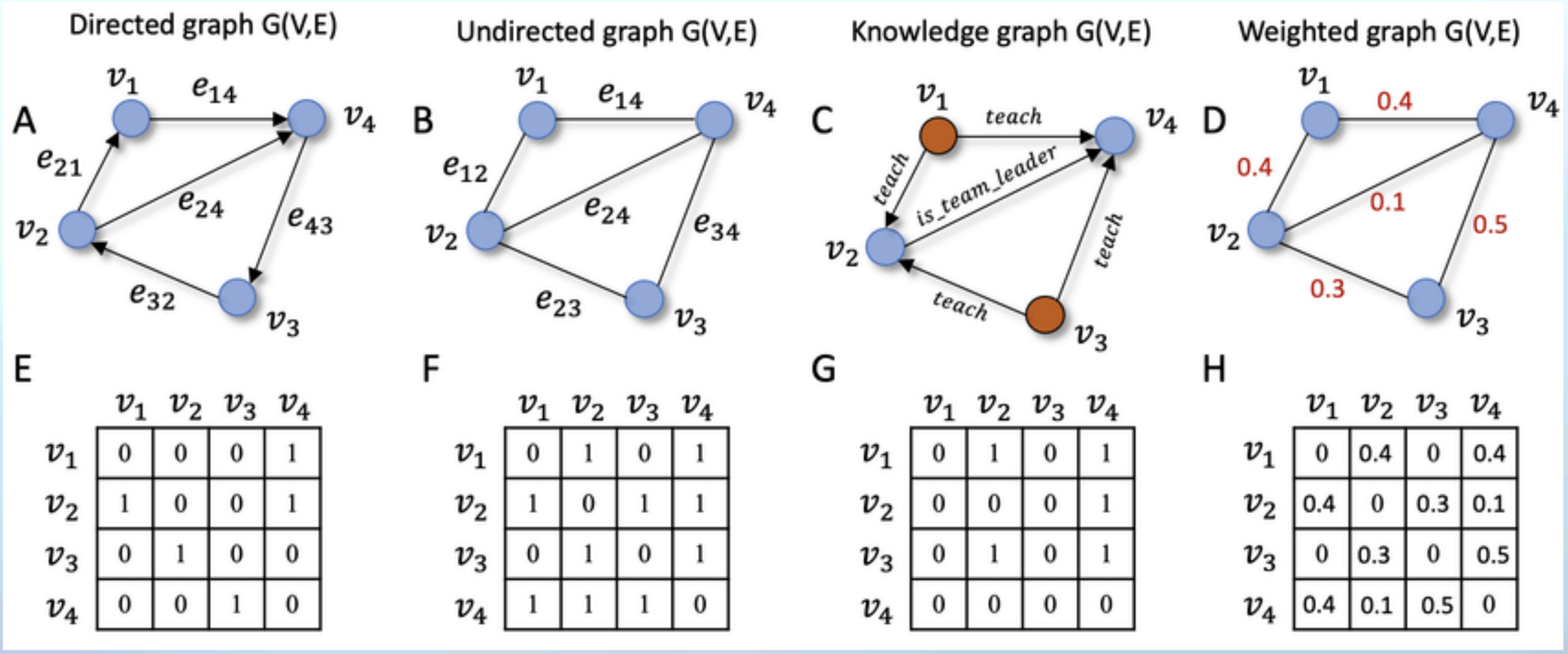
Background

THE UNIVERSITY of EDINBURGH informatics

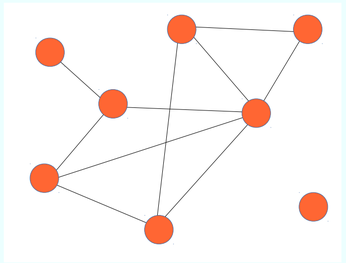
The Anatomy of a Network



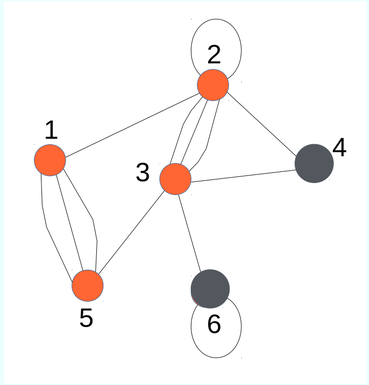
- The study of networks is concerned with understanding and modelling the behaviour of real-world systems.
  - In mathematics called Graph theory.
  - First network Euler's Königsberg bridge problem (1736).



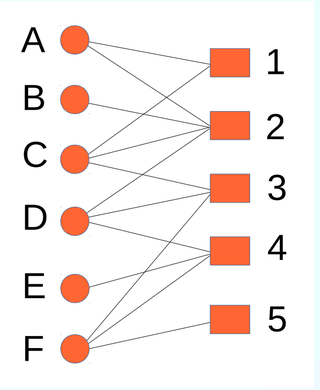
Types of Network



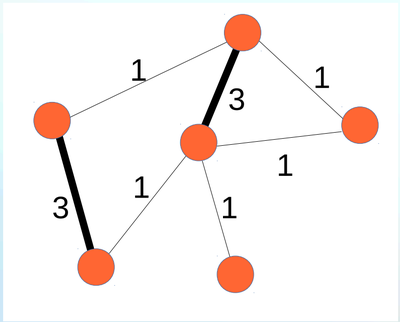
Simple



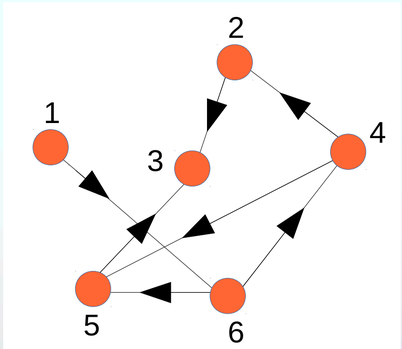
Multigraph



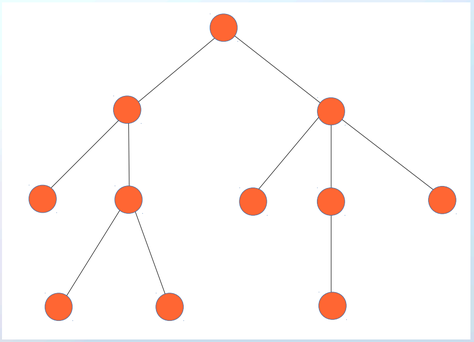
Bipartite



Simple (weighted)



Directed

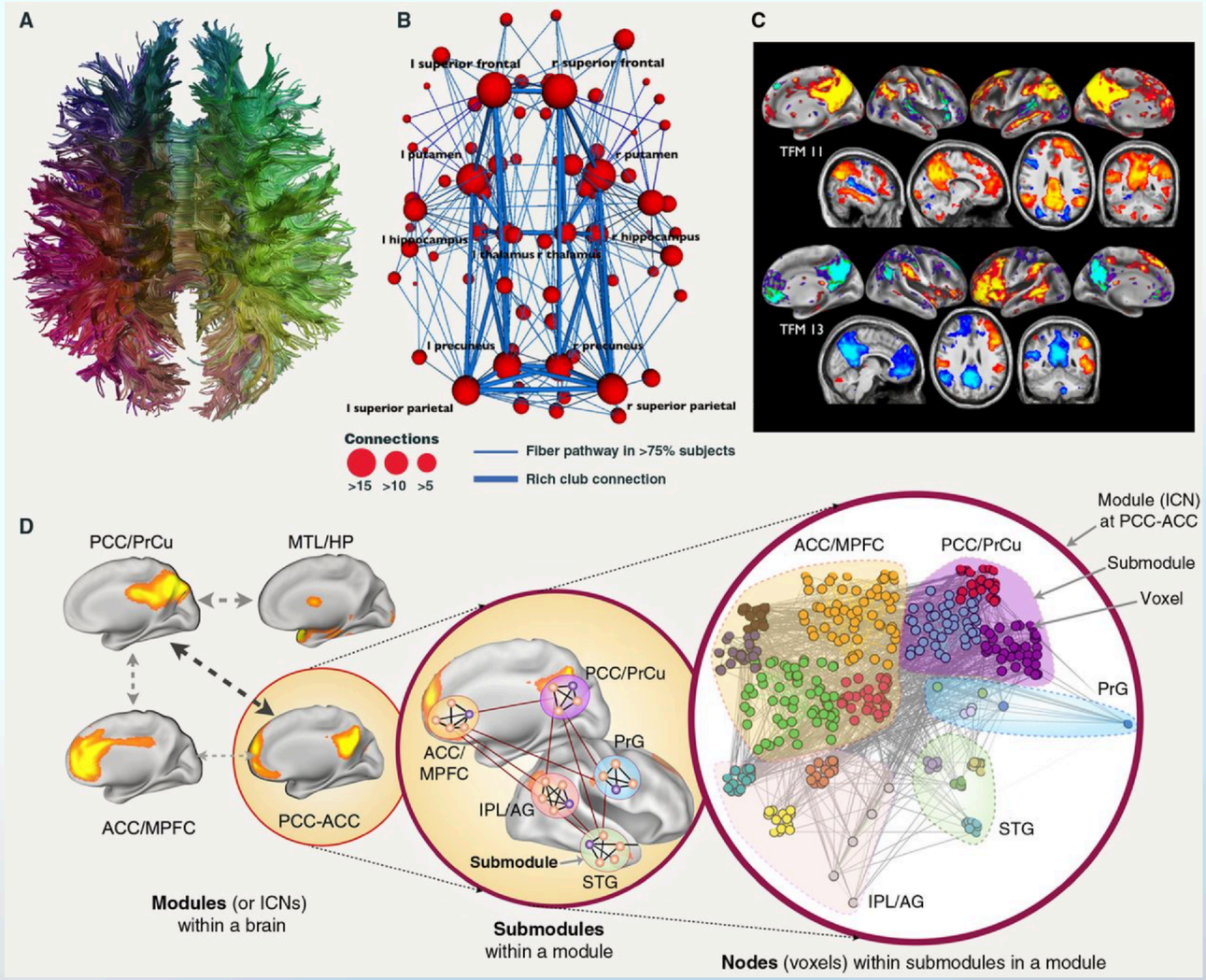


Tree

Simple Network - Connectivity Maps in the Human Brain

These networks have a single type of node, in biomedicine these are very often genes and proteins, but many different modalities can be represented in a uni-partite graph.

