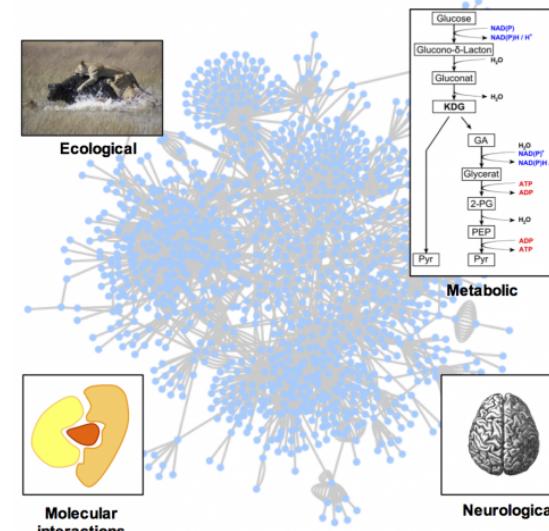




Networks can be used to model many types of biological data



## TODAYS MENU:

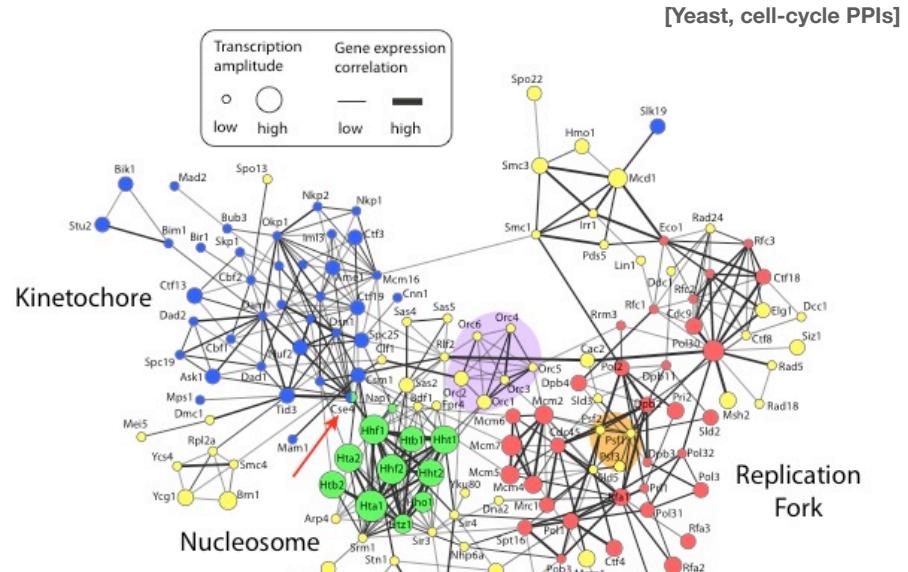
- ▶ Network introduction
- ▶ Network visualization
- ▶ Network analysis
- ▶ Hands-on:  
Cytoscape and R (igraph) software tools  
for network visualization and analysis

## TODAYS MENU:

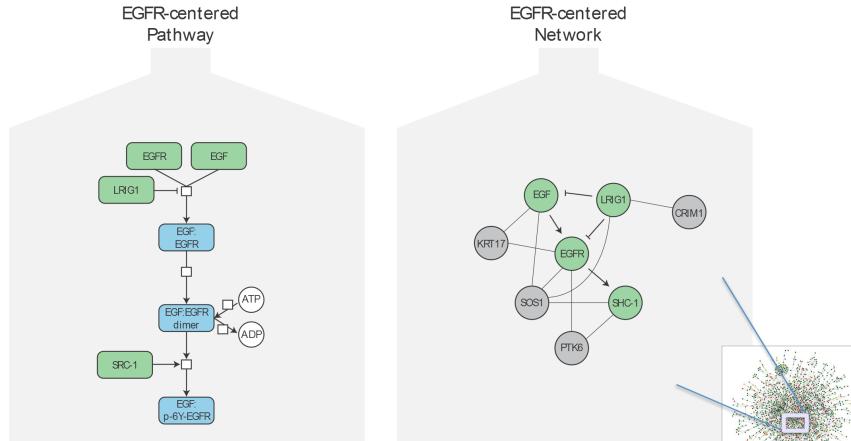
- ▶ Network introduction
- ▶ Network visualization
- ▶ Network analysis
- ▶ Hands-on:  
Cytoscape and R (igraph) software tools  
for network visualization and analysis

# Biological Networks

- Represent biological interactions
  - Physical, regulatory, genetic, functional, etc.
- Useful for discovering relationships in big data
  - Better than tables in Excel
- Visualize multiple heterogeneous data types together
  - Help highlight and see interesting patterns
- Network analysis
  - Well established quantitative metrics from graph theory

