

# Programming for Biomedical Informatics

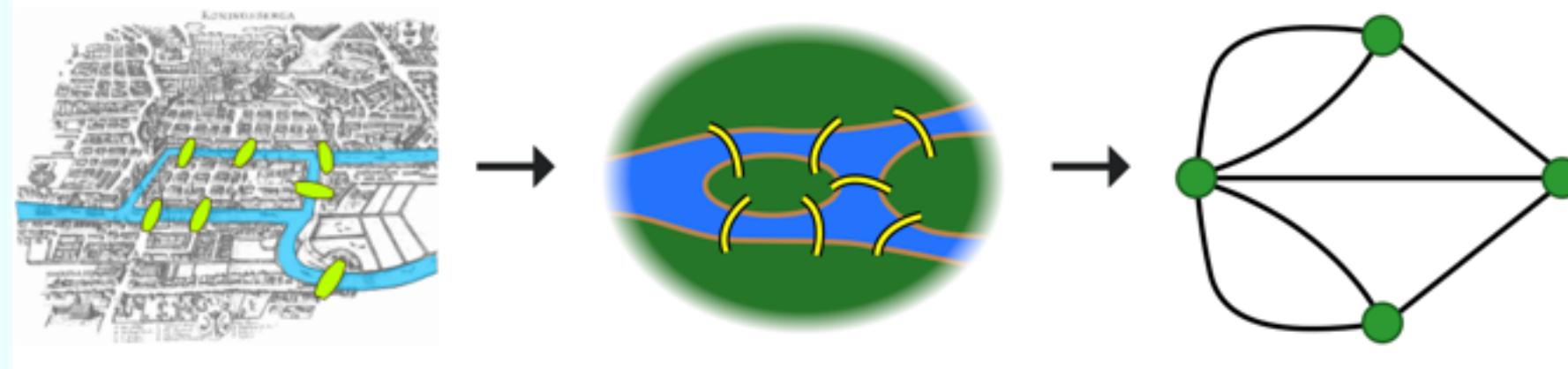
## Lecture 11 “Biological Networks”

<https://github.com/tisimpson/pbi>

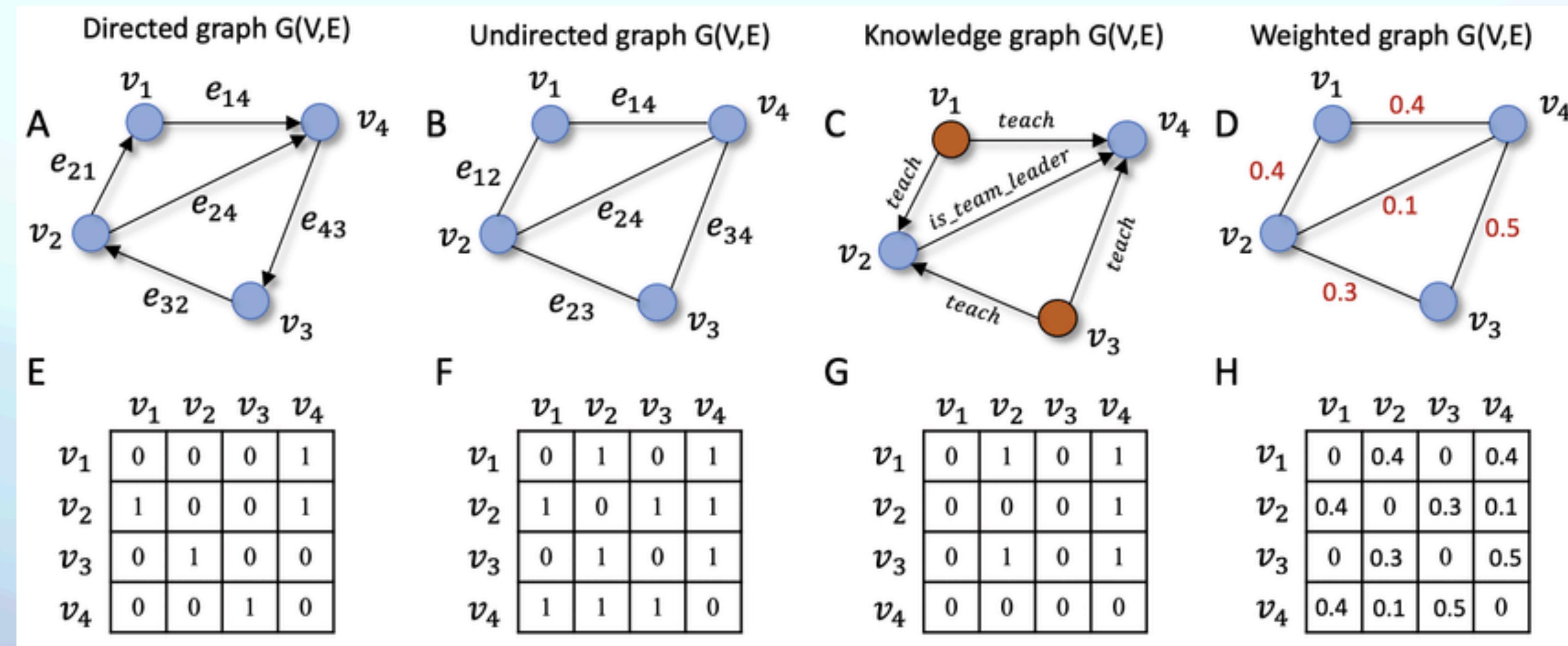
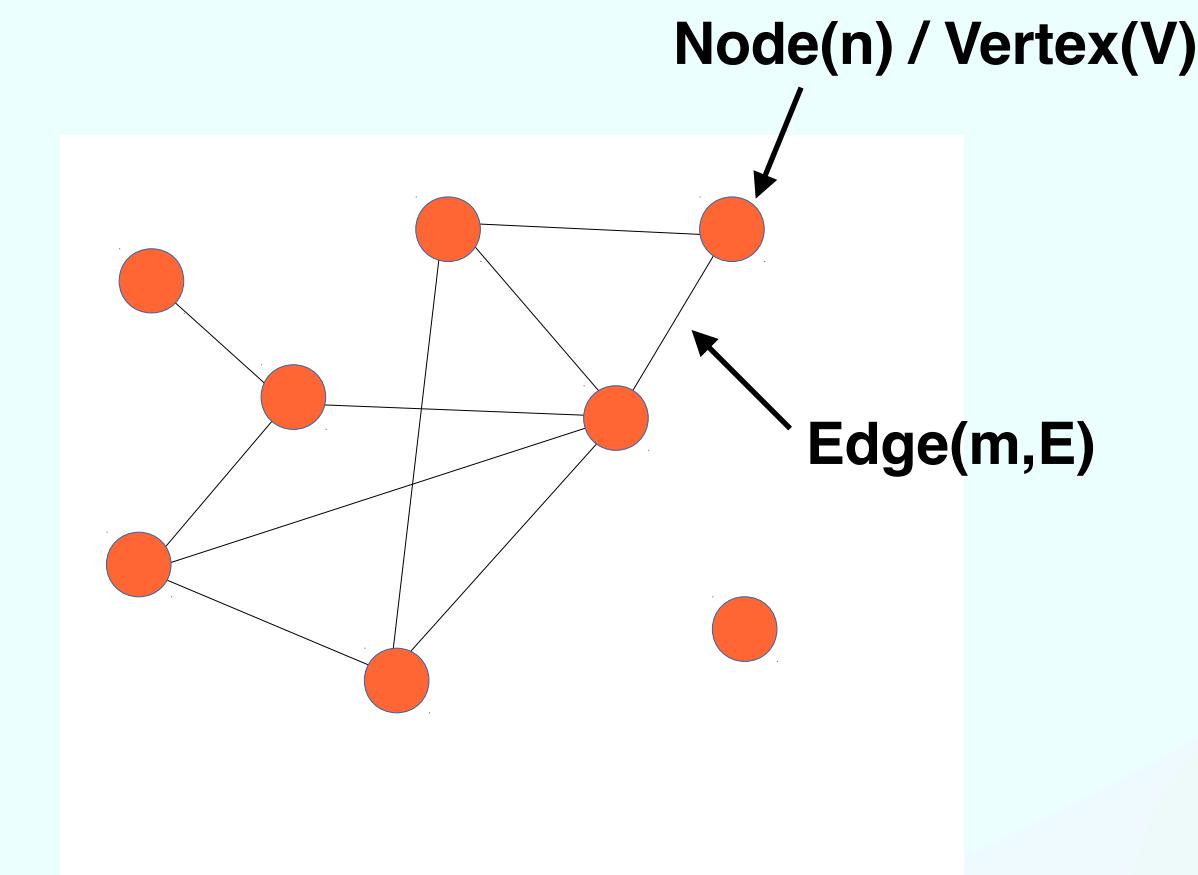
Ian Simpson  
ian.simpson@ed.ac.uk

Background

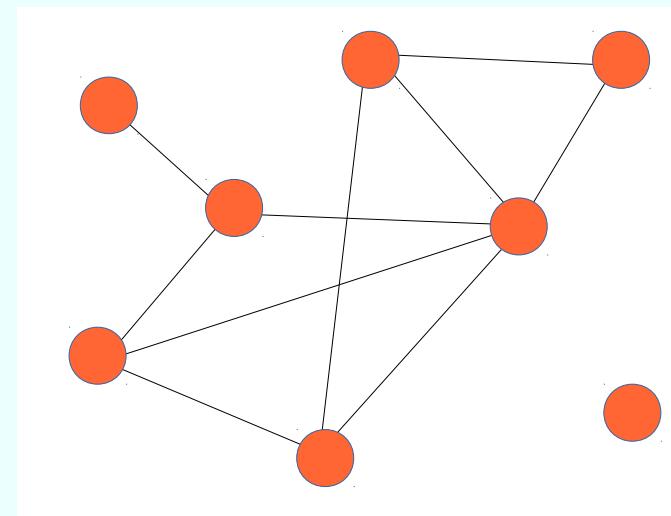
# The Anatomy of a Network



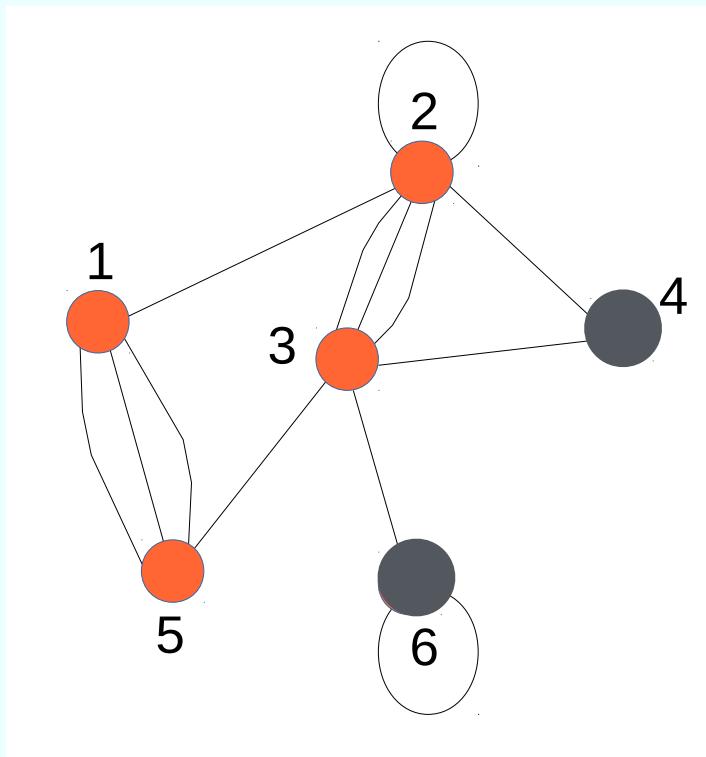
- The study of networks is concerned with understanding and modelling the behaviour of real-world systems.
- In mathematics called Graph theory.
- First network Euler's Königsberg bridge problem (1736).