In this example a functional network of the brain is modelled based on fMRI data.

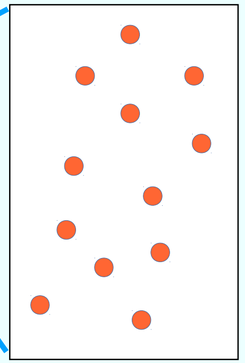


Park HJ, Friston K. Structural and functional brain networks: from connections to cognition. Science. 2013 Nov 1;342(6158):1238411. doi: 10.1126/science.1238411.

Simple Network - Protein Interactions at the Synapse

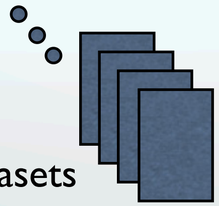
Brain Regions

Forebrain, midbrain, (CA1) hippocampus, cerebellum, (frontal/neo/visual/motor) cortex...

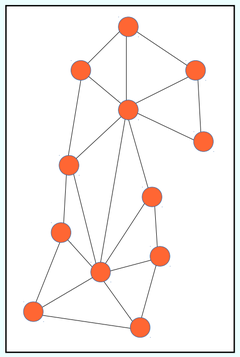


~7000 proteins

35 synaptic proteomic studies (Human, Mouse, Rat) from 2000-2018.

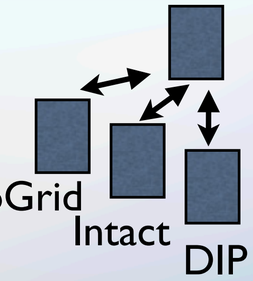


Datasets



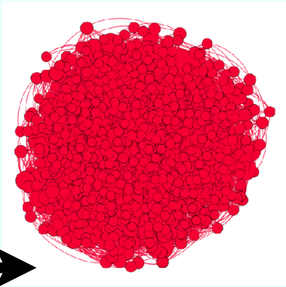
205460 PPIs

PPI databases (BioGrid, Intact, DIP), direct & physical interactions only.



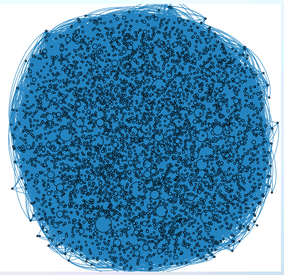
BioGrid Intact DIP

pre-synaptic proteome (pre)



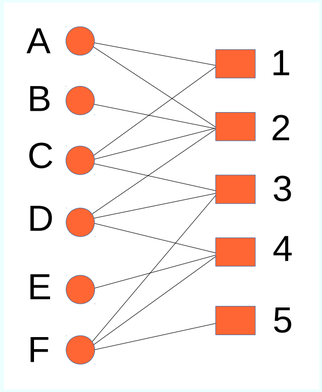
$n = 1589,$
$m = 4321,$
$\rho = 0.003.$

Post-synaptic proteome (PSP)

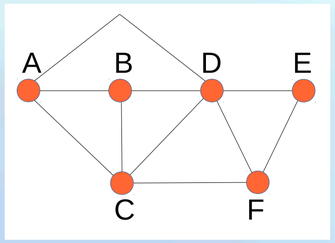


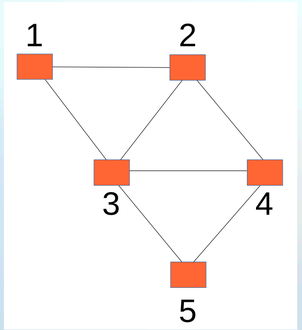
$n = 4752,$
$m = 20444,$
$\rho = 0.002.$

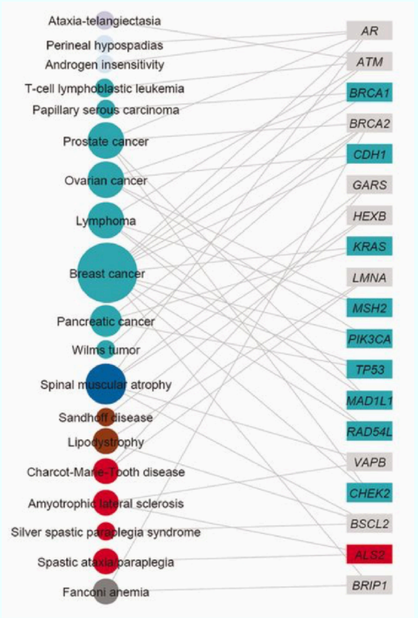
Bipartite Networks



projections

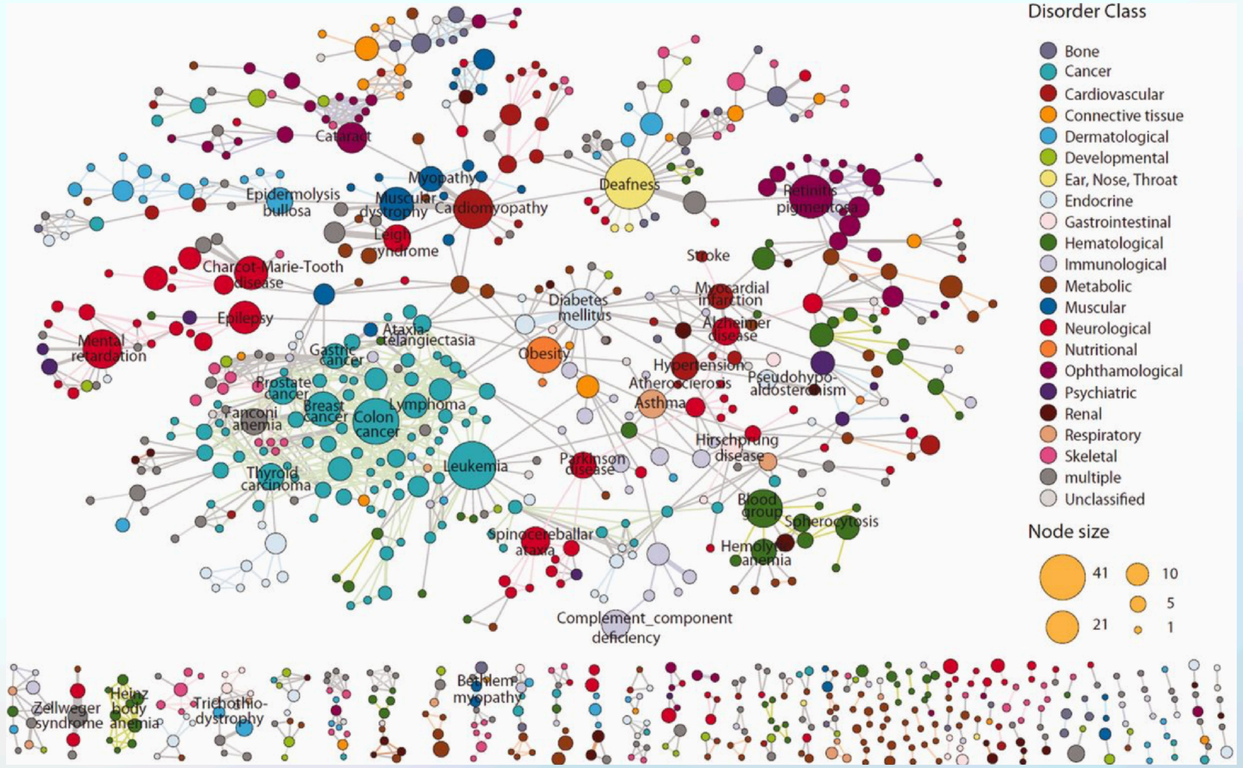






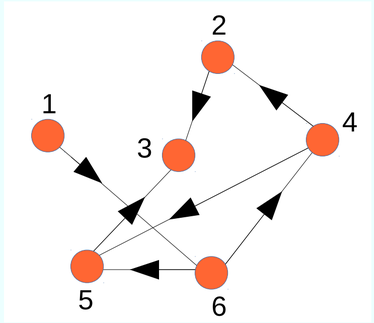
Disease Gene

The Human Diseaseome



Goh KI, Choi IG. Exploring the human diseasome: the human disease network. Brief Funct Genomics. 2012 Nov;11(6):533-42. doi: 10.1093/bfgp/els032.