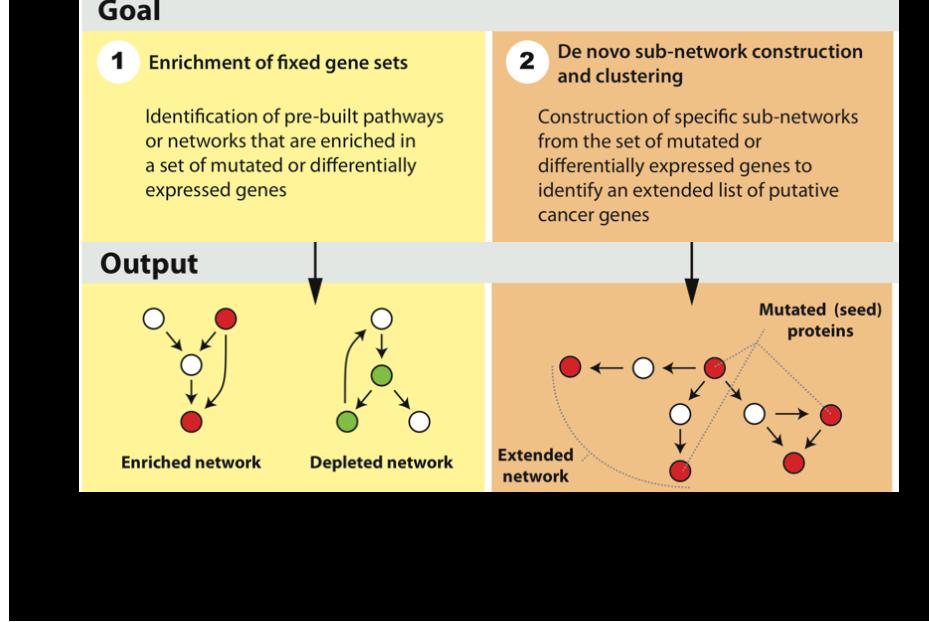


## Pathways vs Networks



- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration



## Goal

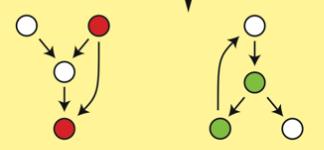
### 1 Enrichment of fixed gene sets

Identification of pre-built pathways or networks that are enriched in a set of mutated or differentially expressed genes

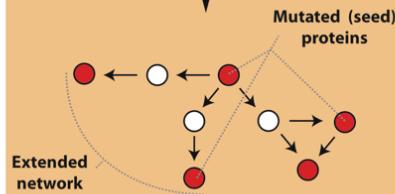
### 2 De novo sub-network construction and clustering

Construction of specific sub-networks from the set of mutated or differentially expressed genes to identify an extended list of putative cancer genes

## Output



What biological process is altered in this cancer?



Are NEW pathways altered in this cancer? Are there clinically relevant tumor subtypes?

**Network analysis is complementary to pathway analysis and can be used to show how key components of different pathways interact.**

This can be useful for identifying regulatory events that influence multiple biological processes and pathways

## Network analysis approaches

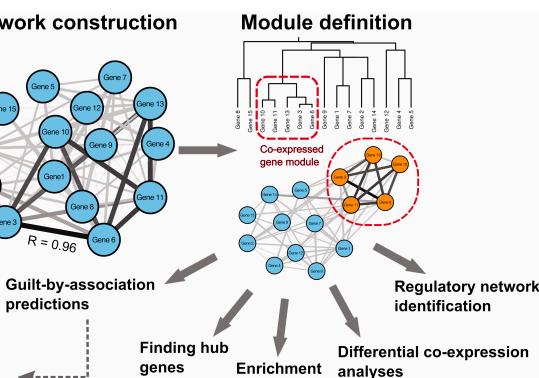
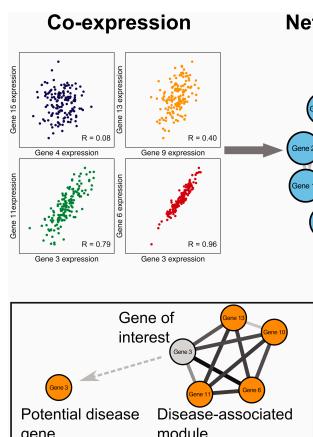
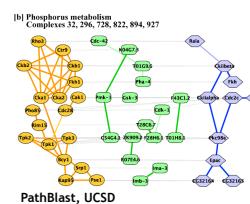
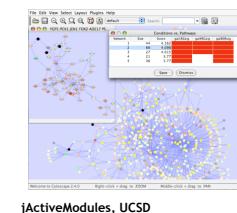


Image from: van Dam et al. (2017) <https://doi.org/10.1093/bib/bbw139>

## Applications of Network Biology

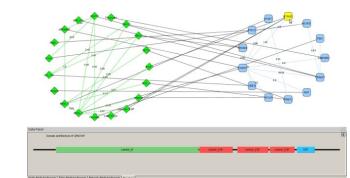
- Gene Function Prediction** – shows connections to sets of genes/proteins involved in same biological process



- Detection of protein complexes/other modular structures** – discover modularity & higher order organization (motifs, feedback loops)



- Network evolution** – biological process(es) conservation across species



- Prediction of new interactions and functional associations** – Statistically significant domain-domain correlations in protein interaction network to predict protein-protein or genetic interaction; allostery in molecular networks

Slide from: [humangenetics-amc.nl](http://humangenetics-amc.nl)

# What's missing

- **Dynamics**
  - Pathways/networks represented as static processes
  - Difficult to represent a calcium wave or a feedback loop
  - More detailed mathematical representations exist that handle these e.g. Stoichiometric modeling, Kinetic modeling (VirtualCell, E-cell, ...)
- **Detail** – atomic structures & exclusivity of interactions.
- **Context** – cell type, developmental stage

