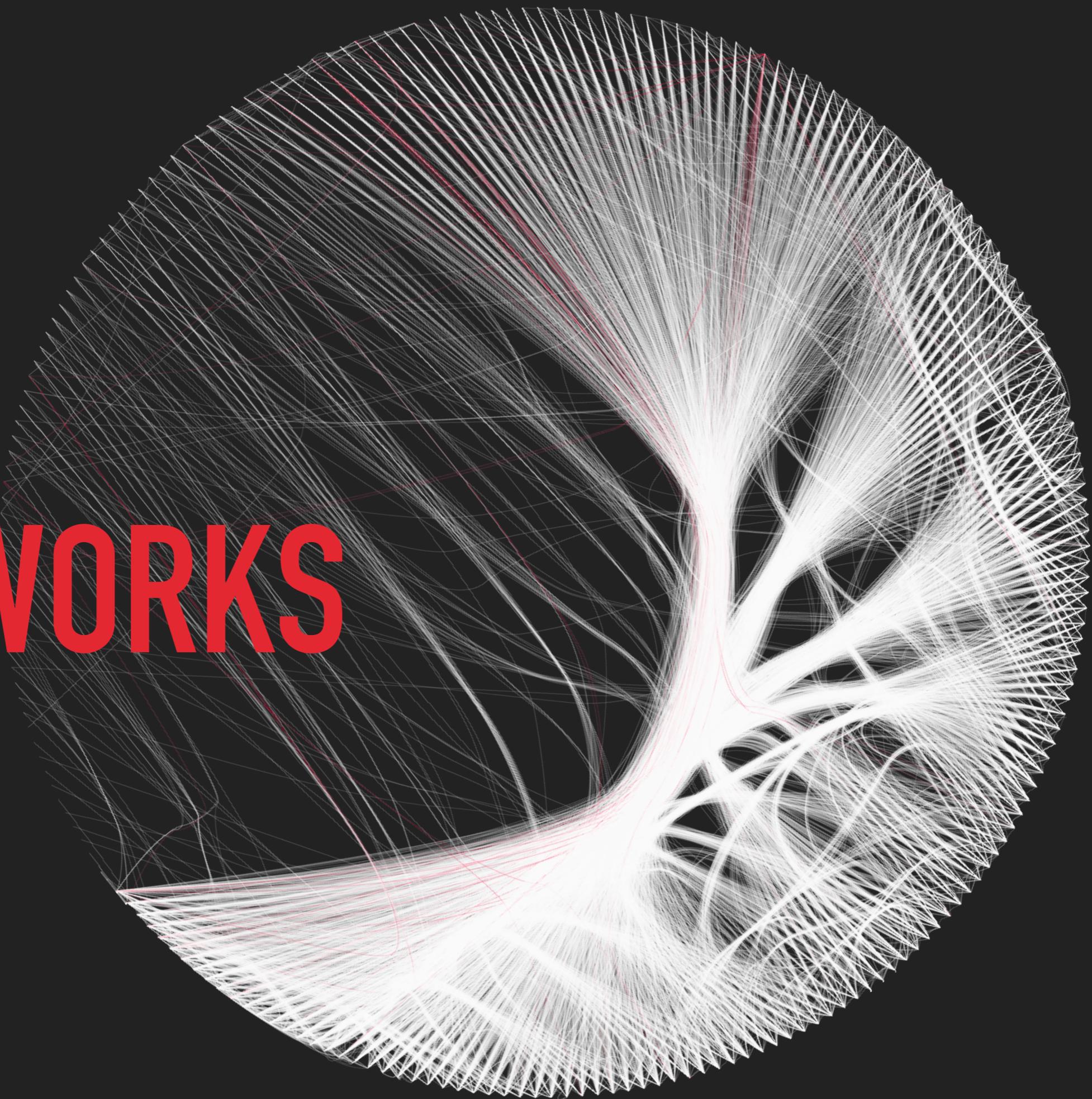


NETWORKS



NETWORK THEORY IN A NUTSHELL

Local similarity between two OTU time series: Suppose the two OTU normal transformed series are O_1 and O_2 with both of length n , i.e. $O_1 = O_{11}, O_{12}, \dots, O_{1n}$ and $O_2 = O_{21}, O_{22}, \dots, O_{2n}$. The positive score matrix $P_{n \times n}$ and negative score matrix $N_{n \times n}$ are calculated as follows:

(1) For $i, j = 1, \dots, n$,

$$P_{0,i} = P_{j,0} = 0 \text{ and } N_{0,i} = N_{j,0} = 0.$$

(2) For $i, j = 1, \dots, n$ with $|i - j| \leq D$,

$$\begin{aligned} P_{i+1,j+1} &= \max\{0, P_{i,j} + O_{1,i+1} \cdot O_{2,j+1}\} \quad \text{and} \\ N_{i+1,j+1} &= \max\{0, N_{i,j} - O_{1,i+1} \cdot O_{2,j+1}\}. \end{aligned}$$

(3) $P(O_1, O_2) = \max_{1 \leq i, j \leq n} P_{i,j}$ and

$$N(O_1, O_2) = \max_{1 \leq i, j \leq n} N_{i,j}.$$

(4) $\text{MaxScore}(O_1, O_2) = \max(P(O_1, O_2), N(O_1, O_2))$ and

$$\text{Flag}(O_1, O_2) = \text{sgn}(P(O_1, O_2) - N(O_1, O_2)).$$

$$\sum_{k=K}^{\infty} \left(\frac{1}{x^2} + \frac{1}{(2k+1)^2 \pi^2} \right) \exp\left(-\frac{(2k+1)^2 \pi^2}{2x^2}\right)$$

$$< \frac{2}{x^2} \sum_{k=K}^{\infty} \exp\left(-\frac{(2k+1)\pi^2}{2x^2}\right)$$

$$= \frac{2 \exp\left(-\frac{(2K+1)\pi^2}{2x^2}\right)}{x^2(1 - \exp\left(-\frac{2\pi^2}{2x^2}\right))}.$$

Thus, for an error threshold β , we can choose K so that

$$\frac{16 \exp\left(-\frac{(2K+1)\pi^2}{2x^2}\right)}{x^2(1 - \exp\left(-\frac{2\pi^2}{2x^2}\right))} \leq \beta. \quad (5)$$

Then we approximate $P(R_n / (\sigma\sqrt{n}) \geq x)$ by

$$1 - 8 \sum_{k=0}^{K-1} \left(\frac{1}{x^2} + \frac{1}{(2k+1)^2 \pi^2} \right) \exp\left(-\frac{(2k+1)^2 \pi^2}{2x^2}\right). \quad (6)$$

Trump Connections

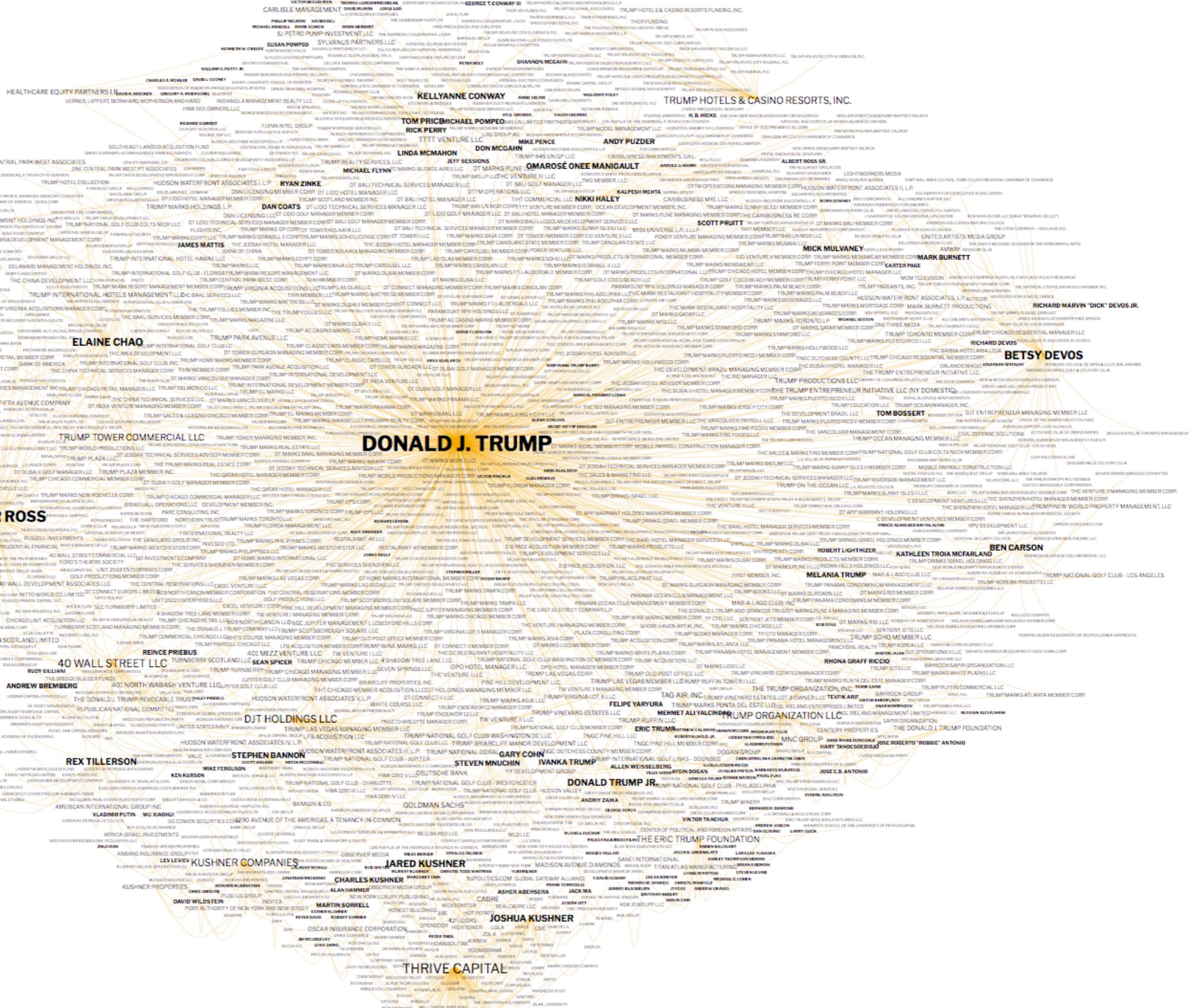
Visualization of 1500 individuals and organizations connected directly and indirectly to Donald Trump. To explore the network further select one of the connections on the right of the graph or search below.

Share: [Facebook](#) | [Twitter](#) | [Reddit](#)

Visualization by Kim Albrecht

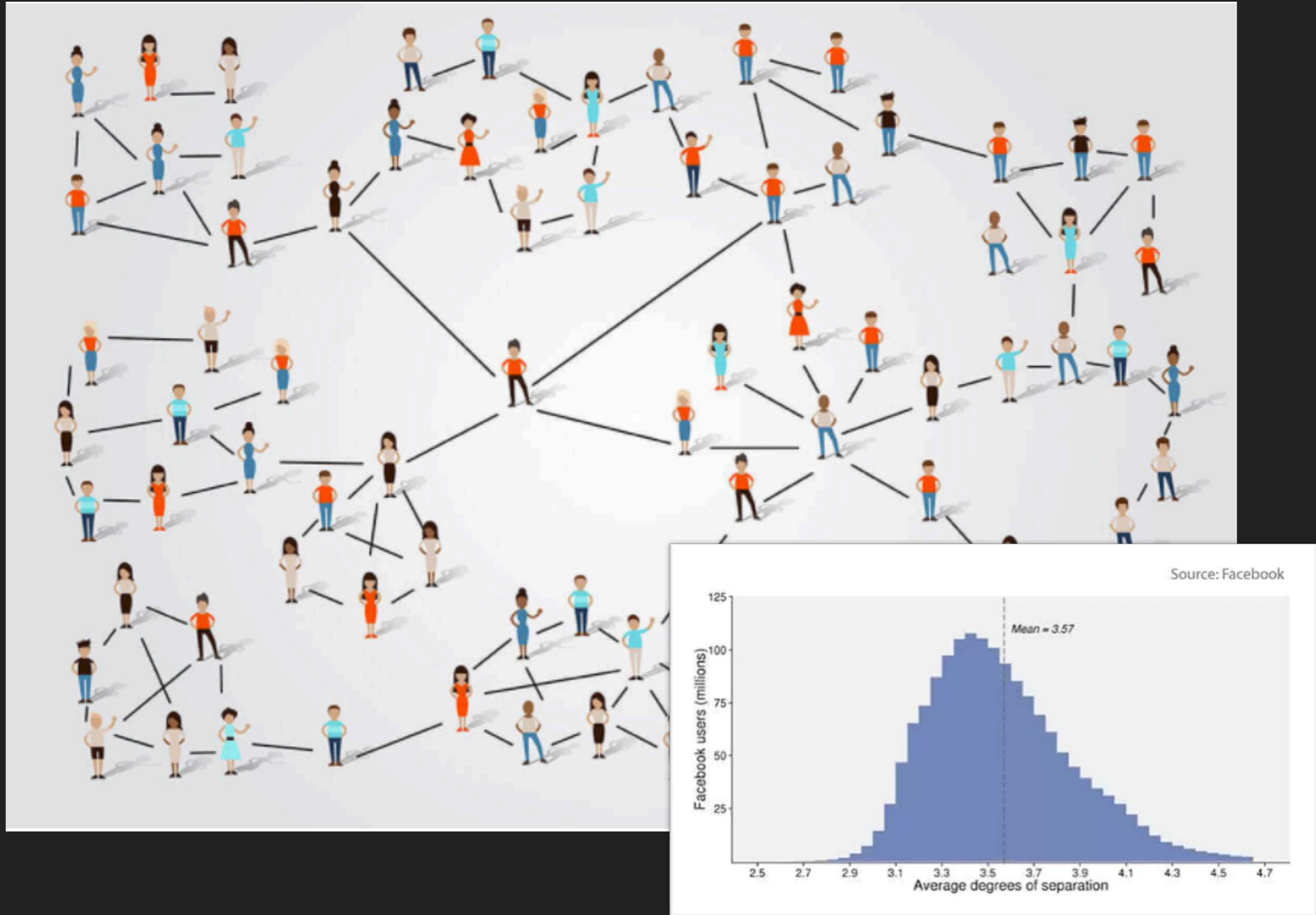
Data and investigation by Buzzfeed

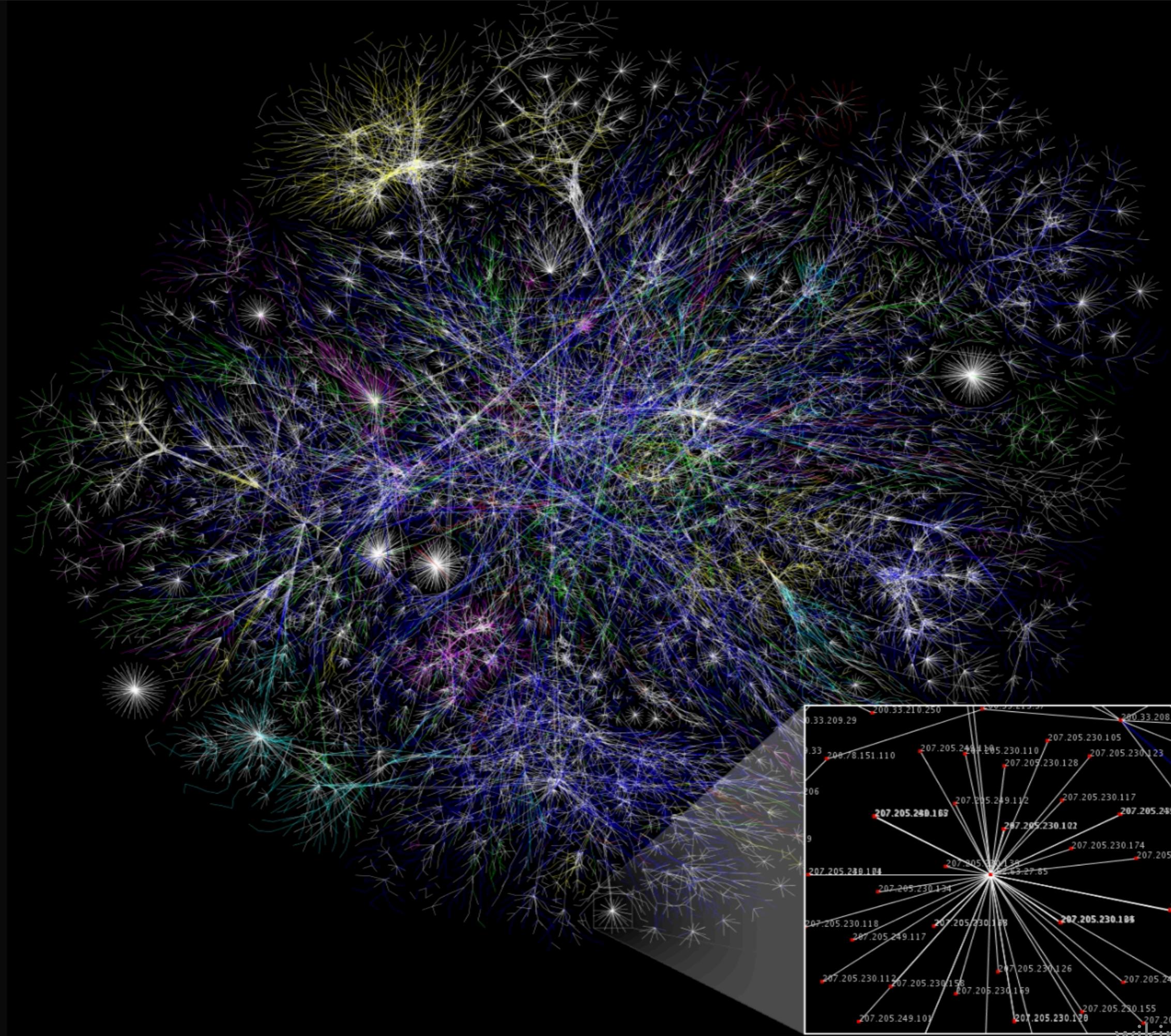
Back to interactive visualization

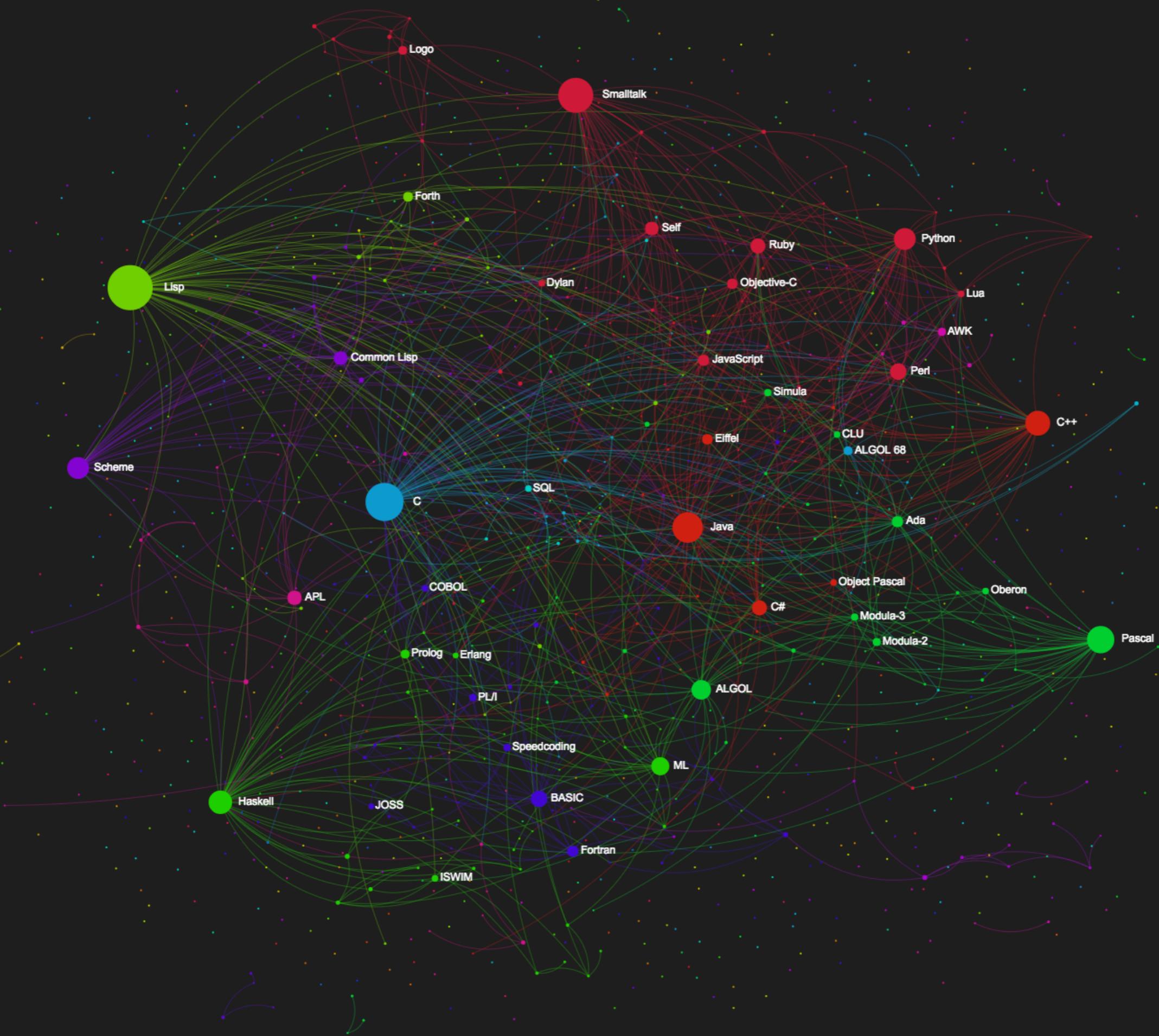


WHY NETWORKS?

