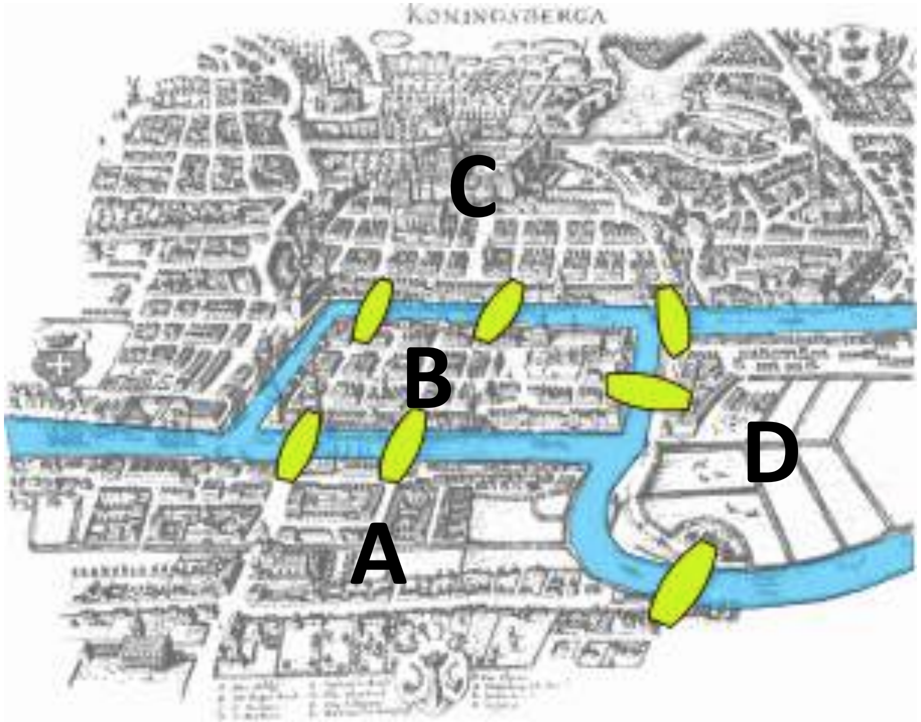
STATISTICS ON BIOLOGICAL NETWORKS

André Fujita

Associate Professor	Alexander von Humboldt Fellow
Dept. of Computer Science	Institut für Informatik
Inst. of Math and Statistics	Bioinformatik
University of São Paulo	University of Leipzig

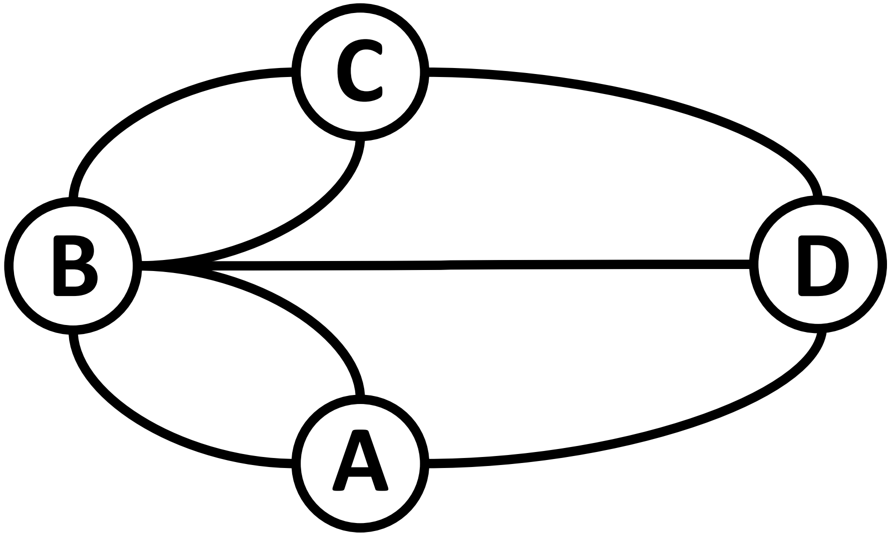
33rd TBI Winter Seminar in Bled
February 11th – 16th, 2018.

NETWORK

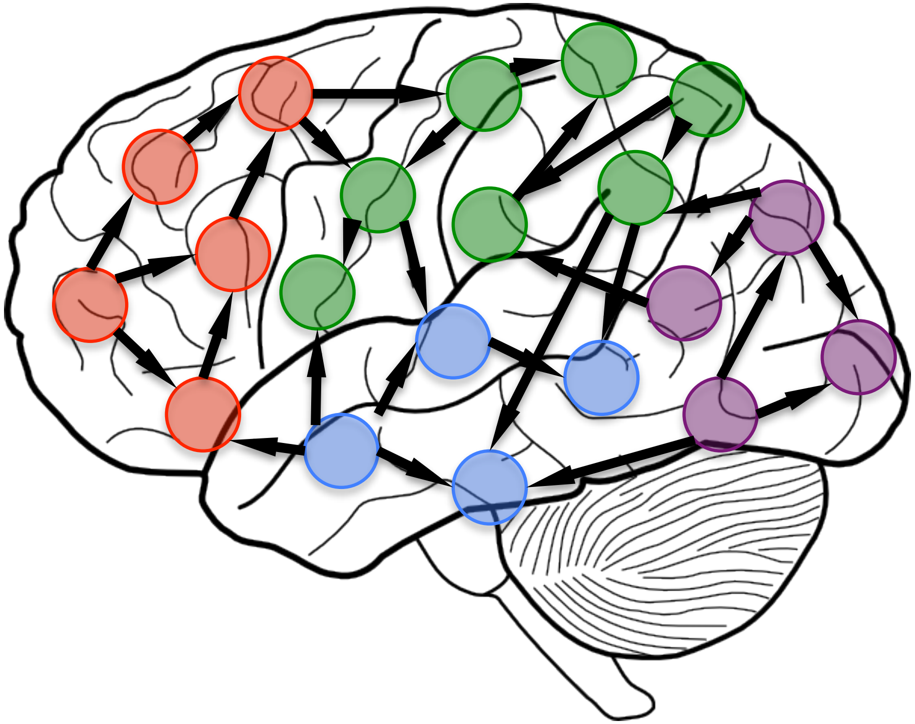


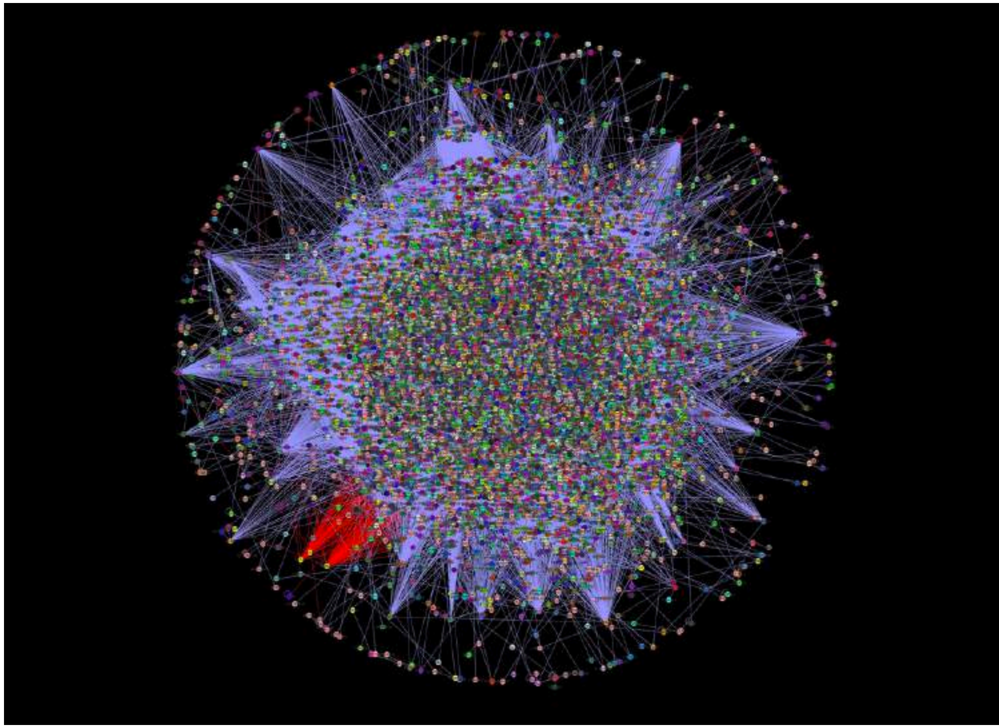
Seven bridges of Konigsberg (1736)

Leonhard P. Euler
1707 - 1783

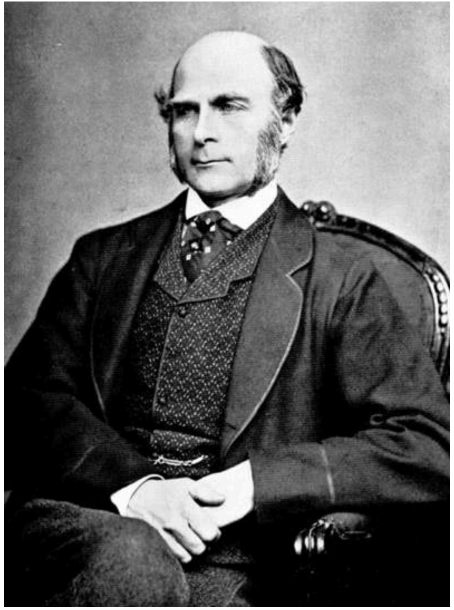


VARIABILITY

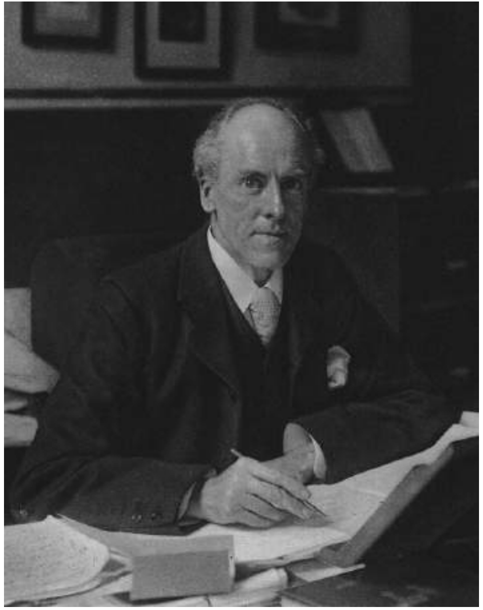




STATISTICS



Francis Galton
1822 - 1911



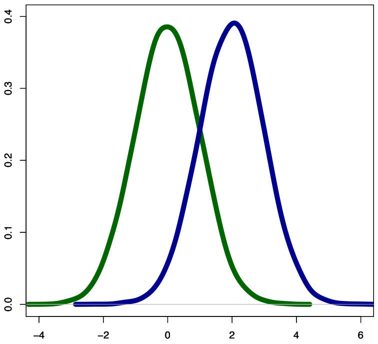
Karl Pearson
1857 - 1936

William Sealy Gosset
1876 - 1937

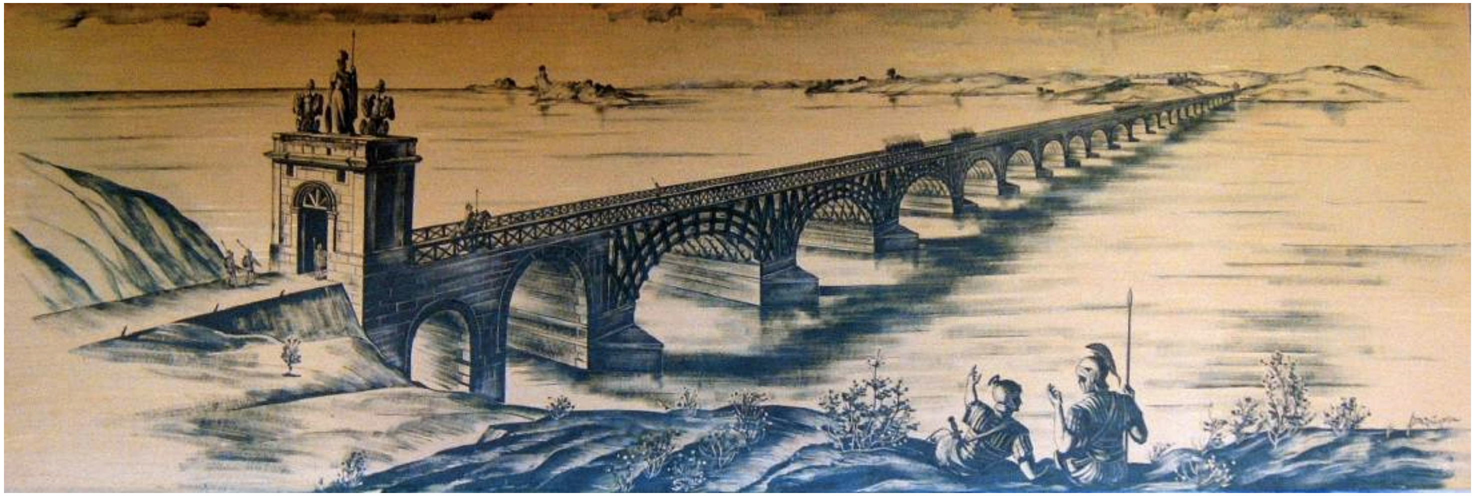
Ronald Aylmer Fisher
1890 - 1962

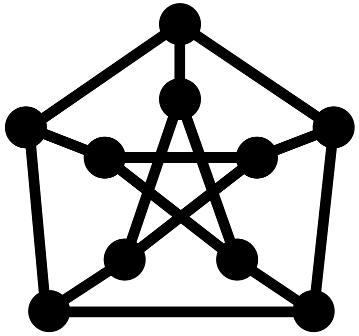
STATISTICS ON NETWORKS

1. Parameter estimation
2. Model selection
3. T-test
4. ANOVA
5. Correlation

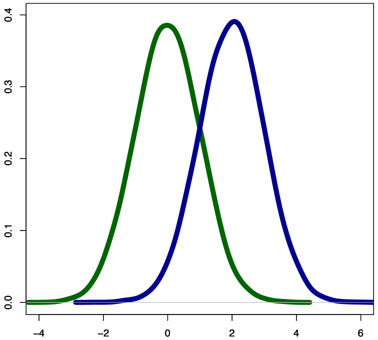


Statistics

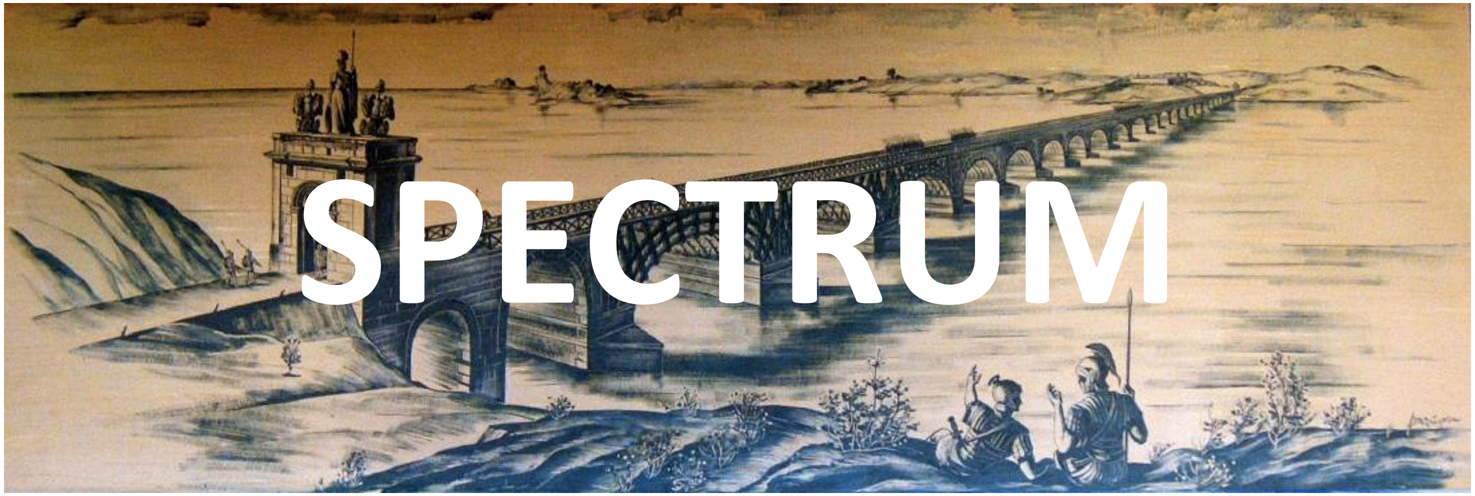




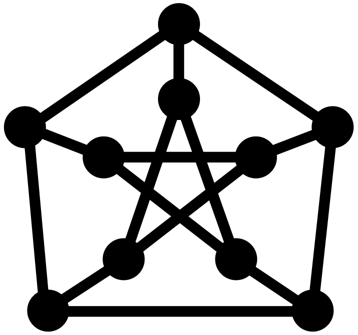
Graph theory



Statistics

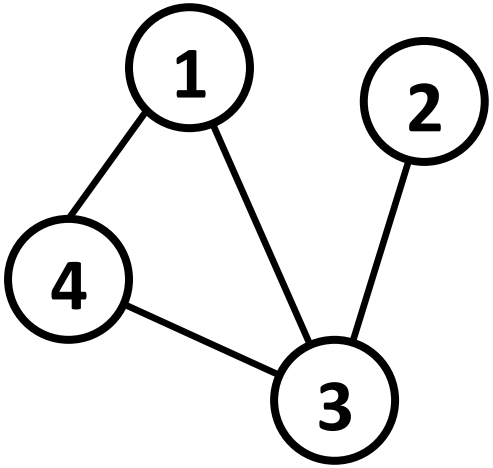


SPECTRUM



Graph theory

Graph *G*

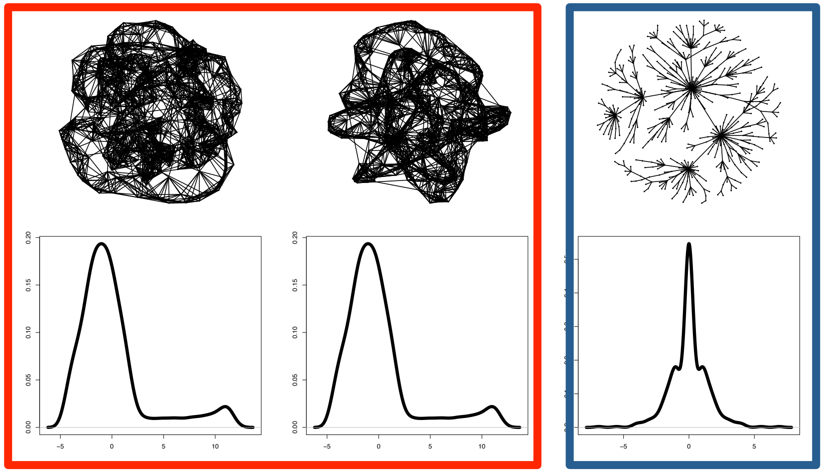


Adjacency matrix A(*G*)

$$\begin{array}{c|cccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \hline \mathbf{1} & 0 & 0 & 1 & 1 \\ \mathbf{2} & 0 & 0 & 1 & 0 \\ \mathbf{3} & 1 & 1 & 0 & 1 \\ \mathbf{4} & 1 & 0 & 1 & 0 \end{array}$$

