Many objective optimization and complex network analysis

Maulana, A.

4

Community Detection in NK-Landscapes -An Empirical Study of Complexity Transitions in Interactive Networks

NK-Models (or NK-Landscapes) were introduced in Kauffman and Levin, 1987 [30] as models of how the fitness of an organism is related to gene interaction. They also gained popularity in the study of complex organizations (see Anderson, 1999 [3]) and innovation networks (see Frenken, 2000 [22]).

In the classical NK-model the parameter N describes the number of components in a network (genes, agents, or nodes in an distributed system) and each component is associated with a control variable x_i and a trait function or output value f_i. K describes the degree of interaction between these components. For each component of the system ($i \in \{1,\ldots,N\}$) , the value of f_i depends on the value of the variable x_i that is associated with it and k other variables $x_{e(i,1)},\ldots,x_{e(i,k)}$. These k variables are called the *epistatic variables*. The fitness of the NK-landscape is the sum of these values:

$$F(x_1,\ldots,x_N) = \sum_{i=1}^{N} f_i(x_i, x_{e(i,1)}, ..., x_{e(i,k)}) \tag{4.1}$$

In Equation 4.1 the values of $E = e(i,j), i \in \{1,\ldots,N\}, j \in \{1,\ldots,K\}$ are the epistatic matrix that determines which variable interacts with which other variable.

Based on the locality of the epistatic matrix two variants of the NK-landscapes

are distinguished: The epistatic variables of gene i can be *adjacent* with respect to the index i (local interactions) or their choice can be *random* (global interactions). Note, that in case of adjacent epistatic genes the indices are mapped cyclically, i.e. gene N is a direct neighbor with gene 1 (wrap around). This way every gene has two direct neighbors. If more than two epistatic genes need to be defined we collect the genes in an increasingly big radius (2-step neighbors, 3-step neighbors and so on). The notion of neighborhood stems here from the idea of physical location on a DNA, that, for the sake of simplicity is viewed as a ring.

Figure 4.1 shows a visualization of the epistasis structure that arises based on these choices.

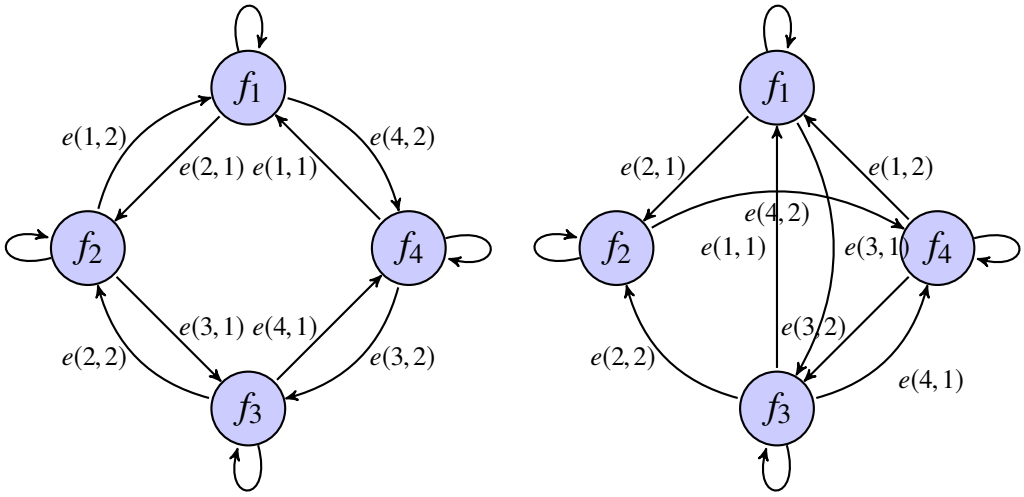

Figure 4.1 (Left hand side) Example of an NK-Landscape epistasis network for $N = 4$, $k = 2$ and adjacent epistatic genes. (Right hand side) Example of an NK-Landscape epistasis network for $N = 4$, $k = 2$ and randomly assigned epistatic genes. The arrows labeled with $e(i, j) \in \{1, \ldots, N\}$ indicate the epistatic genes that influence the gene with index $i \in \{1, \ldots, N\}$ for $j \in \{1, ..., k\}$.

