Answers to exercises of Chapter 13 (Correlation Networks)

The following exercise section is intended to demonstrate basic statistical steps regarding generation and analyses of correlation based probability networks. It will mainly focus on statistical points. For biological interpretation the critical reader is refereed to the sections 12.5. The examples in the section 12.5 may serve as a basis for repeating the analyses by using different publicly available web resources as cited in section 12.4.3.

1. One of the crucial steps in correlation based network analyses represents the choice of association measure. The first example will briefly illustrate the different results which may be observed depending on the coefficient of association. This choice has advantages and disadvantages depending on the researcher's question.
A data matrix with 5 variables (var.1 – var.5) and 10 units (exp.1 – exp.10) is given in Table 12.3.
Table 12.3: Artificial m x n data set.

variables \ units	**exp.1**	**exp.2**	**exp.3**	**exp.4**	**exp.5**	**exp.6**	**exp.7**	**exp.8**	**exp.9**	**exp.10**
var.1	1.11	0.08	0.11	0.78	-0.33	-0.67	-0.15	0.56	-0.35	1.95
var.2	0.83	-0.25	-0.43	0.60	-0.82	-1.56	-0.91	-0.03	-0.92	1.36
var.3	0.26	-1.37	-1.25	-0.54	-1.59	-2.00	-1.23	-0.81	-1.73	0.60
var.4	-0.65	-0.07	0.17	-0.66	0.59	0.84	0.22	-0.54	0.97	-0.98
var.5	0.12	-0.10	0.24	0.23	-0.02	0.32	0.14	-0.16	-0.10	2.33

a) Generate a (line) plot for each of the variables by plotting the units (x-axis) against their respective observations (y-axis) to get a first impression of the data matrix. You may add lines to aid interpretation.
b) Compute the Pearson correlation of var.1 versus all other variables (5 results). In Microsoft (MS) Excel you can use the function 'pearson'; in the statistical software environment R you can use the function 'cor' or 'cor.test'.
c) Compute the Spearman correlation of var.1 versus all other variables (5 results) as Pearson correlation on ranked data. In MS Excel you can use the function 'rank' to rank the observations and than run the 'pearson' function. In R you can use the function 'cor' or 'cor.test' by changing the method parameter to 'spearman'.
d) Compute the Euclidean distance of var.1 versus all other variables (5 results). In R you can use the function 'dist'. In MS Excel you have to calculate it by using the following formula: $d_e(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$, i.e. calculate the square difference of var.1 and e.g. var.2 for each unit, sum over all units, and calculate the square root of this sum.
e) Compare the calculated similarity and dissimilarity coefficients for each variable. Generate scatterplot(s) of var.1 (x-axis) versus all other variables to aid interpretation.

Results and Interpretation:
a) Line plot of units against their respective observation for each variables, var.1 to var.5 (figure 12.6).