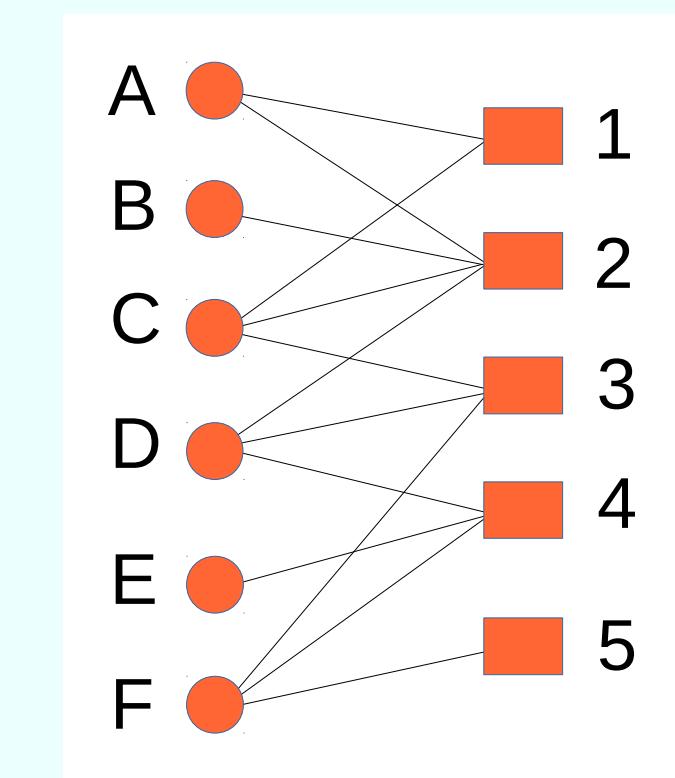
# Types of Network



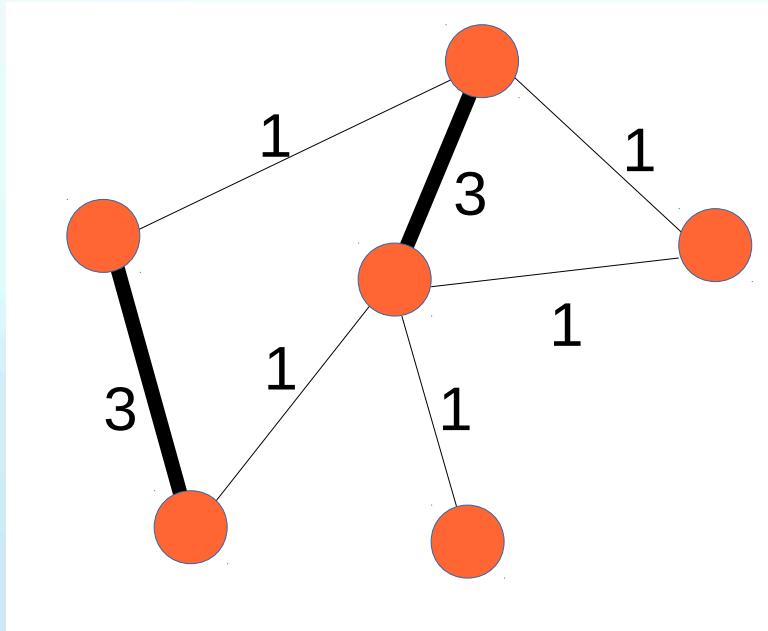
Simple



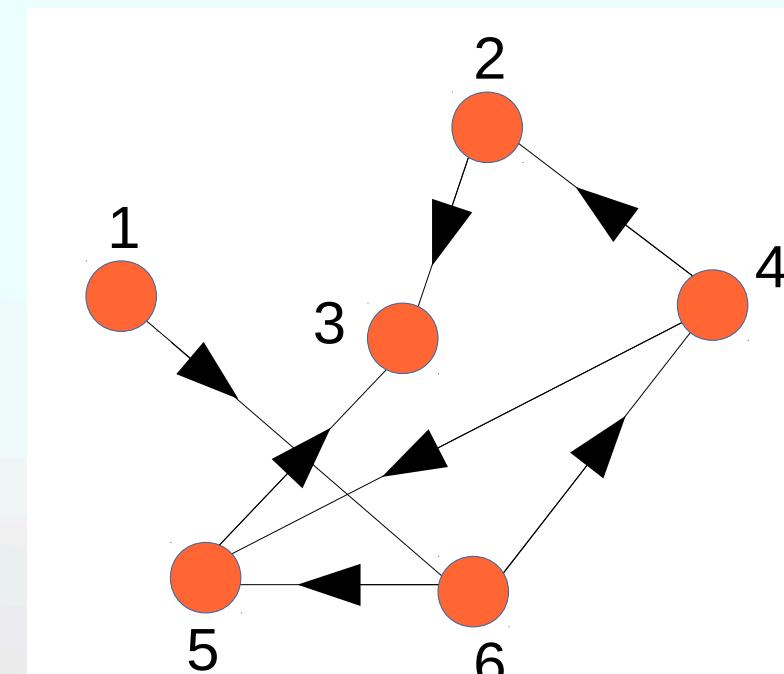
Multigraph



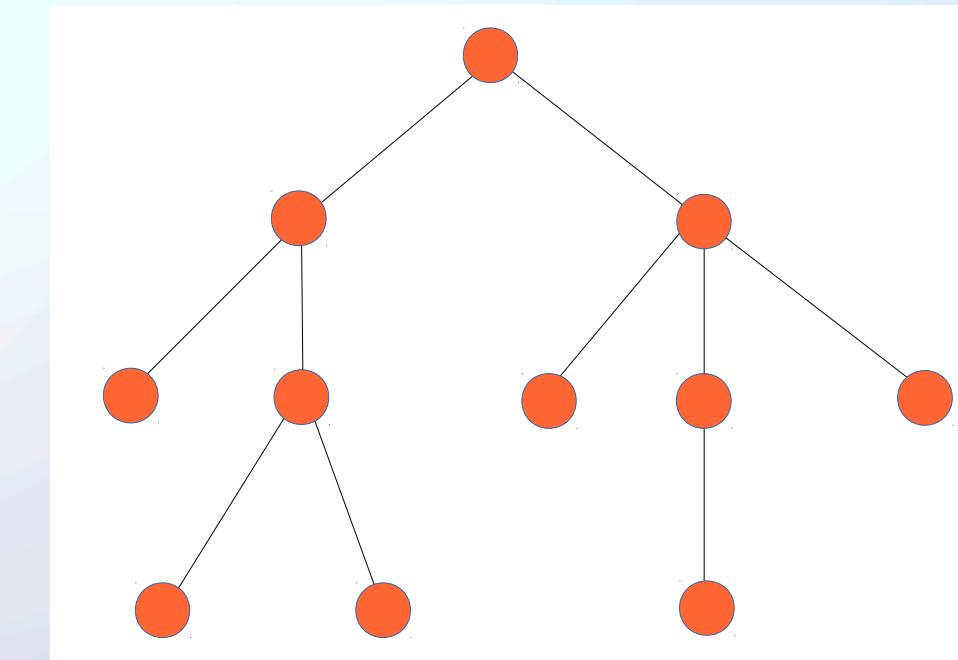
Bipartite



Simple (weighted)



Directed

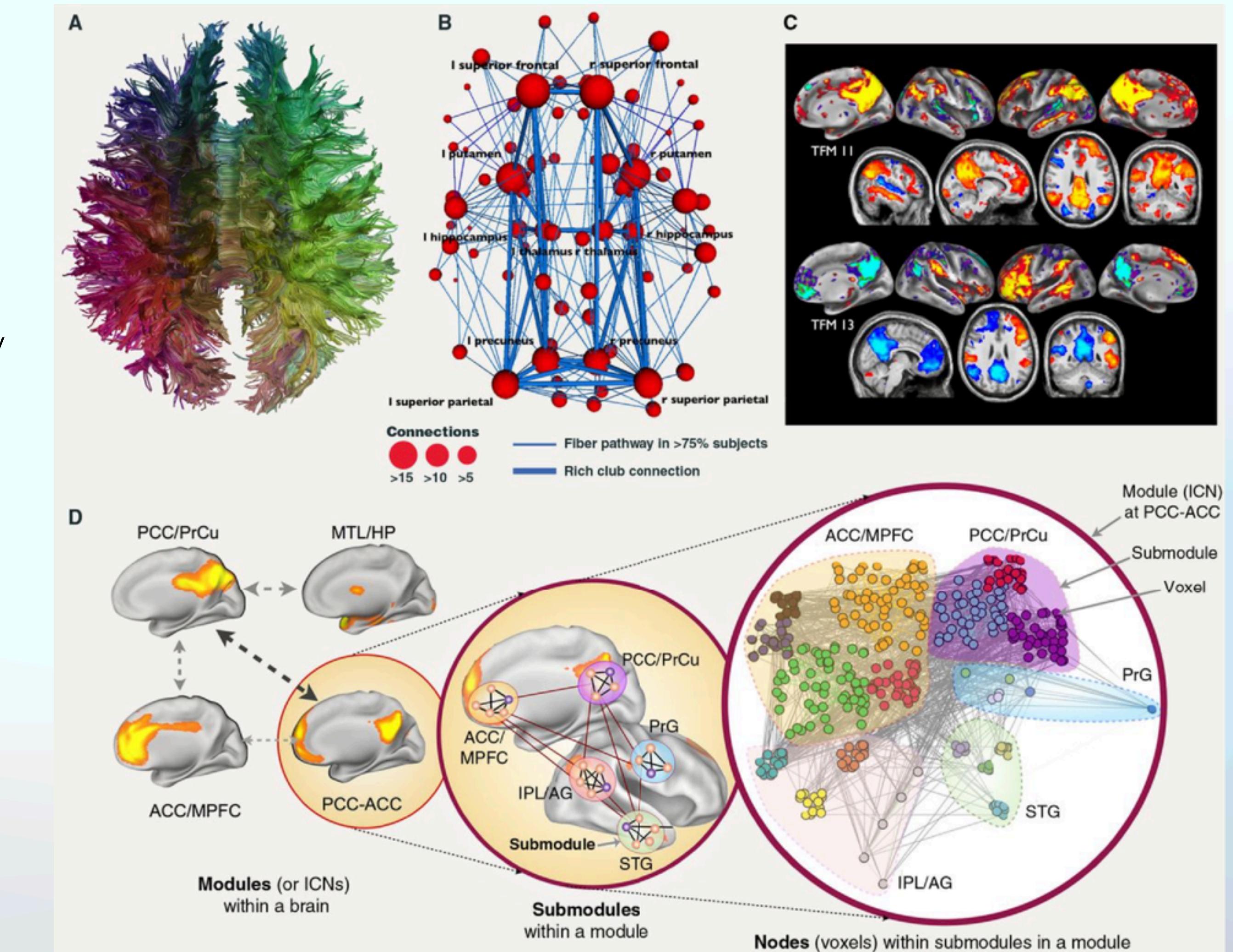


Tree

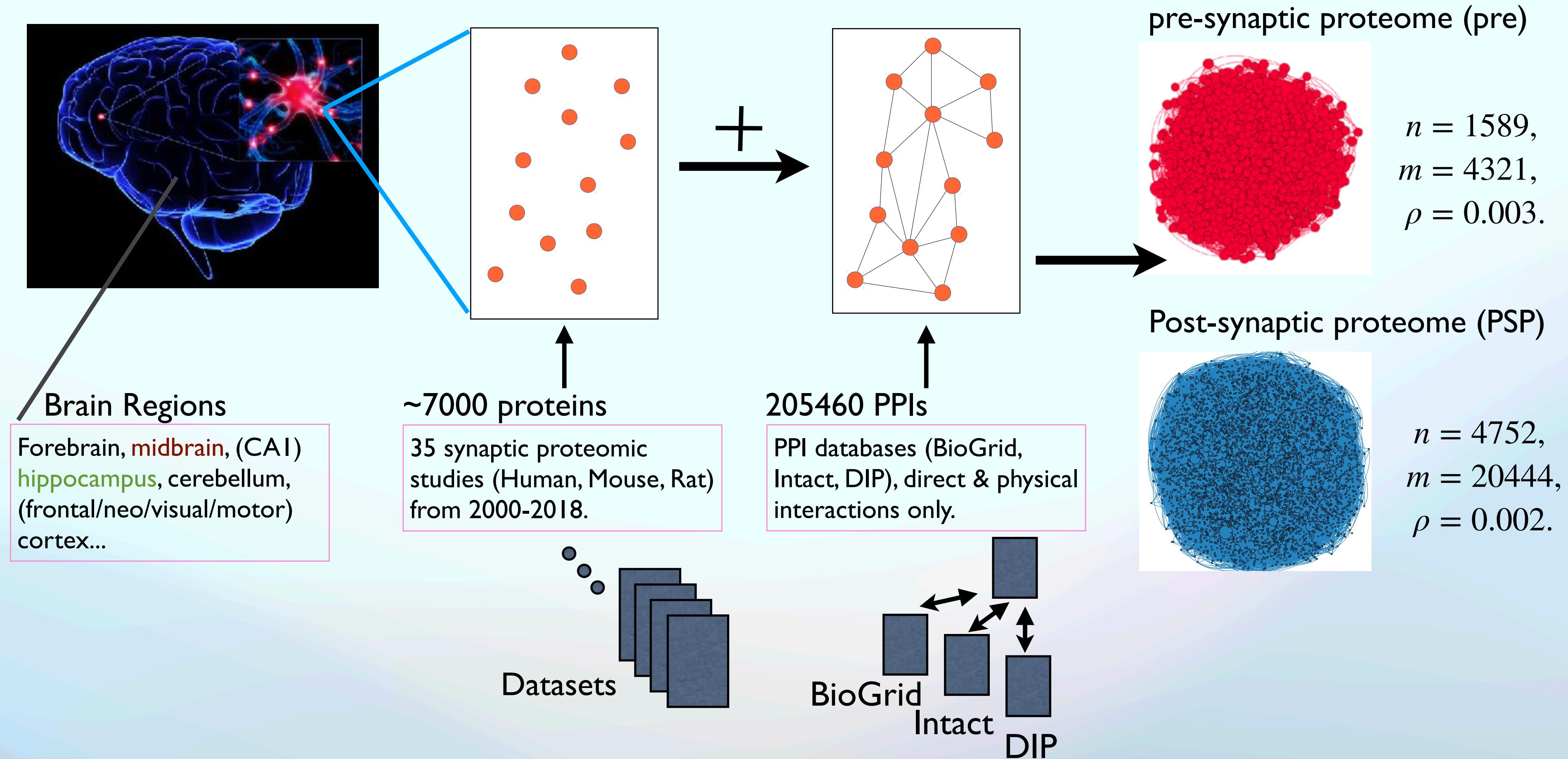
# Simple Network - Connectivity Maps in the Human Brain

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.

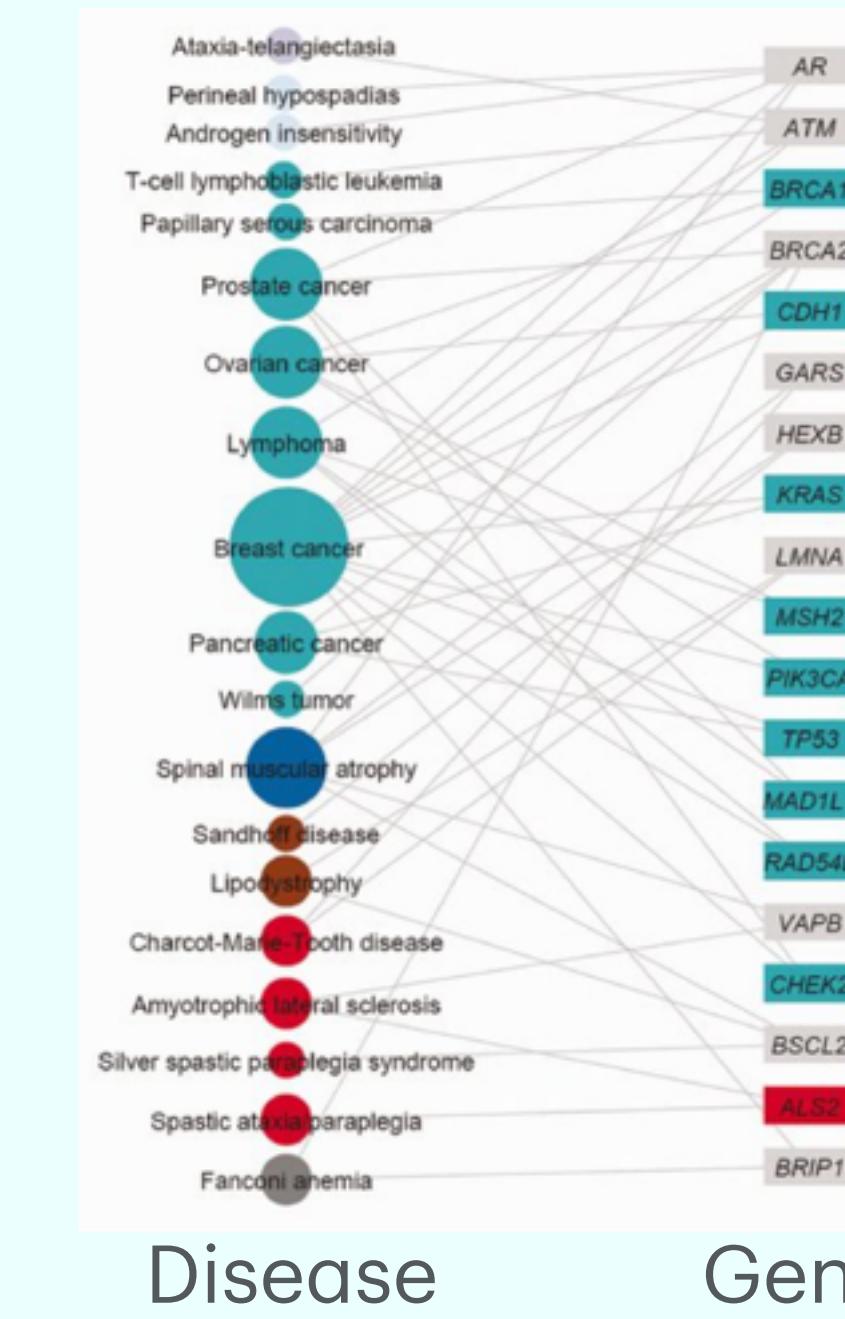
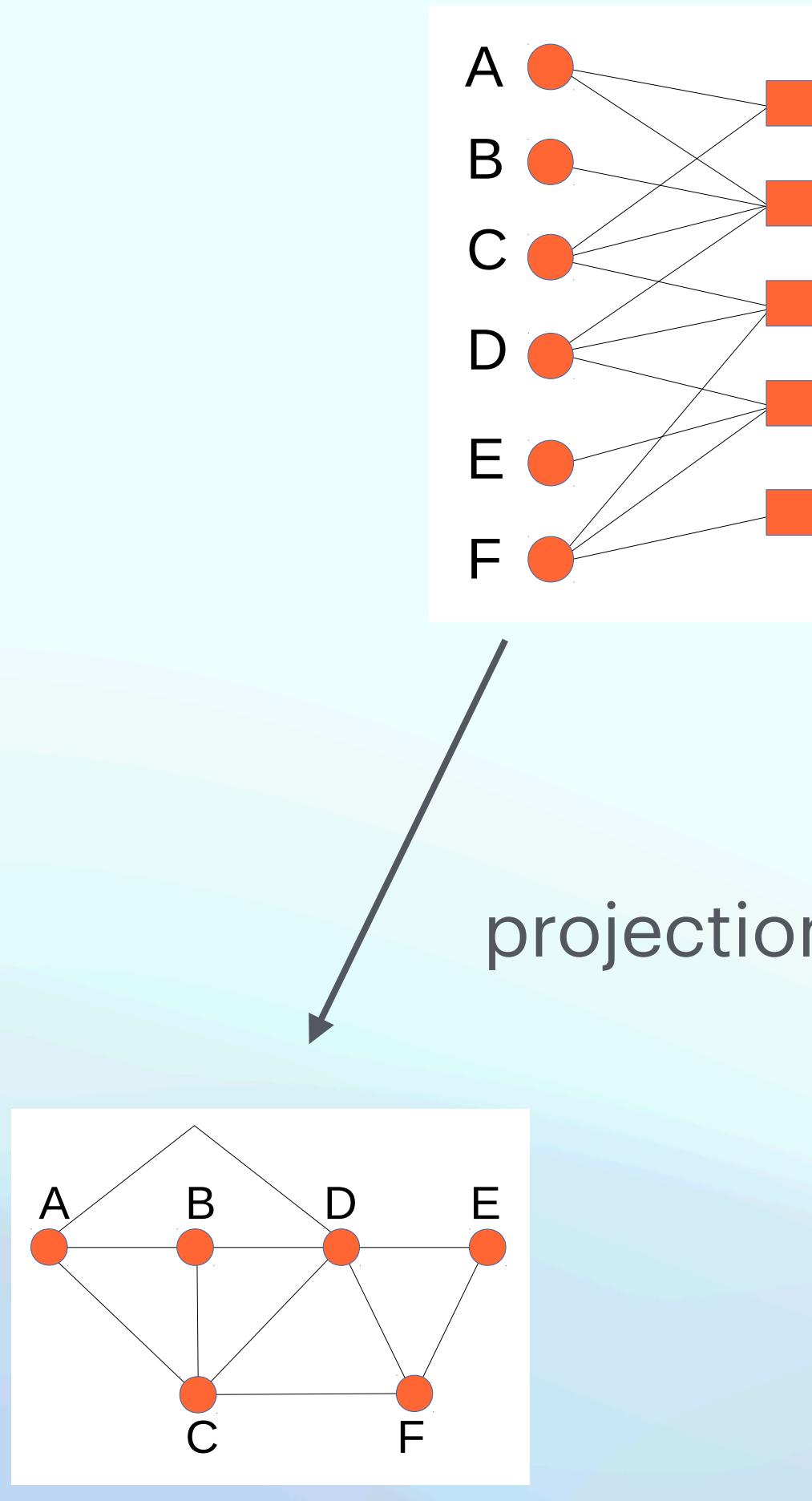
In this example a functional network of the brain is modelled based on fMRI data.