- ▶ Networks are perfect for exploring relational data
 - ▶ Facebook friends
 - ▶ Kevin Bacon game
 - ▶ Erdöz number
- ▶ Visualise and analyse high-dimensional datasets with complex distribution
- ▶ Exploration of network topology can give insight into associations and interactions between the parts (nodes) in the network







BACKGROUND

npg

Top-down controls on bacterial communities
C-ET Chow et al

824

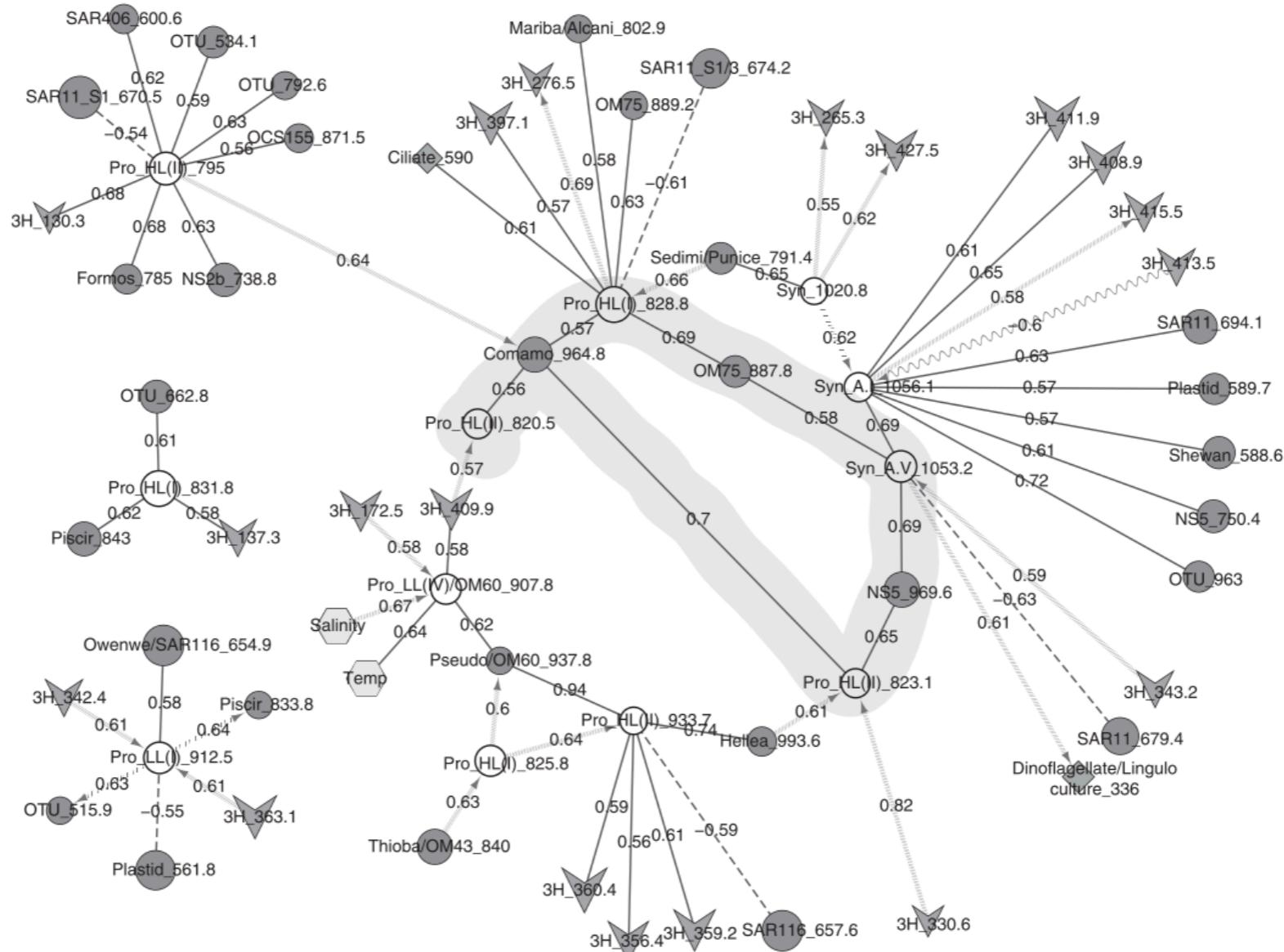


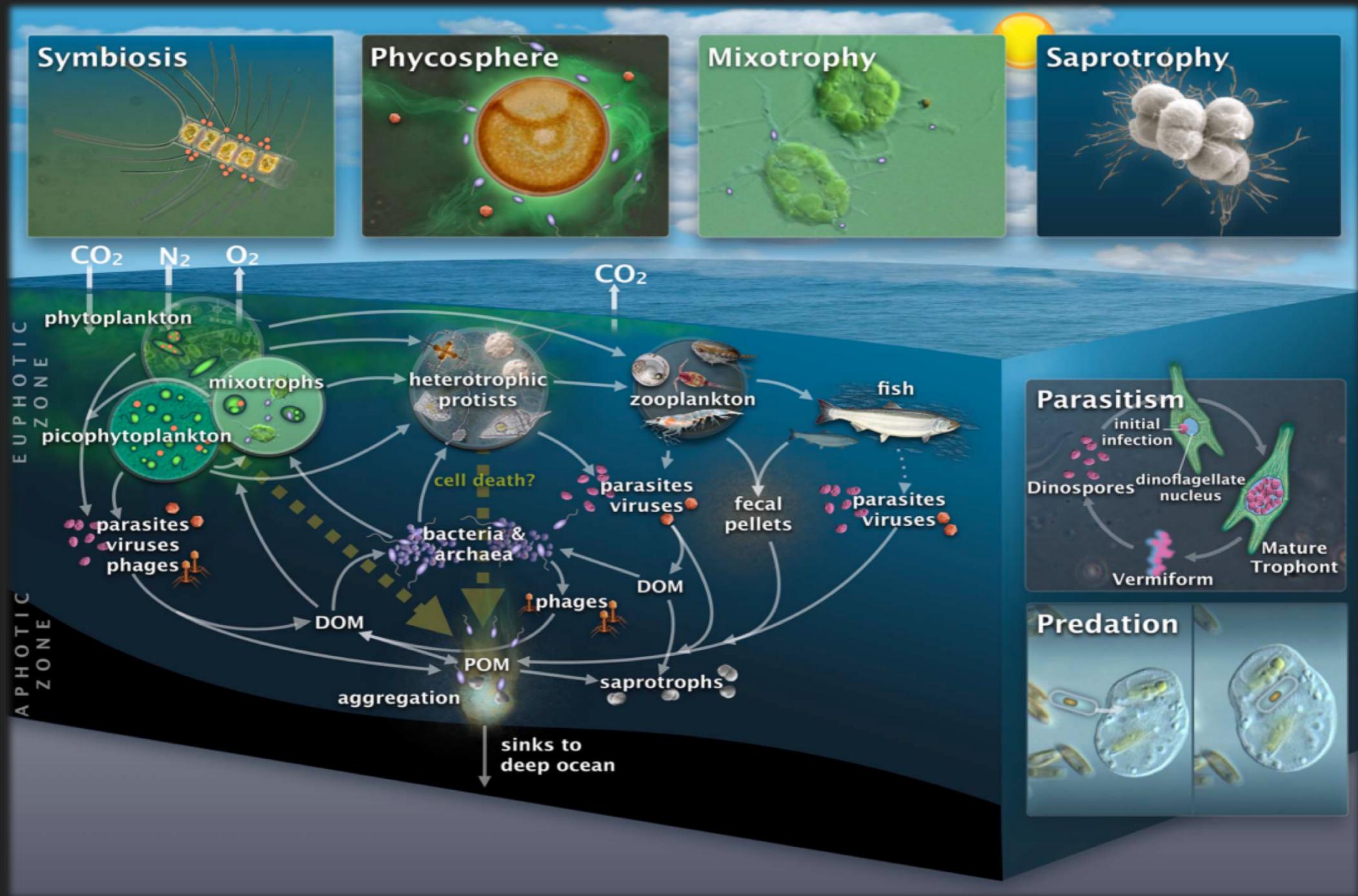
Figure 5 Cyanobacterial OTU correlations to other microbial OTUs reveal potential lytic virus–host relationships, grazing and temporal trends. *Cyanobacteria* OTUs are noted as white circles and labeled as *Prochlorococcus* (Pro) or *Synechococcus* (Syn), followed by ecotype designation (HL: high light; LL: low light; A/B: *Synechococcus* group). All other nodes are bacteria, circles; protists, diamonds; viruses, v-shapes; abiotic, hexagons. Node labels indicate an abbreviated identity (where available) and fragment length. Solid lines are positive correlations with no delay; dashed lines, negative correlations with no delay; sine-wave lines, negative-delayed correlations; and forward-slash lines, positive-delayed correlations. Arrows point toward the lagging OTU.

WHY NETWORKS?

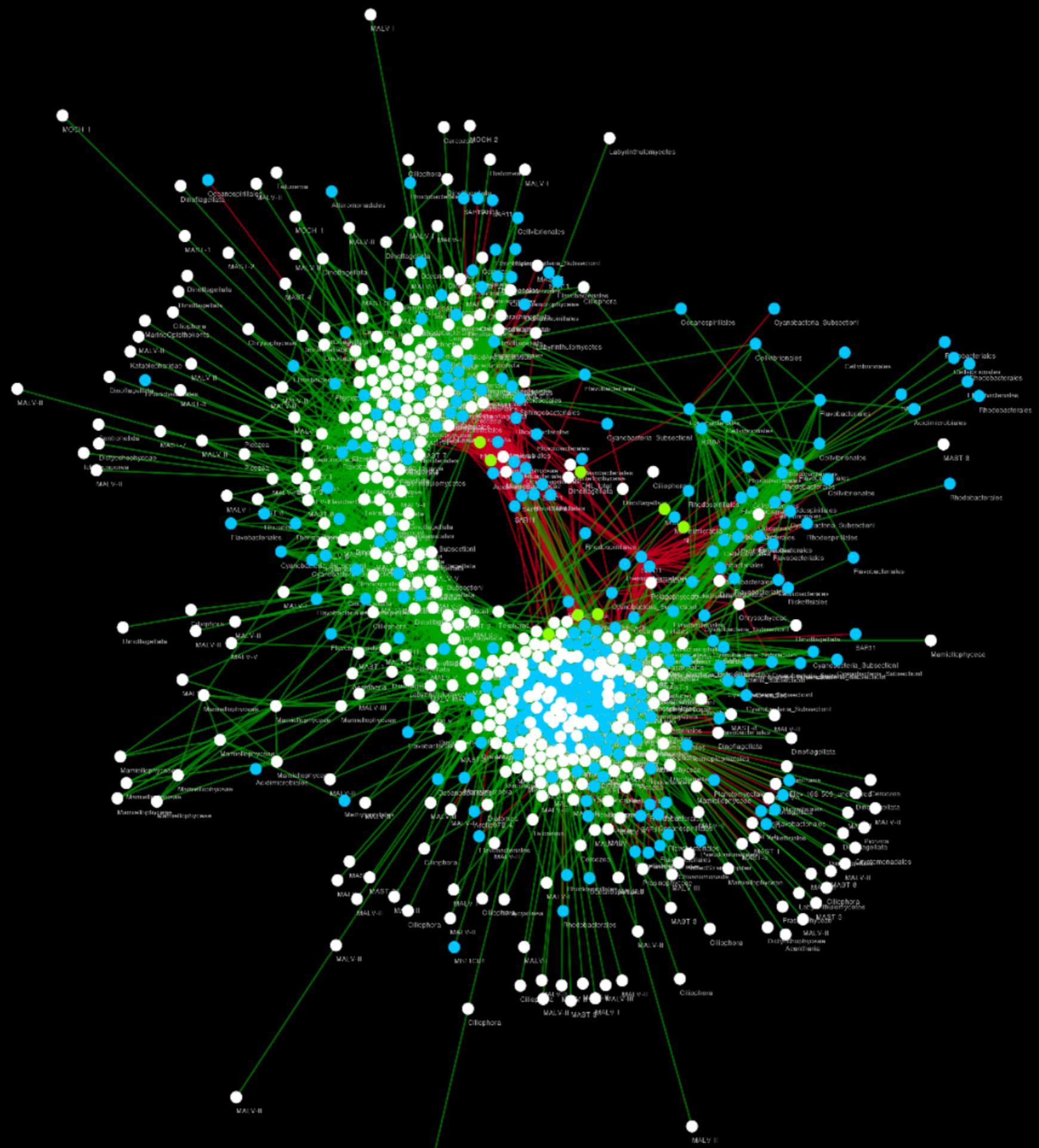
- ▶ For microbial data
- ▶ Networks can help us find patterns behind small- and large-scale ecological and evolutionary processes
- ▶ Moving from identification of parts (species classification, diversity, community composition, geographical distribution etc.) to connections of parts.

BACKGROUND

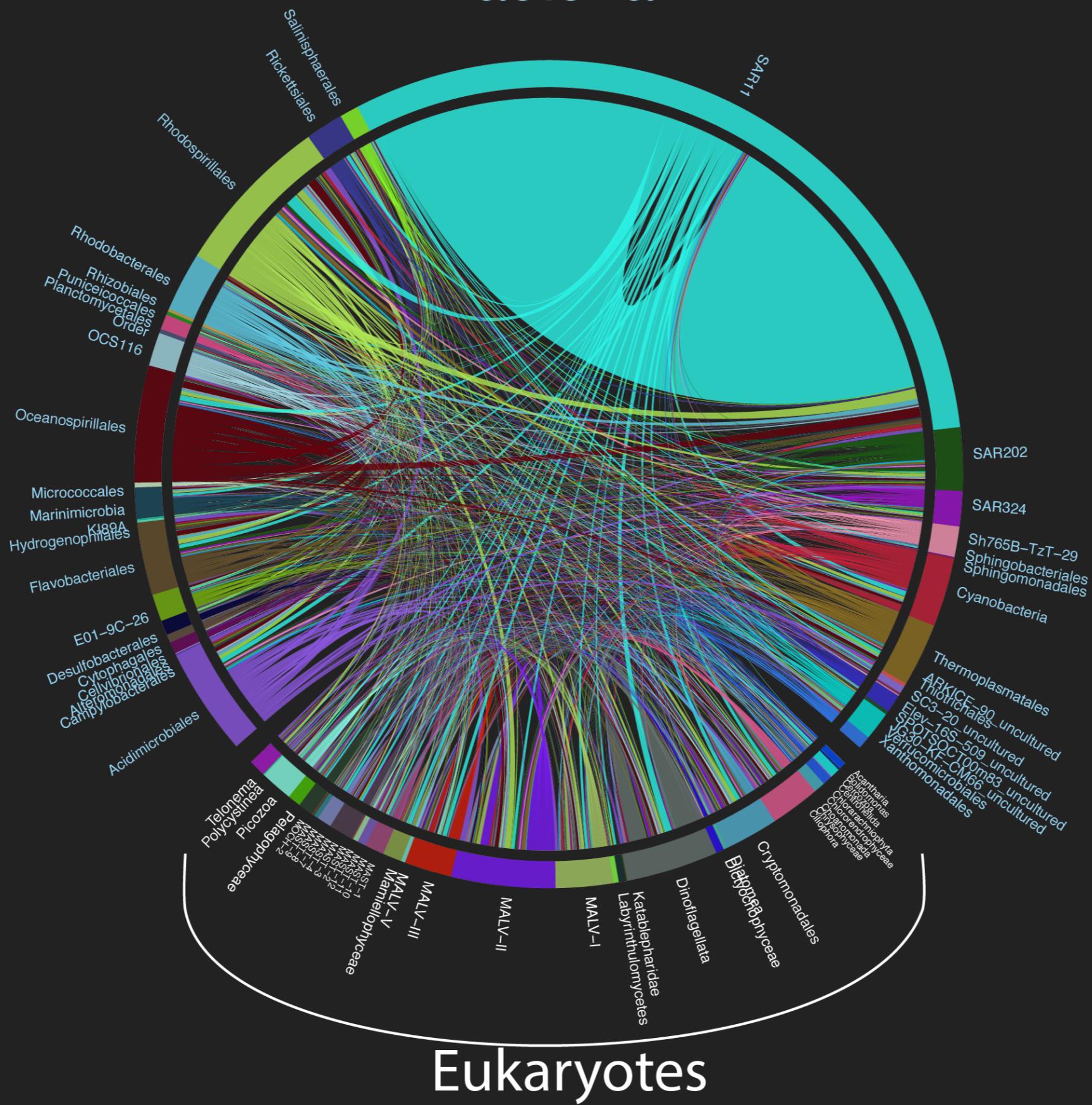
FOODWEBS



Worden et al (2015)



Bacteria

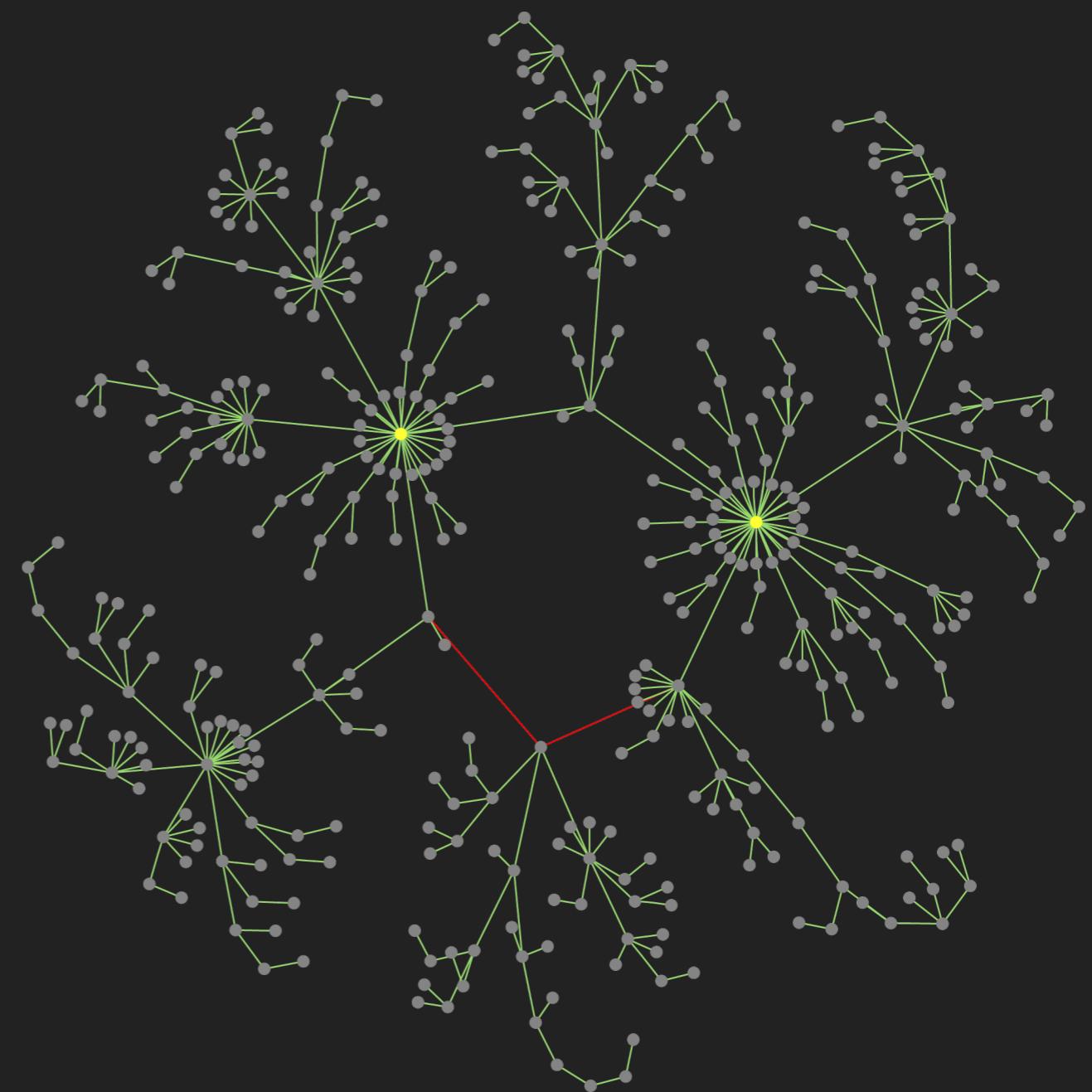


MICROBIAL NETWORKS

- ▶ Associations/co-occurrence between taxa can reveal:
 - ▶ similar response to **environmental** variation
 - ▶ similar niche preference
 - ▶ **direct interactions** between organisms, e.g. host/symbiont, prey/predator
 - ▶ important relationships in the community
 - ▶ key species (**hubs**) and assemebalges of interacting species (**modules**)

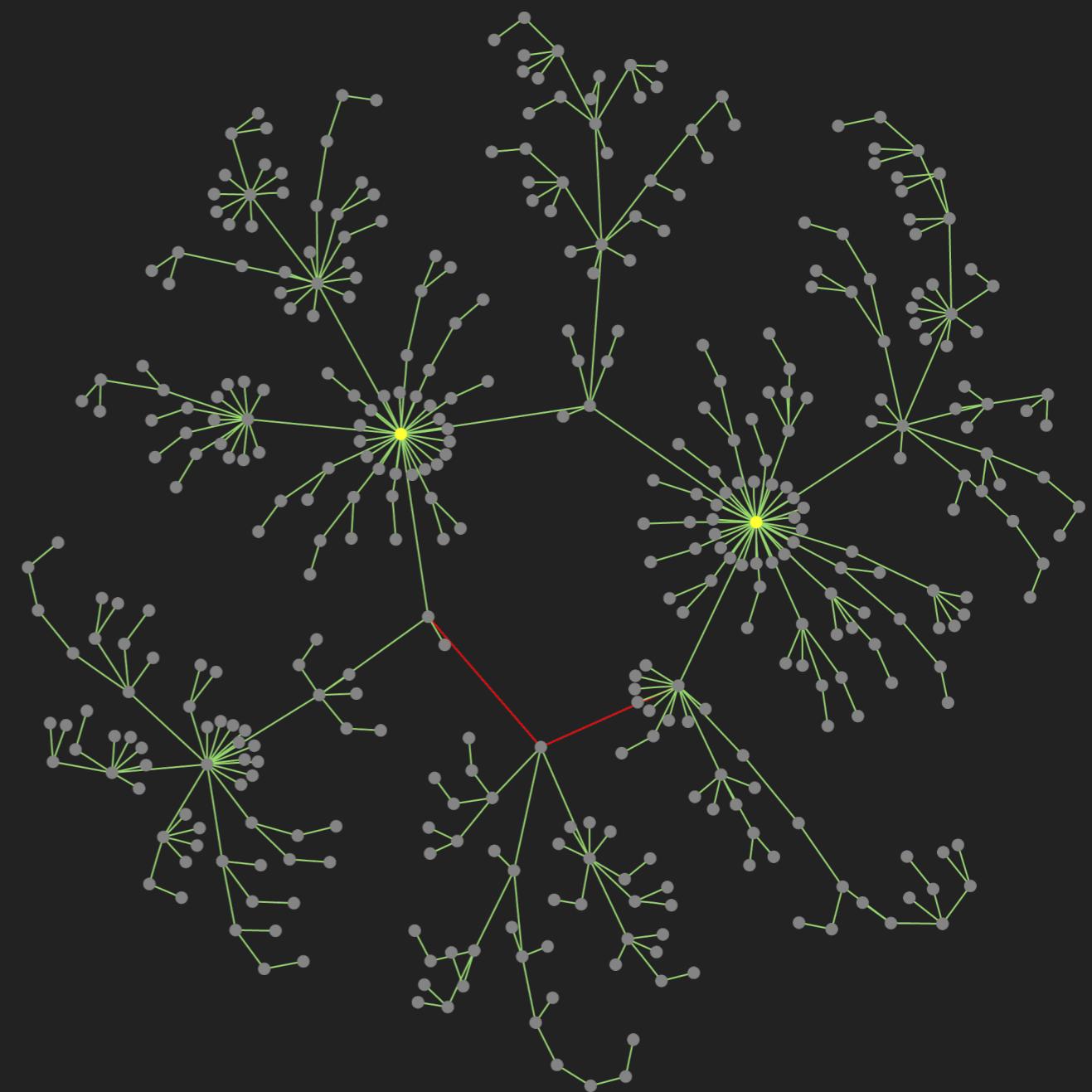
NETWORK FEATURES – SOME TERMINOLOGY

- ▶ **Nodes (dots)** represents OTUs and environmental variables
- ▶ **Edges (lines)** connects the OTUs, signifying co-occurrence (or interaction)
- ▶ By analysing the distribution of the nodes, the edges and the topology of the network we can hopefully say something about the co-occurrence or interactions between species in the network.



