Similarity-based clustering using a network analysis approach

Leandro Ariza-Jiménez

PhD student in Mathematical Engineering

Advisors:
Olga Lucía Quintero Montoya
Nicolás Pinel Peláez

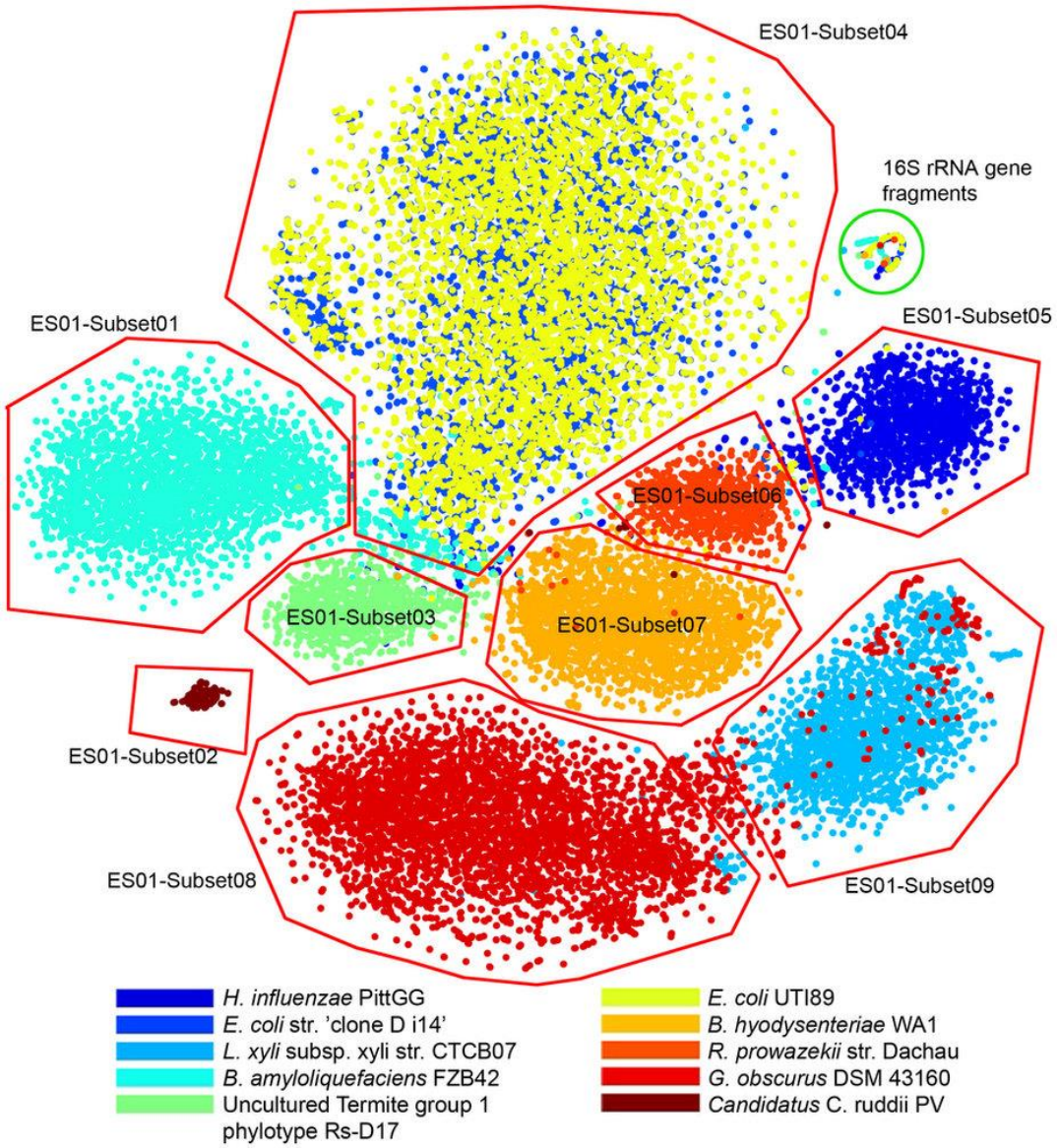
Outline

- Motivation
- Problem statement
- Networks and communities
- Similarity-based networks
- Application examples
- Future work
- Conclusions

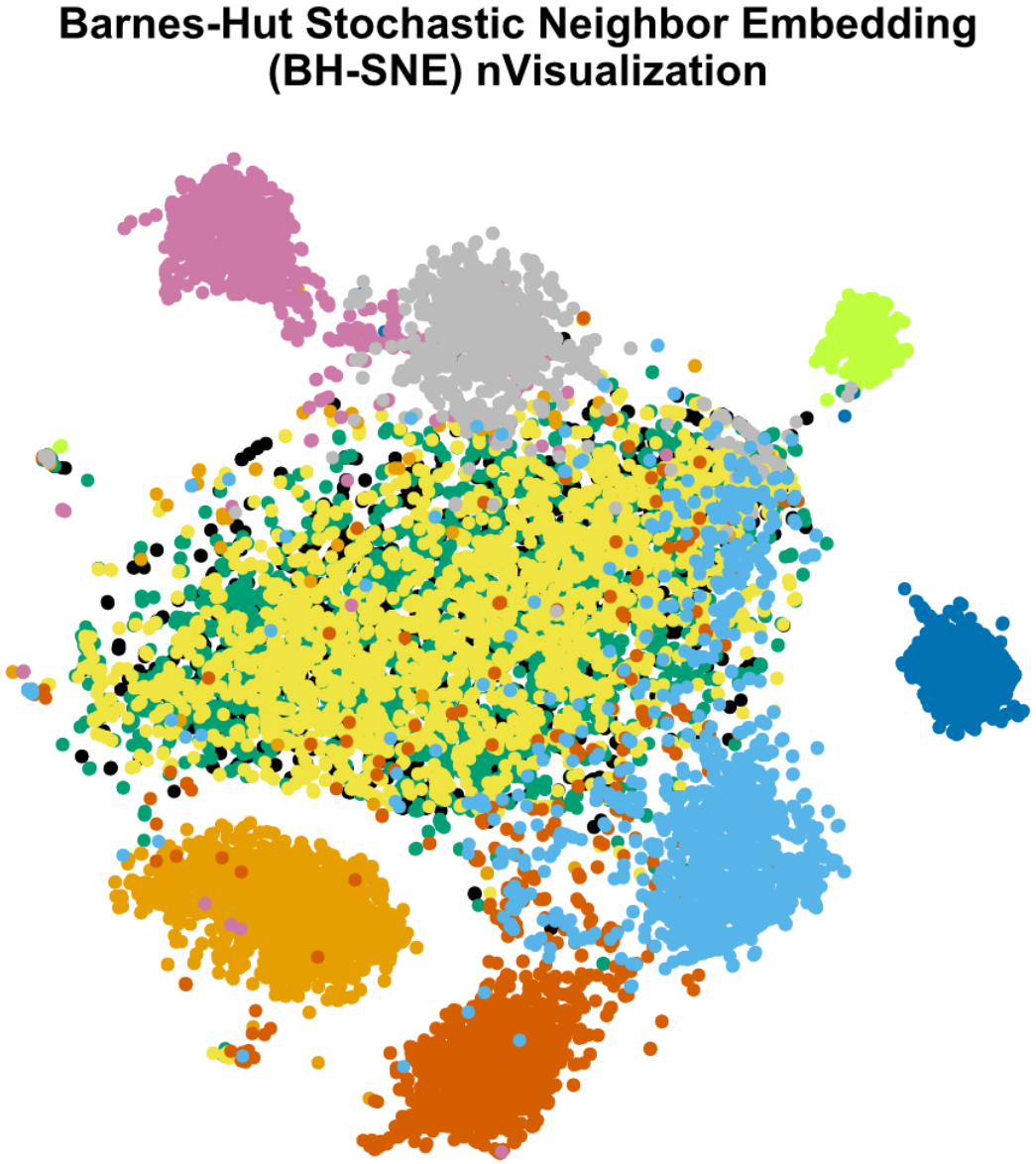
Motivation

Laczny et al. (2014).
Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction.
Scientific Reports, *4*(1).

Metagenomic data visualization of a simulated microbial community.



Motivation


Barnes-Hut Stochastic Neighbor Embedding (BH-SNE) nVisualization
Microorganism
Arcobacter butzleri
Bacteroides caccae
Bacteroides intestinalis
Bacteroides xylanisolvens
Enterobacter cloacae
Helicobacter pylori
Lachnospiraceae saccharolyticum
Lactobacillus fermentum
Lactobacillus reuteri
Ruminococcus obeum

Motivation

- Major challenges and issues in data clustering:
  - *A priori* unknown number of clusters
  - “Dimensionality curse”
  - Convergence
  - Heuristics

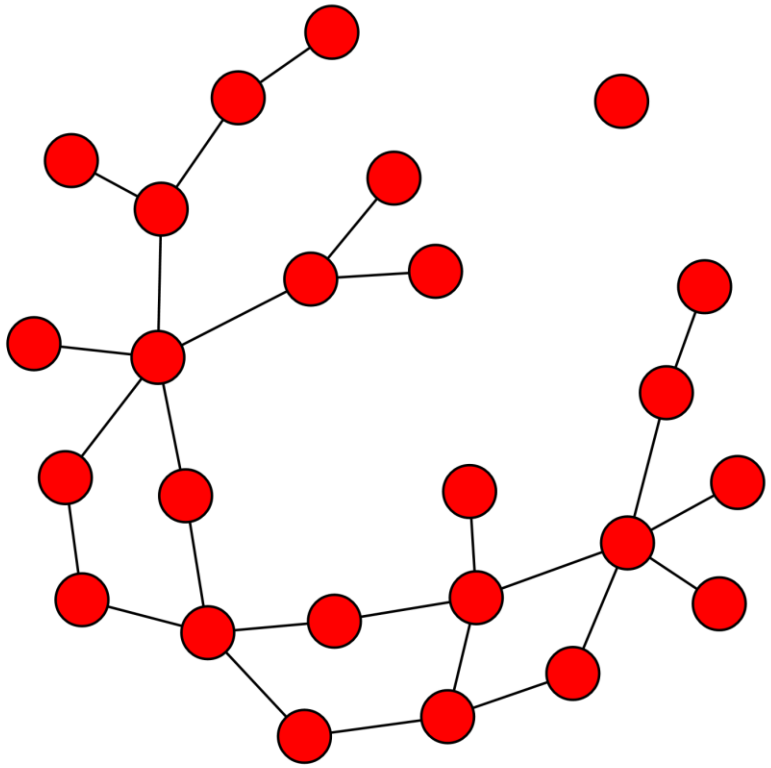
Problem statement

Conventional approaches to data clustering may not always succesfully retrieve the underlying structure of the data due to their inherent issues