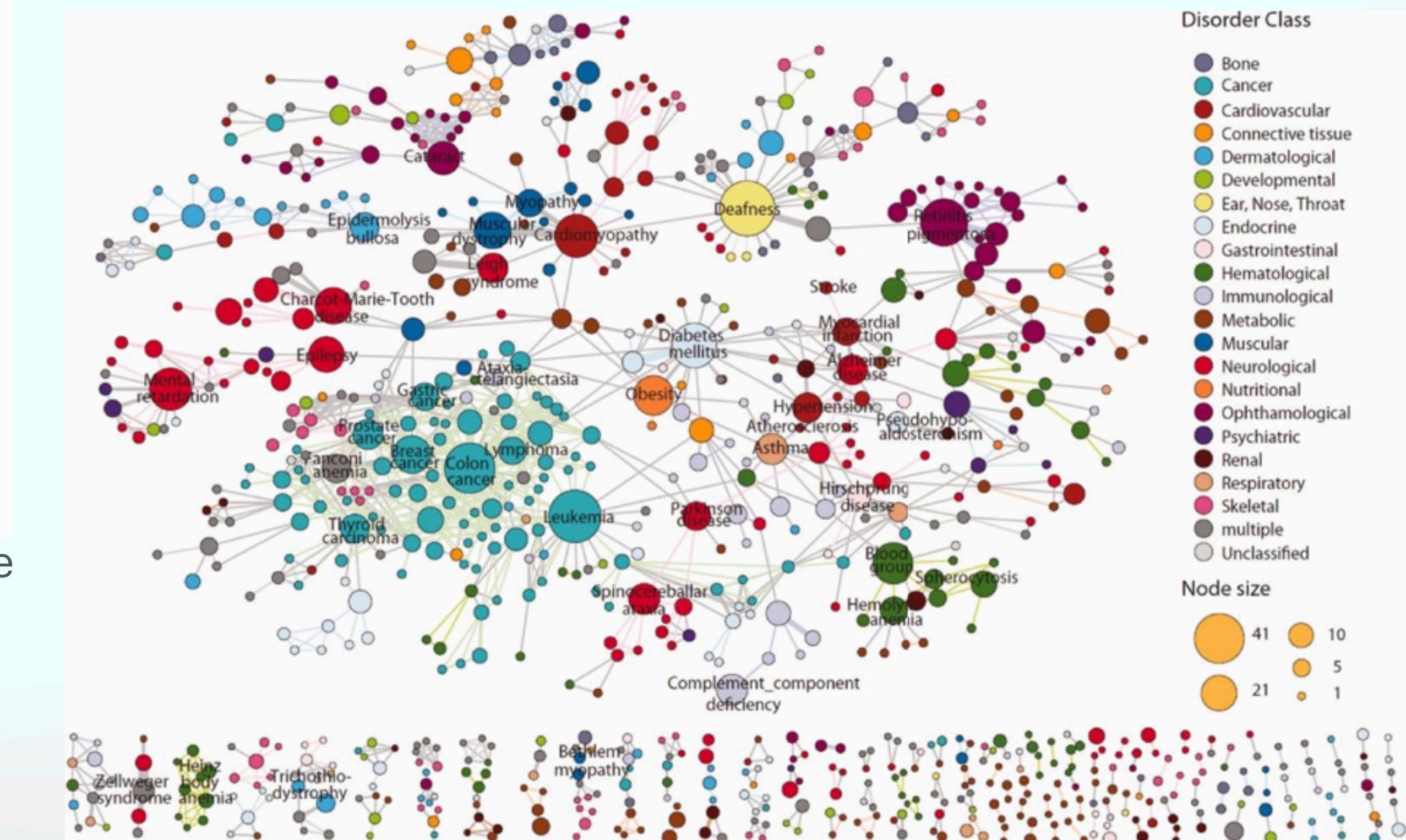
# Simple Network - Protein Interactions at the Synapse



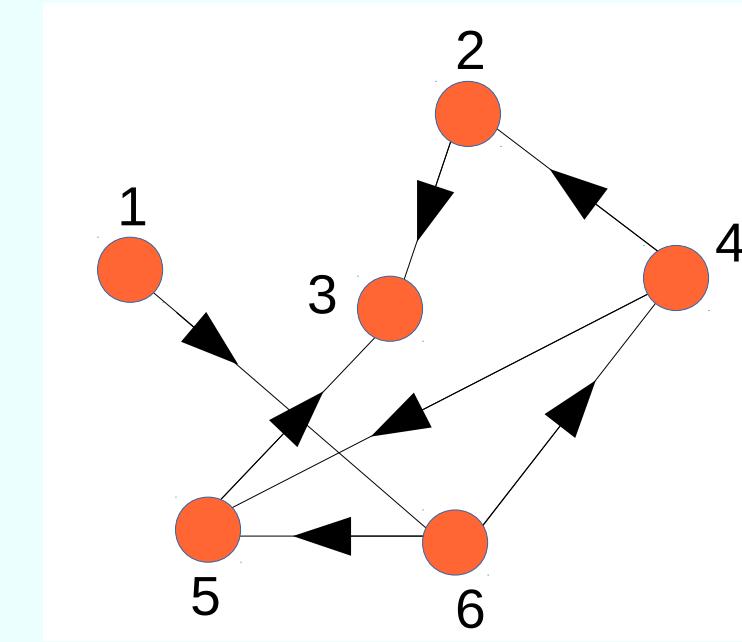
# Bipartite Networks



The Human Diseaseome



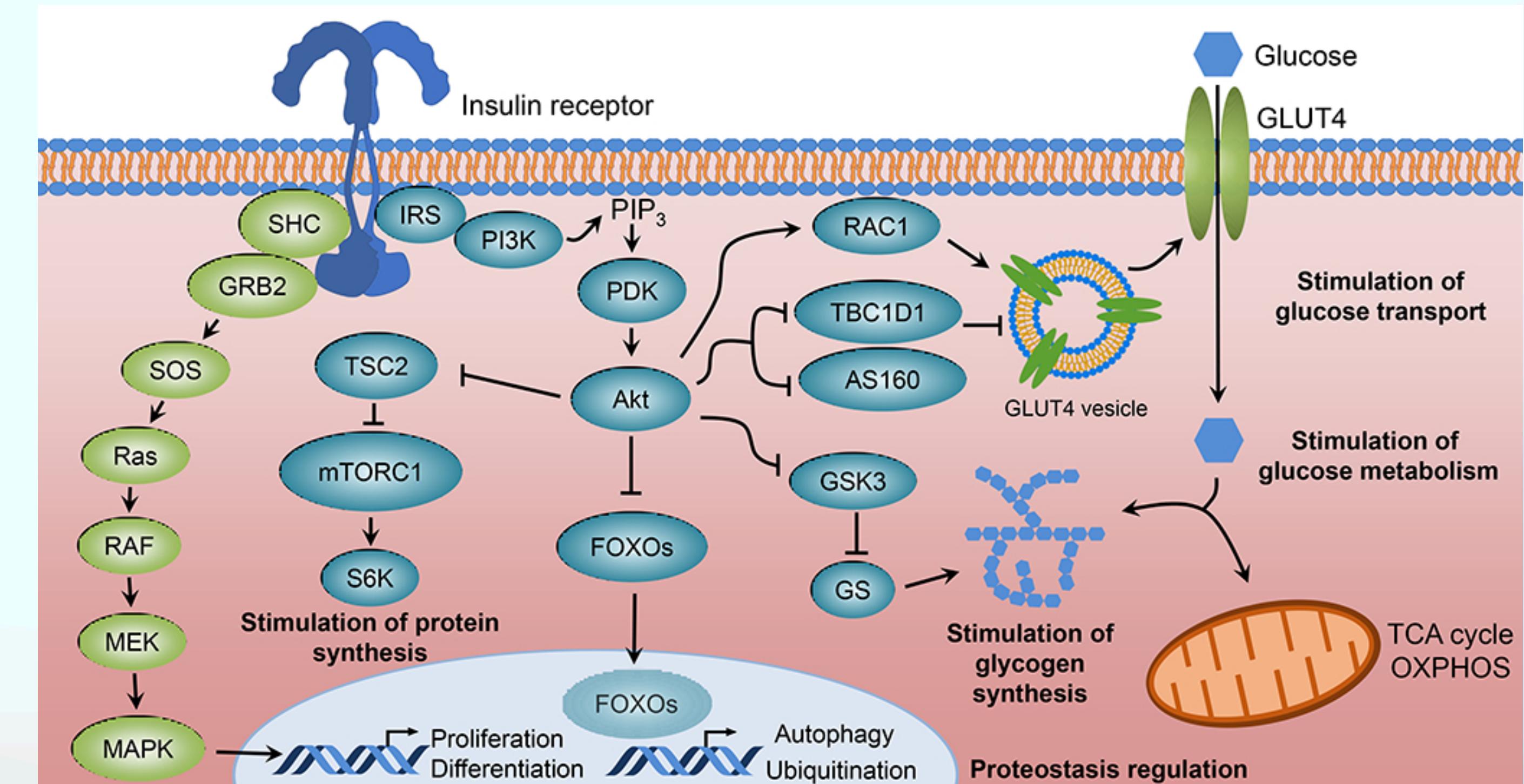
# Directed Networks



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.

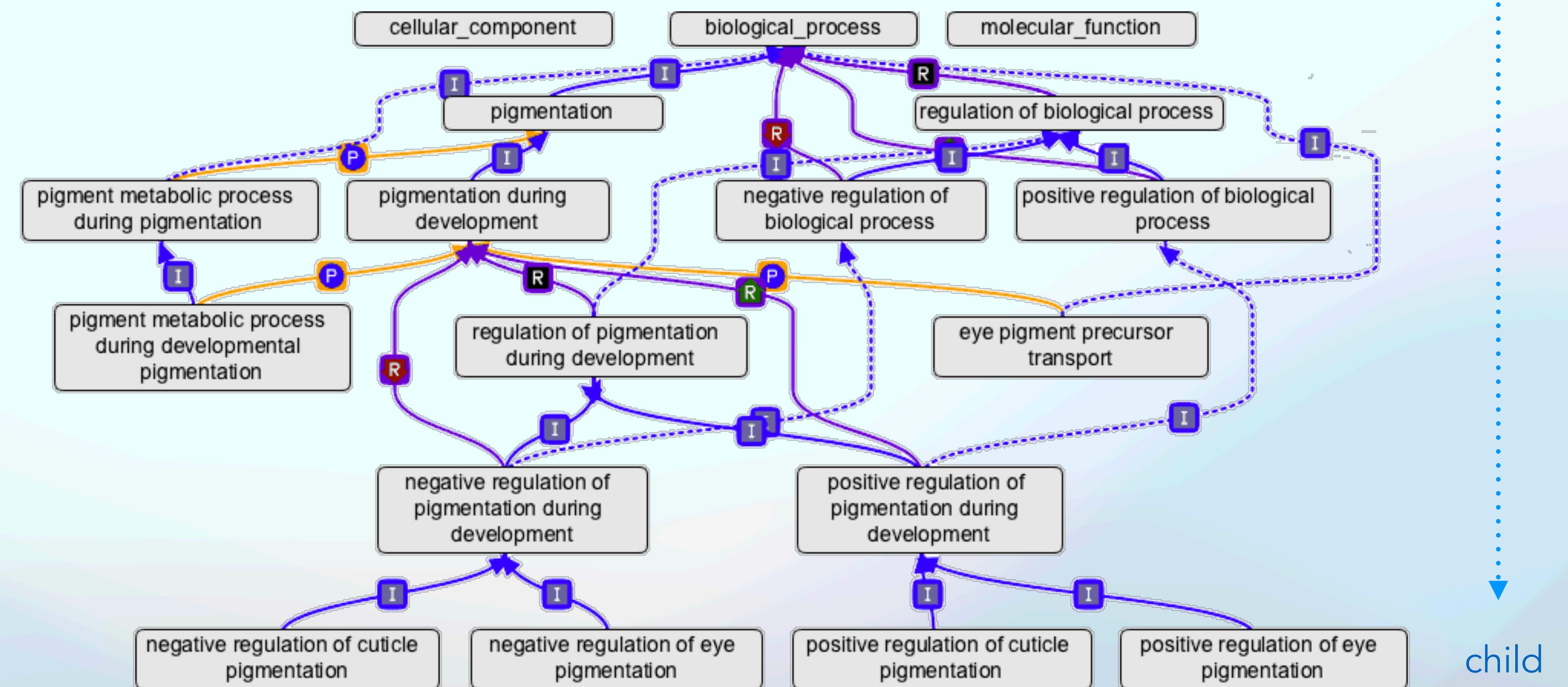
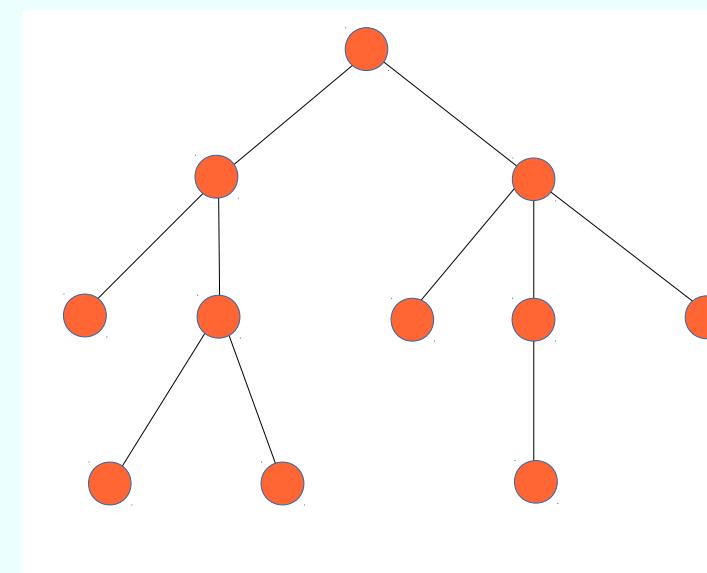