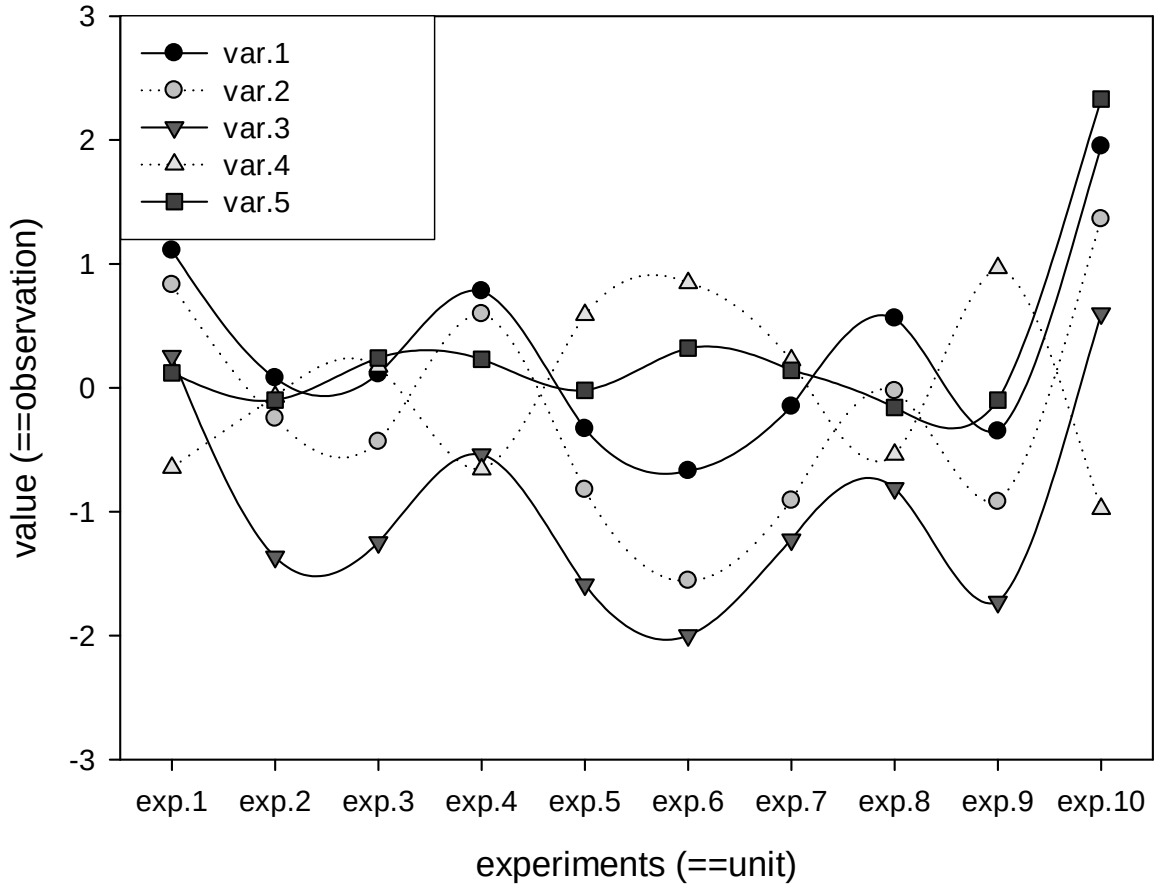

Fig. 12.6 Line plot of the 5 variables of data set table 12.3.

b-d)

Table 12.4: Illustration of the results obtained for few association measures on data set table 12.3.

variables	Pearson Correlation	Spearman Correlation	Euclidean Distance
var.1-var.1	**1**	**1**	**0.00**
var.1-var.2	**0.98**	**0.98**	**1.77**
var.1-var.3	**0.98**	**0.96**	**4.06**
var.1-var.4	**-0.93**	**-0.96**	**4.47**
var.1-var.5	**0.7**	**0.21 (0.20*)**	**1.79**

Remark: The results may differ slightly depending if you use MS Excel or R(*).

e) The results for Spearman and Pearson correlation are basically identical except for the case var.1-var.5. In this case the strong correlation for Pearson is observed due to a bivariate outlier in the data (see scatter plot (marked in brackets), observation in exp.10 for the xy-pair [var.1, var.5]).