# What have we learned so far...

- **Networks are useful for seeing relationships in large data sets**
  - Important to understand what the nodes and edges mean
  - Important to define the biological question - know what you want to do with your gene list or network
- **Many methods available for network analysis**
  - Good to determine your question and search for a solution
  - Or get to know many methods and see how they can be applied to your data

## TODAYS MENU:

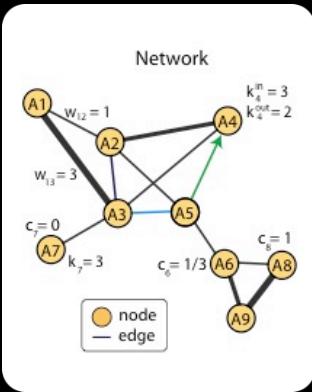
- Network introduction
- **Network visualization**
- Network analysis
- **Hands-on:**  
Cytoscape and R (igraph) software tools  
for network visualization and analysis

## Network Visualization Outline

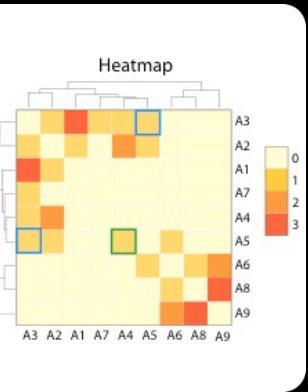
- Network representations
- Automatic network layout
- Visual features
- Visually interpreting a network

## Network representations

Relationships	Optional weight
A1 ↔ A2	1
A1 ↔ A3	3
A2 ↔ A3	1
A2 ↔ A4	2
A2 ↔ A5	1
A3 ↔ A4	1
A3 ↔ A5	1
A3 ↔ A7	1
A5 ↔ A4	1
A5 ↔ A6	1
A6 ↔ A8	1
A6 ↔ A9	2
A8 ↔ A9	3



1  
List of relationships

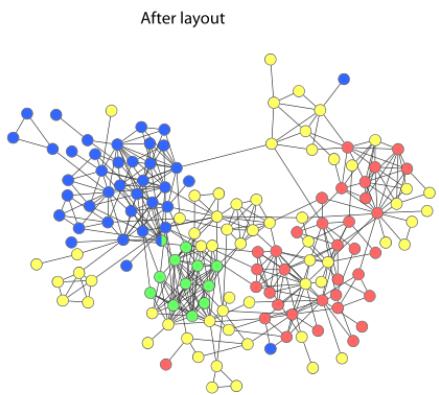
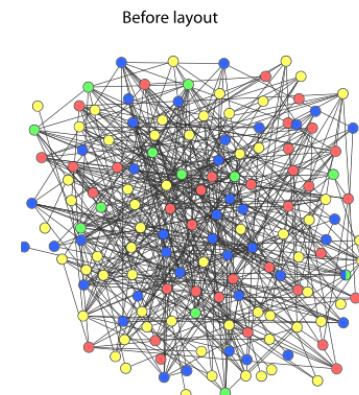


2  
Network view

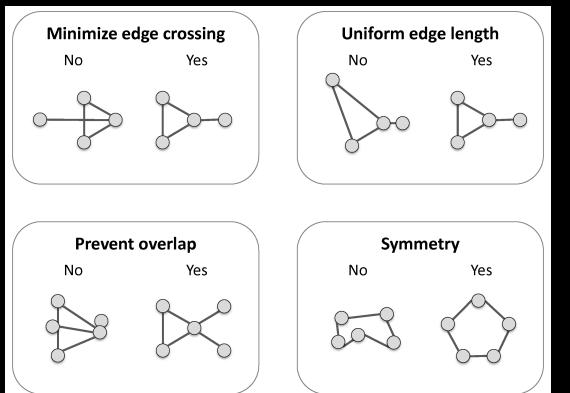
3  
Adjacency matrix view

Network view is most useful when network is sparse!

## Automatic network layout



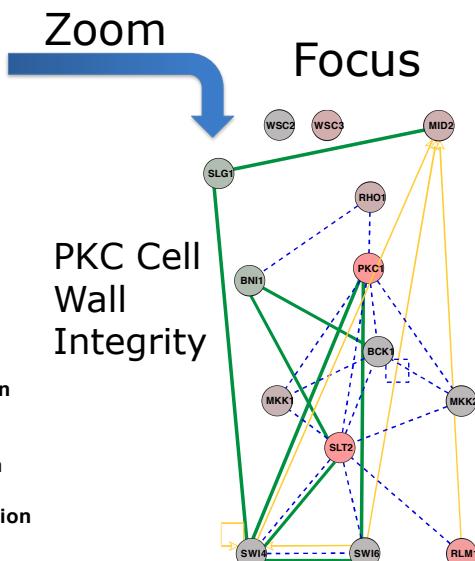
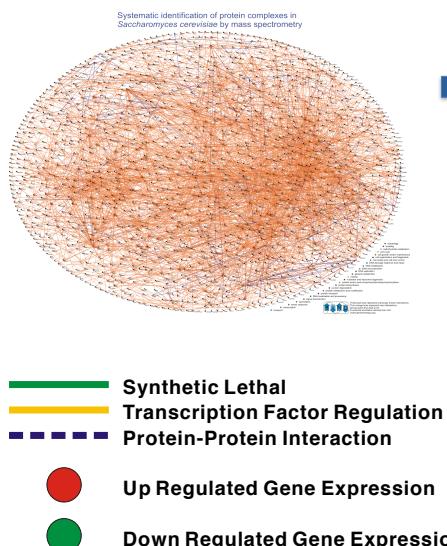
- Modern **graph layouts** are optimized for speed and aesthetics. In particular, they seek to minimize overlaps and edge crossing, and ensure similar edge length across the graph.



