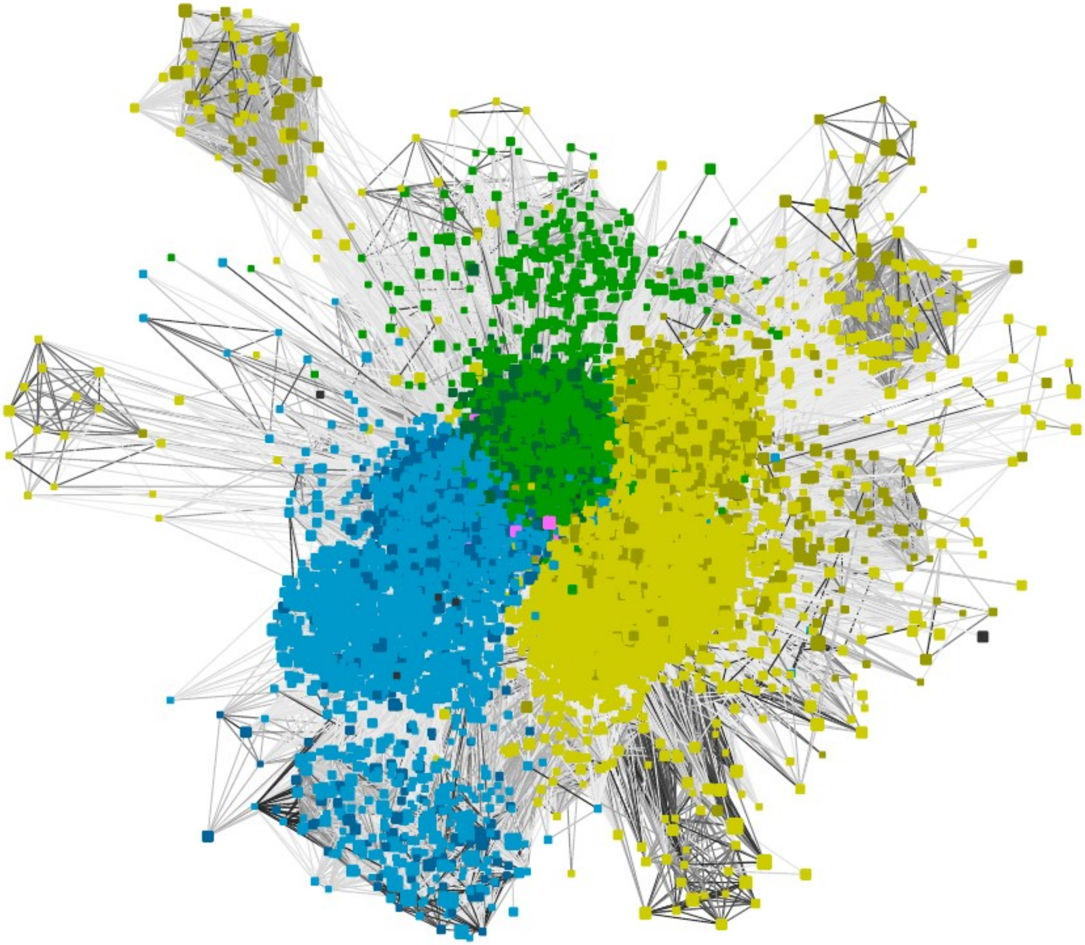
Introduction to Sequence Similarity Networks



Miguel M. Sandin

miguelmendezsandin@gmail.com

2021-05-06

Why networks?

And not yet another phylogenetic tree?

Because we want to answer different questions.

Phylogenies:

What are the evolutionary relationships among taxa?

Based in:
-*Species* selection
-Alignment
-Evolutionary model
-Phylogenetic inference
-Bifurcating phylogenetic tree

SSN:

Where is this protein coming from? How do genomes interact? ...

A→G
=
A→C
=
A→T

Based in:
-*Species* selection
-Similarity search
-Threshold(s) selection
-Network analysis/representation

And what about ecological analysis? Why not yet another ordination analysis?

Because we want to answer different questions.

Ordination:

How do samples correlate?...

Based in:
-Abundance/Presence-Absence

SSN:

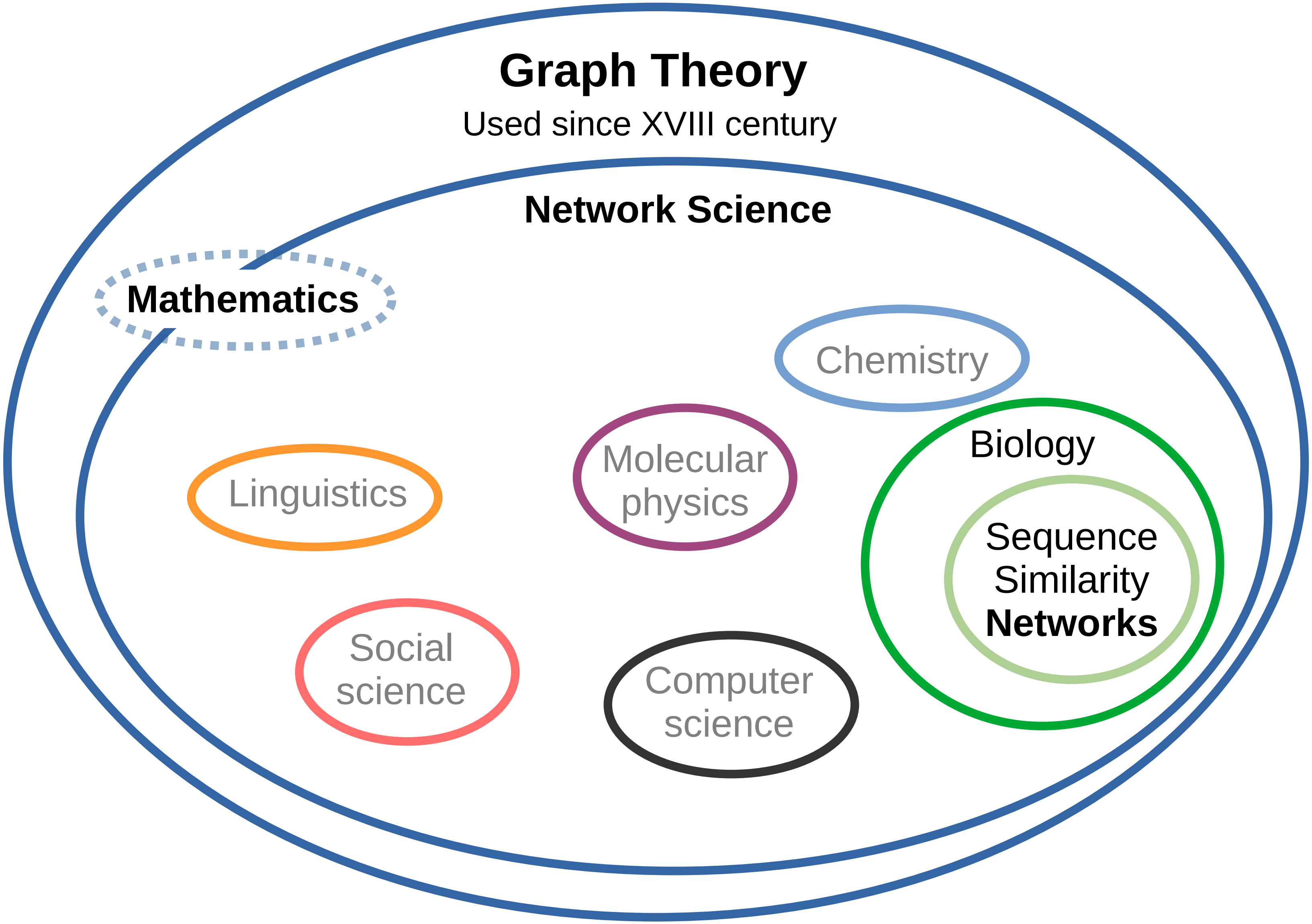
What are the most central sequences?

Based in:
-Sequence similarity

Networks COMPLEMENT previous well-established methods

What is a network?

Before starting!



Wikipedia: Graph Theory

What is a network?

How is it built?

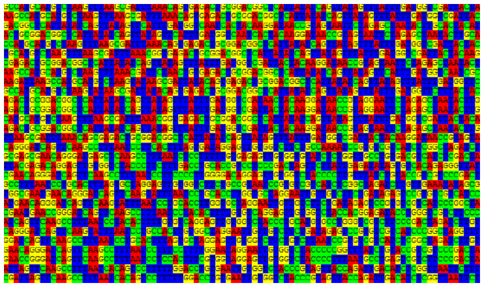
Species selection

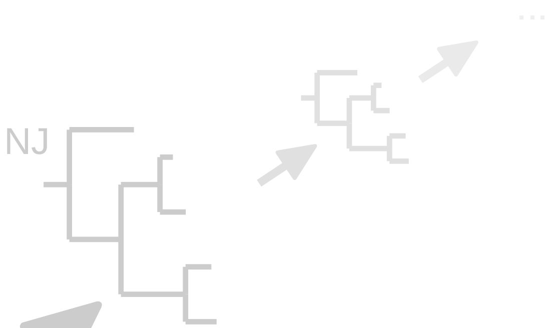
>seq1
ACGATCGATTAC...
>seq2
TGGAGATCATAC...
>seq3
GCAGTCGATTAC...
>seq4
ACGATGCTAGCT...
...

Alignment

TGAAATA
-TAAATA
TAAA-TA
TAAGCTA
TAAGCTA

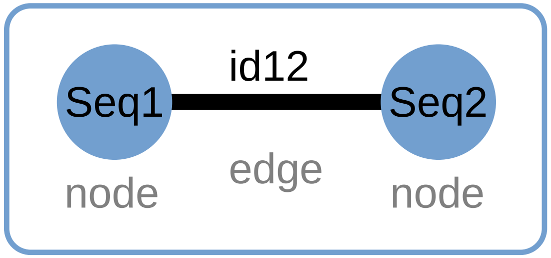
Local alignment





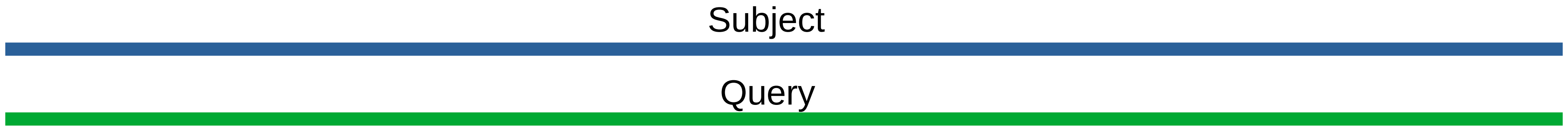
	Seq1	Seq2	Seq3	...	Seqi
Seq1	1	id12	id13	...	id1i
Seq2	id21	1	id21	...	id2i
Seq3	id31	id32	1	...	id3i
...	...	...	...	...	...
Seqi	idi1	idi2	idi3	...	1

Seq1	Seq2	id12
Seq1	Seq3	id13
Seq1	Seq4	id14
...	...	...
Seqi	Seqj	idij



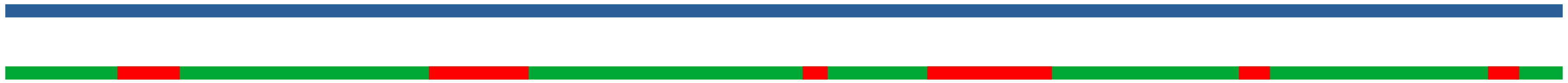
Local Alignment

BLAST

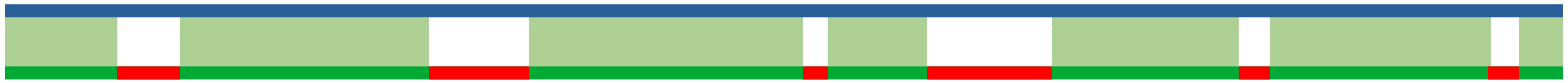


qseqid	**sseqid**	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscare

Identity percentage



Coverage



Expect (E) value

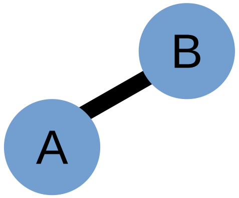
“Significancy” of the hit: How likely
you get the same score by chance.