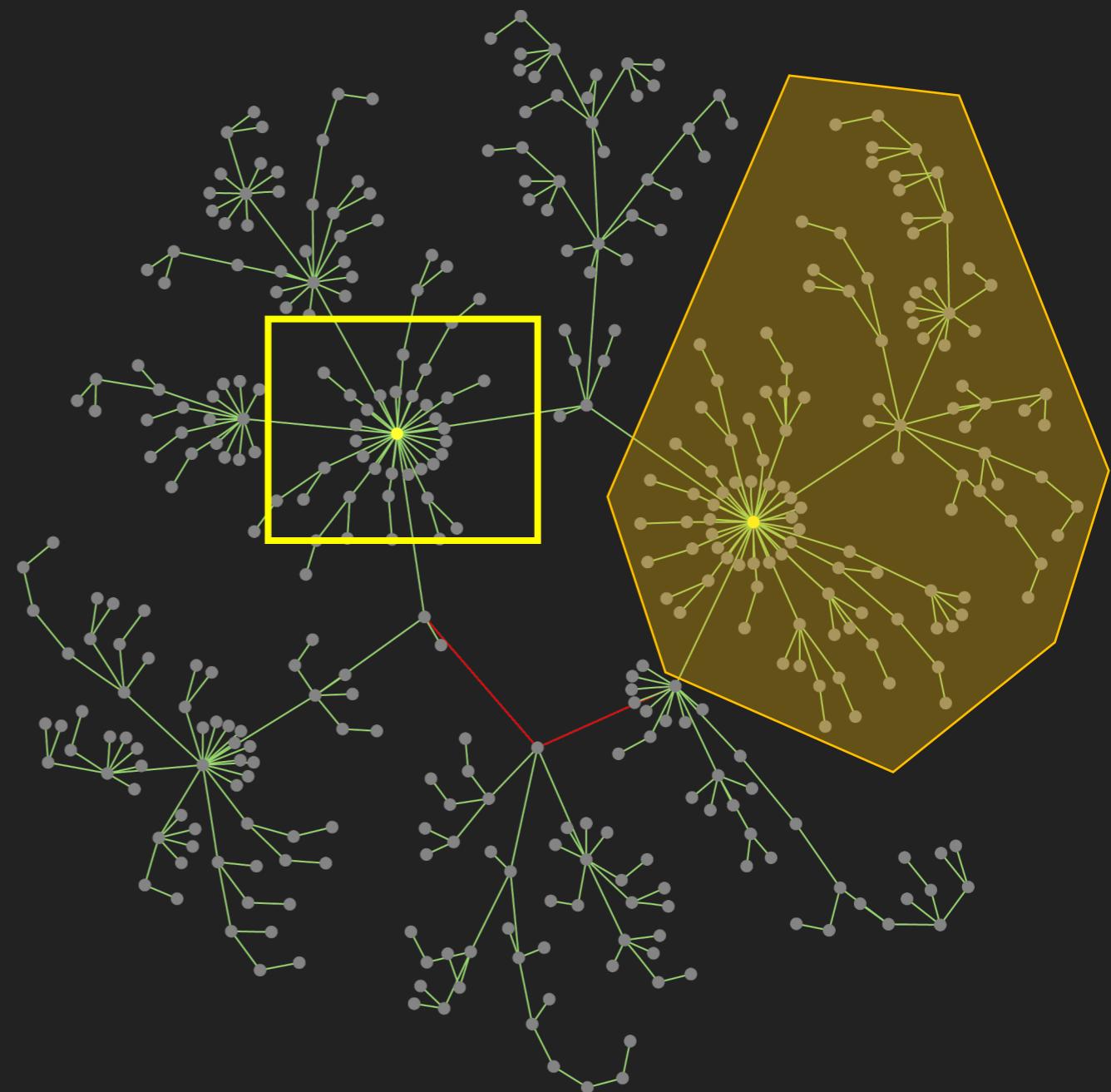
NETWORK FEATURES – SOME TERMINOLOGY

- ▶ **Association network:** network built using positive or negative correlation between taxa
- ▶ **Degree:** number of edges connected to a node
- ▶ **Degree distribution:** cumulative distribution of node's degrees
- ▶ **Distance:** shortest path between two nodes
- ▶ **Characteristic path length:** mean number of steps along the shortest paths for all node-pair in a network (distance average)
- ▶ **Clustering coefficient:** a measure of the clustering of nodes in a network.



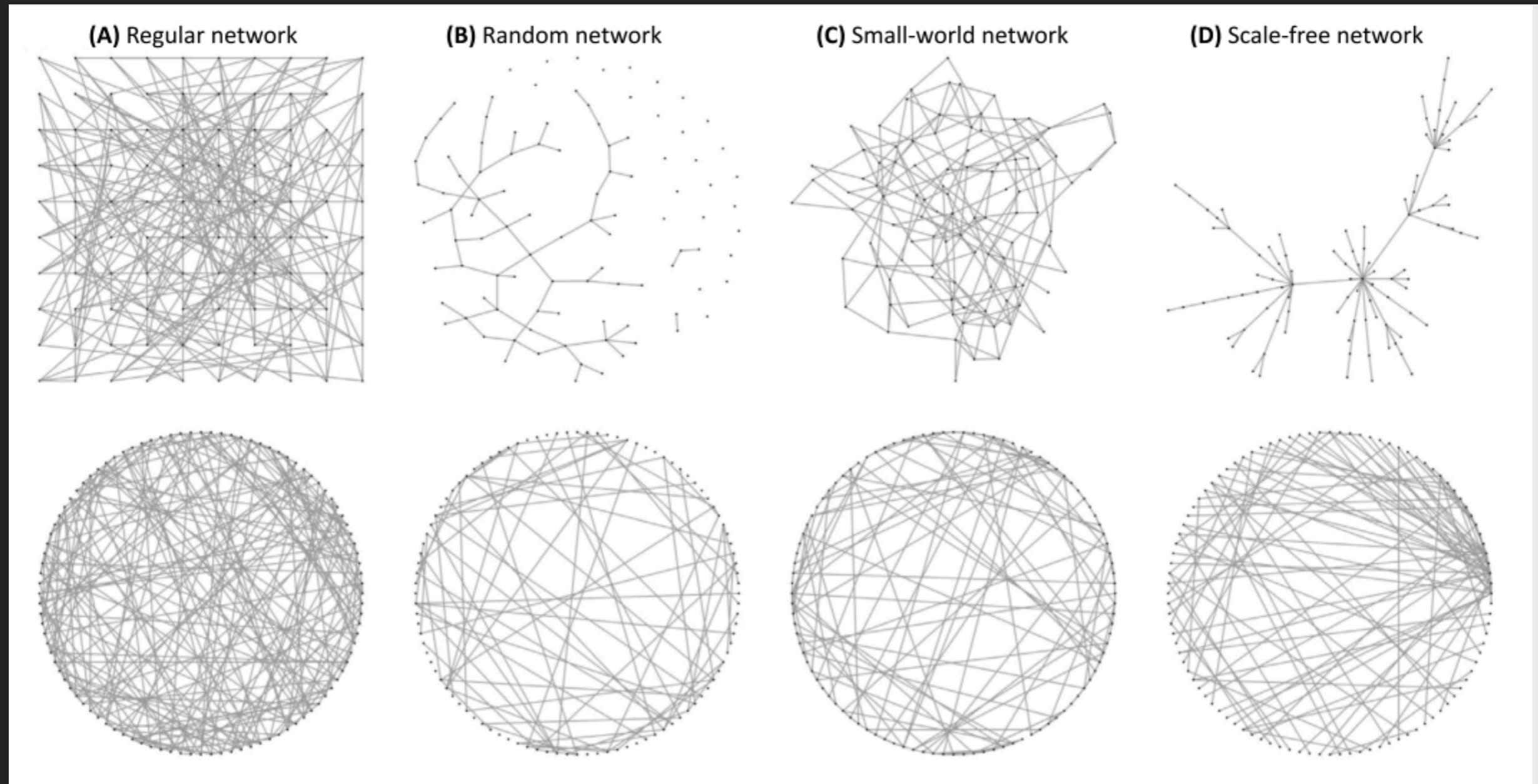
NETWORK FEATURES – SOME TERMINOLOGY

- ▶ **Hubs** are highly connected nodes, often key species in the network.
- ▶ **Modules:** sets of nodes that are more linked to each other than the rest of the network.



BACKGROUND

NETWORK TOPOLOGY



NETWORKS CONSTRUCTION

- ▶ Methods for network construction are many and varies in speed, accuracy and application,
 - ▶ for more see Faust et al (2012), Layeghifard et al. (2017)
 - ▶ Dissimilarity based, for instance Bray-Curtis
 - ▶ Correlation-based; Pearson, Spearman
 - ▶ Regression-based
 - ▶ Probabilistic Graphical Models

SOFTWARE

- ▶ Software for inferring networks
 - ▶ SparCC (Friedman and Alm, 2012) - For compositional data
 - ▶ CoNet (Faust et al. 2012)
 - ▶ WGCNA - Weighted correlation network analysis, often used to find clusters of co-expressed genes. (Langfelder and Horvath, 2008)
 - ▶ eLSA - Extended local similarity analysis (Ruan et al. 2006, Xia et al 2015)
 - For time series
 - ▶ etc.
 - ▶ etc.

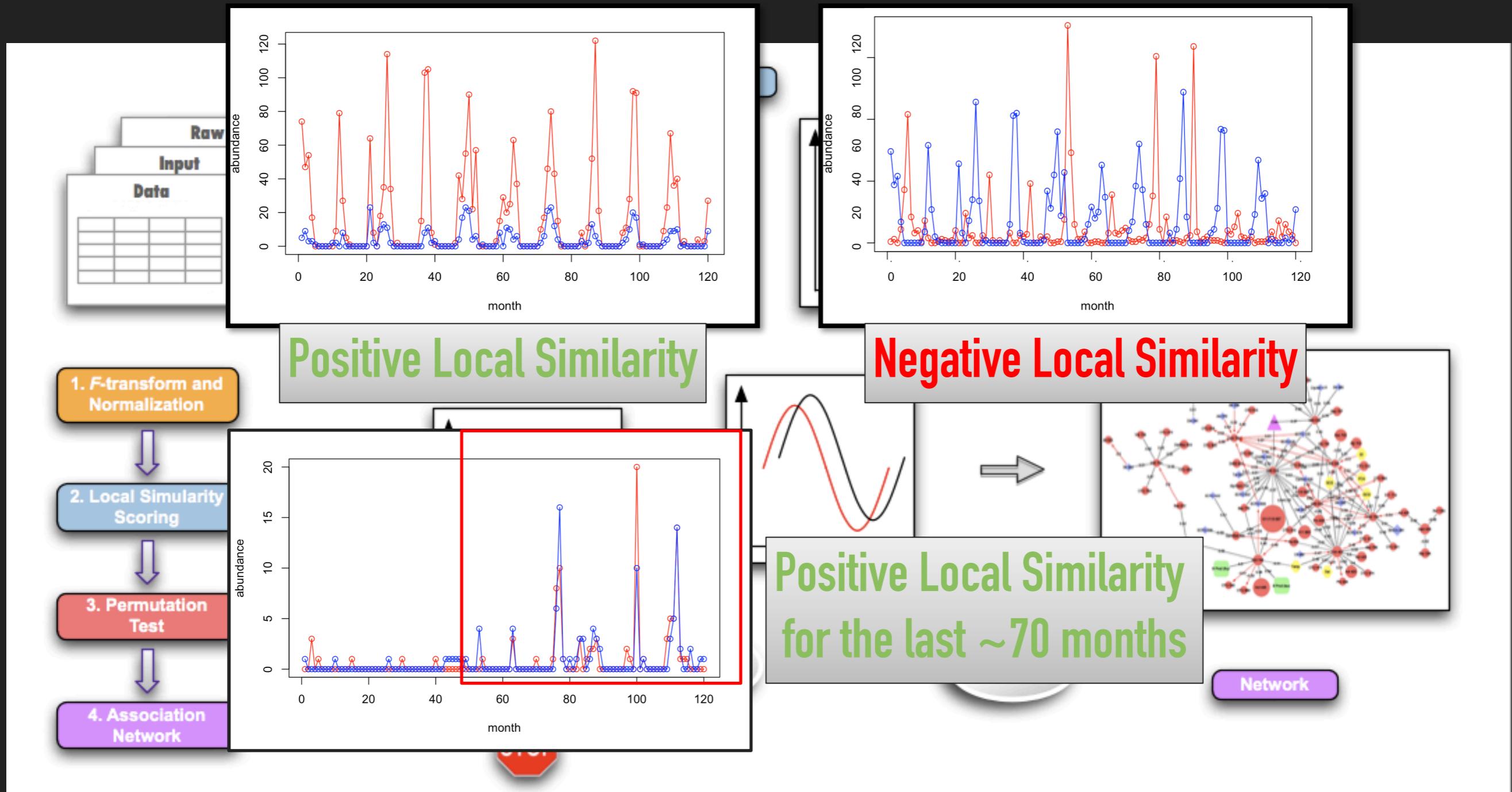
TYPES OF NETWORKS

- ▶ Spatial sampling:
 - ▶ larger geographic coverage but unless samples are taken at the same time, each sample is a snapshot.
- ▶ Temporal
 - ▶ Smaller geographic coverage, but multiple data points for the same location. Allows following microbial communities, and their changes, in time.

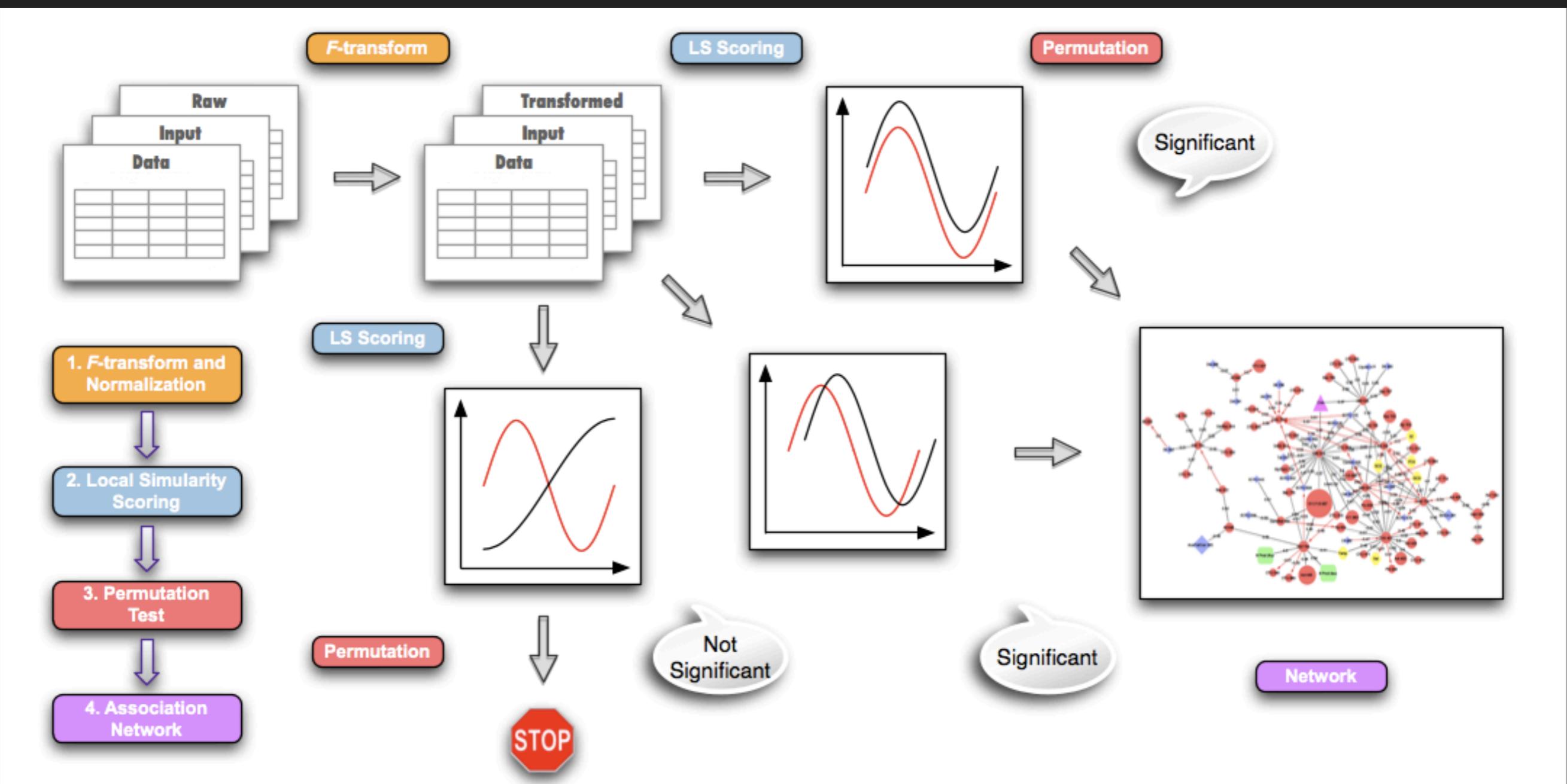
ELSA – EXTENDED LOCAL SIMILARITY ANALYSIS

- ▶ Designed for time series
- ▶ Pairwise comparison between OTUs (and OTUs and environmental variables; n)
a total of $n(n-1)/2$ comparisons (edges)
- ▶ Multiple co-occurrences and co-exclusions in the same community
- ▶ Can incorporate time delays (the maximum allowed time delay has to be specified)
- ▶ <https://bitbucket.org/charade/elsa/wiki/FAQ.wiki>

