ROBIN:

(ROBUSTNESS IN NETWORK)

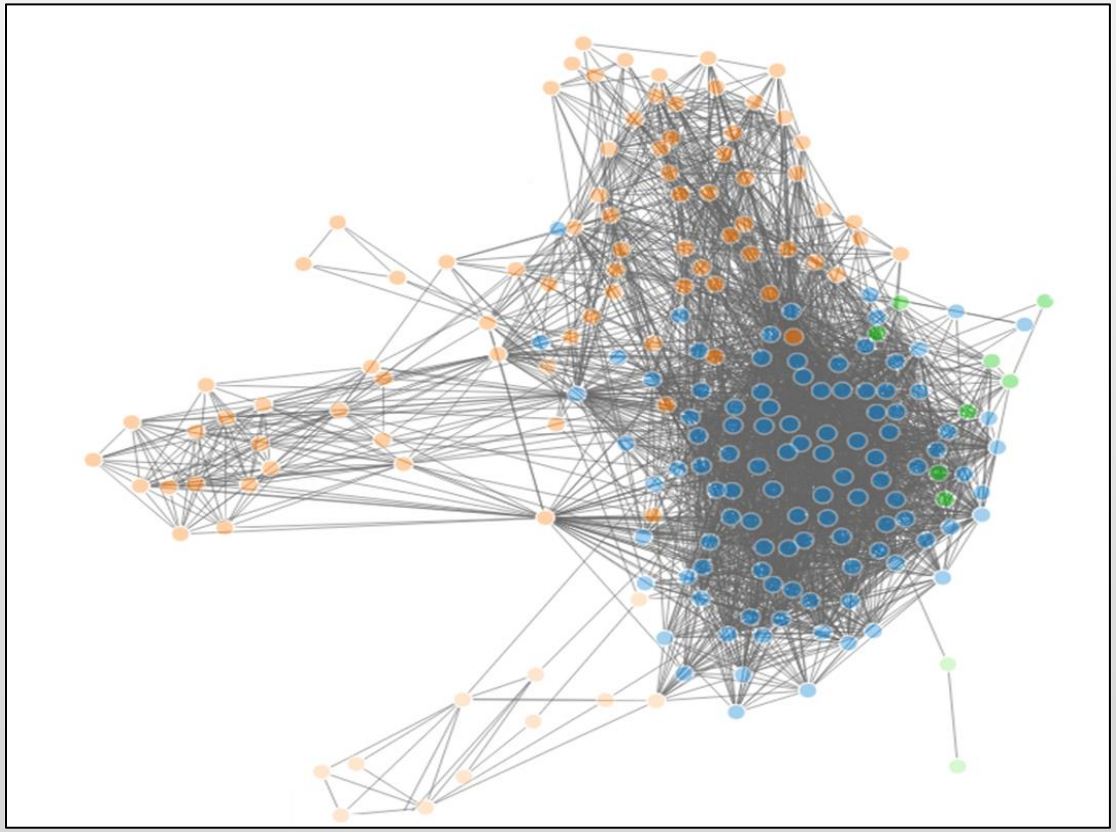
AN R PACKAGE FOR VALIDATION OF COMMUNITY ROBUSTNESS

Valeria Policastro

Istituto per le Applicazioni del Calcolo M.Picone CNR- Naples (Italy)

COMMUNITY DETECTION

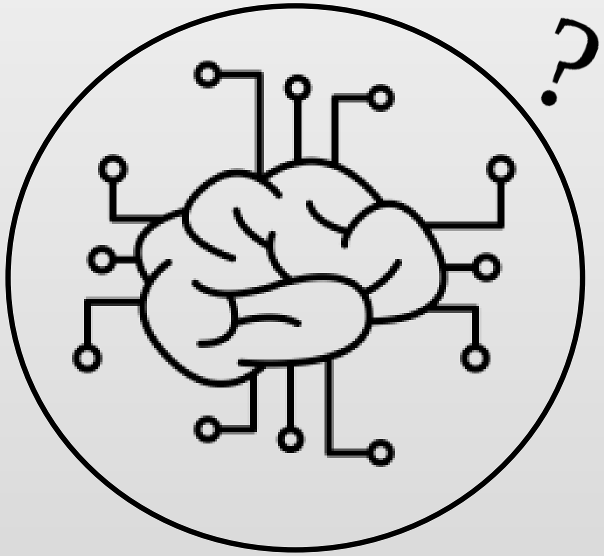
- One of the most relevant features of graphs representing real systems is their **community structure**



- **How can we say that is statistically robust?**

Mc Auley, J., Leskovec, J., 2012. Learning to Discover Social Circles in Ego Networks. NIPS. pp. 548–556

ROBUSTNESS



Are the detected communities significant or are they a result of chance only due to the positions of edges in the network?

ROBIN: (ROBustness In Network)

- An R package that gives a statistical answer to the ***validation of the community structure*** by looking at the robustness of the network

Carissimo A., Cutillo L., De Feis I., Validation of community robustness Computational Statistics and Data Analysis 2017

ROBUSTNESS AGAINST RANDOM PERTURBATION

❖ If a partition is significant, it will be recovered even if the structure of the graph is modified

ROBIN

- ***ROBIN*** gives the possibility to analyze community detection in all different aspects:

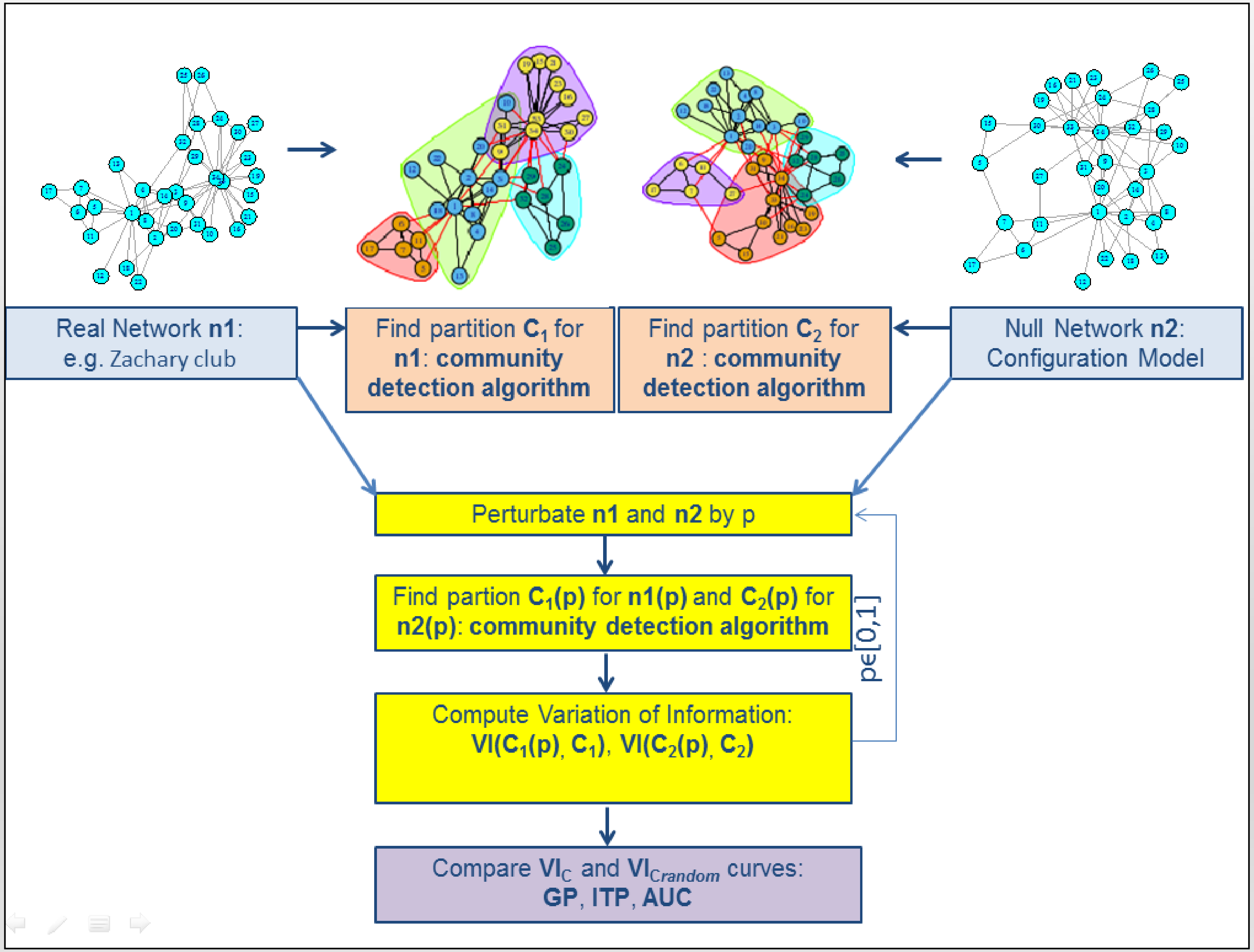
  - ✓ Community detection algorithms
  - ✓ Validation of the community structure
  - ✓ Comparison of different community algorithms
  - ✓ Graphical interactive representation of 3D networks

PROCEDURE

- Given a **community detection method** and a **network** of interest ROBIN analyses the stability of the partitions