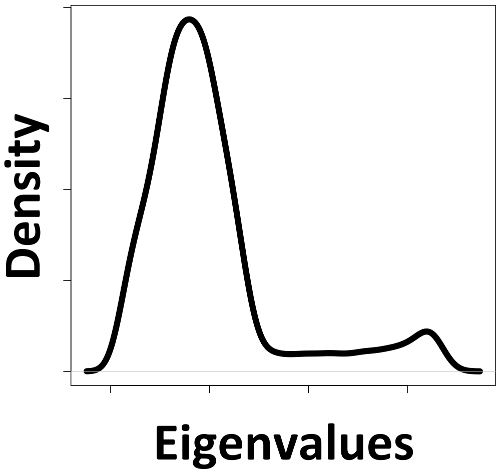
$$\mathbf{Av} = \lambda \mathbf{v}$$

Structural properties of graphs

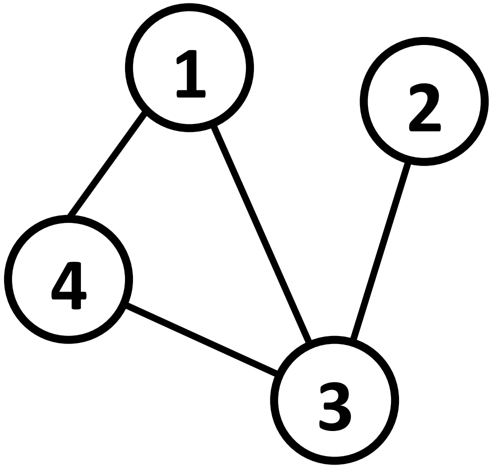


Takahashi et al., 2012

Spectral distribution



Graph *G*

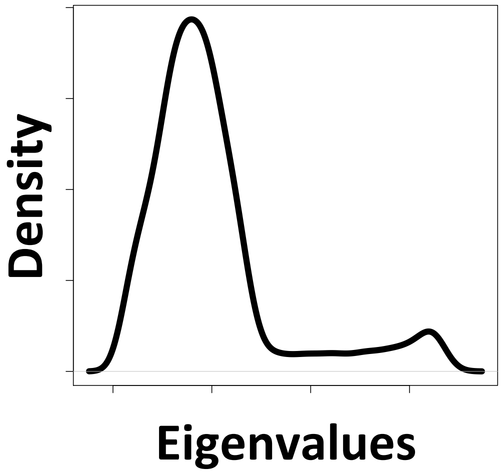


Adjacency matrix A(*G*)

$$\begin{array}{c|cccc} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} \\ \hline \mathbf{1} & 0 & 0 & 1 & 1 \\ \mathbf{2} & 0 & 0 & 1 & 0 \\ \mathbf{3} & 1 & 1 & 0 & 1 \\ \mathbf{4} & 1 & 0 & 1 & 0 \end{array}$$

$$\mathbf{Av} = \lambda\mathbf{v}$$

Spectral distribution

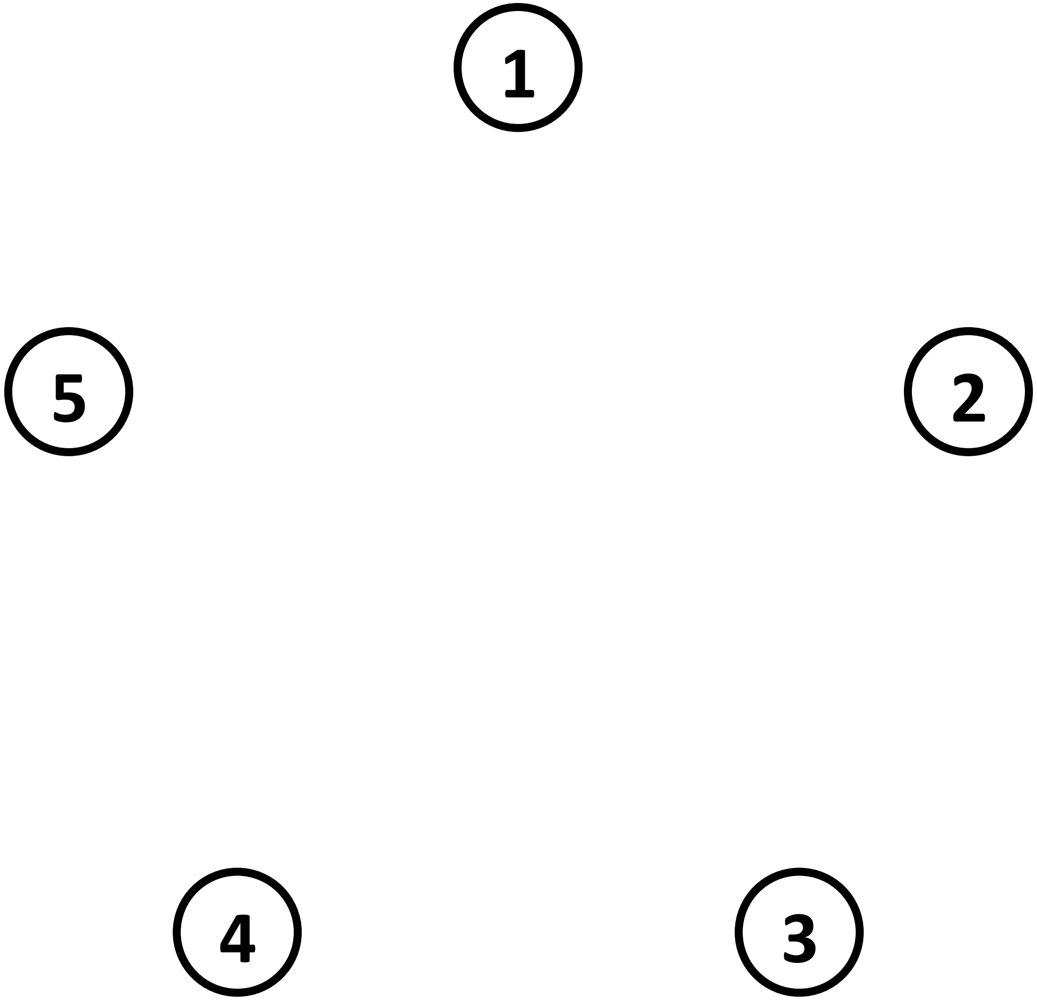


Graph entropy

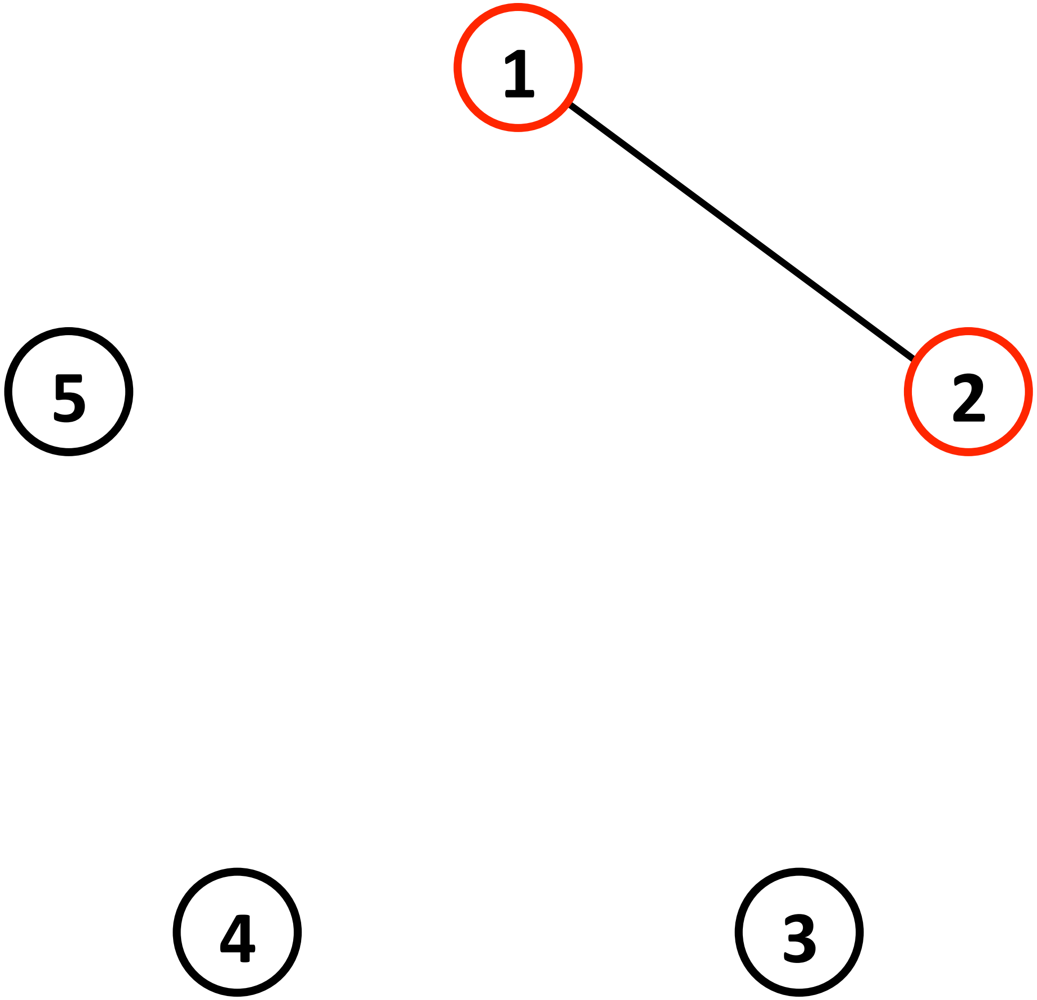
$$\mathrm{H}(\rho) = -\int_{-\infty}^{\infty} \rho(\lambda) \log \rho(\lambda)\, \mathrm{d}\lambda$$