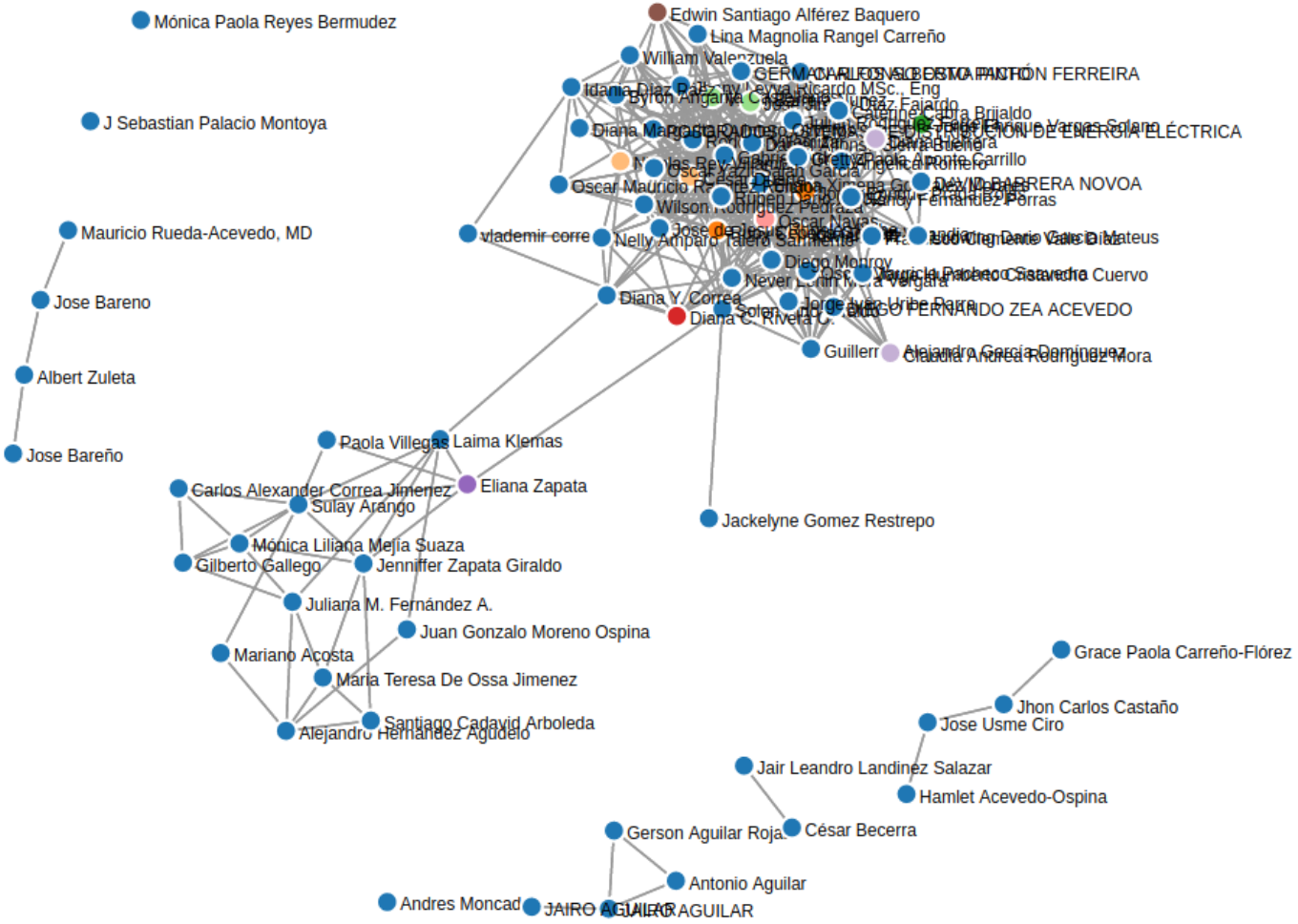
Research question:

Can we overcome these issues by performing data-clustering based on a network analysis approach?