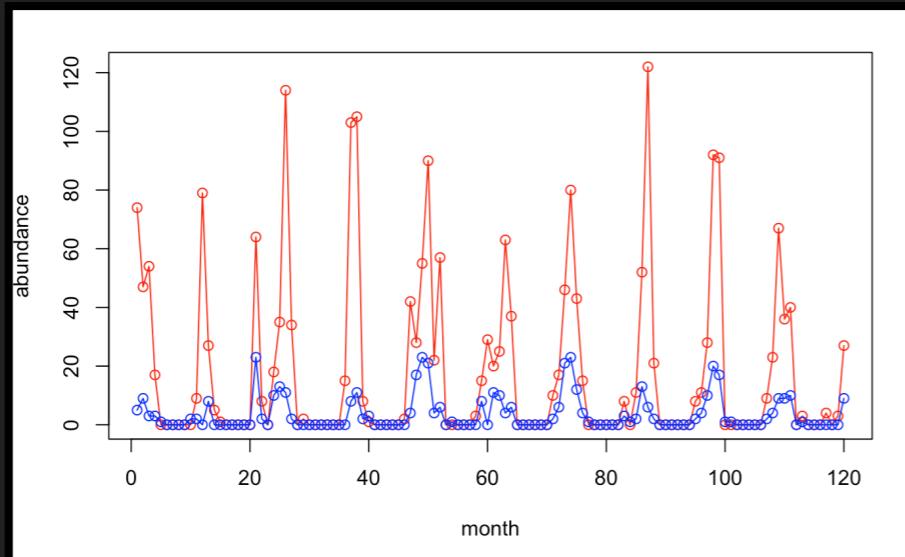
WORKFLOW



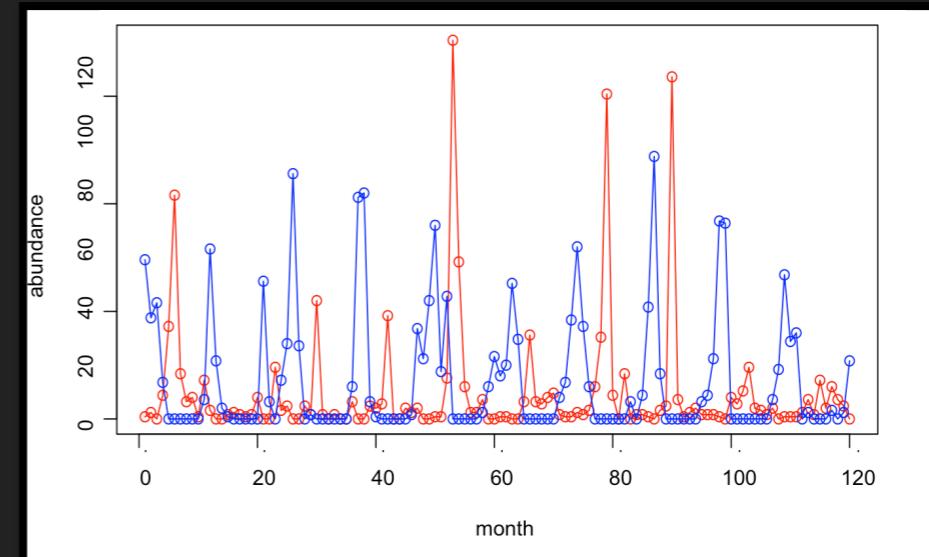
WORKFLOW



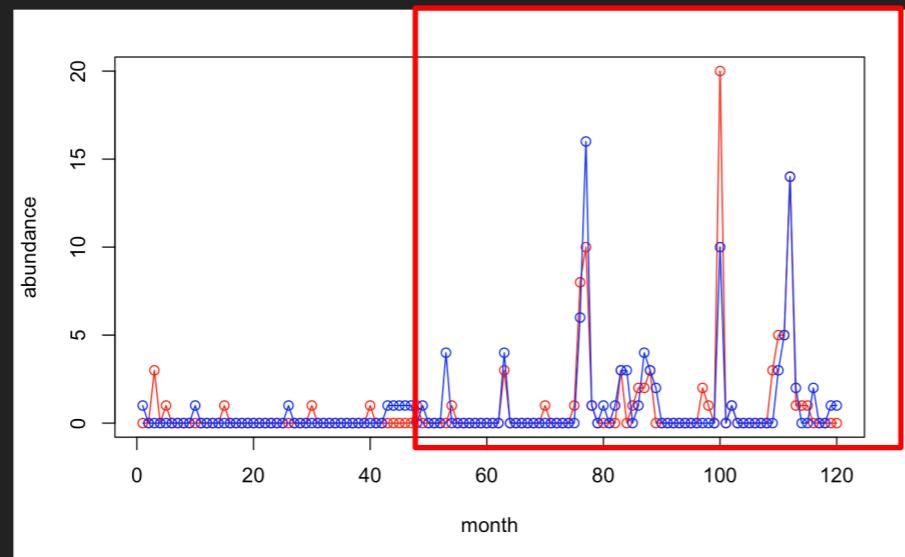
WORKFLOW



Positive Local Similarity



Negative Local Similarity



Positive Local Similarity
for the last ~70 months

ELSA - THE PARAMETERS

- ▶ Specify the number of time spots (*s*) and if you have replicates (*r*)
- ▶ Normalization method (*n*; default robustZ)
- ▶ Delay-limit (*d*; default 3)
- ▶ P-value calculation (*p*; perm, theo or mix)
- ▶ Permutation /precision (*x*; the number of permutations)
- ▶ Missing values (*f*; none, zero, linear, quadratic...)
- ▶ If you have replicates you can set the number of bootstraps for 95% confidence interval estimation (*b*; default 100)
- ▶ Example command for 24 months
- ▶ `lsa_compute <in> <out> -r 1 -s 24 -d 0 -p perm-x 1000 -f linear -n robustZ`

ELSA - THE OUTPUT

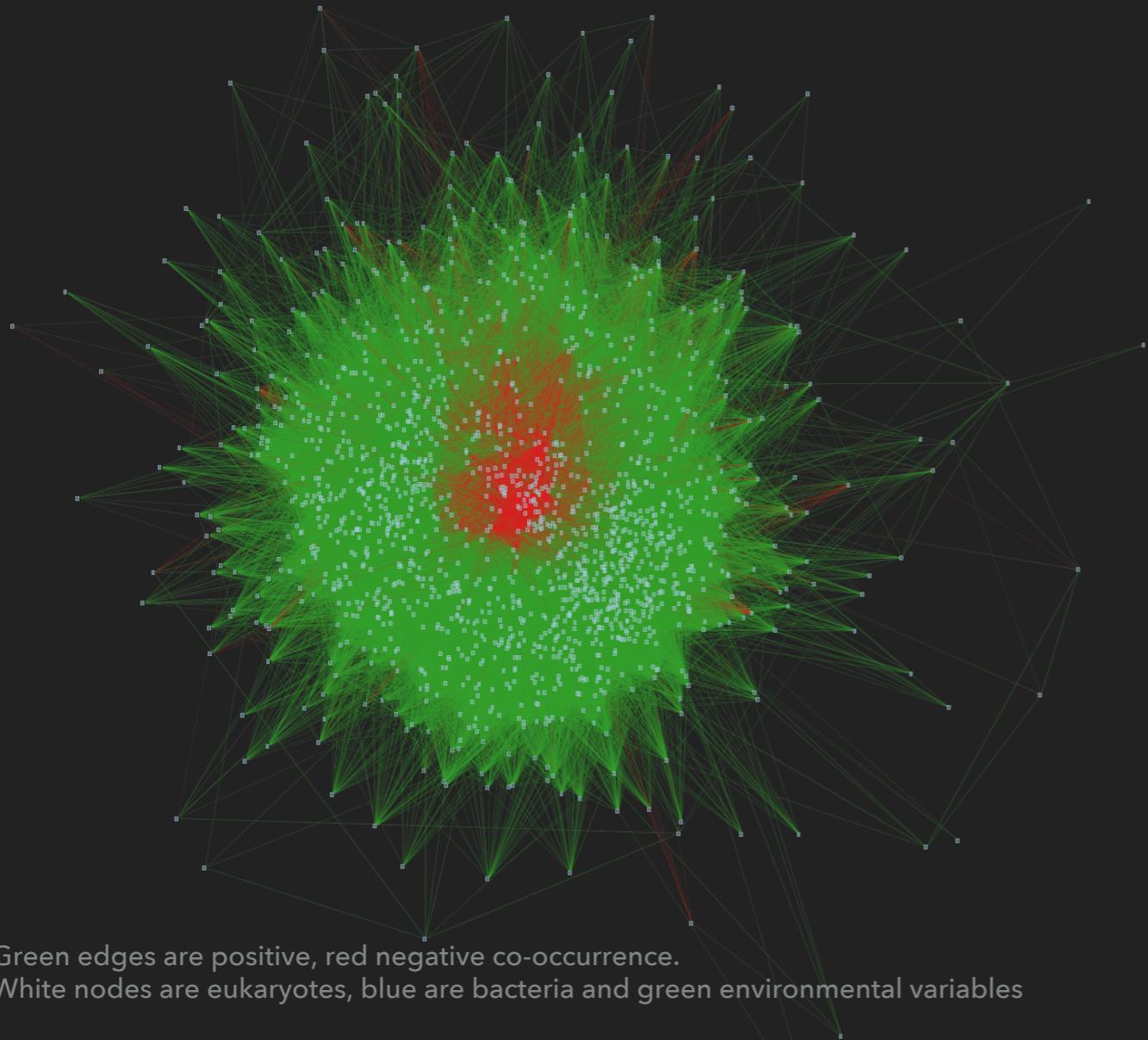
- ▶ X: The name of the first OTU/variable in the pairwise comparison
- ▶ Y: The name of the second OTU/variable in the pairwise comparison
- ▶ LS: Local similarity score. It is defined as "For two normal transformed sequences of the same length, the local similarity score is defined as the maximal sum of the product of the corresponding entries of all their subsequences within some predefined time delay D." Ruan et al 2006.
- ▶ P and Q values, Spearman
- ▶ etc.
- ▶ An explanation of the output is on Github.

BLANES BAY DATA – 10YEARS MONTHLY

- Results from pico fraction (<3 um)
 - Illumina amplicons from 18S (V4) and 16S (V3)
- 16 Millions sequences
- Clustered at 99%
 - Bacteria: 16 196 OTUs
 - Eukaryotes: 36 020 OTUs
- Rarified
- Removed OTUs that were not present in at least 10% of the samples
- 515 bacteria OTUs, 1065 eukaryote OTUs, and 18 Environmental var.

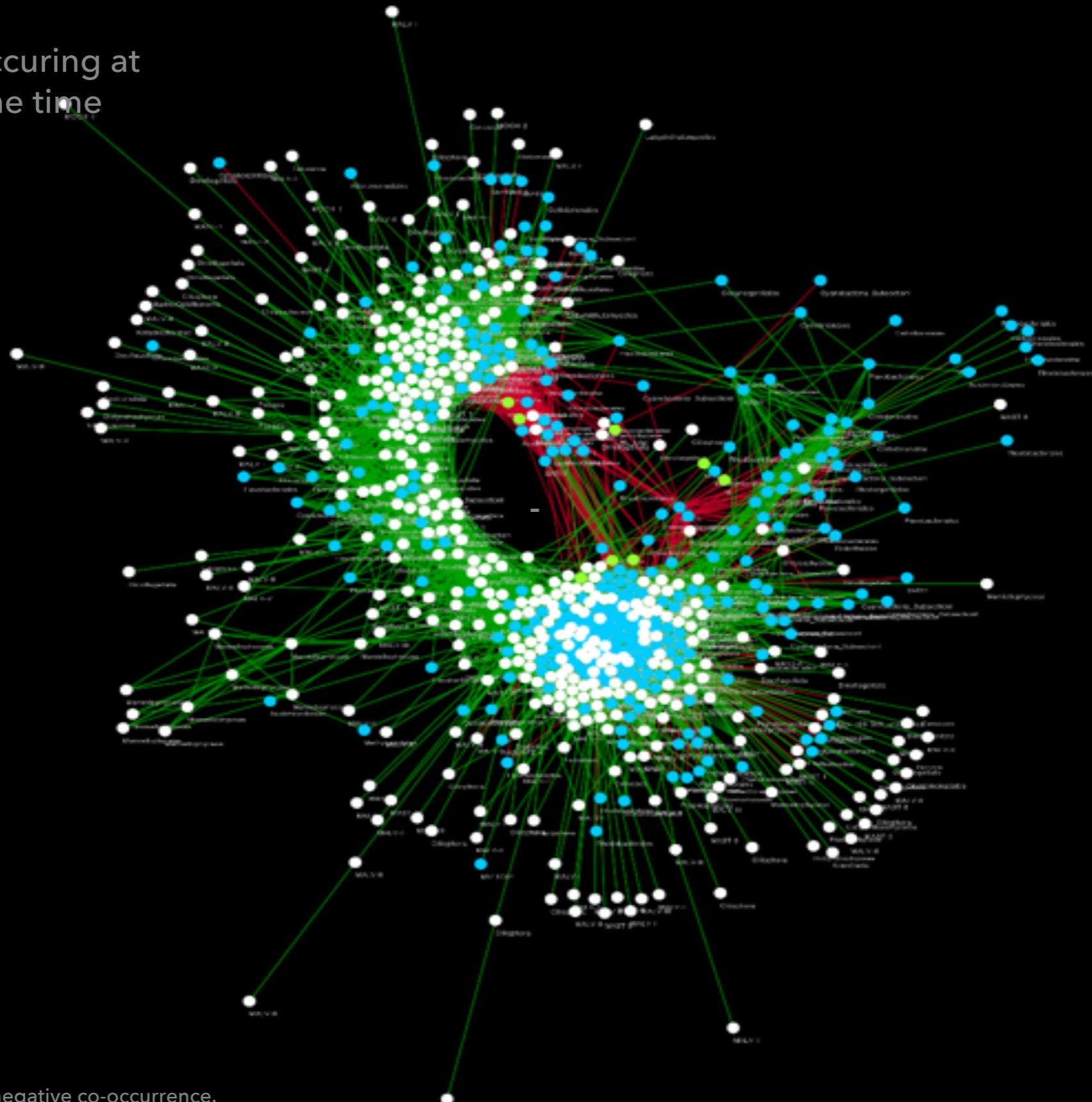
RESULTS - EXAMPLE

- With 1598 OTUs and ENVs for pairwise comparisons there are more than 1.2 million possible edges in the network.
- After removing non-significant edges ($P & Q < 0.05$)
 - 222199 edges (17,4 %)
 - **197541 positive** (88%) and **24658 negative** (12%)



- Which are the most important species, the core community?
- Some of the edges are due to organisms reacting to the environment.
- What about interactions between organisms (symbiosis, parasitism, predation) ?

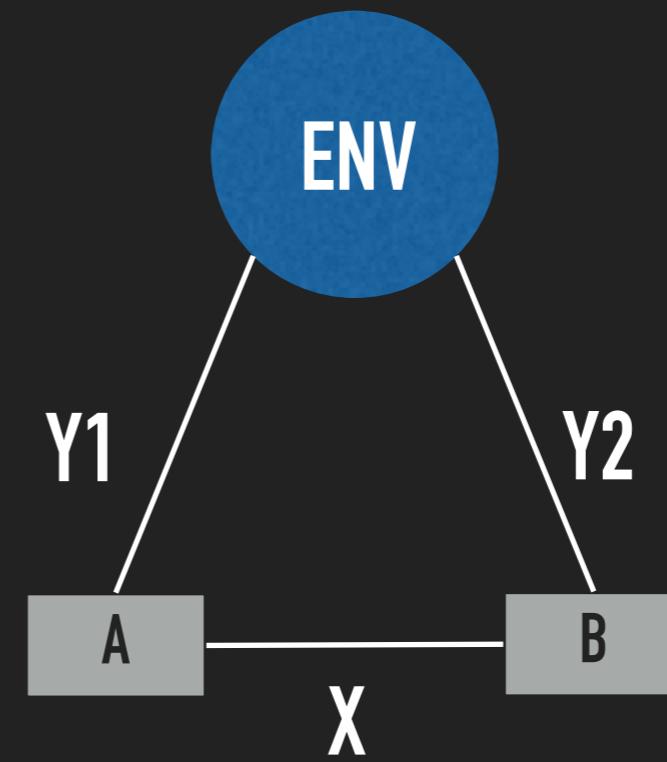
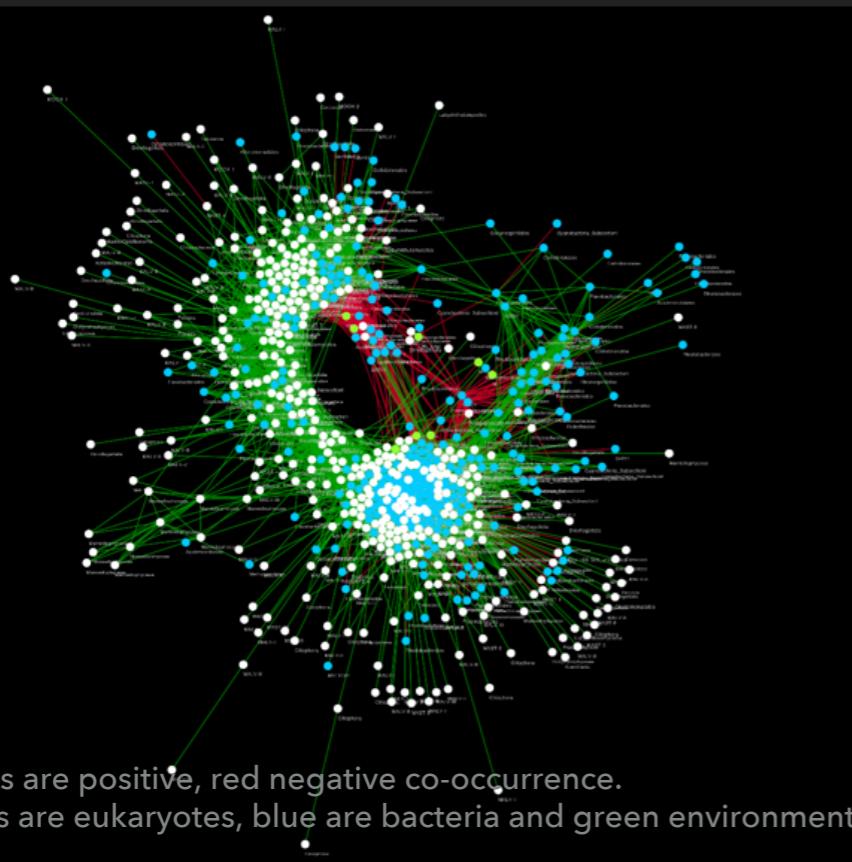
- Interactions occurring at least 60% of the time
- P&Q<0.0001



Green edges are positive, red negative co-occurrence.
White nodes are eukaryotes, blue are bacteria and green environmental variables

PRUNING INDIRECT EDGES

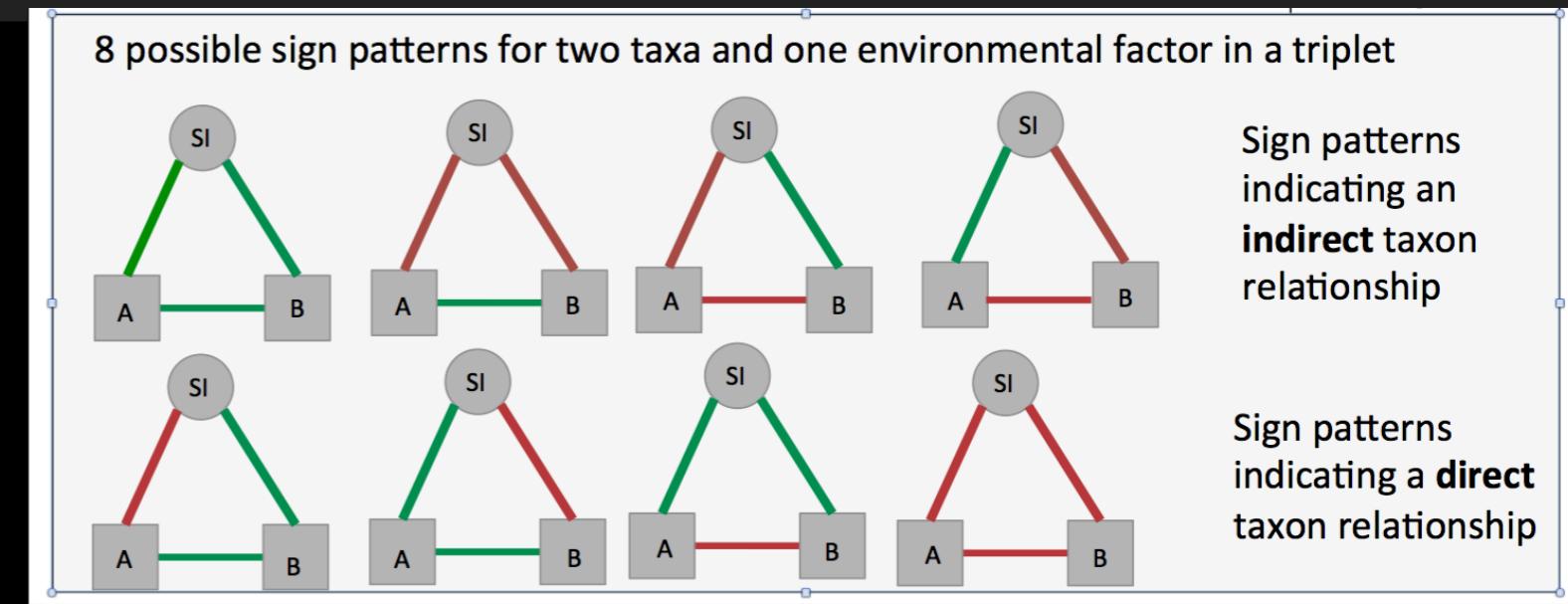
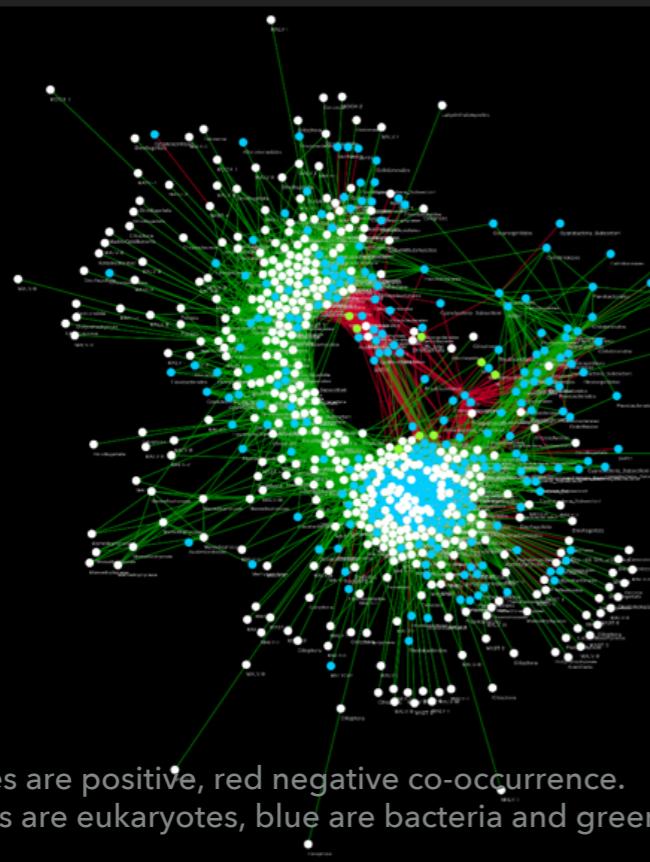
- Remove environment-driven co-occurrence
- Identification of triplets were the environmental factor is stronger than OTU-OTU interactions