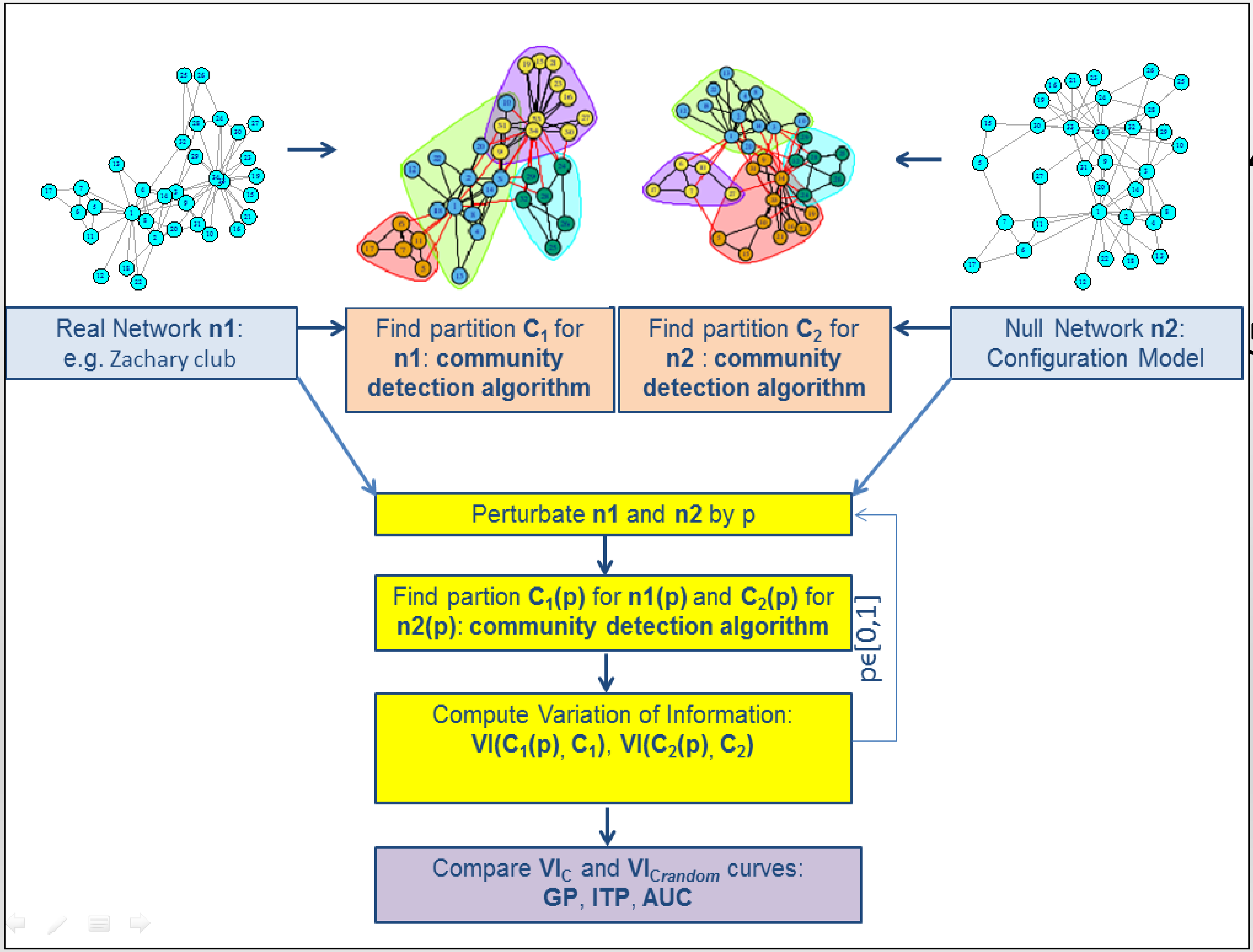
- It implements a **perturbation strategy** and a **null model** to build a procedure based on **Variation of Information**

ROBIN WORKFLOW



1. Given a network, find a partition ***C*** with some algorithm
2. Perturb the network to create a new network, find the partition ***C'*** for the perturbed network, compare ***C*** and ***C'*** computing the Variation of Information ***VI***
3. Repeat this calculation for different perturbation levels ***p***

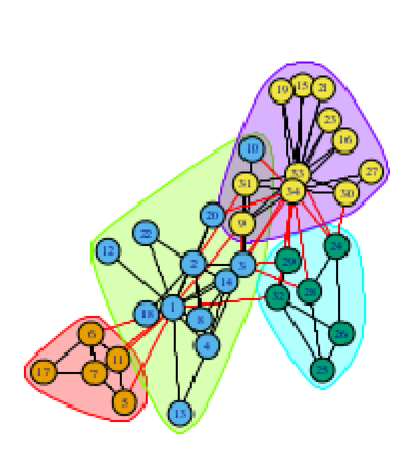
ROBIN WORKFLOW



4. Perform 2. and 3. on a random graph (*null model*)

5. Compare the curves obtained plotting the average ***VI*** versus ***p*** of the original and the null model

ROBIN PERTURBATION STRATEGY



❖The perturbed network has the same number of vertices and edges as the original unperturbed network

p*ϵ*[0,1]

p=0 The original unperturbed graph

p=1 The maximal perturbation level (*random graph*)

ROBIN PERTURBATION STRATEGY

- **Rewire algorithm**: chooses two arbitrary edges in each step (e.g. (a,b) and (c,d))and substitutes them with (a,d) and (c,b) if they do not already exist in the graph

- **keeping_degseq**: preserving the original graph's degree distribution

rewireComplete(graph, number)

ROBIN PERTURBATION STRATEGY

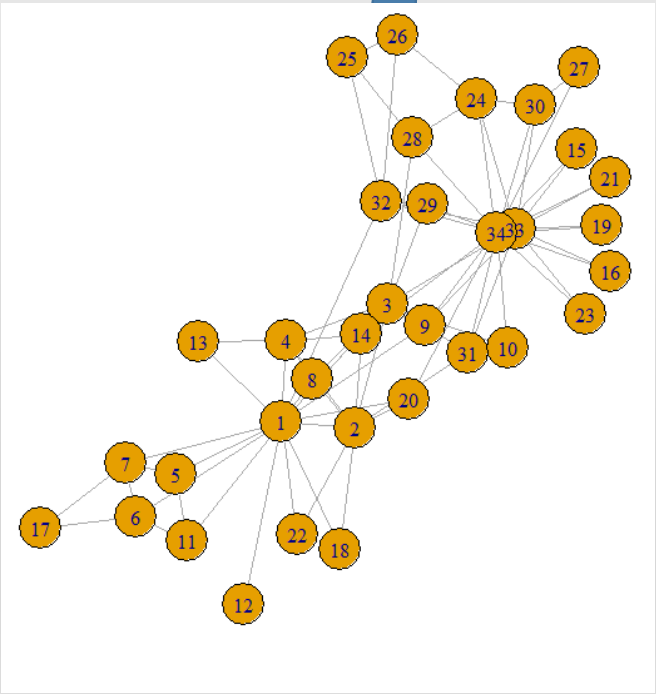
I° level of p

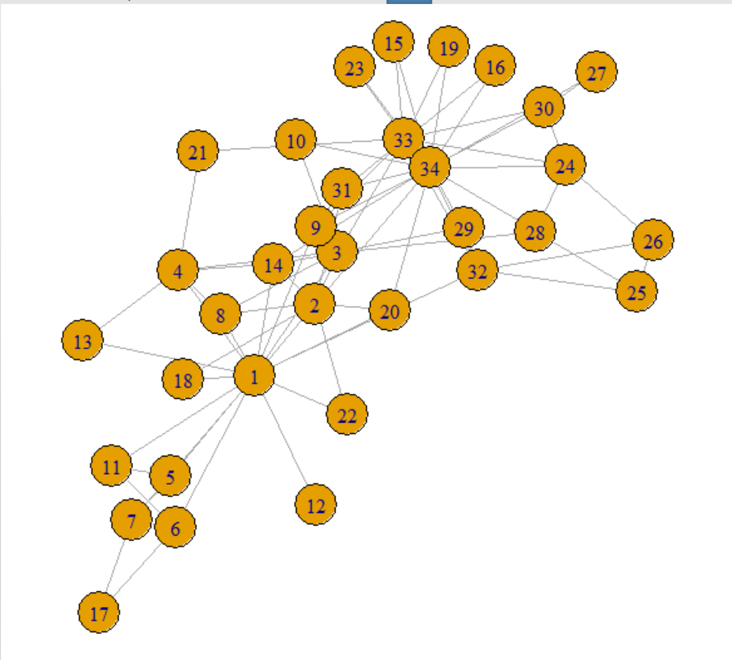
5%

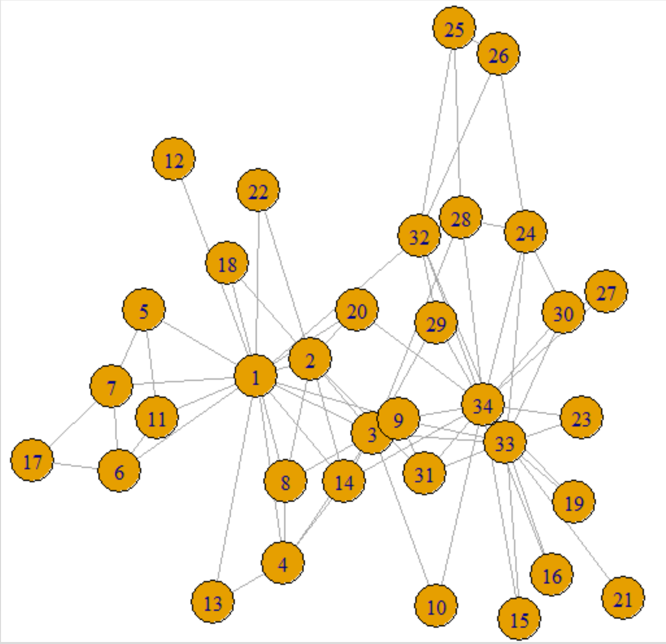
Remove edges rewired

II° level of p

5%







…until is all perturbed

ROBIN PERTURBATION STRATEGY

Generates 20 levels of *p*

- For each different level it generates 10 perturbed graph. Then, from each of the obtained graphs, it generates other 10 graphs rewiring 1% of edges each time

Result: 100 graphs for each level !!!