Takahashi et al., 2012

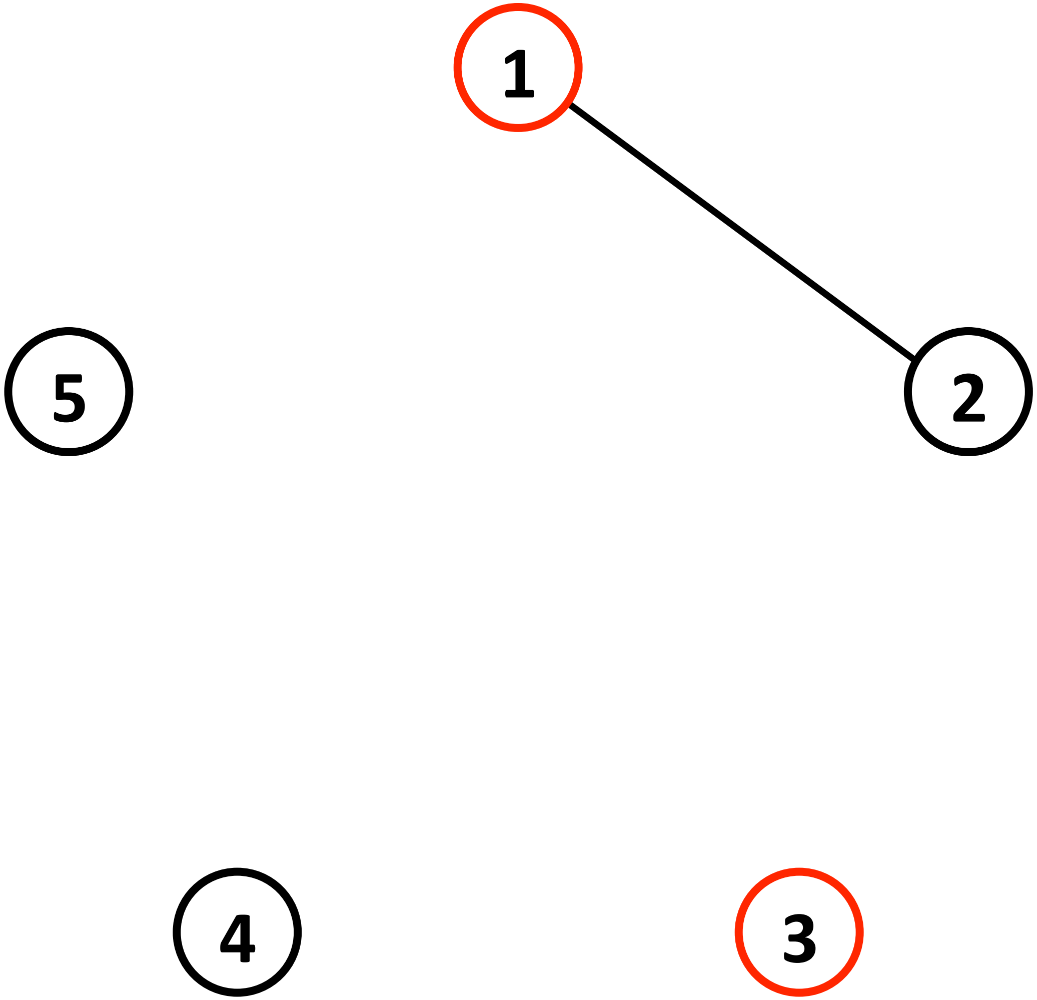
ERDÖS-RÉNYI RANDOM GRAPH MODEL



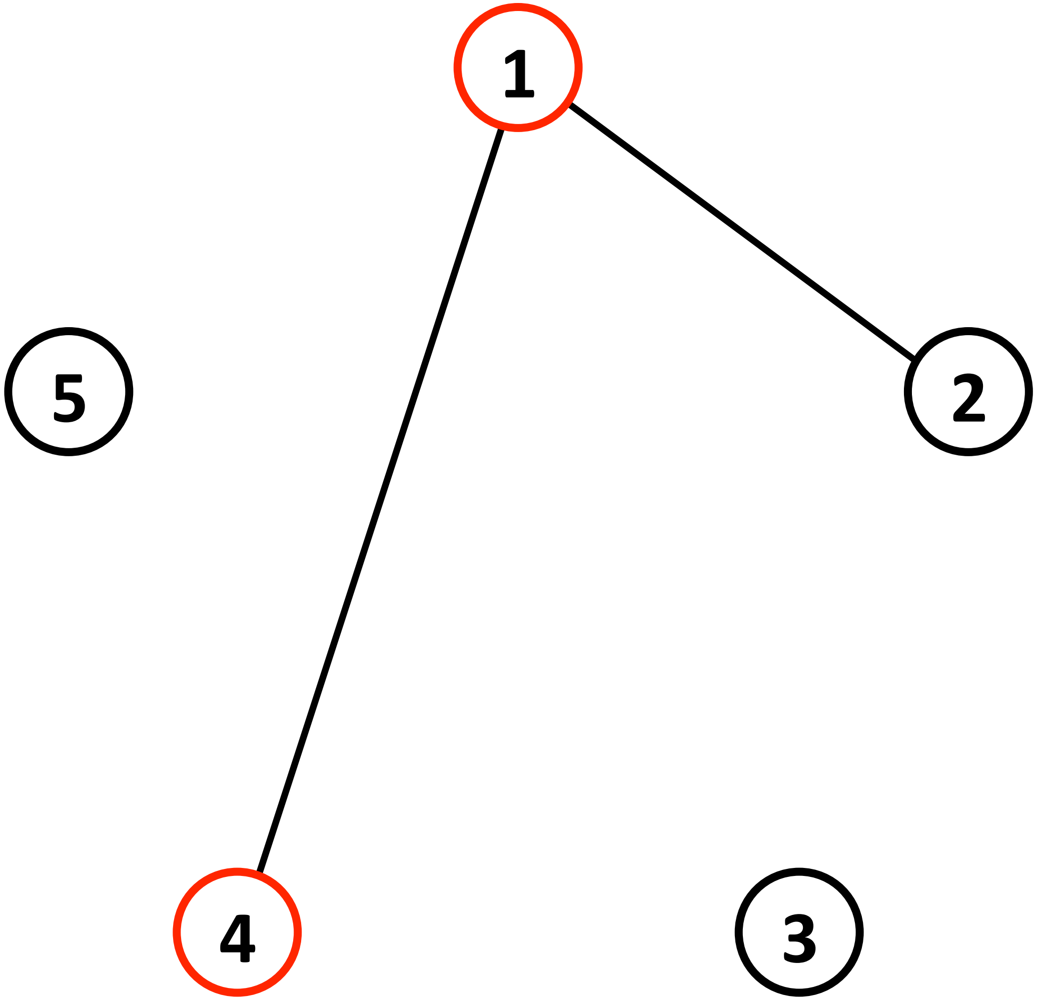
ERDÖS-RÉNYI RANDOM GRAPH MODEL



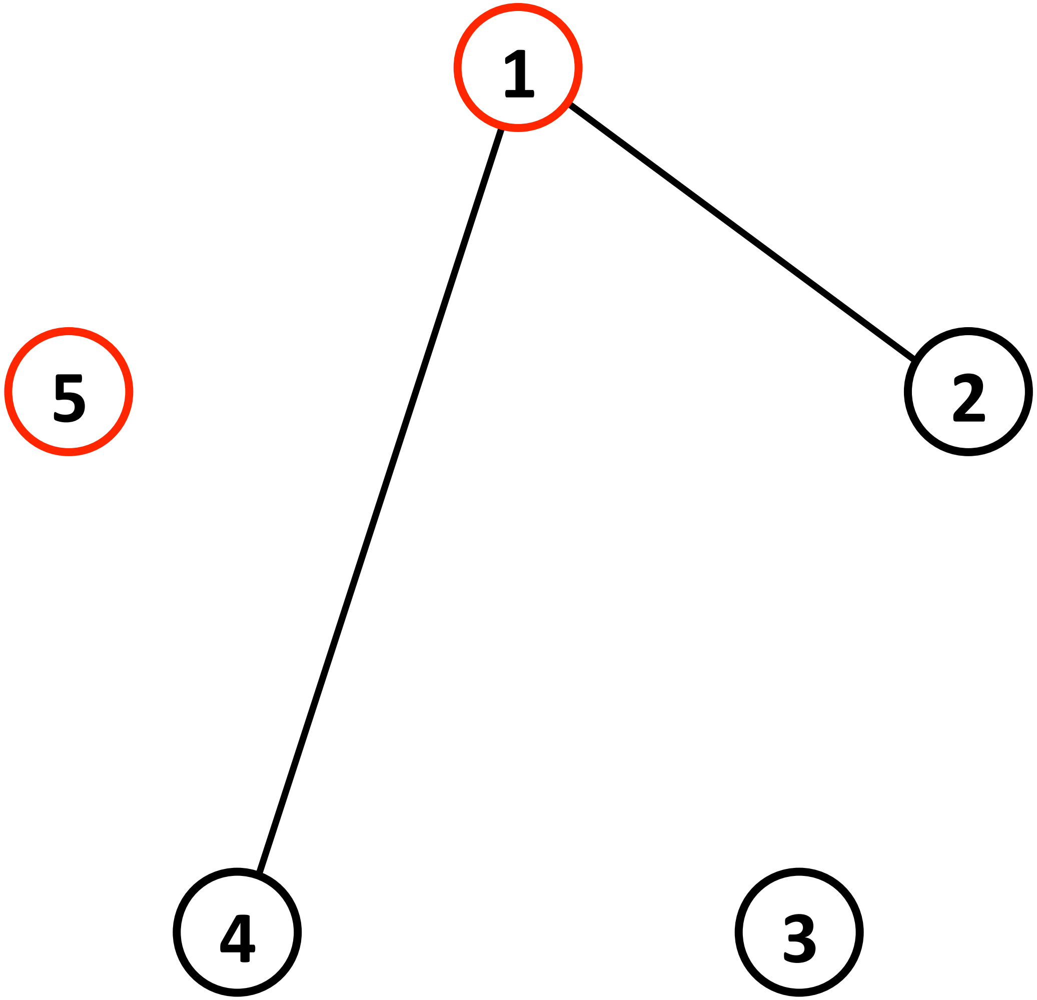
ERDÖS-RÉNYI RANDOM GRAPH MODEL



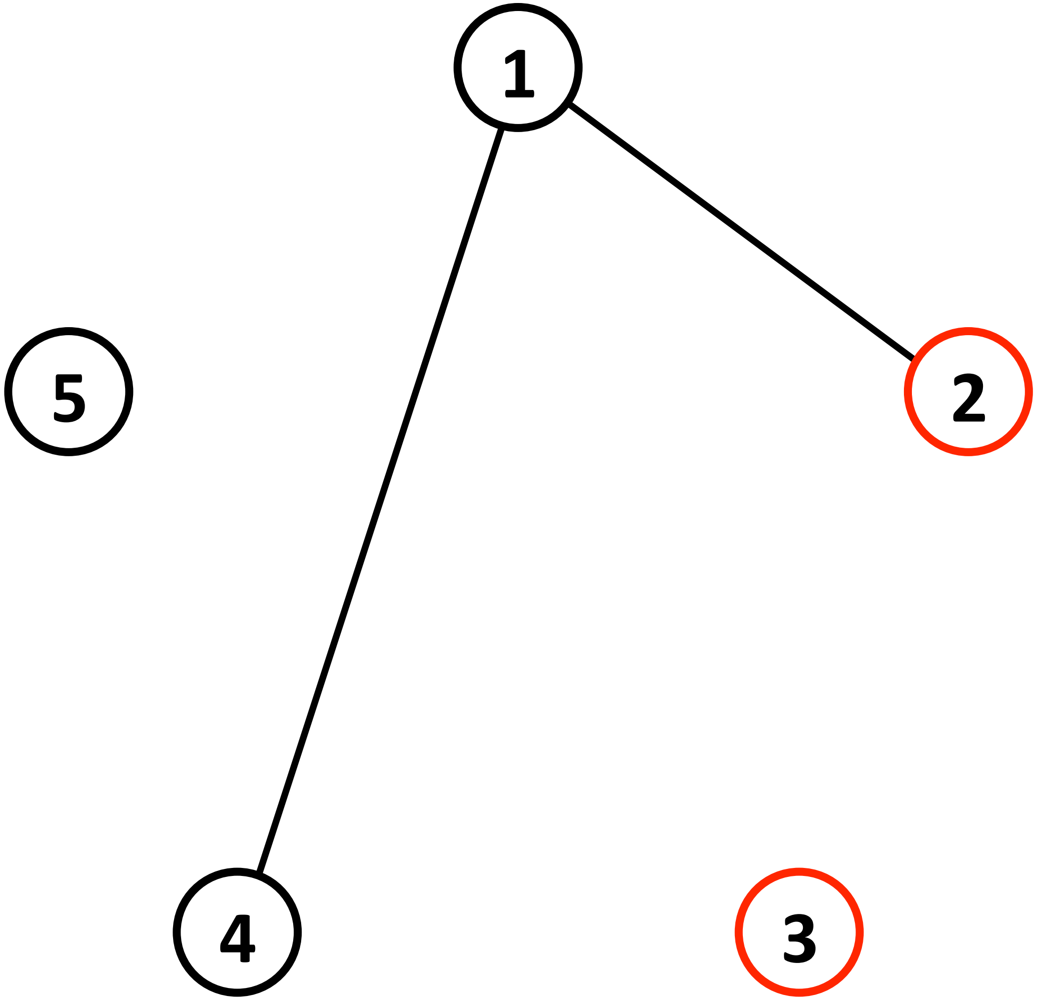
ERDÖS-RÉNYI RANDOM GRAPH MODEL



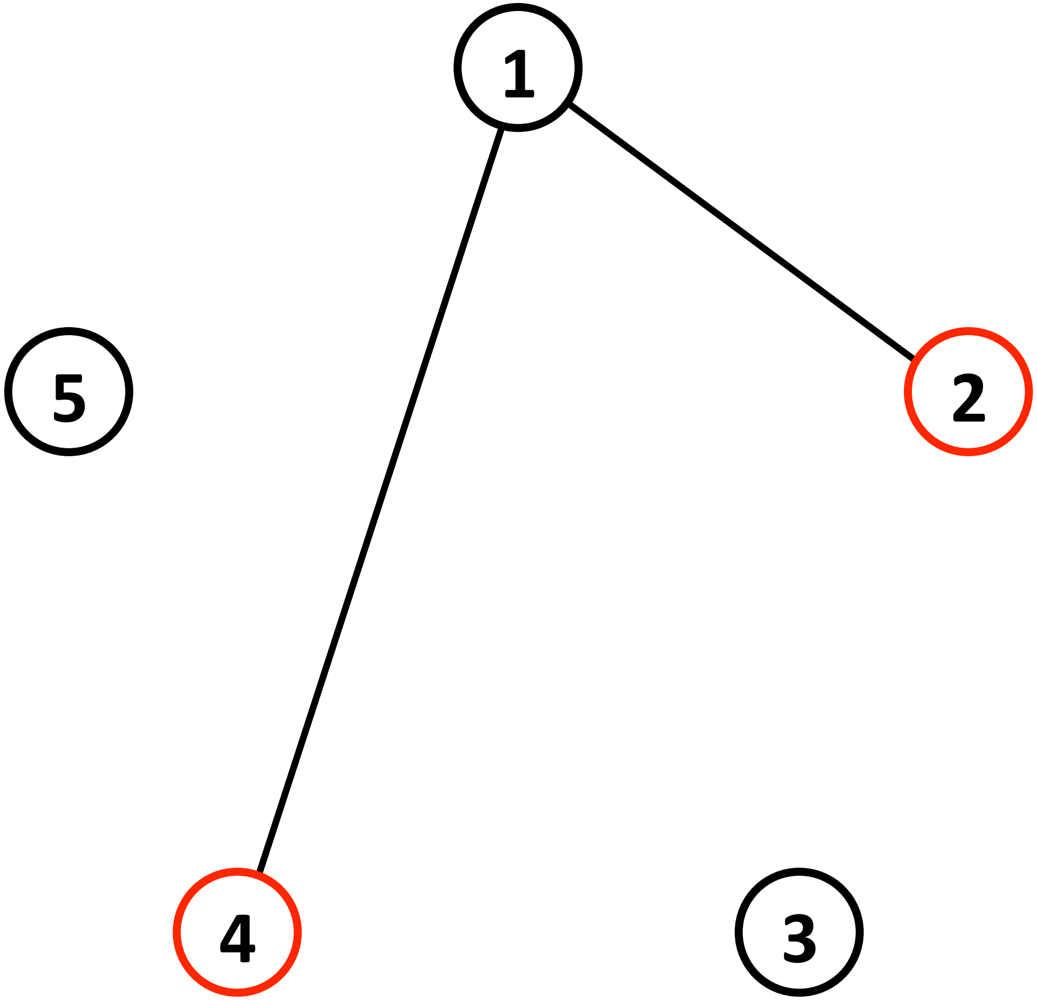
ERDÖS-RÉNYI RANDOM GRAPH MODEL



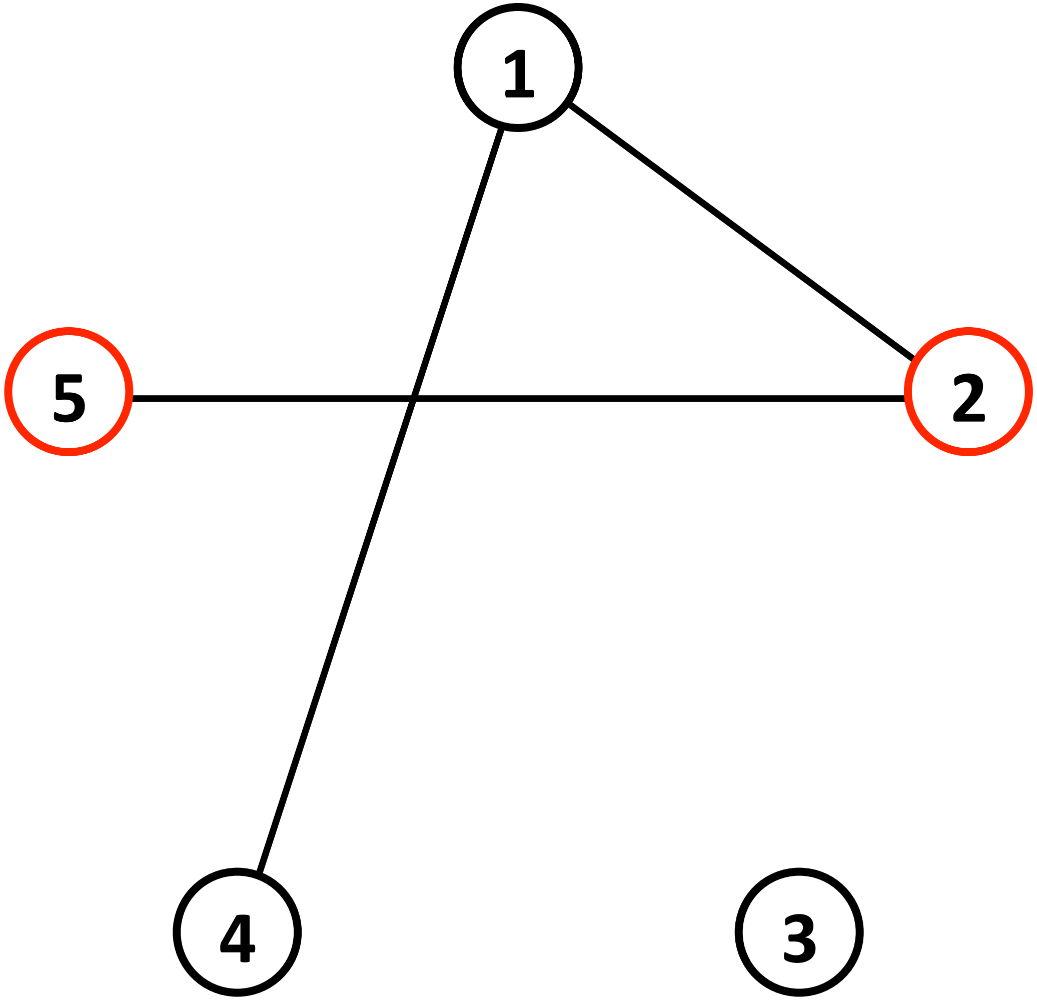
ERDÖS-RÉNYI RANDOM GRAPH MODEL



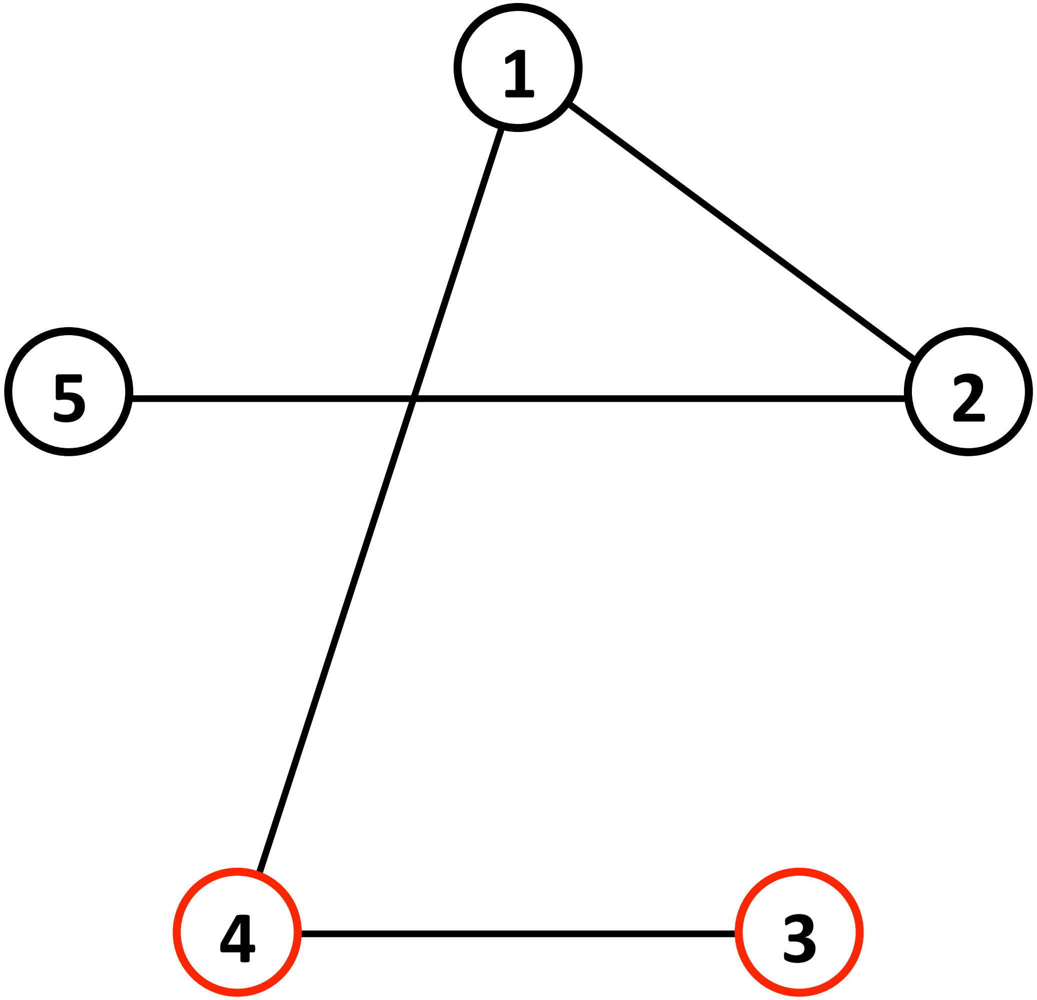
ERDÖS-RÉNYI RANDOM GRAPH MODEL



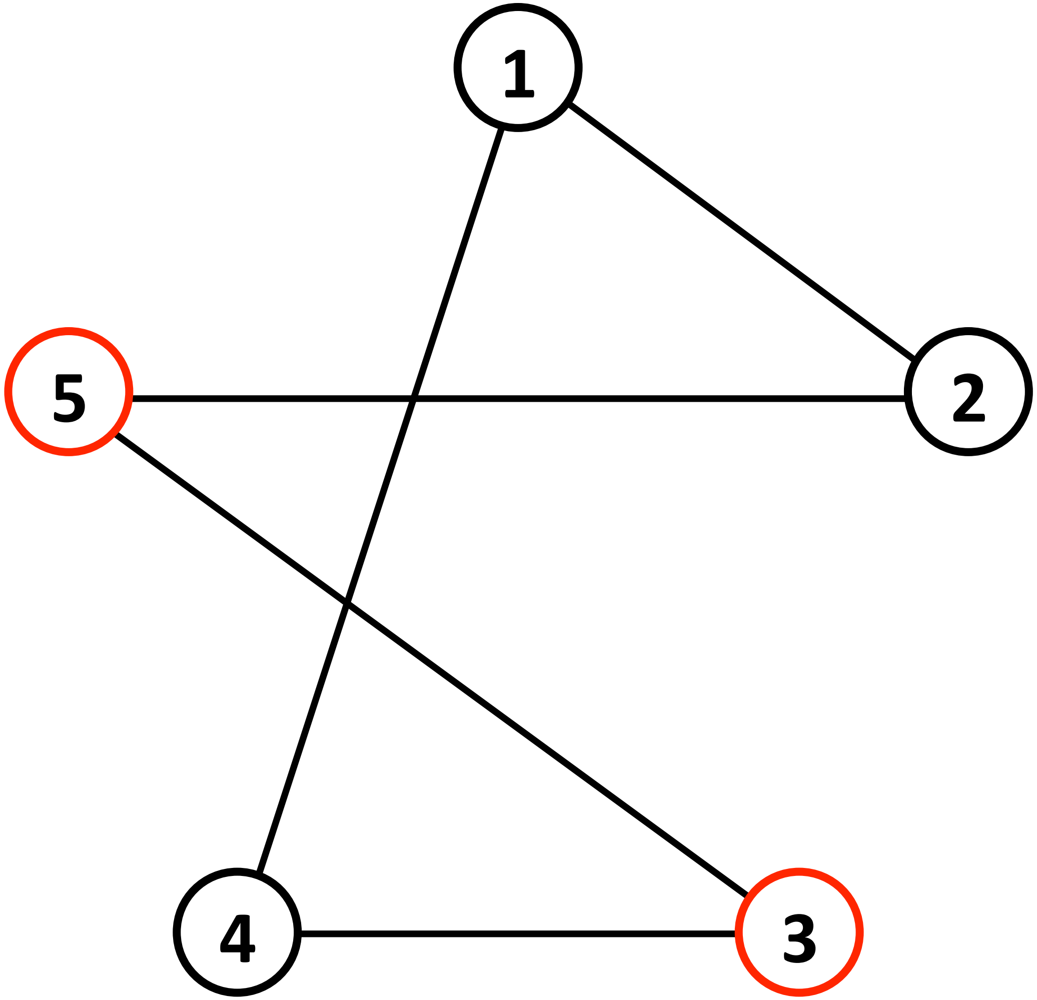
ERDÖS-RÉNYI RANDOM GRAPH MODEL



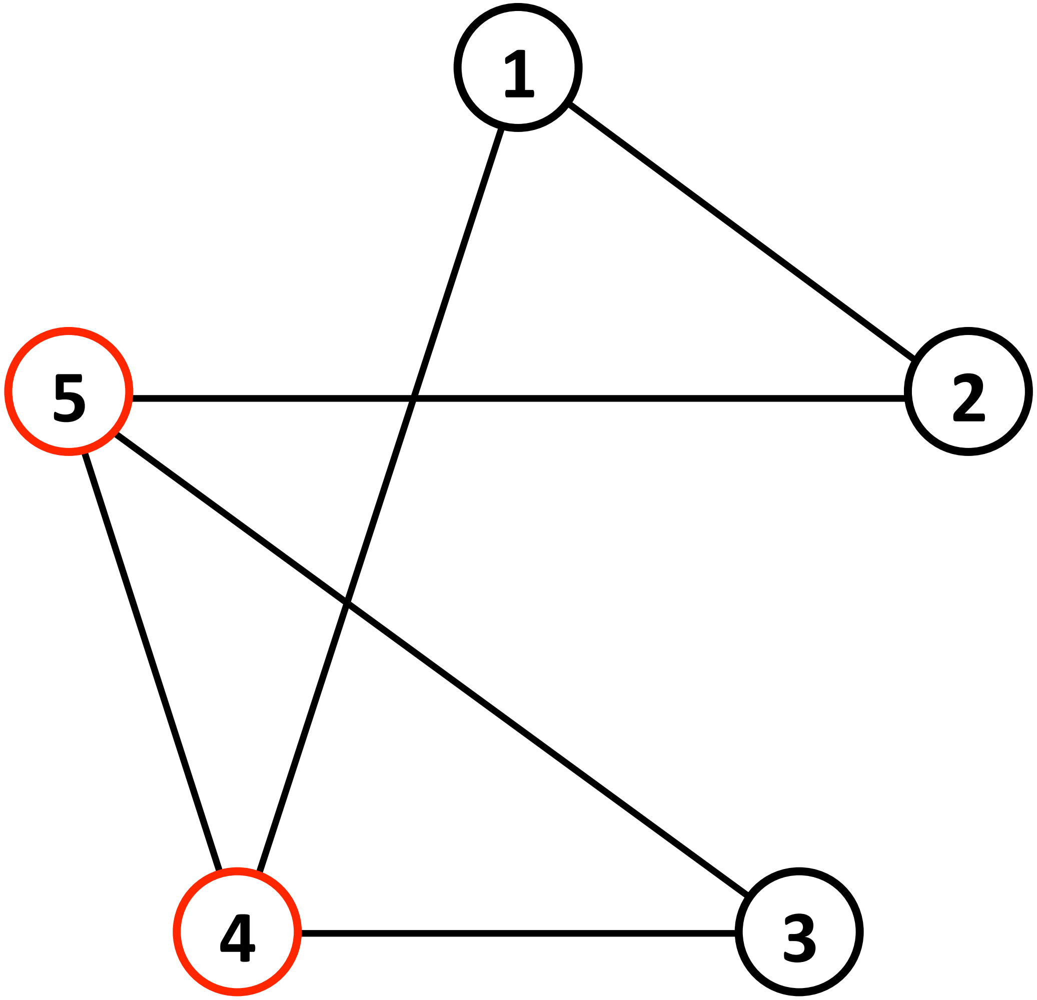
ERDÖS-RÉNYI RANDOM GRAPH MODEL



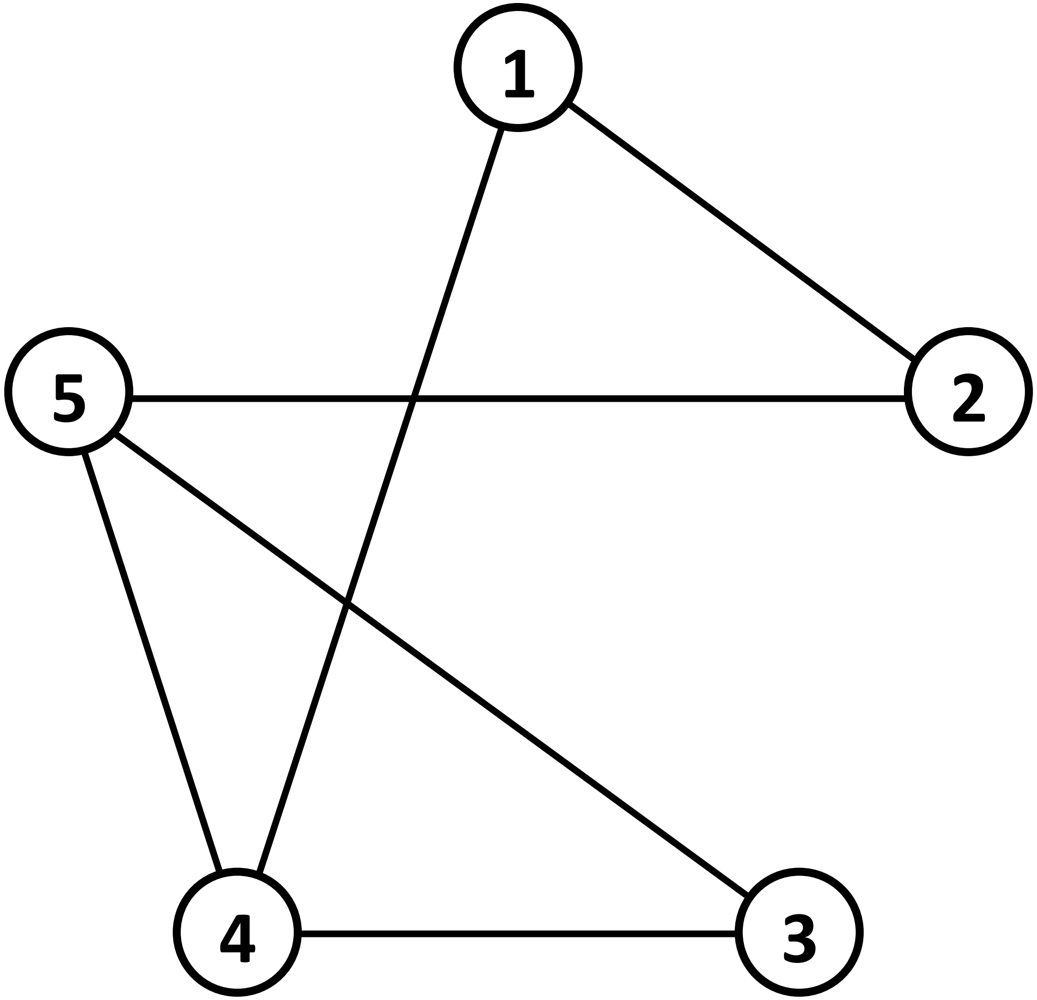
ERDÖS-RÉNYI RANDOM GRAPH MODEL



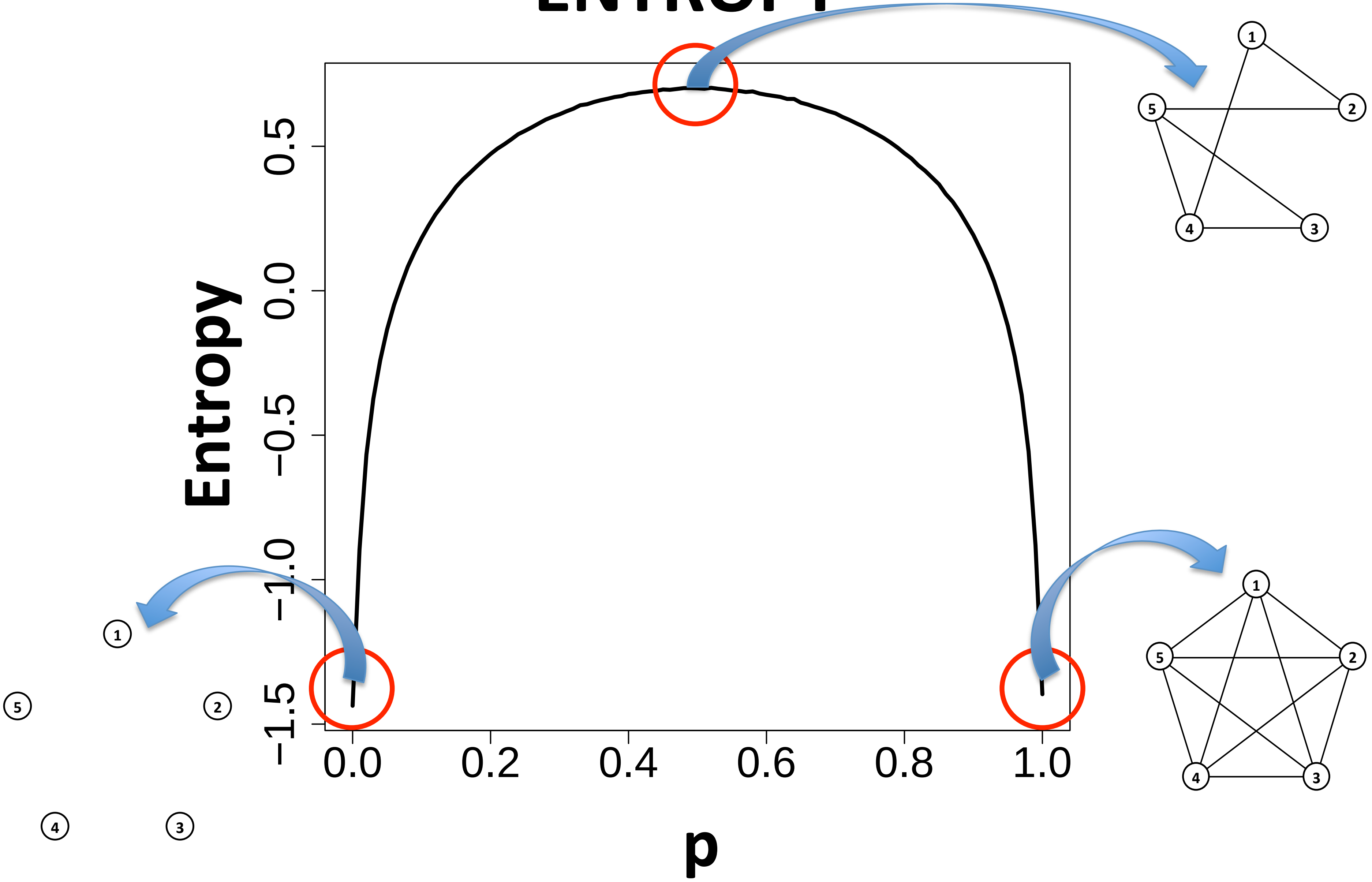
ERDÖS-RÉNYI RANDOM GRAPH MODEL



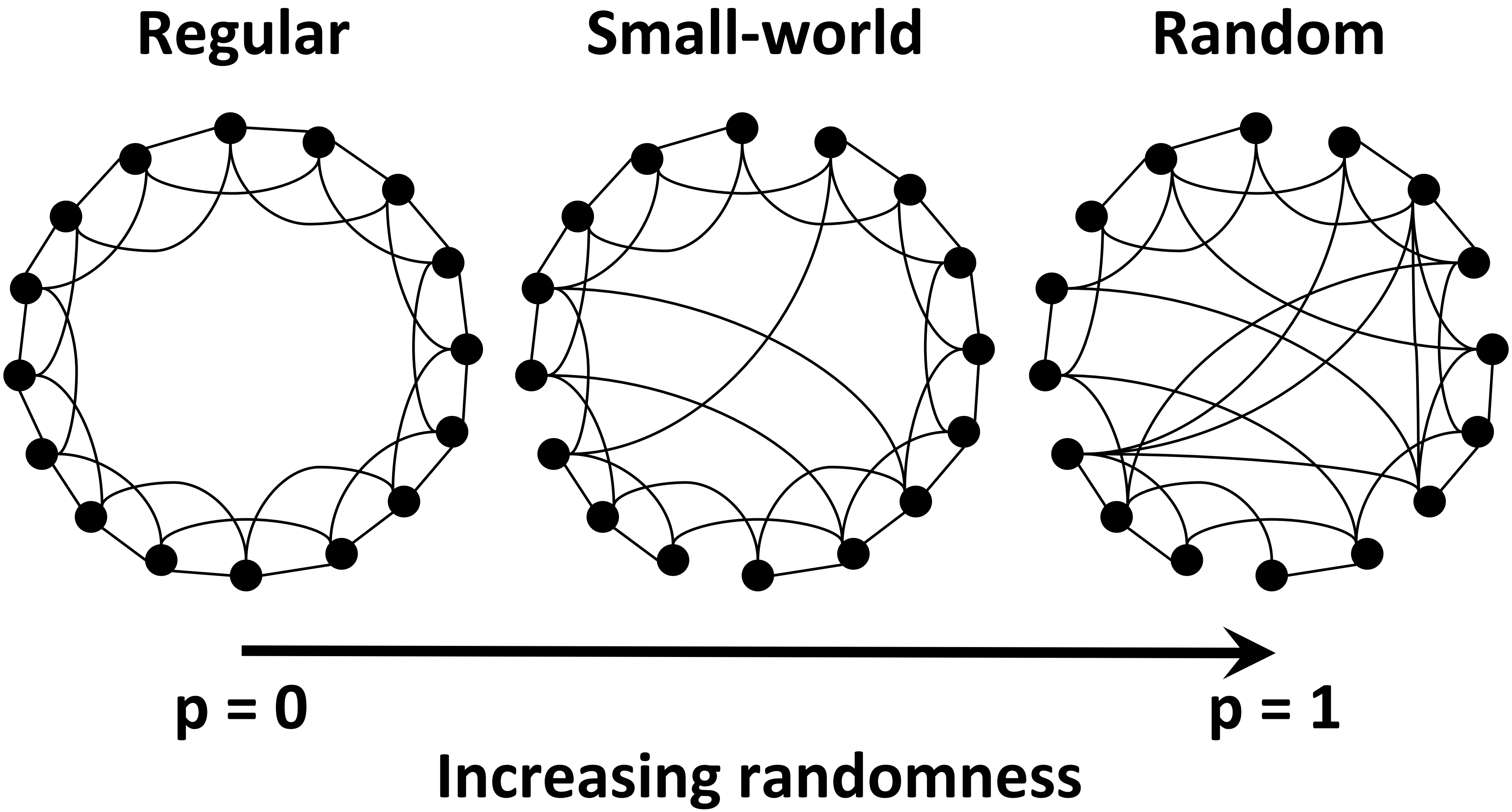