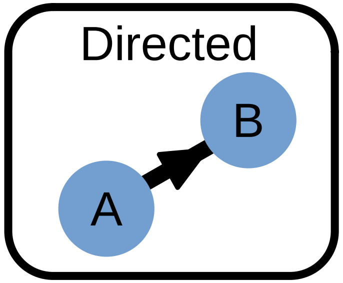
How are networks visualized?

	A	B	C	D	E	F
A	100	91	76	72	92	64
B	91	100	91	96	82	80
C	76	98	100	94	78	84
D	72	96	94	100	62	86
E	94	82	78	62	100	79
F	64	80	84	86	79	100

We choose a threshold: id
Is the **pairwise** similarity above?
If yes, there is a connection

Undirected connection

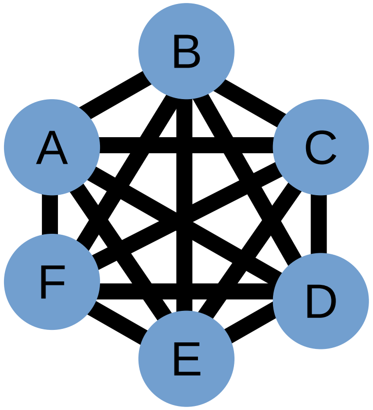




Id: 60

How are networks visualized?

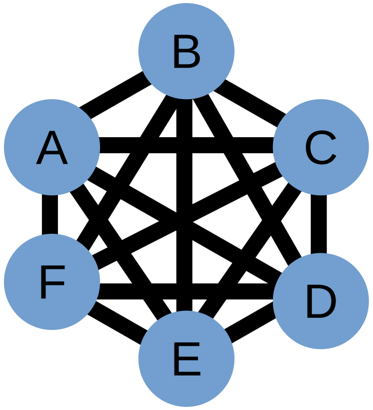
	A	B	C	D	E	F
A	100	91	76	72	92	64
B	91	100	91	96	82	80
C	76	98	100	94	78	84
D	72	96	94	100	62	86
E	94	82	78	62	100	79
F	64	80	84	86	79	100



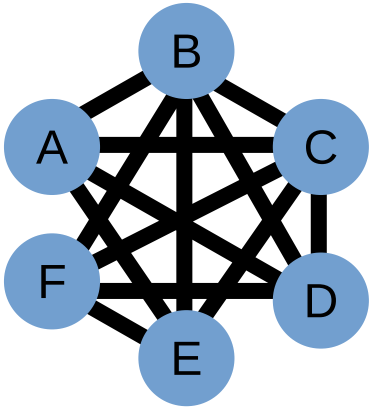
Id: 60

How are networks visualized?

	A	B	C	D	E	F
A	100	91	76	72	92	64
B	91	100	91	96	82	80
C	76	98	100	94	78	84
D	72	96	94	100	62	86
E	94	82	78	✗62	100	79
F	✗64	80	84	86	79	100



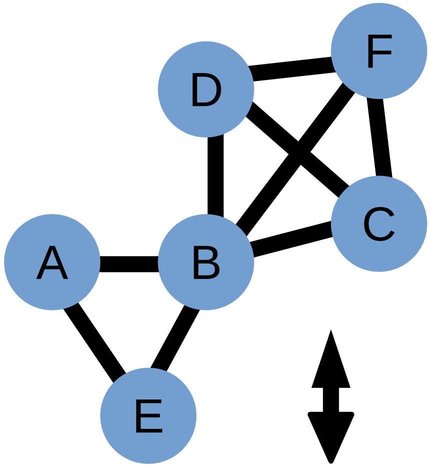
Id: 60

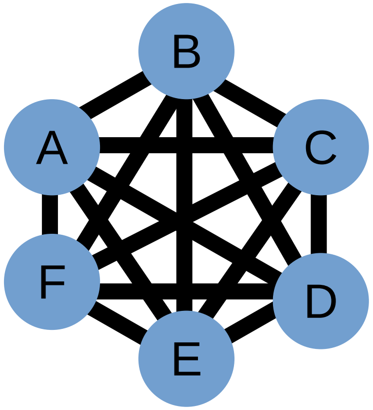


Id: 70

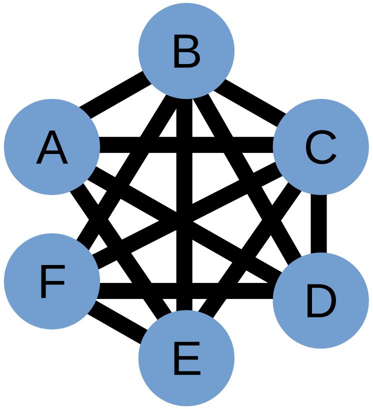
How are networks visualized?

	A	B	C	D	E	F
A	100	91	76	72	92	64
B	91	100	91	96	82	80
C	76	98	100	94	78	84
D	72	96	94	100	62	86
E	94	82	78	62	100	79
F	64	80	84	86	79	100

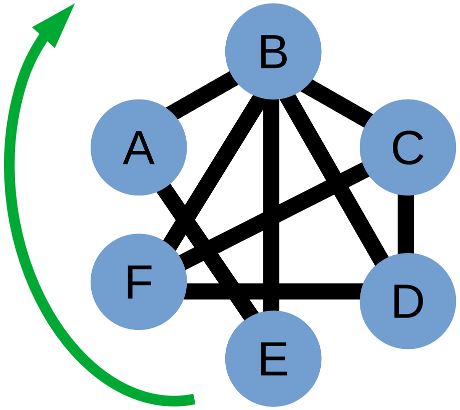




Id: 60



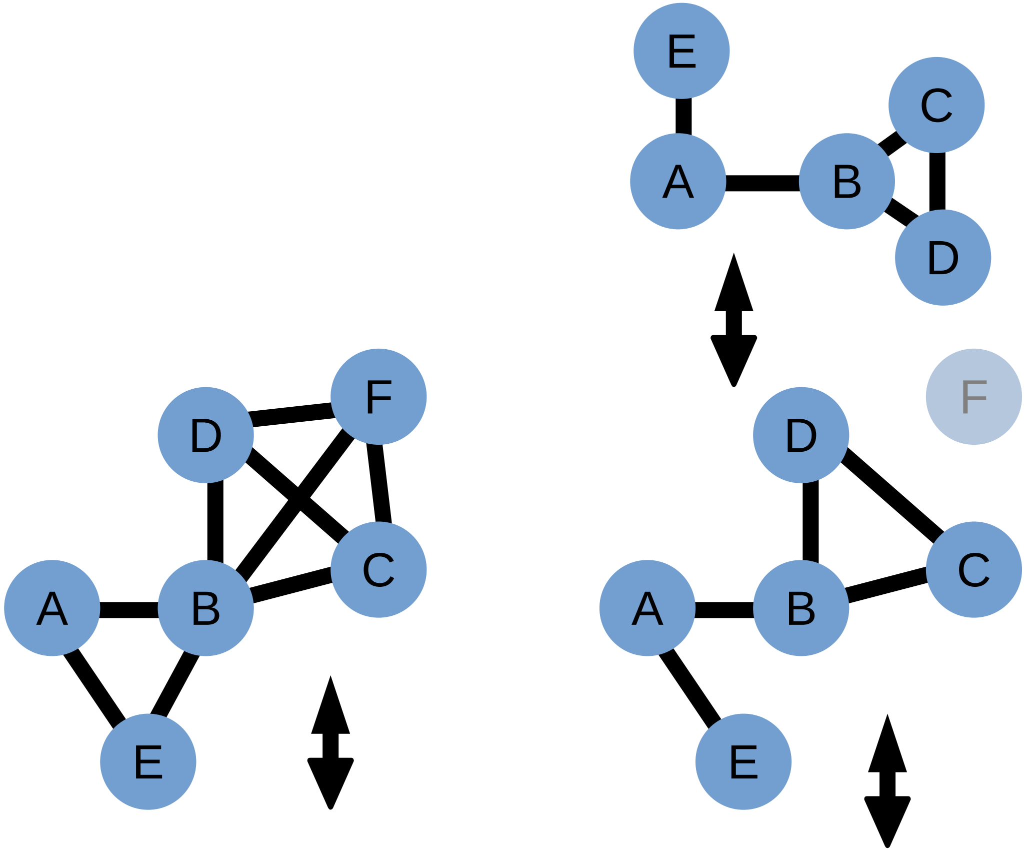
Id: 70

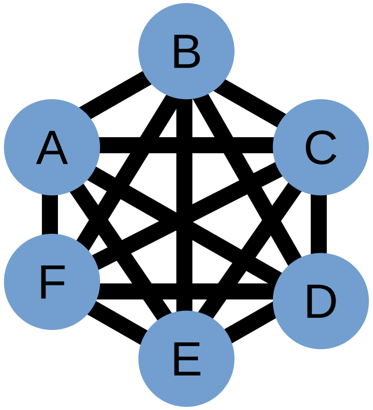


Id: 80

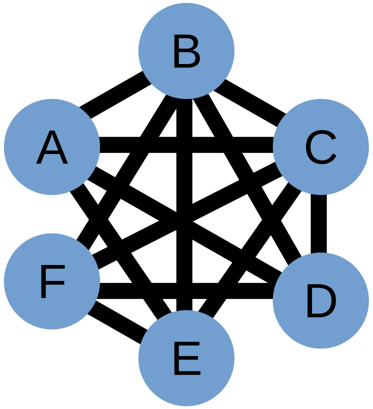
How are networks visualized?

	A	B	C	D	E	F
A	100	91	76	72	92	64
B	91	100	91	96	82	80
C	76	98	100	94	78	84
D	72	96	94	100	62	86
E	94	82	78	62	100	79
F	64	80	84	86	79	100