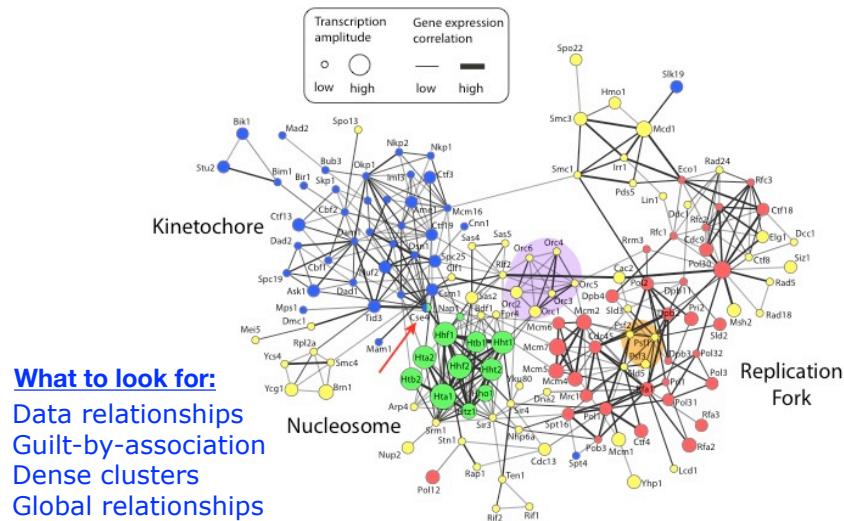
## Force-directed layout: Nodes repel and edges pull

- Good for up to 500 nodes
  - Bigger networks give hairballs
  - Reduce number of edges
  - Or just use a heatmap for dense networks
- Advice: try force directed first, or hierarchical for tree-like networks
- Tips for better looking networks
  - Manually adjust layout
  - Load network into a drawing program (e.g. Illustrator) and adjust labels

## Dealing with 'hairballs': zoom or filter

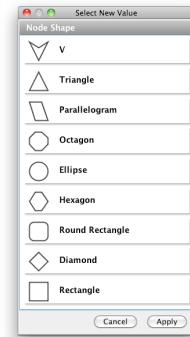
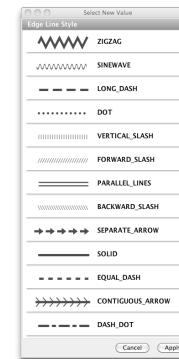
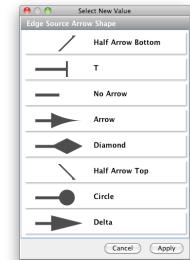


## Visually Interpreting a Network



## Visual Features

- Node and edge attributes
  - Text (string), integer, float, Boolean, list
  - E.g. represent gene, interaction attributes
- Visual attributes
  - Node, edge visual properties
  - Color, shape, size, borders, opacity...



What have we learned so far...

- Automatic layout is required to visualize networks
- Networks help you visualize interesting relationships in your data
- Avoid hairballs by focusing analysis
- Visual attributes enable multiple types of data to be shown at once – useful to see their relationships

## TODAYS MENU:

### › Network introduction

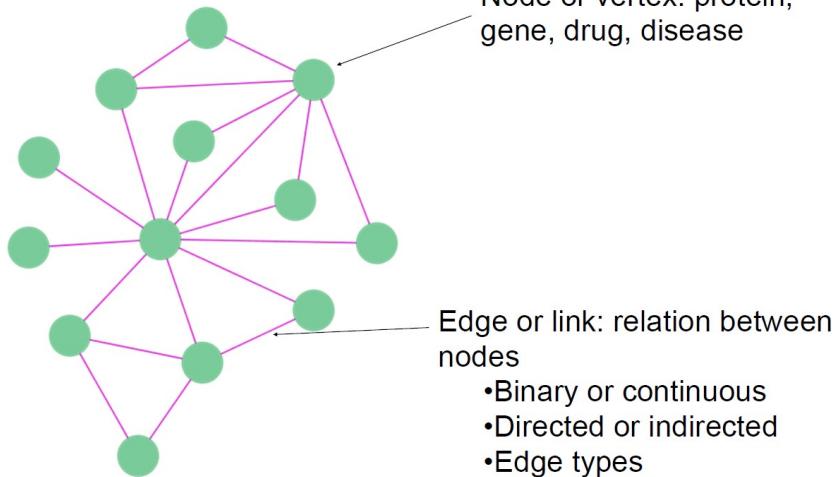
### › Network visualization

### › Network analysis

### › Hands-on:

- Cytoscape and R (igraph) software tools for network visualization and analysis

Network or graph



## Introduction to graph theory

- Biological network analysis historically originated from the tools and concepts of **social network analysis** and the application of **graph theory** to the social sciences.
- Wikipedia defines graph theory as:
  - "[...] the study of graphs used to model pairwise relations between objects. A graph in this context is made up of **vertices** connected by **edges**".
- In practical terms, it is the set of concepts and methods that can be used to visualize and analyze networks