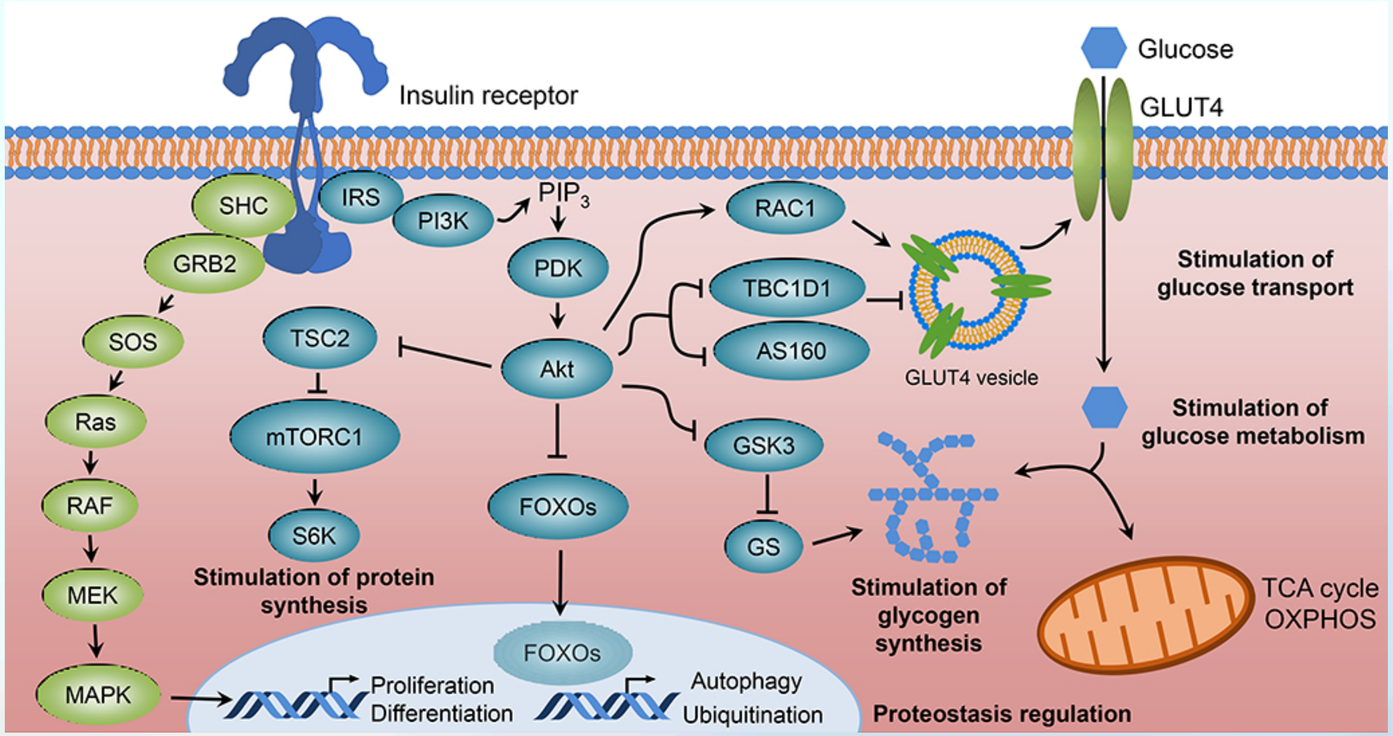
Directed Networks



Many biological processes and systems have inherent directionality

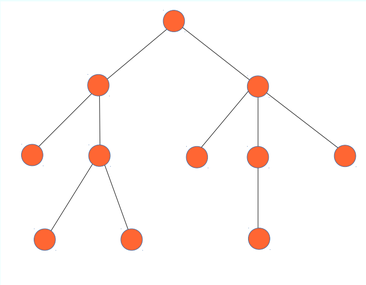
- metabolic pathways composed of reactions in an ordered sequence
- signalling pathways where signal reception occurs and is transduced through a cascade resulting in a cellular response
- gene regulatory networks where master transcriptional regulators control the expression of downstream genes

We can explicitly model directionality (including things like causation) in network models.



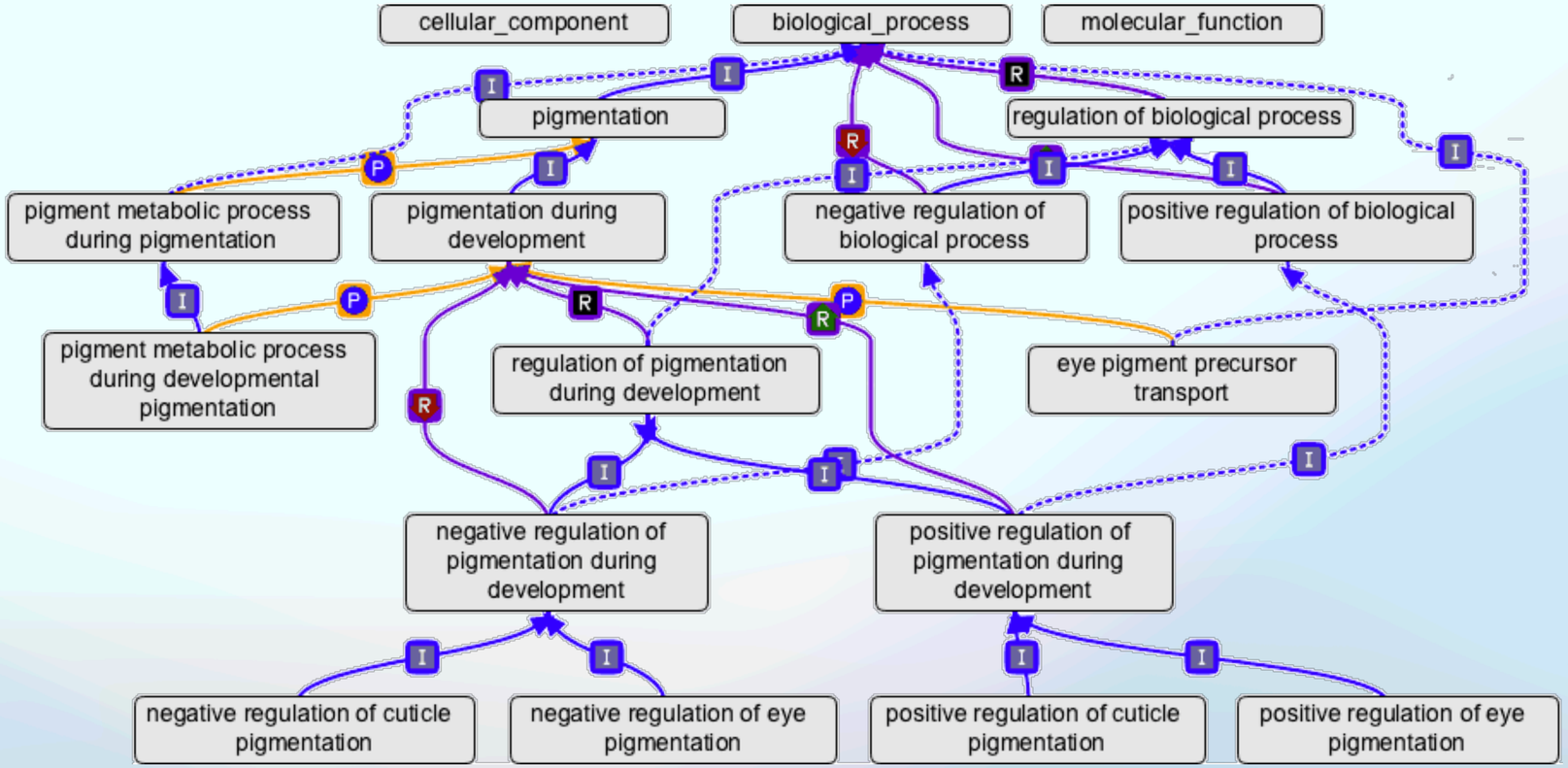
Hierarchical Networks

- nodes are “Terms”
- edges are “Relations”



low specificity roots

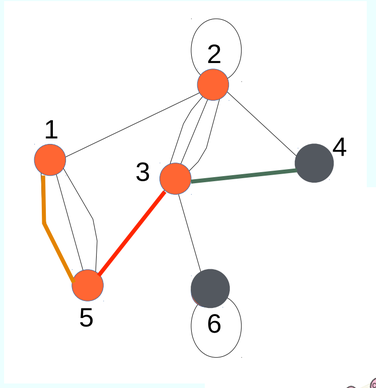
parent

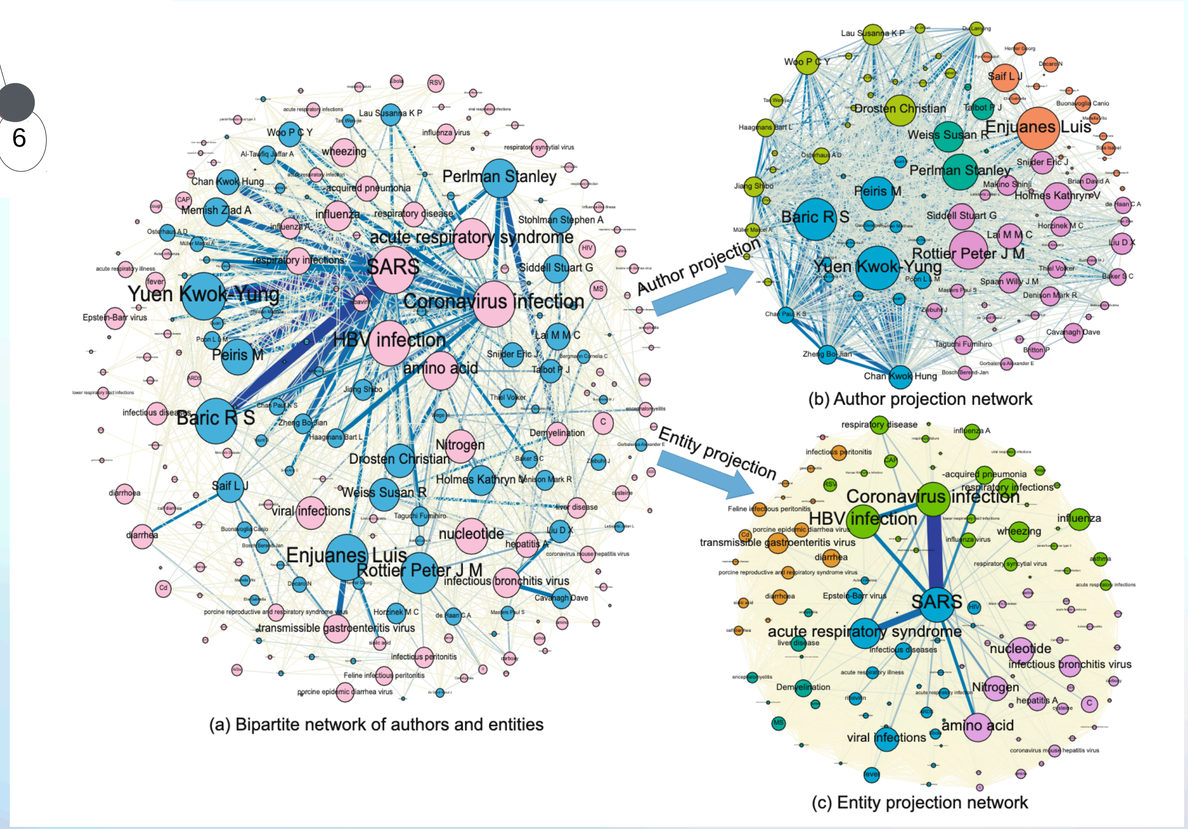


child

high specificity leaves

Multigraph Networks (Knowledge Graphs)





Entities and Relationships
Handle multiple types of nodes and edges allowing for a much richer and flexible knowledge representation

Semantic Context
Commonly have embedded semantic context by using meaningful labels and relationships

Data Integration
As they can handle different node and edge types it is possible to integrate information from different sources

Scalability
By design they can handle vast amounts of interconnected data, are scalable, enabling complex use with real world data

Flexibility and Evolvability
Implicitly modular and can be readily expanded or modified without requiring major structural change

Inference Capabilities
Combination of semantic context and specification allows for inference especially using graph based algorithms and approaches such as GNNs.

Querying and Retrieval
Normally implemented in RDF, OWL or other structured forms that have bespoke software and hardware implementations and languages for complex query and retrieval including semantic query.

Human and Machine Interpretability
These network structures can aid explainability and even support formal reasoning approaches including neuro-symbolic and rules-based logic/ reasoning.