# Hierarchical Networks

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

low specificity roots

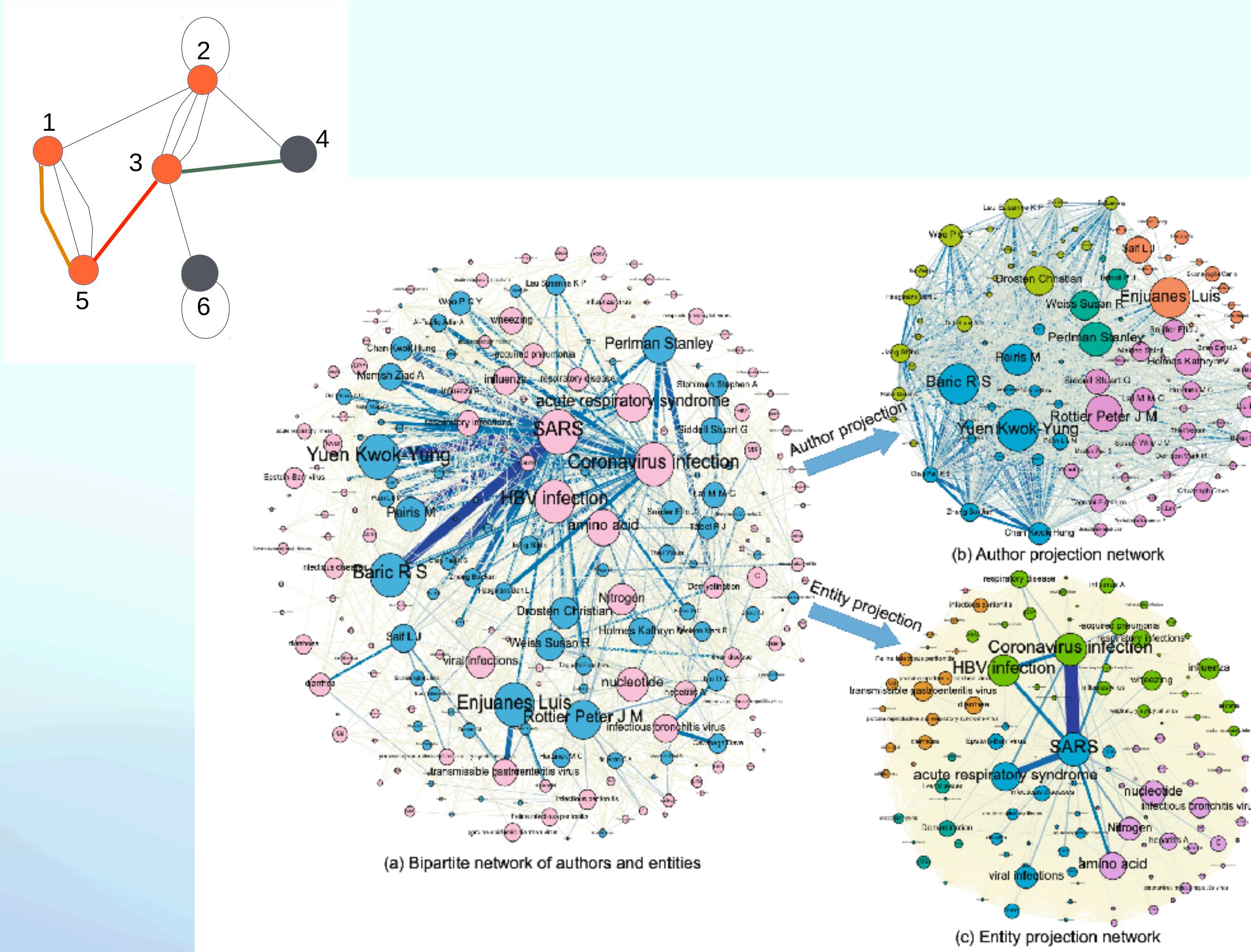
parent



high specificity leaves

child

# 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.

# 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

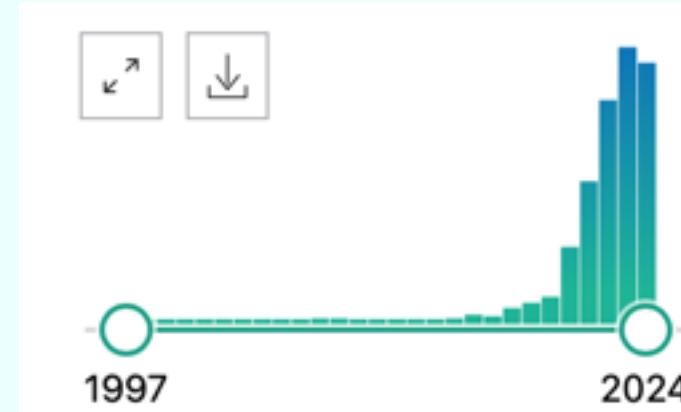
## **MONDO Disease Ontology - <https://mondo.monarchinitiative.org/>**

An integrative disease ontology consolidating information from multiple sources and species, linking diseases across databases like OMIM, ICD, and Orphanet

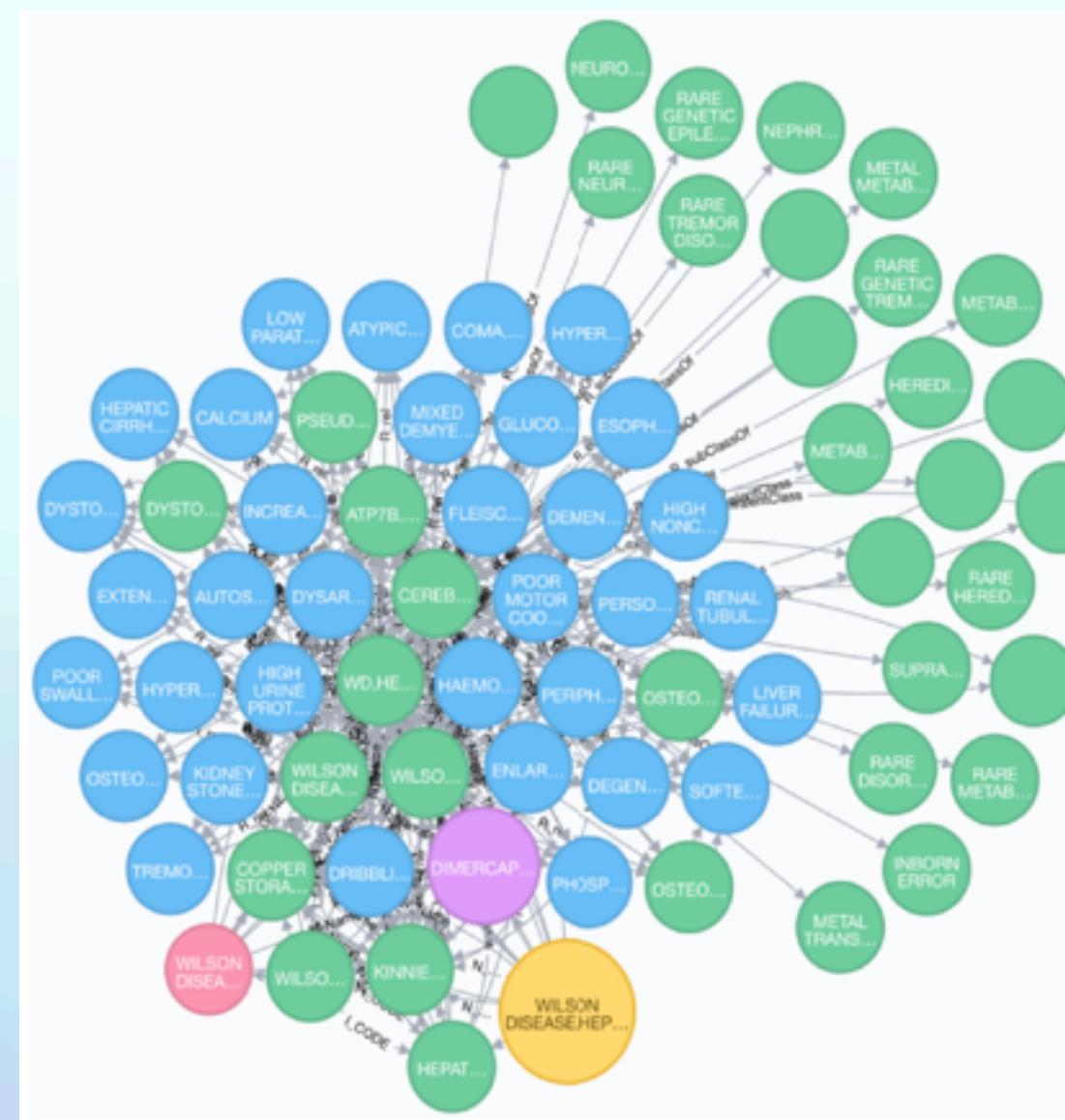
## **PharmGKB (Pharmacogenomics Knowledge Base) - <https://www.pharmgkb.org/>**

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/btz800  
Advance Access Publication Date: 1 August 2019  
Original Paper

OXFORD

Systems biology  
**Discovering protein drug targets using knowledge graph embeddings**

Sameh K. Mohamed <sup>1,2,\*</sup>, Vít Nováček <sup>1,2</sup> and Aayah Nounou <sup>3</sup>

<sup>1</sup>Data Science Institute, College of Engineering and Informatics, <sup>2</sup>Insight Centre for Data Analytics, NUI Galway, Galway, Ireland and <sup>3</sup>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 23, 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 <sup>1\*</sup>, Dac-Trung Nguyen <sup>1†</sup>, Ivan Grishagin <sup>1</sup>, Noel Southall <sup>1</sup>, Eric Sid <sup>2</sup> and Anne Pariser <sup>2</sup>

Brigitte et al. Bioinformatics, 22(3), 2021, 1–14  
doi: 10.1093/bioinformatics/btaa110  
Method Review

OXFORD

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: +86-020-86318825; Email: renyafeng@uiaf.edu.cn

ARTICLES  
<https://doi.org/10.1038/s41587-021-01145-6>

nature biotechnology

Check for updates

## OPEN A knowledge graph to interpret clinical proteomics data

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

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 <sup>1,2,3</sup>, Chengkun Wu <sup>4</sup>, Goran Nenadic <sup>3\*</sup>, Wei Wang <sup>1</sup> and Kai Lu <sup>1</sup>

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

Bioinformatics, 37(23), 2021, 4587–4598  
doi: 10.1093/bioinformatics/btab694  
Advance Access Publication Date: 6 October 2021  
Applications Note

OXFORD

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

Chuming Chen <sup>1,\*</sup>, Karen E. Ross <sup>2</sup>, Sachin Gavali <sup>1</sup>, Julie E. Cowart <sup>1</sup> and Cathy H. Wu <sup>1,2</sup>

<sup>1</sup>Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA and <sup>2</sup>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:

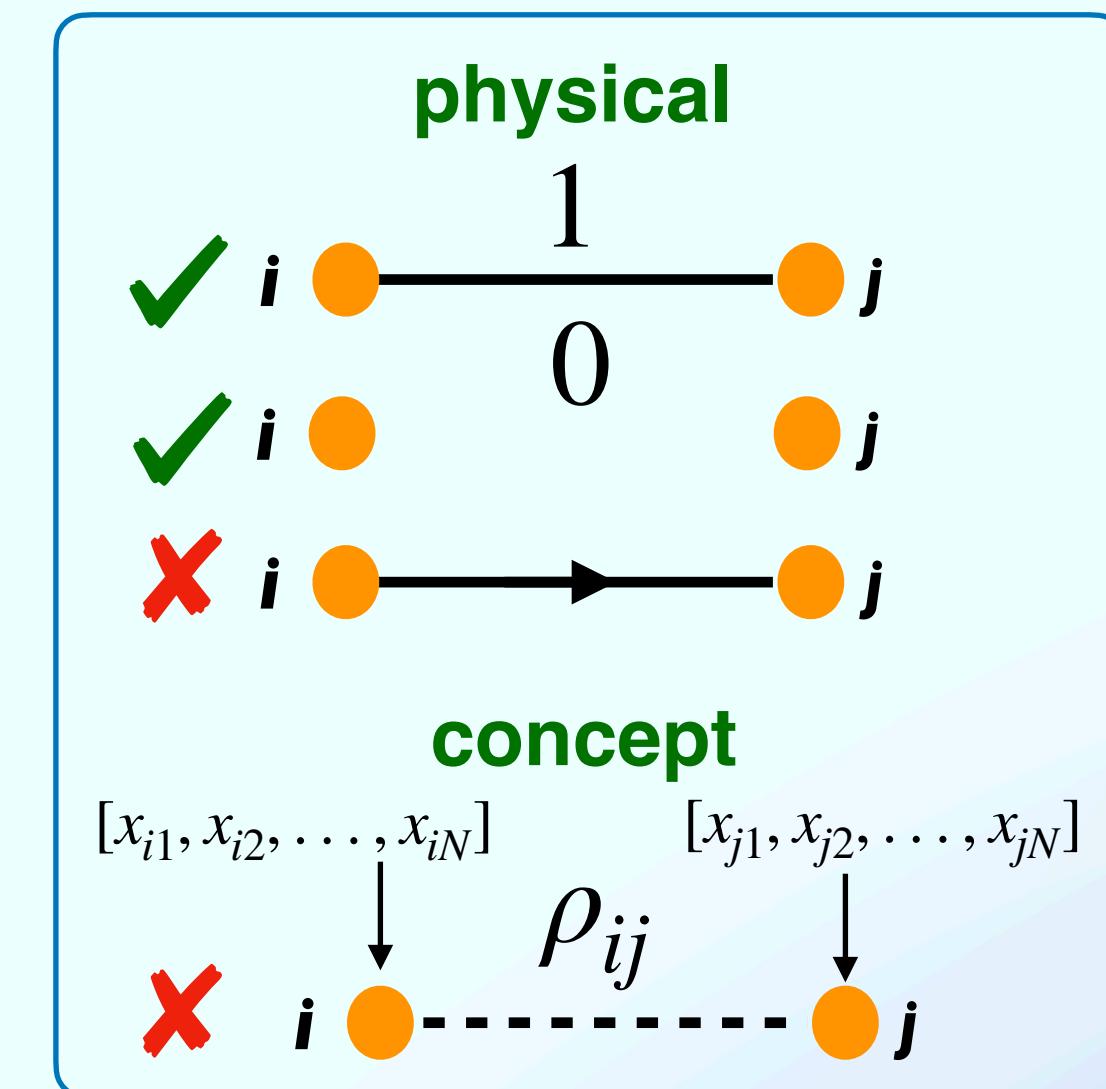
- every edge in network assigned either 1 “measured” or 0 “not measured”

## undirected:

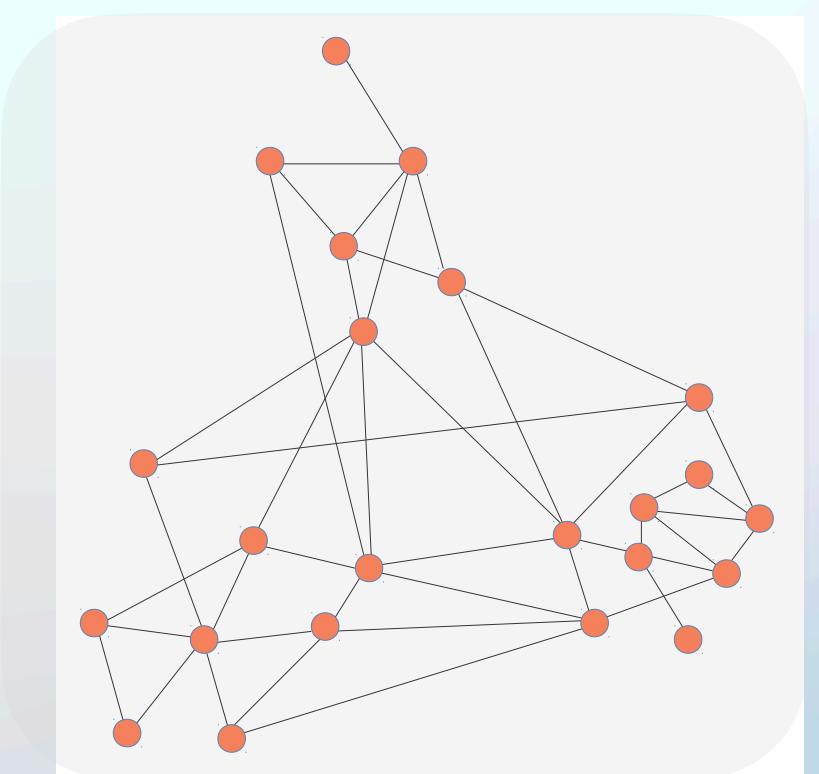
- every measured edge in network can go from  $i \rightarrow j$  or  $j \rightarrow i$  with equal magnitude.