## Types of network edges

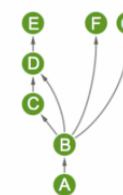
Undirected



Connection,  
without a given  
'flow' implied

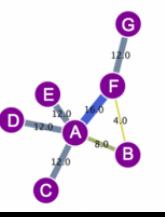
(e.g. protein A  
binds protein B)

Directed



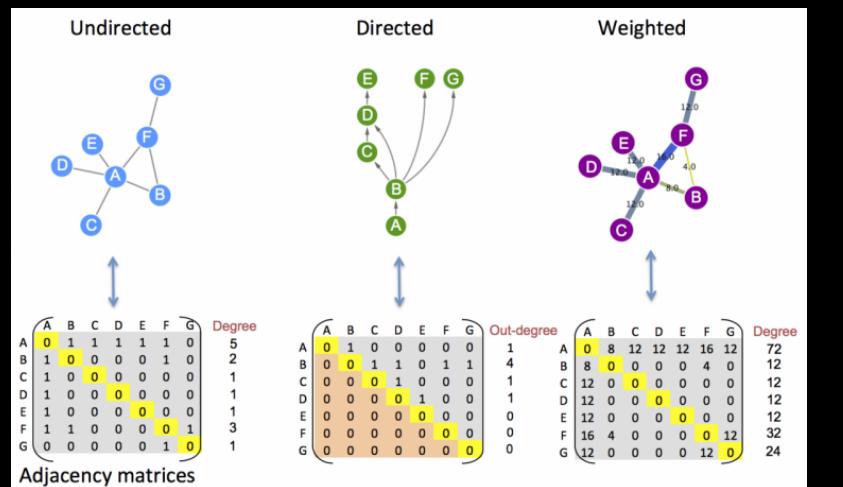
There is directional  
flow/signal implied  
(e.g. metabolic or  
gene networks)

Weighted



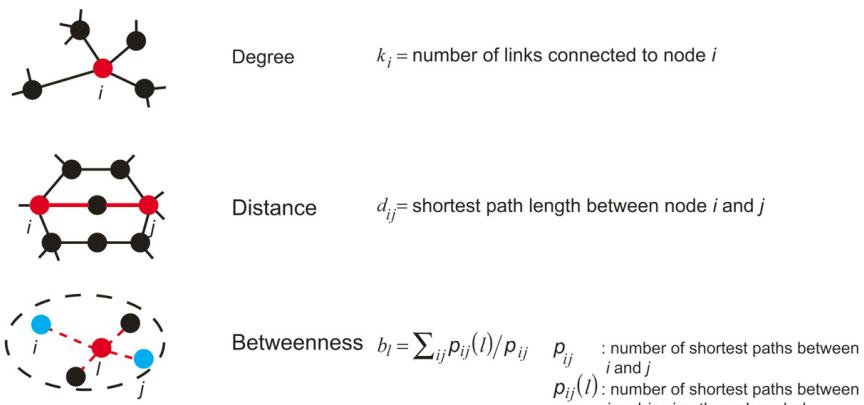
Edges can also  
have weight  
(i.e. a 'strength' of  
interaction).

- Every network can be expressed mathematically in the form of an adjacency matrix

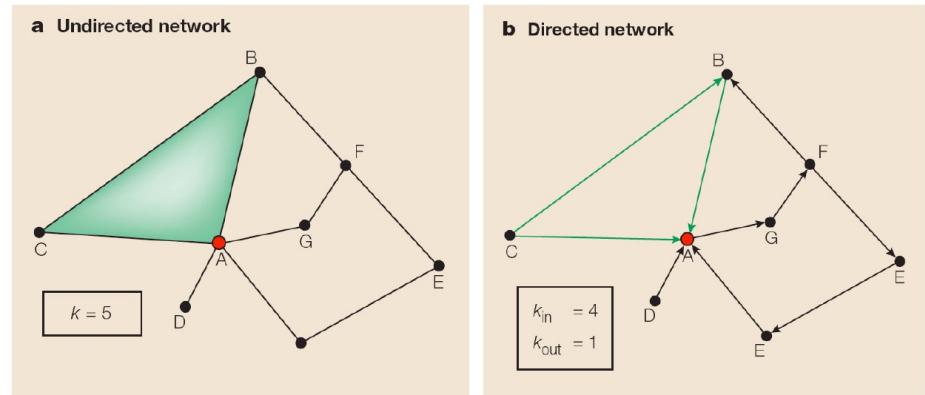


# Network topology

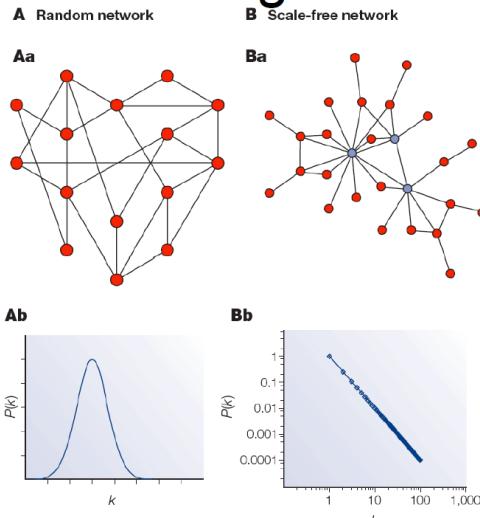
- Topology is the way in which the nodes and edges are arranged within a network.
- The most used topological properties and concepts include:
  - Degree** (i.e. how many node neighbors)
  - Communities** (i.e. clusters of well connected nodes)
  - Shortest Paths** (i.e. shortest distance between 2 nodes)
  - Centralities** (i.e. how ‘central’ is a given node?)
  - Betweenness** (a measure of centrality based on shortest paths)



## Network Measures: Degree



# Degree Distribution



$P(k)$  is probability of each degree  $k$ , i.e fraction of nodes having that degree.