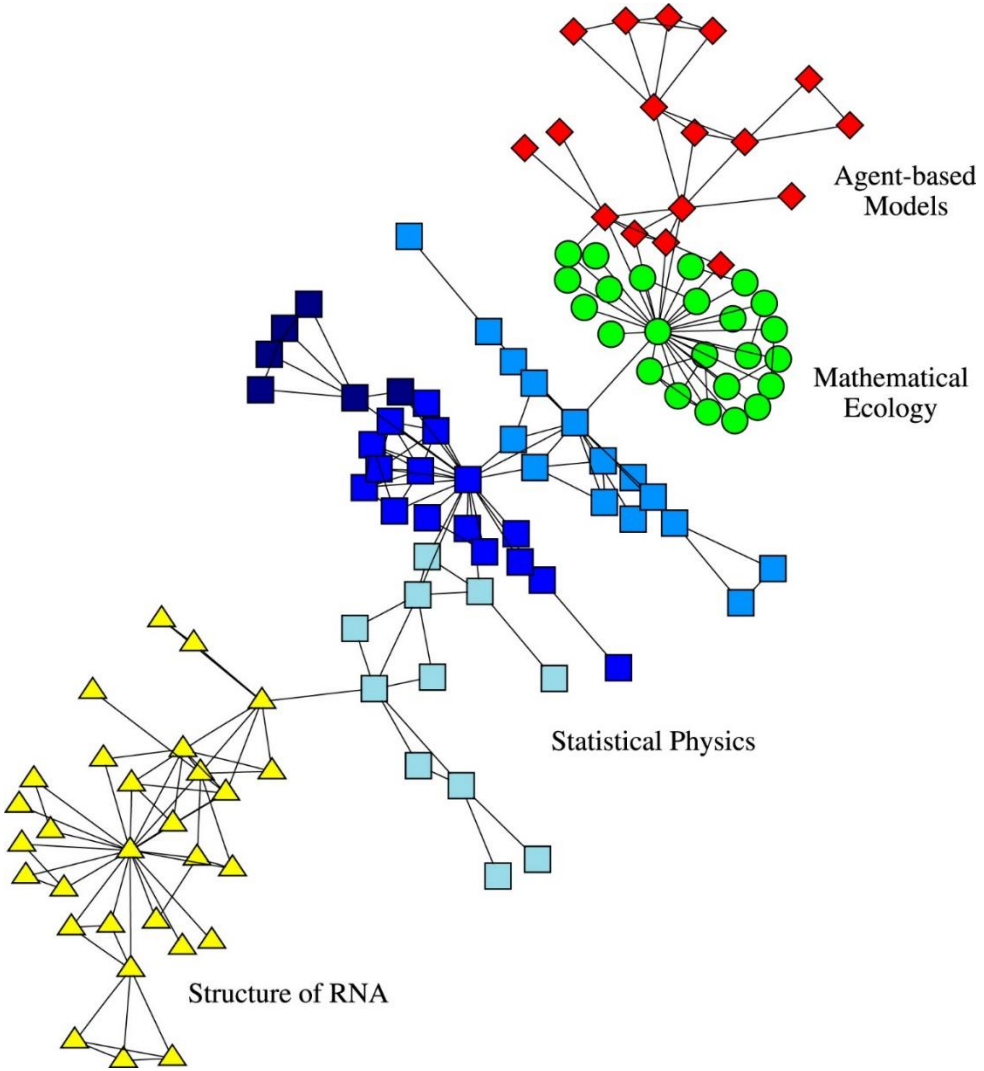
Networks



LinkedIn Social Network



Collaboration network of scientists



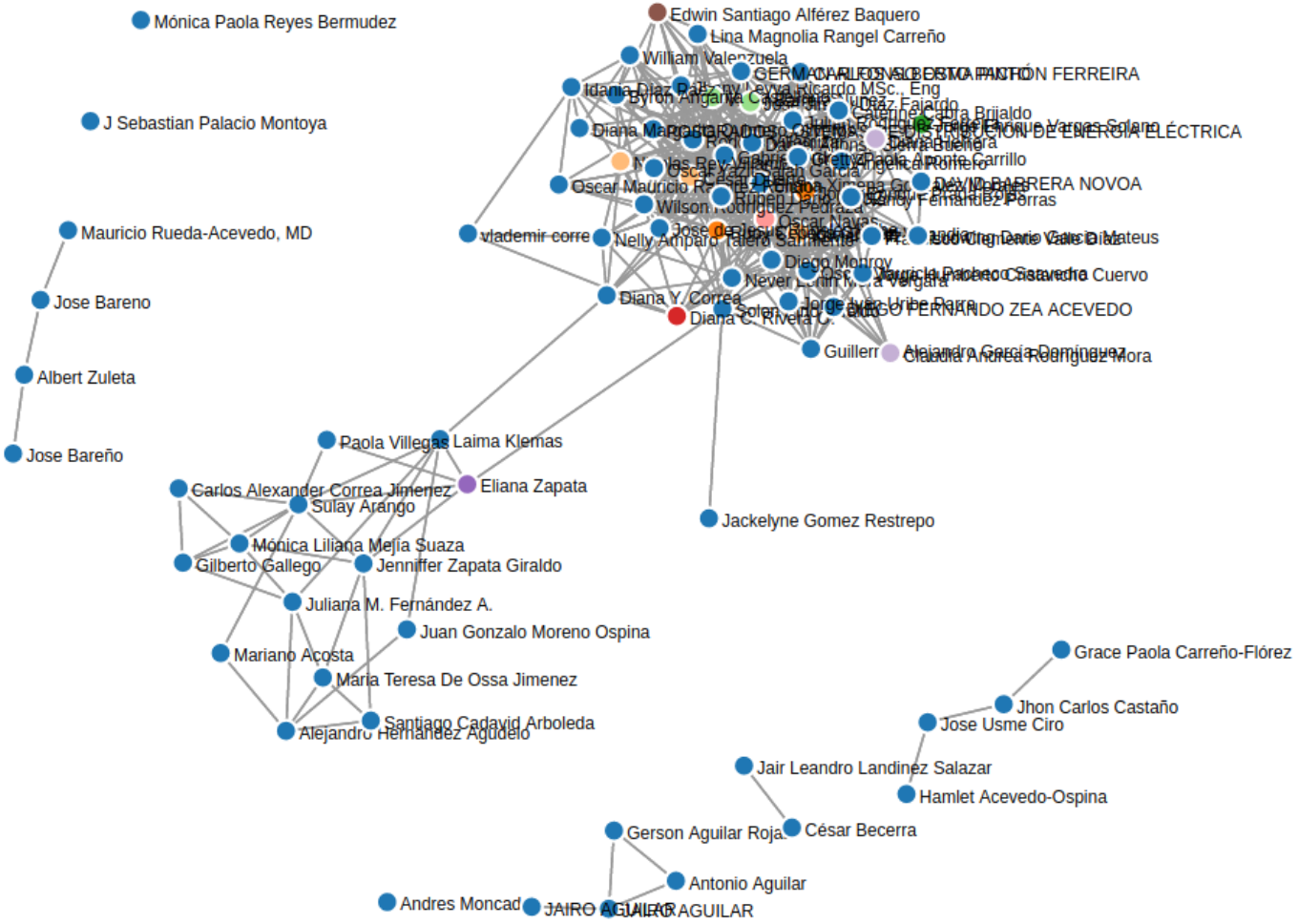
Collaboration network of scientists working at the Santa Fe Institute (SFI). Edges connect scientists that have coauthored at least one paper. Symbols indicate the research areas of the scientists. Naturally, there are more edges between scholars working on the same area than between scholars working in different areas.

Fortunato, S., & Hric, D. (2016). **Community detection in networks: A user guide.** *Physics Reports*, *659*, 1-44.

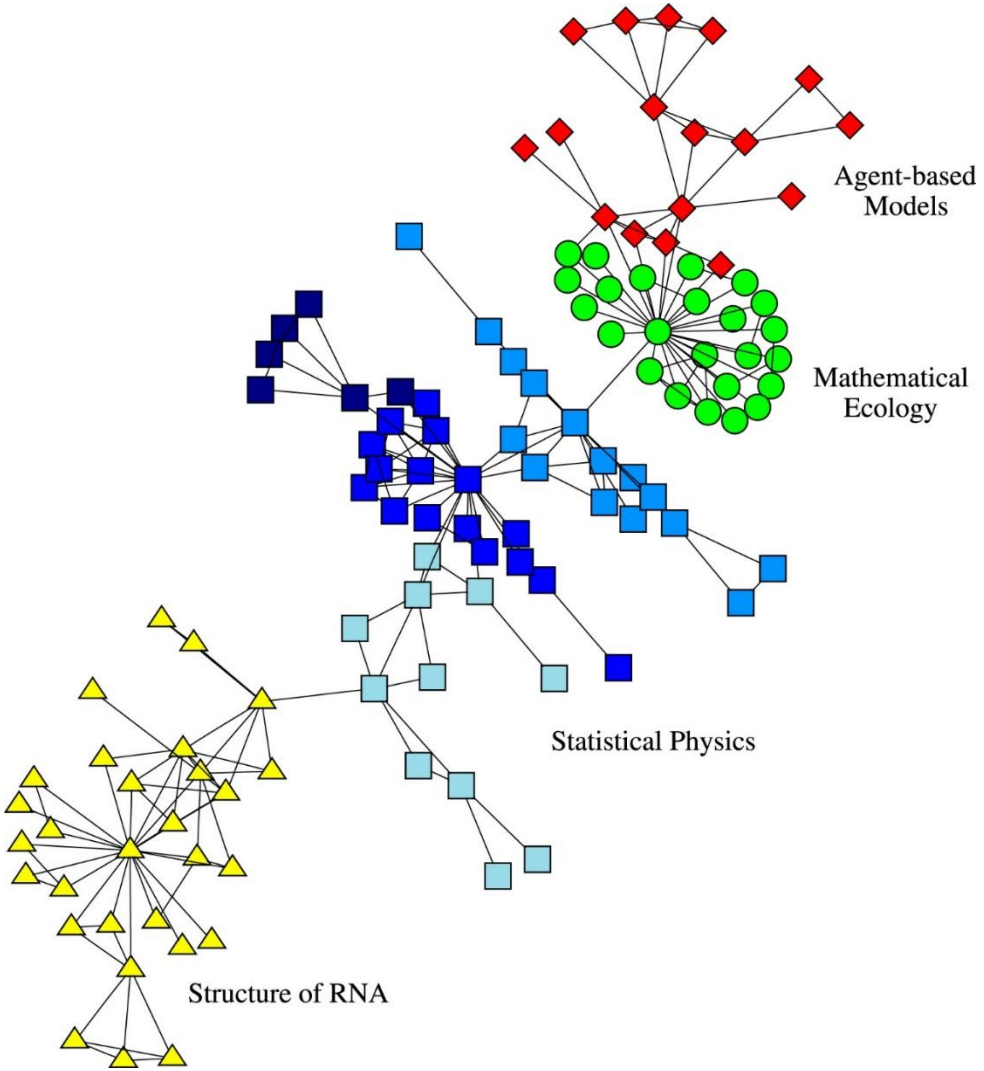
Community

- Networks can have community structure.
  - Network vertices are organized into groups
- No definition is universally accepted, but there is a intuitive definition.
- Its definition often depends on the target application.
- It is a group of vertices which probably…
  - share common properties
  - play similar roles

LinkedIn Social Network



Collaboration network of scientists



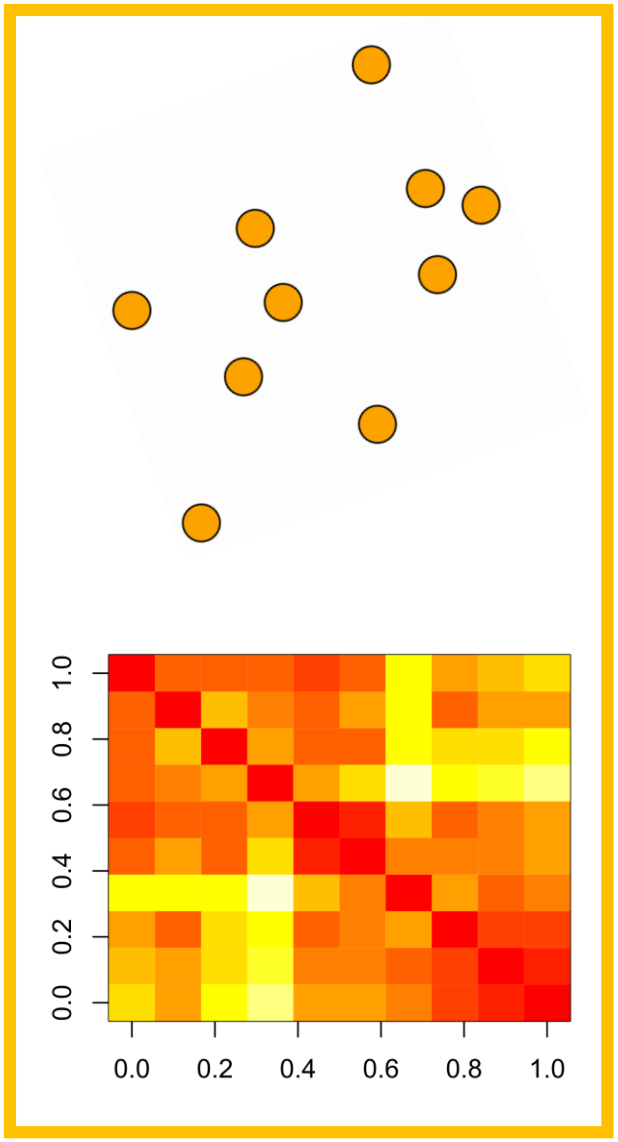
Collaboration network of scientists working at the Santa Fe Institute (SFI). Edges connect scientists that have coauthored at least one paper. Symbols indicate the research areas of the scientists. Naturally, there are more edges between scholars working on the same area than between scholars working in different areas.

Fortunato, S., & Hric, D. (2016). **Community detection in networks: A user guide.** *Physics Reports*, *659*, 1-44.