Ina Deutschmann

PRUNING INDIRECT EDGES

- Remove environment-driven co-occurrence
- Identification of triplets where the environmental factor is stronger than OTU-OTU interactions

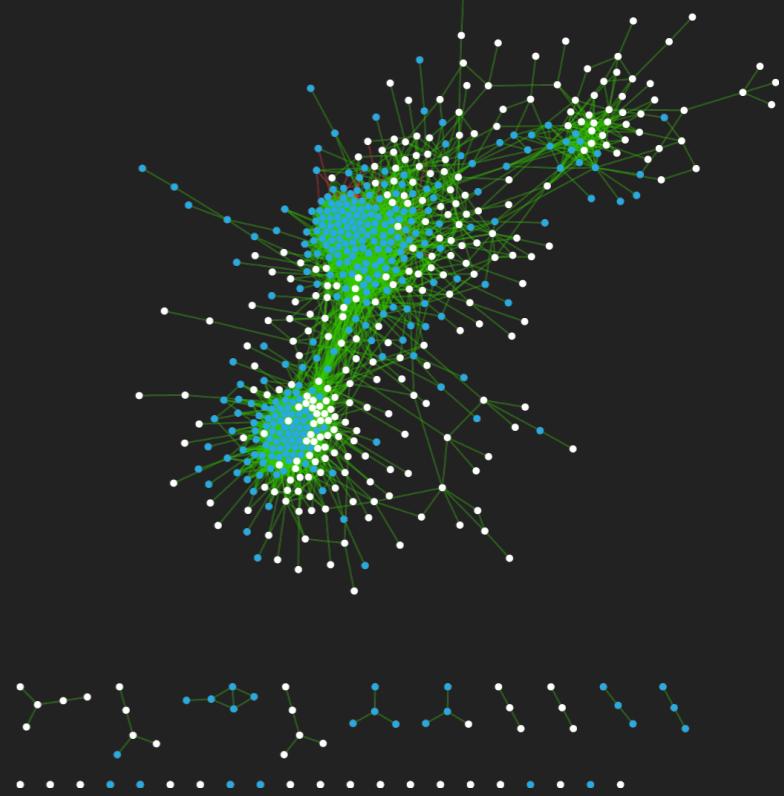
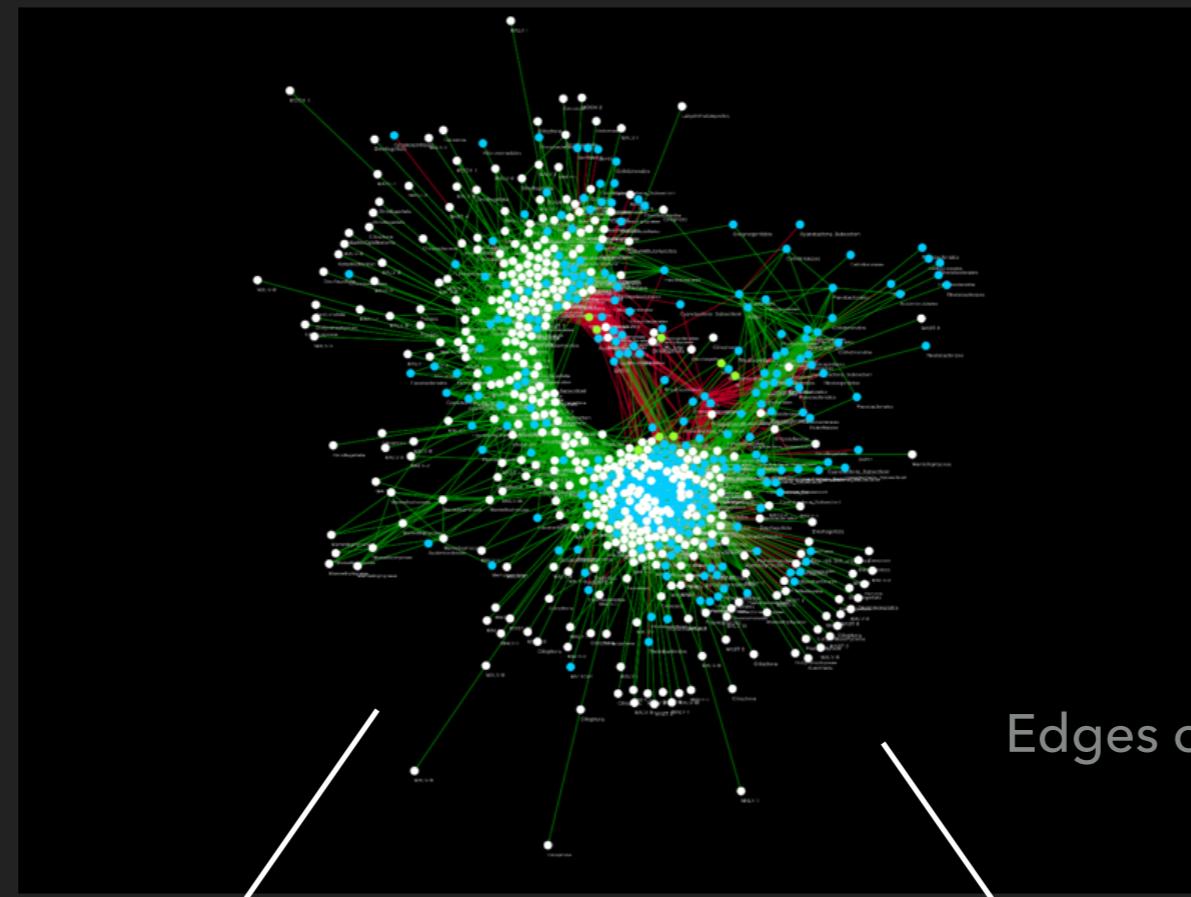
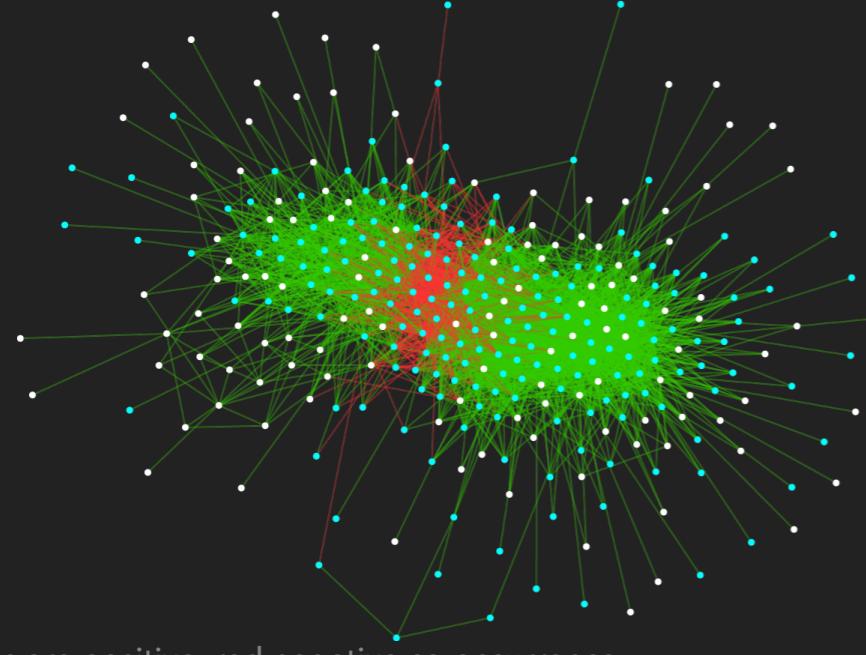


Lima-Mendez et al. 2015, Karoline Faust

Green edges are positive, red negative co-occurrence.
White nodes are eukaryotes, blue are bacteria and green environmental variables

Edges that depend on the environment

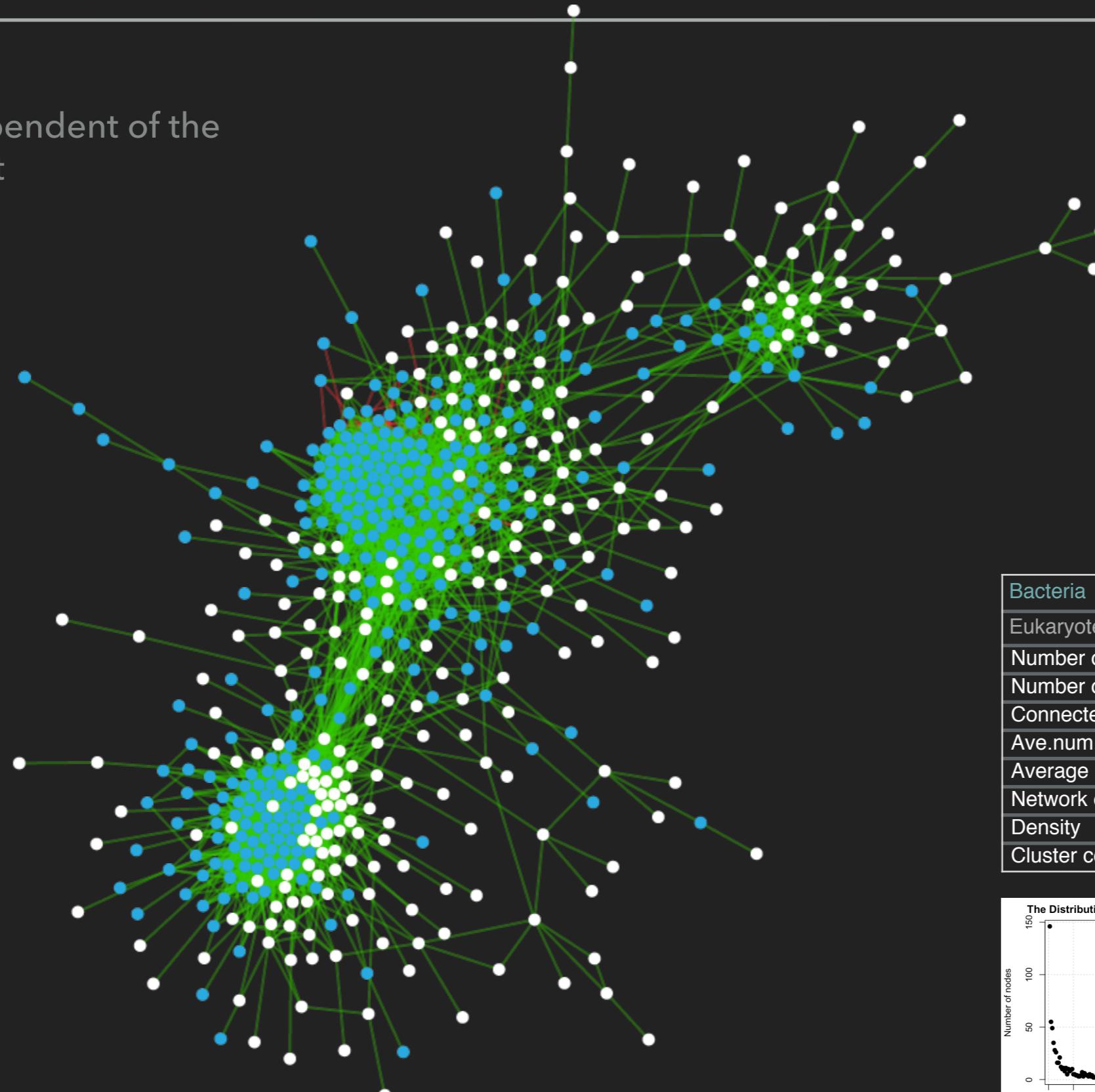
Edges of potential interactions



Green edges are positive, red negative co-occurrence.

White nodes are eukaryotes, blue are bacteria and green environmental variables

Edges independent of the environment

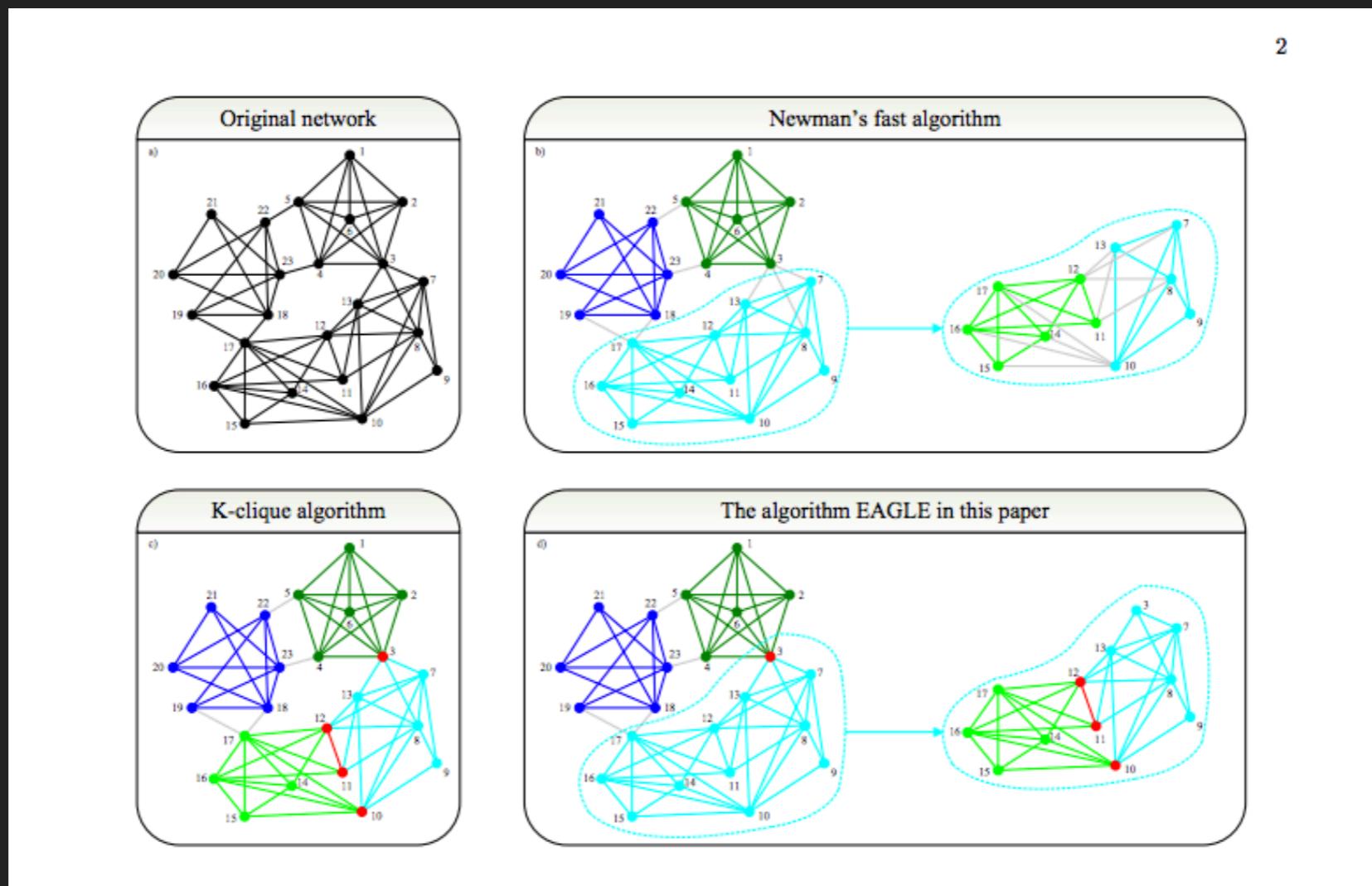


Green edges are positive, red negative co-occurrence.

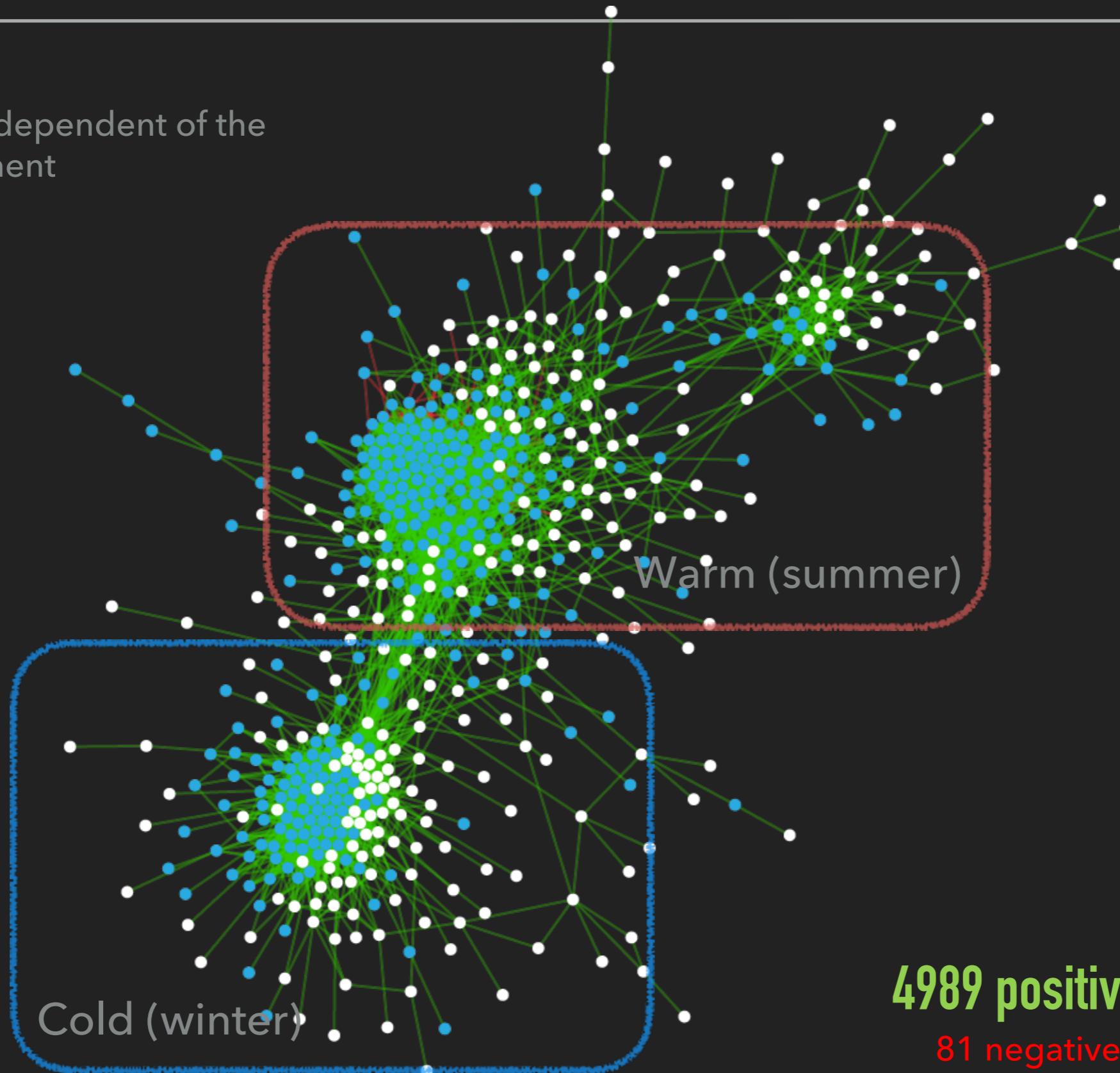
White nodes are eukaryotes, blue are bacteria and green environmental variables

FINDING MODULES

- ▶ MCODE (Bader et al 2002), Eagle, K-clique...