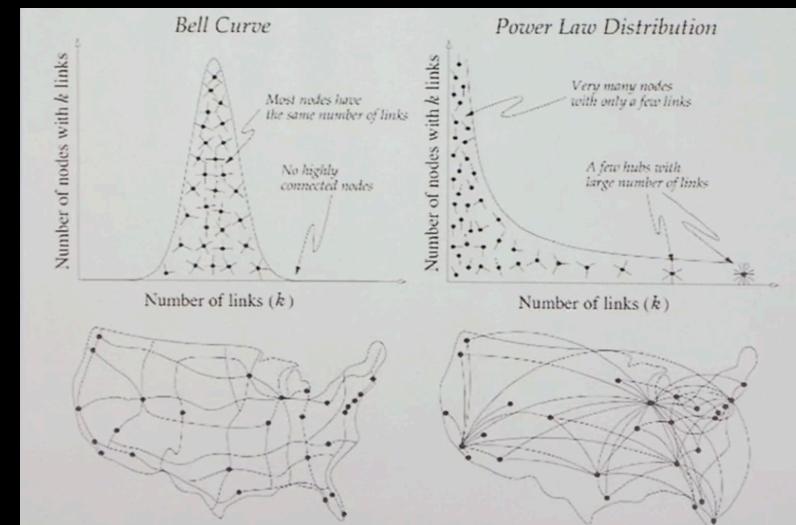
For random networks,  $P(k)$  is normally distributed.

For real networks the distribution is often a power-law:

$$P(k) \sim k^{-\gamma}$$

Such networks are said to be **scale-free**

## Random graphs vs scale free



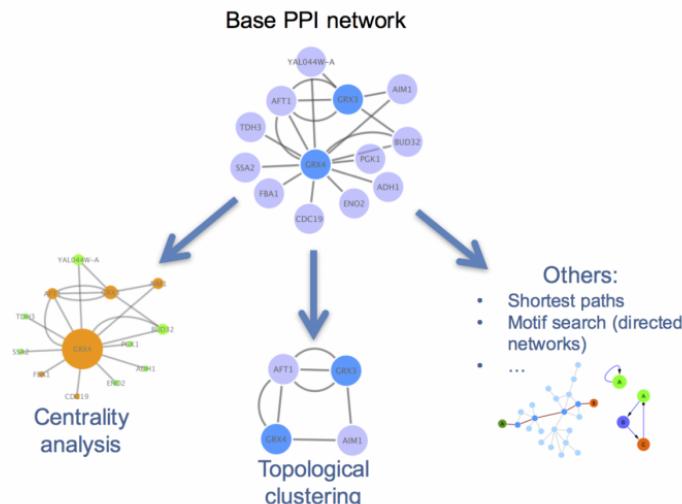
## Scale-Free Networks are Robust

- Complex systems (cell, internet, social networks), are resilient to component failure
- Network topology plays an important role in this robustness
  - Even if ~80% of nodes fail, the remaining ~20% still maintain network connectivity
- Attack vulnerability* if hubs are selectively targeted
- In yeast, only ~20% of proteins are lethal when deleted, and are 5 times more likely to have degree  $k>15$  than  $k<5$ .

## Implications

- Many biological networks (protein-protein interaction networks regulatory networks, etc...) are thought to have hubs, or nodes with high degree.
- For protein-protein interaction networks (PPIs) these hubs have been shown to be older [1] and more essential than random proteins [2]
  - [1] Fraser et al. *Science* (2002) 296:750
  - [2] Jeoung et al. *Nature* (2001) 411:41

Analyzing the topological features of a network is a useful way of identifying relevant participants and substructures that may be of biological significance.



Bigger, redder nodes have higher **centrality values** in this representation.

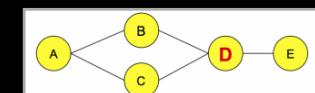


## Centrality analysis

- Centrality gives an estimation on how important a node or edge is for the connectivity or the information flow of the network
- It is a useful parameter in signalling networks and it is often used when trying to find drug targets.
- Centrality analysis in PPINs usually aims to answer the following question:
  - ➡ Which protein is the most important and why?

## Betweenness centrality

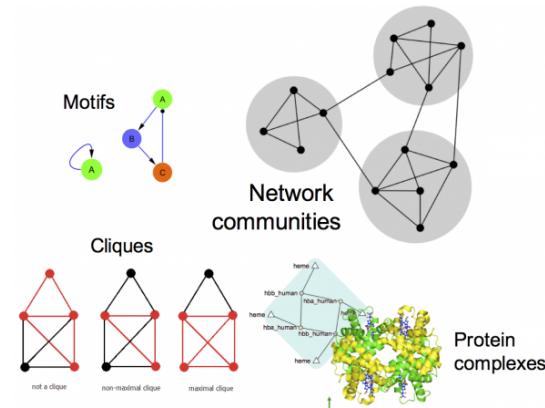
- Nodes with a high betweenness centrality are interesting because they lie on communication paths and can control information flow.
- The number of shortest paths in the graph that pass through the node divided by the total number of shortest paths.
- Betweenness centrality measures how often a node occurs on all shortest paths between two nodes.