Xu J, Kim S, Song M, Jeong M, Kim D, Kang J, Rousseau JF, Li X, Xu W, Torvik VI, Bu Y, Chen C, Ebeid IA, Li D, Ding Y. Building a PubMed knowledge graph. Sci Data. 2020 Jun 26;7(1):205. doi: 10.1038/s41597-020-0543-2

Multigraph Networks (Knowledge Graphs)

Unified Medical Language System (UMLS) - https://www.nlm.nih.gov/research/umls/index.html
A biomedical knowledge graph that integrates and standardises various medical terminologies, helping link different concepts

Open Biological and Biomedical Ontology (OBO) Foundry - https://obofoundry.org/
A collection of ontologies covering various biomedical domains, such as anatomy, genes, diseases, and phenotypes, designed to support data integration and interoperability in the life sciences

DrugBank -https://go.drugbank.com/
Links drug information, chemical structures, pharmacological data, and interactions with drug targets, enzymes, and other molecules

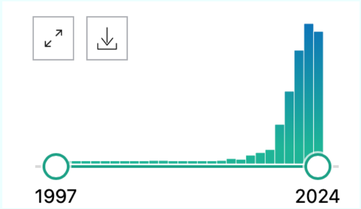
SNOMED CT (Systematised Nomenclature of Medicine Clinical Terms) - https://www.nlm.nih.gov/healthit/snomedct/
Organises healthcare concepts and terms, to aid data integration, electronic health record interoperability, and clinical research

Human Phenotype Ontology (HPO) - https://hpo.jax.org/
A knowledge graph that links human phenotypic abnormalities with associated diseases and genes, used in clinical genetics and precision medicine to facilitate diagnosis

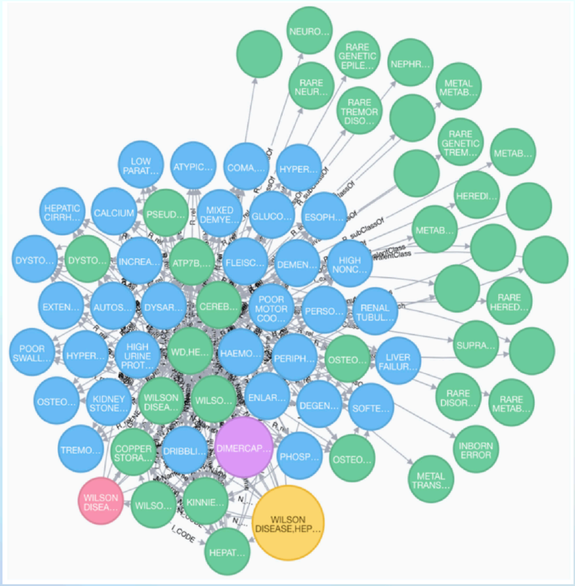
MONDO Disease Ontology - https://mondo.monarchinitiative.org/
An integrative disease ontology consolidating information from multiple sources and species, linking diseases across databases like OMIM, ICD, and Orphanet

PharmGKB (Pharmacogenomics Knowledge Base) - https://www.pharmgkb.org/
Links information on the impact of genetic variation on drug responses, supporting precision medicine through understanding the genetics of drug efficacy and safety.

An Explosion of Biomedical Knowledge Graphs



Over 50% of biomedical papers on KGs in last 2 years



Rare Disease KG

Bioinformatics, 36(2), 2020, 603–610
doi: 10.1093/bioinformatics/btz600
Advance Access Publication Date: 1 August 2019
Original Paper

Systems biology

Discovering protein drug targets using knowledge graph embeddings

Sameh K. Mohamed [1,2,*], Vít Nováček[1,2] and Aayah Nounu[3]

[1]Data Science Institute, College of Engineering and Informatics, [2]Insight Centre for Data Analytics, NUI Galway, Galway, Ireland and [3]MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK

*To whom correspondence should be addressed.
Associate Editor: Lenore Cowen

Received on February 11, 2019; revised on July 20, 2019; editorial decision on July 25, 2019; accepted on July 27, 2019

Zhu *et al. Journal of Biomedical Semantics* (2020) 11:13
https://doi.org/10.1186/s13326-020-00232-y

Journal of Biomedical Semantics

RESEARCH Open Access

An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD)

Qian Zhu[*†], Dac-Trung Nguyen[1†], Ivan Grishagin[1], Noel Southall[1], Eric Sid[2] and Anne Pariser[2]

Briefings in Bioinformatics, 22(3), 2021, 1–14
doi: 10.1093/bib/bbaa110
Method Review

Enriching contextualized language model from knowledge graph for biomedical information extraction

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji and Xiaohui Liang

Corresponding author: Yafeng Ren, Laboratory of Language and Artificial Intelligence, Guangdong University of Foreign Studies, Guangzhou, China. Tel.: +86-13397124322; Fax: +86-020-86318925; Email: renyafeng@whu.edu.cn

ARTICLES
https://doi.org/10.1038/s41587-021-01145-6
nature biotechnology

OPEN

A knowledge graph to interpret clinical proteomics data

Alberto Santos[1,2,3✉], Ana R. Colaço[1], Annelaura B. Nielsen[1], Lili Niu[1], Maximilian Strauss[1,4], Philipp E. Geyer[1,4,5], Fabian Coscia[1,5], Nicolai J. Wewer Albrechtsen[1,6,7], Filip Mundt[1], Lars Juhl Jensen[1] and Matthias Mann[1,5✉]

Yang *et al. BMC Bioinformatics* 2021, **22**(10):387
https://doi.org/10.1186/s12859-021-04292-4

BMC Bioinformatics

RESEARCH Open Access

Mining a stroke knowledge graph from literature

Xi Yang[1,2,3], Chengkun Wu[2], Goran Nenadic[3*], Wei Wang[1] and Kai Lu[1]

From The 19th Asia Pacific Bioinformatics Conference (APBC 2021) Tainan, Taiwan. 3-5 February 2021

Bioinformatics, 37(23), 2021, 4597–4598
doi: 10.1093/bioinformatics/btab694
Advance Access Publication Date: 6 October 2021
Applications Note

Databases and ontologies

COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases

Chuming Chen [1,*], Karen E. Ross[2], Sachin Gavali [1], Julie E. Cowart[1] and Cathy H. Wu[1,2]

[1]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA and [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC 20007, USA

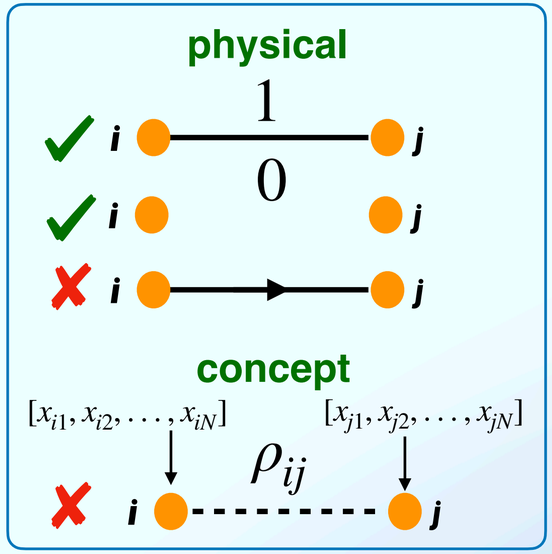
Network - Constraints

physical:
- useful thinking about our networks as being either physical or conceptual

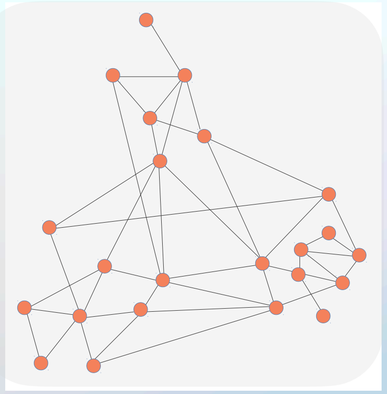
unweighted:
- every edge in network assigned either 1 “measured” or 0 “not measured”

undirected:
- every measured edge in network can go from $i \rightarrow j$ or $j \rightarrow i$ with equal magnitude.

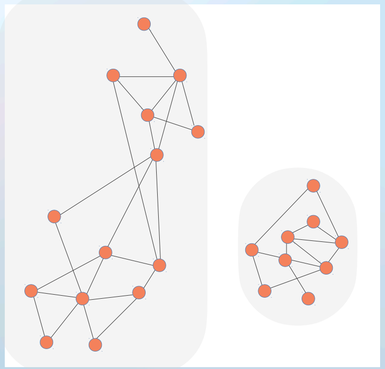
single component:
- fraction of nodes (S) in the largest component is always 1.



$S = 1$



$S = 0.6$

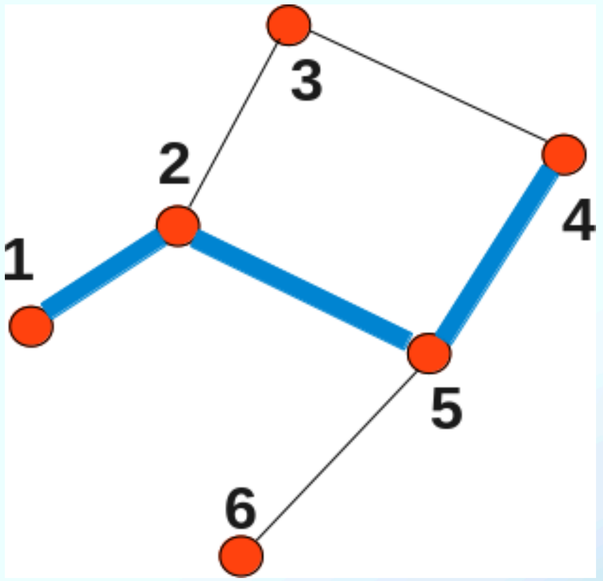


Network - Representations

$A_{ij} \in \mathbb{R}$ (unweighted = 0,1)

$A_{ij} = A_{ji}$ (undirected)

$A_{ii} = 2$ (if self-edge)



Adjacency Matrix

$$\mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$k_i = \sum_{j=1}^{n} A_{ij}$$ (degree of node i)

(edges in network)

$$\frac{1}{2}\sum_{i=1}^{n} k_i = 6$$

Degree Matrix

$$\mathbf{D} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 2 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 2 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 3 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Laplacian Matrix