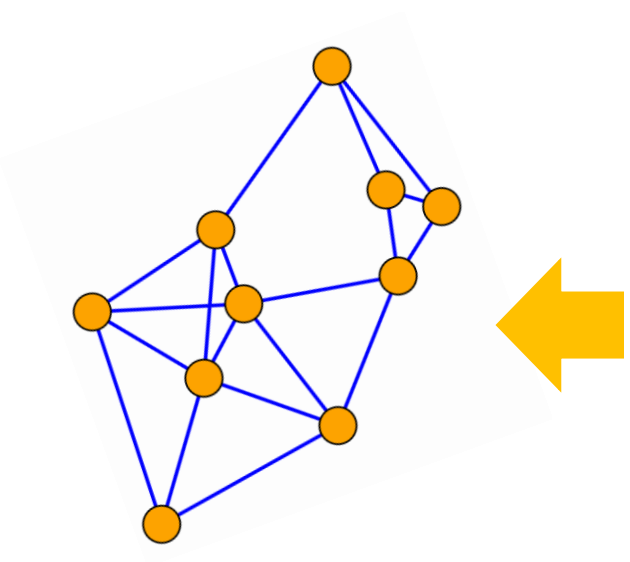
Similarity-based network

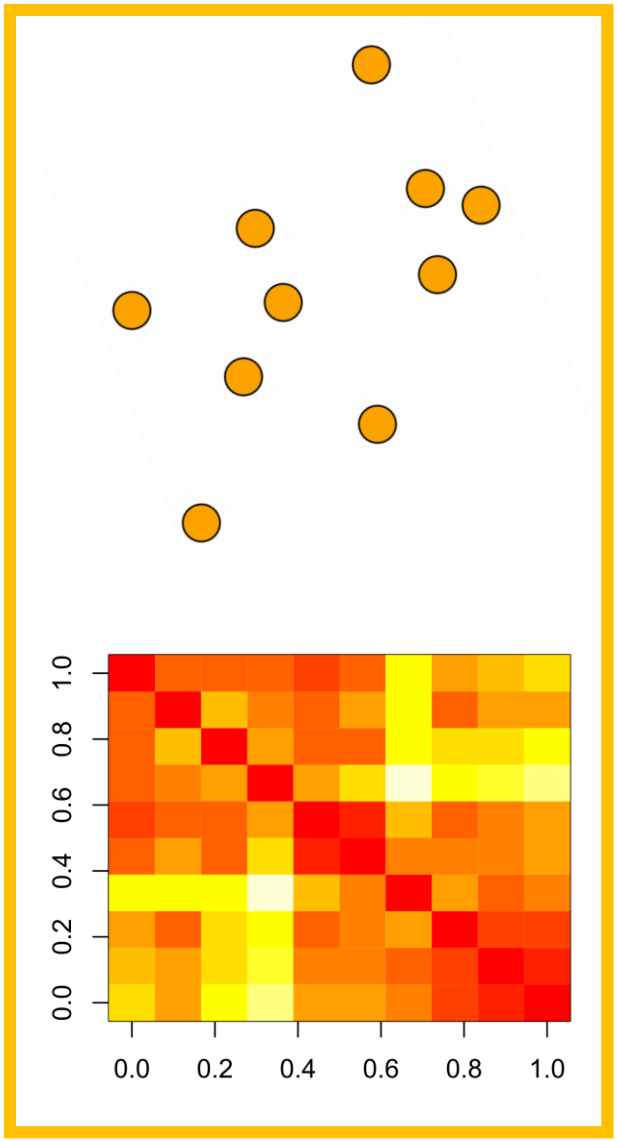
- Networks can represent similarity relationships between objects.
- Have to compute pair-wise similarities as a prerequisite.
- Then, obtain a representative network adjacency matrix.
- Adjacency matrix construction approaches:
  - Knn >> Binary (sparse) matrix
  - Heat kernel>> Weighted (fully connected) matrix