## single component:

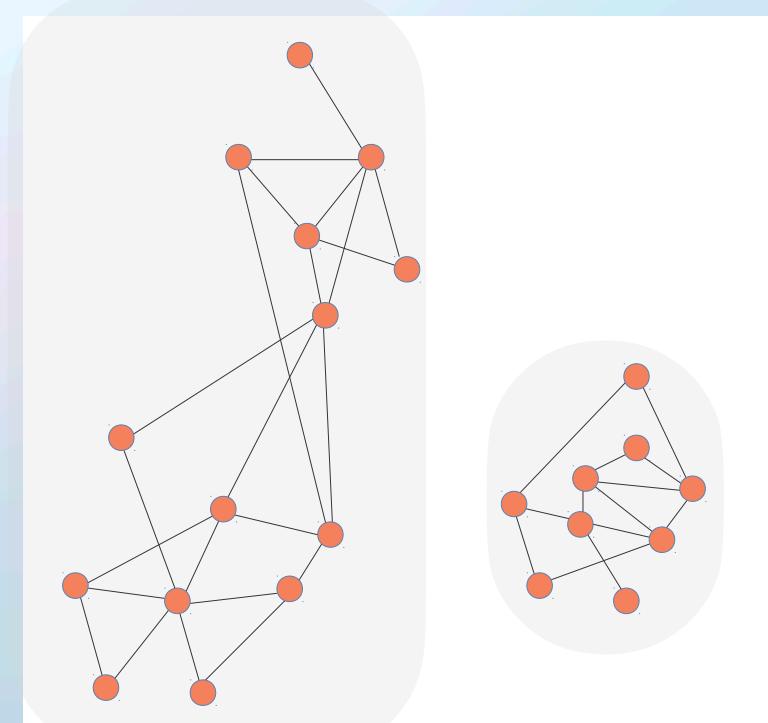
- fraction of nodes ( $S$ ) in the largest component is always 1.



$S = 1$



$S = 0.6$



# Network - Representations

Adjacency Matrix

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
<b>1</b>	0	1	0	0	0	0
<b>2</b>	1	0	1	0	1	0
<b>3</b>	0	1	0	1	0	0
<b>4</b>	0	0	1	0	1	0
<b>5</b>	0	1	0	1	0	1
<b>6</b>	0	0	0	0	1	0

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

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

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

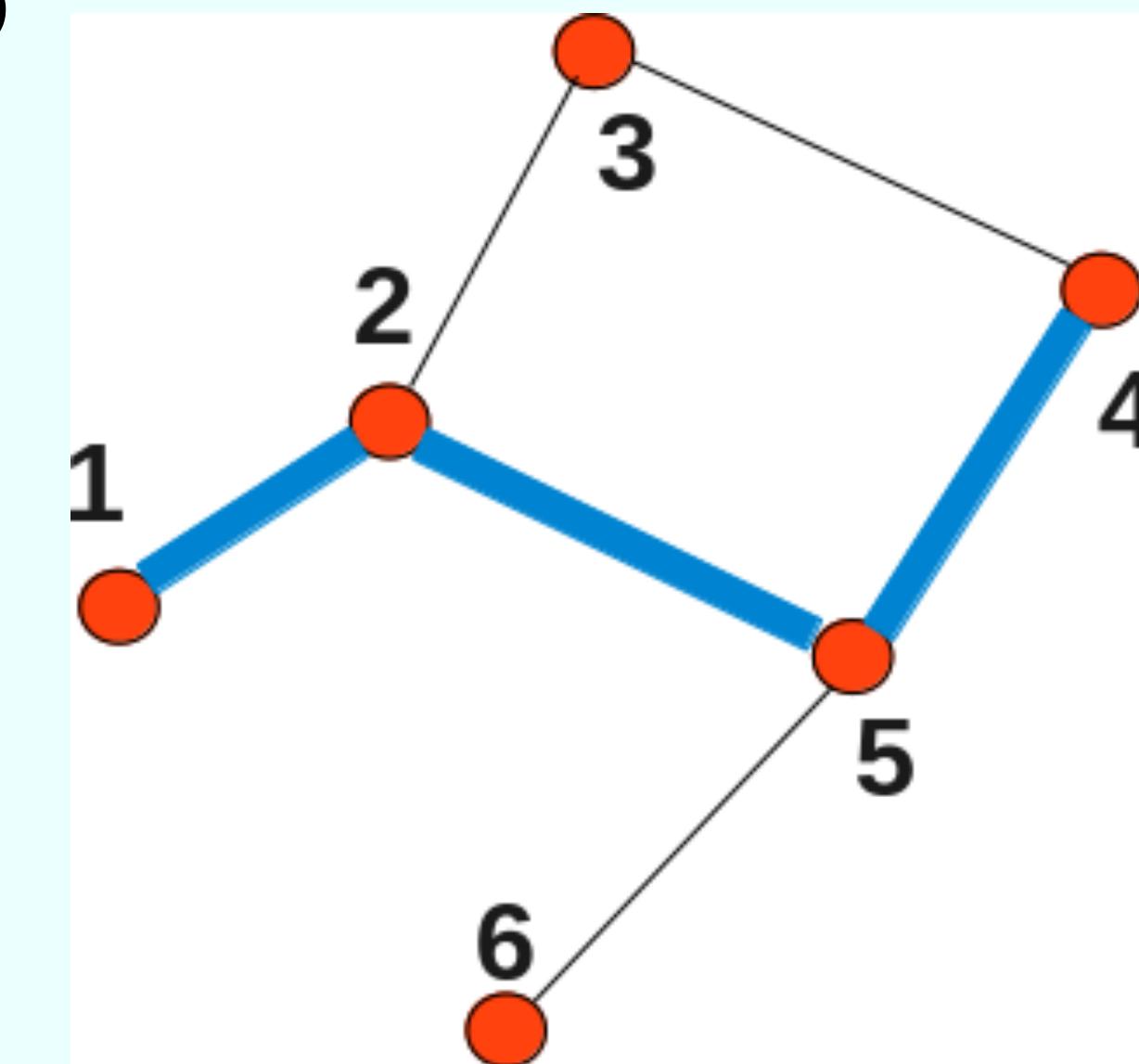
(edges in network)

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

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

Degree Matrix

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
<b>1</b>	1	0	0	0	0	0
<b>2</b>	0	3	0	0	0	0
<b>3</b>	0	0	2	0	0	0
<b>4</b>	0	0	0	2	0	0
<b>5</b>	0	0	0	0	3	0
<b>6</b>	0	0	0	0	0	1



Laplacian Matrix

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
<b>1</b>	1	-1	0	0	0	0
<b>2</b>	-1	3	-1	0	-1	0
<b>3</b>	0	-1	2	-1	0	0
<b>4</b>	0	0	-1	2	-1	0
<b>5</b>	0	-1	0	-1	3	-1
<b>6</b>	0	0	0	0	0	1

$$\mathbf{L} = \mathbf{D} - \mathbf{A} =$$

# Network - Measures

## Measures:

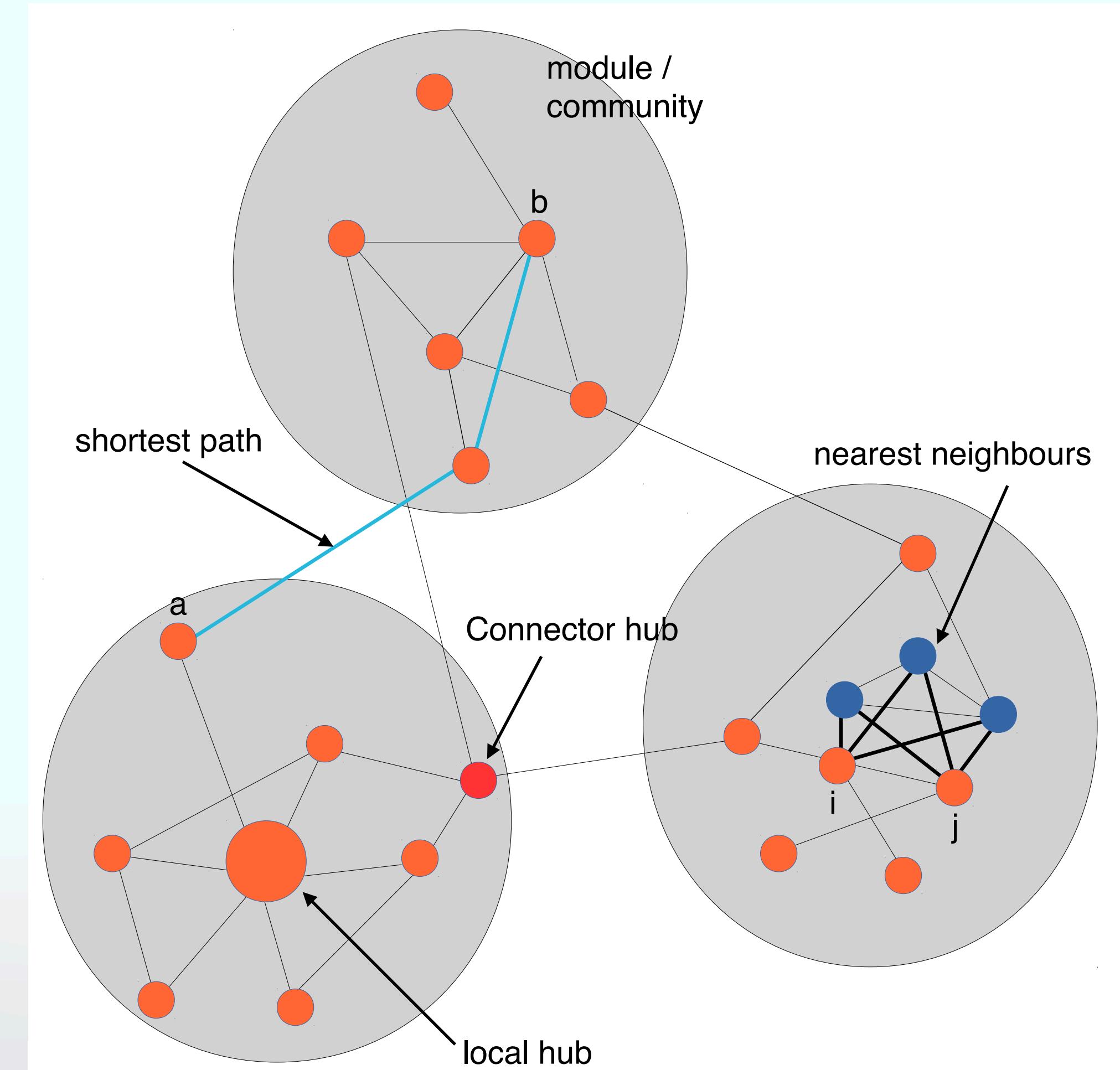
- From our mathematical representation, we can define many concepts which reveal a networks underlying character.

## 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:

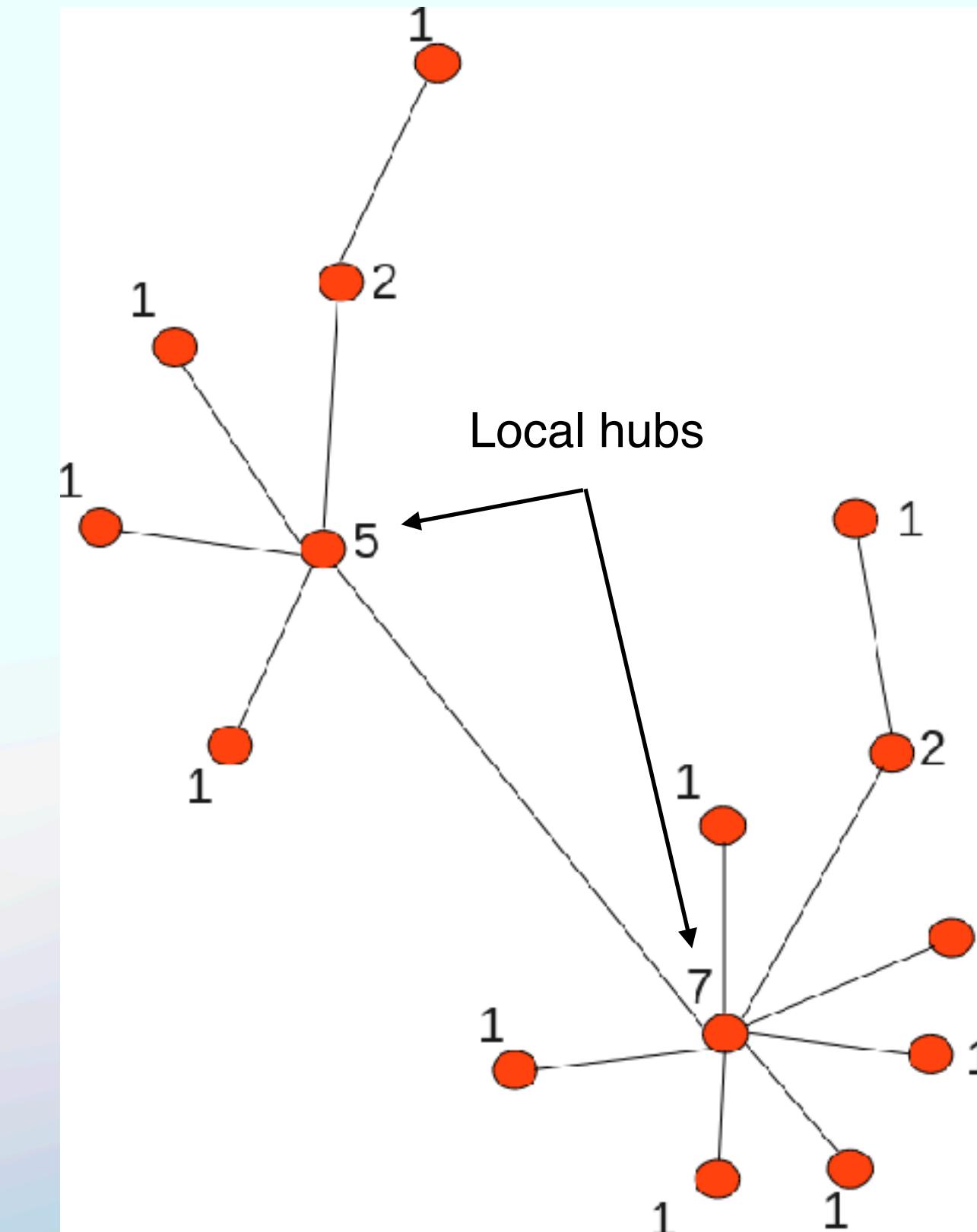
- can be used as a centrality score
- to generate random networks
- (a test) degree distribution of many real-world networks follows a power-law:  $\propto k^{-\alpha}$

Node Degree Centrality:

- Number of neighbours, denoted by  $k$
- local hubs  $\implies \begin{cases} \text{highest degree nodes} \\ \text{disease related proteins} \end{cases}$