ERDÖS-RÉNYI RANDOM GRAPH MODEL



ERDÖS-RÉNYI RANDOM GRAPH ENTROPY



WATTS-STROGATZ RANDOM GRAPH MODEL



Regular

Small-world

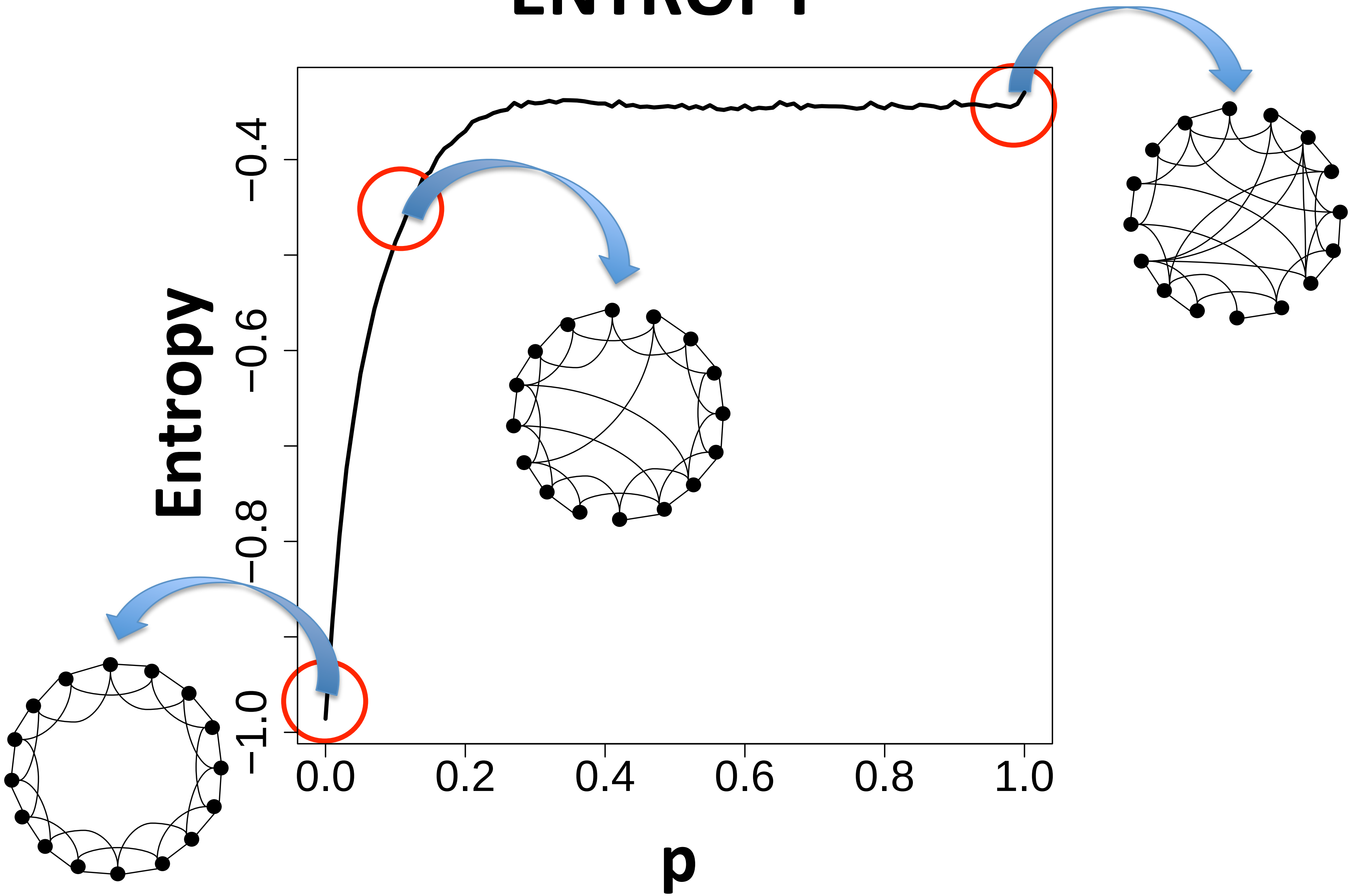
Random

p = 0

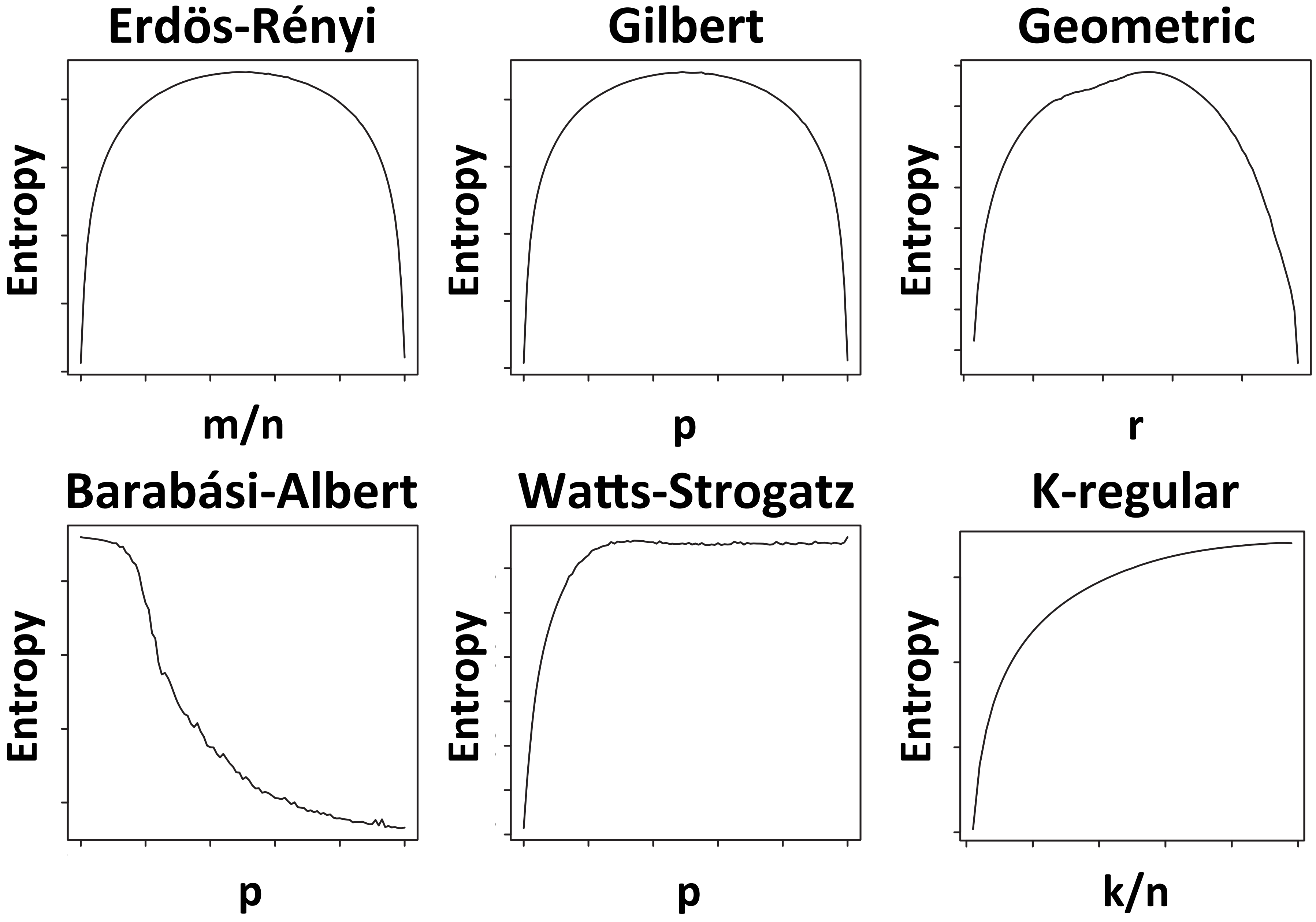
p = 1

Increasing randomness

WATTS-STROGATZ RANDOM GRAPH ENTROPY



GRAPH ENTROPY



Erdös-Rényi

Gilbert

Geometric

Barabási-Albert

Watts-Strogatz

K-regular

DATASETS

ADHD-200 Consortium

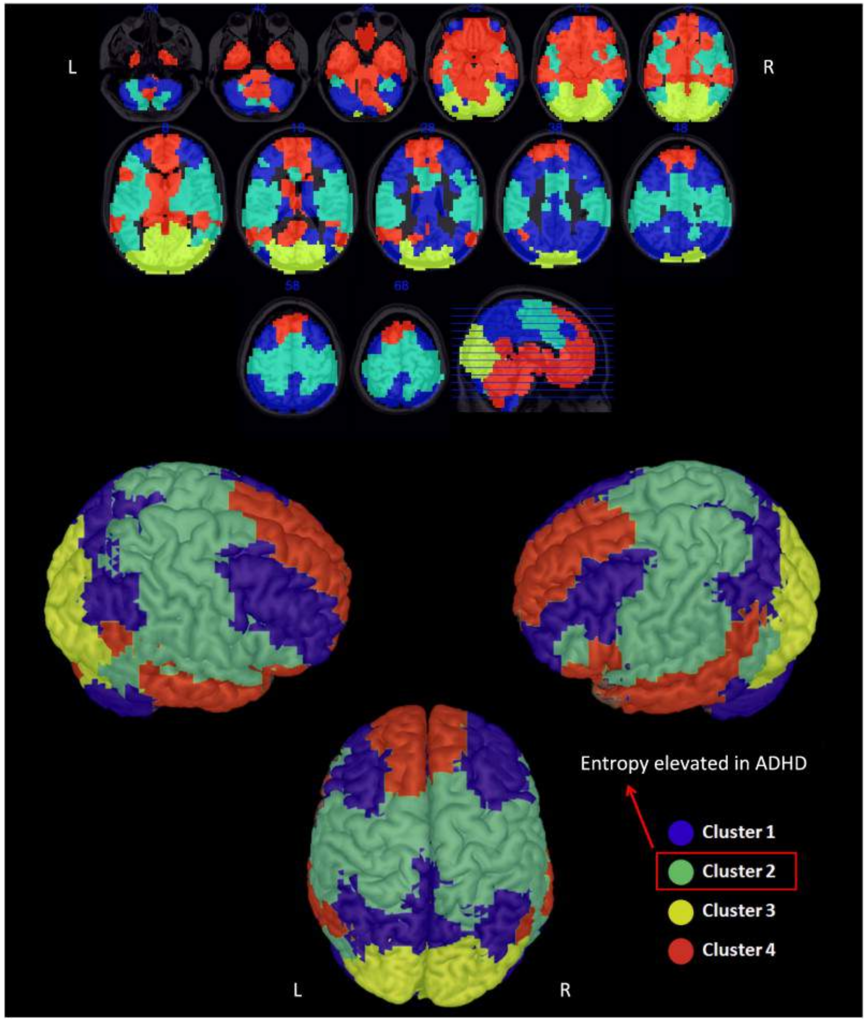
- 759 subjects
- 479 controls (253 males, 12.23±3.26 y.o.)
- 159 combined hyperactive/impulsive and inattentive (130 males, 11.24±3.05 y.o.)
- ~~11 hyperactive/impulsive (9 males, 13.40±4.51 y.o.)~~
- 110 inattentive (85 males, 12.06±2.55 y.o.)
- Pre-processing: Athena pipeline

ABIDE I Consortium

- 814 subjects
- 529 controls (430 males, 17.47±7.81 y.o.)
- 285 autism patients (255 males, 17.53±7.13 y.o.)
- Pre-processing: Athena pipeline