Configuration Model :

- Able to preserve strongly heterogeneous degree distribution of the real network.
- It randomly assign edges between vertices with a given degree distribution.

graphRandom(graph)

Bender,E.A., Canfield,E.R., 1978. The asymptotic number of labelled graphs with given degree sequences. J.Combin. Theory A24, 296–307

VARIATION OF INFORMATION (VI)

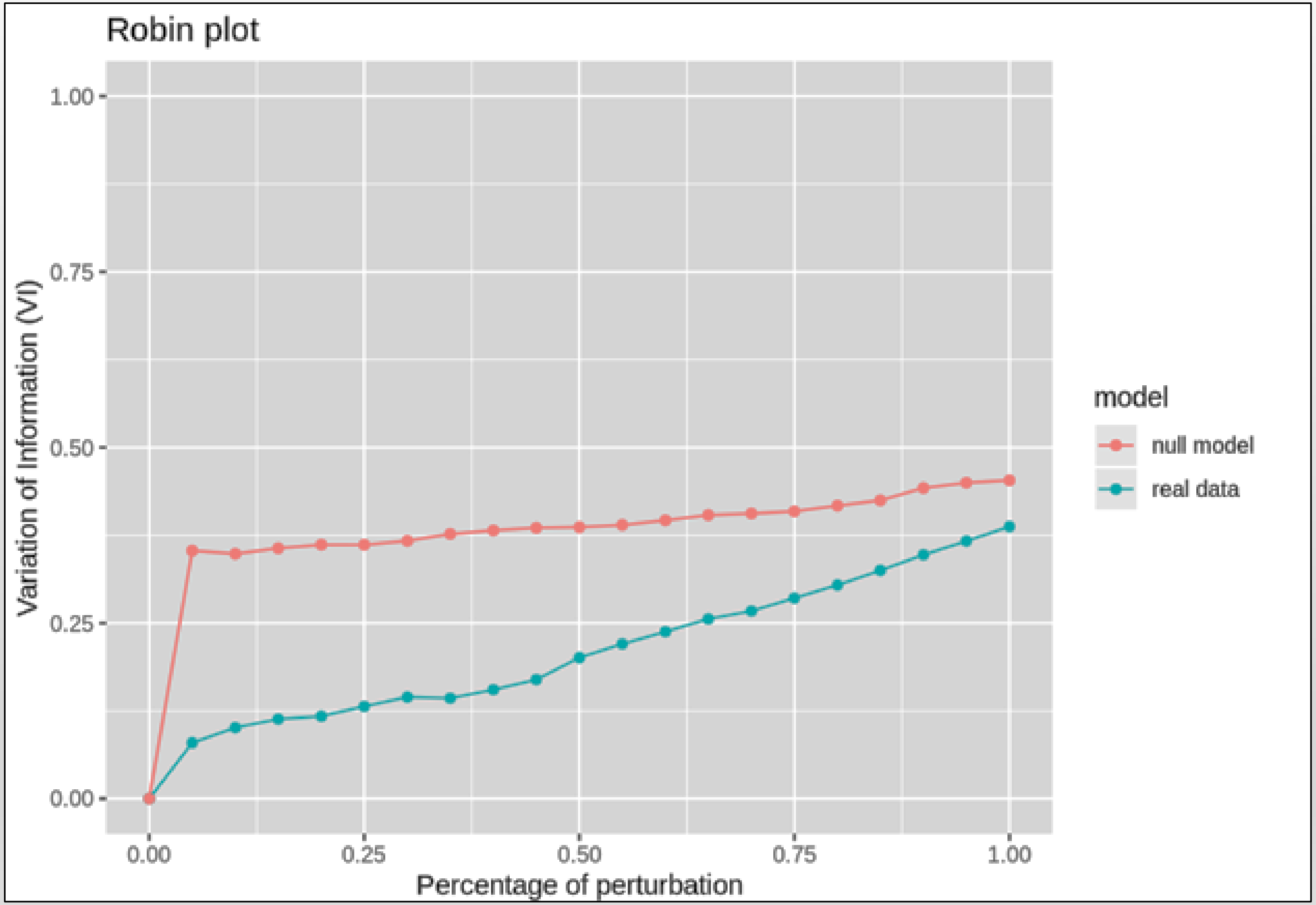
- At each level of perturbation ***p*** ROBIN compares the partition obtained from the original graph with the partition obtained from the perturbed graph computing Variation of Information (***VI***)

compare(comReal, comR, method="vi")

comReal = Community Real
ComR = Community Rewire at a specific level

Meilă M.,2007. Comparing clusterings-an information based distance. J. Multivariate Anal.98,873-895

VI AS FUNCTION OF AMOUNT OF PERTURBATION *P*



plotRobin (graph,model1,model2, legend)

model1= mean vi real data
model2= mean vi null model

How can ROBIN test the differences between the two curves?

ROBIN TESTS

- The percentage of perturbation as time points of two time series

Gaussian Process(GP)

- Functional data analysis

Interval Testing Procedure (ITP)

- Compare the area under the curves

AUC

GAUSSIAN PROCESS

❖Are the two curves from the same process or not?

Gprege package: Gaussian Process Ranking and Estimation of Gene Expression time-series

callGp (ratio)

ratio= log2(viMean/viMeanRandom) *for each level*

Bayes Factor: score based on the log-ratio of marginal likelihoods

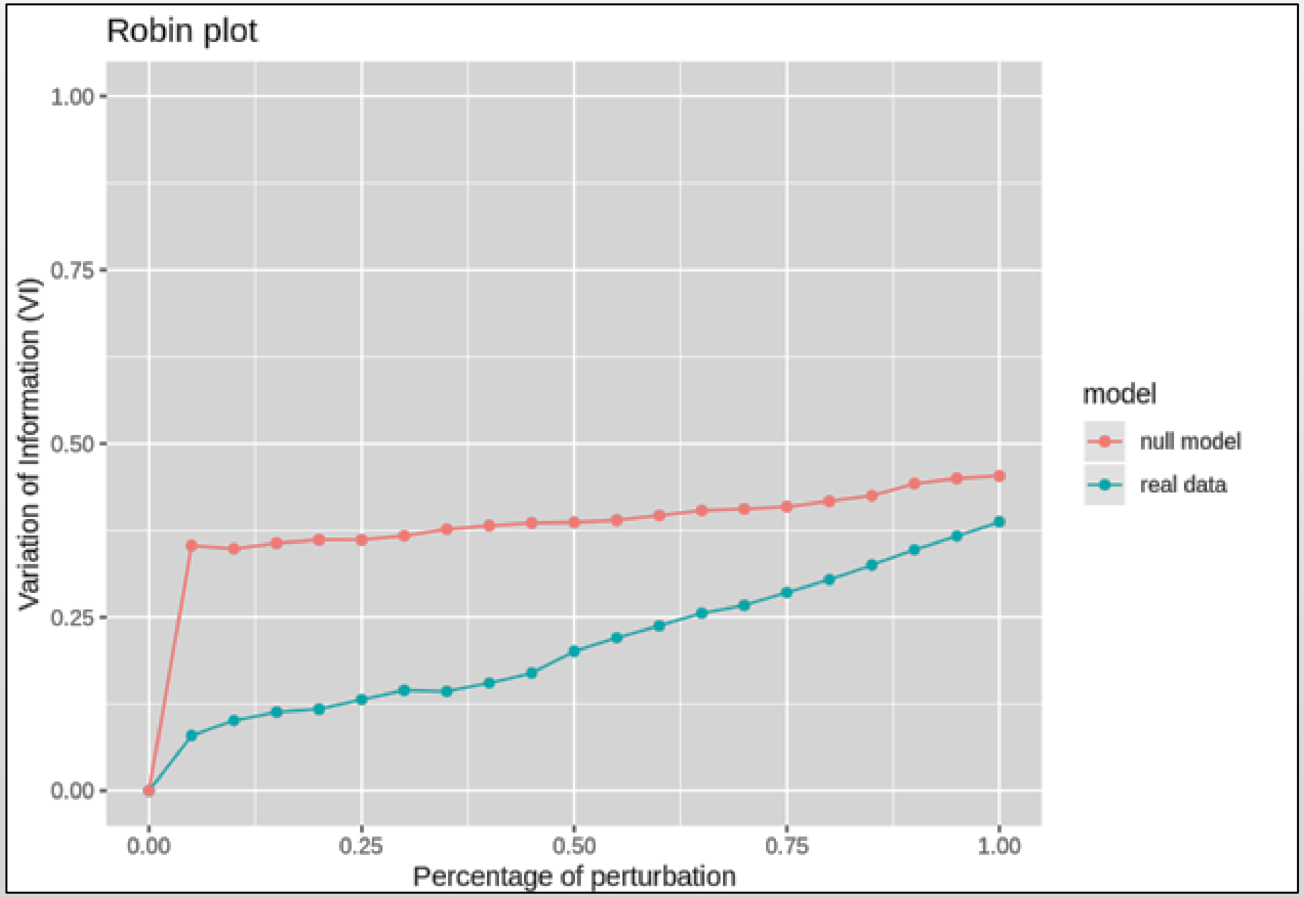
bf=gpregeOutput$rankingScores[1]

Alfredo A. Kalaitzis and Neil D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. BMC Bioinformatics, 12(180), 2011.