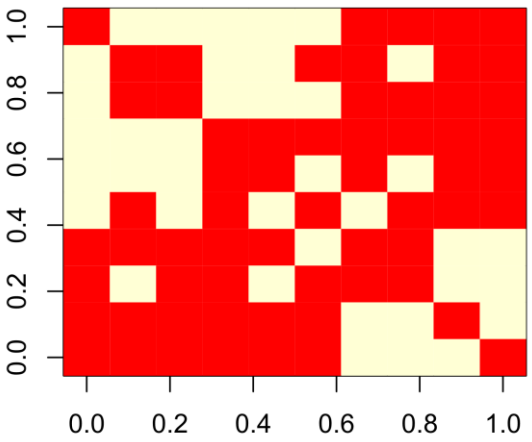
Similarity-based network



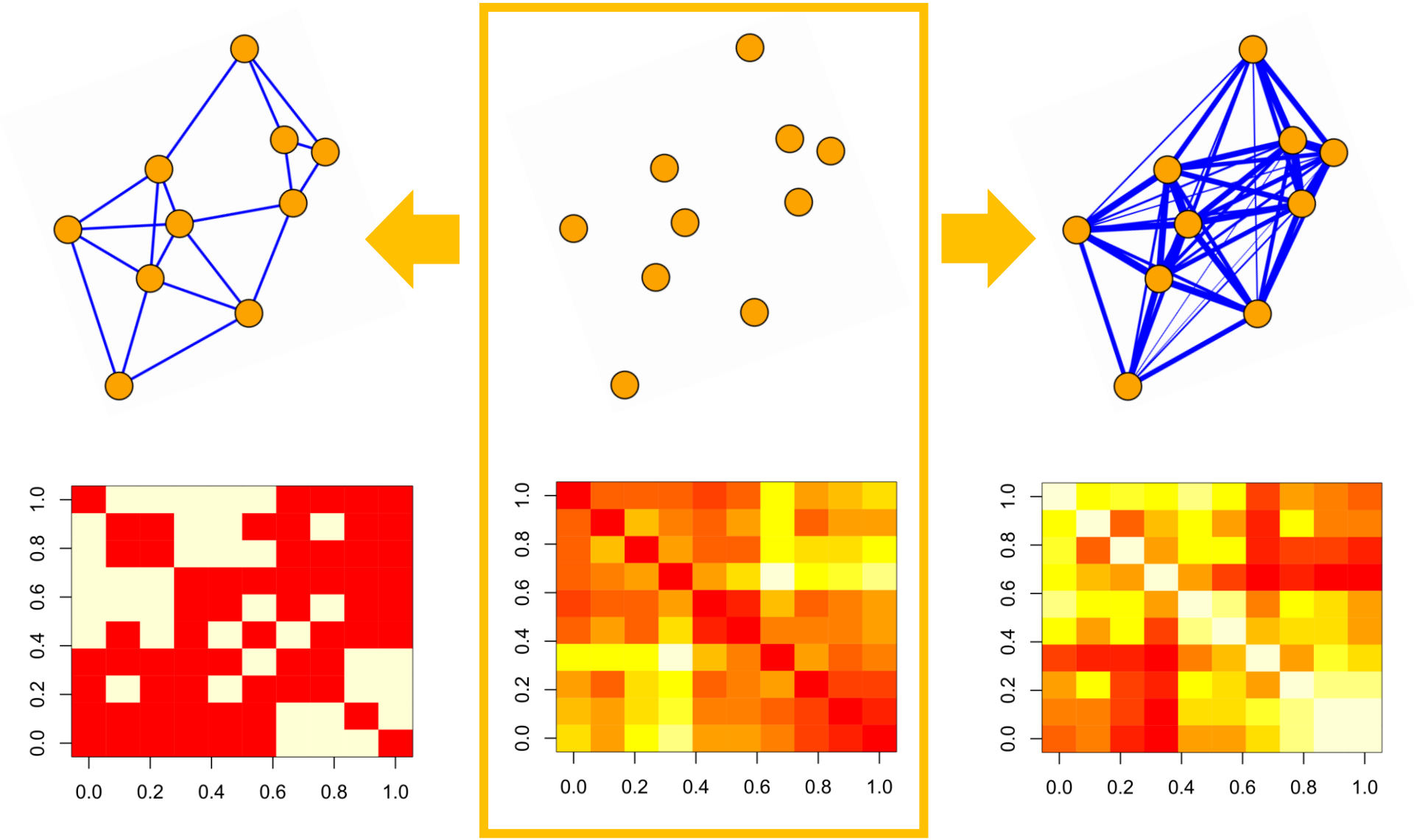
Similarity-based network



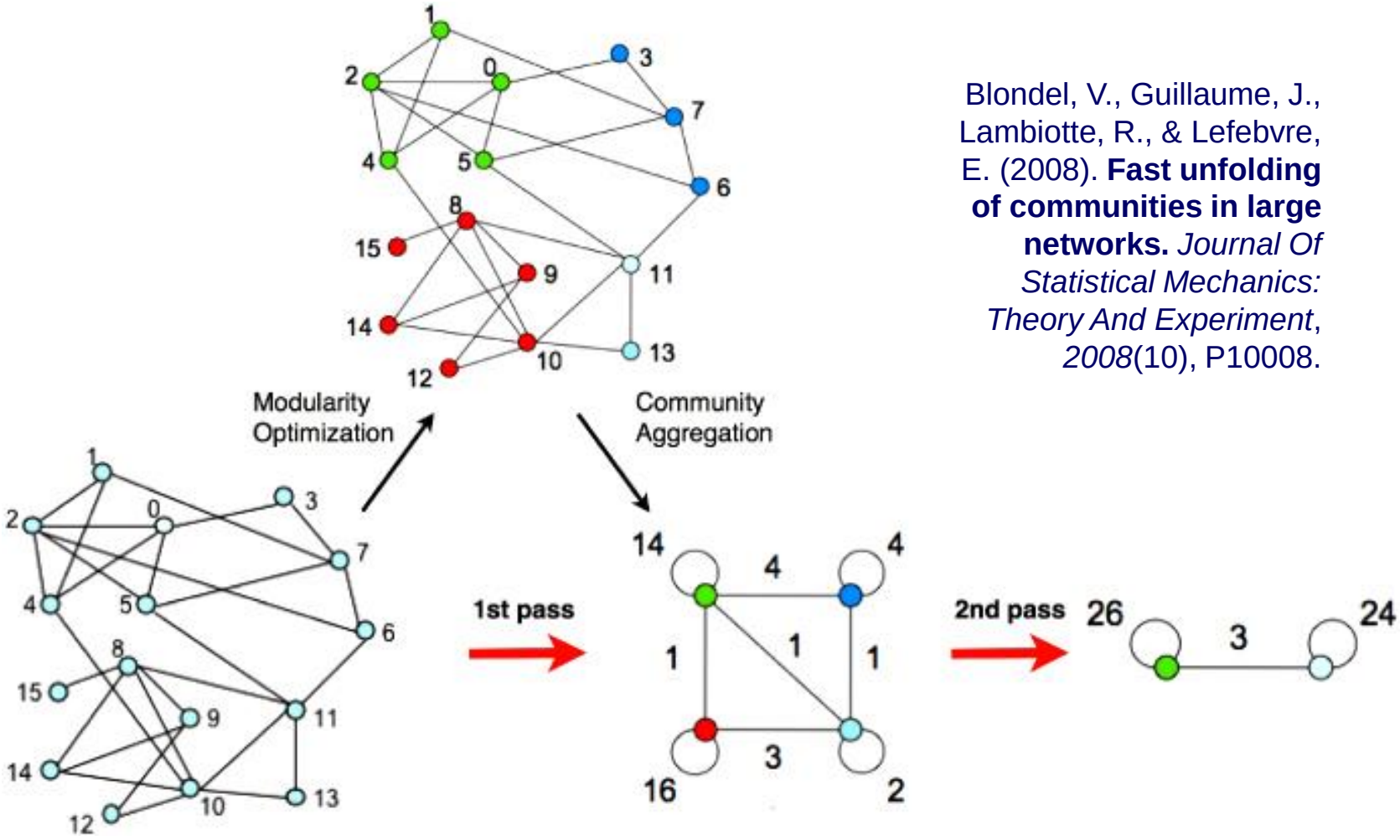




Similarity-based network

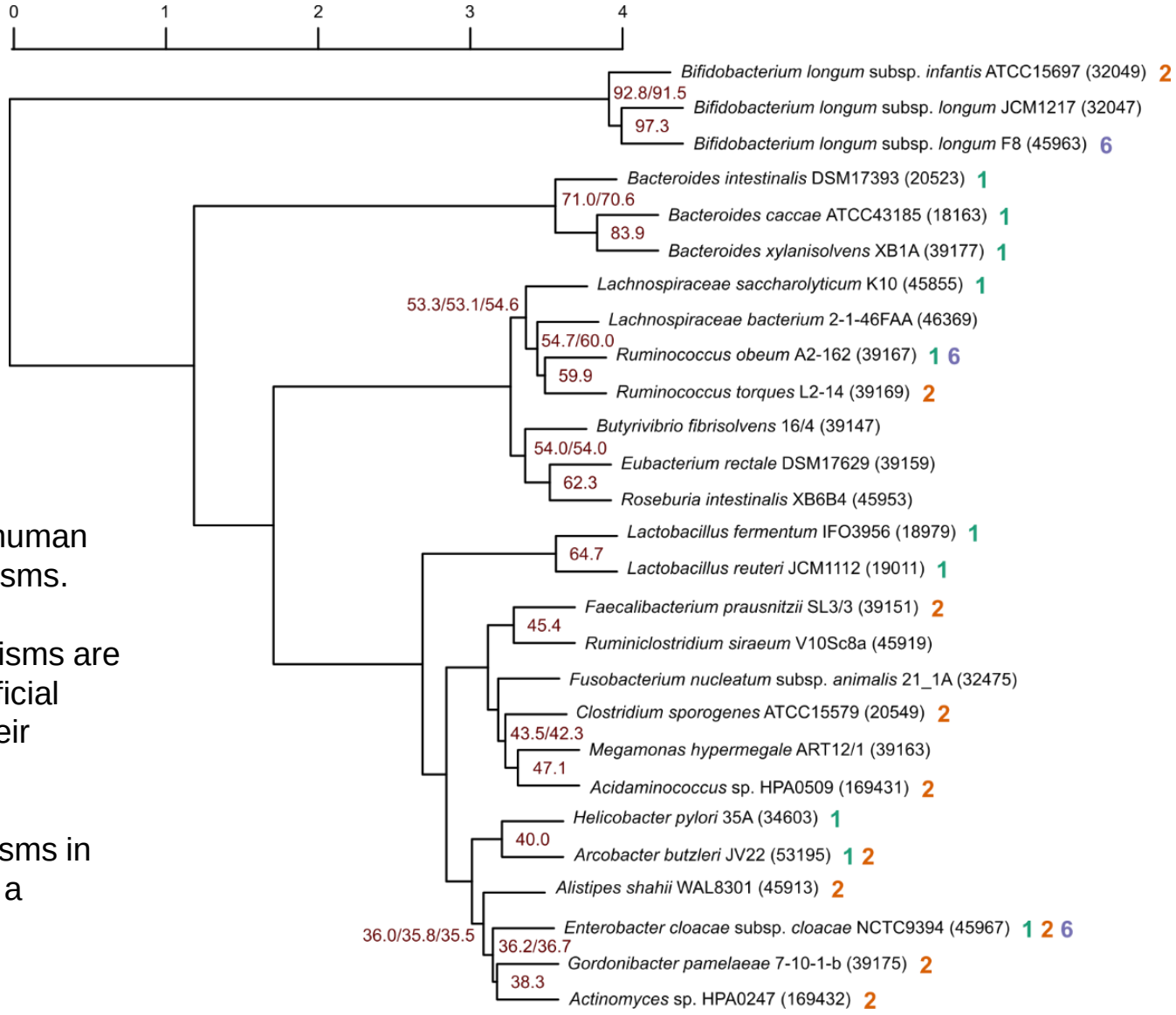


Community detection in networks



Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). **Fast unfolding of communities in large networks.** *Journal Of Statistical Mechanics: Theory And Experiment*, *2008*(10), P10008.

Application example



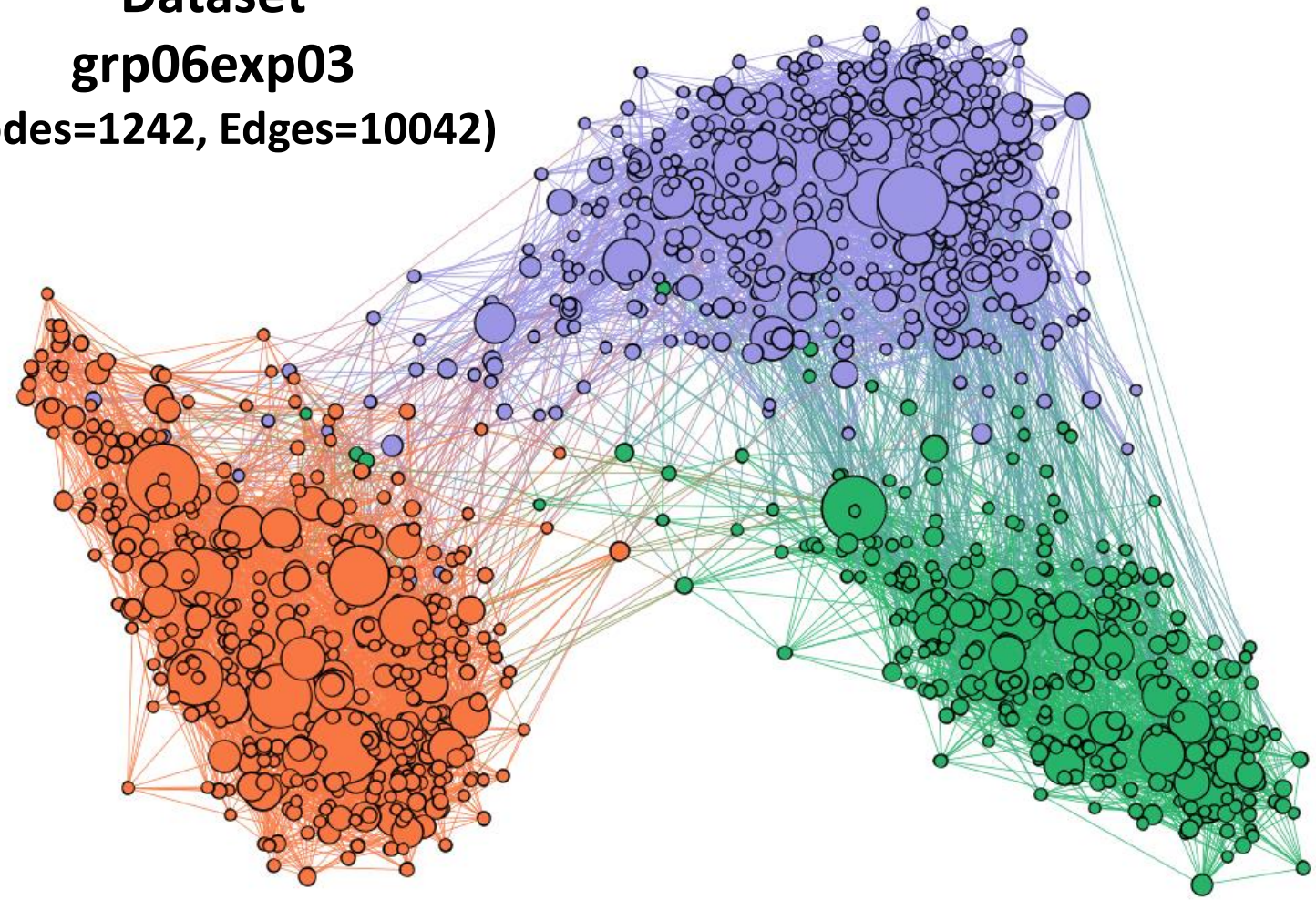
Phylogenetic tree of human intestinal microorganisms.

Numered microorganisms are used to construct artificial datasets based on their genome sequences.

Adjacent microorganisms in the tree are similar in a Phylogenetic sense.