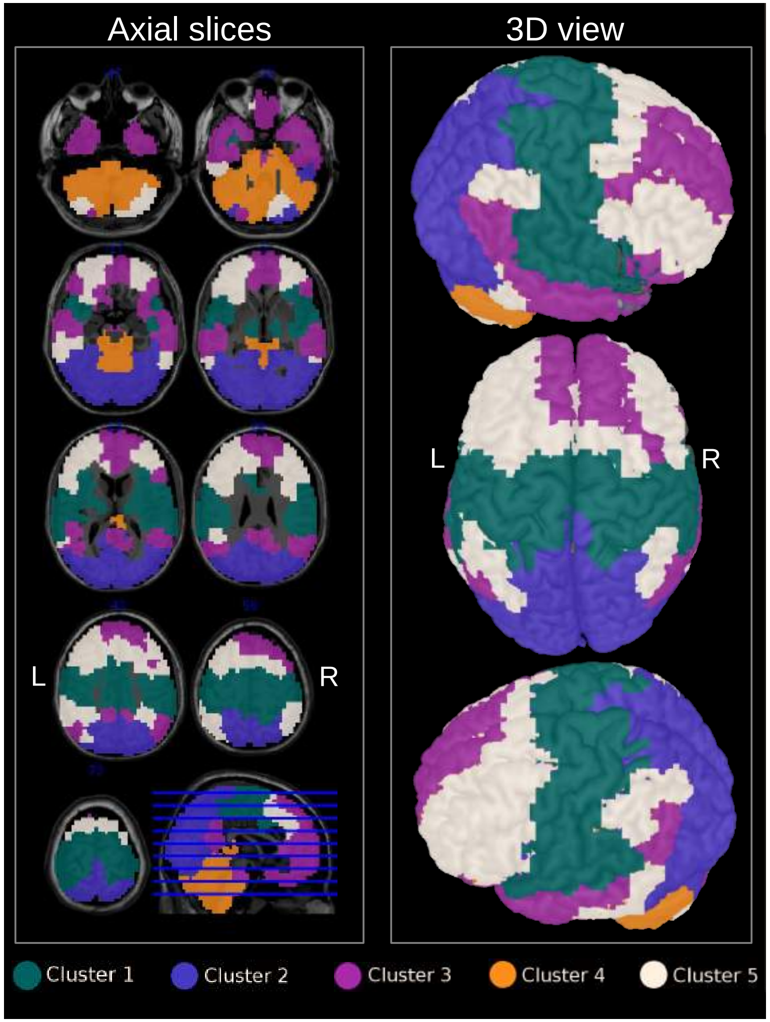
GRAPH ENTROPY

Attention Deficit Hyperactivity Disorder



Sato et al., 2013

Autism Spectrum Disorder

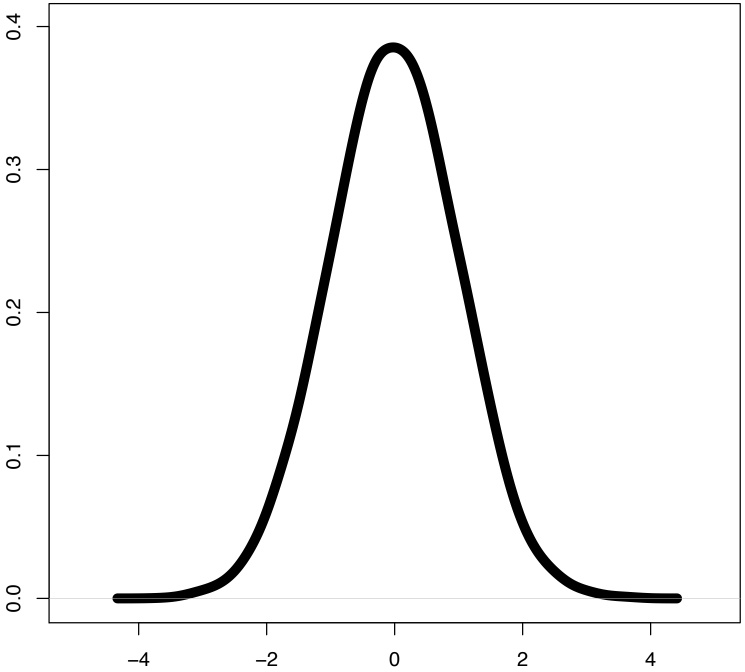


Sato et al., 2015

PARAMETER ESTIMATION

Data

0.018
-0.184
-1.371
-0.599
0.294
0.389
-1.208
-0.363
-1.626
-0.256
1.101
0.755
-0.238
0.987
0.741
0.089

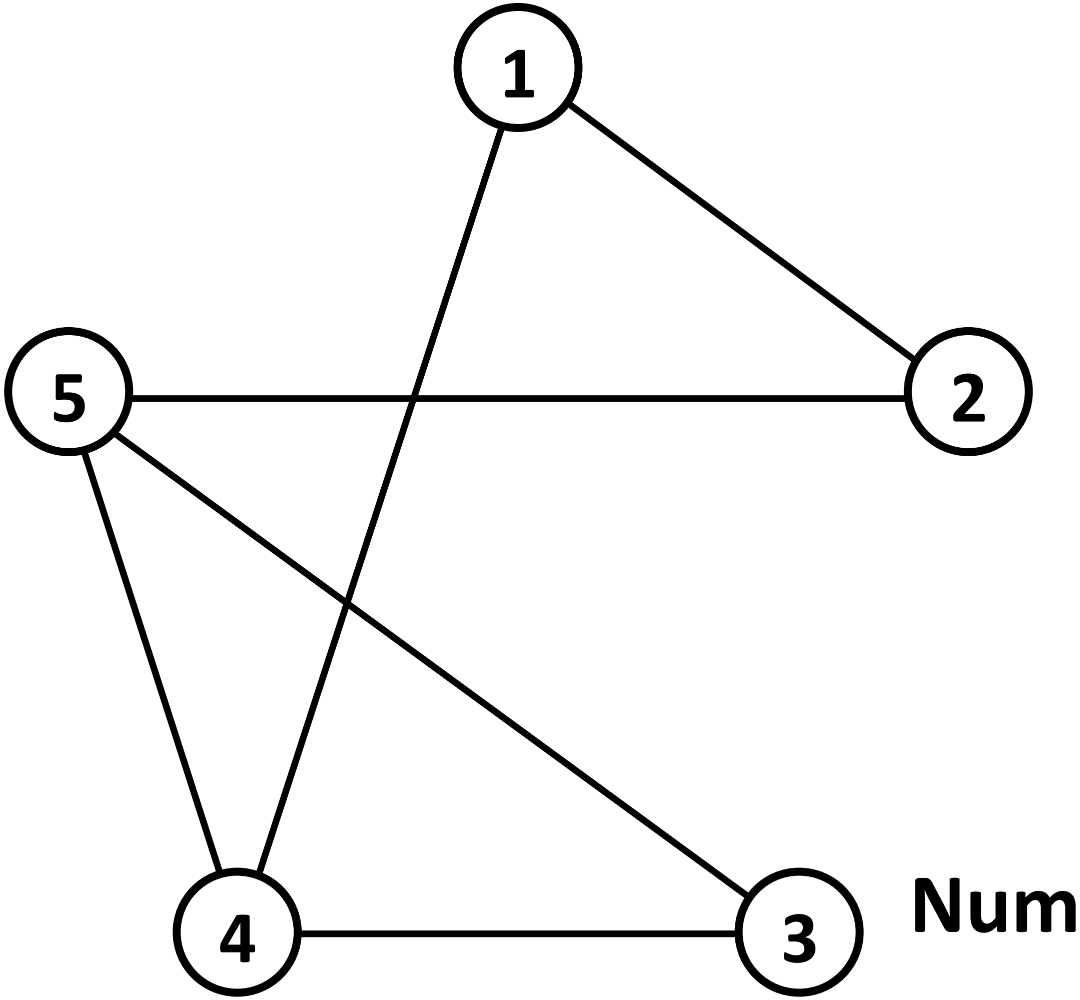


$$N(\mu, \sigma^2)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = -0.091$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.671$$

ERDÖS-RÉNYI RANDOM GRAPH MODEL



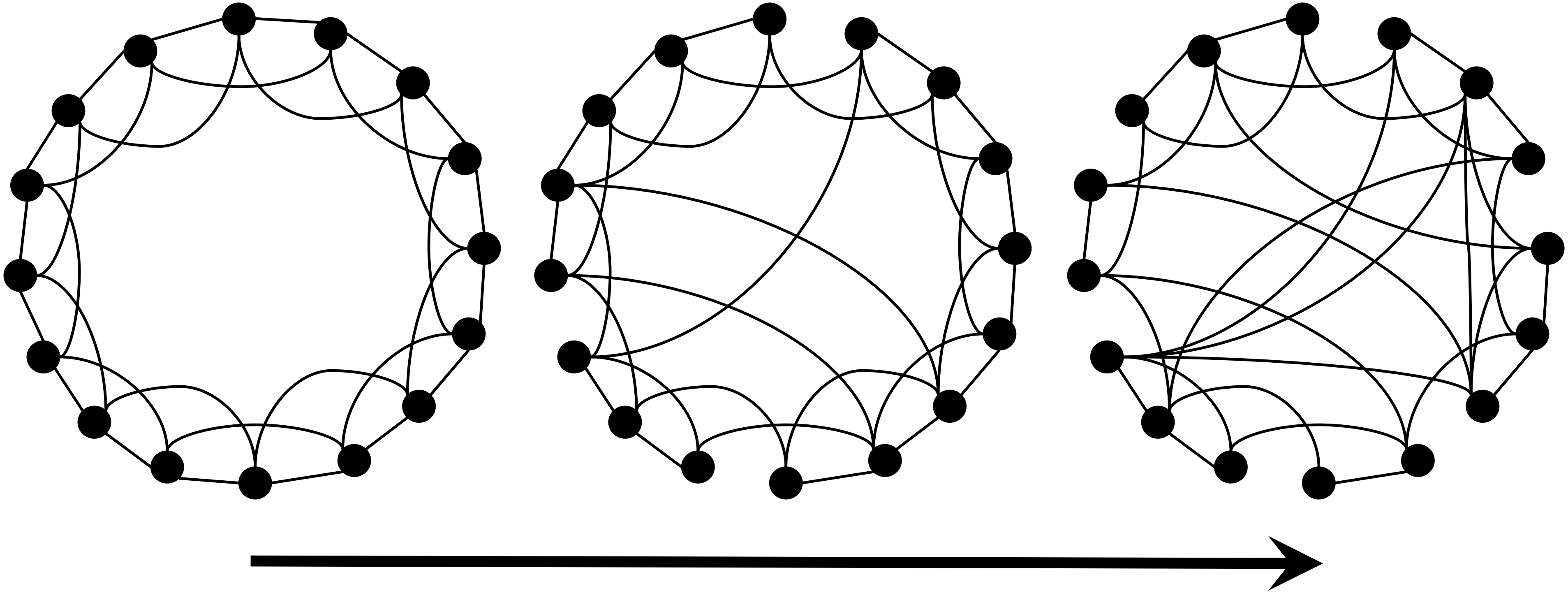
Number of edges

$$\hat{p} = \frac{6}{\binom{5}{2}} = 0.6$$

Number of possible edges

$$\binom{n}{2}$$

WATTS-STROGATZ RANDOM MODEL



Regular

Small-world

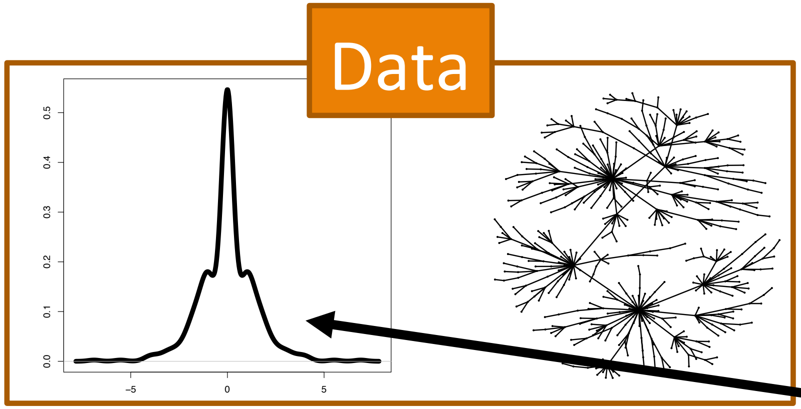
Random

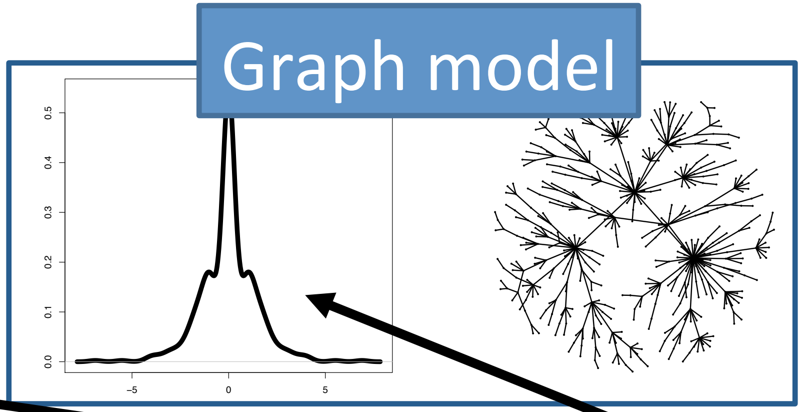
p = 0

p = 1

Increasing randomness

PARAMETER ESTIMATION





Kullback-Leibler divergence

$$\mathrm{KL}(\rho_1|\rho_2) = \int_{-\infty}^{\infty} \rho_1(\lambda) \log \frac{\rho_1(\lambda)}{\rho_2(\lambda)} \mathrm{d}\lambda$$

$$\boxed{\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathrm{KL}(\hat{\rho}_g|\rho_\theta)}$$

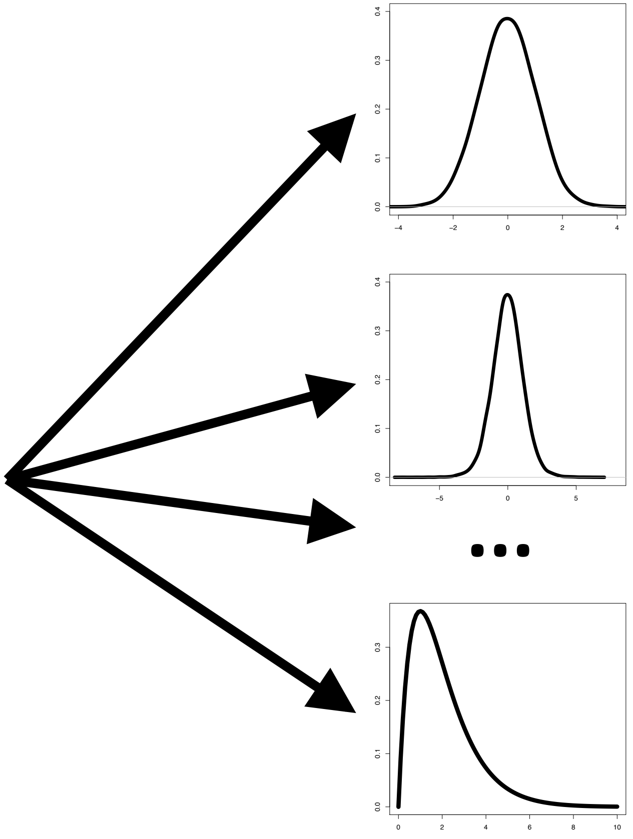
	Size			
Model	20	50	100	500
ER (*m*=0.5)	0.503±0.013	0.500±0.002	0.500±0	0.499±0.003
GI (*p*=0.5)	0506±0.039	0.501±0.014	0.501±0.008	0.499±0.003
GE (*r*=0.5)	0.493±0.061	0.506±0.037	0.502±0.022	0.500±0.010
BA (*p*=1)	1.128±0.309	1.044±0.125	1.026±0.047	1.020±0.025
WS (*k*=0.25)	0.129±0.155	0.069±0.011	0.071±0.008	0.070±0.003
KR (*k*=0.25)	0.264±0.013	0.245±0.005	0.250±0	0.249±0.004

ER: Erdös-Rényi
GI: Gilbert
GE: Geometric
BA: Barabasi-Albert
WS: Watts-Strogatz
KR: K-Regular

Takahashi et al., 2012
de Siqueira Santos et al., 2016

MODEL SELECTION

Data
0.018
-0.184
-1.371
-0.599
0.294
0.389
-1.208
-0.363
-1.626
-0.256
1.101
0.755
…

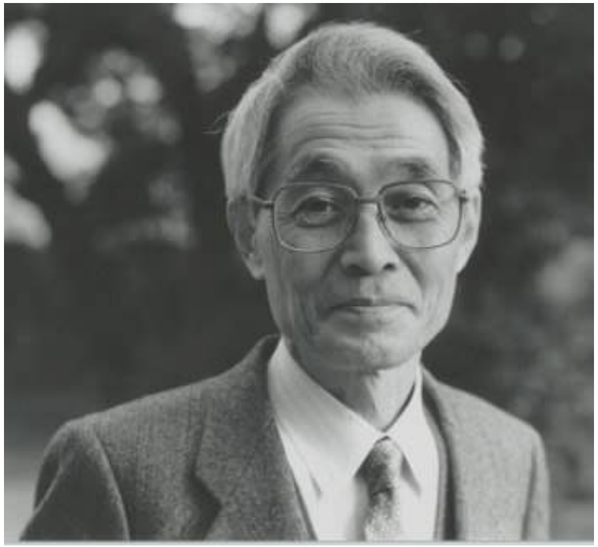


Normal

t

…

Gamma



Hirotugu Akaike
1927 - 2009

Akaike Information Criterion - AIC

$$\hat{L} = P(x|\hat{\theta}, M)$$

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

MODEL SELECTION

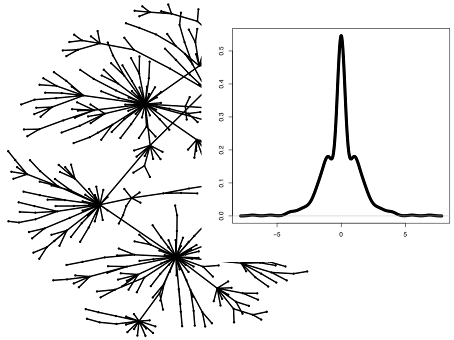
Kullback-Leibler divergence

$$\mathrm{KL}(\rho_{g1}|\rho_{g2}) = \int_{-\infty}^{\infty} \rho_1(\lambda) \log \frac{\rho_{g1}(\lambda)}{\rho_{g2}(\lambda)} \mathrm{d}\lambda$$