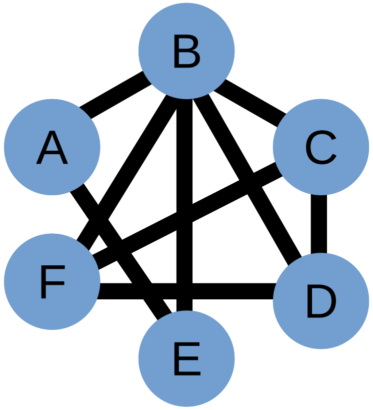




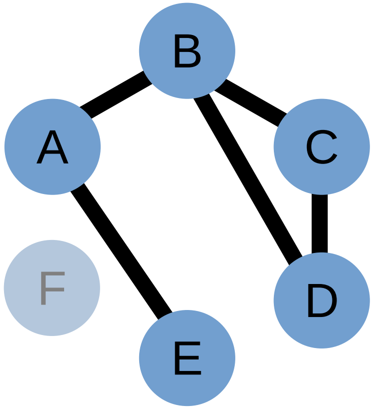
Id: 60



Id: 70



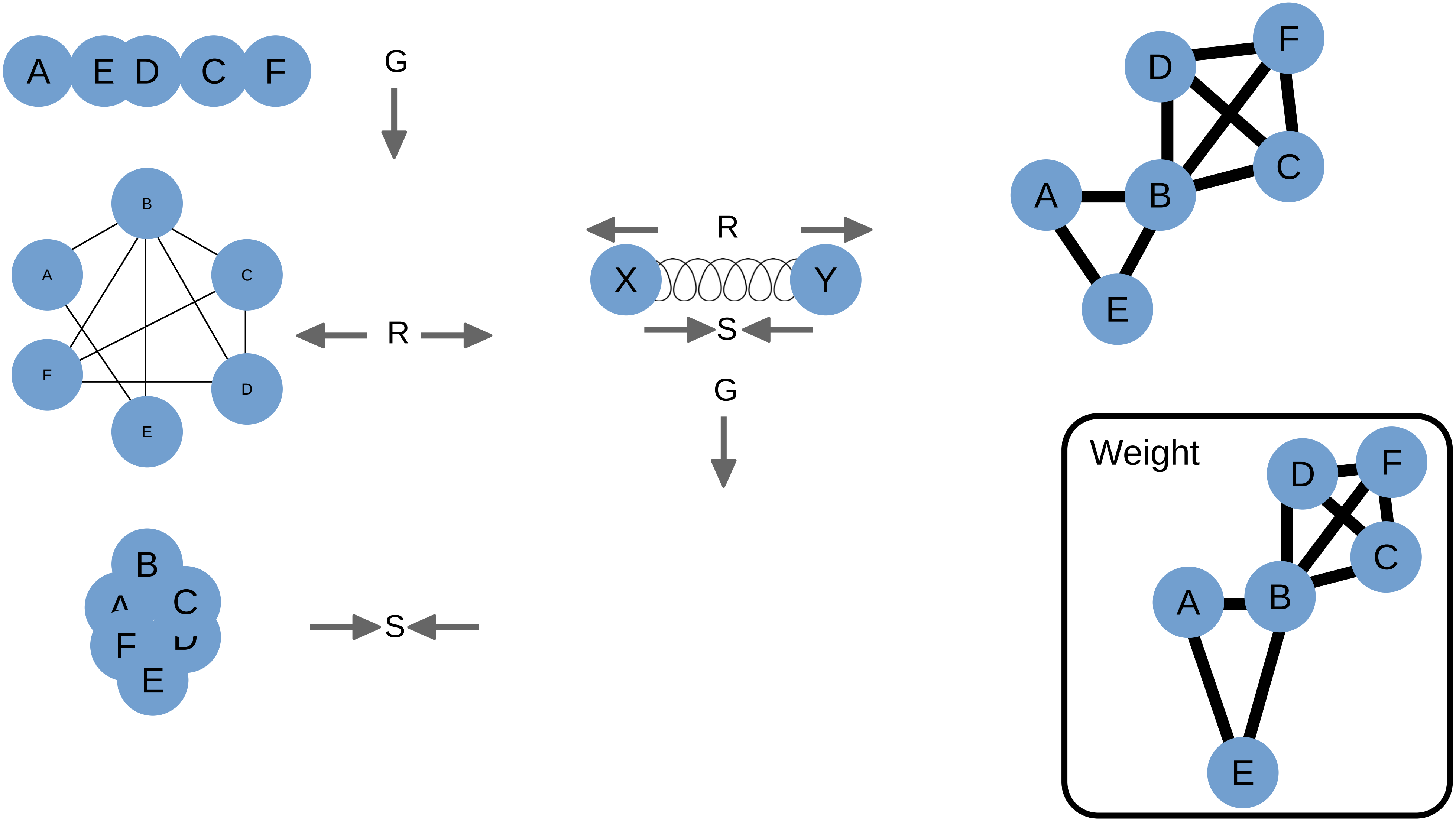
Id: 80



Id: 90

How are networks visualized?
The layout

2D representation: Gravity, Repulsion and Spring forces
(and some more...)



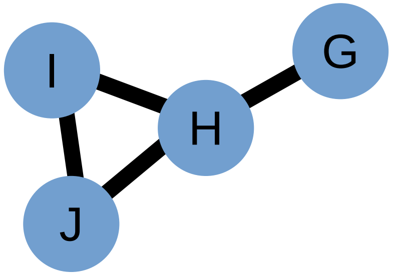
Properties of networks

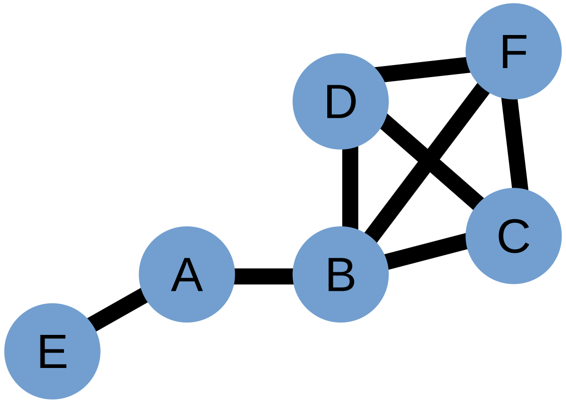
Nnodes: 4

Nedges: 4

Connectivity: 2

Density: 0.66





Nnodes: 6

Nedges: 8

Connectivity: 2.6

Density: 0.53

Connected components: 2

Number of nodes: 10

Number of edges: 12

Connectivity: $\frac{2+4+3+3+1+3+1+3+2+2}{10}$ = 2.4

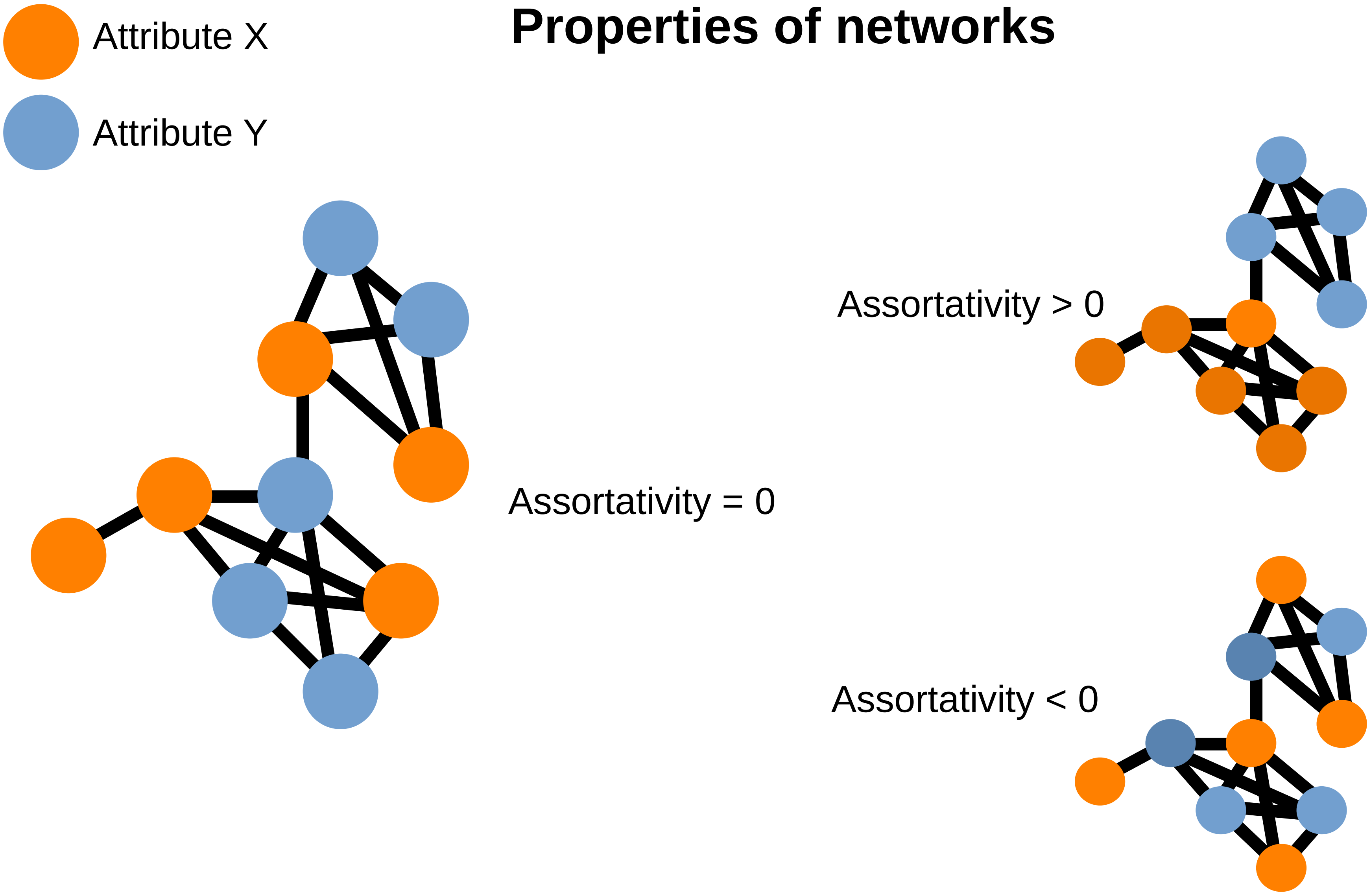
Clustering coefficient: $\frac{2e}{n(n-1)}=\frac{12}{10(10-1)}$ = 0.26

Connected components: A subgraph in which any pair of nodes is connected, and that is not connected to the rest of the graph

Connectivity: Average number of neighbors

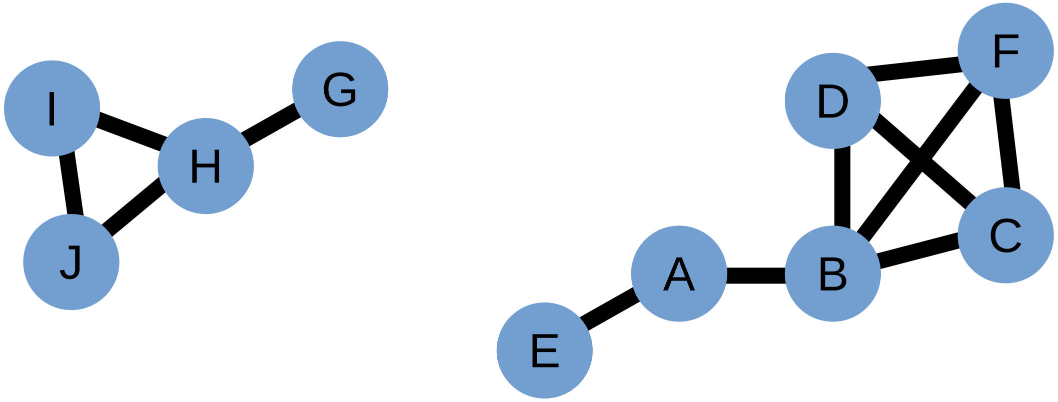
Clustering coefficient (**density**): Proportion of number of edges with respect to the maximum possible edges.

Properties of networks



Assortativity: A measure of the preference for labeled nodes in a network to attach to other nodes with identical labels.

Properties of the nodes



Degree:	B
Closeness:	H
Eccentricity:	E, D, C or F
Betweenness:	B or H

Degree: Number of edges that a node is connected to.
Closeness: Average shortest distance between a node and all the other nodes.
Eccentricity: Average longest distance between a node and all other nodes.
Betweenness: Frequency at which a node is found in all the possible shortest paths between any two nodes in the network.