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

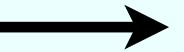
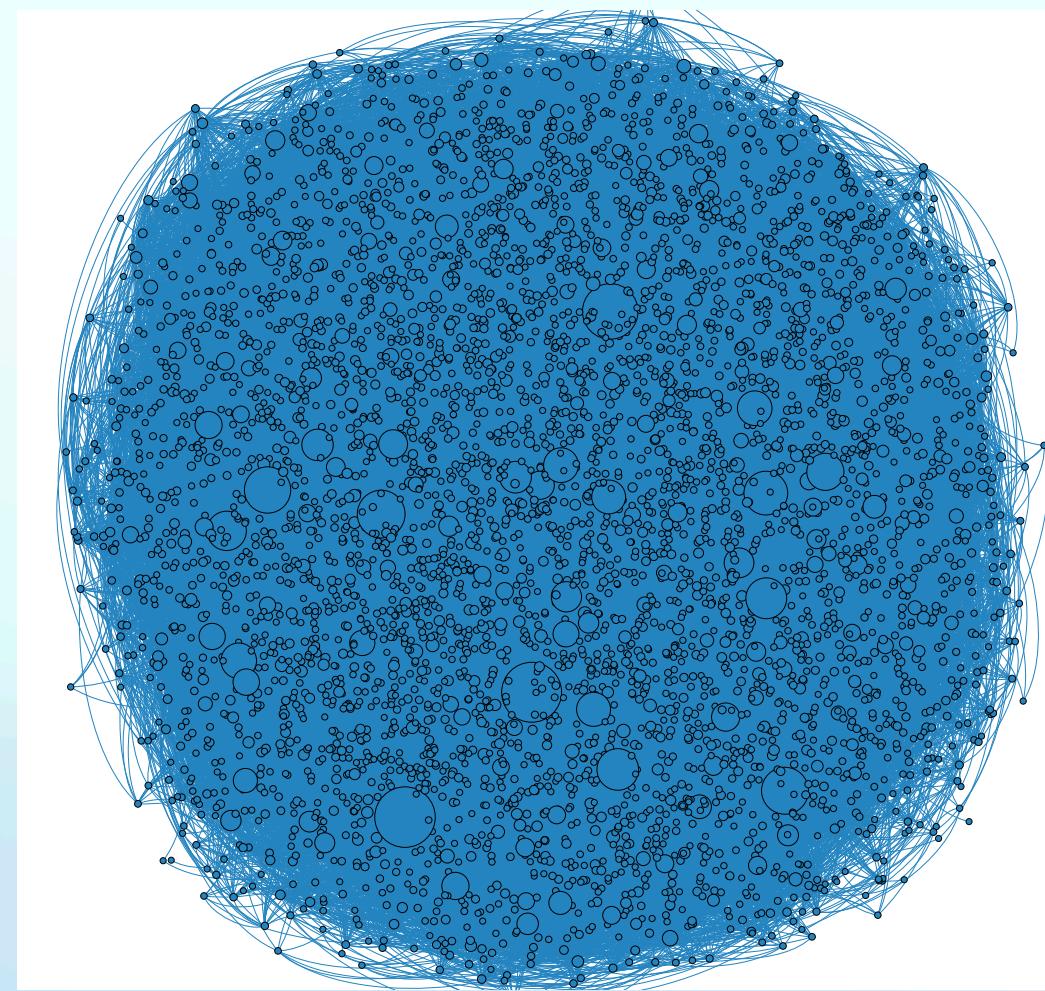


# Network - Degree

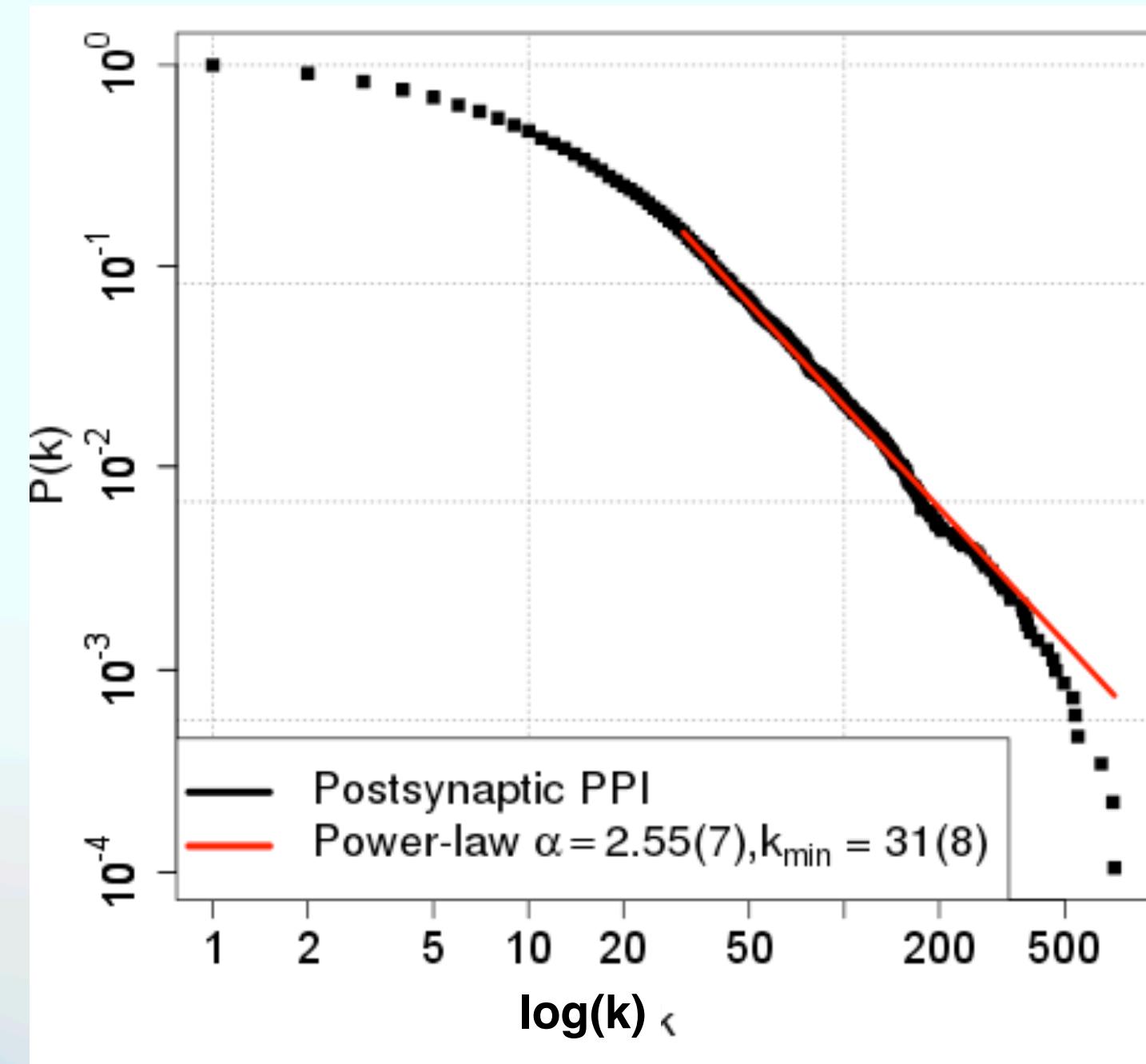
One of the most useful concepts, encoding a network's local structure:

- can be used as a centrality score.
- generate random networks...
- (at least) degree distribution of many real-world networks follows a power-law:  $\propto k^{-\alpha}$

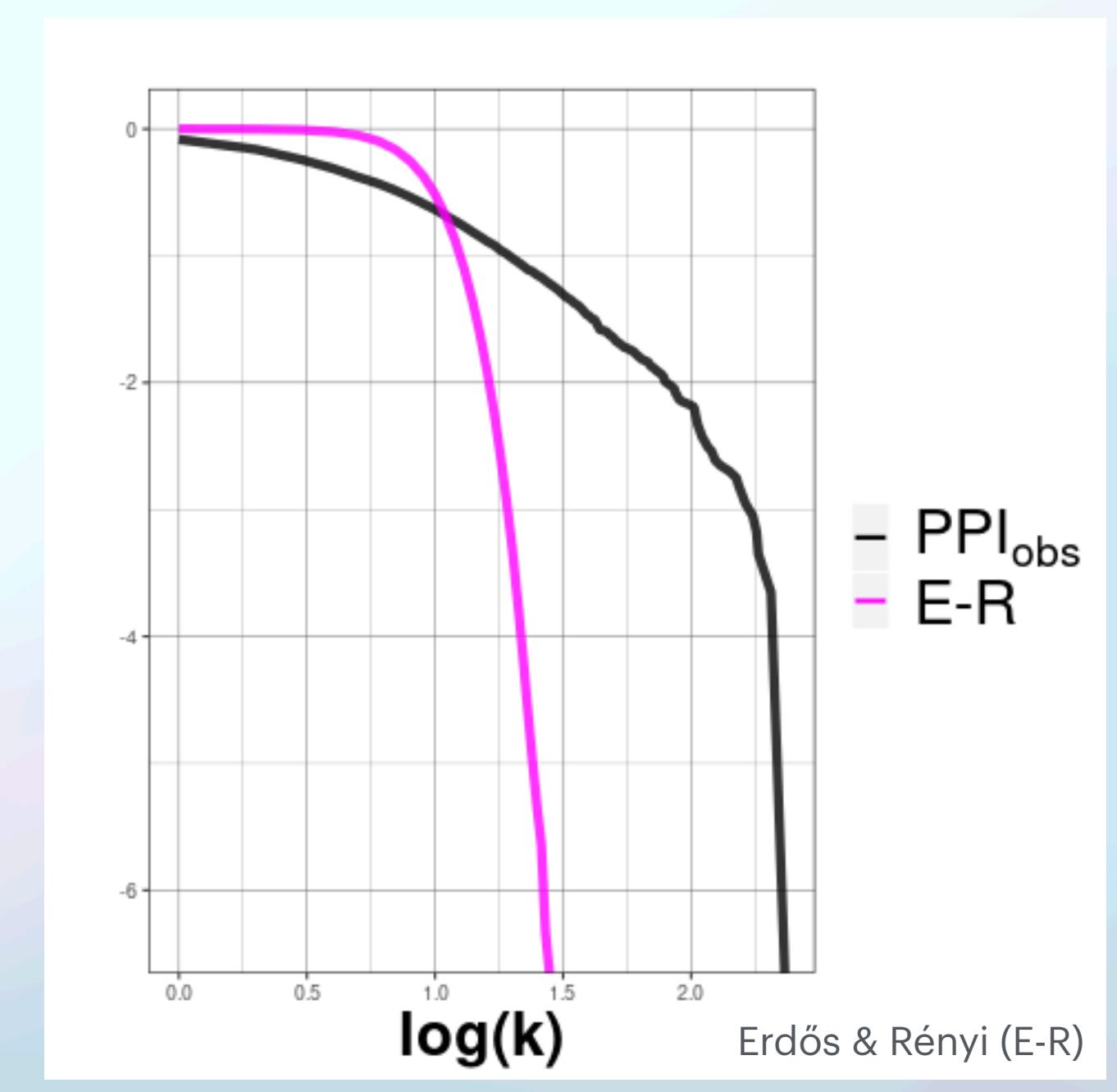
Network



Degree Distribution



Random Distribution

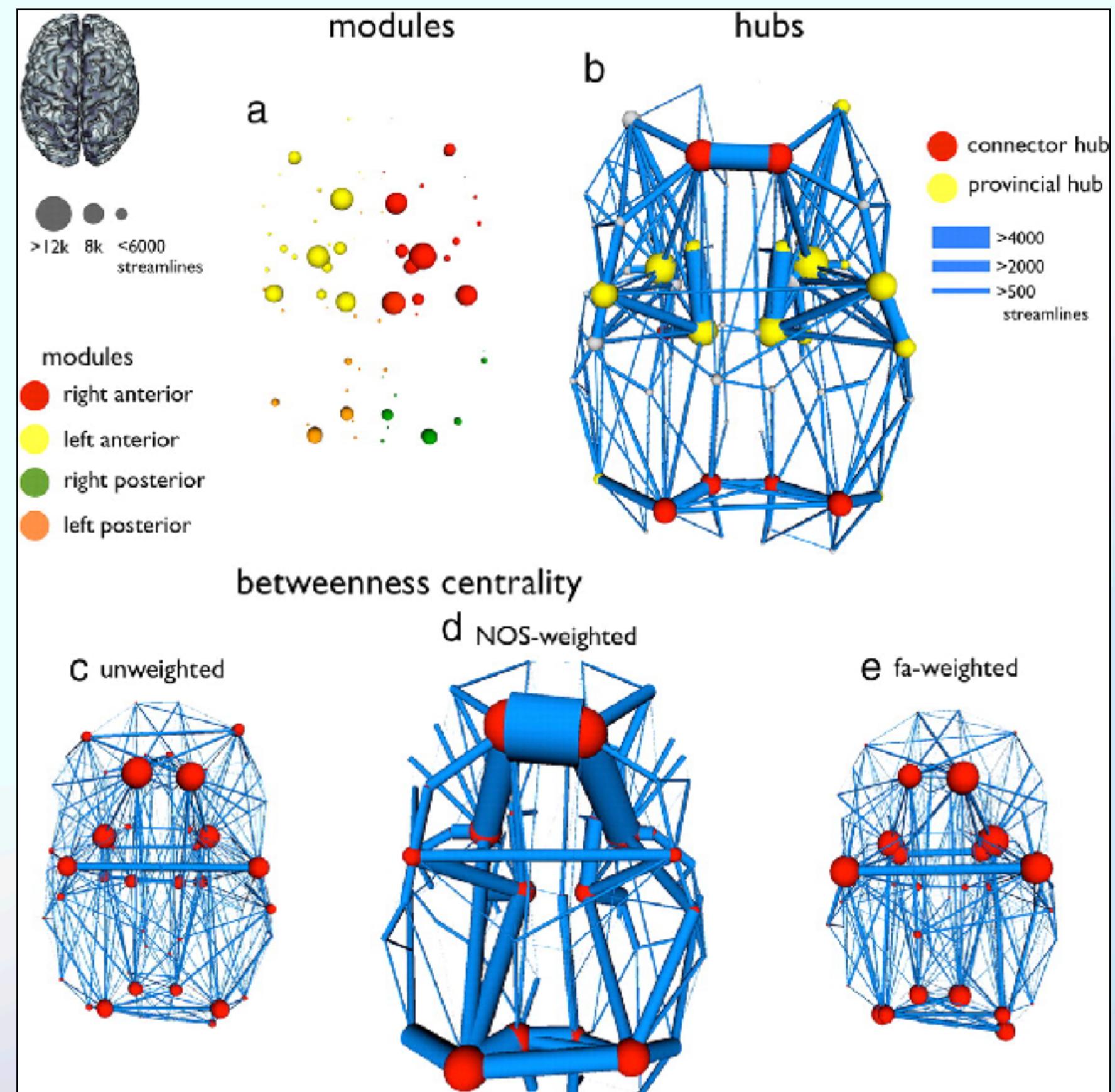
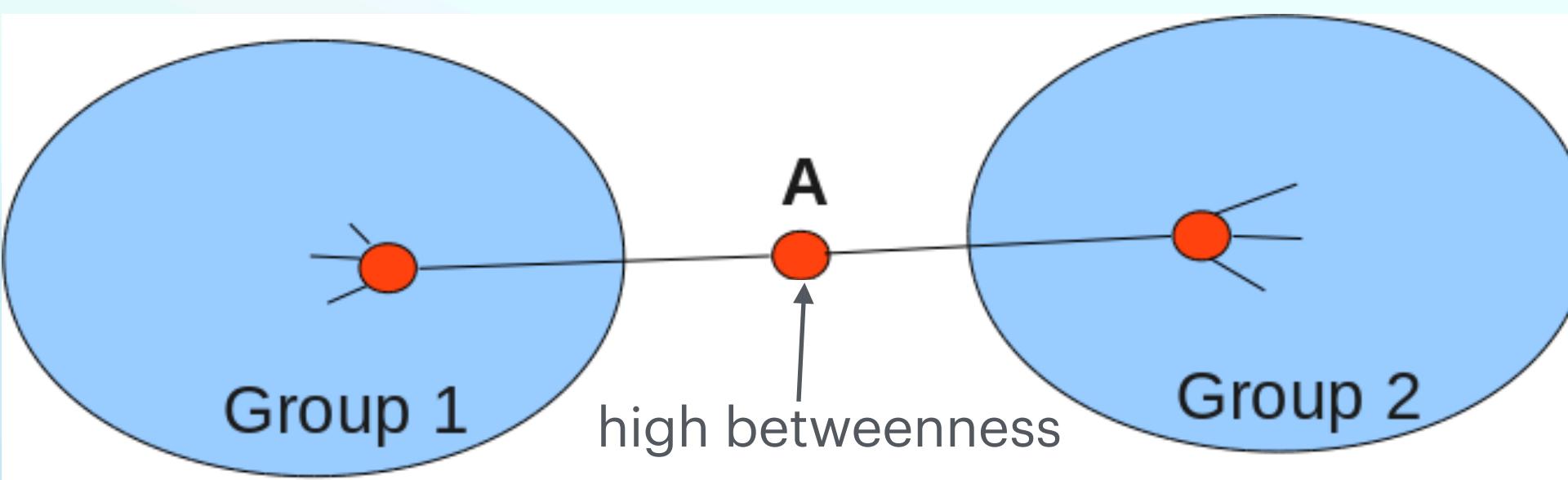


# Network - Betweenness

## 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  $\Rightarrow$   $\begin{cases} \text{high betweenness nodes} \\ \text{controlling information flow} \end{cases}$