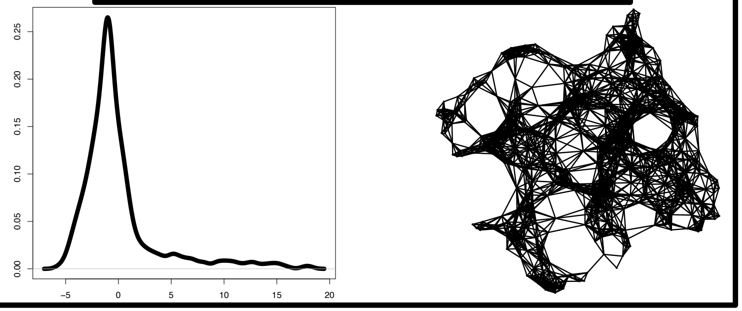
Unknown graph spectrum

Reference spectrum

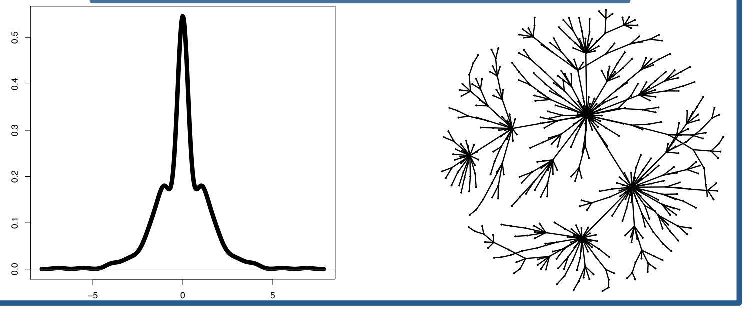
Unknown graph



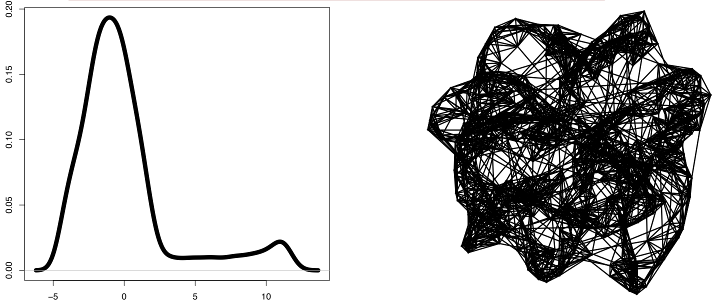
Reference graph model 1



Reference graph model 2



Reference graph model 3

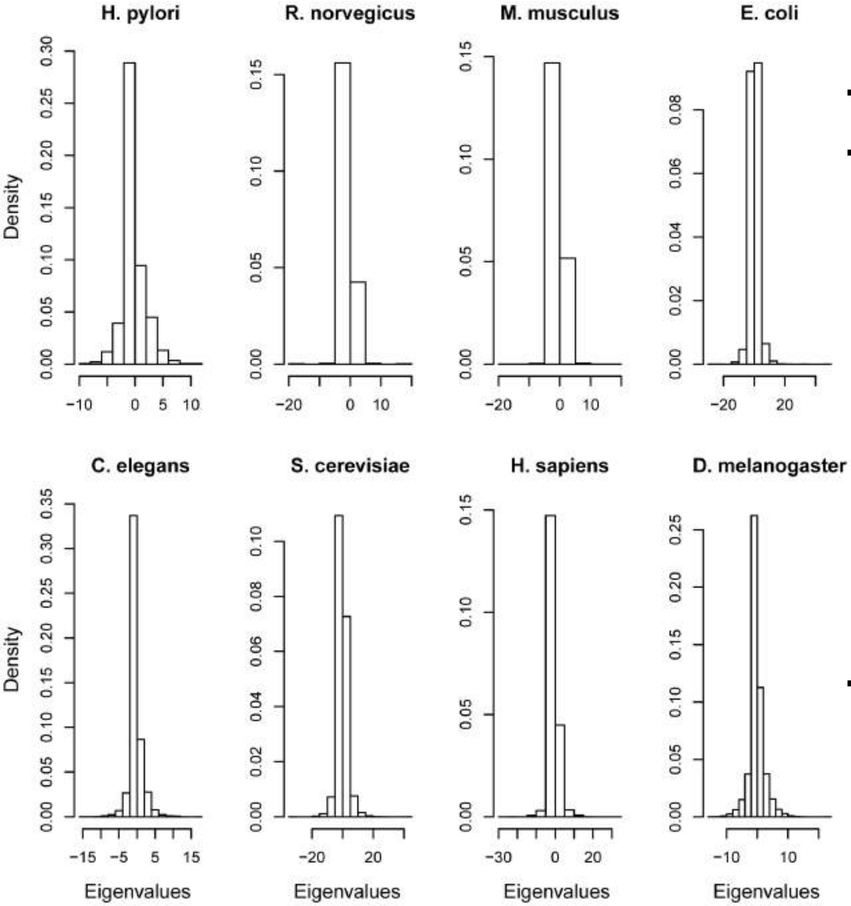


$$j = \underset{i}{\mathrm{argmin}}\, 2\mathrm{KL}(\hat{\rho}|\rho_{\hat{\theta}_i}) + \underbrace{2|\hat{\theta}_i|}$$

Penalization to avoid overfitting

Takahashi et al., 2012

Species	Number of nodes	Number of edges	Average degree	Diameter	Clustering coefficient	Average path length
H. pylori	714	1,393	3.9	9	0.016	4.139
R. norvegicus	758	691	1.82	9	0.001	3.651
M. musculus	1,868	1,895	2.03	20	0.006	6.28
E. coli	2,997	12,348	8.24	12	0.115	3.986
C. elegans	3,183	5,068	3.18	13	0.012	4.803
S. cerevisiae	5,213	25,073	9.62	10	0.058	3.86
H. sapiens	5,940	14,144	4.76	17	0.017	4.755
D. melanogaster	7,931	23,386	5.9	12	0.012	4.468

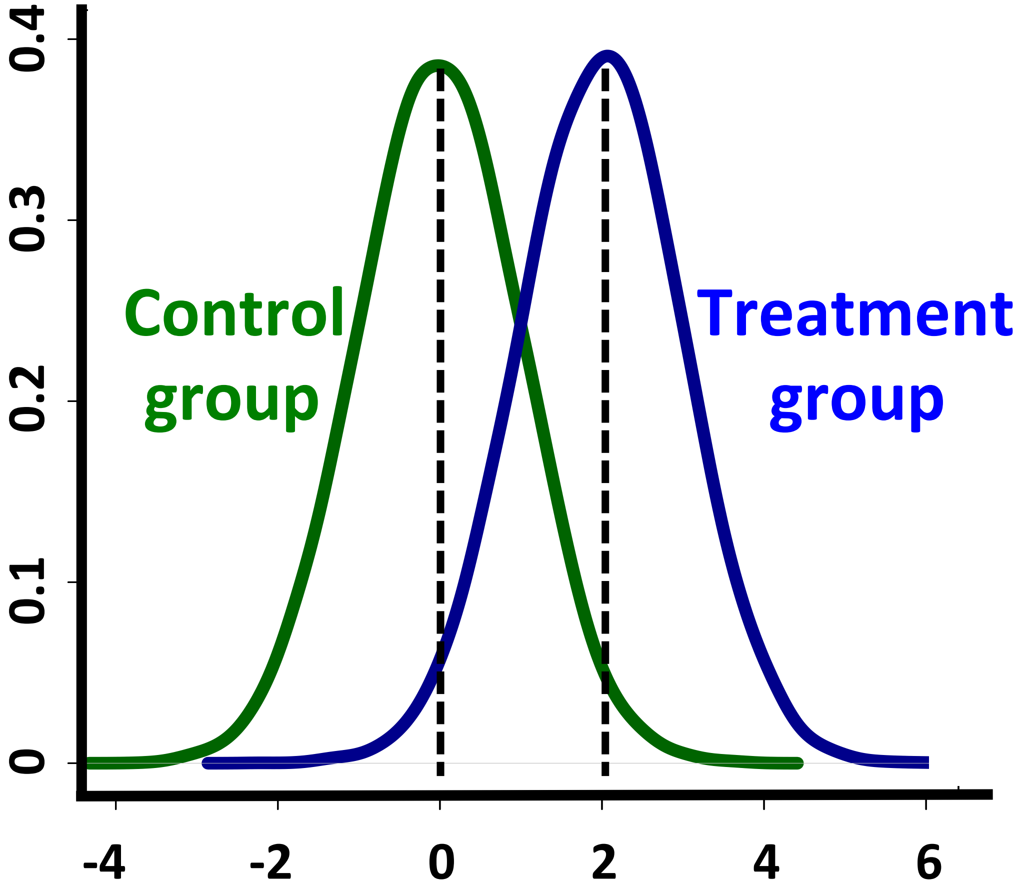


Species	Erdös-Rényi	Scale-free	Small-world
H. pylori	15.07	**1.46**	11.36
R. norvegicus	134.67	**100.47**	118.67
M. musculus	14.1	**6.93**	24.51
E. coli	21.15	**1.91**	17.9
C. elegans	30.48	**2.66**	30.23
S. cerevisiae	24.21	**0.87**	18.25
H. sapiens	47.1	**11.31**	44.04
D. melanogaster	17.4	**0.39**	18.06

Takahashi et al., 2012

T test

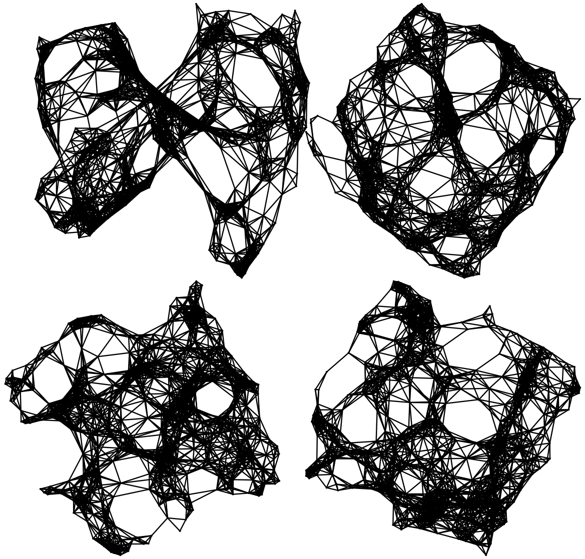
Control	Treatment
0.018	2.974
-0.184	1.993
-1.371	3.567
-0.599	2.474
0.294	1.055
0.389	0.456
-1.208	4.654
-0.363	-0.148
-1.626	0.231
-0.256	1.612
1.101	4.254
0.755	3.035
-0.238	4.236
0.987	3.263
0.741	3.138
0.089	1.571
...	...



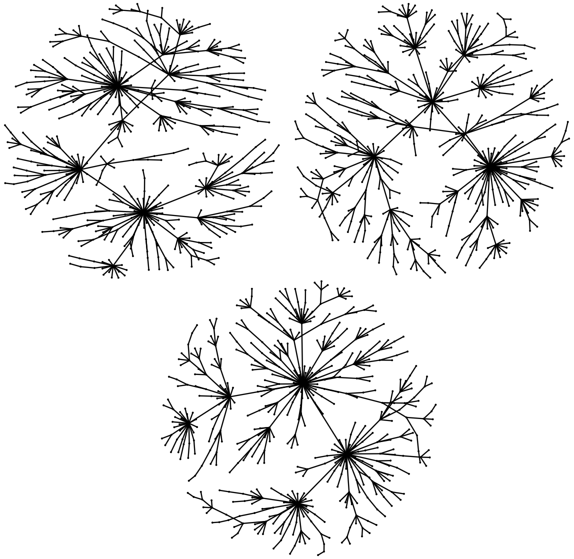
H_0: the means of the two populations are equal

H_1: the means of the two populations are not equal

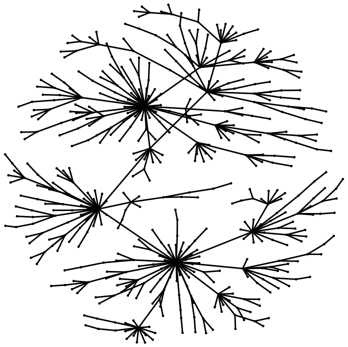
H_0:



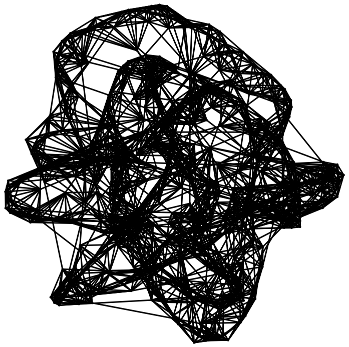
=

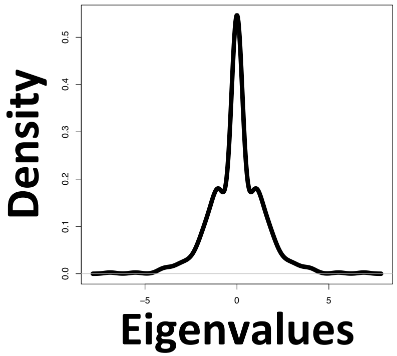


COMPARISON TEST

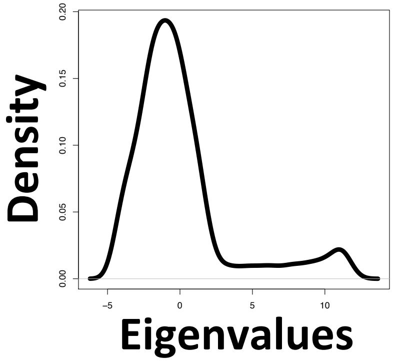


“distance”





Jensen-Shannon divergence



$$\mathrm{JS}\left(\rho_{g_1},\rho_{g_2}\right)=\frac{1}{2}\mathrm{KL}\left(\rho_{g_1}|\rho_{g_M}\right)+\frac{1}{2}\mathrm{KL}\left(\rho_{g_2}|\rho_{g_M}\right)$$

where $\rho_{g_M}=\frac{1}{2}(\rho_{g_1}+\rho_{g_2})$

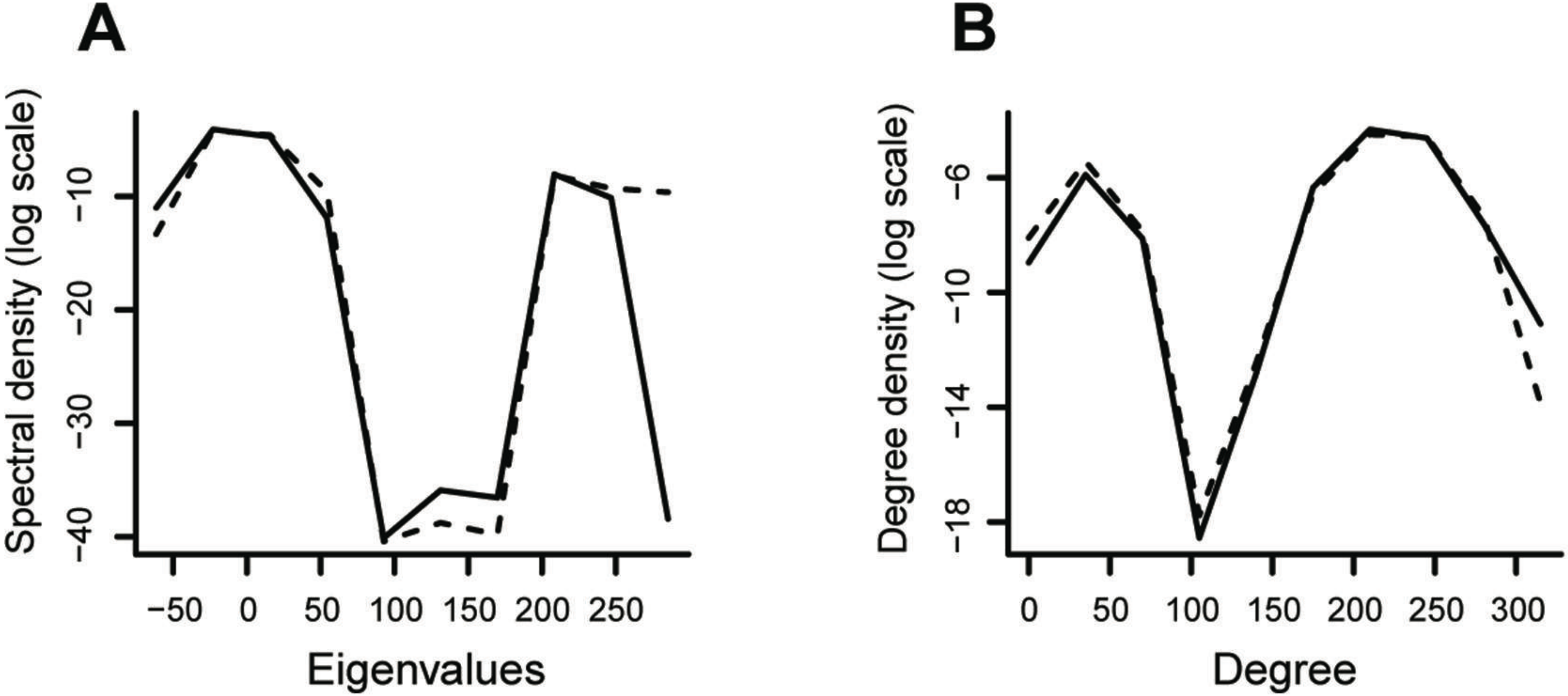
Hypothesis test

$\mathrm{H}_0\colon \mathrm{JS}\left(\rho_{g_1},\rho_{g_2}\right)=0$

$\mathrm{H}_1\colon \mathrm{JS}\left(\rho_{g_1},\rho_{g_2}\right)>0$

Takahashi et al., 2012

ADHD

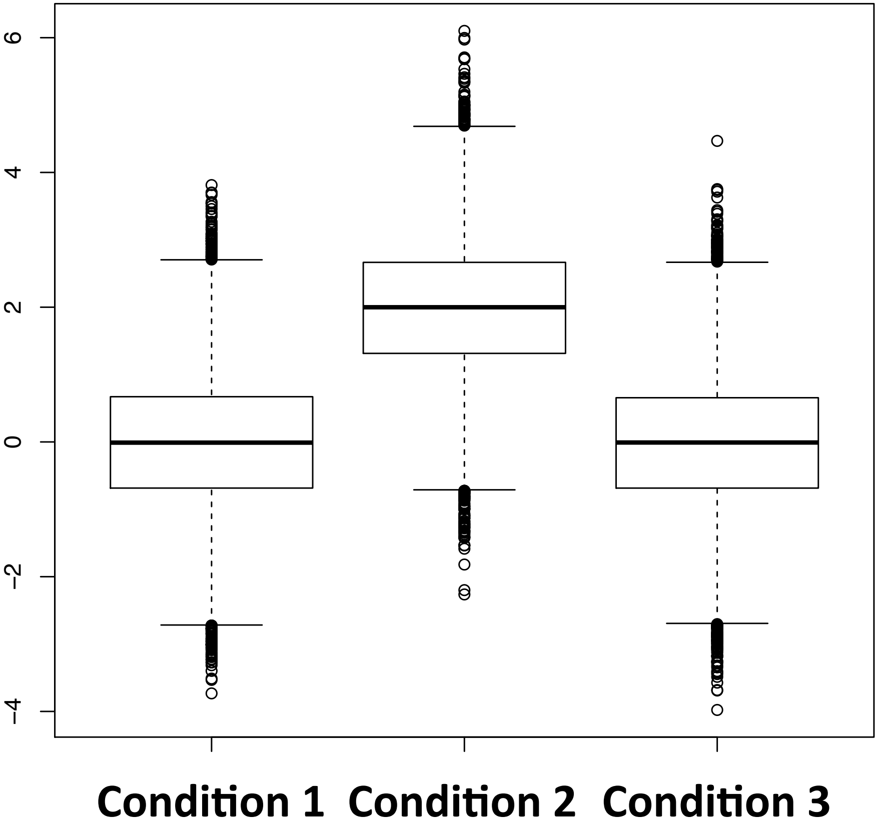


	Number of edges	Clustering coefficient	Average path length	Degree distribution	Spectrum
Normal vs ADHD	0.82	0.85	0.87	0.031	0.024

Takahashi et al., 2012

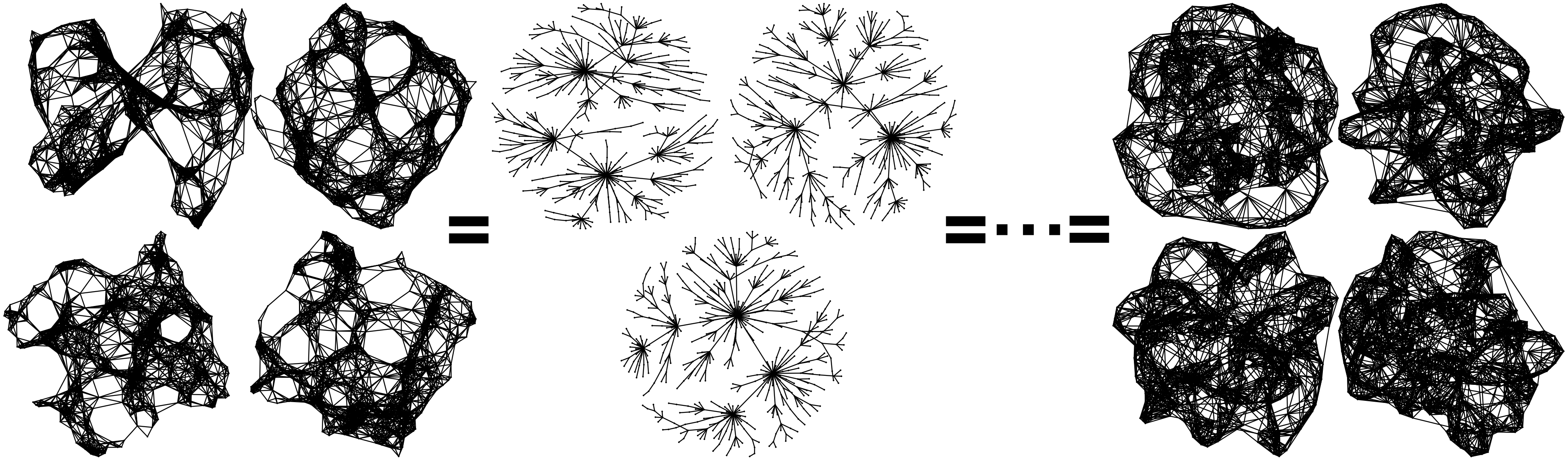
ANOVA
(Analysis of Variance)

Condition 1	Condition 2	Condition 3
0.018	2.974	1.729
-0.184	1.993	-1.071
-1.371	3.567	-2.339
-0.599	2.474	-0.379
0.294	1.055	2.511
0.389	0.456	-0.044
-1.208	4.654	0.929
-0.363	-0.148	-0.891
-1.626	0.231	-0.892
-0.256	1.612	1.204
1.101	4.254	-0.077
0.755	3.035	-1.944
-0.238	4.236	-0.816
0.987	3.263	-1.103
0.741	3.138	0.623
0.089	1.571	-0.104
…	…	…