We can now introduce a concrete realization of a function F, for instance by using the classical binary NK-landscape where each variable can only obtain the values 0 and 1. For this we will define 'upper case' F_i component functions that will accept a bit string the components of which correspond to the values of x (Goedel encoding of

the vector):

$$F(x_1, \ldots, x_N) = \sum_{i=1}^{N} F_i(2^0 x_i + 2^1 x_{e(i,1)} + \cdots + 2^k x_{e(i,k)}), \quad \mathbf{x} \in \{0,1\}^N \tag{4.2}$$

Each one of the functions F_i is looked up in a table of size 2^{k+1}. Each of these tables comprises 2^{k+1} random numbers that are sampled from a uniform distribution in $[0, 1]$ (see, e.g. [2]). This was done in order to make the model simple and to not introduce additional complexity in its construction [30]. Subsequent analysis revealed interesting, emergent behavior of NK-Landscapes already on this very basic level. One interesting feature of NK-landscapes is that their properties critically depend on the choice of k. Some interesting properties that depend on the choice of k are summarized in the following list (see also [2]):

- $k = 0$ (no epistasis):
  - The problem is separable.
  - There exists a unique global and local optimum[1].
- $k = 1$
  - A global optimum can be found in polynomial time.
- $k \geq 1$
  - Adjacent epistatic genes: Time complexity for finding a global optimum is in $O(N^k)$.
  - Randomly assigned epistatic genes: Finding a global optimum becomes NP complete; time complexity is in $O(2^N)$ under the assumption that $P \neq NP$.
- $k = N - 1$
  - Random function value assignment; causality is lost and finding the global optimum takes $\Omega(2^N)$ time.

An interesting research question is: What happens to the structure of the problem at the critical transitions from simple ($k = 0$), via polynomially optimizable $k = 1$

[1]Except for degenerate cases, which occur with probability of zero-measure.

or problems with adjacent epistatic genes for greater k, to complex networks $k = 2$ (global interactions) and how does this differ from complete random function value assignments at the level of $k = N - 1$.

In order to study this, we propose to study in more depth the interaction between the components of F, namely the functions F_i. A new perspective to look at this question is to view the F_i as objective functions in a many-objective optimization problem. The novel perspective taken is to view these trait functions as objective functions that seek to contribute to F with the highest possible contribution, or, that seek to obtain the best adaptive value. This yields a multi-objective optimization problem, which can be written as:

$$F_1(\mathbf{x}) \rightarrow \max, \ldots, F_N(\mathbf{x}) \rightarrow \max, \quad \mathbf{x} \in \{0,1\}^N \tag{4.3}$$

The correlation between trait functions can be determined if the input vector is viewed as a random sample. Different trait functions F_i can support each other (positive correlation), be neutral with respect to each other (zero correlation), or conflict with each other (negative correlation).

It is noted that the maximization of each single component function takes time $\Omega(2^k)$ due to the random assignment process. By introducing interaction the complexity of the optimization task grows. So far it is unknown what exactly happens at the transition from binary to ternary interactions, that is from polynomially time solvable to NP-complete problems, and we hope that correlation and community analysis will shed some new light on this.

In summary, the contributions of our work will be as follows. In order to better understand the transitions in complexity in NK-landscapes from the perspective of communities of component functions we will

1. visualize the community structure among the different F_i trait functions using state-of-the-art algorithms from community detection.

2. provide statistics on number of communities and modularity for different values of k, and

3. discuss correlation and squared correlation as a measure of connectedness in both adjacent and random NK-landscapes.

In Section 4.1 the approach will be discussed in detail, i.e. how to perform community detection among the different F_i trait functions, and how to use the squared correlation measure for global statistics. The results will be discussed in Section 4.2.