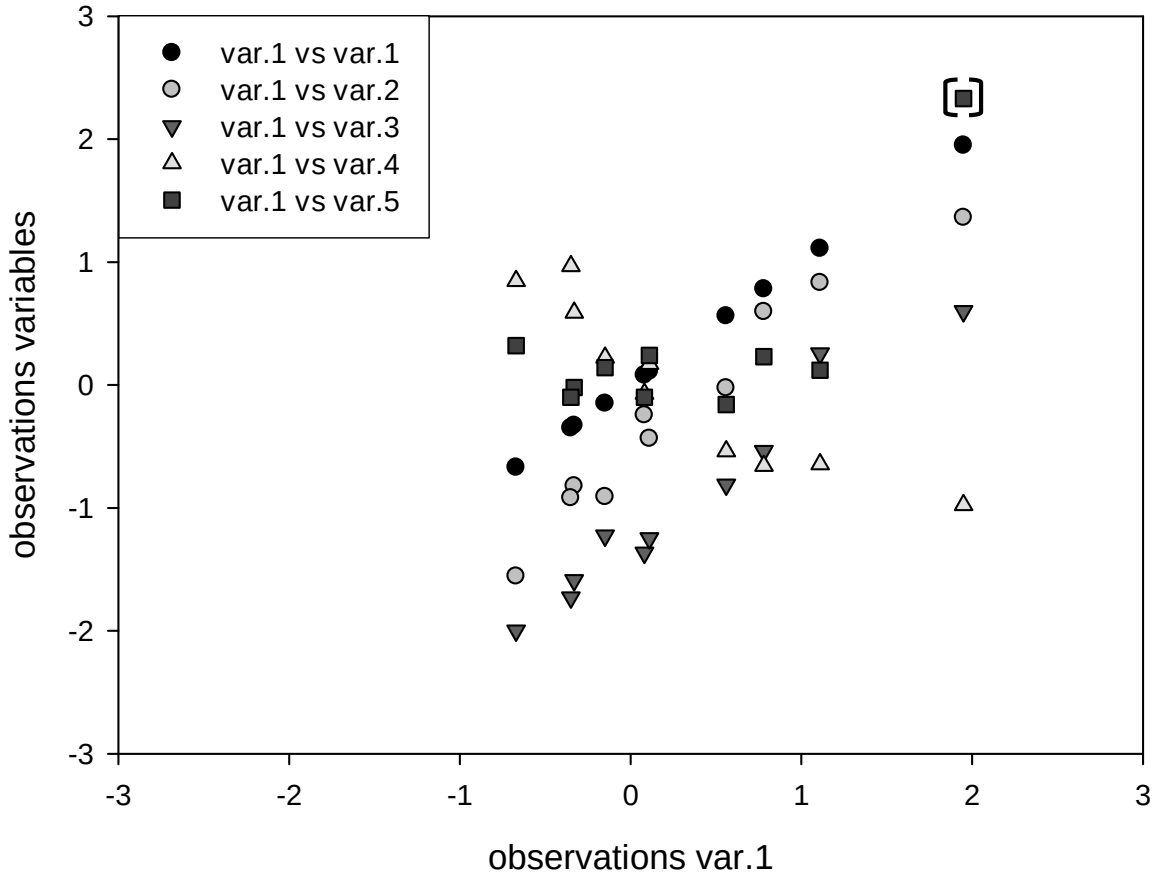

Fig. 12.7 Scatter plot of var.1 against the 5 variables of data set table 12.3.

Therefore, an outlier can drive high and statistically significant correlation which may lead to drawing an edge between a pair of nodes, while in truth the correlation is quite weak. You may test this by re-running the evaluation without the outlying point (or by omitting exp. 10 from the data matrix).

Comparison of the Pearson correlations obtained for var.1-var.2 and var.1-var.3 reveal strong difference in the Euclidean distance despite their identical correlation. The reason for this is found in the differences of the absolute values of var.2 and var.3 compared to var.1. The choice of association measure depends on what you can assume or are looking for: similarity in behavior or similarity in observed values.

Currently, no public convention exists as to which numerical approach is best applied to detect and validate correlations. Each coefficient has own advantages and disadvantages, which must be keep in mind by interpretation. The choice should strongly depend on the biological question.

2. A Pearson correlation between two variables of 0.7 is observed on the basis of 3, 10, 50 and 100 paired observations n. Compute the degree of freedom (df), the t statistic (t_S, acc. equation (7)) and the p-value. In MS Excel the p-value can be computed by the function 'tvert' and using the parameter t_S, df, and 2 for a two-sided test. In R you can use the function 'pt' (Note: A one-sided value will be returned.).

Results and Interpretation:

Table 12.5: Statistical results obtained for identical correlations derived for data sets with different size n.

pearson's r	n	df	t_S	p-value

0.7	3	1	0.9802	0.5064
0.7	10	8	2.7724	0.0242
0.7	50	48	6.7910	1.54E-08
0.7	100	98	9.7034	5.33E-16

Remark: The results may differ slightly depending on the software that you used (especially regarding the probability).
The choice of cutoff for drawing an edge between two nodes should not exclusively depend on the observed correlation. The correlation should be at least significant at p-value < 0.05 (95%) which depends on the underlying number of paired observations (n). But keep in mind, that the p-value may also be influenced by outliers (see Exercise 1). For multiple comparisons the p-value should be adjusted. The easiest way is to apply the Bonferroni Correction, i.e. the new adjusted acceptance value for 0.05 is 0.05 divided by the number of comparisons.