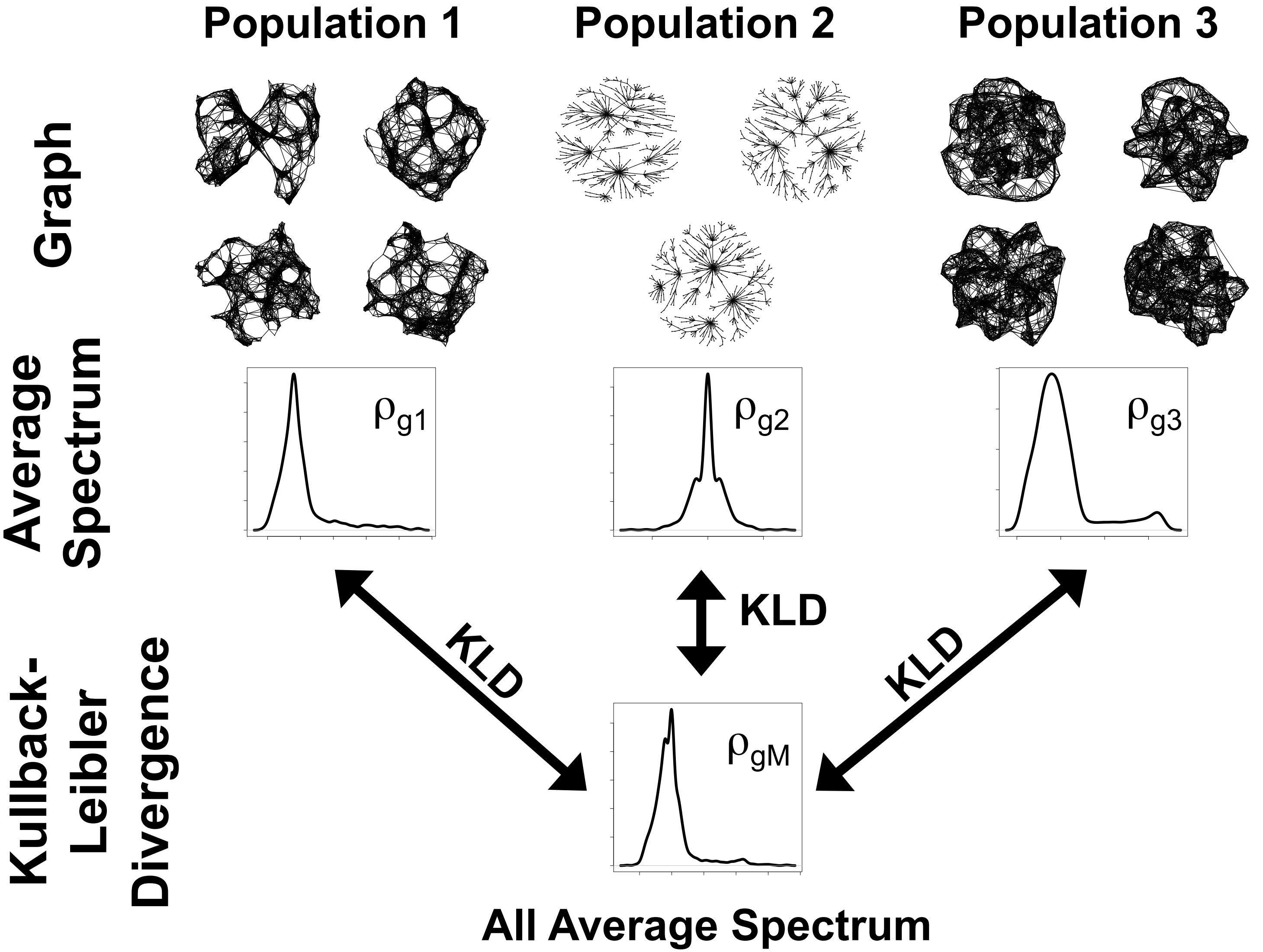


H_0: all the means are equal

H_1: at least one of the means is not equal



ANOGVA: Analysis of Graph Variability


Population 1

Population 2

Population 3

Graph

Average Spectrum

Kullback-Leibler Divergence

KLD

KLD

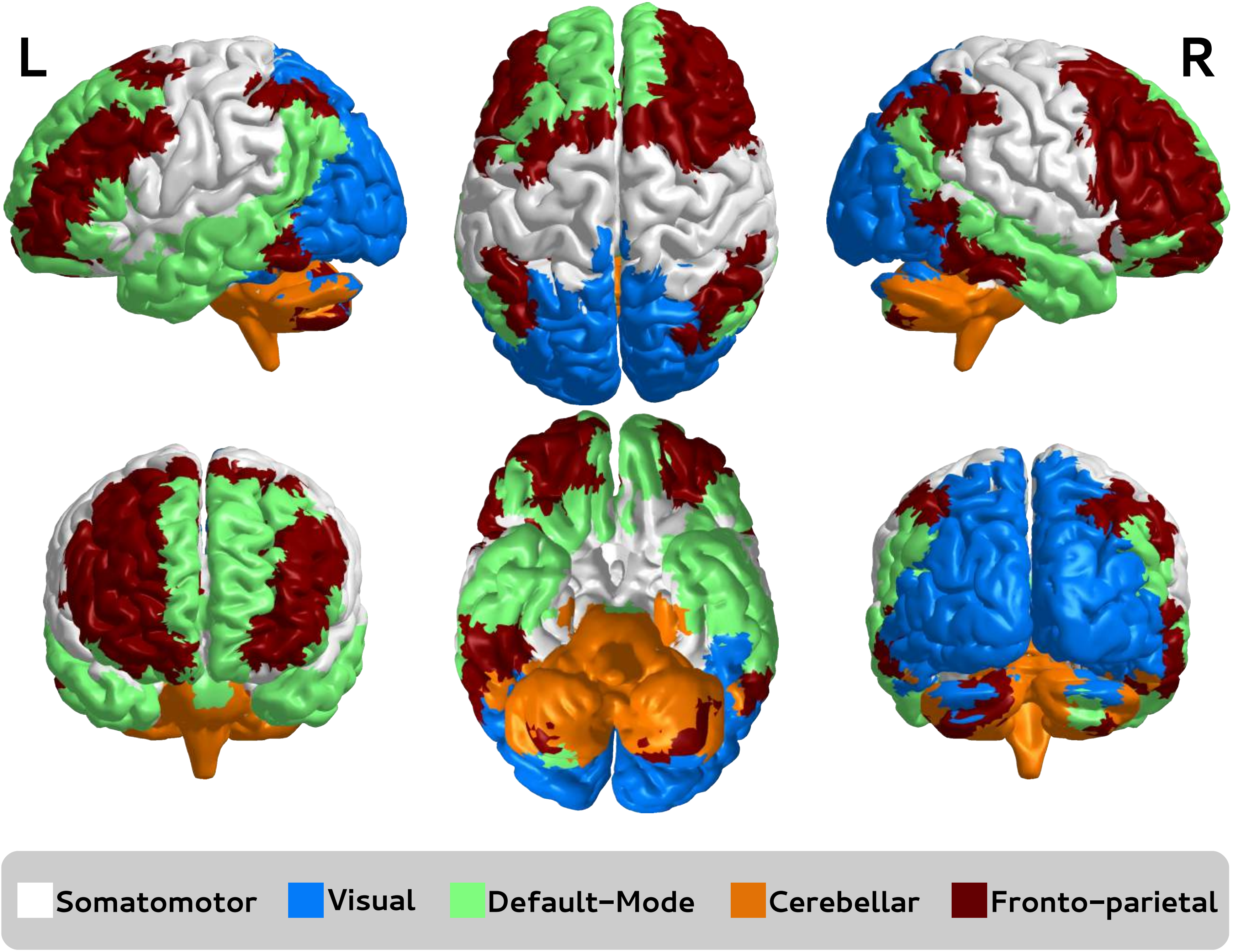
KLD

All Average Spectrum

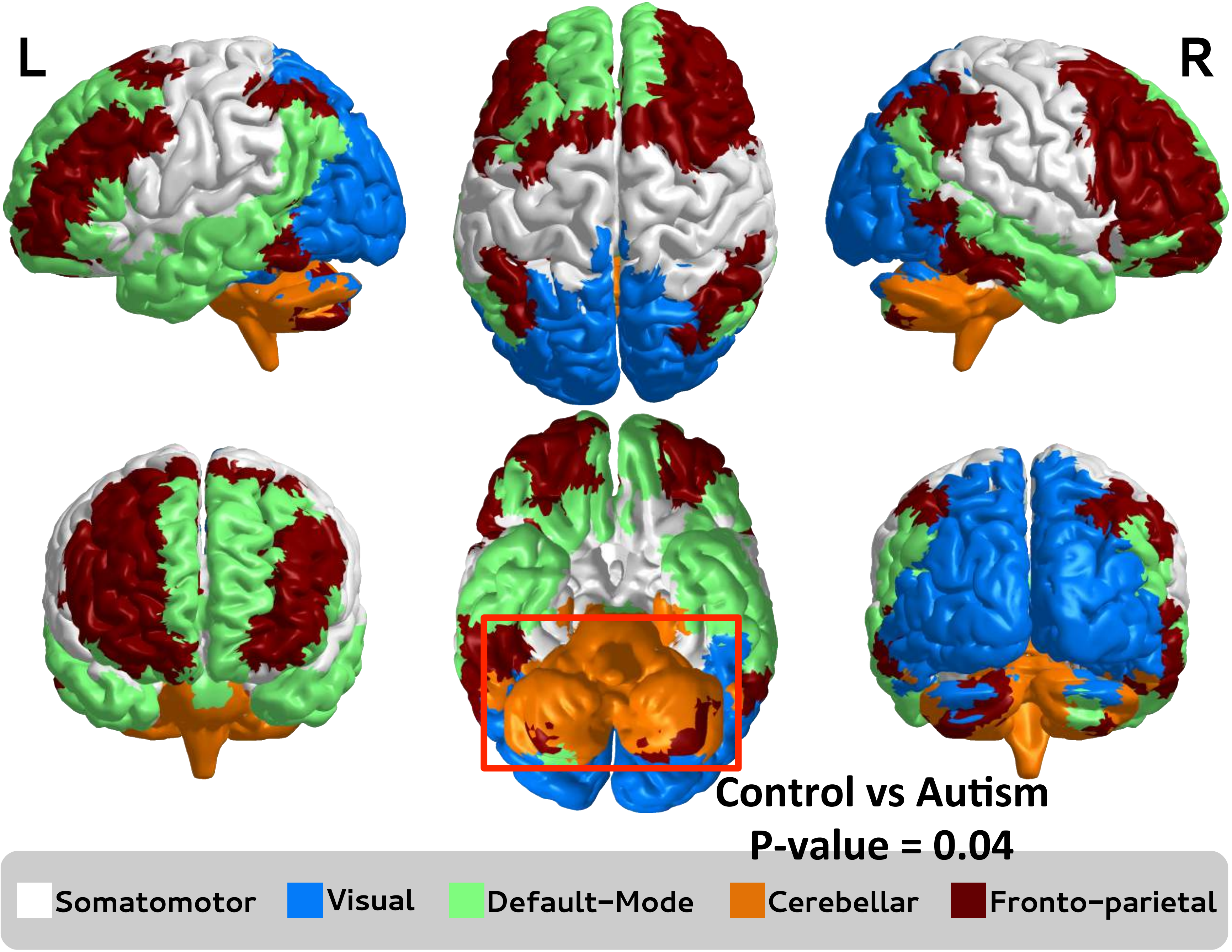
H_0: $\mathrm{KL}(\rho_{g_1}, \rho_{g_\mathrm{M}}) = \mathrm{KL}(\rho_{g_2}, \rho_{g_\mathrm{M}}) = \cdots = \mathrm{KL}(\rho_{g_k}, \rho_{g_\mathrm{M}}) = 0$

H_1:At least one population of graphs is generated in a different manner

Fujita et al., 2017



Fujita et al., 2017

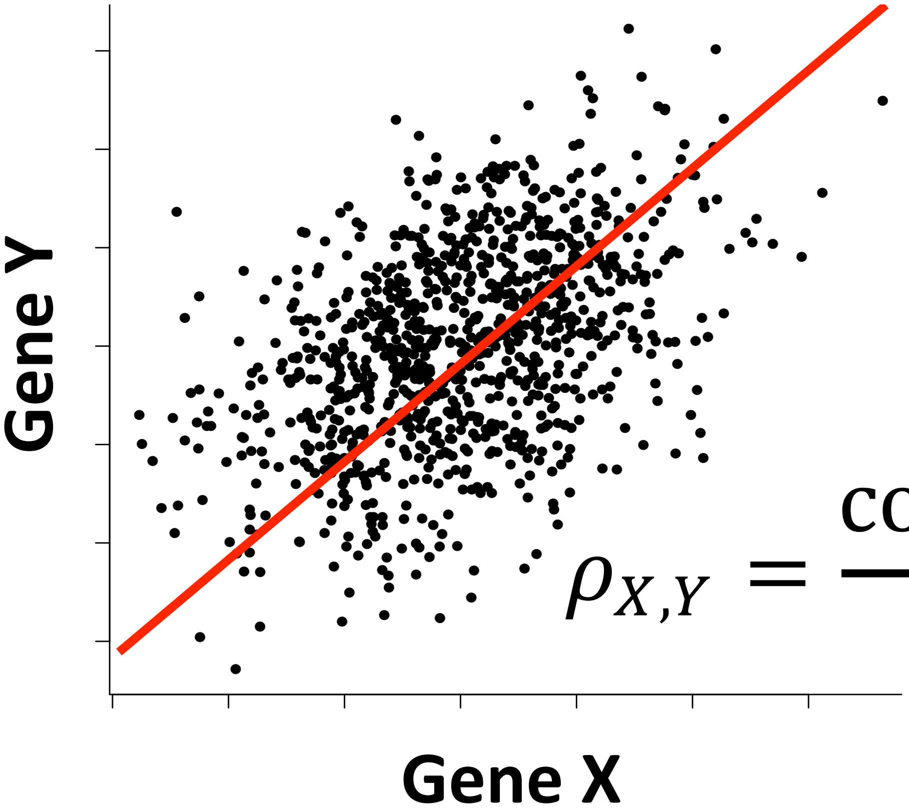


Control vs Autism
P-value = 0.04

Fujita et al., 2017

Correlation

Gene X	Gene Y
-0.508	0.541
-0.302	-0.016
-1.302	-1.067
-0.935	-0.102
0.366	0.143
0.193	0.481
-0.876	-1.216
-0.845	0.215
-0.804	-2.013
-0.748	0.304
1.015	0.892
0.280	1.028
-0.966	0.553
0.926	0.783
0.882	0.401
0.559	-0.404
…	…

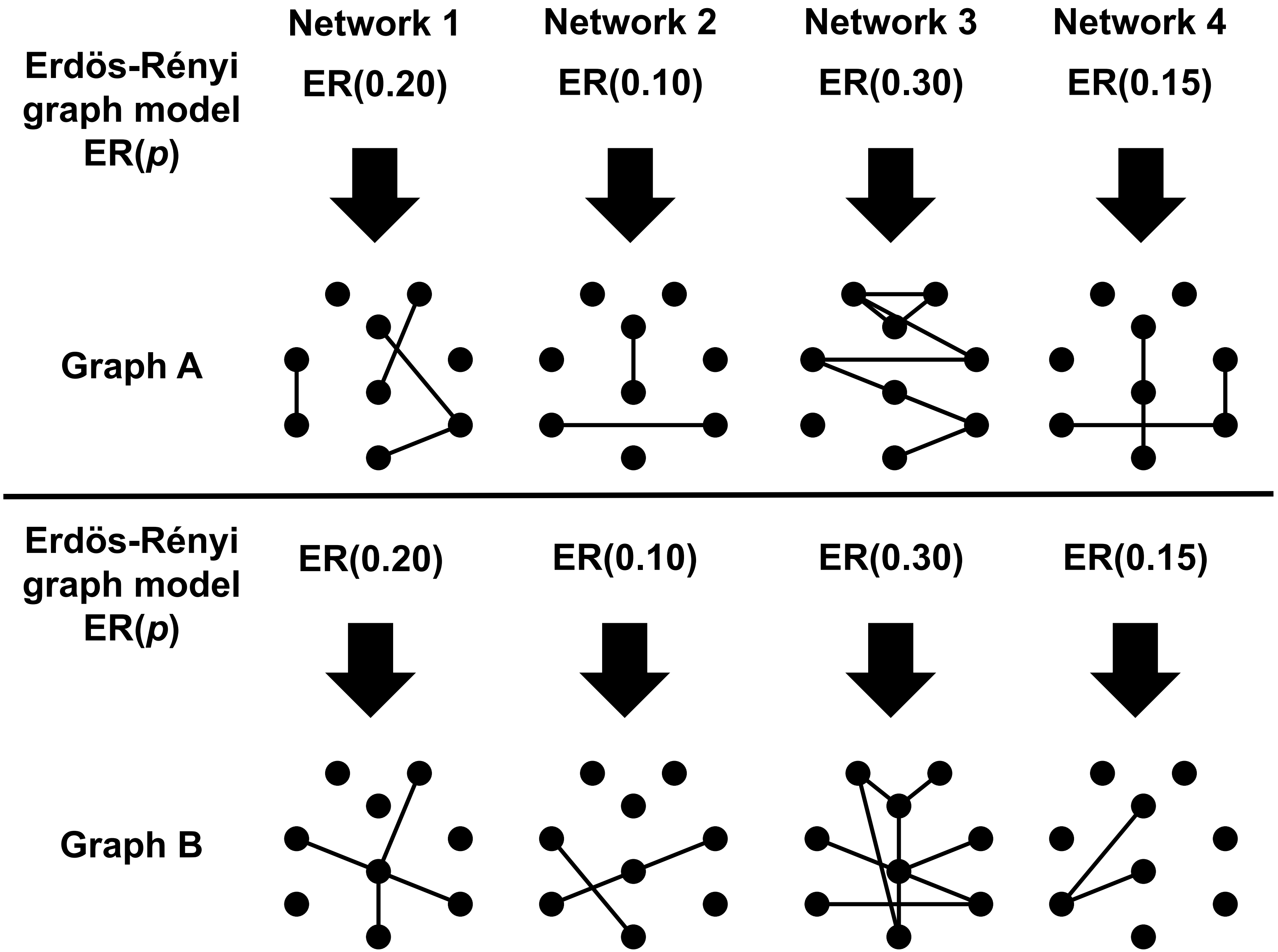


$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Hypothesis test

$\mathrm{H}_0 : \rho_{X,Y} = 0$ **X and Y are not linearly dependent**

$\mathrm{H}_1 : \rho_{X,Y} \neq 0$ **X and Y are linearly dependent**



Network 1

Network 2

Network 3

Network 4

Erdös-Rényi graph model ER(*p*)

ER(0.20)

ER(0.10)

ER(0.30)

ER(0.15)

Graph A

Erdös-Rényi graph model ER(*p*)

ER(0.20)

ER(0.10)

ER(0.30)

ER(0.15)