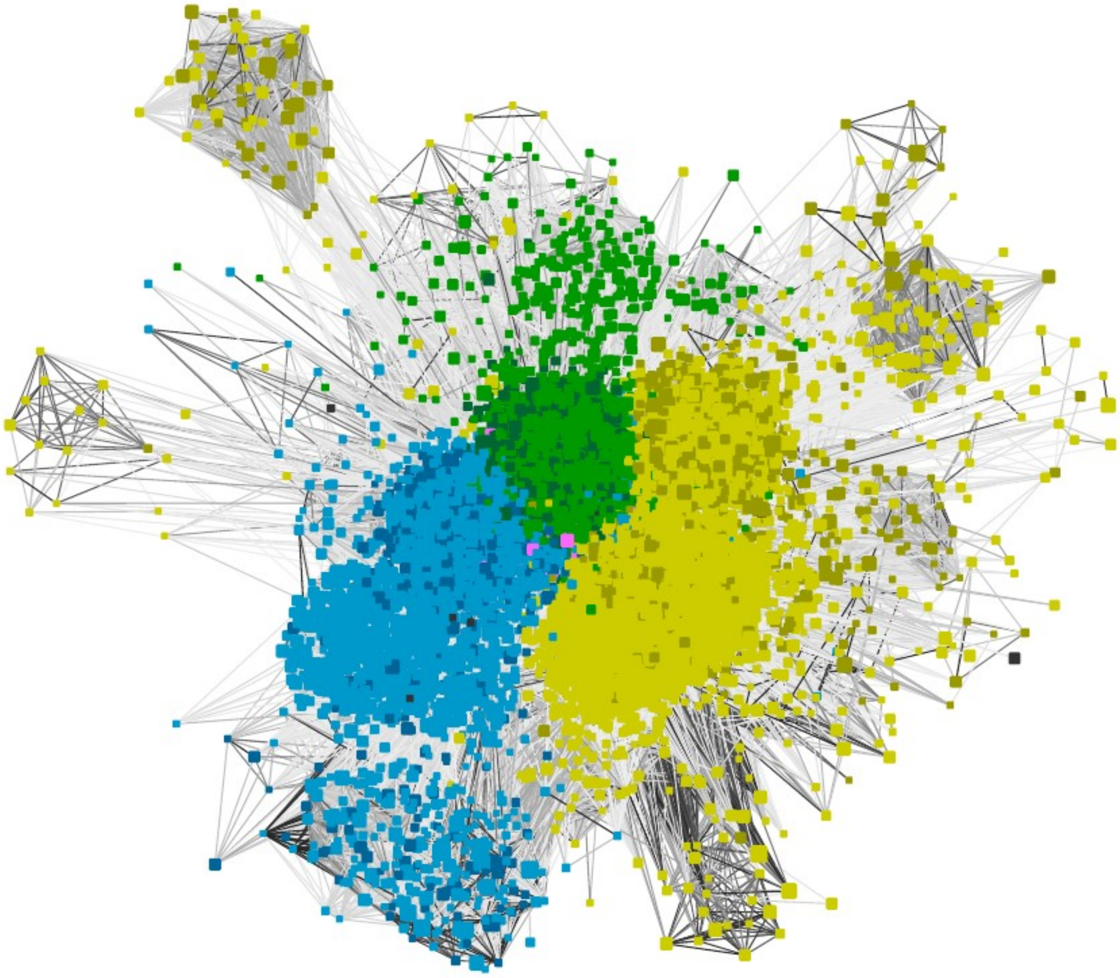
What is the difference between “betweeness” and “closeness” or “eccentricity”?

The betweenness describe the relative position of the node, whereas the closeness and eccentricity is telling how central or peripheral, respectively, the given node is.

A betweenness close to 1 is indicative of a highly central gene, whereas close to 0 is more peripheral.

Hands on!

Let’s get to build some networks



Hands on!

Let’s get to build some networks

1. BLAST all against all

    1.1 Clean the blast output

2. Visualize the network

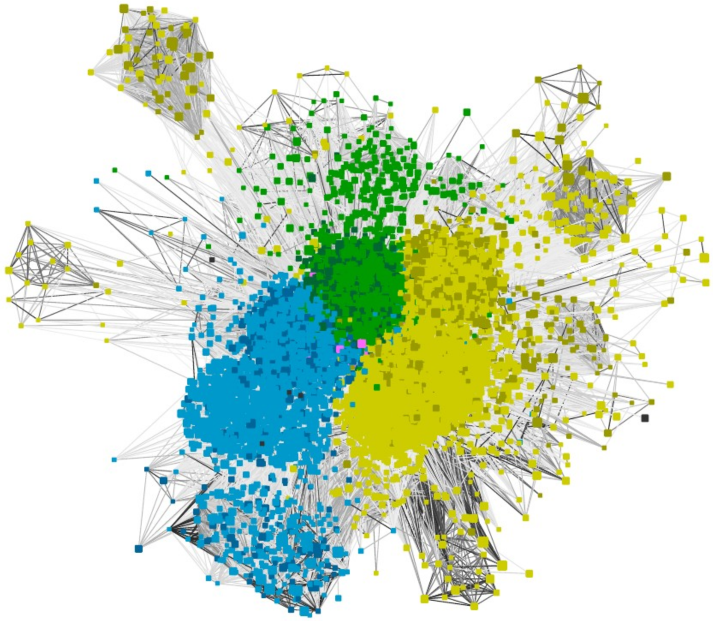
    2.1 Build the network

    2.2 Prepare some attributes

3. Analyze the network

4. Explore assortativity of the attributes

5. Other analysis (shortest path analysis)



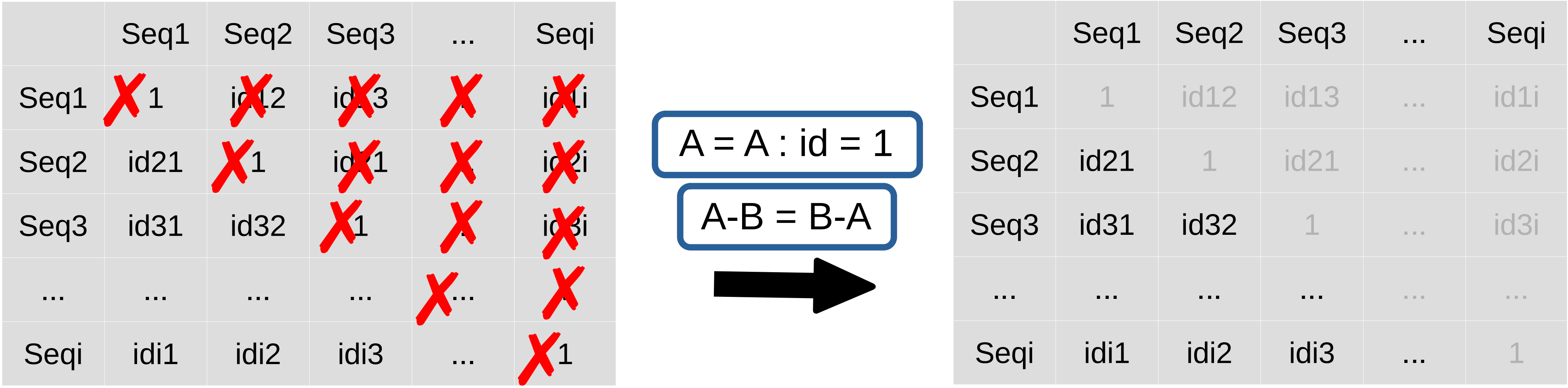
1. BLAST all against all

1. Create a database of the fasta file and run blast against the database:

```
1_blastn_allAgainstAll.sh
```

1.1 Clean the blast output

```
1.1_blastnClean.py
```



	Seq1	Seq2	Seq3	...	Seqi
Seq1	✗ 1	✗ id12	✗ id13	✗	✗ id1i
Seq2	id21	✗ 1	✗ id21	✗	✗ id2i
Seq3	id31	id32	✗ 1	✗	✗ id3i
...	...	...	...	✗ ...	✗
Seqi	idi1	idi2	idi3	...	✗ 1

A = A : id = 1

A-B = B-A

	Seq1	Seq2	Seq3	...	Seqi
Seq1	1	id12	id13	...	id1i
Seq2	id21	1	id21	...	id2i
Seq3	id31	id32	1	...	id3i
...	...	...	...	...	...
Seqi	idi1	idi2	idi3	...	1

2. Visualize the network

2.1 Build the network

`2.1_buildNetwork.py`

2.2 Prepare some attributes

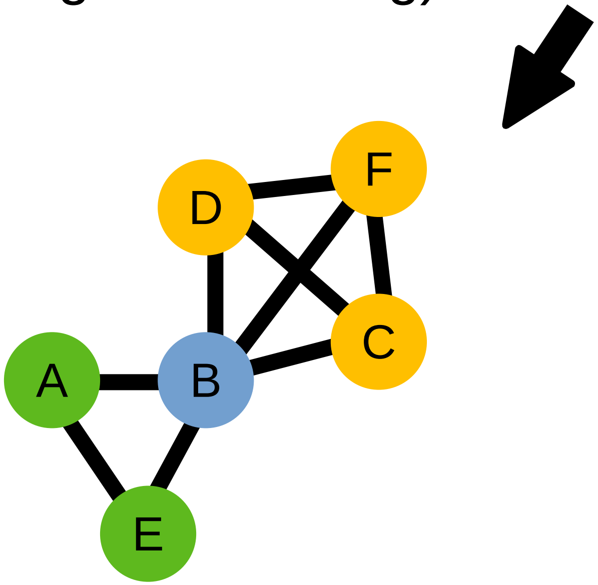
`2.2_attributes_file.R`

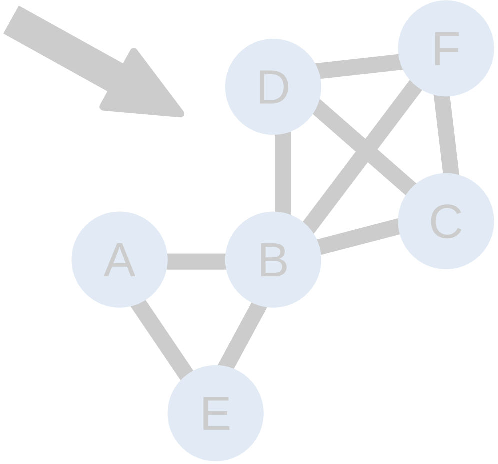
	Seq1	Seq2	Seq3	...	Seqi
Seq1	1	id12	id13	...	id1i
Seq2	id21	1	id21	...	id2i
Seq3	id31	id32	1	...	id3i
...	...	...	...	...	...
Seqi	idi1	idi2	idi3	...	1

Seq1	Seq2	id12
Seq1	Seq3	id13
Seq1	Seq4	id14
...	...	...
Seqi	Seqj	idij

Attributes
(Biological meaning)