GAUSSIAN PROCESS

Bayes_Factor 383.8757

K	dHart	bits	Strength of evidence
$< 10^0$	0	—	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive



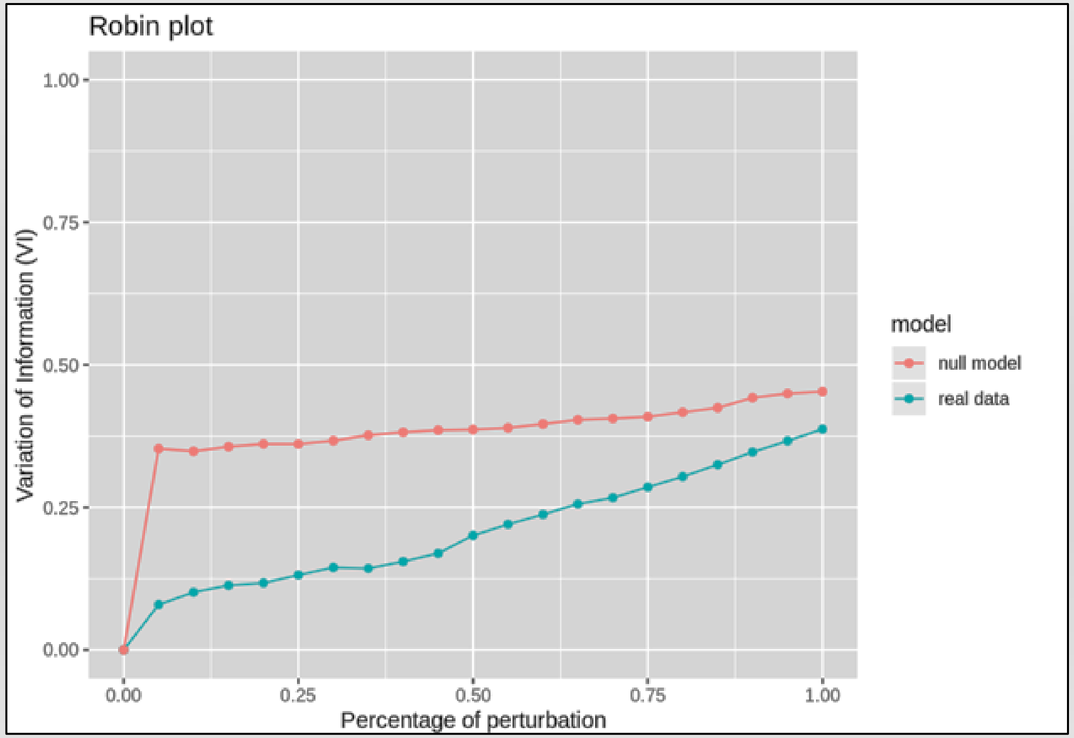
AUC

❖Calculate the area under the curve with a spline approach

DescTools package: Tools for Descriptive Statistics

AUC(x=percPerturb, y=mvimeanmodel1, method ="spline")

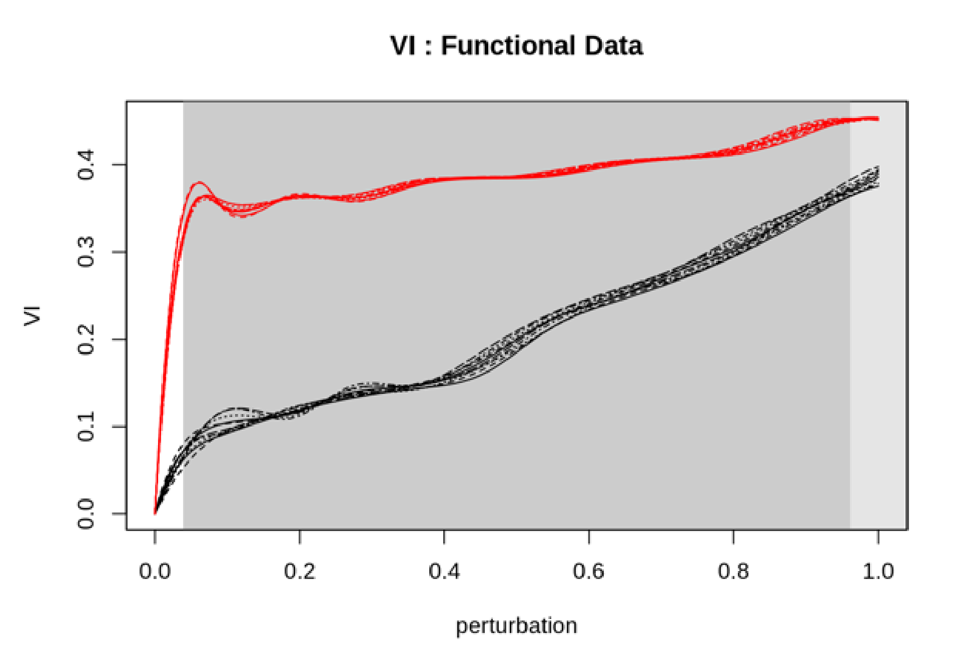
```
Area Under the Curve:
      Area1 0.2082588
      Area2 0.3840378
```

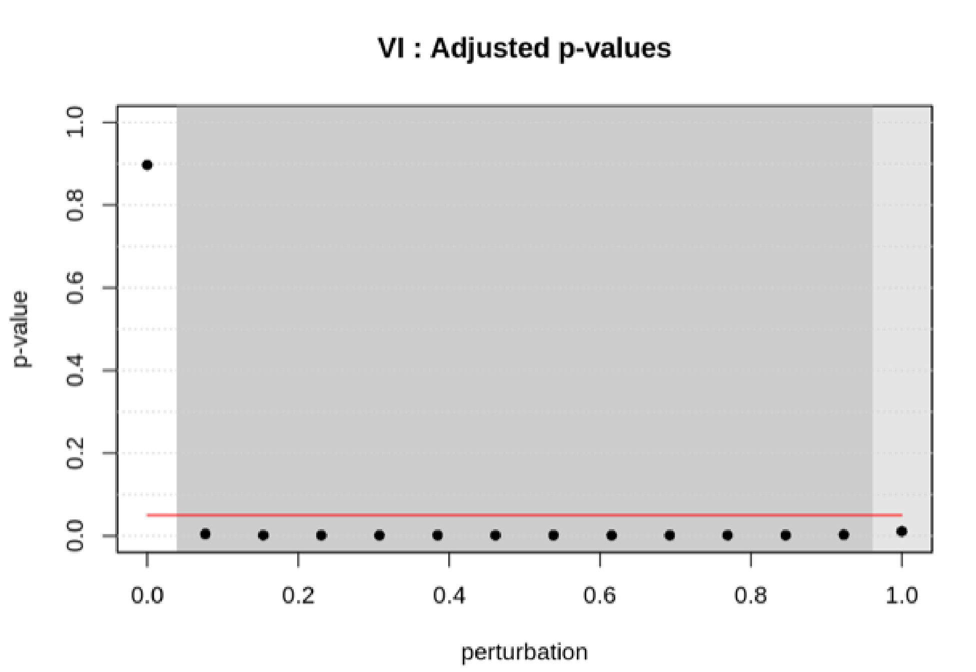


INTERVAL-WISE FUNCTIONAL TESTING PROCEDURE

- Provides an interval-wise non parametric functional testing able to point out specific differences

createITPSplineResult (graph, model1, model2)





The community structure found is statistically significant!!

Pini A.,Vantini S.,2016.The interval testing procedure: A general framework for inference in functional data analysis. Biometrics.

IGRAPH COMMUNITY DETECTION ALGORITHMS

- **Modularity**
  - Fast greedy
  - Louvain
  - Optimal
  - Leading eigenvector
- **Random walk**
  - Walktrap
  - Infomap
- **Node**
  - Propagating labels
- **Edges**
  - Edge betweenness
  - Springlass

methodCommunity(graph, method, directed,weights, steps, spins, e.weights, v.weights, nb.trials)

HOW CAN WE SAY WHICH METHOD IS THE BEST?

- A procedure to compare different algorithms

To check which algorithm fits better your network!