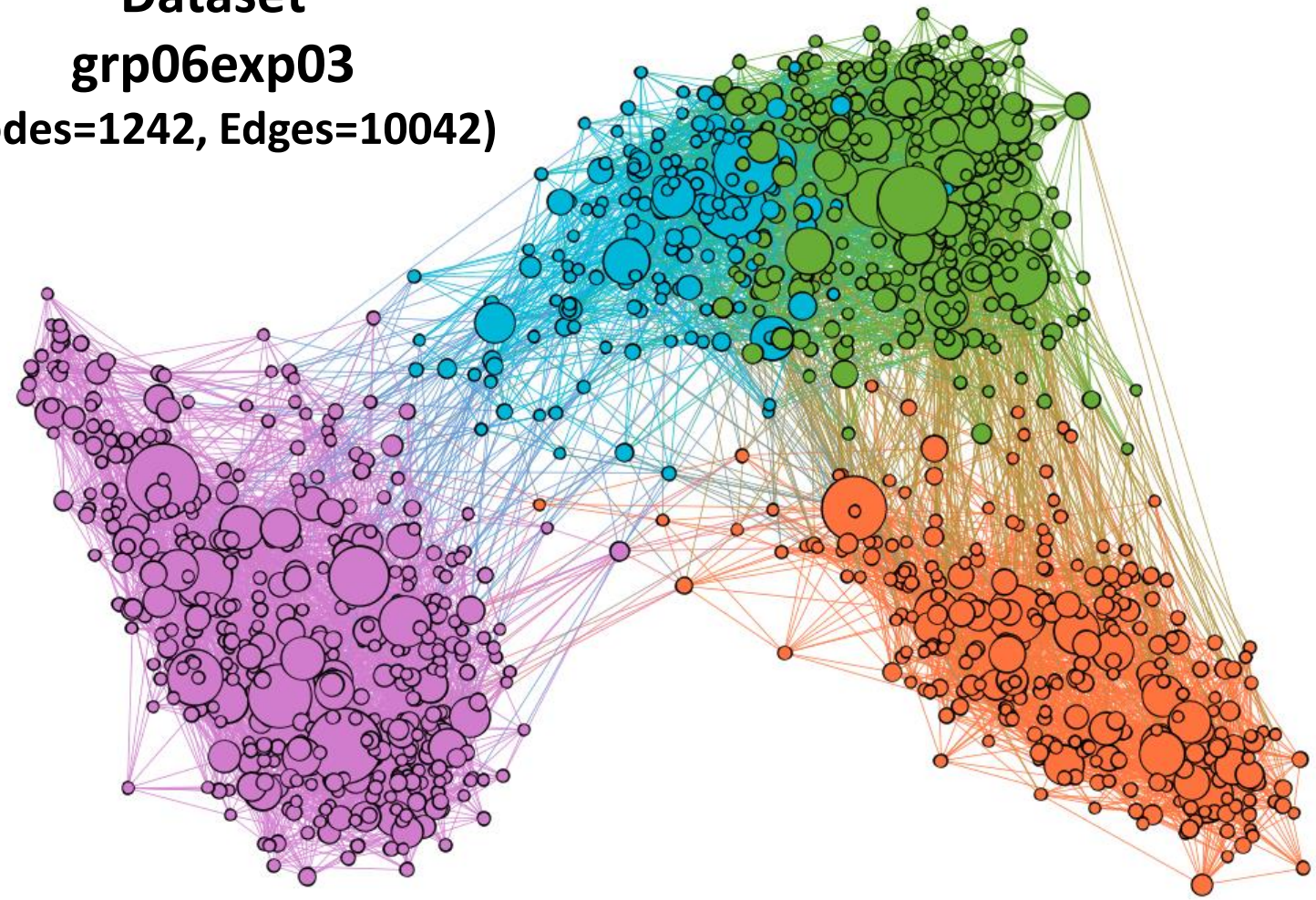
Application example: True comms (3)

Dataset
grp06exp03
(Nodes=1242, Edges=10042)

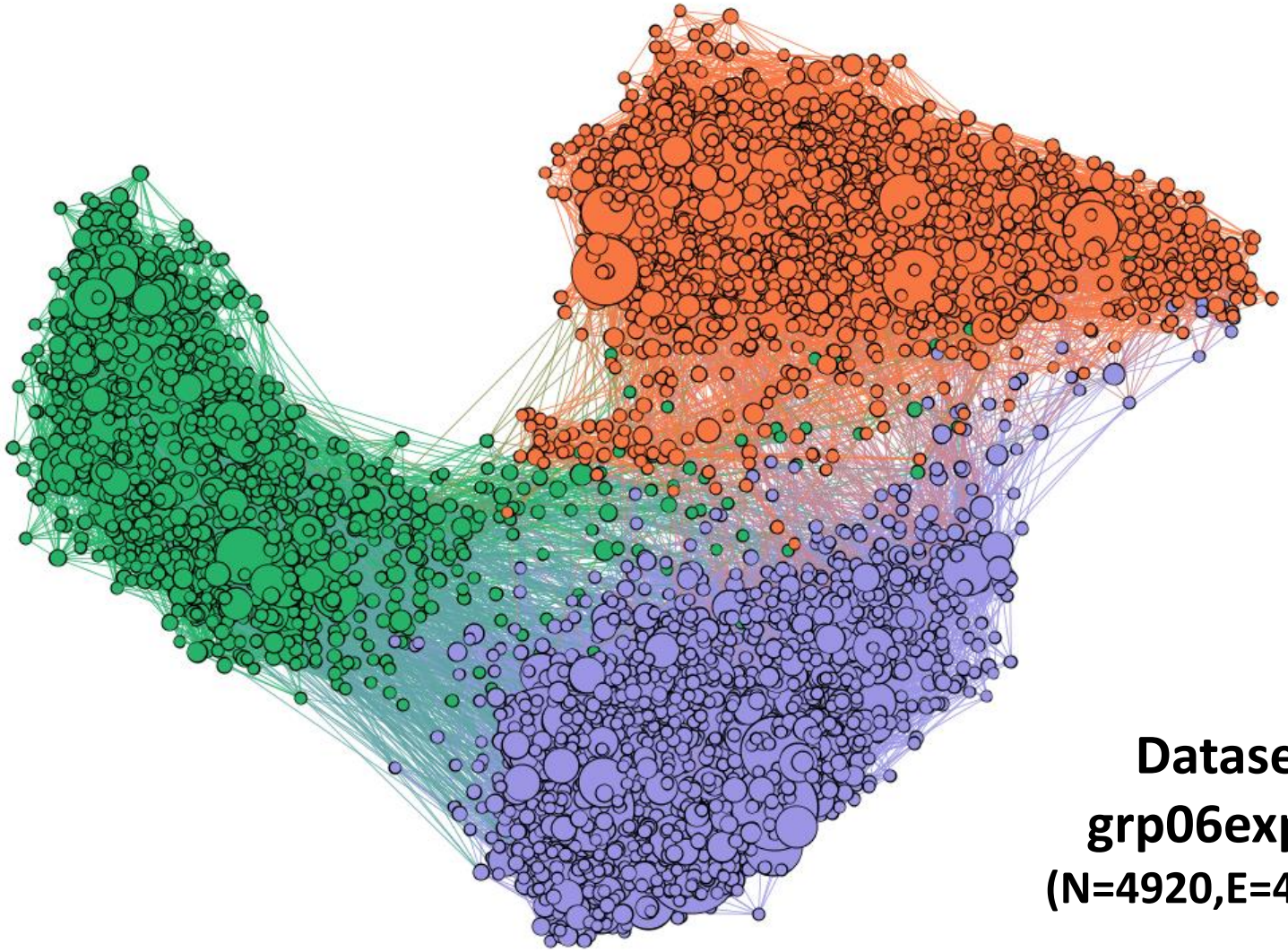


Application example: Louvain comms (4)

Dataset
grp06exp03
(Nodes=1242, Edges=10042)