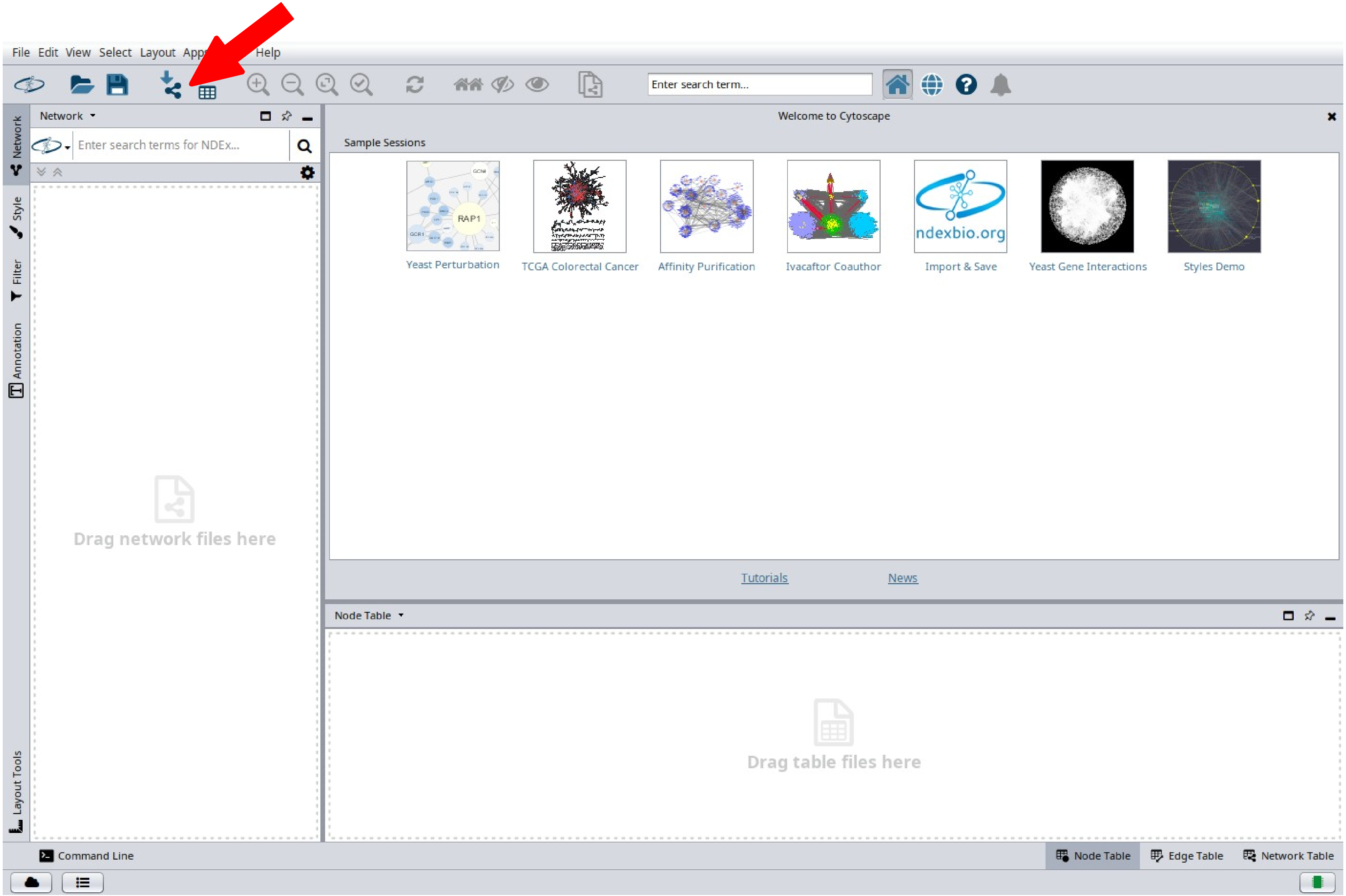
+





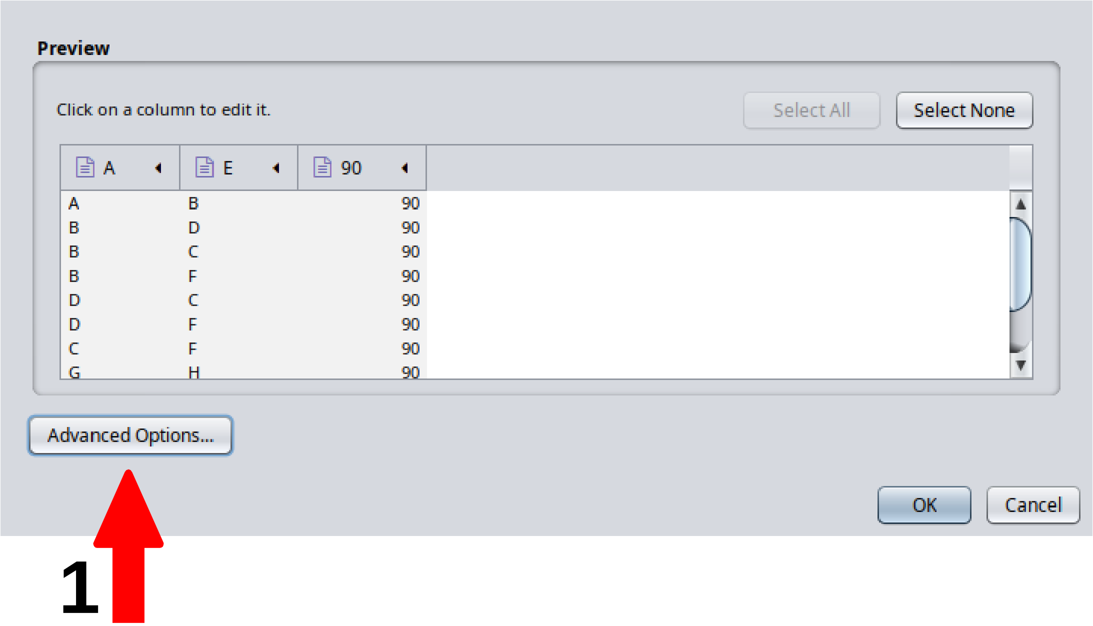
2. Visualize the network

Cytoscape



2. Visualize the network

Cytoscape



Preview

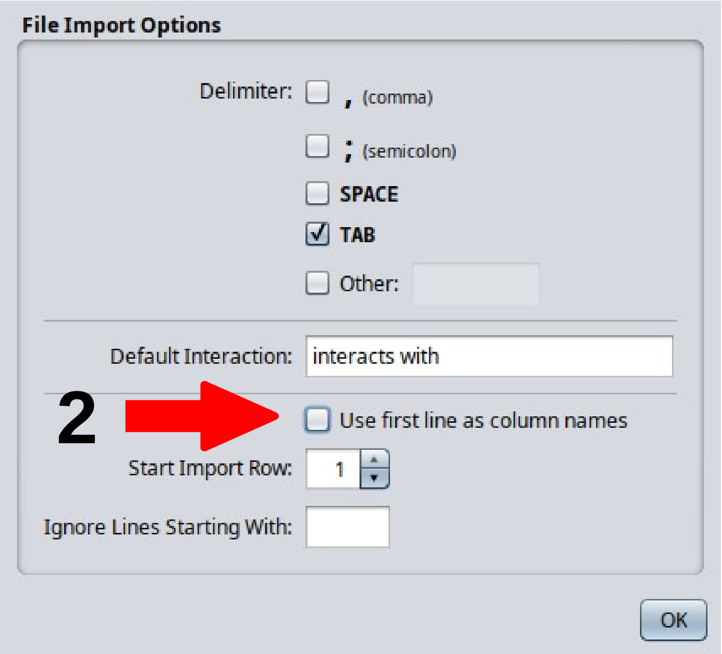
Click on a column to edit it.

Select All | Select None

A	E	90
A	B	90
B	D	90
B	C	90
B	F	90
D	C	90
D	F	90
C	F	90
G	H	90

Advanced Options...

OK | Cancel



File Import Options

Delimiter: □ , (comma)
□ ; (semicolon)
□ SPACE
☑ TAB
□ Other:

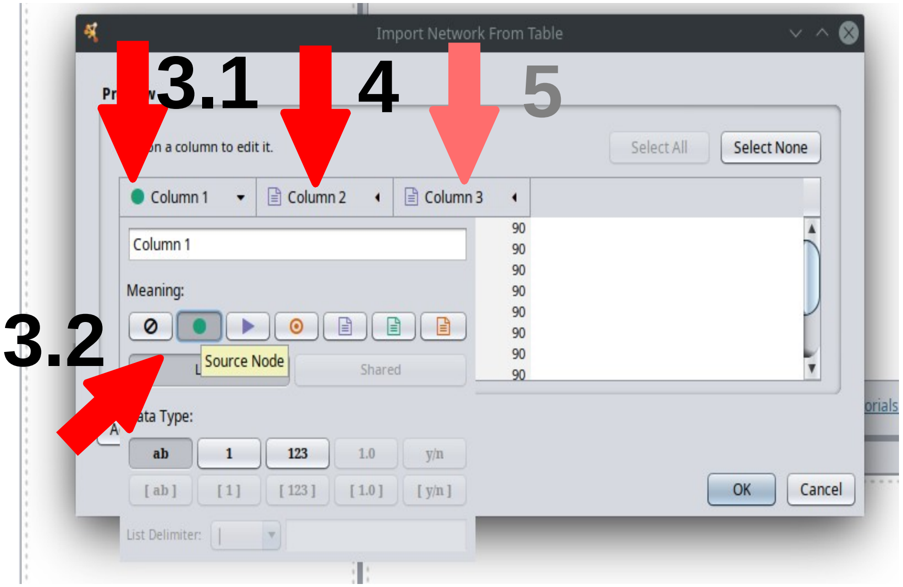
Default Interaction: interacts with

□ Use first line as column names

Start Import Row: 1

Ignore Lines Starting With:

OK



Import Network From Table

Pr...w

...n a column to edit it.

Select All | Select None

● Column 1	Column 2	Column 3
		90
		90
		90
		90
		90
		90
		90
		90

Column 1

Meaning:

Source Node

Shared

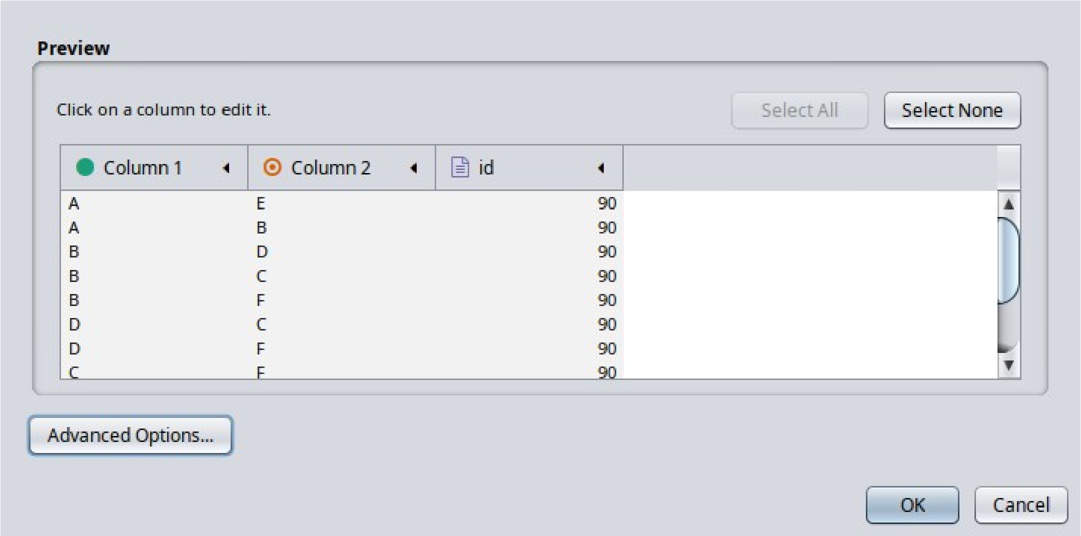
...ata Type:

ab | 1 | 123 | 1.0 | y/n

[ab] | [1] | [123] | [1.0] | [y/n]

List Delimiter: |

OK | Cancel



Preview

Click on a column to edit it.