# Network - Similarity

**common neighbours ( $n_{ij}$ ):**

- nearest neighbours of nodes  $i$  and  $j$  is:

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

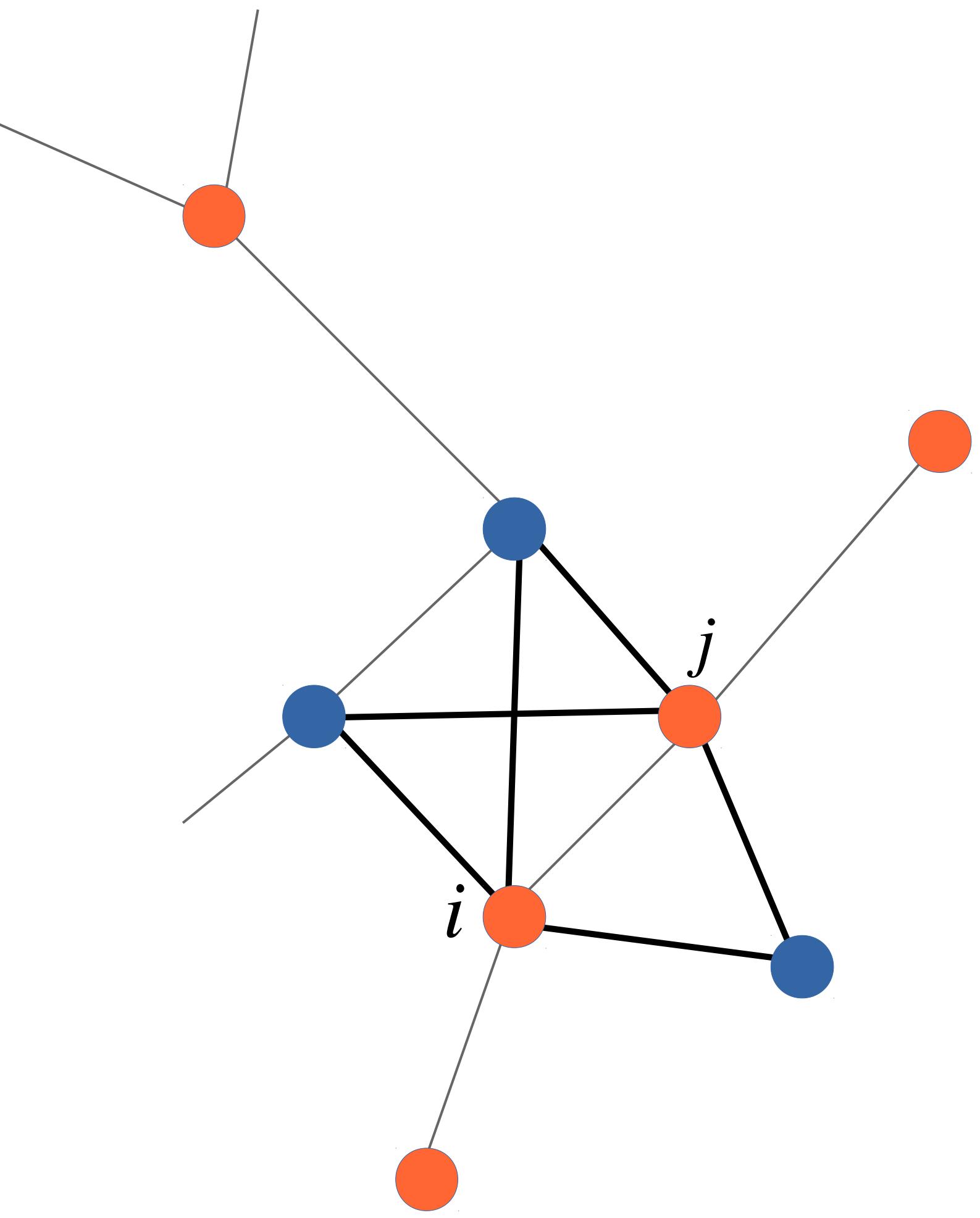
↓      ↓  
each row(column) of  $\mathbf{A}$  is a vector

**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:

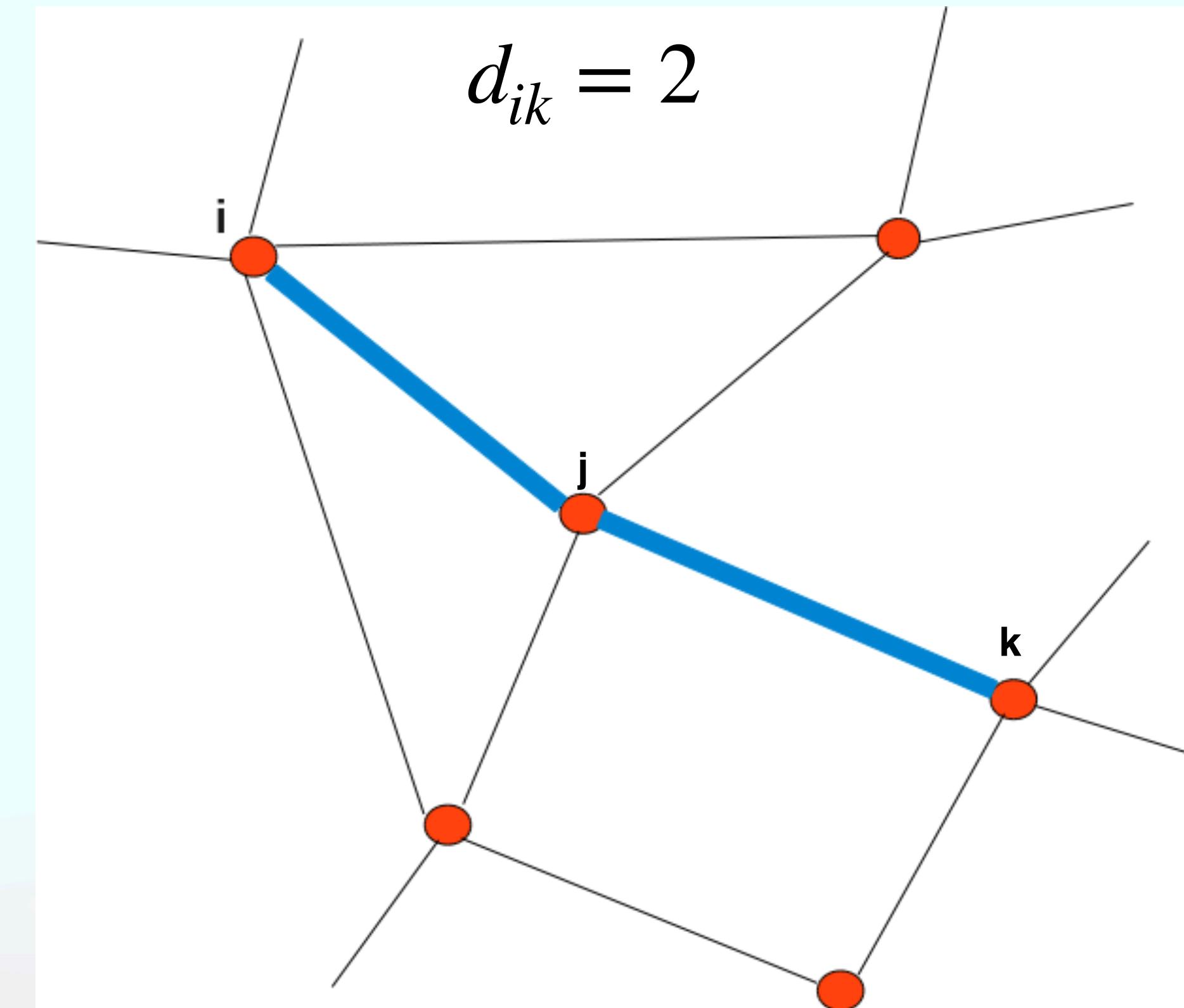
- Transition matrix  $\mathbf{P}$ , where probability going from  $i$  to  $j$  is:  $P_{ij} = \frac{1}{k_i}$
- the steady state, i.e. after making many random walks:

$$\begin{array}{c} \mathbf{P}\vec{\pi} = \vec{\pi} \\ \mathbf{L}[\mathbf{D}^{-1}\vec{\pi}] = 0 \\ \uparrow \\ \mathbf{D}^{-1}\vec{\pi} = a\vec{1} \\ \therefore \pi_i = \frac{k_i}{2m} \end{array}$$

Initial prob. dist. of  
nodes in network

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

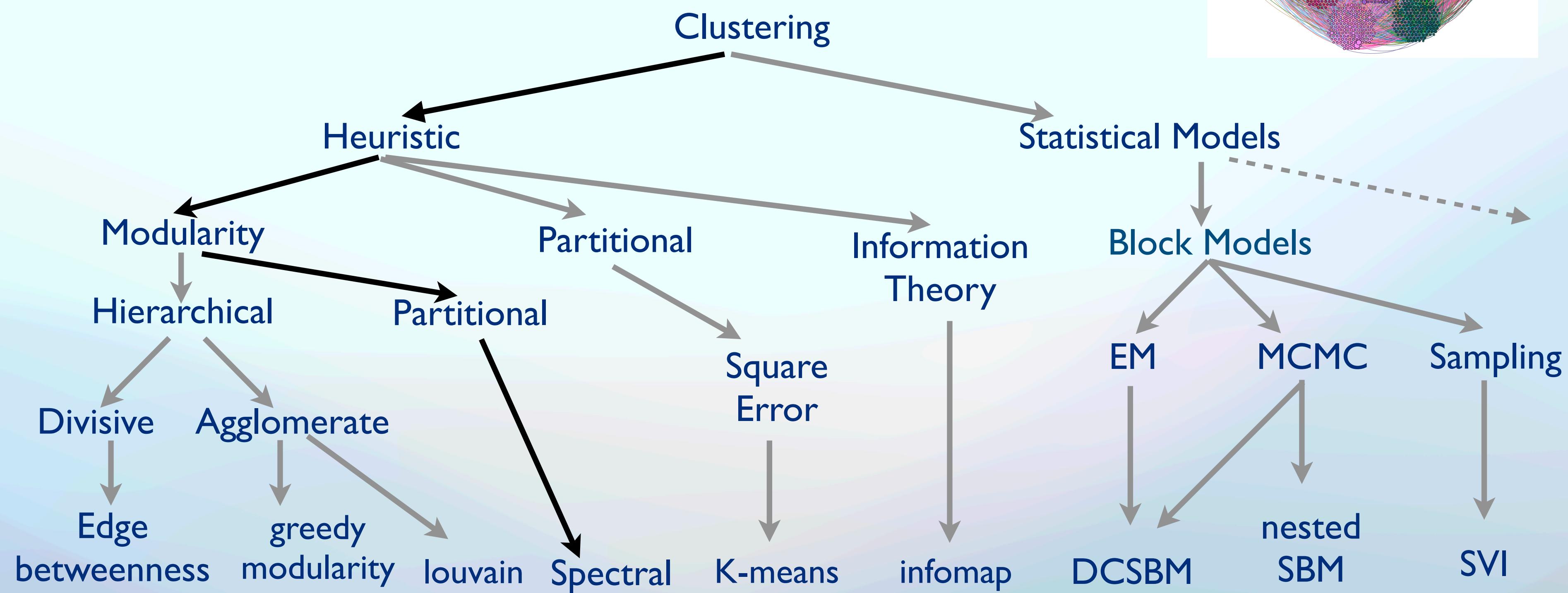
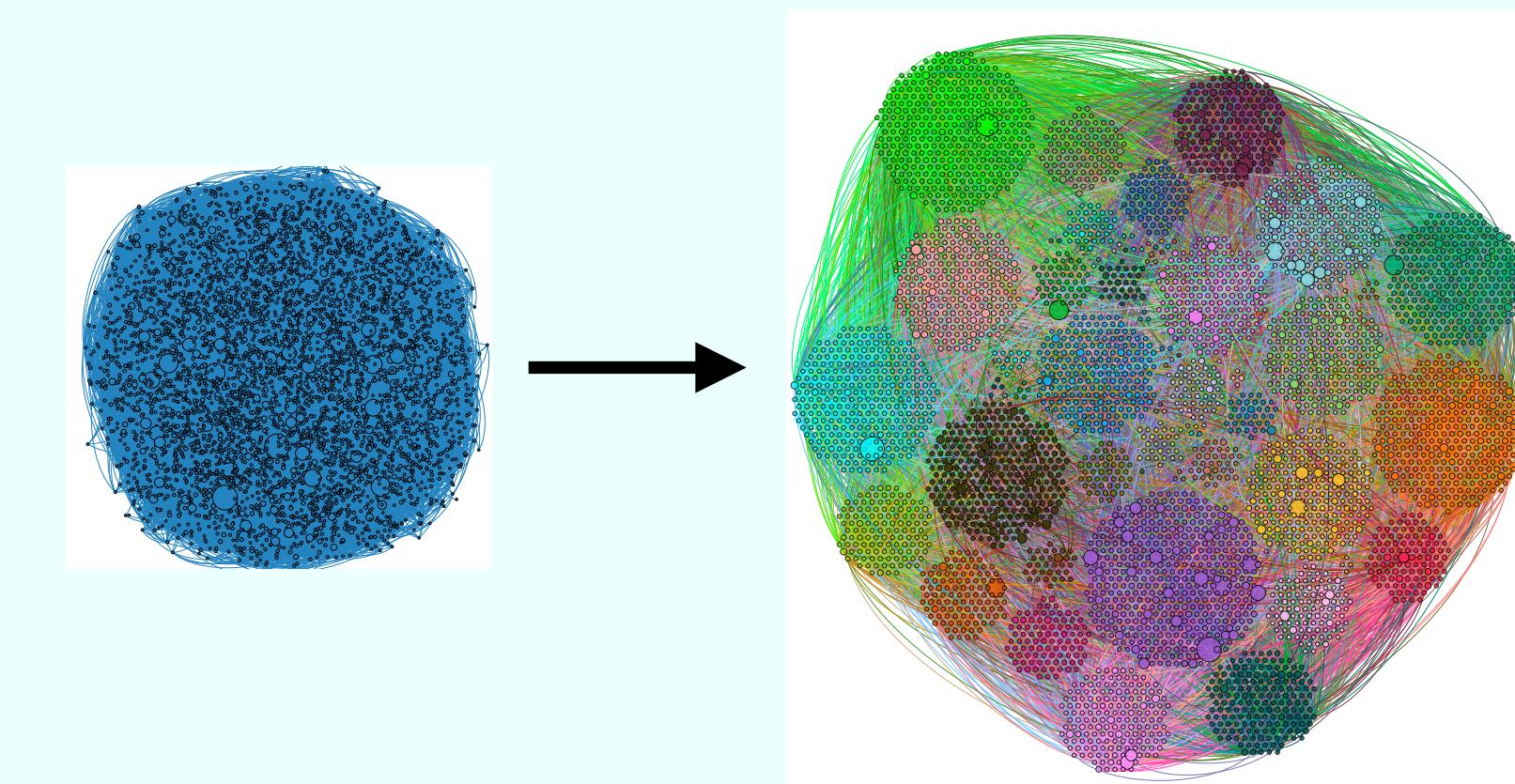
$$p(i \rightarrow j) = \frac{1}{k_i} \frac{k_i}{2m}$$