# Community analysis

- **Community:** A general, catch-all term that can be defined as a group (i.e. *cluster*) of nodes that are more connected within themselves than with the rest of the network. The precise definition for a community will depend on the method or algorithm used to define it.

Looking for communities in a network is a nice strategy for reducing network complexity and extracting functional modules (e.g. protein complexes) that reflect the biology of the network.



## TODAYS MENU:

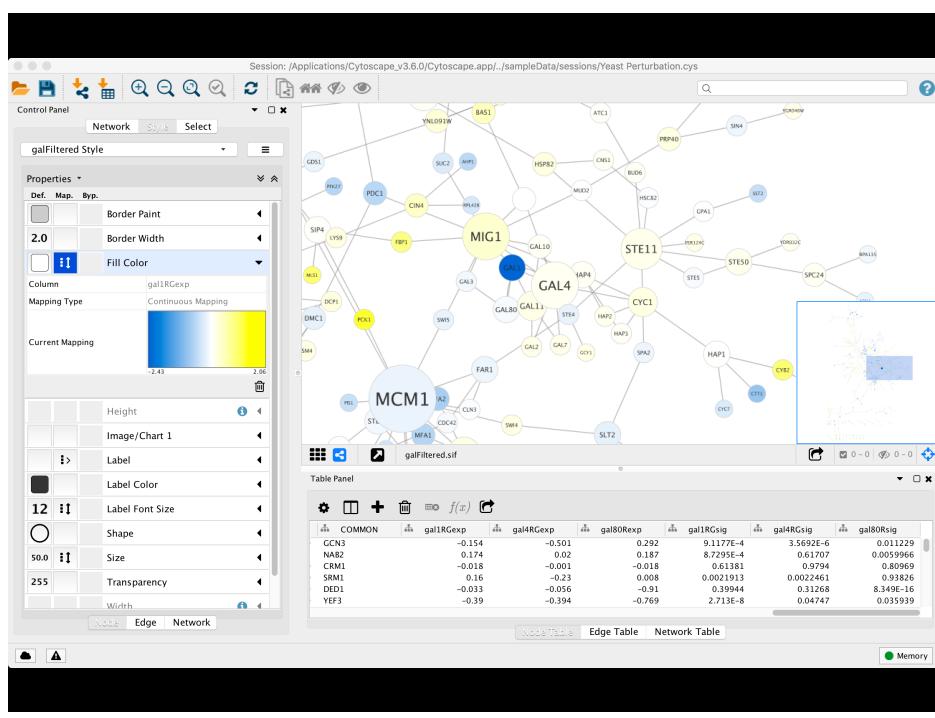
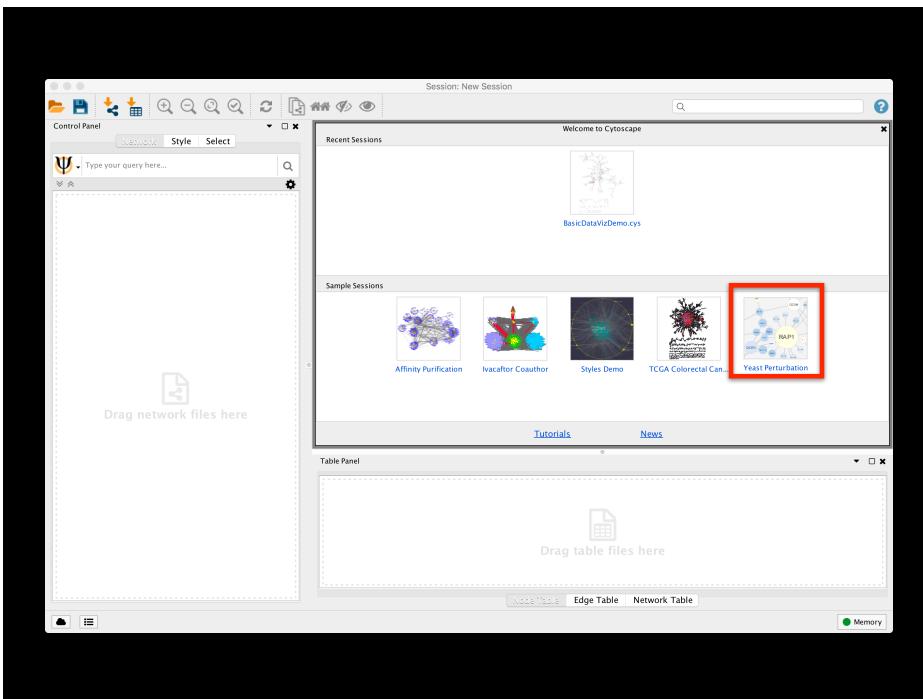
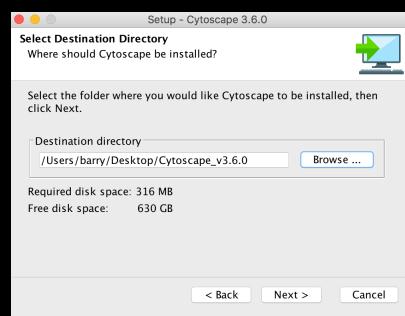
- **Network introduction**
- **Network visualization**
- **Network analysis**
- **Hands-on:**
  - Cytoscape and R (igraph) software tools for network visualization and analysis