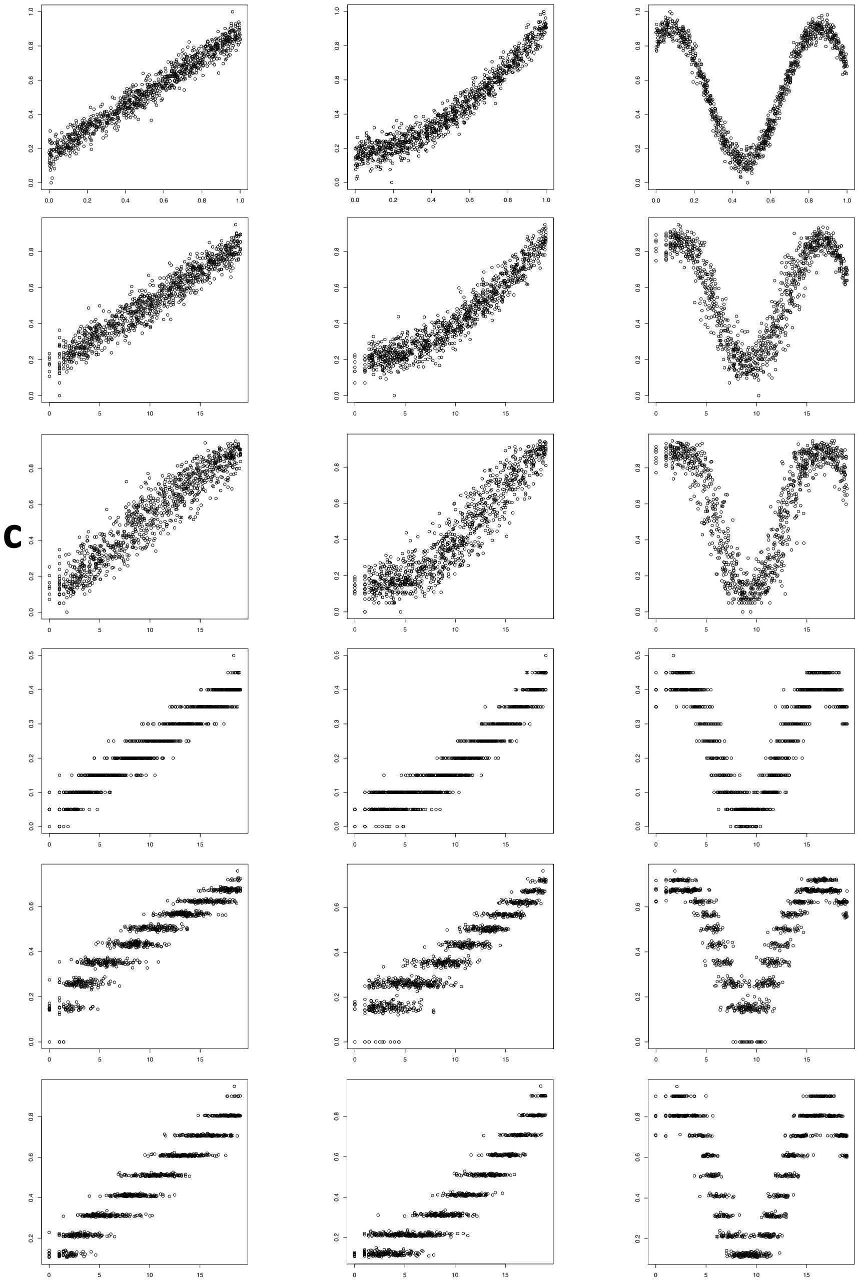
Graph B

Fujita et al., 2016

Linear

Monotonic

Non-monotonic



True

Erdös-Rényi

Geometric

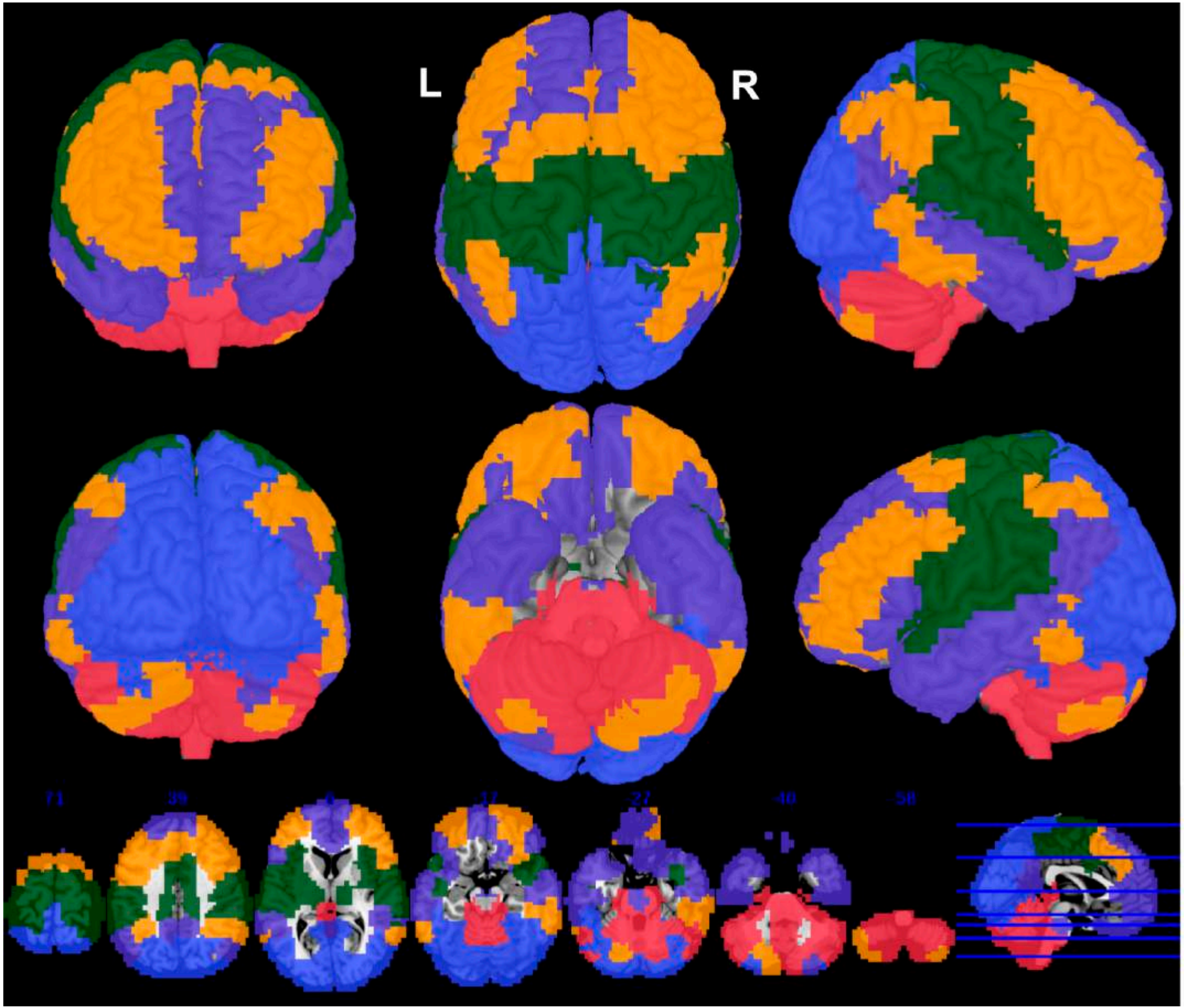
K-Regular

Barabási-Albert

Watts-Strogatz

Fujita et al., 2016

AUTISM SPECTRUM DISORDER



● Somatomotor

● Visual

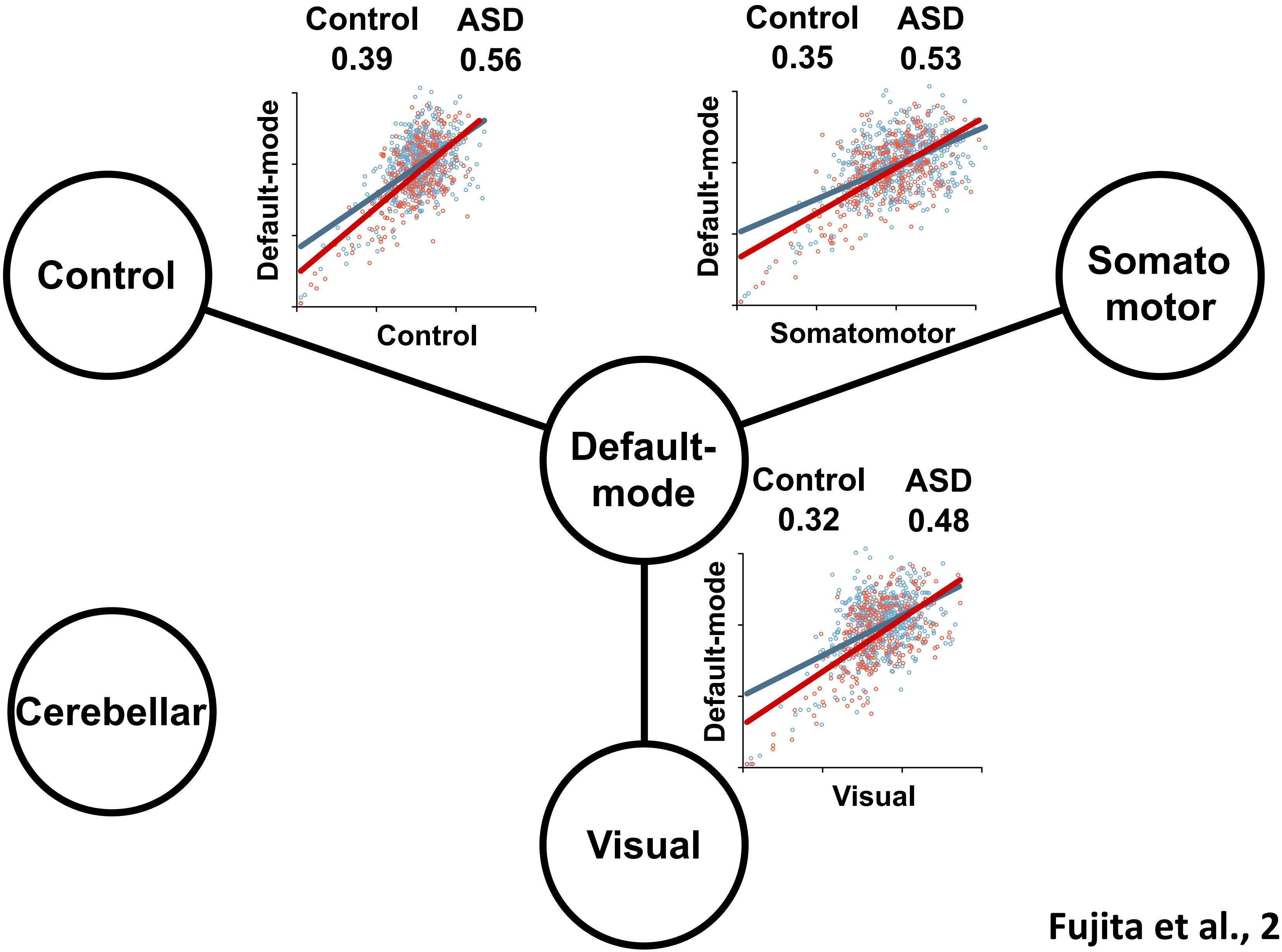
● Default-mode

● Cerebellar

● Control

Fujita et al., 2016

AUTISM SPECTRUM DISORDER



Control

Somato motor

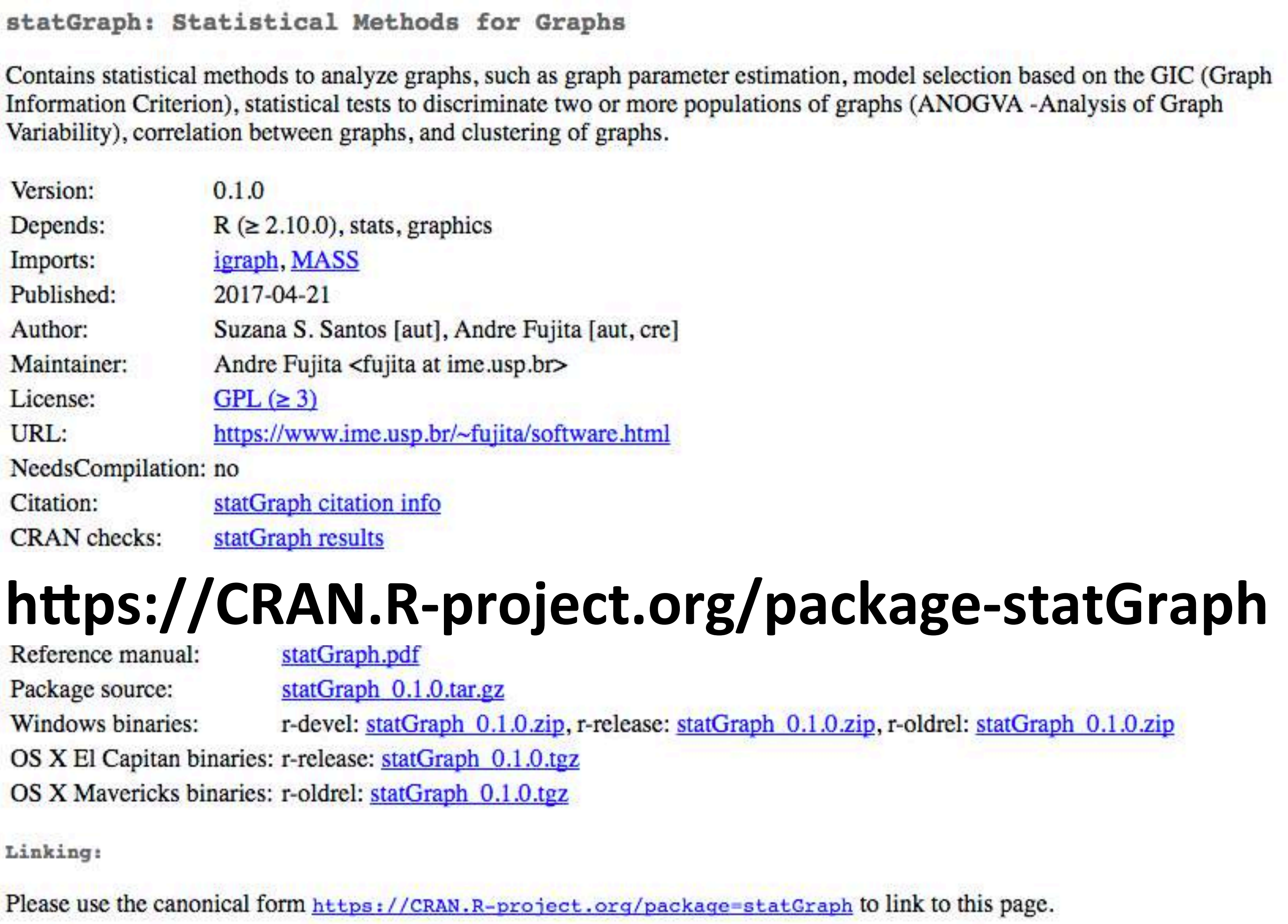
Default-mode

Cerebellar

Visual

Fujita et al., 2016

statGraph



statGraph: Statistical Methods for Graphs

Contains statistical methods to analyze graphs, such as graph parameter estimation, model selection based on the GIC (Graph Information Criterion), statistical tests to discriminate two or more populations of graphs (ANOGVA -Analysis of Graph Variability), correlation between graphs, and clustering of graphs.

Version:	0.1.0
Depends:	R (≥ 2.10.0), stats, graphics
Imports:	igraph, MASS
Published:	2017-04-21
Author:	Suzana S. Santos [aut], Andre Fujita [aut, cre]
Maintainer:	Andre Fujita <fujita at ime.usp.br>
License:	GPL (≥ 3)
URL:	https://www.ime.usp.br/~fujita/software.html
NeedsCompilation:	no
Citation:	statGraph citation info
CRAN checks:	statGraph results

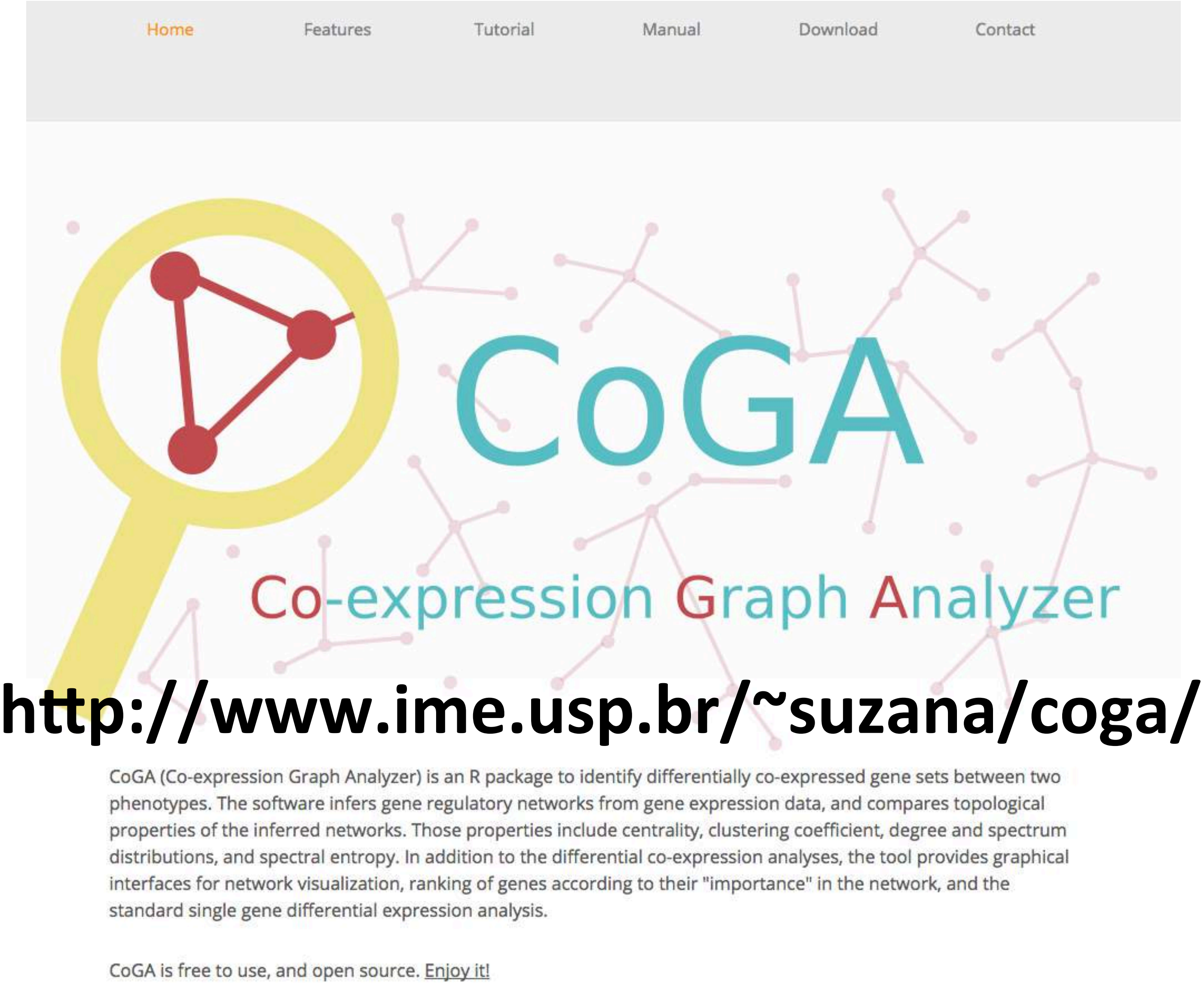
https://CRAN.R-project.org/package-statGraph

Reference manual:	statGraph.pdf
Package source:	statGraph_0.1.0.tar.gz
Windows binaries:	r-devel: statGraph_0.1.0.zip, r-release: statGraph_0.1.0.zip, r-oldrel: statGraph_0.1.0.zip
OS X El Capitan binaries:	r-release: statGraph_0.1.0.tgz
OS X Mavericks binaries:	r-oldrel: statGraph_0.1.0.tgz

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=statGraph` to link to this page.

CoGA: Co-expression Graph Analyzer



http://www.ime.usp.br/~suzana/coga/

CoGA (Co-expression Graph Analyzer) is an R package to identify differentially co-expressed gene sets between two phenotypes. The software infers gene regulatory networks from gene expression data, and compares topological properties of the inferred networks. Those properties include centrality, clustering coefficient, degree and spectrum distributions, and spectral entropy. In addition to the differential co-expression analyses, the tool provides graphical interfaces for network visualization, ranking of genes according to their "importance" in the network, and the standard single gene differential expression analysis.

CoGA is free to use, and open source. Enjoy it!

de Siqueira Santos et al., 2015

University of São Paulo (students)

Abner C.R. Neto (Post-Doc, CAPES)
Adèle H. Ribeiro (Ph.D. candidate, CAPES)
Bruno Yamada (undergrad student)
Carlos Farfan (Master candidate)
Carlos Relvas (Ph.D. candidate)
Eduardo Lira (Ph.D. candidate, CAPES)
Gabriela E. Soares (Ph.D. candidate, CAPES)
Grover E.C. Guzman (Ph.D. candidate, CAPES)
João Madeira (undergrad student)
Juliana C. Cavalcanti (Master candidate)
Maciel C. Vidal (Ph.D. candidate, CAPES)
Suzana S. Santos (Ph.D. candidate, FAPESP)
Taiane C. Ramos (Ph.D. candidate, CNPq)
Vinicius J. Carvalho (Master candidate, CAPES)

University of São Paulo

Carlos Eduardo Ferreira
Suely Kazue Nagahashi Marie

Federal University of ABC

João Ricardo Sato

Albert Einstein Research and Education Institute

Joana Bisol Balardin

Friedrich Alexander Universität Erlangen-Nürnberg

Lars Schewe

Princeton University

Daniel Yasumasa Takahashi

Sorbonne Université

Catherine Matias

University College London

Janaína Mourão-Miranda

Universität Leipzig

Peter Stadler

FAPESP

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES

JICA

Alexander von Humboldt
Stiftung/Foundation

Thank you for your attention

André Fujita

fujita@ime.usp.br

https://www.ime.usp.br/~fujita