Additional exercise (not in book)
3. A symmetrical correlation matrix with 40 variables is given (table 12.6). The observed correlations are based on a data matrix with 50 units (experiments). The adjusted probability to accept a significant correlation is 0.00125. Compute the connectivity k and the number of variables for each observed connectivity n(k) for the following absolute correlation thresholds:

- abs(r) > 0.3 (p-value = 0.0343)
- abs(r) > 0.5 (p-value = 0.0002)
- abs(r) > 0.7 (p-value = 1.54E-08)

Plot the connectivity k (x-axis) against n(k) with logarithmic scaling of both axes. Compare the obtained distribution against each other and amongst the literature, e.g. open-access paper [Bergmann et al., 2004]. In MS Excel you can use the 'if' and 'abs' function to convert the similarity matrix into an adjacent graph (Boolean values). The connectivity (degree centrality) is the number of connection for a node without the diagonal (in our case).

Table 12.6: Correlation matrix derived from 50 true generated experiments (see also Excel file).

cor	var.1	var.2	var.3	var.4	var.5	var.6	var.7	var.8	var.9	var.10	var.11	var.12	var.13	var.14	var.15	var.16	var.17
Var.1	1.00	0.21	0.38	0.21	-0.03	-0.10	-0.33	0.45	0.08	-0.25	-0.14	0.16	-0.09	-0.07	0.11	-0.13	0.05
Var.2	0.21	1.00	0.87	0.93	-0.40	-0.04	-0.57	0.57	-0.10	-0.69	-0.56	-0.30	-0.43	-0.40	-0.22	-0.33	-0.14
Var.3	0.38	0.87	1.00	0.86	-0.26	-0.15	-0.55	0.68	-0.13	-0.61	-0.47	-0.20	-0.32	-0.38	-0.39	-0.34	-0.04
Var.4	0.21	0.93	0.86	1.00	-0.41	-0.00	-0.59	0.54	-0.03	-0.64	-0.51	-0.28	-0.41	-0.49	-0.30	-0.33	-0.03
Var.5	-0.03	-0.40	-0.26	-0.41	1.00	-0.04	0.47	-0.37	0.14	0.23	0.48	-0.40	0.32	-0.26	-0.07	0.30	0.07
Var.6	-0.10	-0.04	-0.15	-0.00	-0.04	1.00	-0.29	-0.05	-0.14	-0.49	-0.36	-0.11	-0.47	-0.11	-0.40	-0.11	-0.38
Var.7	0.33	0.57	0.55	0.59	-0.47	-0.29	1.00	-0.54	0.13	-0.58	-0.70	-0.35	-0.72	-0.31	-0.09	-0.50	-0.40
Var.8	-0.45	-0.57	-0.68	-0.54	-0.37	-0.05	-0.54	1.00	-0.05	-0.36	-0.48	-0.12	-0.27	-0.47	-0.33	-0.52	-0.00
var.9	-0.08	-0.10	-0.13	-0.03	-0.14	-0.14	-0.13	-0.05	1.00	-0.27	-0.44	-0.40	-0.43	-0.17	-0.10	-0.03	-0.30
var.10	-0.25	-0.69	-0.61	-0.64	-0.23	-0.49	-0.58	-0.36	-0.27	1.00	-0.49	-0.44	-0.59	-0.13	-0.09	-0.34	-0.18
var.11	-0.48	-0.36	-0.70	-0.44	-0.49	-1.00	-0.45	-0.06	-0.39	-0.48							