Select All | Select None

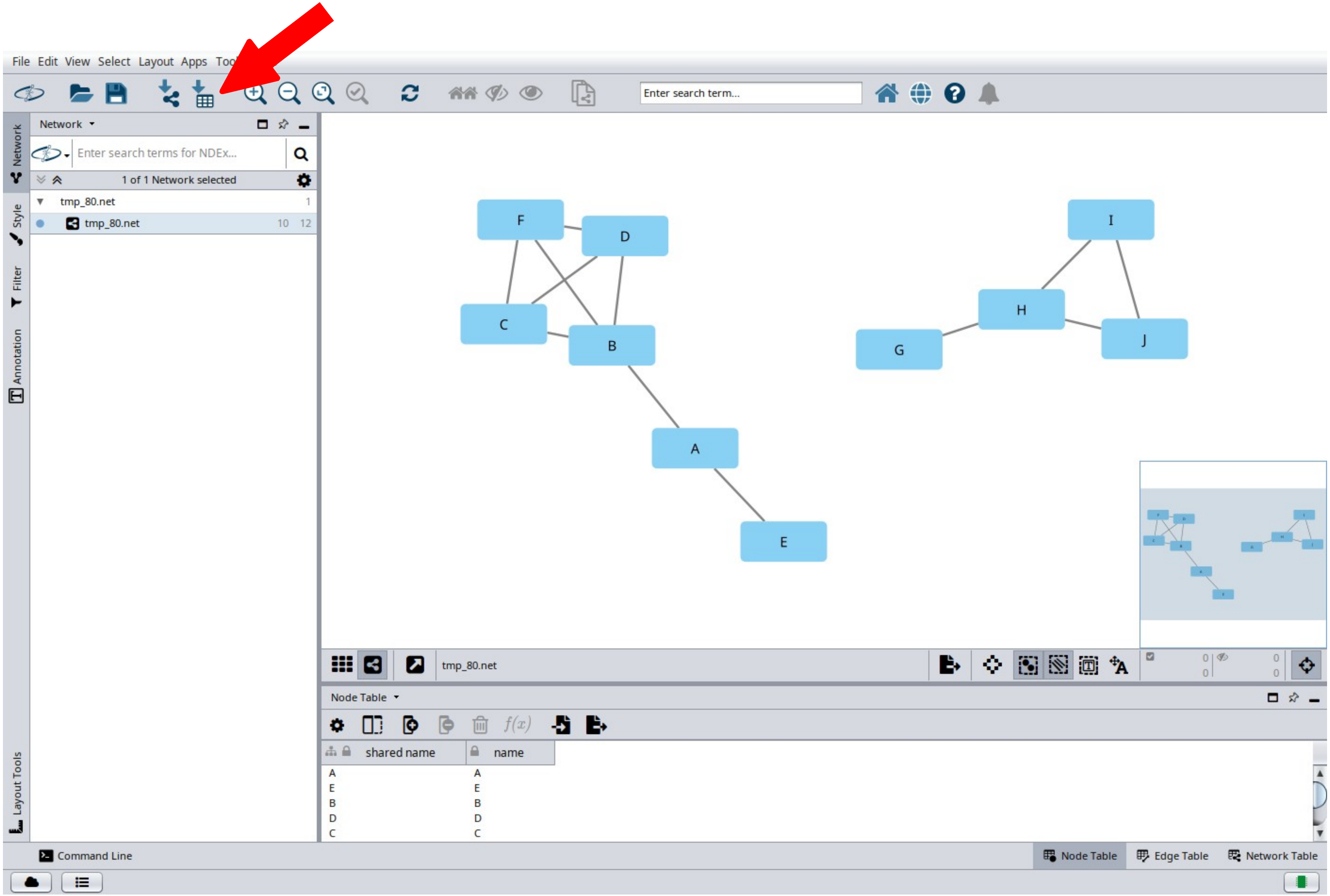
● Column 1	⊙ Column 2	id
A	E	90
A	B	90
B	D	90
B	C	90
B	F	90
D	C	90
D	F	90
C	F	90

Advanced Options...

OK | Cancel

2. Visualize the network

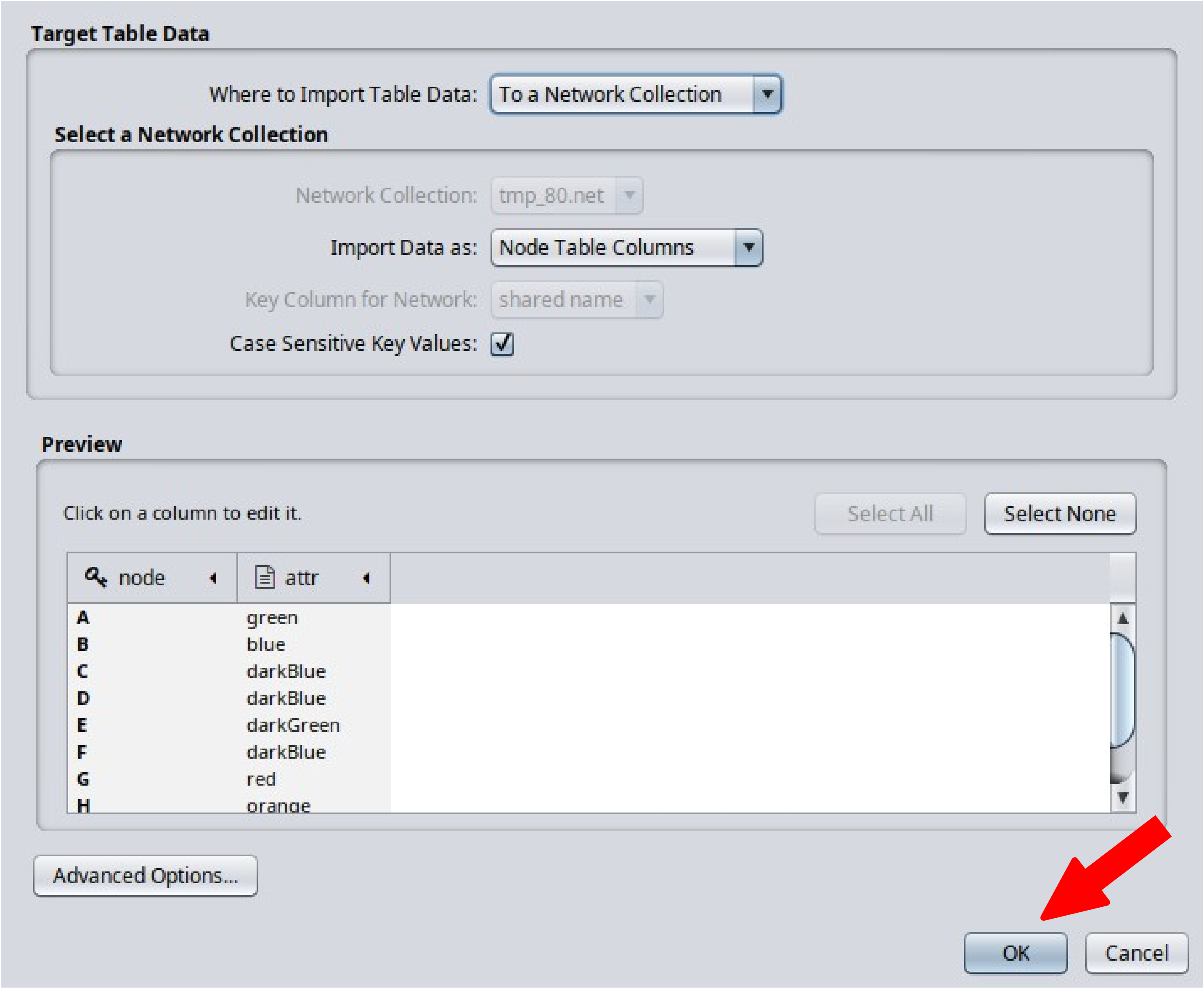
Cytoscape



shared name	name
A	A
E	E
B	B
D	D
C	C

2. Visualize the network

Cytoscape



Target Table Data

Where to Import Table Data: To a Network Collection

Select a Network Collection

Network Collection: tmp_80.net

Import Data as: Node Table Columns

Key Column for Network: shared name

Case Sensitive Key Values: ☑

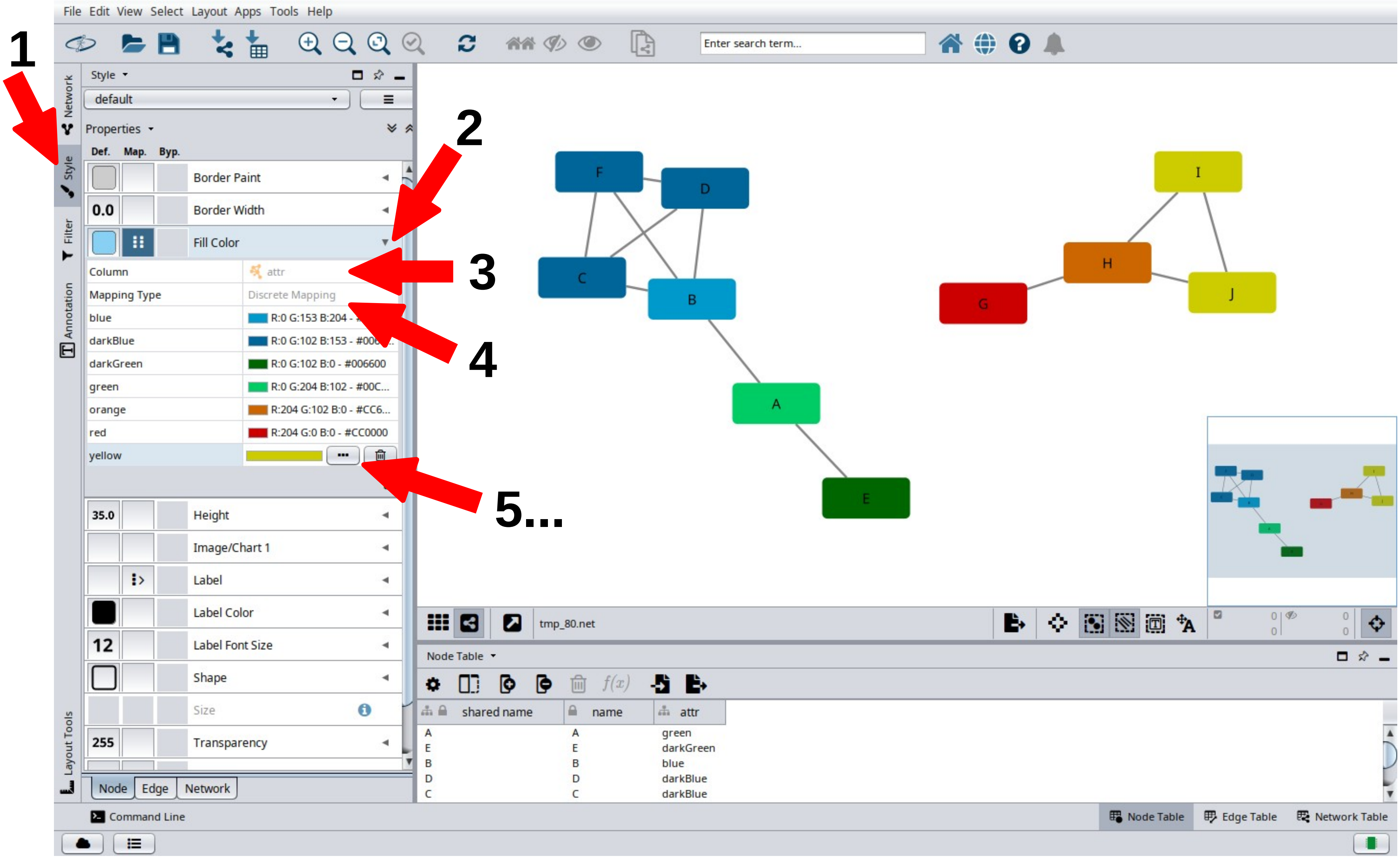
Preview

Click on a column to edit it.

node	attr
A	green
B	blue
C	darkBlue
D	darkBlue
E	darkGreen
F	darkBlue
G	red
H	orange

2. Visualize the network

Cytoscape



Style: default

Def.	Map.	Byp.	Property
			Border Paint
0.0			Border Width
			Fill Color

Column	attr
Mapping Type	Discrete Mapping
blue	R:0 G:153 B:204 -
darkBlue	R:0 G:102 B:153 - #006...
darkGreen	R:0 G:102 B:0 - #006600
green	R:0 G:204 B:102 - #00C...
orange	R:204 G:102 B:0 - #CC6...
red	R:204 G:0 B:0 - #CC0000
yellow	

Def.	Map.	Byp.	Property
35.0			Height
			Image/Chart 1
			Label
			Label Color
12			Label Font Size
			Shape
			Size
255			Transparency