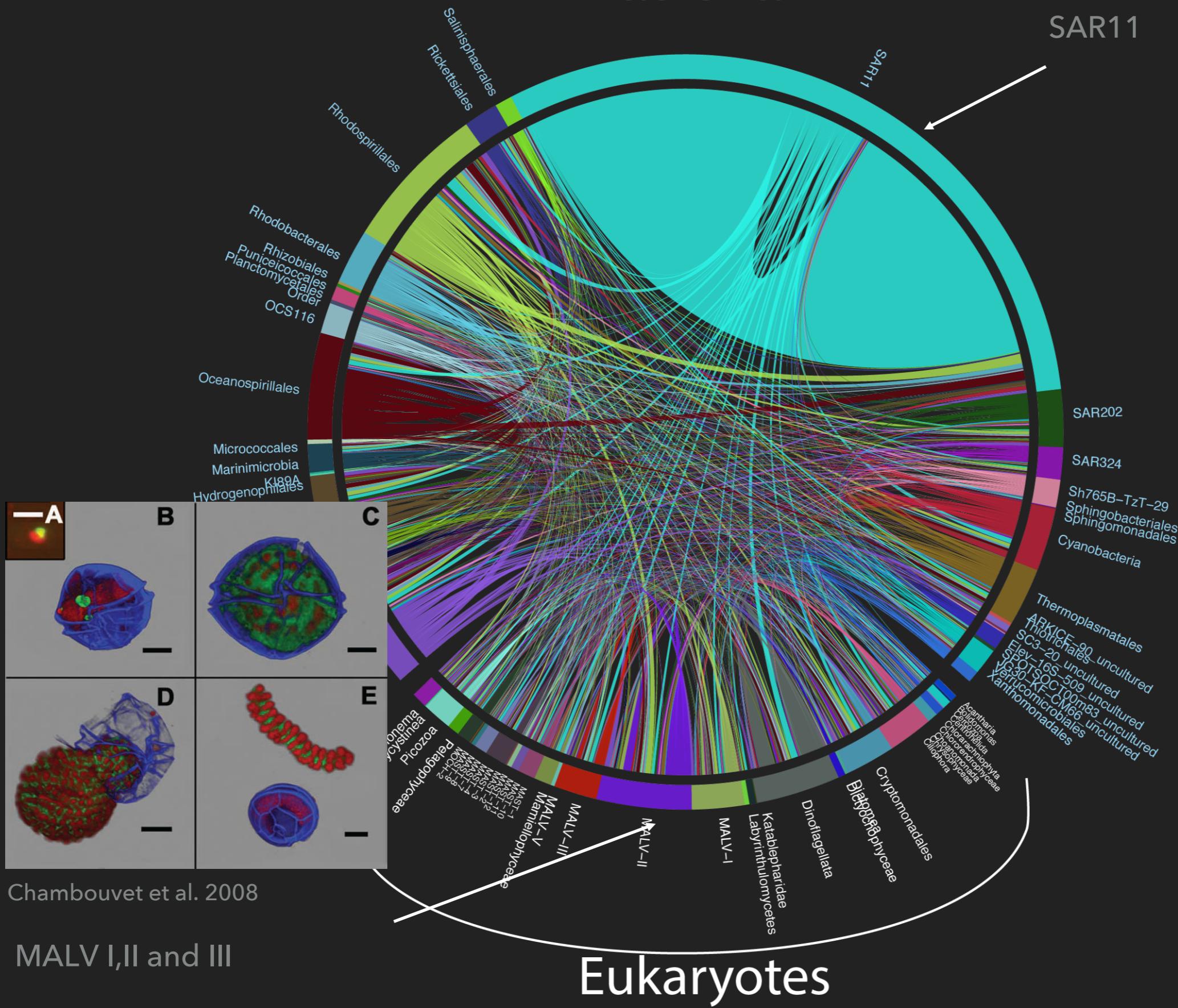
Edges independent of the environment



Green edges are positive, red negative co-occurrence.
White nodes are eukaryotes, blue are bacteria and green environmental variables

Bacteria

SAR11



PROGRAMS FOR VISUALIZATIONS

- ▶ Cytoscape (<http://www.cytoscape.org/>)
- ▶ Gephi
- ▶ Several packages for R
 - ▶ iGraph
 - ▶ ProNet
 - ▶ circlize

SOME RESOURCE

- ▶ The code repository for Schmidt et. al 2017 "A Family of Interaction-Adjusted Indices of Community Similarity" doi:10.1038/ismej.2016.139
 - ▶ https://github.com/defleury/Schmidt_et_al_2016_community_similarity/
- ▶ Tutorial for WGCNA:
<https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html>
- ▶ eLSA: <https://bitbucket.org/charade/elsa/wiki/Home>
- ▶ fast eLSA: <http://www.cmde.science.ubc.ca/hallam/fastLSA/>
- ▶ SparCC: a python module for computing correlations in compositional data (16S, metagenomics, etc'). <https://bitbucket.org/yonatanf/sparcc>
- ▶ iGraph package for R:
 - ▶ <https://www.r-bloggers.com/an-example-of-social-network-analysis-with-r-using-package-igraph/>

CYTOSCAPE – SHORT TUTORIAL

THE NETWORK FILES

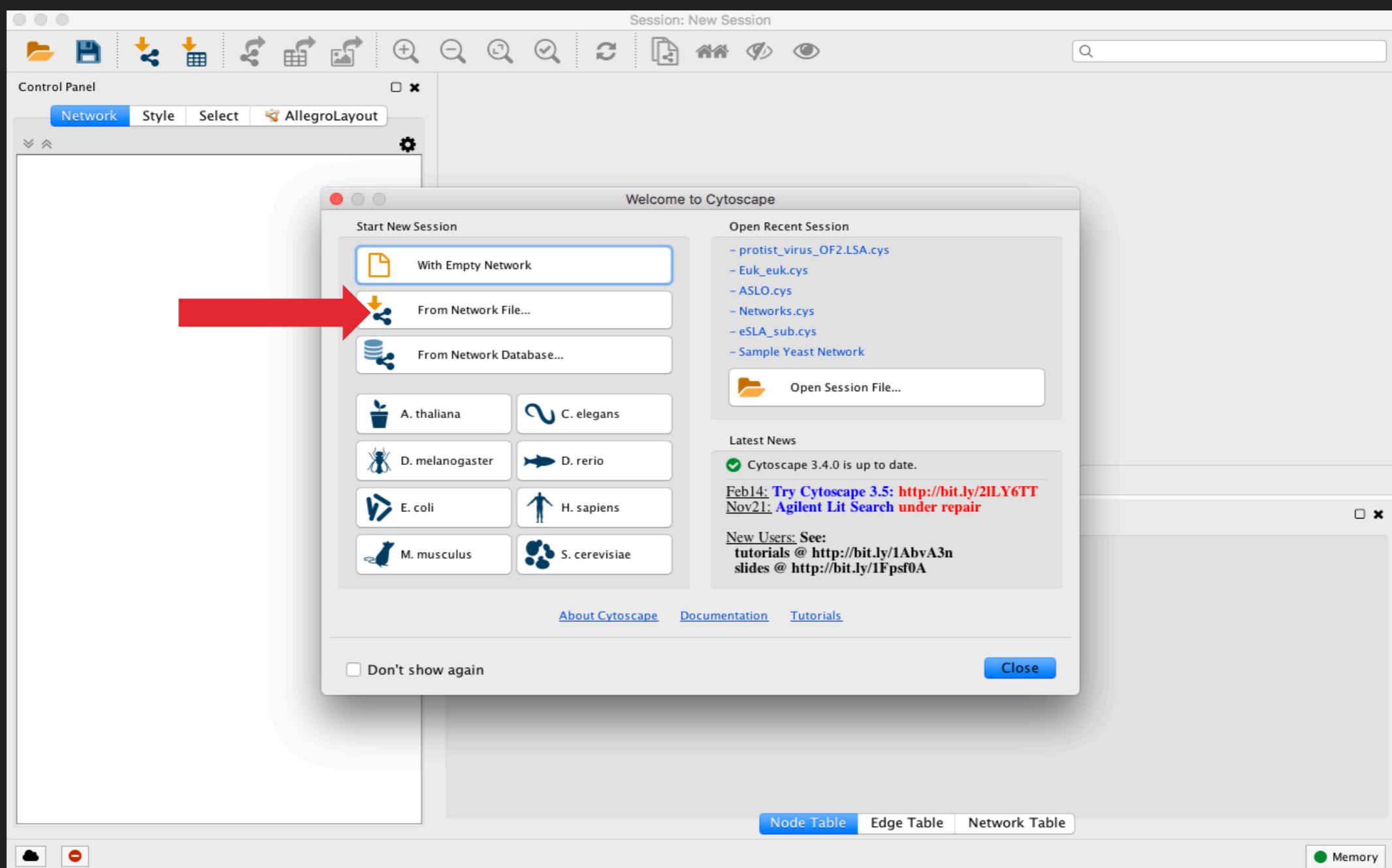
- ▶ <https://github.com/bio9905/course-material>
- ▶ *eLSA_installation_guide_for_abel.sh*
 - ▶ If you want to install on abel for yourself
- ▶ *eLSA_starting_script.sh*
 - ▶ The commands for a typical run
- ▶ *eLSA_otu_table.tsv* -> input data
eLSA_top100_otos.tsv -> input data

THE NETWORK FILES

- ▶ <https://github.com/bio9905/course-material>
- ▶ *Output from eLSA*
 - ▶ *eLSA_network_top100.perm.d0.tsv*
eLSA_network_top100.perm.d1.tsv
eLSA_network.d0.tsv
eLSA_network.d1.tsv
- ▶ *eLSA_node_annotation_relabund.tsv*
- ▶ *eLSA_node_annotation_tax.tsv*
- ▶ *eLSA_for_cytoscape.cys*

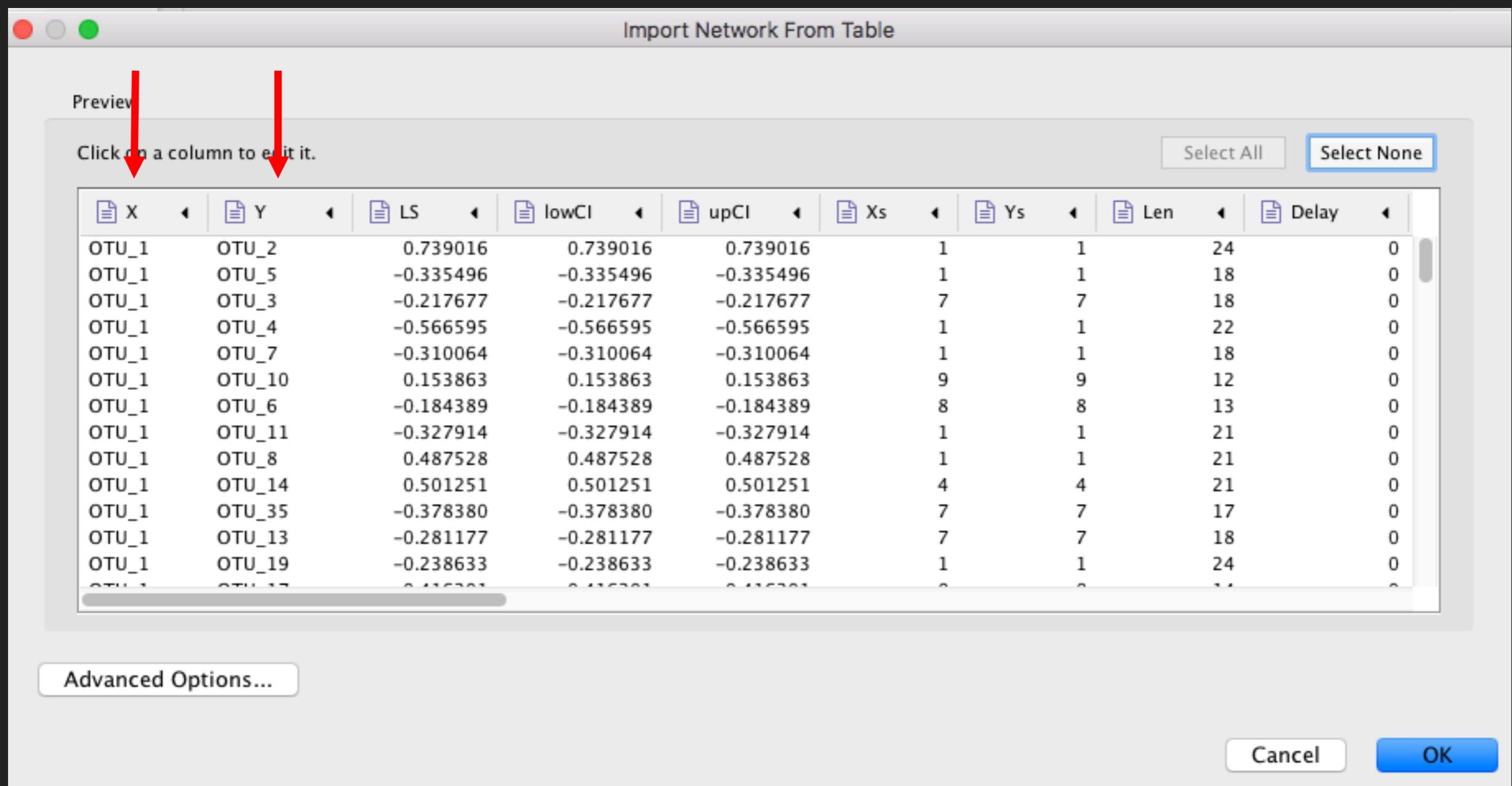
USING CYTOSCAPE

- ▶ open Cytoscape, import *eLSA_network_top100.perm.d0.tsv*

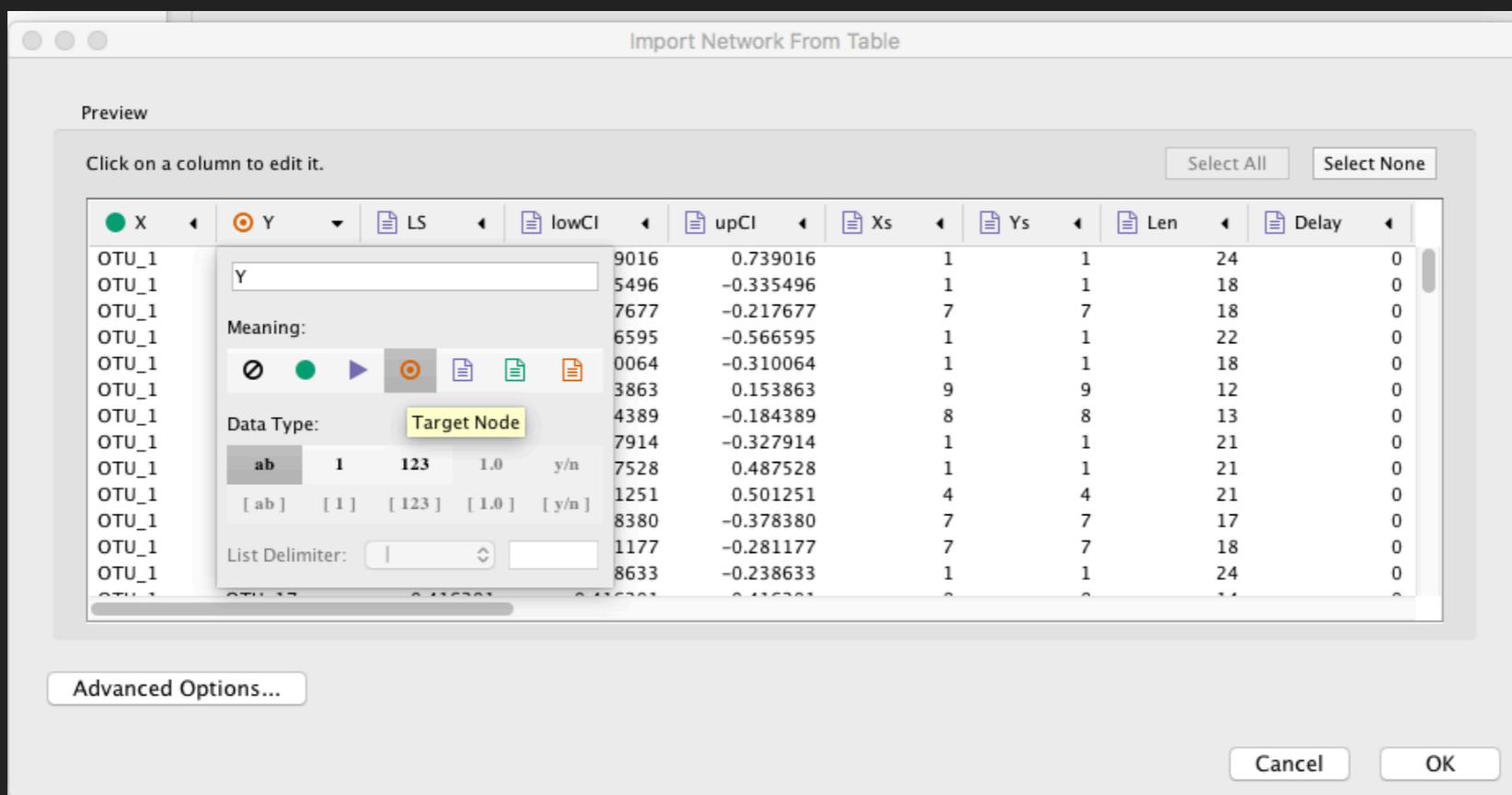
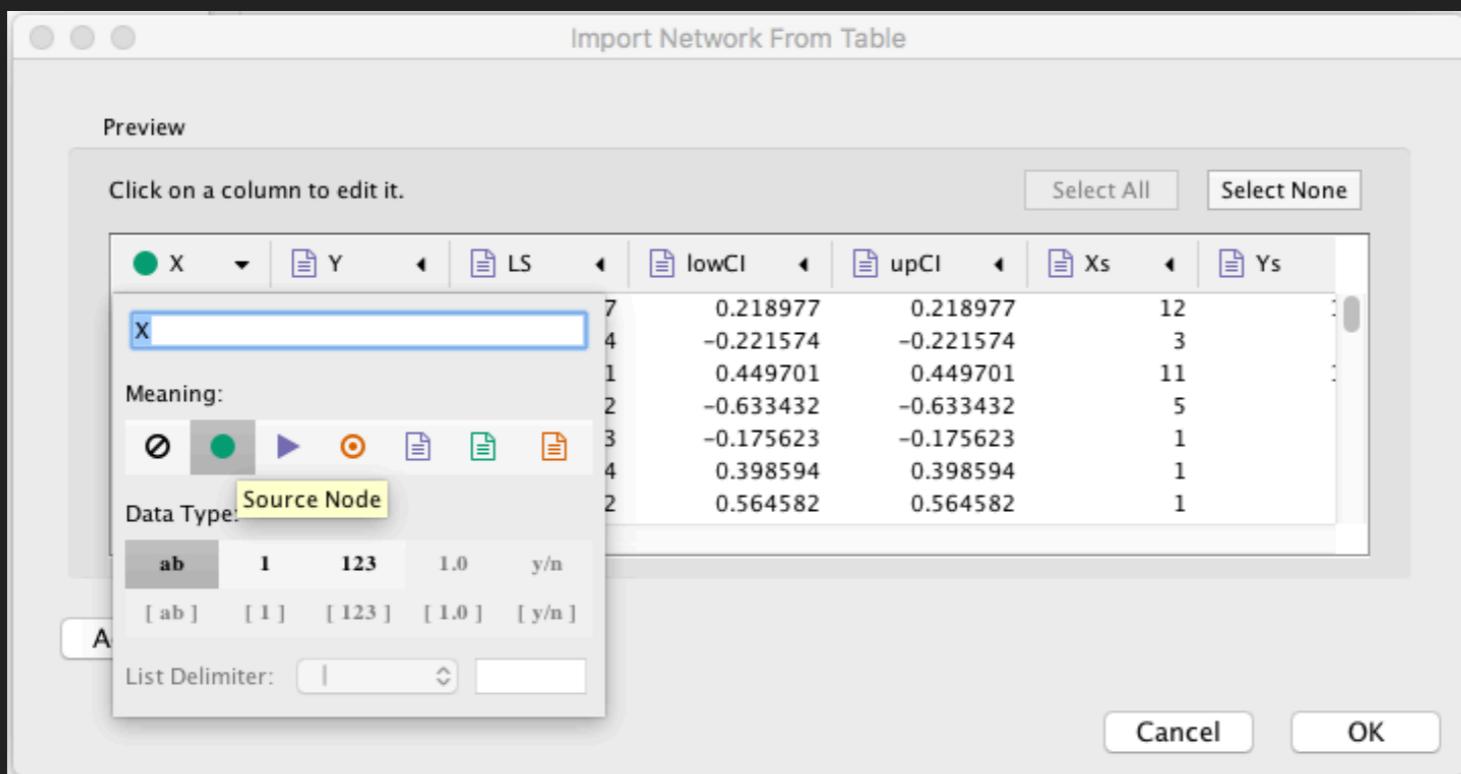


USING CYTOSCAPE

- ▶ Define target and source node (this is an undirected networks so the order of the nodes doesn't matter)



USING CYTOSCAPE



Session: New Session

Control Panel

Network Style Select AllegroLayout

1 of 1 Network selected

eLSA_network_top100.perm.d0.tsv 114 6441

• number of nodes: 114
• number of edges: 6441
• This includes ALL possible connections between nodes.
• We have to remove edges e.g. based on P and Q value to get the significant associations