	0.14	0.56	0.47	0.51				0.48					0.27				
var.12	0.16	0.30	0.20	0.28	0.40	0.11	0.35	0.12	0.40	0.44	0.45	1.00	0.55	0.11	0.32	0.45	0.18
var.13	0.09	0.43	0.32	0.41	0.32	0.47	0.72	0.27	0.43	0.59	0.68	0.55	1.00	0.17	0.18	0.38	0.55
var.14	0.07	0.40	0.38	0.49	0.26	0.11	0.31	0.47	0.17	0.13	0.27	0.11	0.17	1.00	0.61	0.16	0.12
var.15	0.11	0.22	0.39	0.30	0.07	0.40	0.09	0.33	0.10	0.09	0.06	0.32	0.18	0.61	1.00	0.24	0.18
var.16	0.13	0.33	0.34	0.33	0.30	0.11	0.50	0.52	0.03	0.34	0.39	0.45	0.38	0.16	0.24	1.00	0.20
var.17	0.05	0.14	0.04	0.03	0.07	0.38	0.40	0.00	0.30	0.18	0.48	0.18	0.55	0.12	0.18	0.20	1.00
var.18	0.06	0.18	0.19	0.10	0.17	0.11	0.08	0.08	0.47	0.31	0.02	0.01	0.11	0.27	0.07	0.42	0.09
var.19	0.06	0.42	0.42	0.44	0.40	0.03	0.47	0.49	0.04	0.36	0.42	0.63	0.35	0.18	0.29	0.84	0.06
var.20	0.03	0.07	0.09	0.05	0.18	0.11	0.20	0.08	0.50	0.33	0.01	0.19	0.08	0.27	0.11	0.33	0.03
var.21	0.28	0.06	0.07	0.10	0.30	0.14	0.25	0.15	0.13	0.14	0.28	0.33	0.12	0.25	0.09	0.44	0.11
var.22	0.24	0.25	0.09	0.13	0.04	0.03	0.02	0.15	0.50	0.09	0.20	0.14	0.19	0.05	0.05	0.30	0.33
var.23	0.52	0.24	0.21	0.16	0.10	0.25	0.17	0.36	0.07	0.21	0.15	0.18	0.18	0.04	0.09	0.08	0.22
var.24	0.04	0.48	0.39	0.48	0.59	0.28	0.67	0.39	0.29	0.53	0.73	0.68	0.71	0.28	0.10	0.51	0.32
var.25	0.22	0.42	0.37	0.42	0.42	0.01	0.52	0.47	0.25	0.27	0.30	0.12	0.17	0.15	0.09	0.61	0.08
var.26	0.09	0.16	0.07	0.14	0.08	0.35	0.06	0.12	0.06	0.19	0.02	0.34	0.19	0.10	0.01	0.34	0.05
var.27	0.33	0.07	0.02	0.15	0.49	0.27	0.18	0.10	0.04	0.19	0.22	0.33	0.02	0.24	0.11	0.21	0.08
var.28	0.21	0.59	0.66	0.67	0.31	0.11	0.40	0.59	0.05	0.44	0.33	0.37	0.28	0.32	0.12	0.59	0.08
var.29	0.01	0.32	0.28	0.32	0.25	0.17	0.24	0.19	0.10	0.22	0.16	0.56	0.15	0.31	0.11	0.26	0.08
var.30	0.31	0.36	0.32	0.31	0.01	0.35	0.27	0.18	0.40	0.53	0.28	0.17	0.38	0.16	0.23	0.18	0.16
var.31	0.19	0.47	0.48	0.39	0.12	0.09	0.07	0.18	0.03	0.39	0.03	0.22	0.07	0.21	0.30	0.14	0.06
var.32	0.34	0.78	0.76	0.79	0.31	0.08	0.50	0.53	0.16	0.62	0.43	0.33	0.48	0.28	0.16	0.22	0.03
var.33	0.42	0.64	0.73	0.64	0.18	0.13	0.48	0.53	0.05	0.60	0.33	0.16	0.37	0.25	0.11	0.40	0.06
var.34	0.15	0.18	0.06	0.06	0.05	0.03	0.14	0.15	0.57	0.15	0.10	0.39	0.16	0.00	0.03	0.26	0.17
var.35	0.14	0.56	0.56	0.55	0.50	0.05	0.58	0.54	0.37	0.39	0.66	0.66	0.53	0.28	0.06	0.57	0.16
var.36	0.29	0.48	0.55	0.46	0.45	0.03	0.57	0.70	0.07	0.40	0.48	0.50	0.38	0.21	0.06	0.75	0.11
var.37	0.09	0.32	0.24	0.31	0.44	0.11	0.51	0.47	0.22	0.33	0.52	0.70	0.51	0.25	0.24	0.78	0.23
var.38	0.12	0.32	0.39	0.29	0.25	0.02	0.53	0.48	0.48	0.38	0.56	0.61	0.48	0.09	0.12	0.70	0.25
var.39	0.08				0.56	0.06	0.59		0.30	0.44	0.61	0.77	0.57		0.14	0.67	0.21

	-	0.46 -	0.40 -	0.49 -		-		0.47 -						0.31 -	-			
var.40	0.22	0.27	0.38	0.28	0.23	0.12	0.48	0.54	0.36	0.26	0.44	0.49	0.37	0.23	0.12	0.50	0.20	