## Practical issues

- Major tools for the **creation, manipulation** and **visualization** of biological networks include:
  - Cytoscape,
  - Gephi
  - R packages (igraph, graph, tidygraph, ggraph)
- Tools for network analysis and modeling include:
  - Cytoscape apps/plugins
  - R packages (igraph and many others)
  - NetworkX (for Python)
  - ByoDyn, COPASI

<http://cytoscape.org/download.php>

**Note:** If you are on a classroom Mac please check if Cytoscape is already installed. If not then please be sure to install to your **Desktop** directory!



## Cytoscape Memory Issues

- Cytoscape uses lots of memory and doesn't like to let go of it
  - An occasional restart when working with large networks is a good thing
  - Destroy views when you don't need them
- Since version 2.7, Cytoscape does a much better job at "guessing" good default memory sizes than previous versions but it still not great!
  - Java doesn't give us a good way to get the memory right at start time

# Cytoscape Sessions

- Sessions save pretty much everything:
  - Networks
  - Properties
  - Visual styles
  - Screen sizes
- Saving a session on a large screen may require some resizing when opened on your laptop

# Hands-on: Part 1

[https://bioboot.github.io/bggn213\\_S19/lectures/#17](https://bioboot.github.io/bggn213_S19/lectures/#17)

- The data used in **part 1** is from yeast, and the genes Gal1, Gal4, and Gal80 are all yeast transcription factors. The experiments all involve some perturbation of these transcription factor genes.
- In this network view, the following node attributes have been mapped to visual style properties in cytoscape:
  - The "gal80exp" expression values are used for Node Fill Color.
  - The Default Node Color, for nodes with no data mapping, is dark grey.
  - Nodes with expression values that are significant are rendered as rectangles, others are ovals.
  - The common name for each gene is used as the Node Label.

# Hands-on: Part 2

[https://bioboot.github.io/bggn213\\_W19/lectures/#17](https://bioboot.github.io/bggn213_W19/lectures/#17)

- The data used in **part 2** is from an ocean metagenomic sequencing project - where all the genetic material in a sample of ocean water is sequenced.
- We will use the R package **igraph** and the bioconductor package **RCy3** together with Cytoscape.
- Many of these microbial species in these types of studies have not yet been characterized in the lab.
  - Thus, to know more about the organisms and their interactions, we can observe which ones occur at the same sites.
  - One way to do that is by using **co-occurrence networks** where you examine which organisms occur together at which sites.

# Network representations

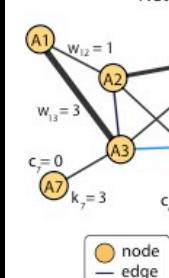
## Relationships

	Optional weight
A1 $\leftrightarrow$ A2	1
A1 $\leftrightarrow$ A3	3
A2 $\leftrightarrow$ A3	1
A2 $\leftrightarrow$ A4	2
A2 $\leftrightarrow$ A5	1
A3 $\leftrightarrow$ A4	1
A3 $\leftrightarrow$ A5	1
A3 $\leftrightarrow$ A7	1
A5 $\rightarrow$ A4	1
A5 $\leftrightarrow$ A6	1
A6 $\leftrightarrow$ A8	1
A6 $\leftrightarrow$ A9	2
A8 $\leftrightarrow$ A9	3

1

List of relationships

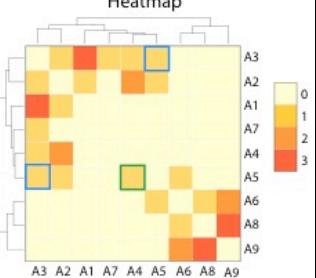
## Network



2

Network view

## Heatmap



3

Adjacency matrix view

Network view is most useful when network is sparse!

# Summary

- Network biology makes use of the tools provided by **graph theory** to represent and analyze complex biological systems.
- Major types of biological networks include: genetic, metabolic, cell signaling etc.
- Networks are represented by **nodes** and **edges**.
- Biological networks have a number of characteristics, mainly:
  - **Scale-free:** A small number of nodes (hubs) are a lot more connected than the average node.
  - **Transitivity:** The networks contain communities of nodes that are more connected internally than they are to the rest of the network.
- Major tools for network analysis include: **Cytoscape**, **igraph**, Gephi and NetworkX.
- Two of the most used topological methods to analyze PPINs are:
  - **Centrality analysis:** Which identifies the most important nodes in a network, using different ways to calculate centrality.
  - **Community detection:** Which aims to find heavily inter-connected components that may represent protein complexes and machineries

# Summary cont...

- **Cytoscape** is a useful, free software tool for network visualization and analysis
  - Provides basic network manipulation features
  - Plugins/Apps are available to extend the functionality
- The R **igraph** package has extensive network analysis functionality beyond that in Cytoscape
- The R bioconductor package **RCy3** package allows us to bring networks and associated data from R to Cytoscape so we can have the best of both worlds.

## Network Analysis Overview