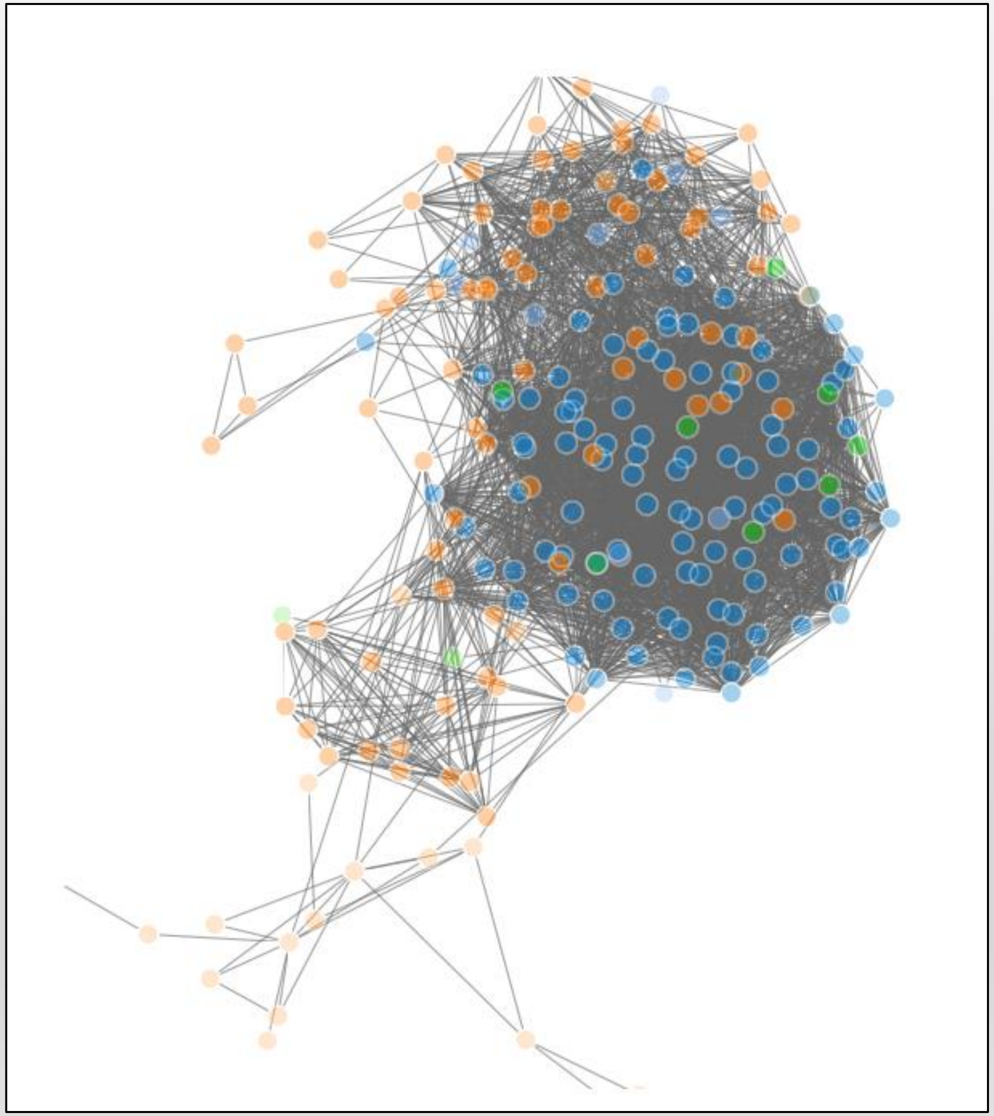
comparison(graph,graphRandom,method1,method2)

- At each perturbation level the procedure calculates the ***VI*** for both methods

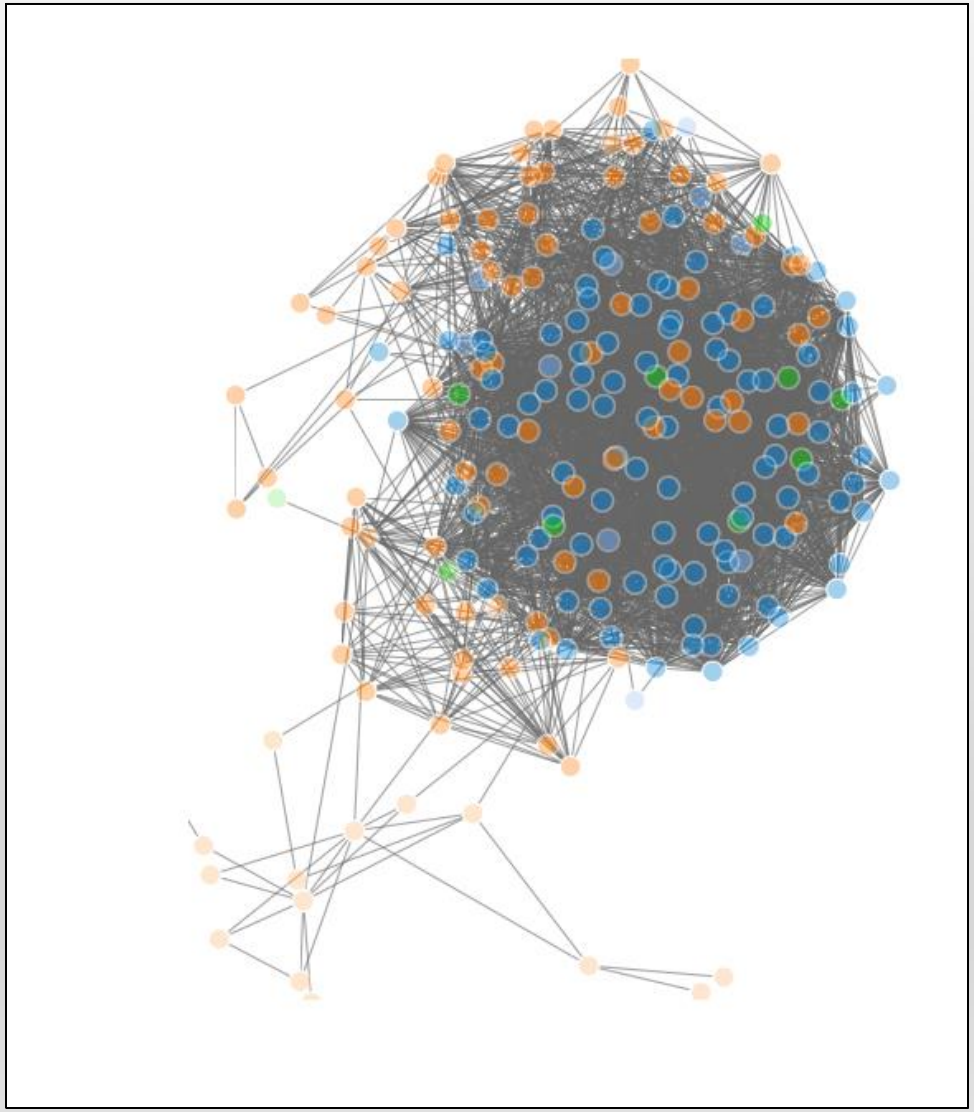
compare(comr1, comReal1, method="vi") *for the first method*
compare(comr2, comReal2, method="vi") *for the second method*

HOW CAN WE SAY WHICH METHOD IS THE BEST?