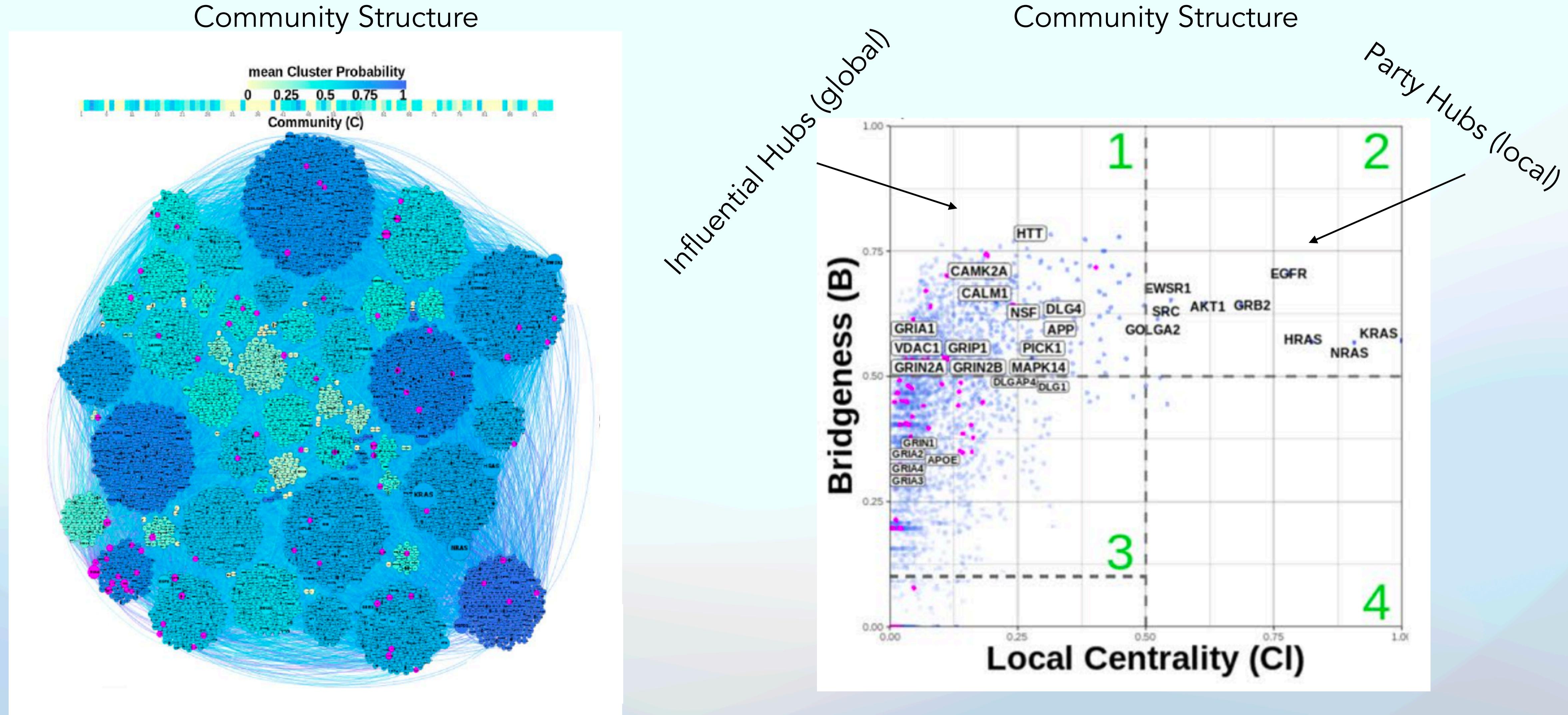
# Network - Clustering

What is a Community?

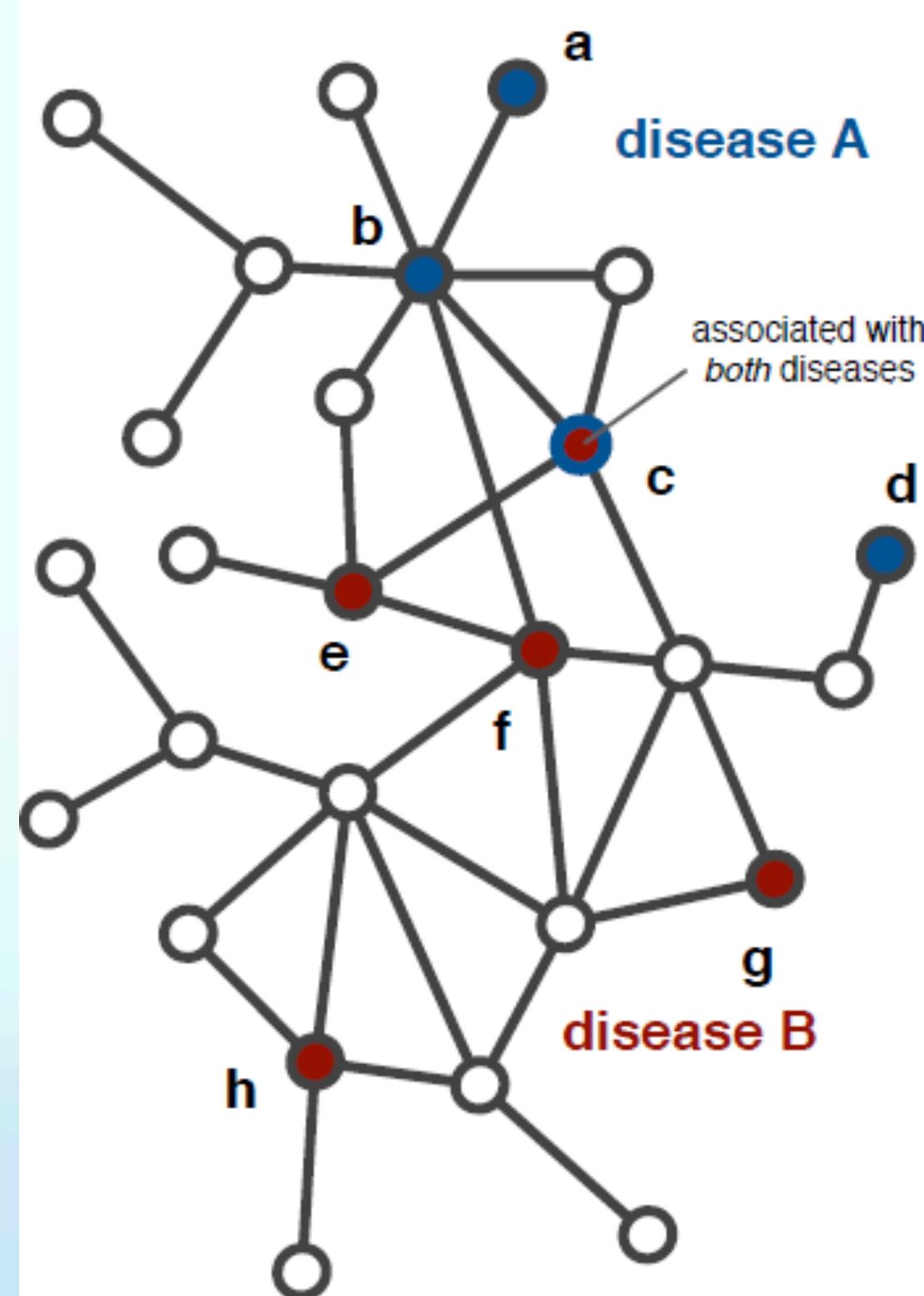
- locally dense regions of nodes in our network that are similar
- modules, complexes, groups, blocks.



# Community Detection in the Synaptic Interactome



# 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 disease's GDA's at random compute  $S_{AB}^{rand}$ .
  - repeat 10,000.

# Disease co-occurrence

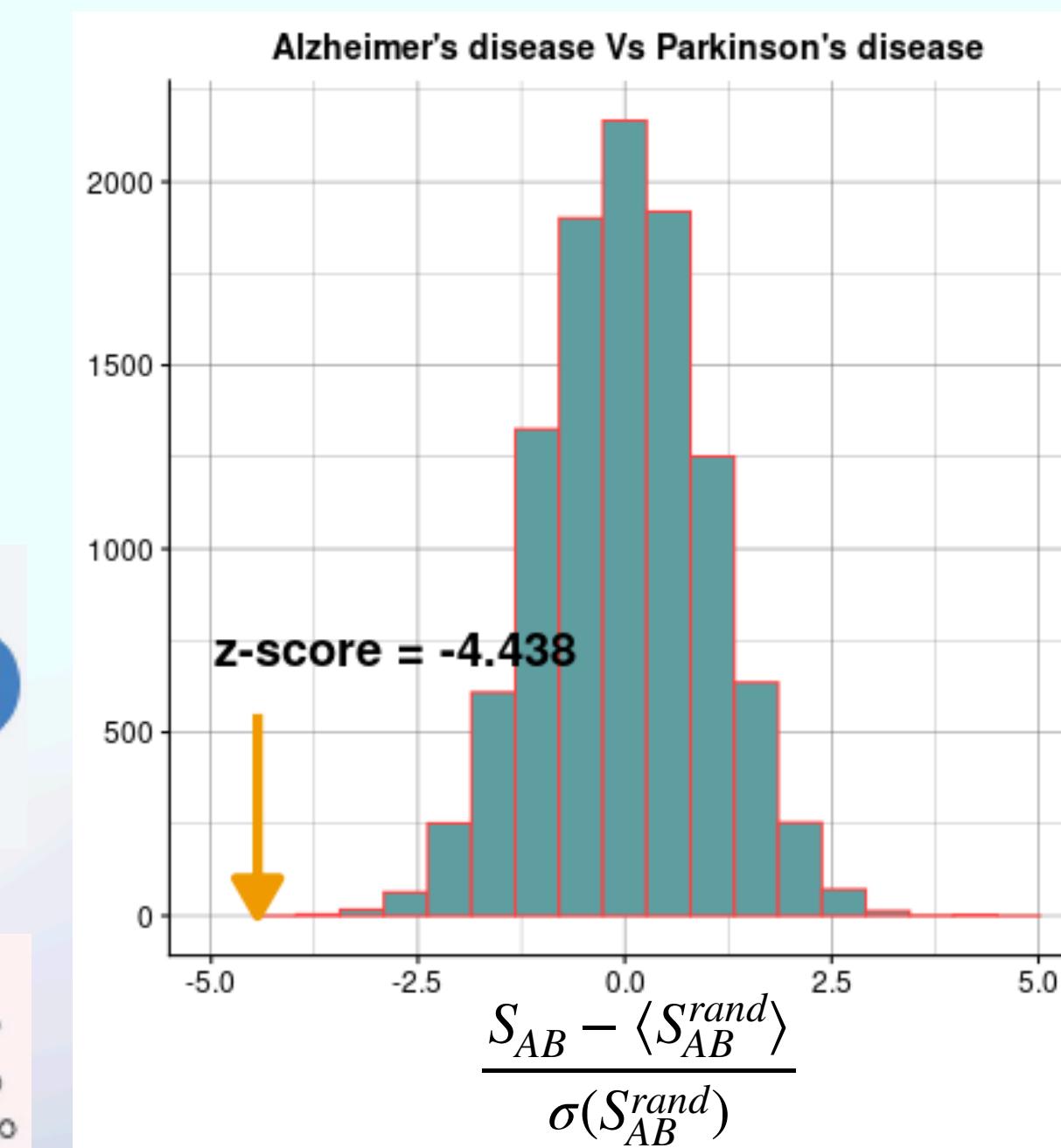
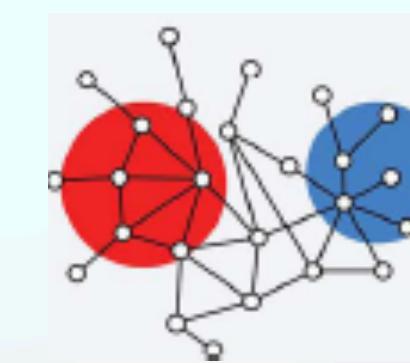
$$S_{AB} = \langle d_{AB} \rangle - \frac{\langle d_{AA} + d_{BB} \rangle}{2}$$

$$S_{AB} \geq 0$$

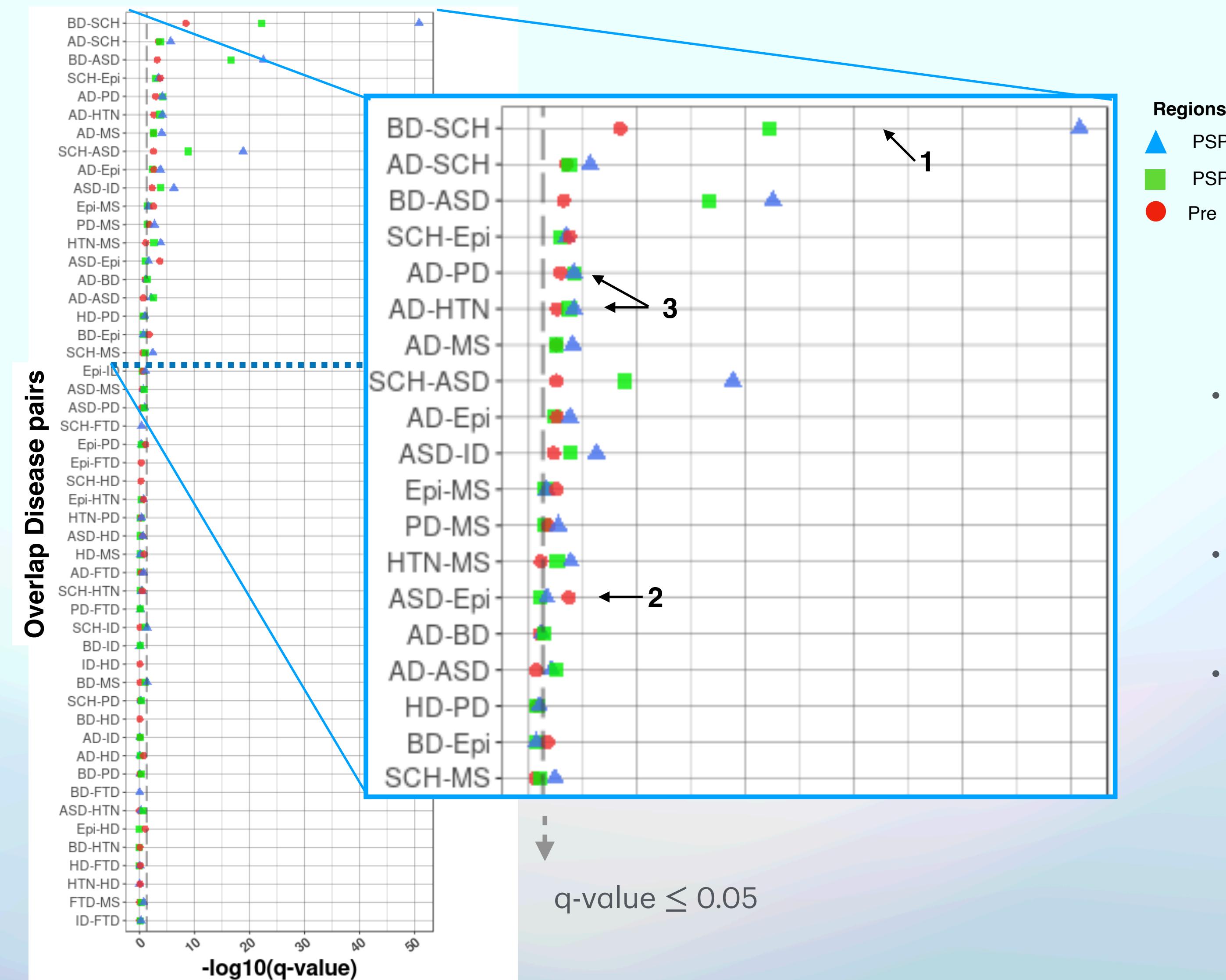
## **Separated disease pair**

$$S_{AB} < 0$$

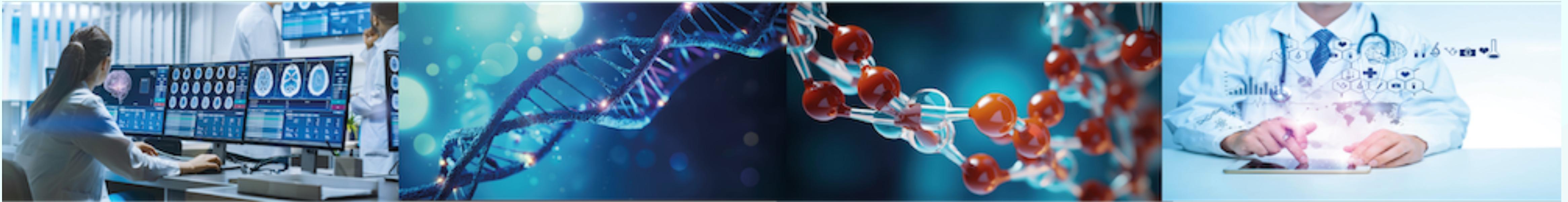
## Overlap disease pain



# 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 EdStem Discussion Board**

**Coding**

<https://github.com/tisimpson/pbi>