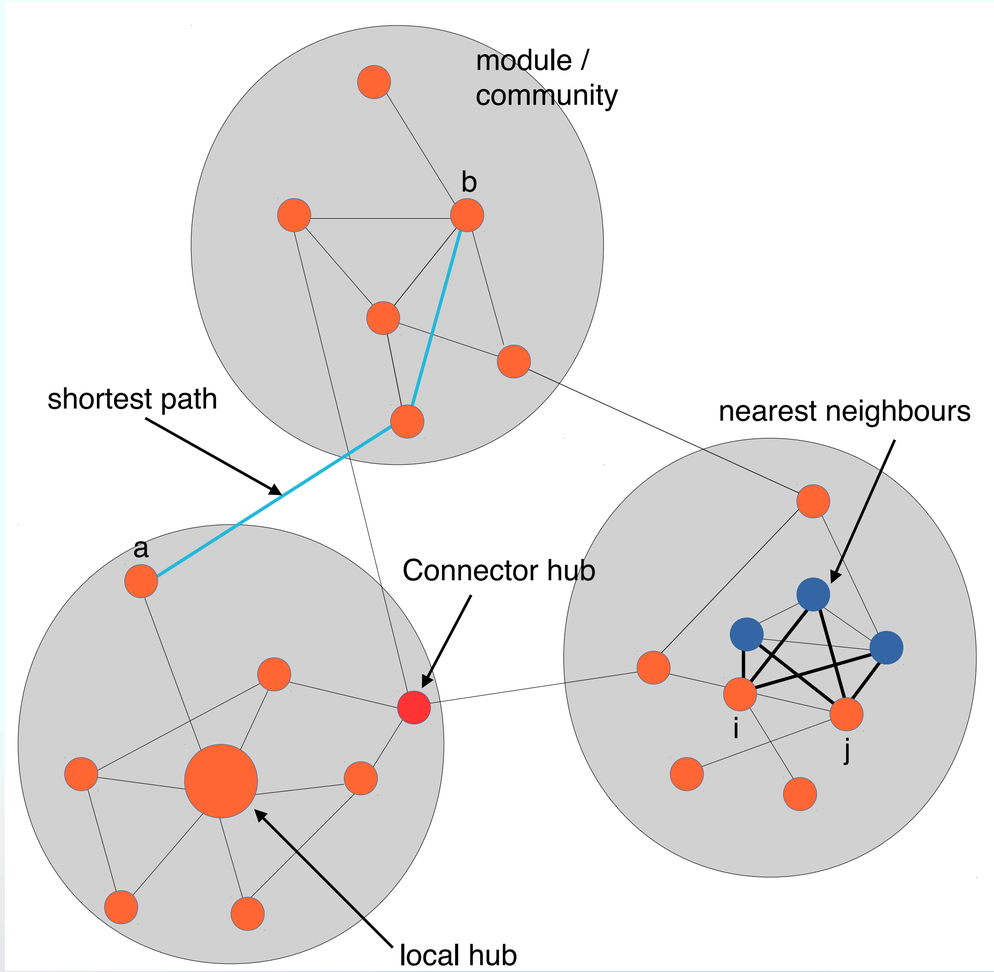
$$\mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 2 & -1 & 3 & -1 & 0 & -1 & 0 \\ 3 & 0 & -1 & 2 & -1 & 0 & 0 \\ 4 & 0 & 0 & -1 & 2 & -1 & 0 \\ 5 & 0 & -1 & 0 & -1 & 3 & -1 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Network - Measures

Measures:
- From our mathematical representation, we can define many concepts which reveal a networks underlying character.

Centrality:
- “What are the most important or central nodes (i.e. influencer, disease hub) in our network.”
- Degree, Betweenness Google’s PageRank, Semi-local, Closeness...

Similarity:
- “How similar two nodes are to each other.”
- how many shared neighbours.
- Pearson correlation, Modularity.

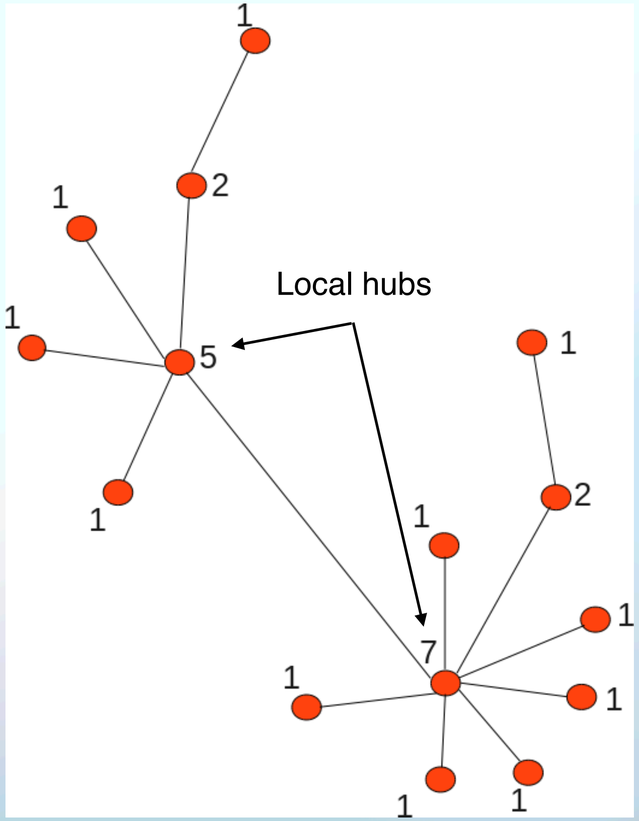


Network - Degree

One of the most useful concepts, encoding a network's local structure:
- can be used as a centrality score
- to generate random networks
- (a test) degree distribution of many real-world networks follows a power-law: $\propto k^{-\alpha}$

Node Degree Centrality:
- Number of neighbours, denoted by k

- local hubs $\Longrightarrow$ { highest degree nodes; disease related proteins



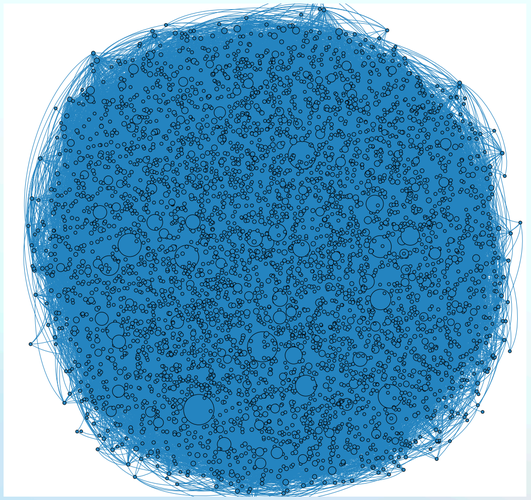
5 highest degree nodes in PPI PSP

Gene	Degree	Diseases
EWSR1	232	MND
GRB2	205	SCH
SPRK2	183	SCH
LMNA	180	AD
EGFR	174	AD,SCH, HD, BD

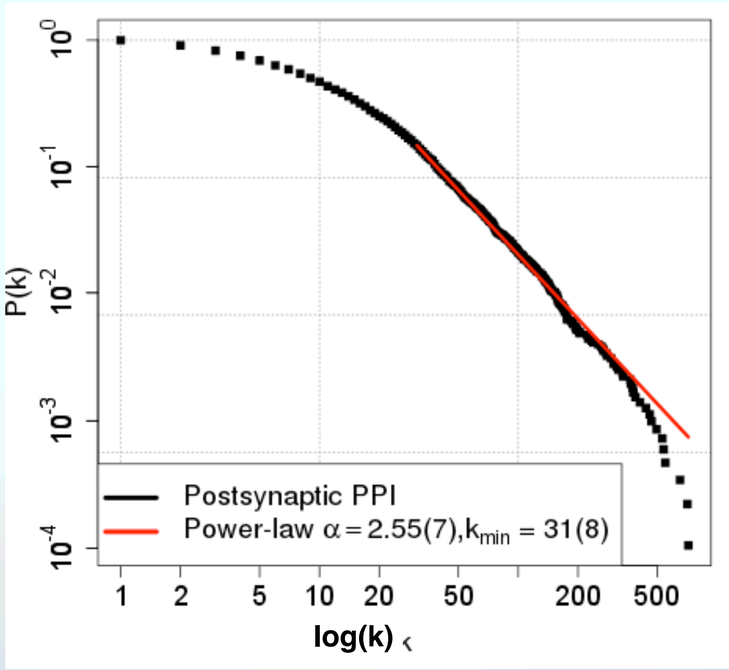
Network - Degree

One of the most useful concepts, encoding a network's local structure:
- can be use as a centrality score.
- generate random networks...
- (a test) degree distribution of many real-world networks follows a power-law: $\propto k^{-\alpha}$