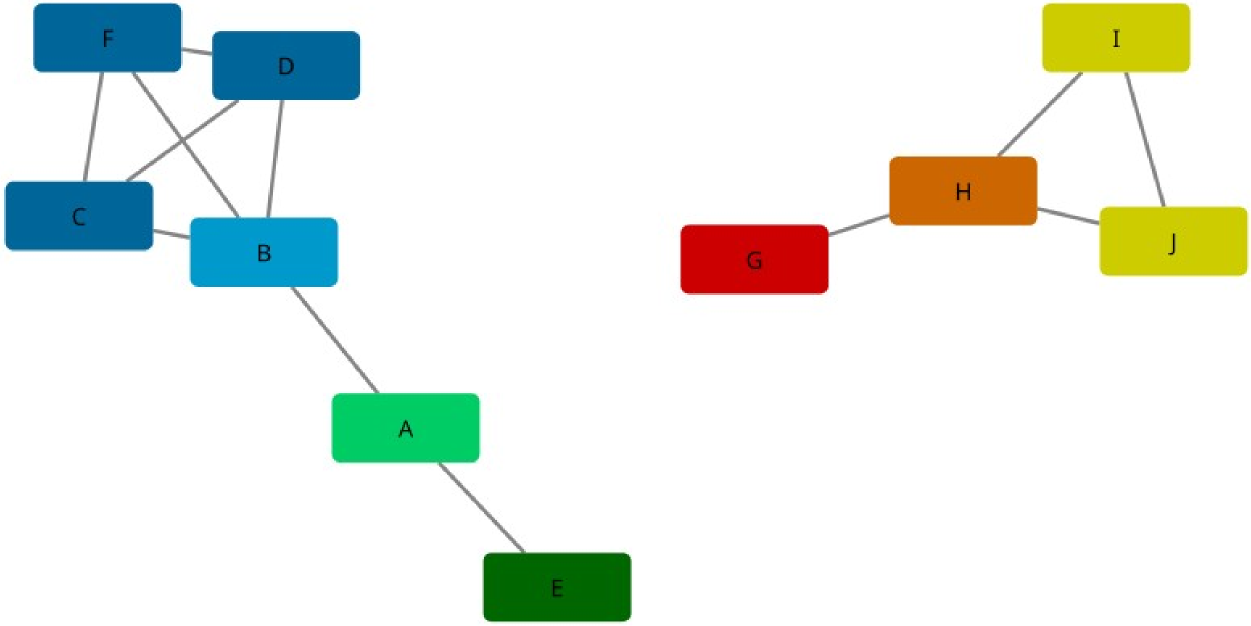
shared name	name	attr
A	A	green
E	E	darkGreen
B	B	blue
D	D	darkBlue
C	C	darkBlue

2. Visualize the network

Cytoscape

Now let’s play and compare our networks.

And try to get some biological meaning out of it!



3. Analyze the network

3. Calculate properties of the network (its Connected Components) and the nodes:

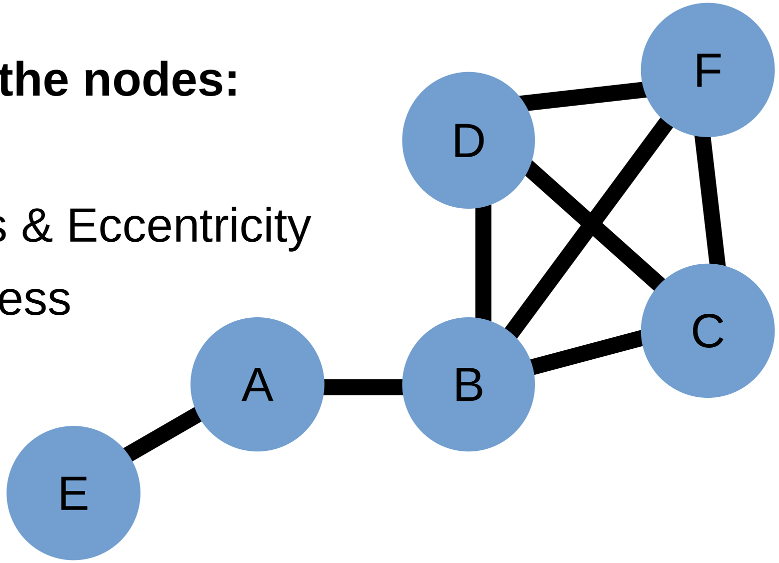
`3_analyzeNetwork.py`

Properties of the network:

- Connected components
- Number of nodes
- Number of edges
- Connectivity
- Clustering coefficient

Properties of the nodes:

- Degree
- Closeness & Eccentricity
- Betweenness

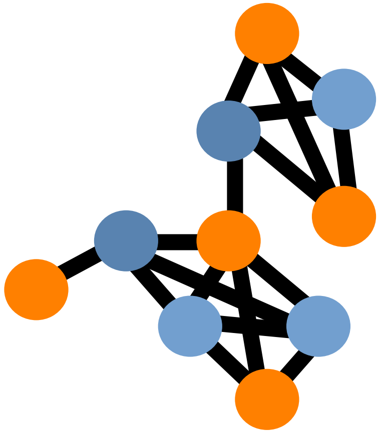


Connected components: A subgraph in which any pair of nodes is connected, and that is not connected to the rest of the graph
Connectivity: Average number of neighbors
Density: Proportion of number of edges with respect to the maximum possible edges.
Degree: Number of edges that a node is connected to.
Closeness: Average shortest distance between a node and all the other nodes.
Eccentricity: Average longest distance between a node and all other nodes.
Betweenness: Frequency at which a node is found in all the possible shortest paths between any two nodes in the network.

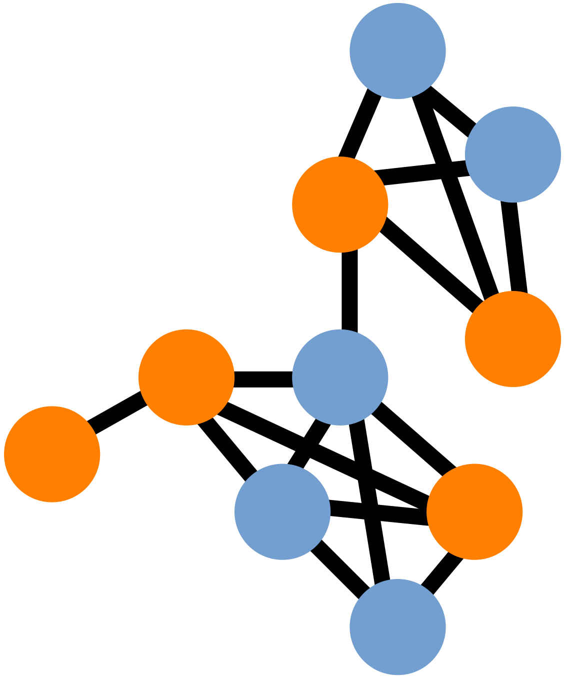
4. Assortativity of the attributes

4. Calculate how your attributes are connected within and between them:

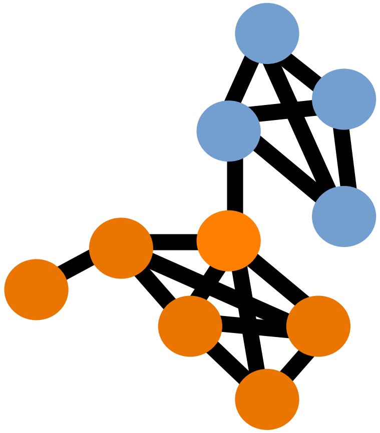
```
4_analyzeNetworkAssortativity.py
```



Assortativity < 0



Assortativity = 0



Assortativity > 0

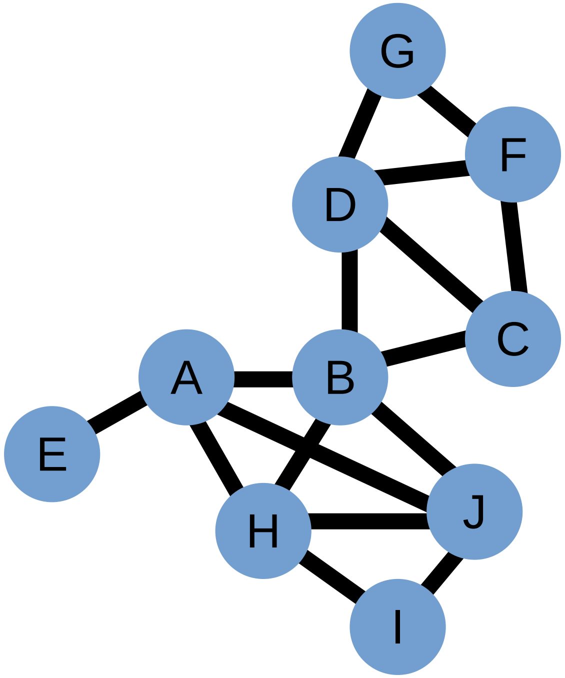
Assortativity:A measure of the preference for labelled nodes in a network to attach to other nodes with identical labels.

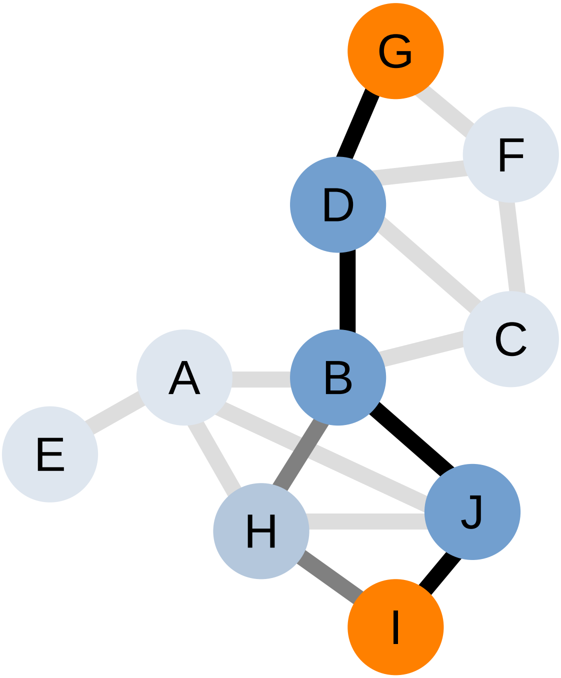
5. Shortest path analysis

Minimum distance (number of edges) between two nodes

5. Calculate shortest path between all pairs of nodes from attribute A and a attrbiute B

```
5_analyzeNetworkShortestPath.py
```





Path through *H* and through *J* are equivalent

We can finally get to see our results

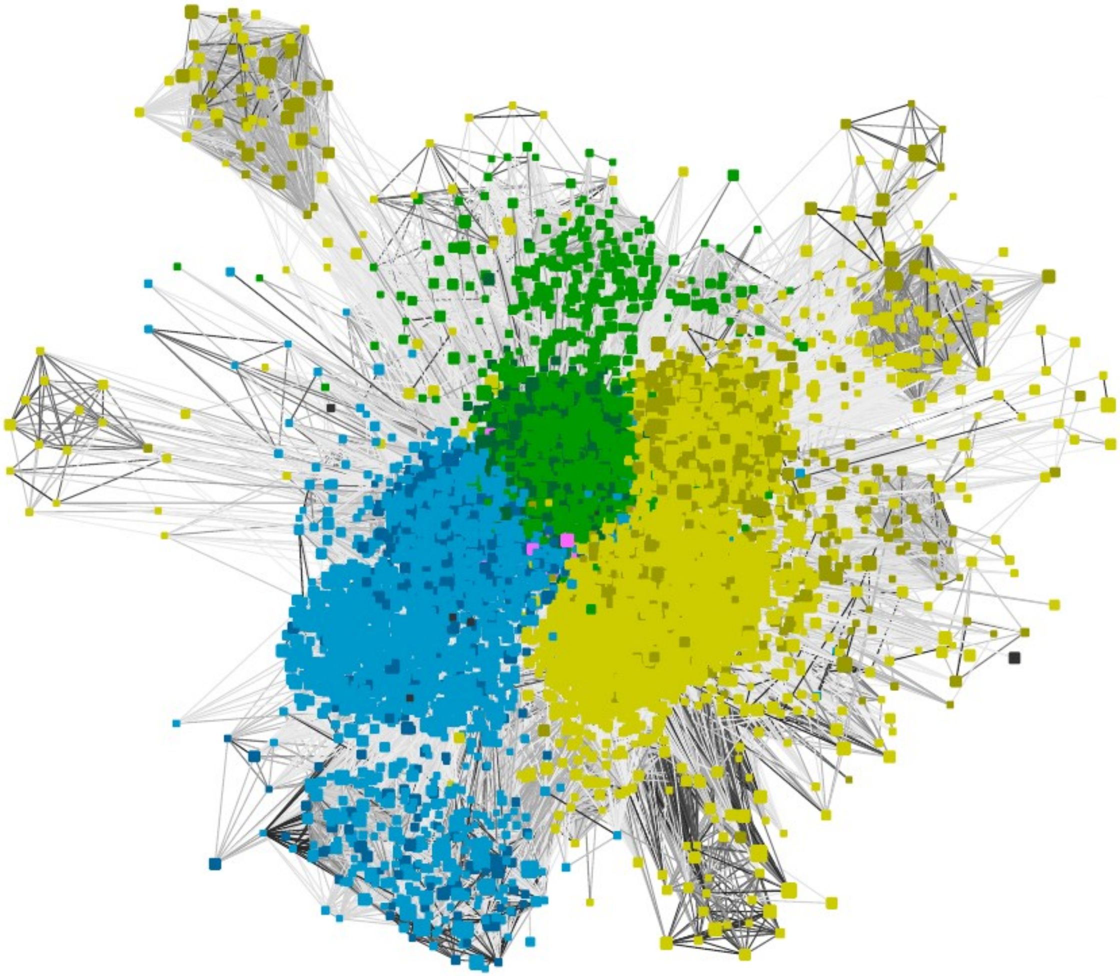
6. Plotting the results

`statsNetworks.R`

Structure of the script:

`1. Libraries`	# Load required packages
`2. Working directory`	# Set your preferred working directory
`3. Network analysis`	# Plot results for every network (≠ ID thresholds)
`4. Connected components`	
`5. Nodes centralities`	
`6. Assortativity`	
`7. Shortest path`	

Tree and network thinking



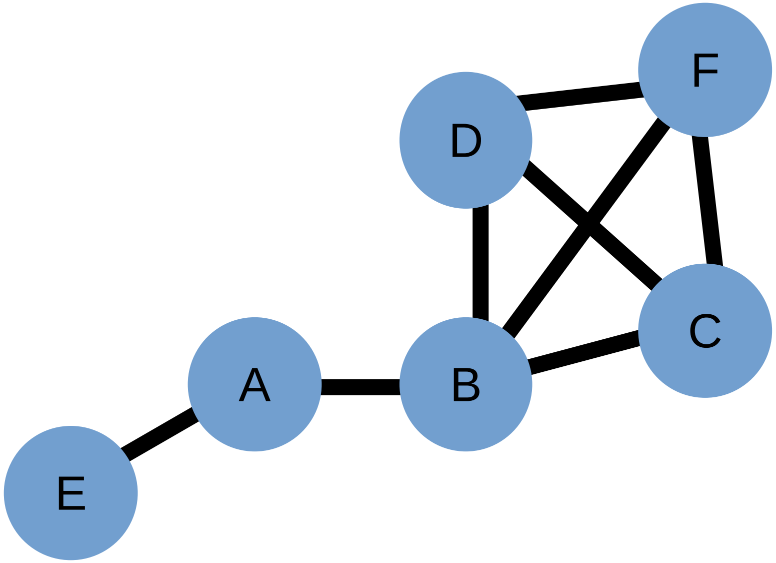
Quick reminder

Properties of the network:

- Connected components
- Number of nodes
- Number of edges
- Connectivity
- Clustering coefficient
- Assortativity

Properties of the nodes:

- Degree
- Closeness & Eccentricity
- Betweenness



Connected components: A subgraph in which any pair of nodes is connected, and that is not connected to the rest of the graph
Connectivity: Average number of neighbors
Density: Proportion of number of edges with respect to the maximum possible edges.
Assortativity:A measure of the preference for labelled nodes in a network to attach to other nodes with identical labels.
Degree: Number of edges that a node is connected to.
Closeness: Average shortest distance between a node and all the other nodes.
Eccentricity: Average longest distance between a node and all other nodes.
Betweenness: Frequency at which a node is found in all the possible shortest paths between any two nodes in the network.

Concluding remarks

The *simplicity* of networks helps tackling issues where phylogenies fail or are limited, and/or give a different perspective in ordination analysis

Phylogenies:

-Alignment dependent

-Evolutionary model

-Phylogenetic inference

-Bifurcating phylogenetic tree

SSN:

-Alignment free

-Similarity search

-Network representation

Ordination:

-Abundance (or presence-Absence)

-Spatial ordination

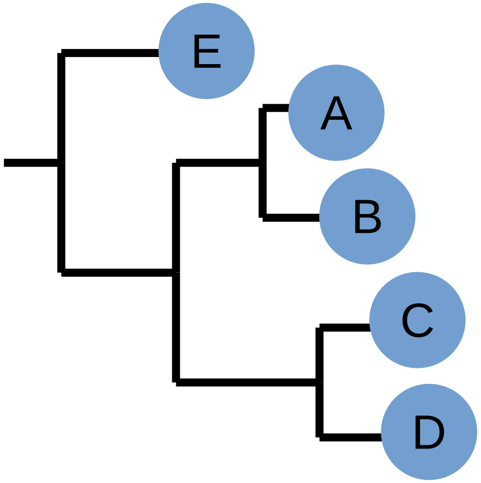
-No genetic information

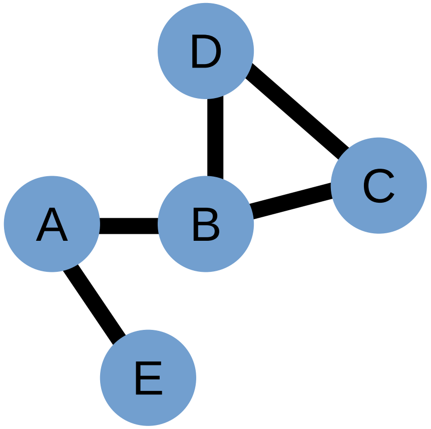
✗✓

Different questions, different approaches

Networks COMPLEMENT previous well-established methods

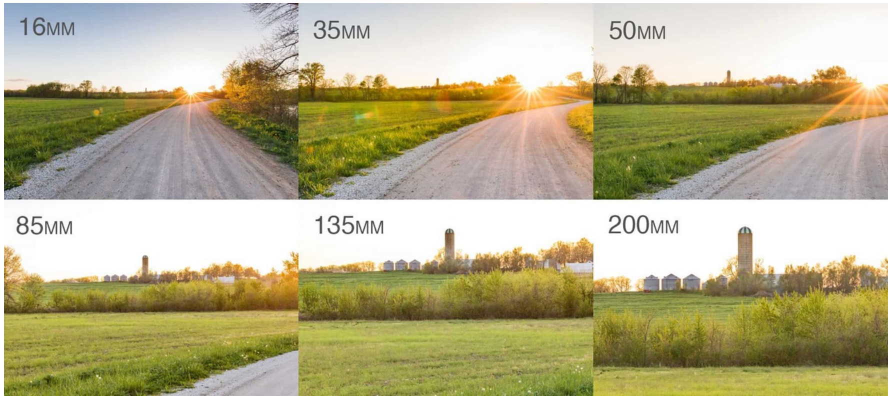
Concluding remarks





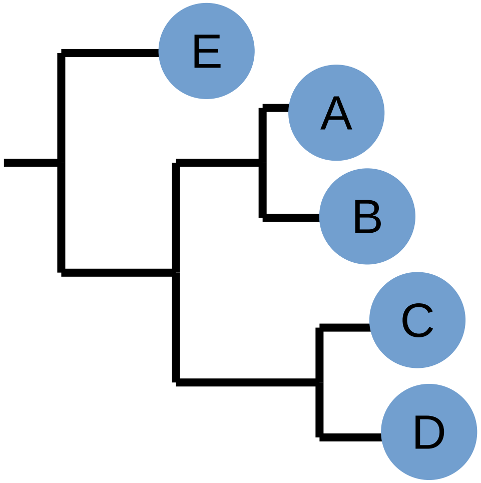
Each tree or network is
an hypothesis for the given data!

Different pictures
of the same “reality”

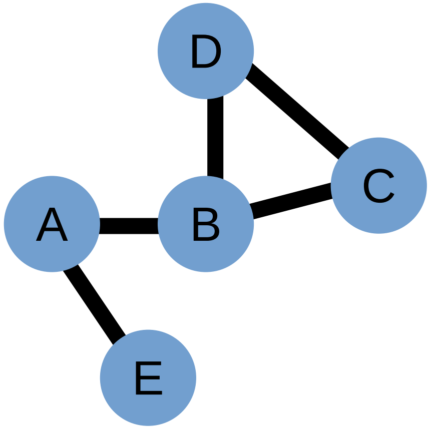


(c) Google images

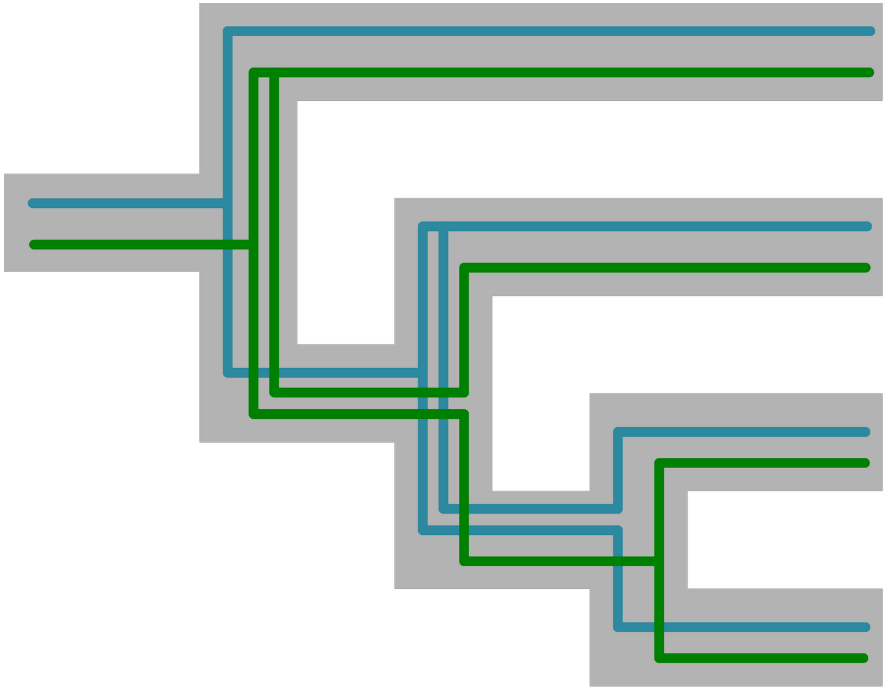
Concluding remarks

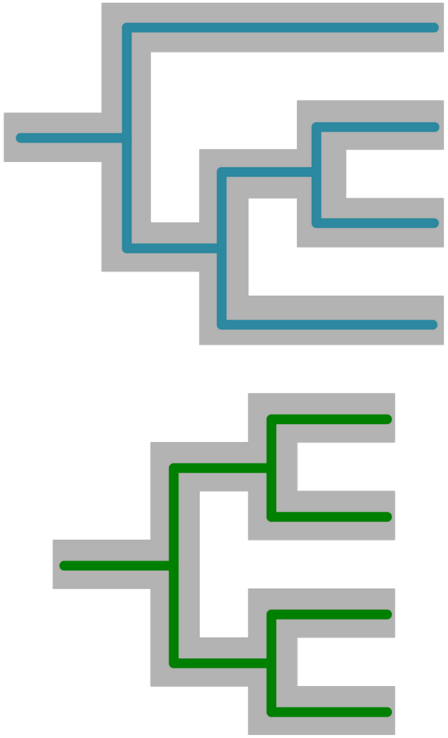


Each tree or network is
an hypothesis for the given data!



Different pictures
of the same "reality"





Only by accessing all truths we can better understand the true patterns:
Yet, what is understanding?