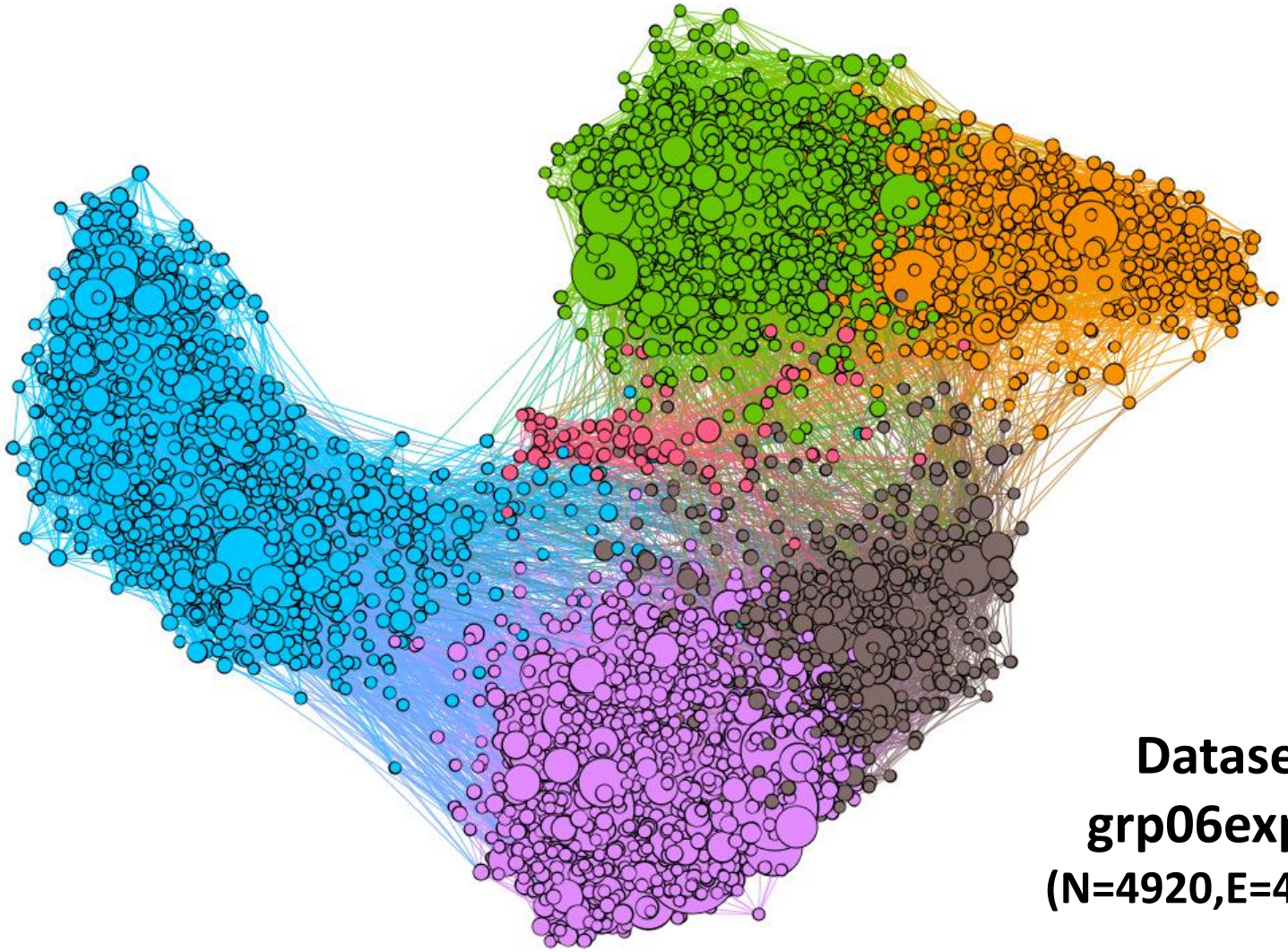


Application example: True comms (3)



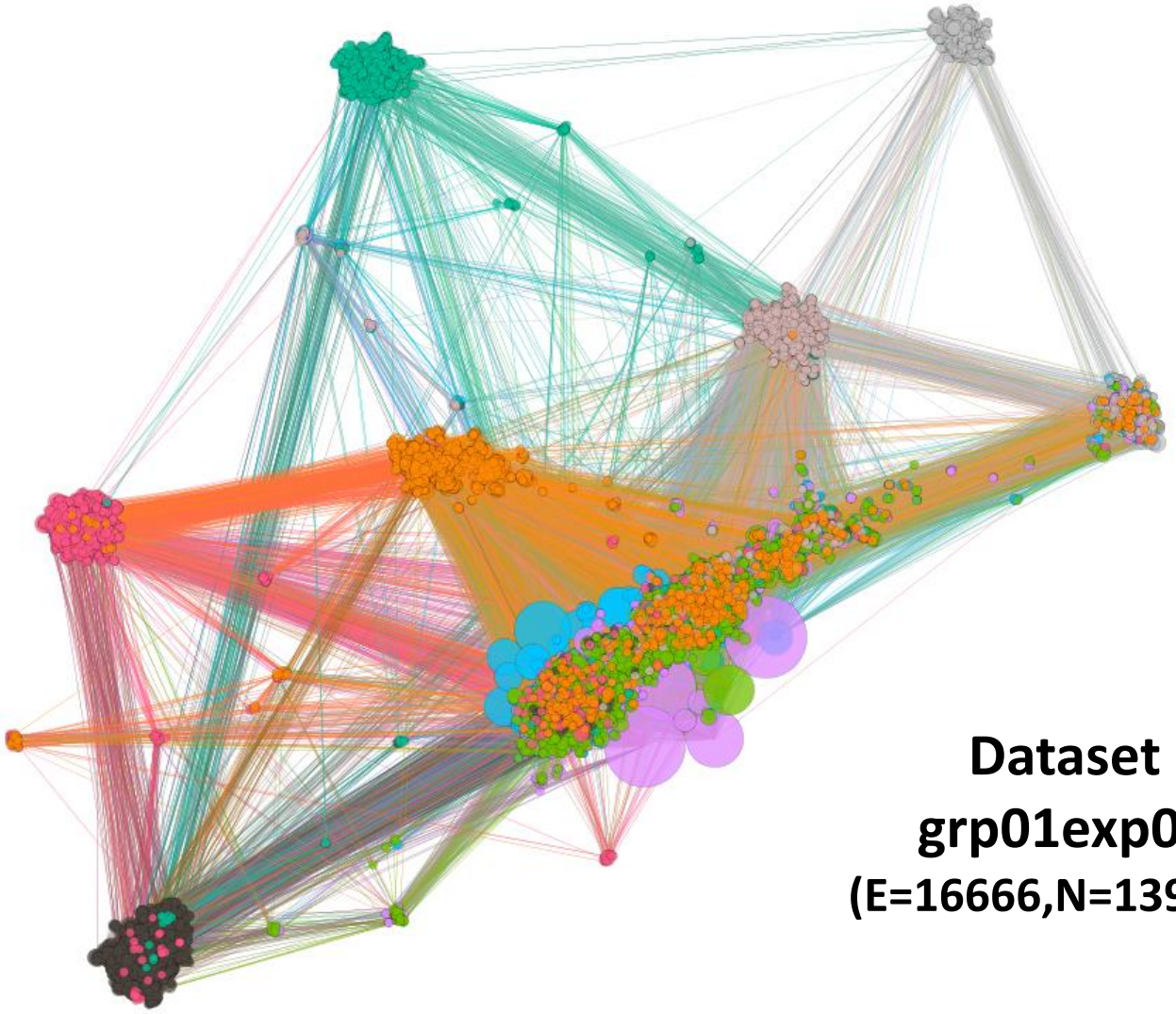
Dataset
grp06exp02
(N=4920,E=40680)

Application example: Louvain comms (7)



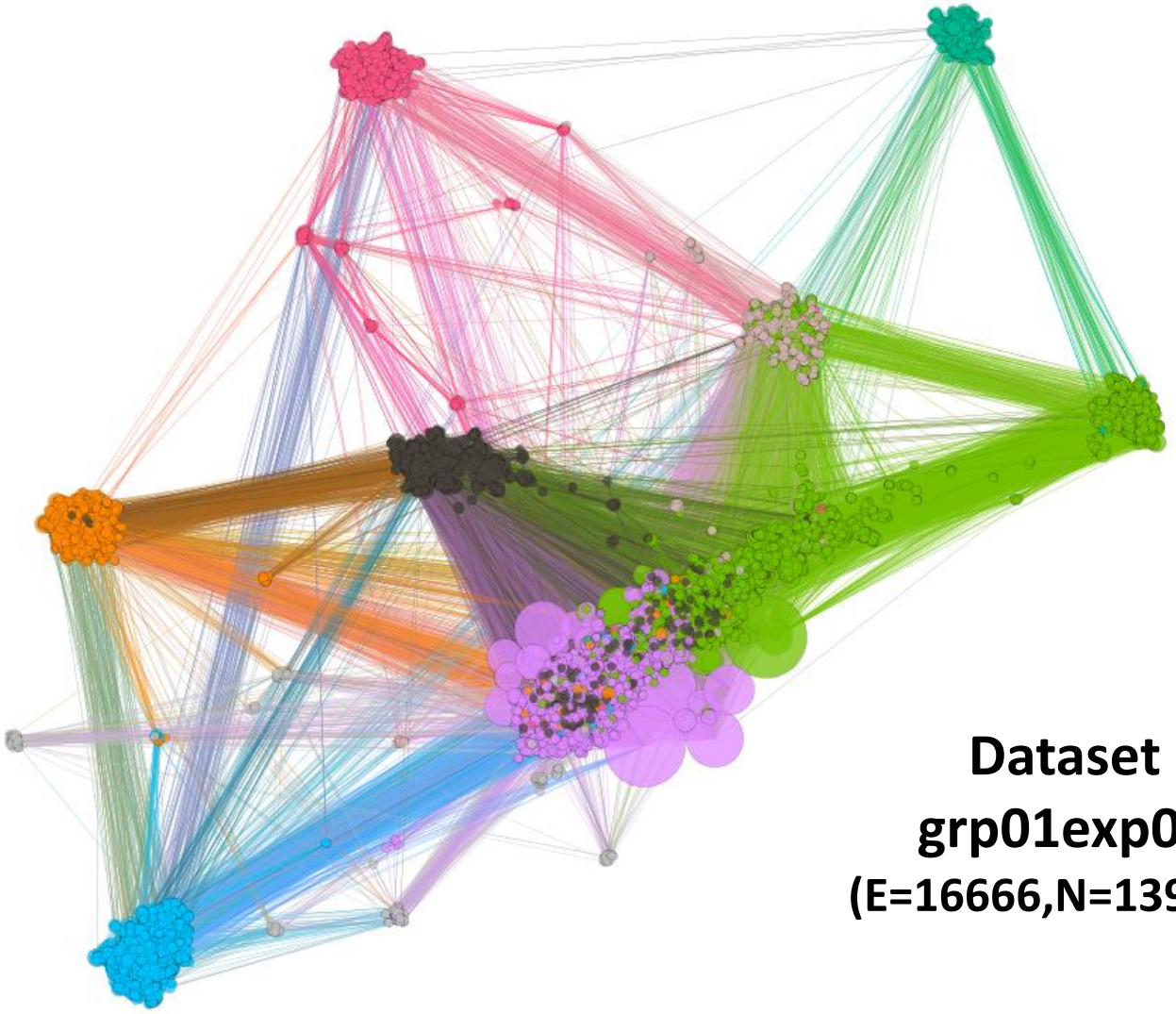
Dataset
grp06exp02
(N=4920,E=40680)

Application example: True comms (10)