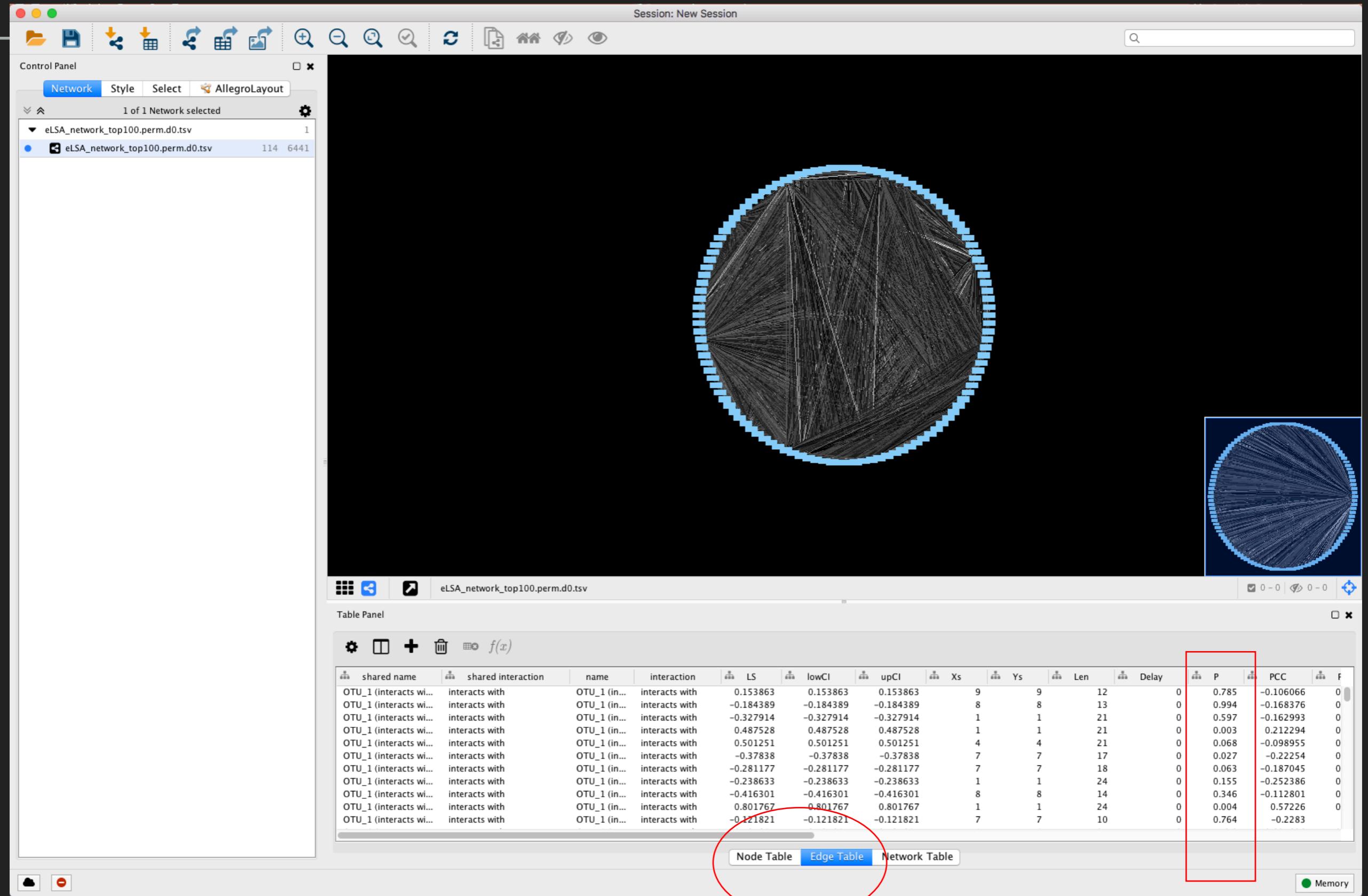
Table Panel

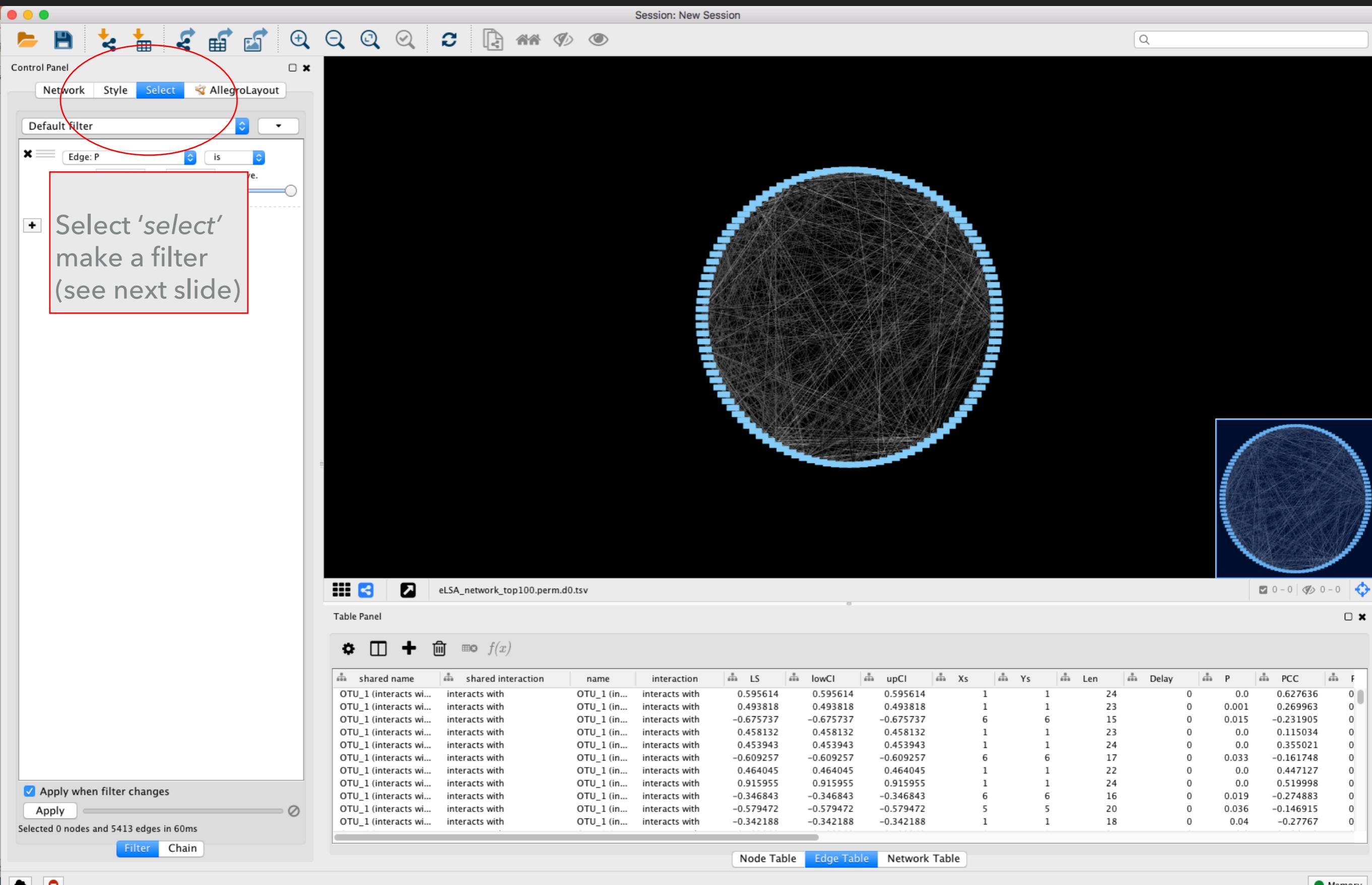
eLSA_network_top100.perm.d0.tsv

Shared name	name
OTU_1	OTU_1
OTU_2	OTU_2
OTU_5	OTU_5
OTU_3	OTU_3
OTU_4	OTU_4
OTU_7	OTU_7
OTU_10	OTU_10
OTU_6	OTU_6
OTU_11	OTU_11
OTU_8	OTU_8
OTU_14	OTU_14
OTU_35	OTU_35

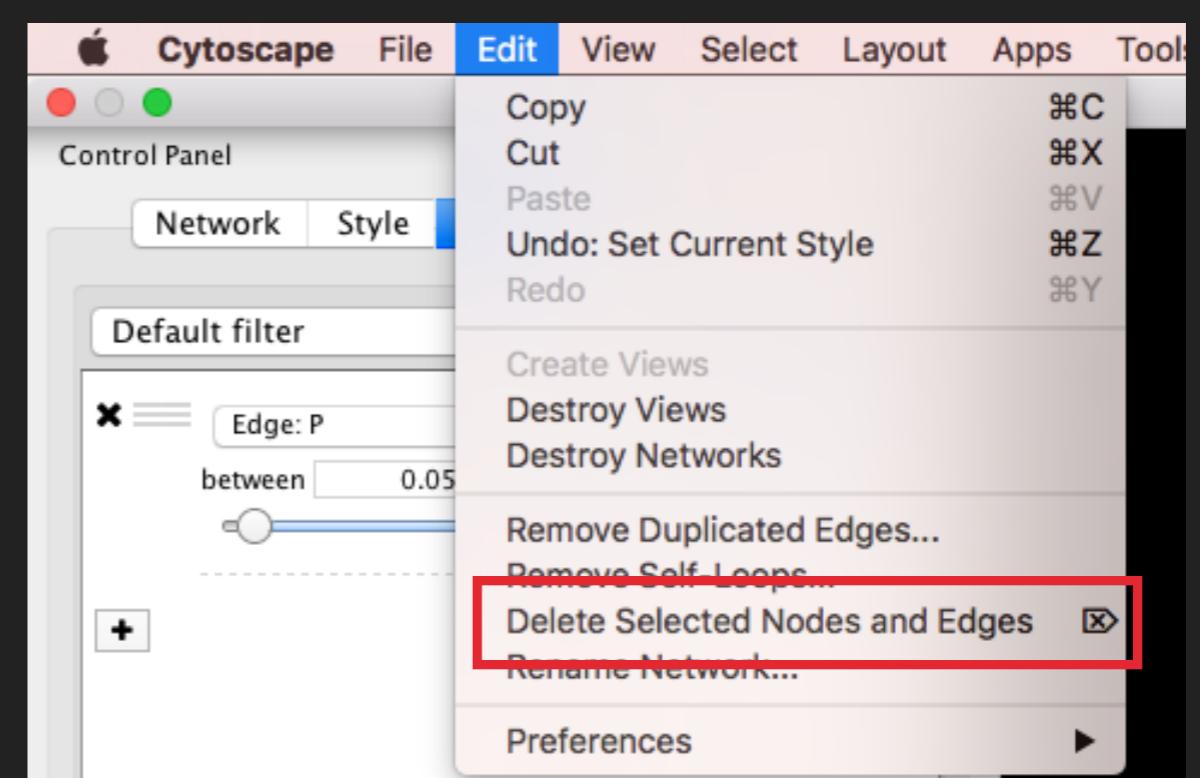
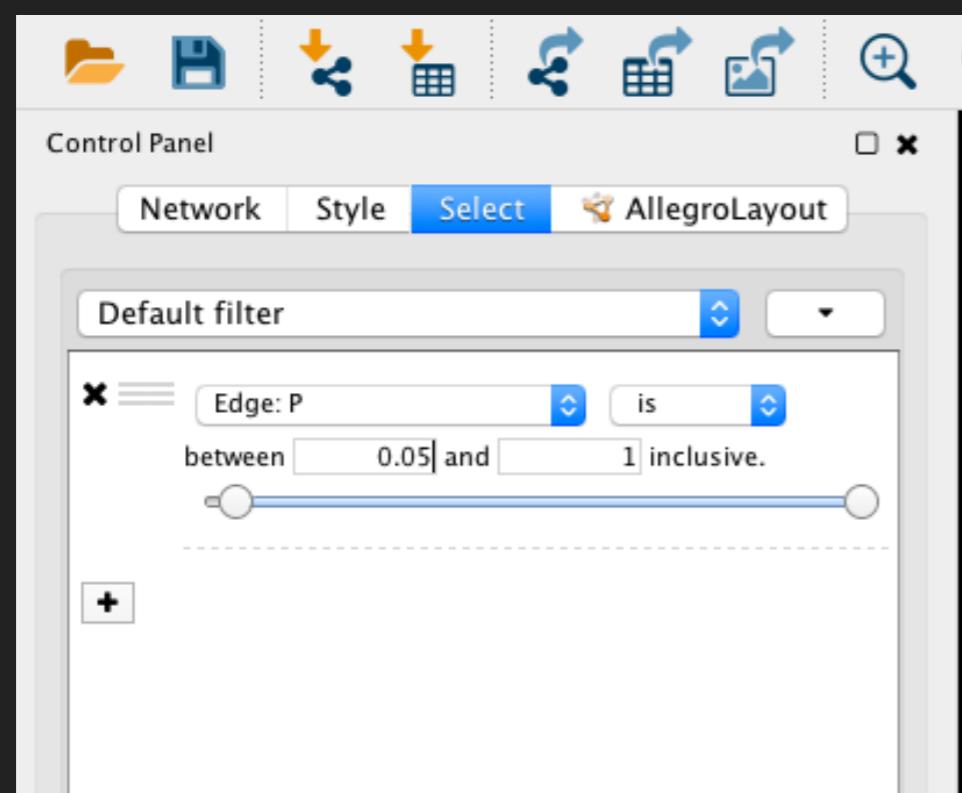
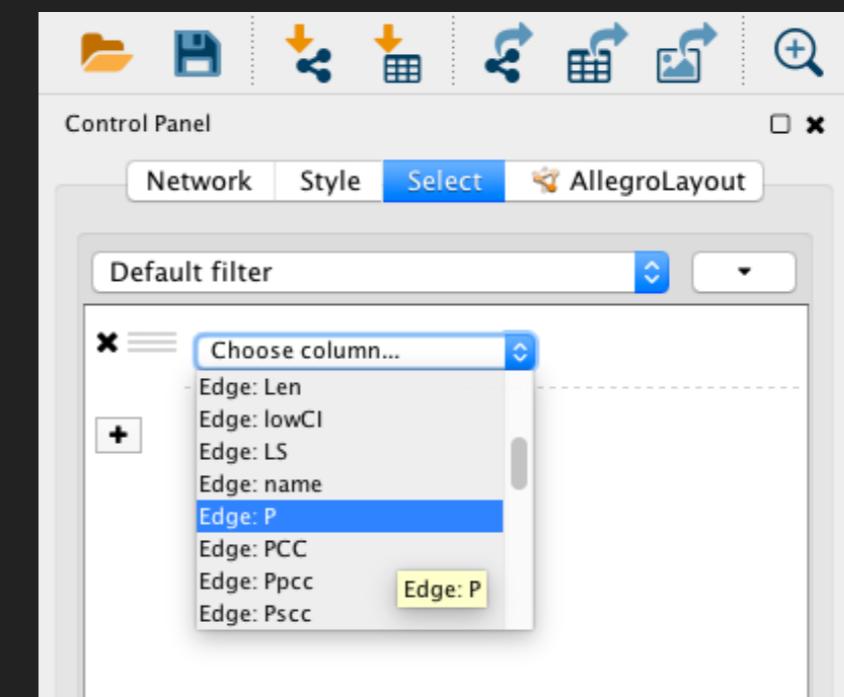
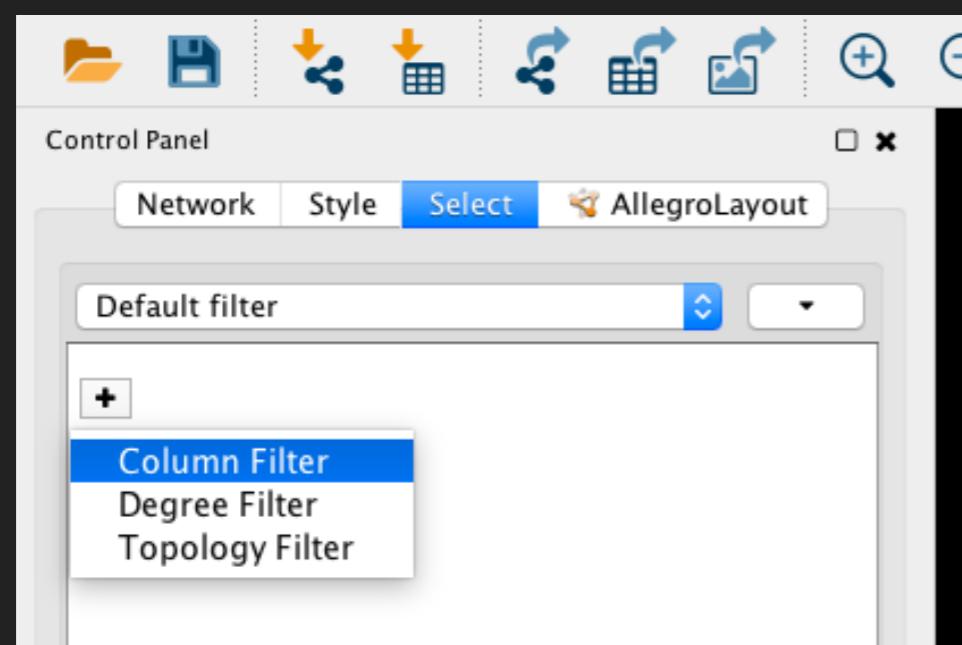
Node Table Edge Table Network Table

Memory





- ▶ Setting up a filter for P value (repeat for Q value)
How many edges are left?



Choose a layout - Circular is fastest
Edge-weighted Spring Embedded Layout (LS) looks better😊

Screenshot of the Cytoscape application interface showing the Layout menu open.

The Layout menu is open, displaying various network layout options:

- Bundle Edges
- Clear All Edge Bends
- DeDaL
- Rotate
- Scale
- Align and Distribute
- Settings...
- Apply Preferred Layout F5
- yFiles Layouts
- Grid Layout
- Hierarchical Layout
- Circular Layout
- Stacked Node Layout
- Attribute Circle Layout
- Prefuse Force Directed OpenCL Layout
- Degree Sorted Circle Layout
- Prefuse Force Directed Layout
- Group Attributes Layout
- Edge-weighted Force directed (BioLayout)
- Edge-weighted Spring Embedded Layout** (highlighted in blue)
- Inverted Self-Organizing Map Layout
- Allegro Spring-Electric
- Allegro Fruchterman-Reingold
- Allegro Weak Clustering
- Allegro Strong Clustering
- Allegro Edge-Repulsive Spring-Electric
- Allegro Edge-Repulsive Fruchterman-Reingold
- Allegro Edge-Repulsive Weak Clustering
- Allegro Edge-Repulsive Strong Clustering
- Relative Entropy Optimization (EntOpt) Layout

The main canvas displays a circular network graph with many edges, and a smaller inset shows the same graph with a different layout.

Now choose 'Style'
and then 'Edge' to change
the color of the edges

Session: New Session

Control Panel

Network Style Select AllegroLayout

default

Properties Def. Map. Byp.

Border Paint
Border Width
Fill Color
Height
Image/Chart 1
Label
Label Color
Label Font Size
Shape
Size
Transparency
Width
 Lock node width and height

Node Edge Network

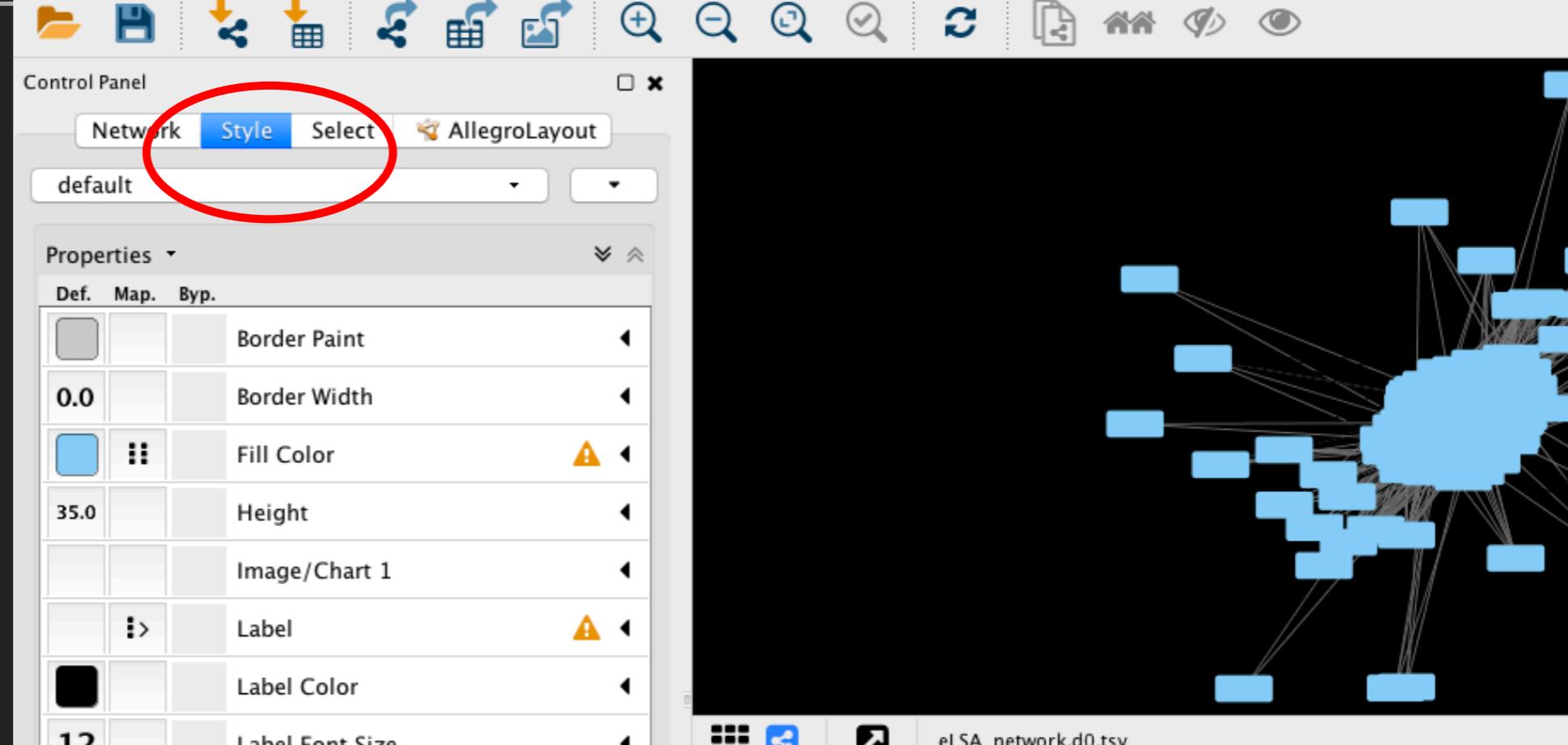
Table Panel

eLSA_network.d0.tsv

shared name name

OTU_521	OTU_521
OTU_522	OTU_522
OTU_526	OTU_526
OTU_529	OTU_529
OTU_5293	OTU_5293
OTU_53	OTU_53
OTU_537	OTU_537
OTU_5382	OTU_5382
OTU_540	OTU_540
OTU_543	OTU_543
OTU_5462	OTU_5462
OTU_55	OTU_55