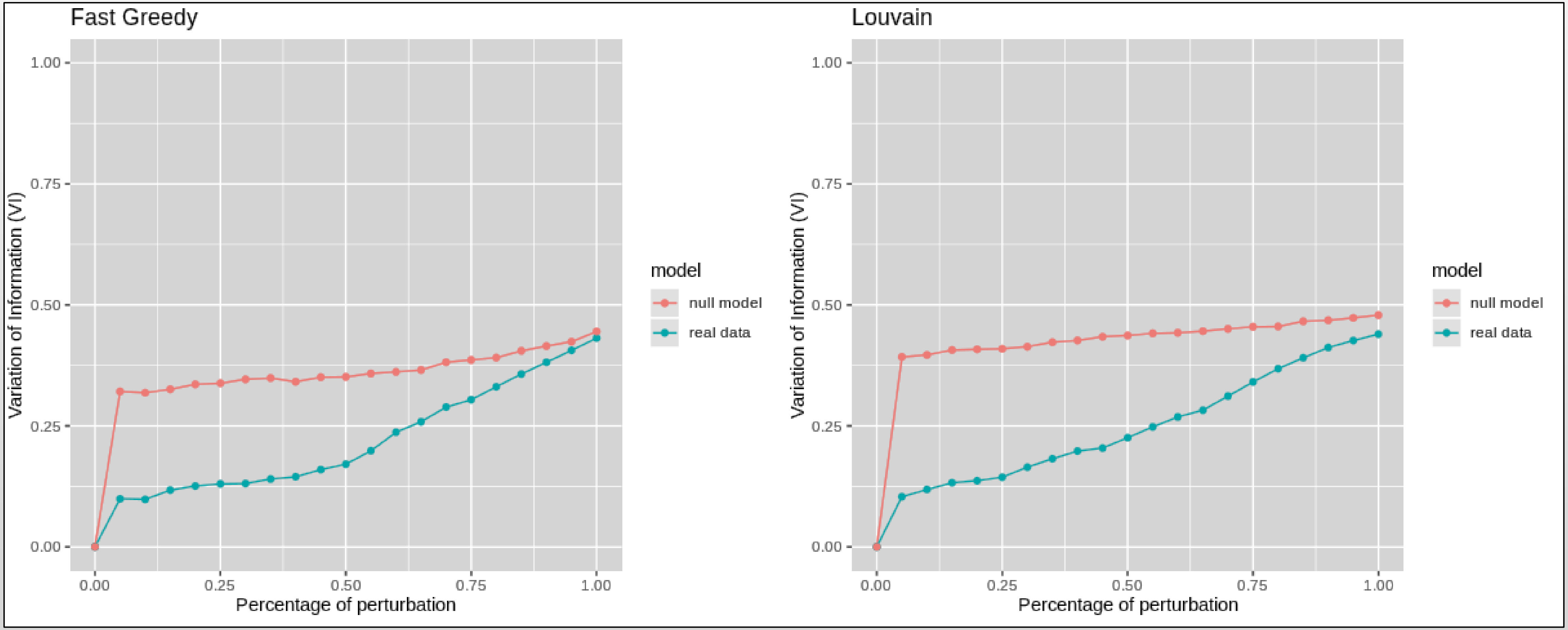
- Fast greedy



- Louvain



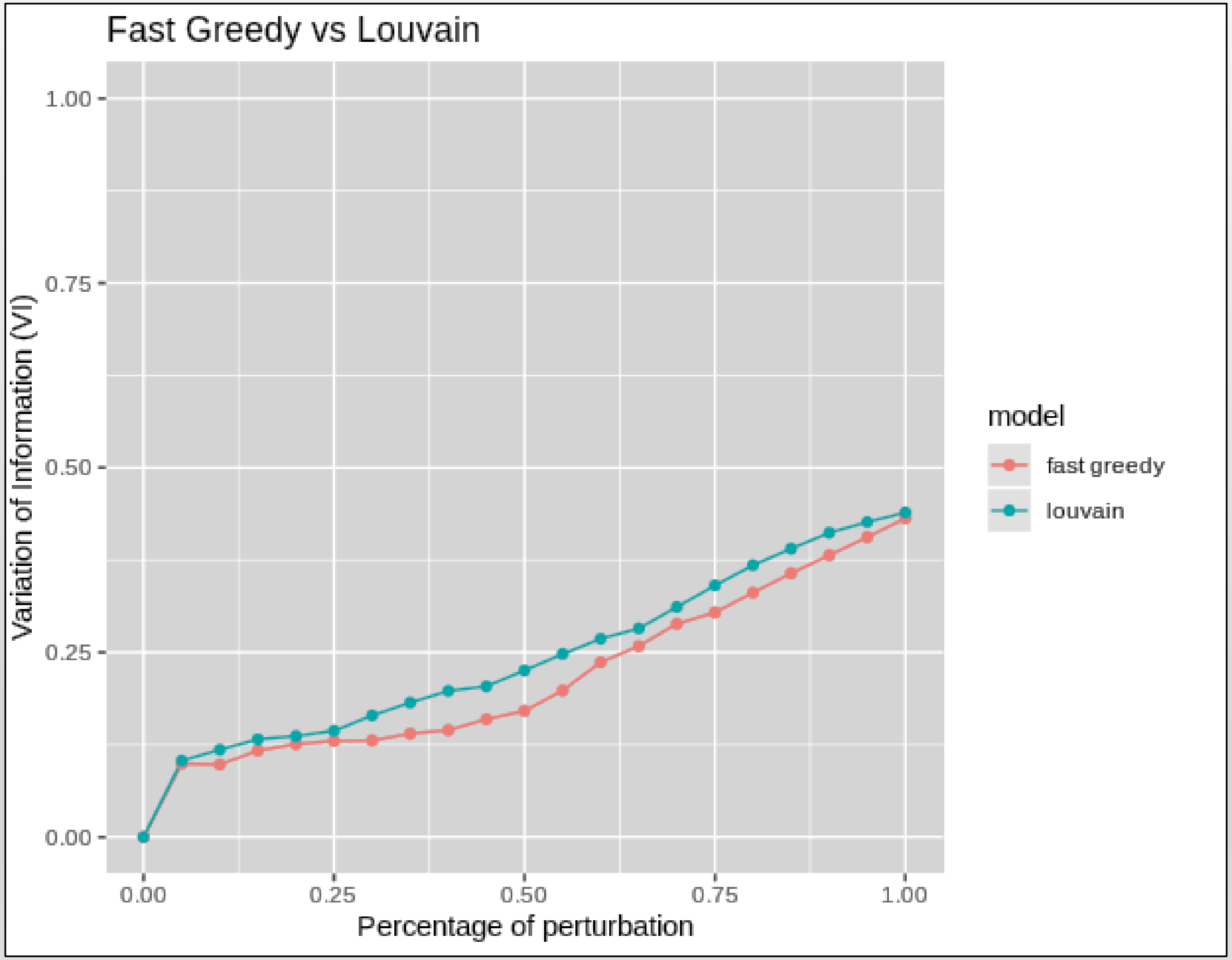
VI CURVES



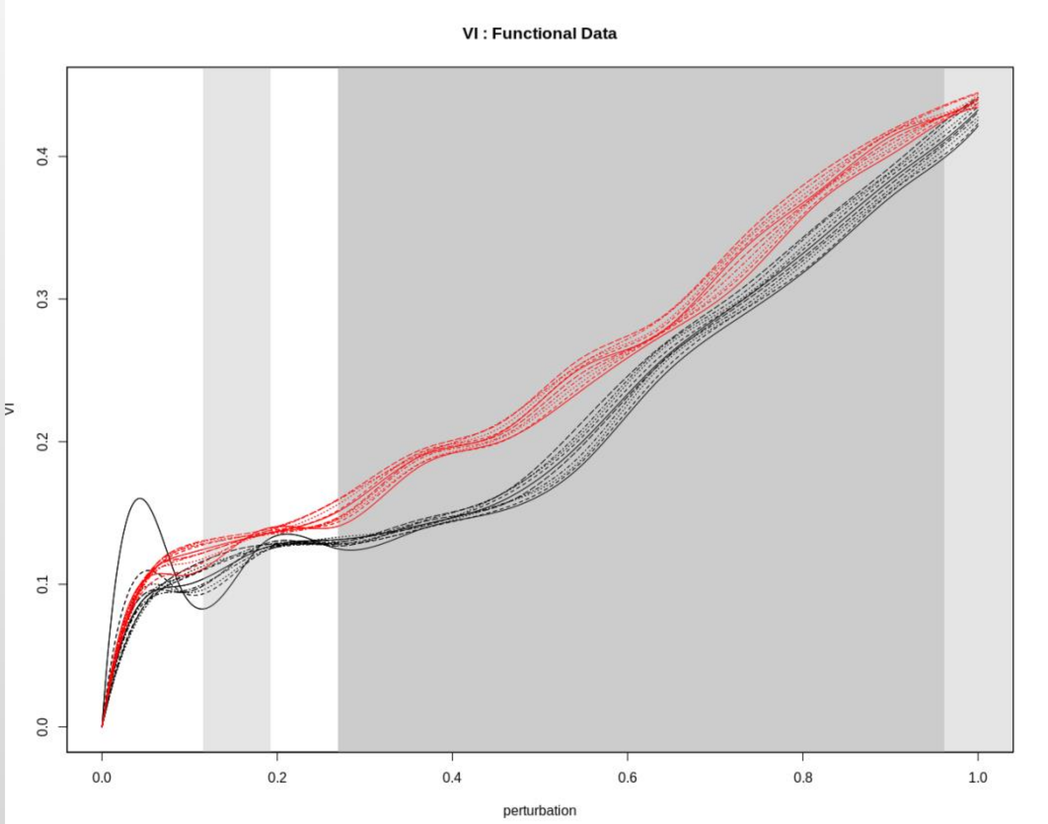
plotRobinCompare(graph, legend, legend1vs2, title1, title2, title1vs2)