4.1 · Approach

The correlation matrix used in this chapter is similar to the statistical concept of correlation explained in chapter 2 but more specific to the problem we investigated. The graph that we consider has N nodes. Each node is associated with a component function F_i. Links between the nodes are weighted by the correlation between the two function values. Now, ρ^e_{ij} serves as an estimate of the correlation ρ_{ij} between $\mathcal{F}_i$ and $\mathcal{F}_j$. The value of the correlation ρ_{ij} ranges from perfect anti correlation (-1), via independence (0), to perfect correlation ($+1$).

In the context of optimization of the F_i functions, we can interpret this correlation as follows:

- Positively correlated trait functions can be interpreted as trait functions that support each other. This means that maximizing one function will imply that also the other function will obtain high values.
- Uncorrelated trait functions are considered to be independent of each other. They can be maximized in isolation from each other. The NK-landscape value can be maximized by independently maximizing these traits.
- Negatively correlated objective functions are in a strong conflict with each other, that is an increment of one value typically will lead to a deterioration of the other value. Intuitively, one might expect that if there are many conflicting trait functions the maximization of the NK-landscapes gets more difficult.

The next two examples of a correlation matrix of NK-Landscapes for $k = 0$ and $k = 1$ with $N = 10$ are provided with Table 4.2 and Table 4.2, respectively. Note, that in case $k = 0$ not all correlations between different trait functions are 0, although from the construction of the NK-Landscape all trait functions should be independent in the case $k = 0$. The finiteness of the sampling space makes it however unlikely that also the statistical correlation is exactly the same. For $k = 1$, clearly, some correlations have values higher than zero and higher than all correlations for the previous case. As we will see, this strong correlation will vanish again quickly for values of $k \gg 1$.

Table 4.1 Correlation coefficient matrix for N = 10 and k = 0

1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.000	0.000	1.000	-0.014	0.013	-0.014	0.012	-0.012	-0.032	0.000
0.000	0.000	-0.014	1.000	0.013	-0.014	0.012	-0.012	-0.032	0.000
0.000	0.000	0.013	0.013	1.000	-0.043	0.035	-0.035	-0.097	0.000
0.000	0.000	-0.014	-0.014	-0.043	1.000	0.058	-0.058	-0.161	0.000
0.000	0.000	0.012	0.012	0.035	0.058	1.000	-0.128	-0.358	0.000
0.000	0.000	-0.012	-0.012	-0.035	-0.058	-0.128	1.000	0.358	0.000
0.000	0.000	-0.032	-0.032	-0.097	-0.161	-0.358	0.358	1.000	0.000
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

In summary, correlation analysis is applied to derive the weights of links between the pairs of component functions. The $N \times N$ correlation matrix is interpreted as a weighted graph with weighs in $[-1, 1]$. This way, it can be analyzed using graph theoretic algorithms and in particular by community detection. Thereafter, statistics on macroscopic properties of the community graphs can be applied to find regularities that might reveal new insights in the critical transition(s) of the landscape's complexity as k grows.

4.2 · Results

Results depicted in Figure 4.4 to 4.8 visualize the concrete results of the community detection obtained and visualized with `Pajek` using Louvain and, respectively, VOS clustering. First, let us summarize results for the Louvain method. Figure 4.4 and, respectively, 4.5 show the transition of community structures for randomly assigned genes and, respectively, for adjacent epistatic genes. The value of k is chosen from 0 to

Table 4.2 Correlation coefficient matrix for N = 10 and k = 1

1.000	-0.237	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.246
-0.237	1.000	0.325	0.015	0.013	-0.011	0.010	0.011	0.030	0.000
0.000	0.325	1.000	0.455	-0.005	0.004	-0.004	-0.004	-0.011	0.000
0.000	0.015	0.455	1.000	-0.033	0.020	-0.020	-0.020	-0.058	0.000
0.000	0.013	-0.005	-0.033	1.000	-0.065	0.040	0.044	0.124	0.000
0.000	-0.011	0.004	0.020	-0.065	1.000	-0.334	-0.065	-0.184	0.000
0.000	0.010	-0.004	-0.018	0.039	-0.334	1.000	0.804	0.319	0.000
0.000	0.011	-0.004	-0.020	0.044	-0.065	0.803	1.000	0.355	0.000
0.000	0.030	-0.011	-0.057	0.124	-0.184	0.319	0.355	1.000	0.000
-0.246	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

$N - 1$.

