Patient heterogeneity assessments via network-based ANOVA

Diane Duroux[1] , Federico Melograna[2] , Kristel Van Steen[1,2]

[1]BIO3-GIGA R Medical Genomics, University of Liège, Belgium, [2]BIO3-Department of Human Genetics, KU Leuven, Belgium

Contact: diane.duroux@uliege.be

POSTER 282

Motivation

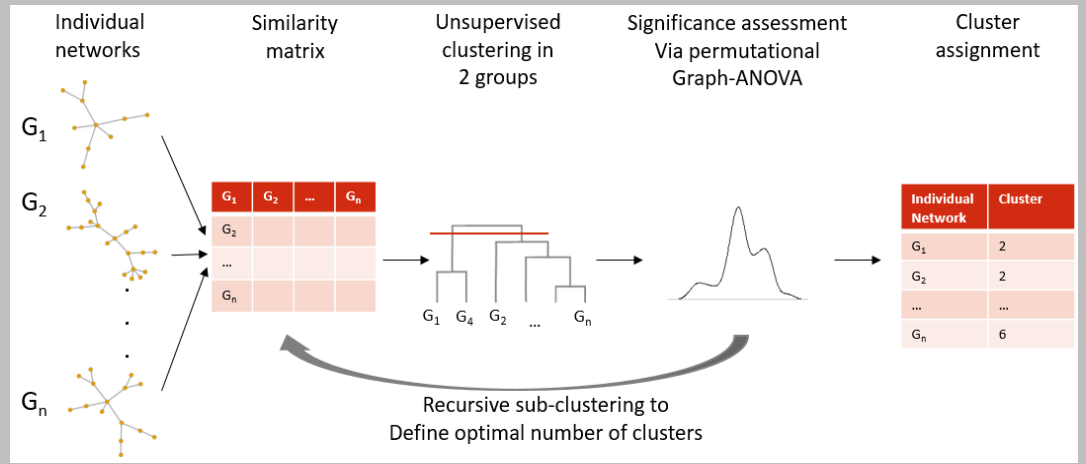
Context:

- Heterogeneity between individuals lies at the core of precision medicine.
- Graph unsupervised classification is relevant to identifying networks that can be aggregated in a cluster and can be used for disease subtyping.

Ideas:

- ➢ Represent information about individuals as networks characterized by individual specific edge and nodes values.
- ➢ Determine the clustering based on notions of statistical significant differences between clusters.
- ➢ Use a statistic relevant for networks.

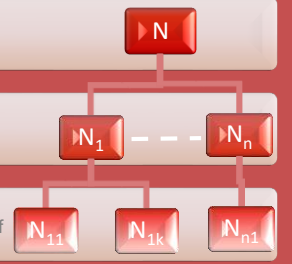
Methodology



Data

Simulations: Baseline

Original network	Erdos-Renyi 100 binary nodes Density 0.01
Group networks via perturbations	4 groups Degree preserving switch of 25% of original edges
Individual networks via perturbations	10 individual networks Degree preserving switch of 25% of edges of group net

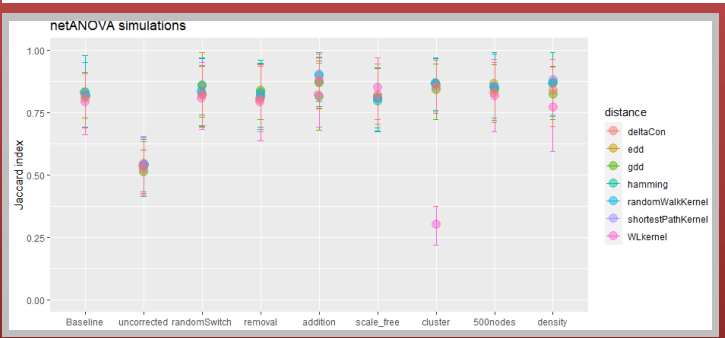


➢ Goal: find the groups to which the networks belong.

Real life: We used *Chronic obstructive pulmonary disease* data. We derived the individual network from the gene expressions of 1000 genes for 100 cases and 100 controls.

➢ Goal: Identify cases and controls

Results



Variations tested:

- p-values not corrected for multiple testing
- Perturbation via removal, addition or random switch of edges
- Network structures: scale-free and cluster networks
- 500 nodes and density 0.005 of the original network

Real data:

The gaussian kernel is computed on the vectorized adjacency matrices with $\sigma = 1000$. The obtained kernel matrix is used in the network-based ANOVA algorithm where the hierarchical clustering algorithm is applied to detect groups. 11 clusters are identified.

As a baseline comparison, we applied the kmeans alogorithm to cluster individuals. Here the input data contains the vectorized adjacency matrices of all the indidivuals networks. The same number of clusters was specified.

$$VI(network-based\ ANOVA) = 2.5$$
$$VI(kmeans) = 3.6$$

The variation of information measures the amount of information lost and gained in changing from the observed (cases and controls) and the inferred clusterings.

The network-based ANOVA algorithm decreases the variation of information and hence, improves the performance.

Conclusion

Future aspects:

a. Evaluate the influence of the covariates on the final clusters.
b. Estimate the impact of the number of individuals, groups and the level of heterogeneity between and within groups.
c. Investigate the added benefit of graph representation.
d. Estimate type I error.

Our strategy pave the way towards deciding which networks can sensibly be aggregated while providing a statistical measure of the difference between groups.

This is not only relevant in stratified medicine, but also in areas such as genetic epistasis detection in which conclusions need to be drawn from multiple statistical epistasis networks across different analysis protocols.

This project has received funding from the European Union's Horizon 2020 research and innovation 421 programme under the Marie Sklodowska-Curie grant agreements No 813533 and 860895