Results and Interpretation:

The connectivity and the number of variables for each observed connectivity n(k) for the respective cutoff are listed and visualized as follows:

k \ n(k)	|r|>0.3	|r|>0.5	|r|>0.7
1		6	4
2		6	5
3		2	2
4		1	4
5		2	1
6	1	1	1
7	3	1	3
8	2	3	1
9	1	3	
10	1	4	
11	2	2	
12			
13		1	
14	1	4	
15	1	1	
16	1	1	
17	3		
18	1		
19	1		
20	1	1	
21			
22	2		
23	3		
24	6		
25	6		
26	2		
27	2		

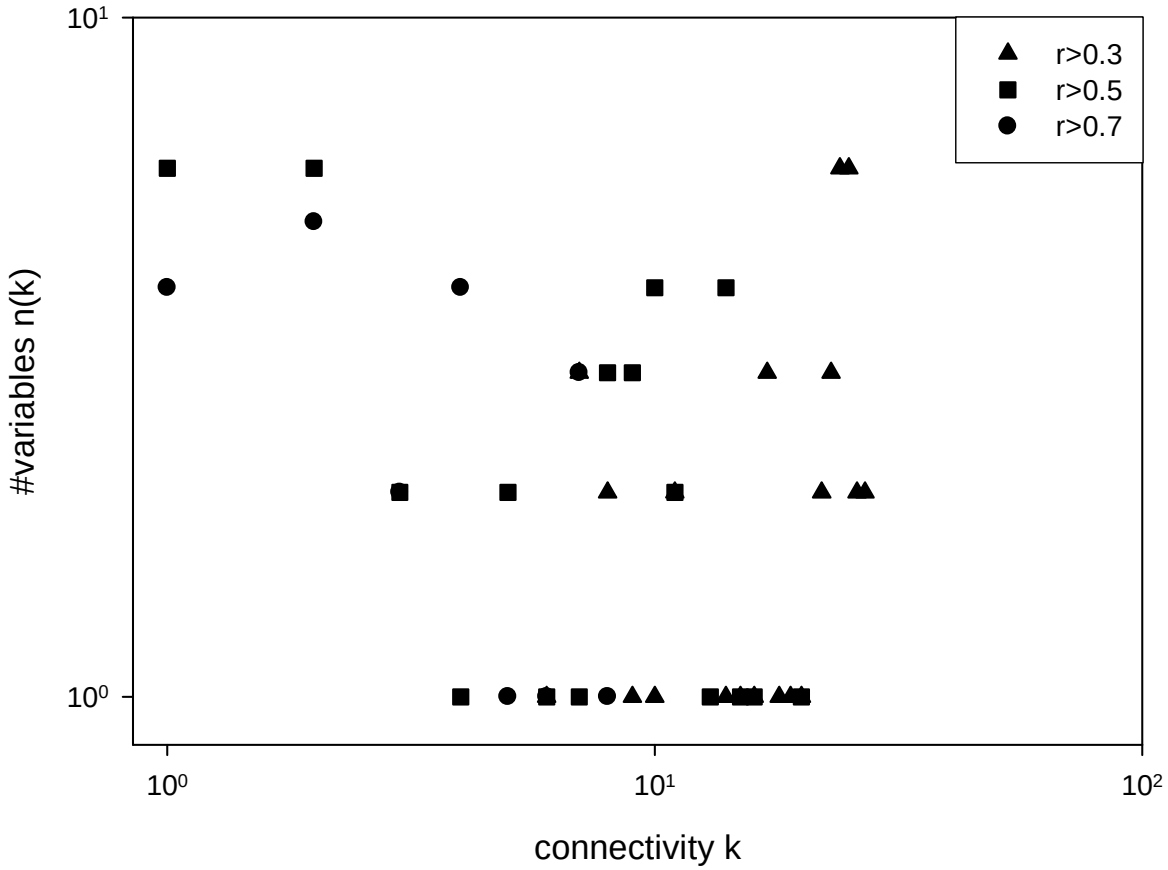

Fig. 12.8 Global property of a small correlation network based on expression values. The number of genes (i.e. variables) n(k) with connectivity k is plotted as afunction of k by using different cutoffs.

The cutoff value to convert a similarity matrix into an adjacent graph determines the resulting graph and therefore the topology.