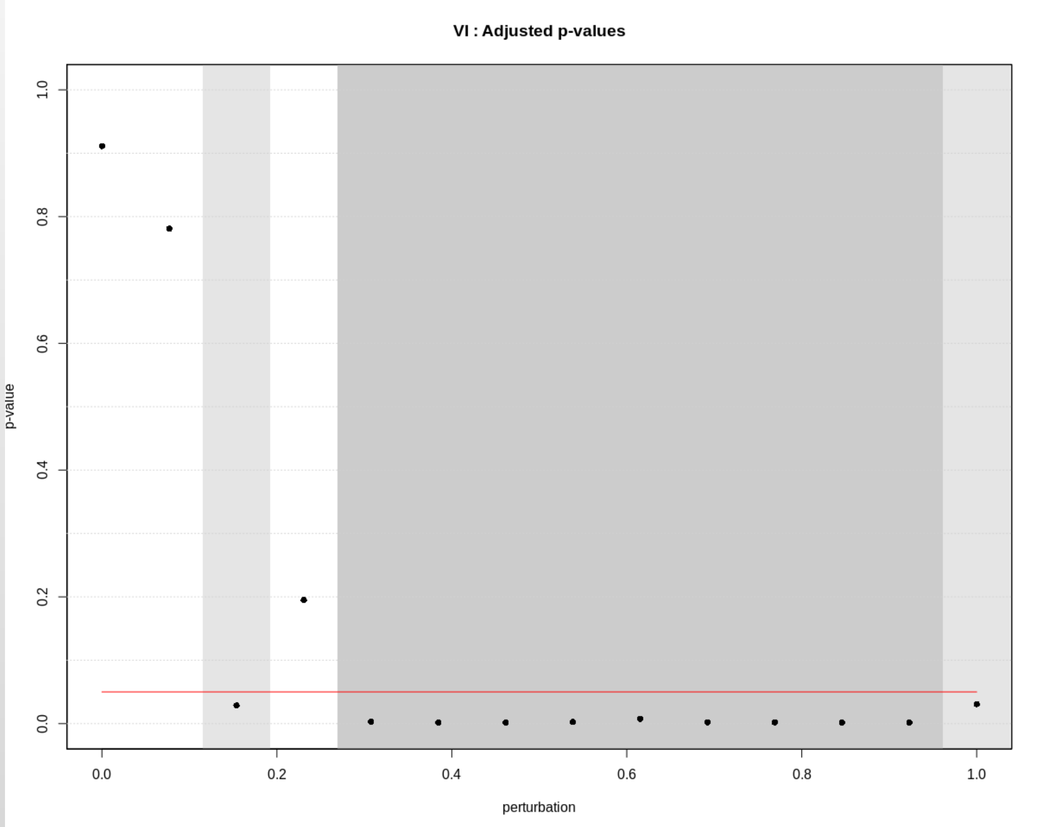
VI CURVES

plotRobinCompare(graph,
legend=c("real data", "null model"),
legend1vs2=c("fast greedy",
"louvain"),
title1="Fast Greedy",
title2="Louvain",
title1vs2="Fast Greedy vs Louvain")



ITP, *GP* and *AUC*





Bayes_Factor 120.642

Area1 0.2151066 Area2 0.2442918

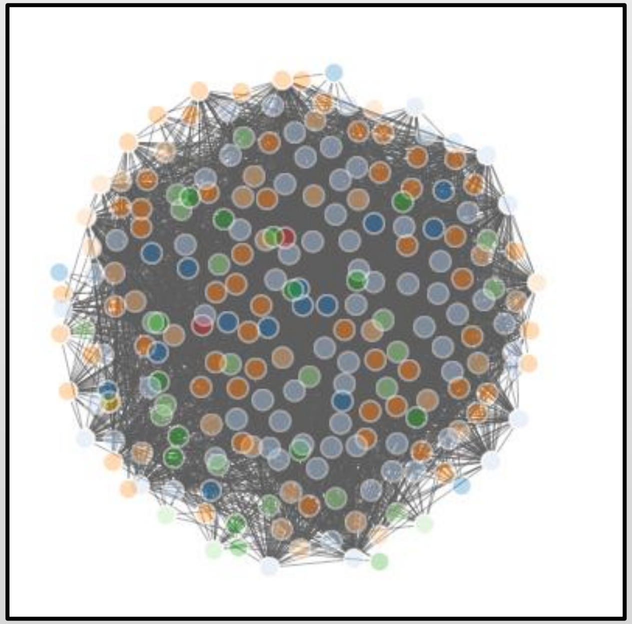
robinTest(graph, model1, model2, ratio, legend)

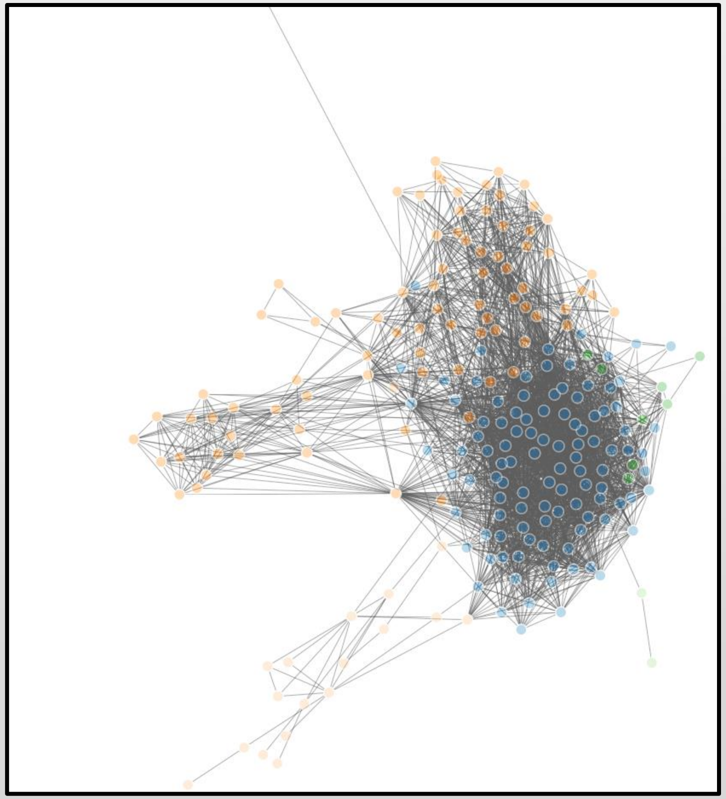
REPRESENTATIONS

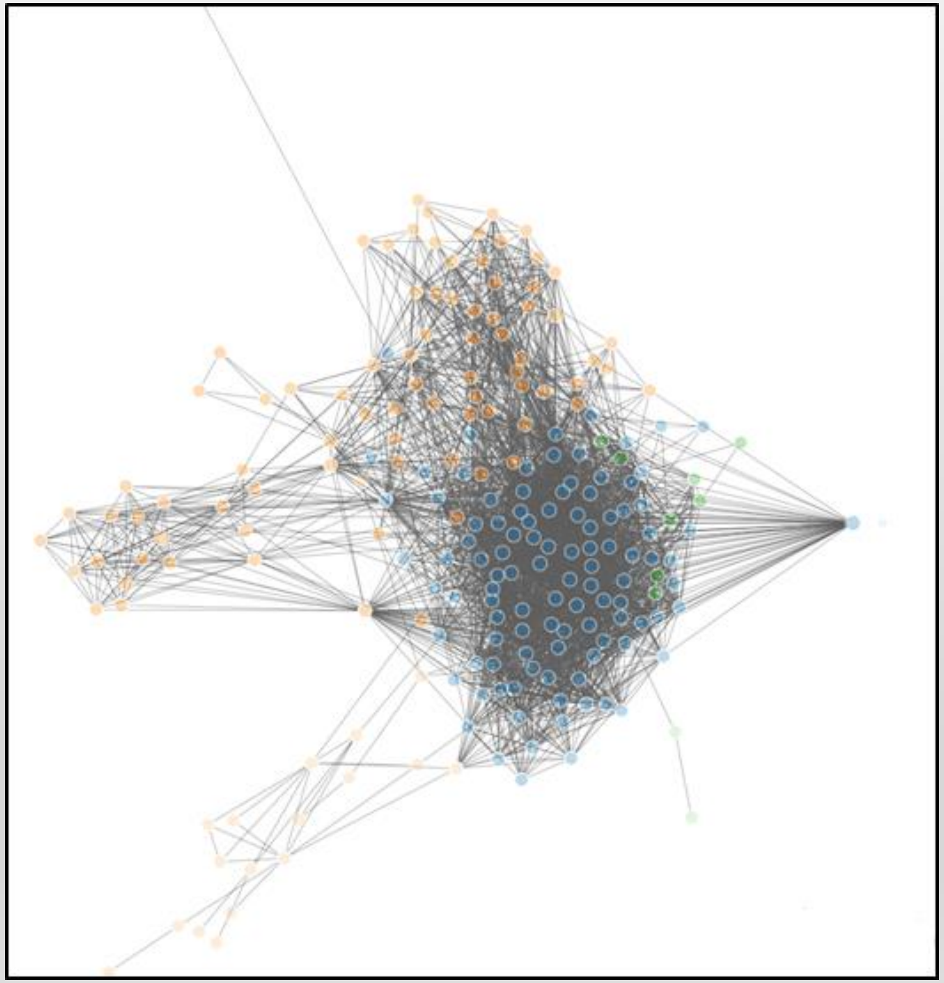
plotCommunity(graph,method)

- Graph communities
- 3D
- Interactive

➢ Fast greedy method







file:///F:/CNR%20Tigem/robin/Analisi_Leedsrip_348.edges.html

ROBIN (ROBustness In Network)

- ✓ ***Community detection algorithms***
- ✓ ***Validation of the community structure***
- ✓ ***Comparison of different community algorithms***
- ✓ ***Graphical interactive representation of 3D networks***

FUTURE

- ***To Publish ROBIN***

- To use a different clustering stability measure

- To compare the performance of different measures for community structure comparison

- To test if the differences of the different methods are only due to the degree distribution of the network or are influenced by other factors

BIBLIOGRAPHY

- Carissimo A., Cutillo L., De Feis I., Validation of community robustness Computational Statistics and Data Analysis 2017
- Karrer B, Levina E and Newman M E J Robustness of community structure in networks 2008 *Phys. Rev.* E **77** 046119
- Alfredo A. Kalaitzis and Neil D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. BMC Bioinformatics, 12(180), 2011
- Mc Auley, J., Leskovec, J., 2012. Learning to Discover Social Circles in Ego Networks. NIPS. pp. 548–556
- Meilă M.,2007. Comparing clusterings-an information based distance. J. Multivariate Anal.98,873-895
- Bender,E.A., Canfield,E.R., 1978. The asymptotic number of labelled graphs with given degree sequences. J.Combin. Theory A24, 296–307
- Pini A.,Vantini S.,2016.The interval testing procedure: A general framework for inference in functional data analysis. Biometrics

ACKNOWLEDGMENT

- *Annamaria Carissimo*
- *Luisa Cutillo*
- *Dario Righelli*
- *Italia De Feis*

Thanks for the attention!