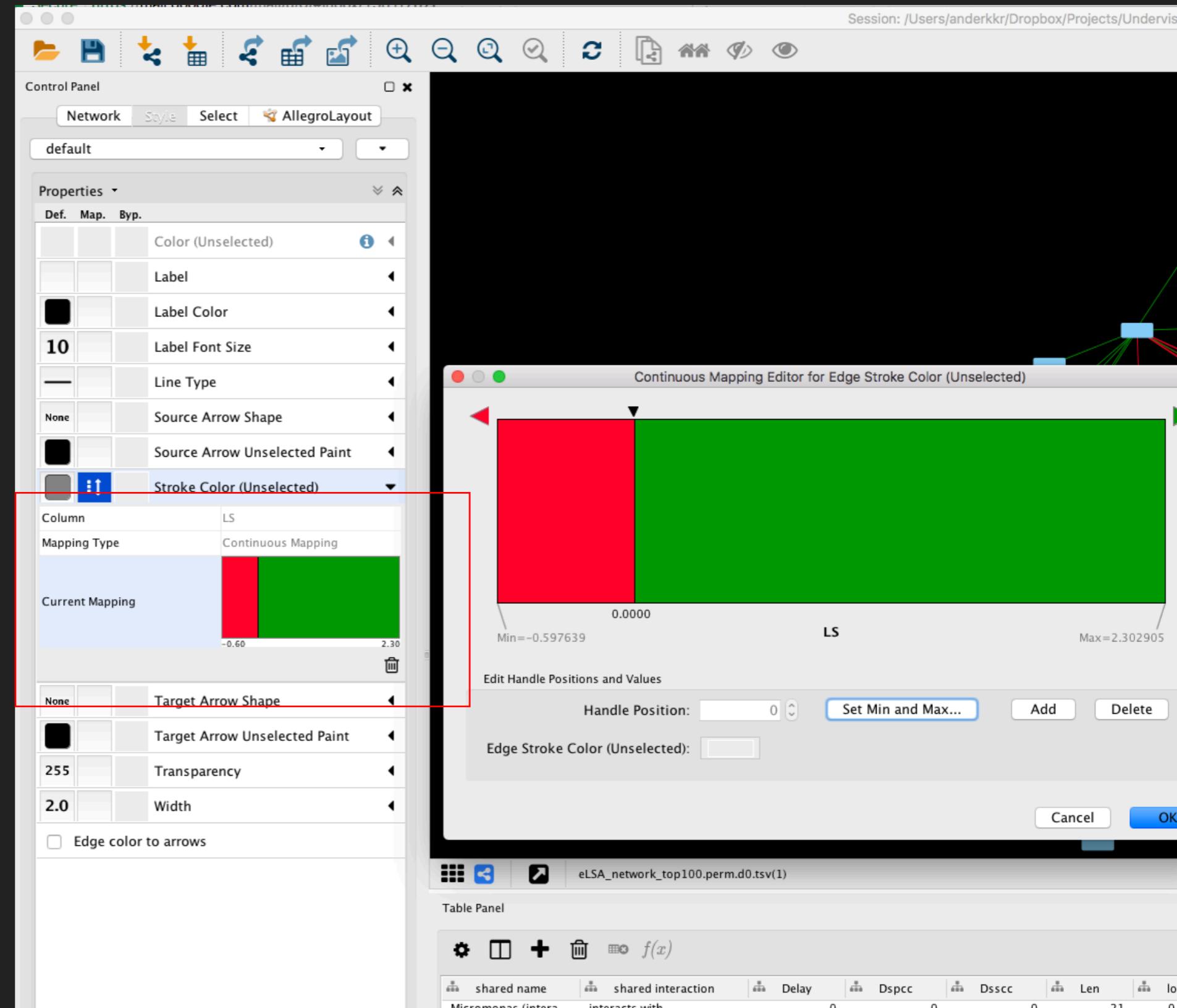
Node Table Edge Table



Use *Stroke Color* to make the positive edges green and the negative edges red

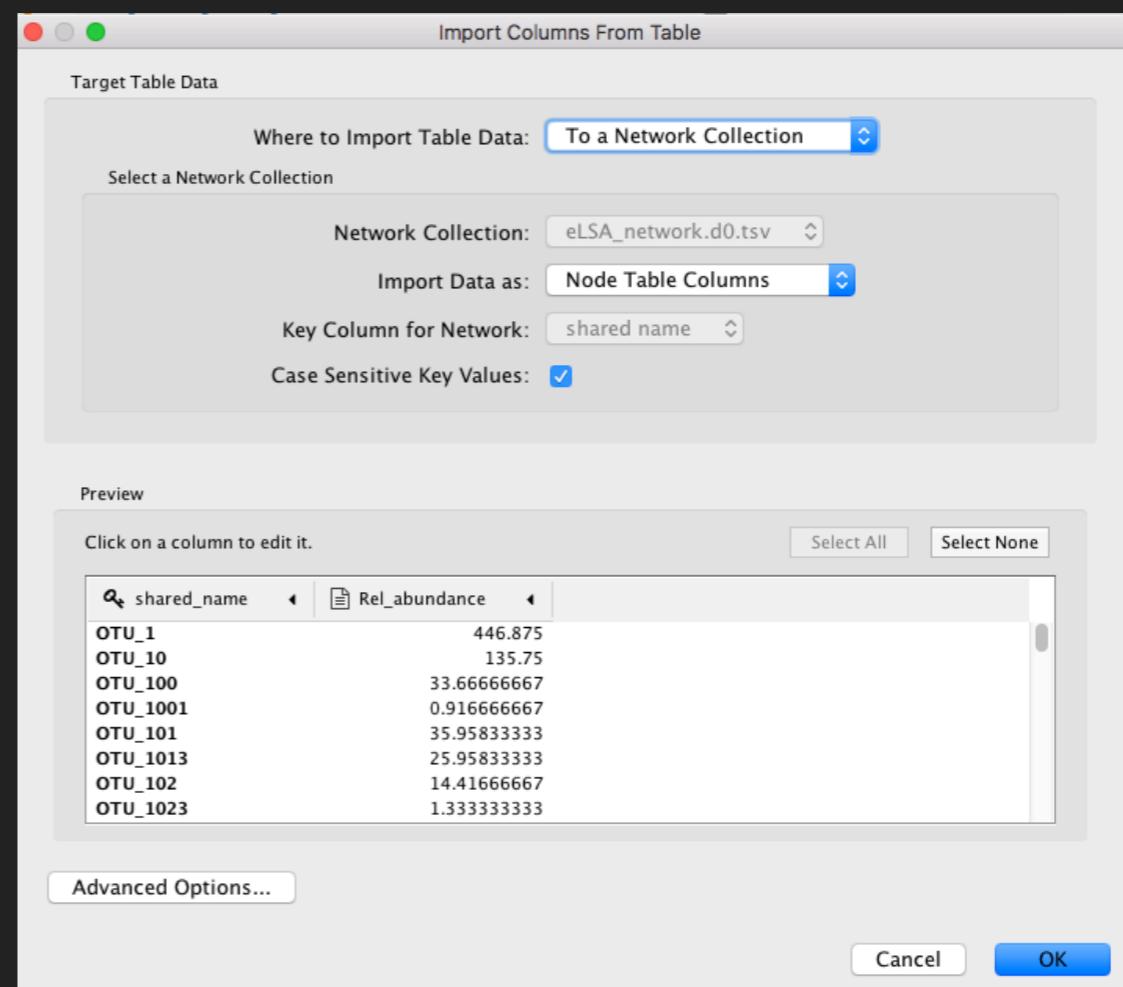
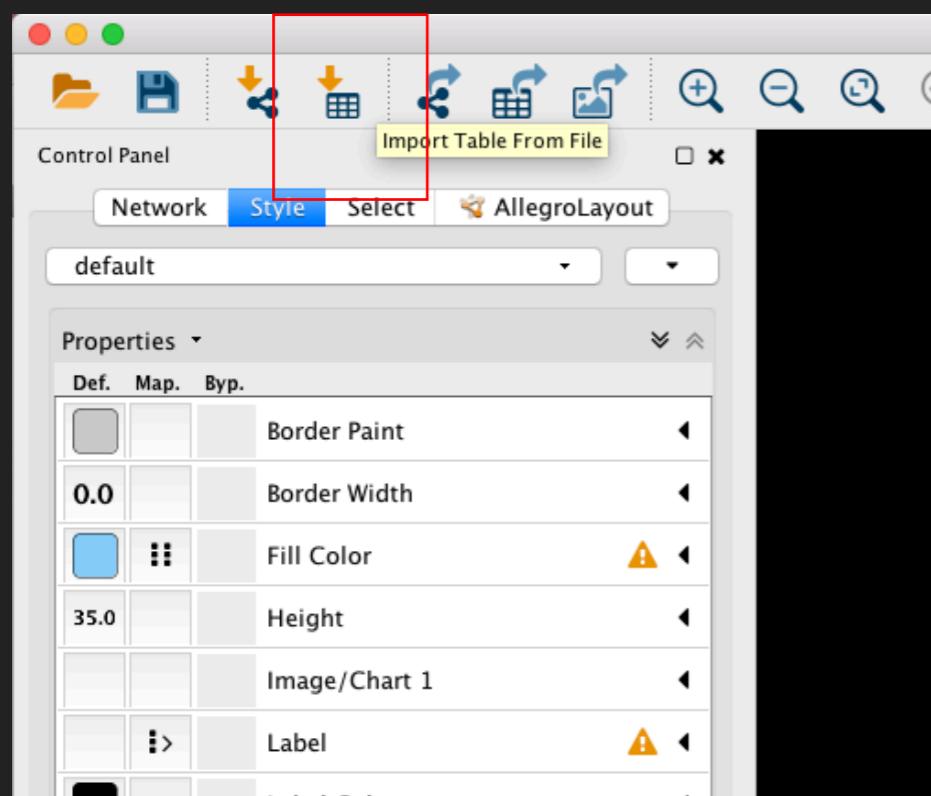
Column: LS

Mapping type: Continuous



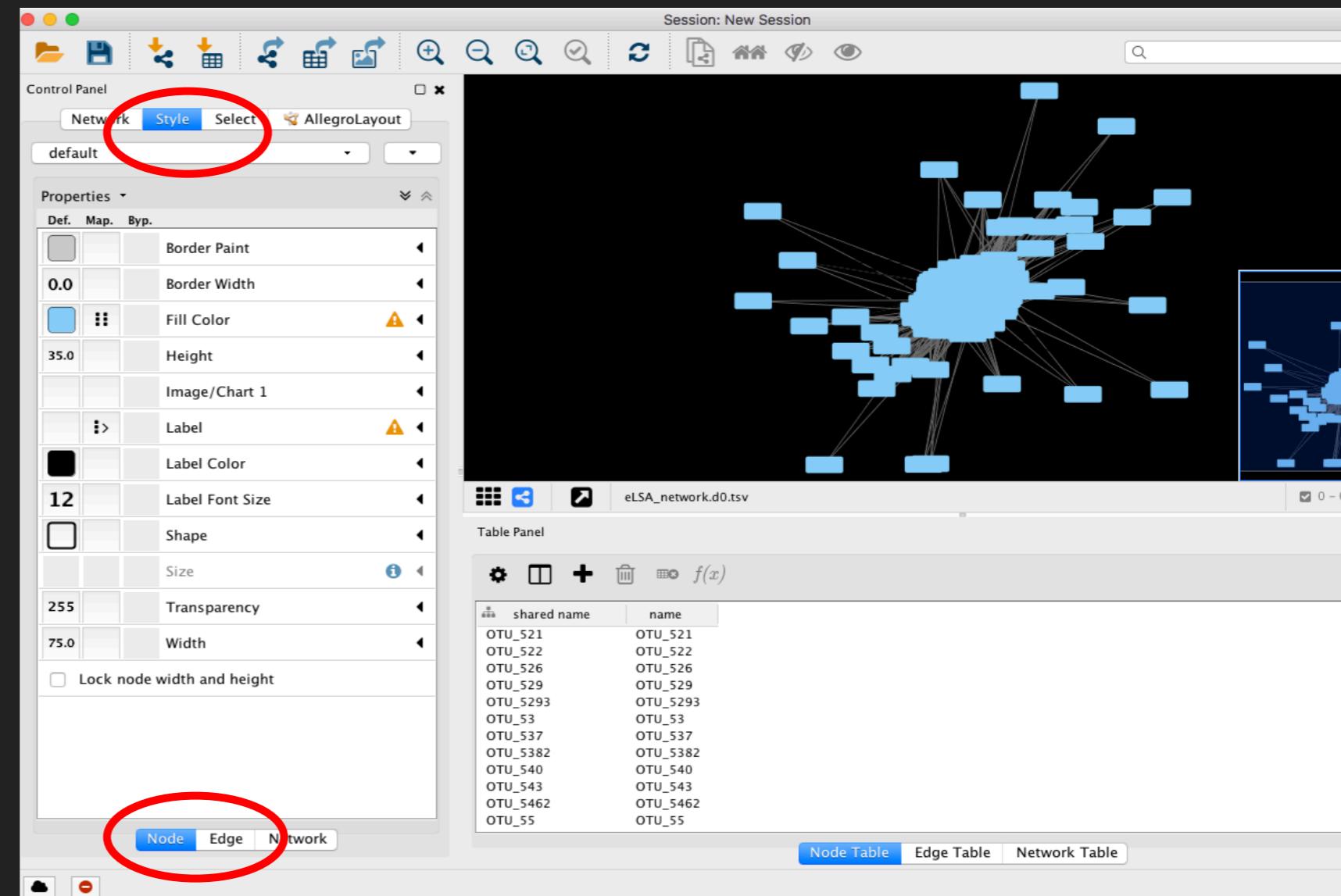
ADD ANNOTATIONS TO THE NODES- IMPORT TABLE

- ▶ Two annotation files:
 - 1) Relative abundance (eLSA_annotation_relabund.tsv)
 - 2) Taxonomic assignment (blastn vs MAS from UParse pipeline)

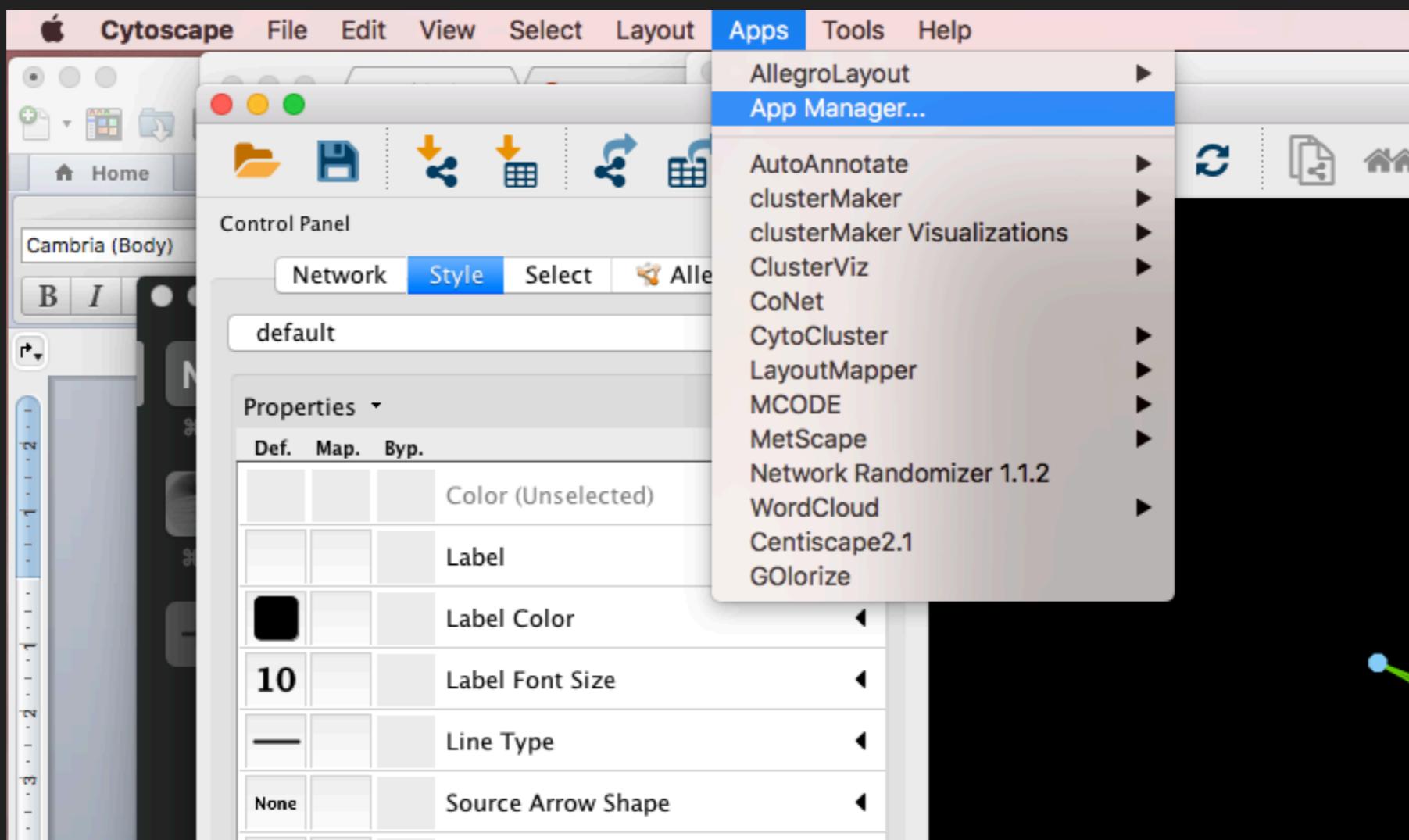


- ▶ Change the name, shape, color and/or size of the nodes based on the relative abundance and the taxonomic assignment of the OTUs

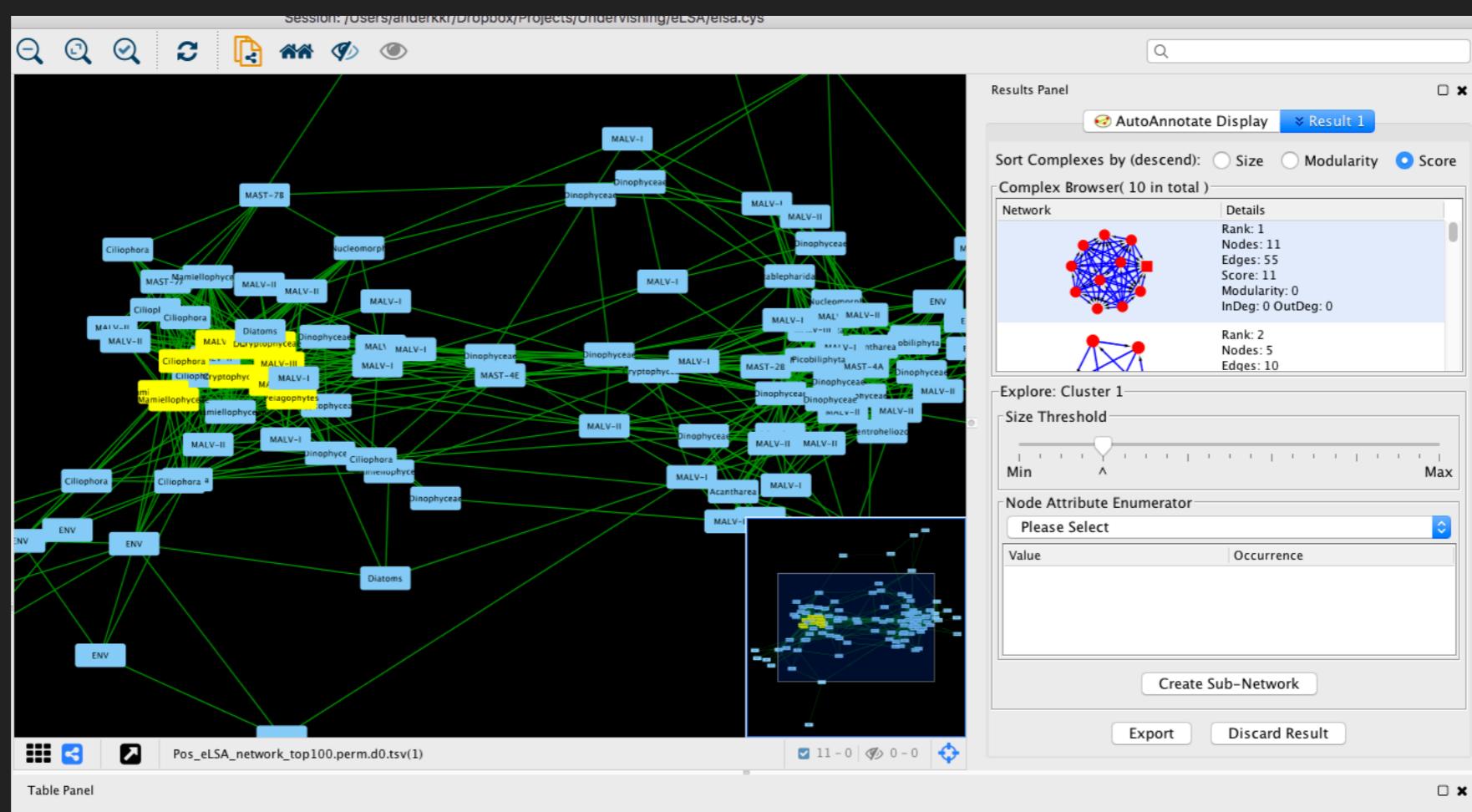
Choose 'select'
and then 'node' to change
the color, shape etc of the nodes



- ▶ Searching for Modules:
- ▶ Install ClusterViz from the App Manager



- ▶ Searching for Modules:
- ▶ (Save your project before moving on, just in case)
- ▶ Remove the negative edges, then search for Modules using MCODE in the ClusterViz App.



- ▶ Searching for Hubs (species that are highly connected, or have a high Degree). MCODE added the Degree for each node to the Node table.

Table Panel

The screenshot shows a table panel with the following interface elements:

- Toolbar icons: gear, refresh, plus, trash, and a search bar labeled $f(x)$.
- Table header: shared name, name, AverageShortestPathLength, BetweennessCentrality, ClosenessCentrality, ClusteringCoefficient, Degree, Eccentricity.
- Data rows (12 rows total):

shared name	name	AverageShortestPathLength	BetweennessCentrality	ClosenessCentrality	ClusteringCoefficient	Degree	Eccentricity
OTU_1	OTU_1	1.65486726	0.03309071	0.60427807	0.37439024	41	3
OTU_2	OTU_2	1.68141593	0.02127558	0.59473684	0.4048583	39	3
OTU_8	OTU_8	1.84955752	0.02086659	0.54066986	0.41269841	28	3
OTU_15	OTU_15	1.92035398	0.00404103	0.52073733	0.63333333	25	3
OTU_16	OTU_16	1.73451327	0.01822451	0.57653061	0.44507576	33	3
OTU_28	OTU_28	1.78761062	0.01807856	0.55940594	0.4516129	32	3
OTU_23	OTU_23	1.83185841	0.00877496	0.54589372	0.54232804	28	3
OTU_30	OTU_30	1.79646018	0.00946158	0.55665025	0.52643678	30	3
OTU_64	OTU_64	1.86725664	0.0054171	0.53554502	0.58465608	28	3
OTU_69	OTU_69	1.78761062	0.01844582	0.55940594	0.48790323	32	3
OTU_62	OTU_62	1.85840708	0.01512415	0.53809524	0.48029557	29	3
- Bottom navigation buttons: Node Table (highlighted in blue), Edge Table, Network Table.

-
- ▶ If there is time you can repeat these steps with the network with allowed delay to see if there are any differences. And see which if any species have a higher LS value when delay is allowed in the LS scoring.
 - ▶ eLSA_network_top100.perm.d1.tsv