From the visual impression, it is clear that the highest degree of separation is obtained in the case $k = 0$ for both epistatic link structures (random, adjacent).

The nodes that belong to the same community are indicated by nodes which share the same color. The number of communities reaches its minimum for $k = 2$. A confirmation of this can be obtained when plotting the number of communities over different values of k, which is done in Figure 4.2 for randomly assigned epistatic genes and in Figure 4.3. For adjacent ones, the value of $k = 2$ is a clear optimum in both cases.

This is surprising because it might be expected that for $k = N-1$ all nodes merge to one big community. This is not the case and – in first approximation – might be explained by the fact that nodes can also negatively influence each other (conflicting nodes). Note, that the number of communities grows at a slower rate for adjacent epistatic genes. This coincides with the slower increase of computational complexity [58].

Analogous findings have been made with the VOS clustering method for community detection. Figure 4.7 and 4.8 show the community structures, whereas Figure 4.9 and 4.10 show the results. The correspondence between the two different approaches for community detection underpins that the findings are not an artifact of the method. More dissonance between Louvain and VOS methods is found for the number of communities for high levels of k. However, in all methods the general trend can be observed that the number of communities first decreases and then grows again.

A conjecture we obtained from the pictures is that the correlations or anti-correlations are first very strong and then weaken again. This can be measured by the squared correlation. It is an indication of how much the results of two nodes depend on each other (either positively or negatively). Values close to zero indicate independence of the results at two different nodes. The average squared correlation between nodes in the network is shown in Figure 4.6 – both for adjacent and randomly assigned epistatic genes.

Clearly, both landscapes have a peak at low values of k. It is striking that the peak for the NK-landscape with low values of k has a sharp decay in average squared correlation, whereas the decay of the adjacent case is gradual. Again this coincides with the finding that for randomly assigned epistatic genes a sharp transition in computational complexity appears whereas the transition is gradual for the case of adjacent epistatic genes.

Viewing the plot one might even speculate that the observed phenomena are a *sawtooth transition*. This is found in other complex systems at the edge of chaos and is conjectured to be a universal law for macroscopic observations at the transition from systems with complex, but still predictable behavior, to chaotic and unpredictable systems (see for instance Adriaans [1]). Further analysis of larger models and the theoretical analysis of analogies between the models will be required to either confirm or reject this interesting hypothesis.

In the plots of Figure 4.2 and 4.3 (Louvain) and Figure 4.9 and 4.10 (VOS) we also show the observed modularity of community components. Here we shift a bit more emphasis on the results of the Louvain method as it explicitly seeks to find communities based on modularity. However, the graphical results are less clear for this and to find a trend we also put the tables with the numerical results in Table 4.3 (Random) and Table 4.4 (Adjacent). From these numerical results, it can be obtained that the modularity of the communities first decreases slightly and then goes up again. Clearly, the highest

average modularity is achieved for $k = 0$ in which case nodes are relatively isolated. Again, the peak is pronounced stronger for randomly assigned epistatic genes.

Table 4.3 Comparison of clustering algorithms applied in community detection, Louvain clustering Algorithm compare to VOS clustering algorithm, both with randomly assigned epistatic genes (NC is a number of community and Q is the Value of modularity for community detection)

Louvain Clustering			VOS Clustering		
k	NC	Q	k	NC	Q
0	9	0.842457	0	8	0.8645009
1	6	0.459726	1	5	0.6995902
2	5	0.613504	2	4	0.5636842
3	7	0.630355	3	7	0.6158102
4	7	0.606203	4	6	0.6416844
5	7	0.741914	5	7	0.6716613
6	10	0.701109	6	8	0.6608056
7	10	0.718037	7	10	0.7180642
8	10	0.717084	8	10	0.7285837
9	9	0.757455	9	10	0.7050733