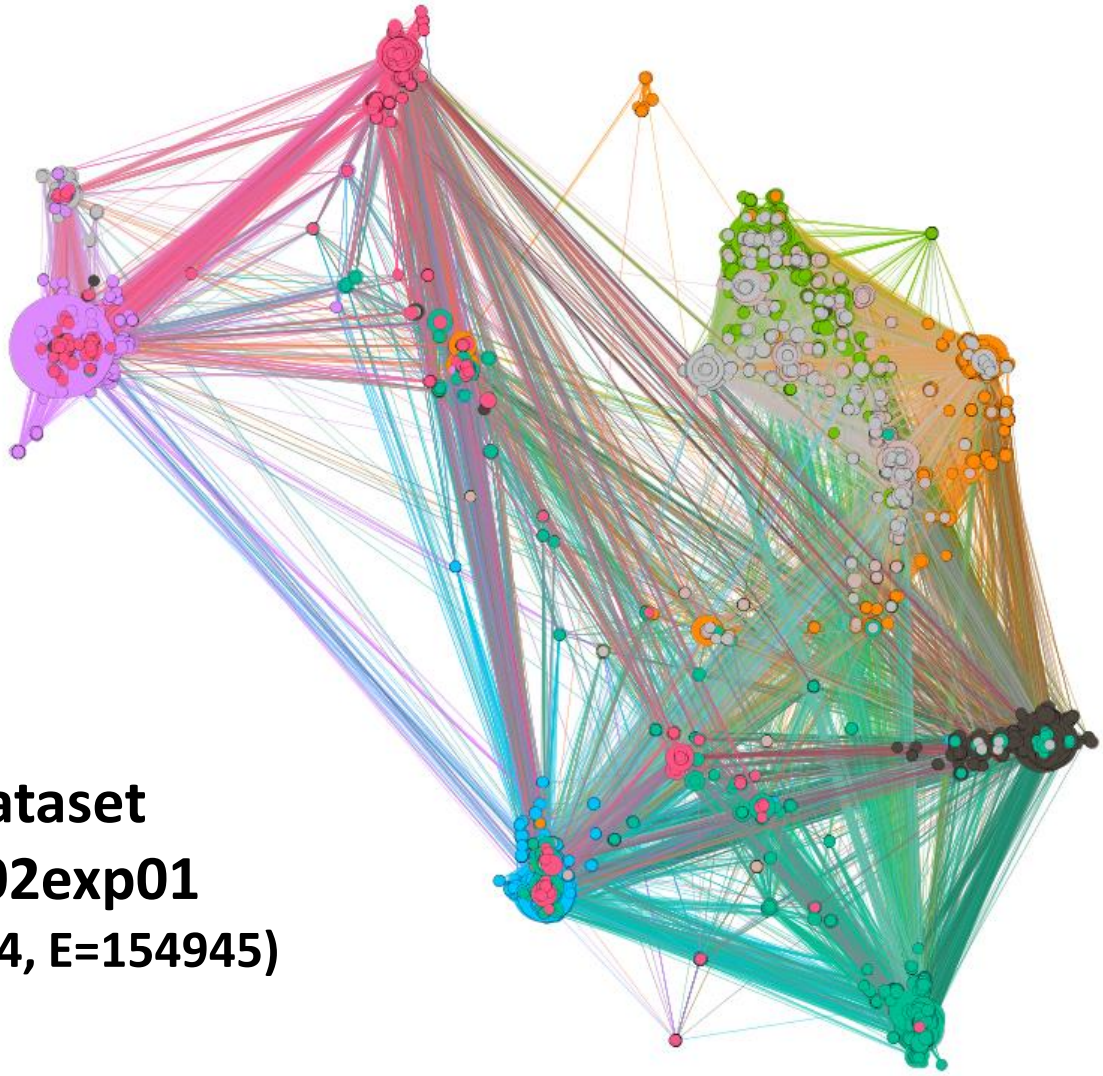
Dataset
grp01exp01
(E=16666,N=139619)

Application example: Louvain comms (10)



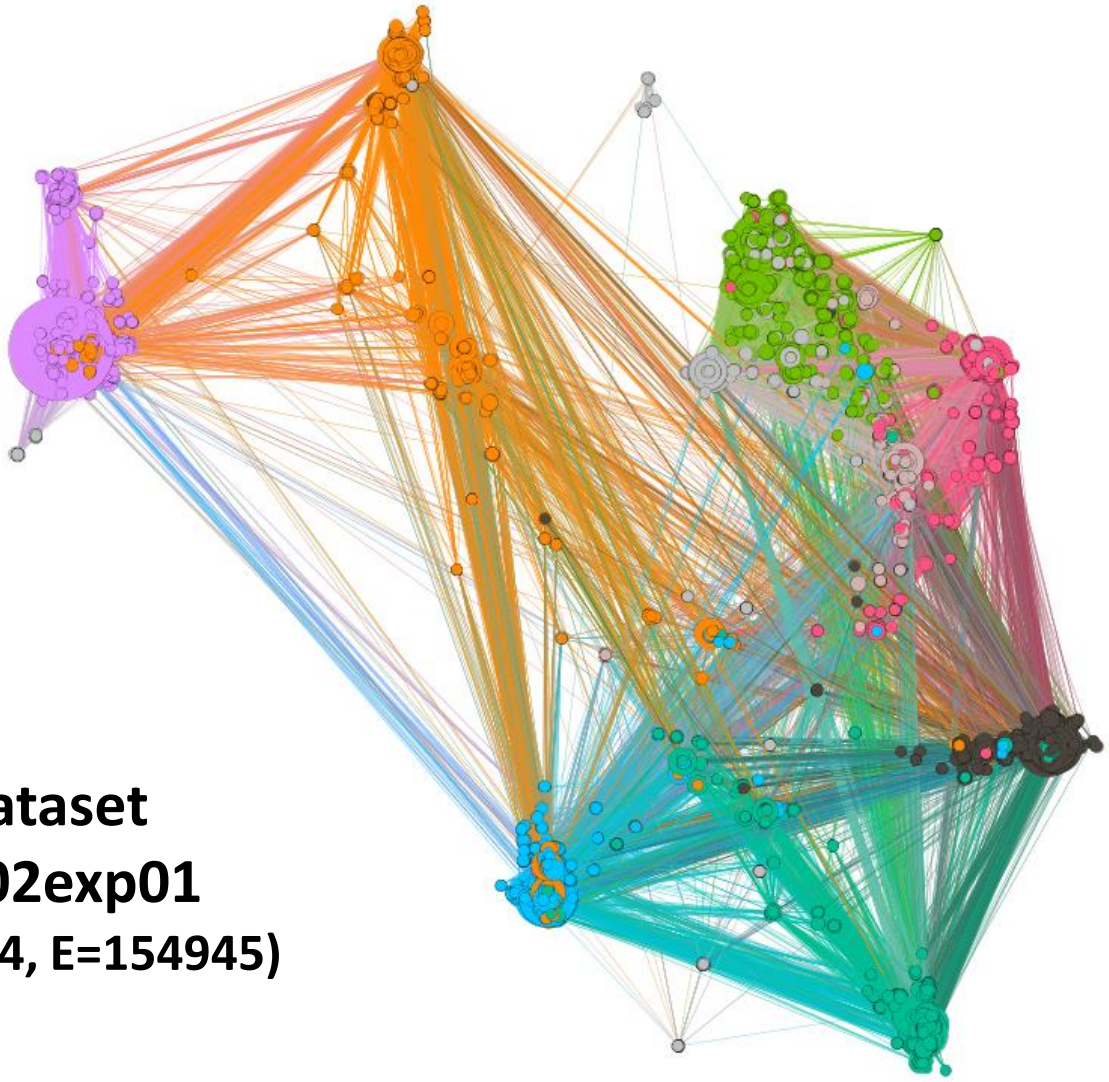
Dataset
grp01exp01
(E=16666,N=139619)

Application example: True comms (10)



Dataset
grp02exp01
(N=18474, E=154945)

Application example: Louvain comms (13)



Dataset
grp02exp01
(N=18474, E=154945)

Application example: Results

Community	Reference		CAA			NAA		
	K	SI	t	K	SI	t	K	SI
grp06exp03	3	0.127	15,439	3	0.133	0,572	4	0.076
grp06exp02	3	0.134	97,347	3	0.140	6,281	7	0.132
grp01exp01	10	0.079	598,244	10	0.090	76,958	10	0.094
grp02exp01	10	0.044	1659,469	10	0.060	94,530	13	0.106

K = Number of communities
t = computation time (s)
SI = Silhouette index

CAA = Cluster analysis approach
NAA = Network analysis approach

Future work

- Generate similarity networks based on different measure functions.
- Explore strategies to obtain sparse and weighted adjacency matrices.
  - E.g. Hybrid approach between Knn and heat kernel.
- Include complementary goodness metrics for community detection.
- Adopt/propose a community definition that represent the behavior of the metagenomic communities.
- Evaluate state-of-the-art algorithms for disjoint and overlapping community detection.

Conclusions

- Clustering of high-dimensional data can be performed following a network analysis approach.
- Network analysis can provide…
  - A direct representation of high-dimensional data
  - Methods for clustering data into communities without supervision
- The success of this approach depends on how is measured the similarity between objects in high-dimensional spaces.

References

- Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal Of Statistical Mechanics: Theory And Experiment, 2008(10), P10008. http://dx.doi.org/10.1088/1742-5468/2008/10/p10008
- Coscia, M., Giannotti, F., & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. Statistical Analysis And Data Mining, 4(5), 512-546. http://dx.doi.org/10.1002/sam.10133
- Geoff Dougherty. Pattern Recognition and Classification. Springer New York, 2013.
- Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. Physics Reports, 659, 1-44. http://dx.doi.org/10.1016/j.physrep.2016.09.002

References

- Laczny, C., Pinel, N., Vlassis, N., & Wilmes, P. (2014). Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. Scientific Reports, 4(1). http://dx.doi.org/10.1038/srep04516
- von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics And Computing*, *17*(4), 395-416. http://dx.doi.org/10.1007/s11222-007-9033-z
- van der Maaten, Laurens. (2013). Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research, 15, 3221−3245.