MODULARITY

- A network with **strong** community structure has **high** modularity

- **Q is not sufficient**, not all networks with high modularity have strong community structure

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys.

SIMULATION

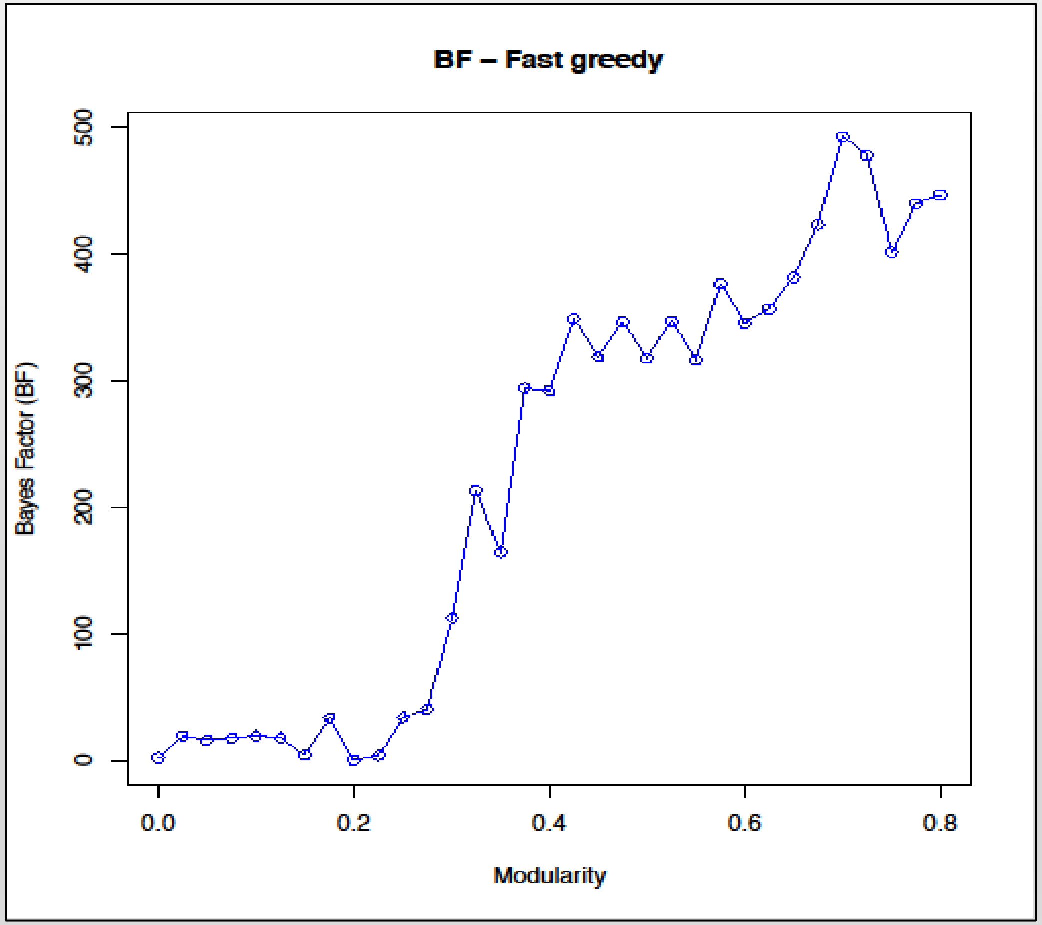
❖We used a literature model that generates undirected, simple and connected graphs with prescribed degree sequences and a specified level of community structure

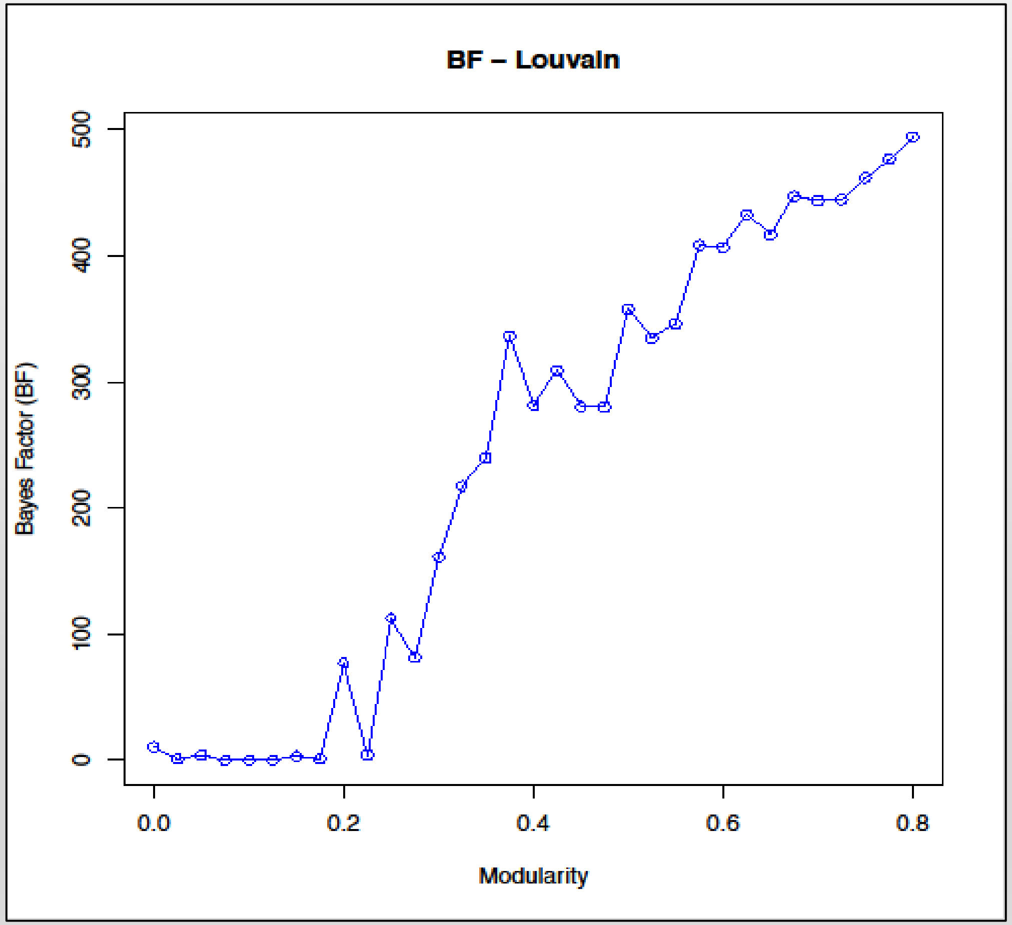
❖We constructed networks with 2000 nodes, 10 communities, an average degree equal to 10 at different value of modularity Q

Sah P, Singh LO, Clauset A, Bansal S. Exploring community structure in biological networks with random graphs. BMC Bioinformatics. 2014 Jun 25;15:220

SIMULATION STUDY

❖Null models with different modularity





VARIATION OF INFORMATION (VI)

❖ At each level of perturbation ***p*** we compared the partition obtained from the original with the partition obtained from the perturbed graph computing Variation of Information (***VI***)

$$\boldsymbol{VI(C,C')=H(C|C') + H(C'|C)}$$

- $0 \leq VI(C,C') \leq 2\log K$

- C and C' are two generic partitions
- K is the number of clusters

Meilă M.,2007. Comparing clusterings-an information based distance. J. Multivariate Anal.98,873-895

GAUSSIAN PROCESS

- ❖ Are the two curves from the same process or not?

- ❖ The hypothesis testing problem can be reformulated over the perturbation interval [0,1] as:

$$H_0 : \log_2 \frac{VIc(x)}{VIc_{random}(x)} \sim \mathcal{GP}(0, k(x, x'))$$

$$H_1 : \log_2 \frac{VIc(x)}{VIc_{random}(x)} \sim \mathcal{GP}(m(x), k(x, x'))$$

Bayes Factor is approximated with a log-ratio of marginal likelihoods of two GPs, each one representing the hypothesis of differential (the profile has a significant underlying signal) and non differential expression (there is no underlying signal in the profile, just random noise).

Alfredo A. Kalaitzis and Neil D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. BMC Bioinformatics, 12(180), 2011.

ITP

1. **Basis Expansion:** functional data are projected on a functional basis (i.e. Fourier or B-splines expansion);

2. **Interval-Wise Testing:** statistical tests are performed on each interval of basis coefficients;

3. **Multiple Correction:** for each component of the basis expansion, an adjusted p-value is computed from the p-values of the tests performed in the previous step.