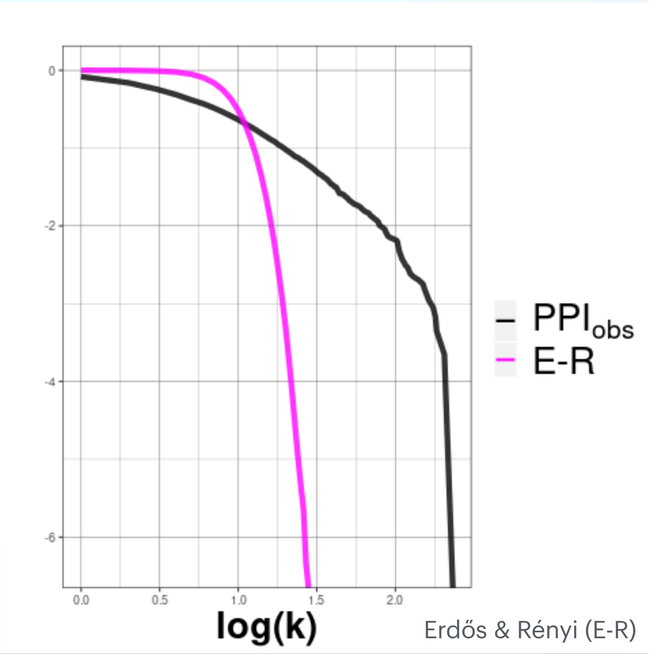
Network



Degree Distribution



Random Distribution

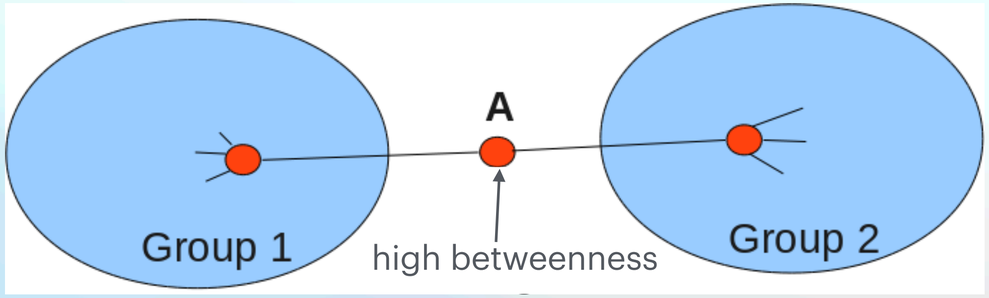


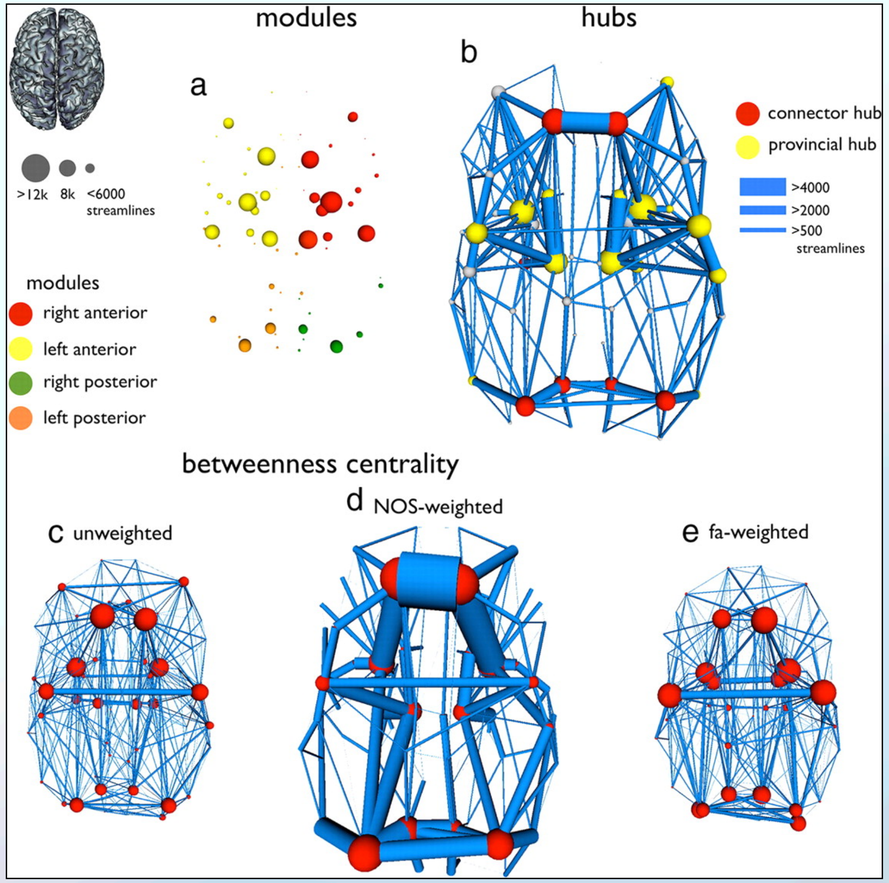
Erdős & Rényi (E-R)

Network - Betweenness

Betweenness:
- Fraction of shortest paths passing through a node.
  - Nodes can have low degree, be a long distance on average from others, and still have high betweenness. For instance ‘A‘ below.

-connector hubs $\Longrightarrow$ { high betwneenness nodes; controlling information flow }





Network - Similarity

common neighbours (n_{ij}):

- nearest neighbours of nodes i and j is:

$$n_{ij} = \sum_k A_{ik}A_{kj} = \vec{A_{i.}} \cdot \vec{A_{j.}} = 3$$

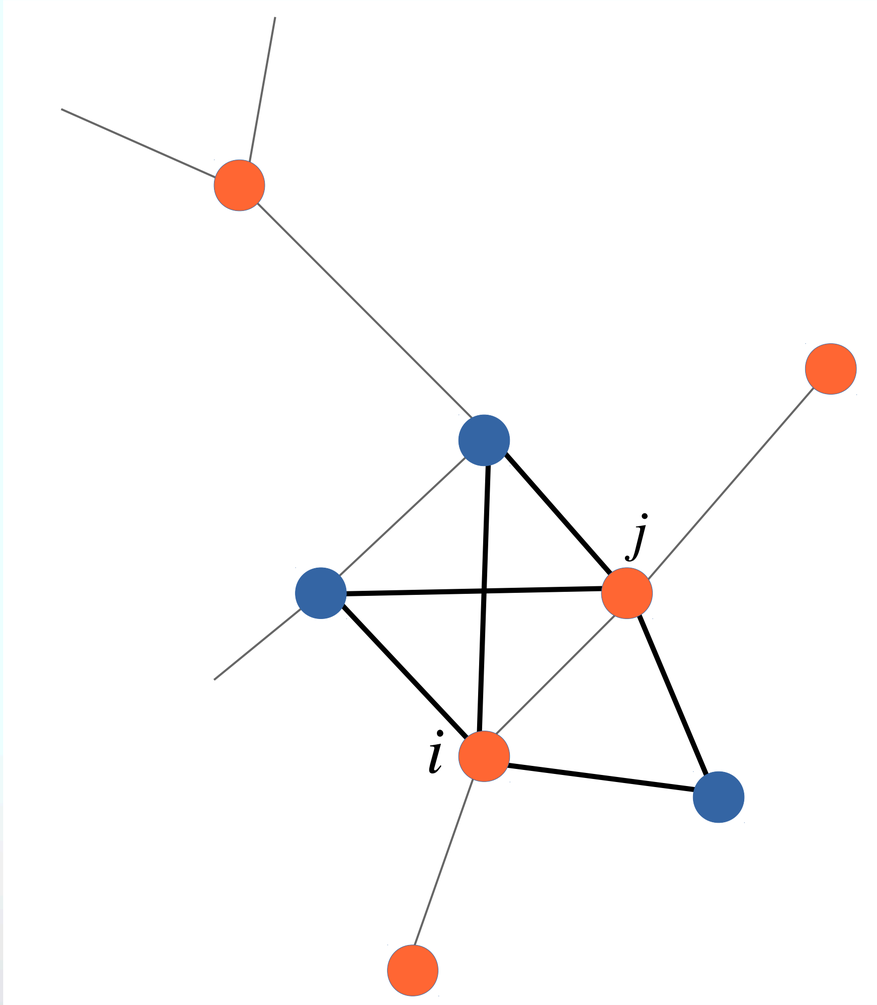
each row(column) of $\mathbf{A}$ is a vector

cosine:

- $\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i}\sqrt{k_j}} = \frac{3}{\sqrt{5}\sqrt{5}} = 0.6$

Pearson correlation:

- $r_{ij} = \frac{\sum_k (A_{ik} - \langle A_i \rangle)(A_{jk} - \langle A_j \rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i \rangle)^2}\sqrt{\sum_k (A_{jk} - \langle A_j \rangle)^2}}$



Network - Paths

Geodesic distance (d):

- Shortest path (excluding loops) between two nodes (d_{ik}) through network.

Random walks:

-Transition matrix $\mathbf{P}$, where probability going from i to j is: $P_{ij} = \frac{1}{k_i}$

- the steady state, i.e. after making many random walks:

$$\mathbf{P}\vec{\pi} = \vec{\pi}$$

Initial prob. dist. of nodes in network

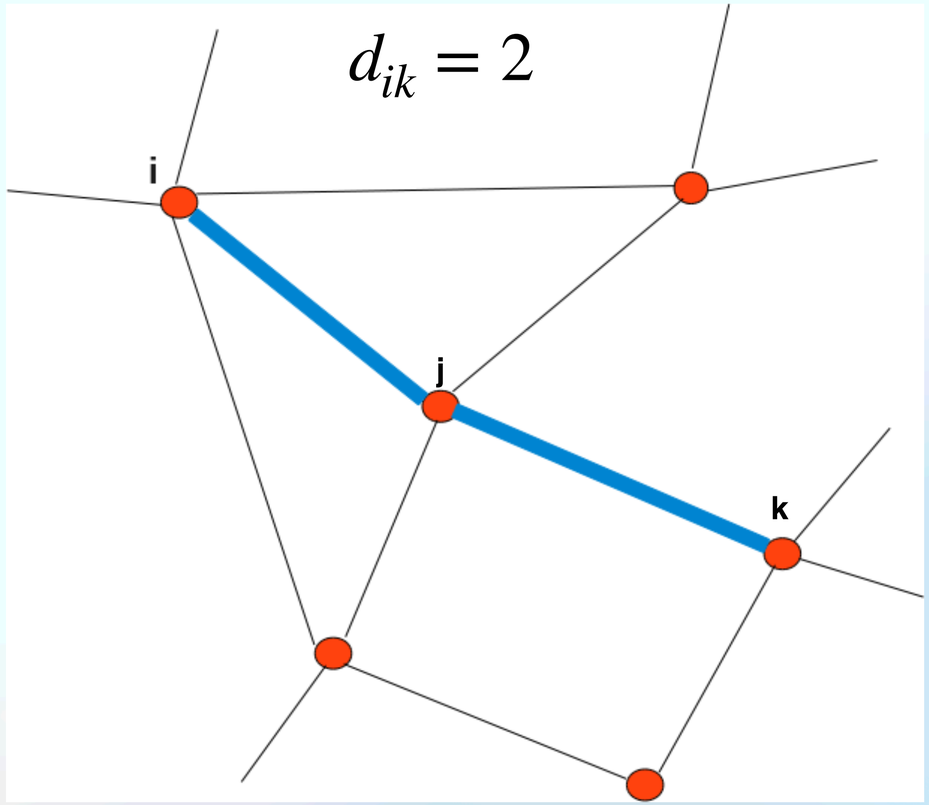
$$\mathbf{L}\left[\mathbf{D}^{-1}\vec{\pi}\right] = 0$$

$$\mathbf{D}^{-1}\vec{\pi} = a\vec{1}$$

$$\therefore \pi_i = \frac{k_i}{2m}$$

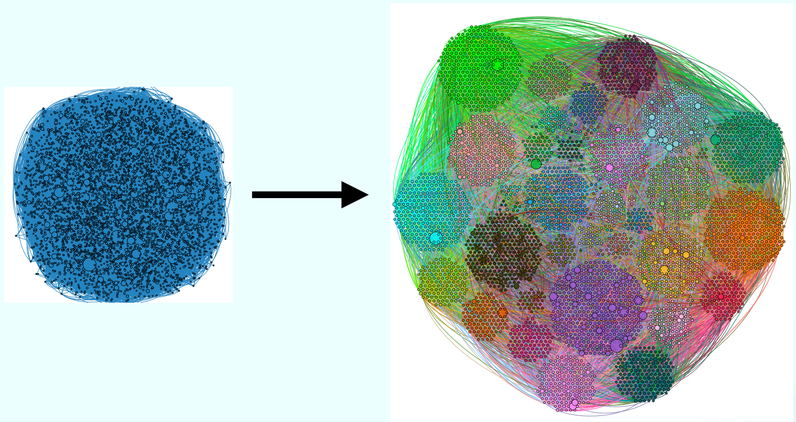
- probability $p(i \to j)$ of walk along edge from i to j at any time of random walk is:

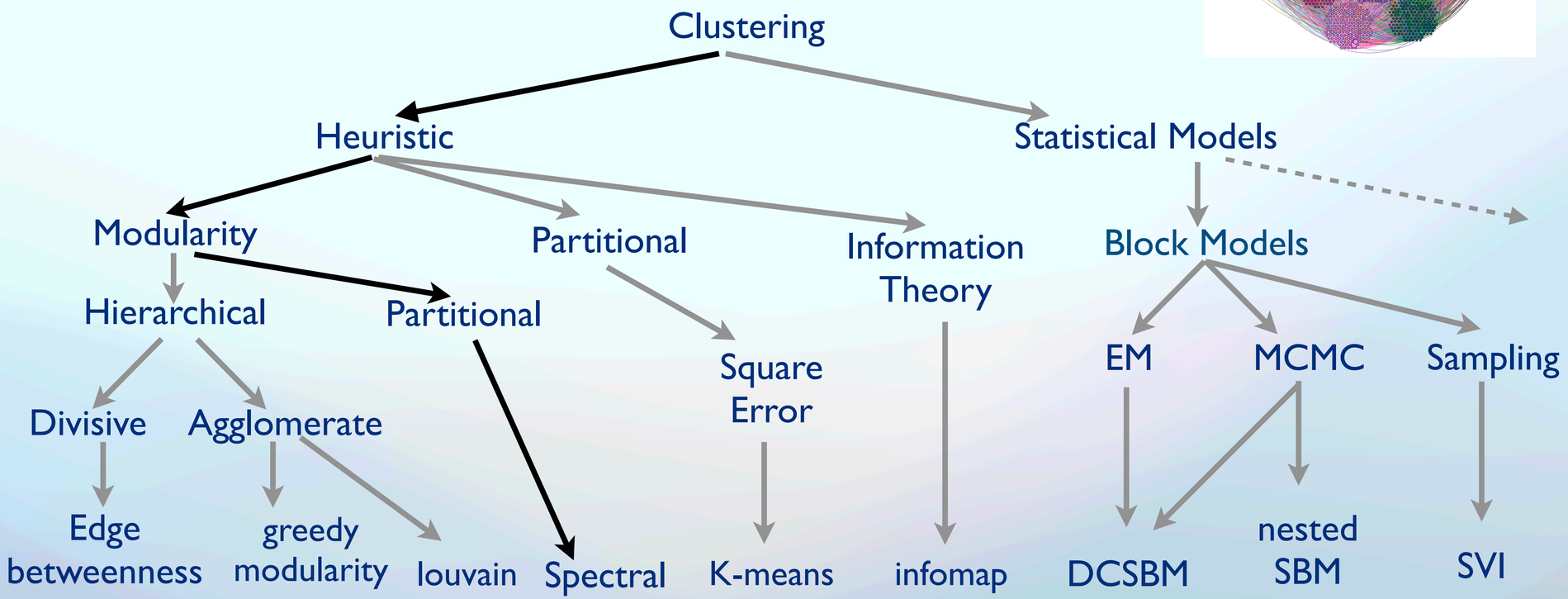
$$p(i \to j) = \frac{1}{k_i}\frac{k_i}{2m}$$



Network - Clustering

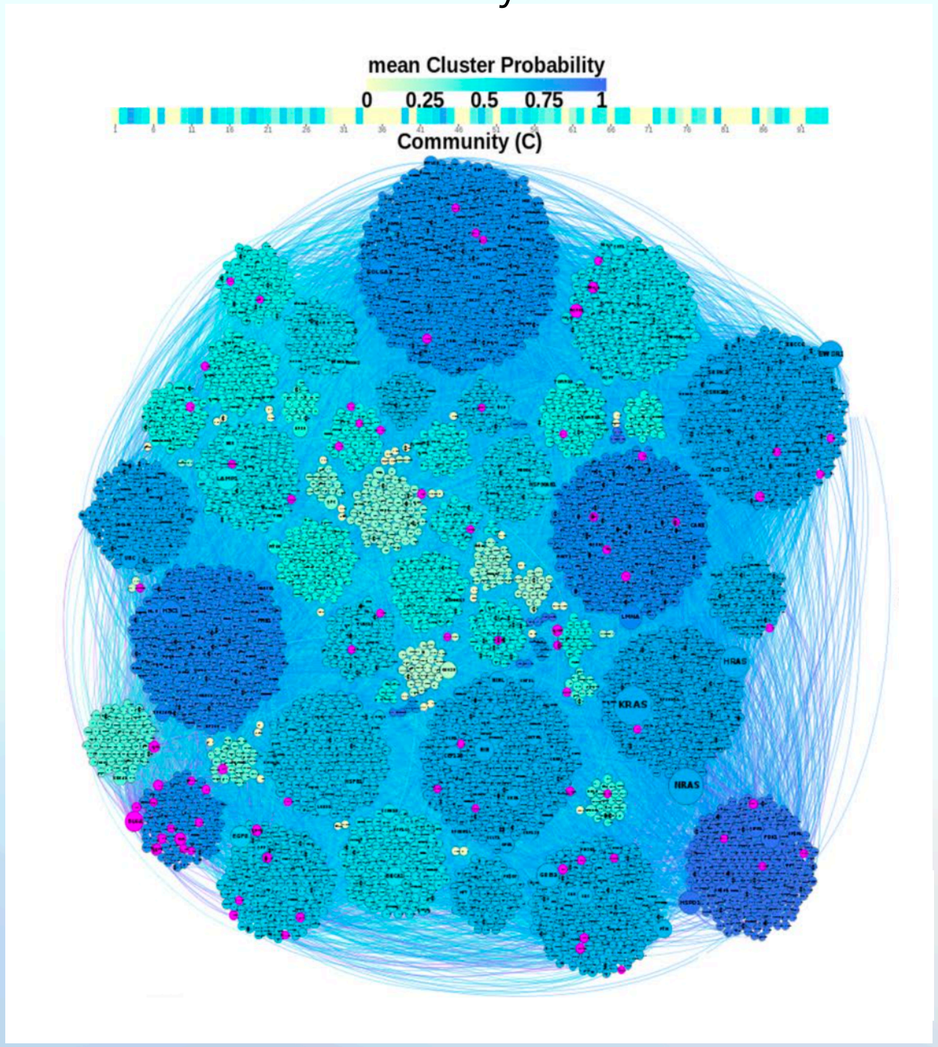
What is a Community?
- locally dense regions of nodes in our network that are similar
- modules, complexes, groups, blocks.



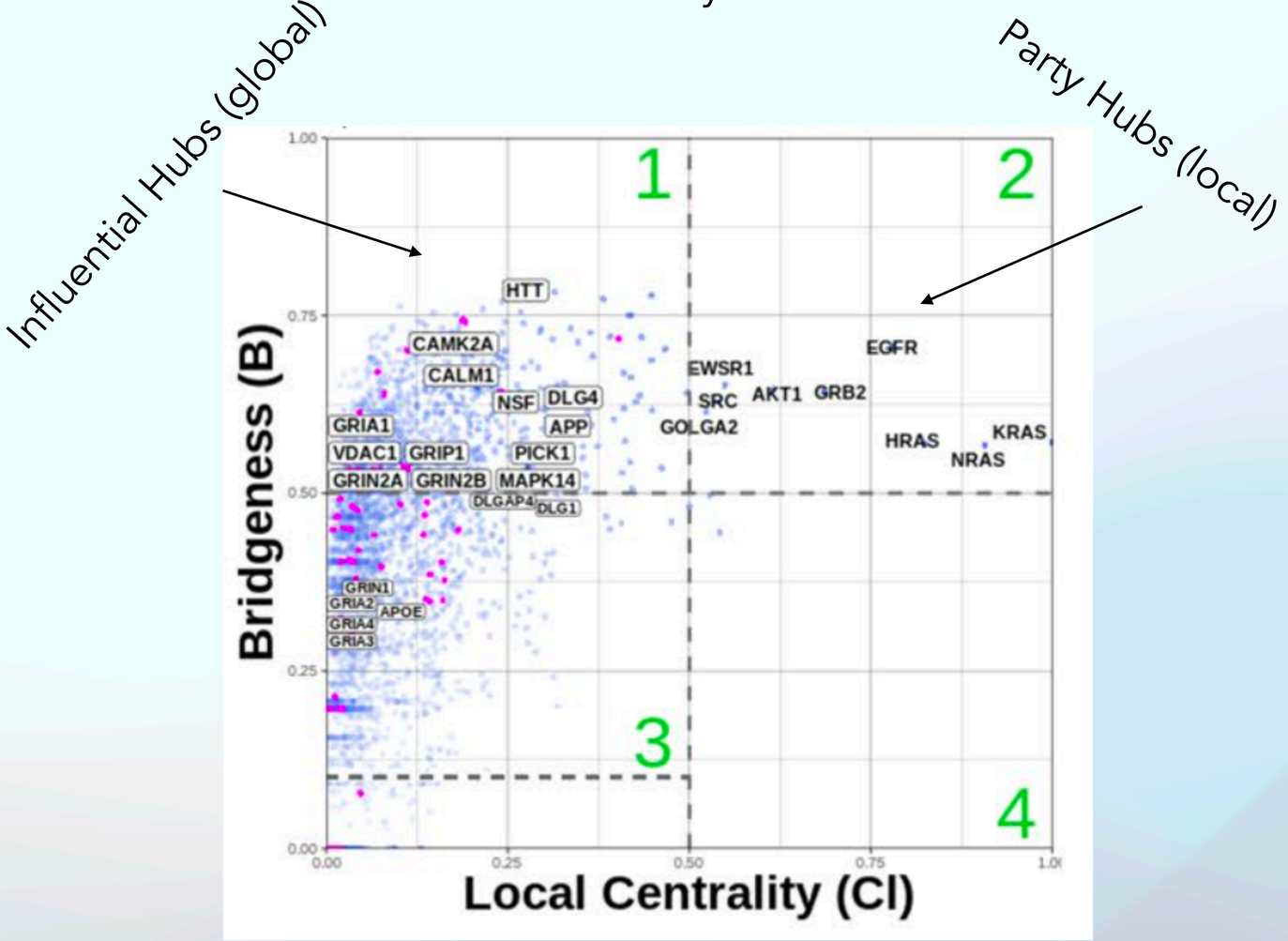


Community Detection in the Synaptic Interactome

Community Structure

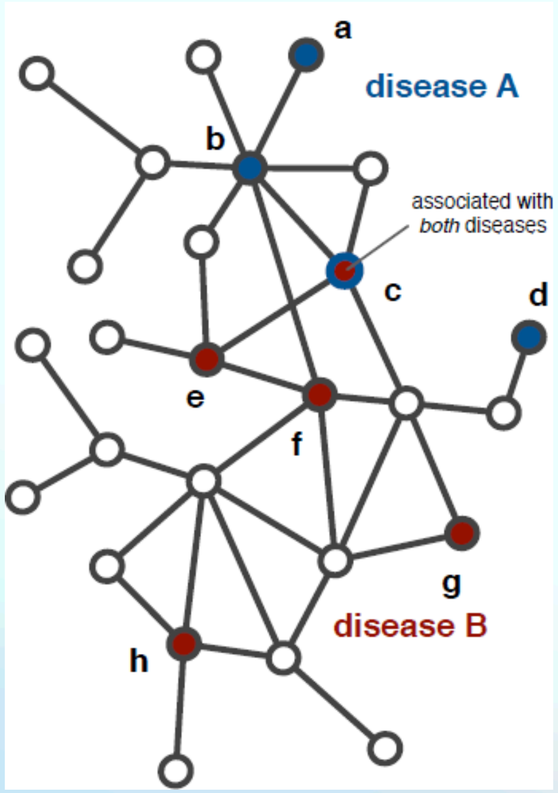


Community Structure



Sorokina O, Mclean C, Croning MDR, Heil KF, Wysocka E, He X, Sterratt D, Grant SGN, Simpson TI, Armstrong JD. A unified resource and configurable model of the synapse proteome and its role in disease. Sci Rep. 2021 May 11;11(1):9967. doi: 10.1038/s41598-021-88945-7. Erratum in: Sci Rep. 2021 Aug 4;11(1):16240. doi: 10.1038/s41598-021-95608-0

Network Example - Disease Comparison



Geodesic distance (d):
- average shortest path between node pairs.
- $\langle d_{AA} \rangle$ with-disease average.
- $\langle d_{AB} \rangle$ between-disease average.

Randomised model:
- place each diseases GDA's at random compute S_{AB}^{rand}.
- repeat 10,000.

Disease co-occurence:

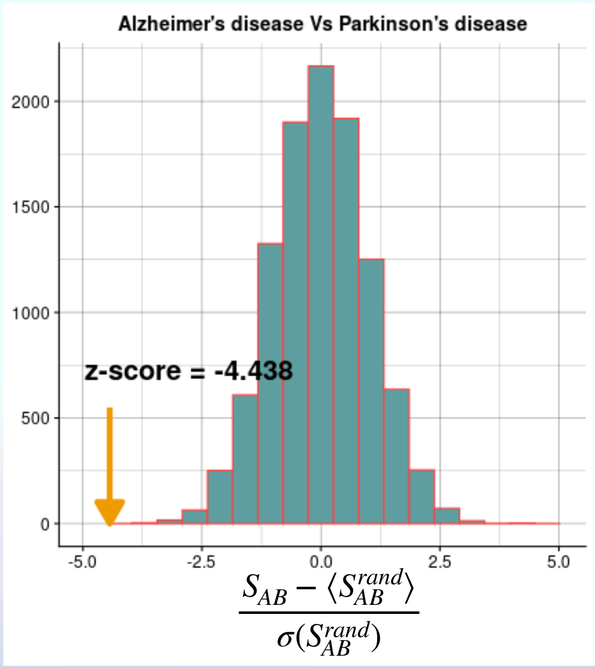
$$S_{AB} = \langle d_{AB} \rangle - \frac{\langle d_{AA} + d_{BB} \rangle}{2}$$

$S_{AB} \geq 0$

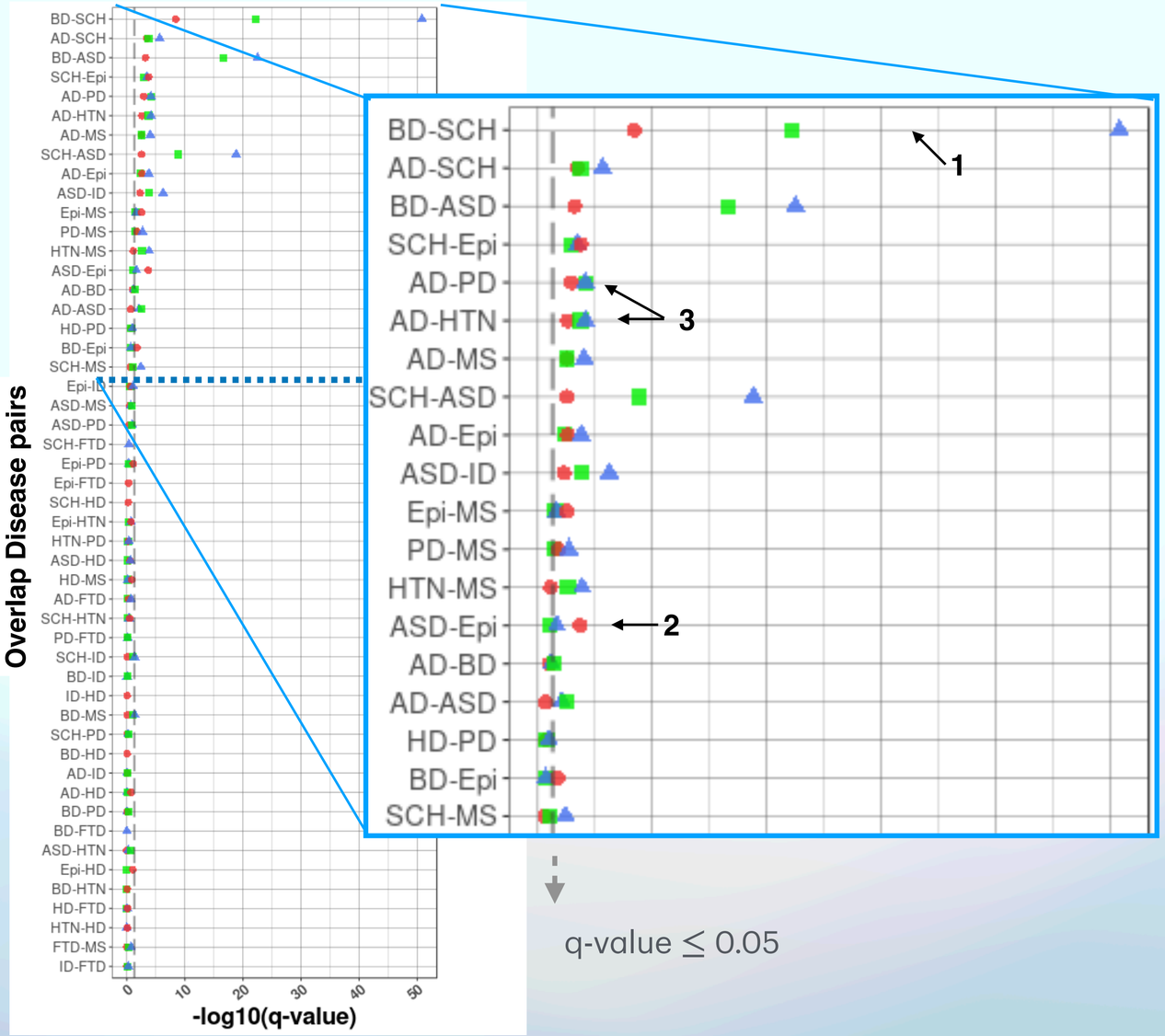
Separated disease pair

$S_{AB} < 0$

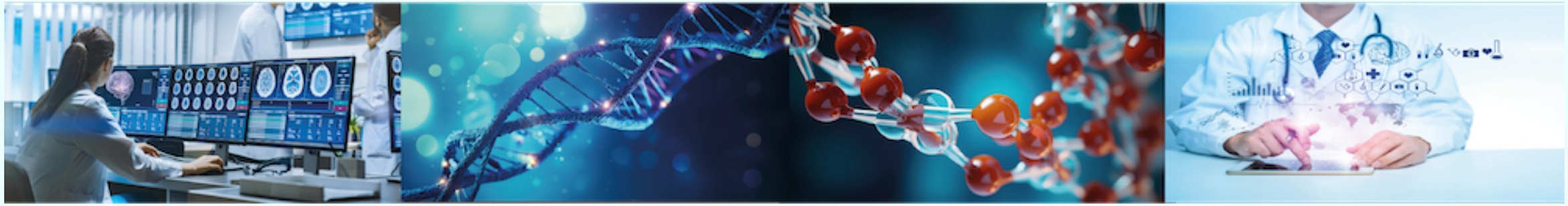
Overlap disease pair



Network Example - Disease Comparison



- clear overlap between neuropsychiatric diseases (e.g. bipolar disorder and schizophrenia)
- tendency for epilepsy to overlap more in the pre-synaptic network.
- expect Alzheimer's and Parkinson's diseases to overlap, but also find overlap with Huntington's



Programming for Biomedical Informatics

Next Lecture this Thursday - “Network Construction Techniques”

Please Bring your Laptop!

Ask Questions on the Piazza Discussion Board

Coding