4.3 • Summary

This chapter looked at the graph derived from the correlation structure among the component functions of an NK-landscape model that were treated as objectives of a many objective optimization problem. The results show that the community structure that is detected for this 'correlation graph' does not correspond with the community structure of the epistatic link network which has many components for small values of k and only one big component for $k = N - 1$ (every gene is linked to every other gene). Instead, the correlation network has the lowest number of components

Table 4.4 Comparison clustering algorithm applied in community detection, Louvain clustering Algorithm compare to VOS clustering algorithm with adjacent epistatic genes (NC is a number of community and Q is the Value of modularity for community detection)

Louvain Clustering			VOS Clustering		
k	NC	Q	k	NC	Q
0	9	0.842457	0	8	0.8645009
1	6	0.677228	1	5	0.7196617
2	6	0.714034	2	4	0.6285909
3	6	0.713723	3	5	0.6452446
4	8	0.663882	4	8	0.6670689
5	7	0.656833	5	7	0.6213816
6	9	0.679317	6	7	0.6804879
7	8	0.691417	7	10	0.6805063
8	9	0.680426	8	9	0.7017337
9	10	0.715158	9	7	0.6856205

for $k = 2$. For values lower and higher the number of communities clearly grows. As the critical transition from polynomial time, solvable maximization problems to NP-complete maximization problems appears at the transition from $k = 1$ to $k = 2$ (for random networks) we suspect that these findings might be not coincidental. We show also that the average squared correlation reaches a sharp peak near this value of k. This peak is less pronounced for adjacent epistatic genes which do not undergo a critical transition but a gradual transition in terms of complexity. So far we have only studied the case $N = 10$ and studies on larger networks are required in the future to improve the generality of the findings. A problem that needs to be solved for such studies is how to tame the 'explosion' in the size of the random number tables needed to generate the NK-landscapes. A useful proposal has been made by Altenberg [2], who suggested to re-generate the random numbers on-the-fly when needed and provided a function that can be used for this.

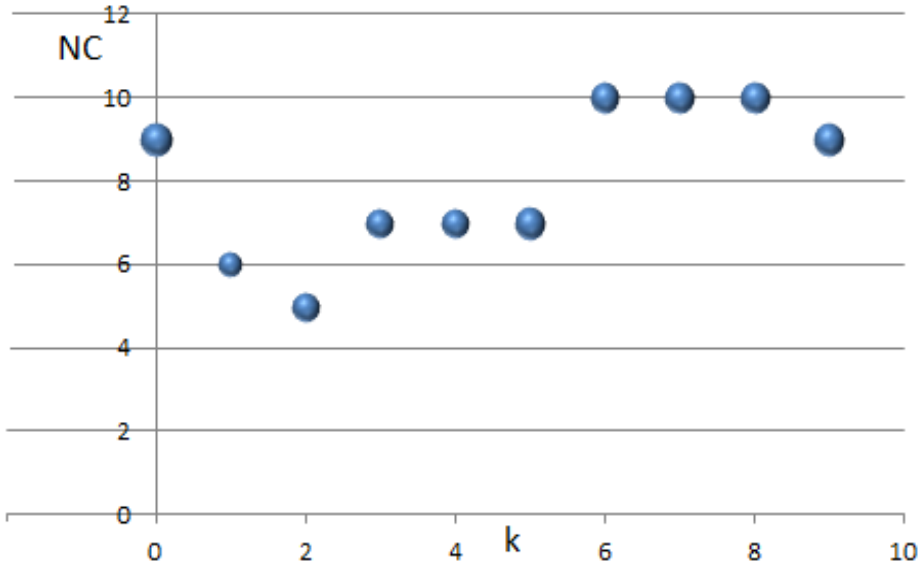

Figure 4.2 Community detection by Louvain clustering algorithm based on randomly assigned epistatic genes.

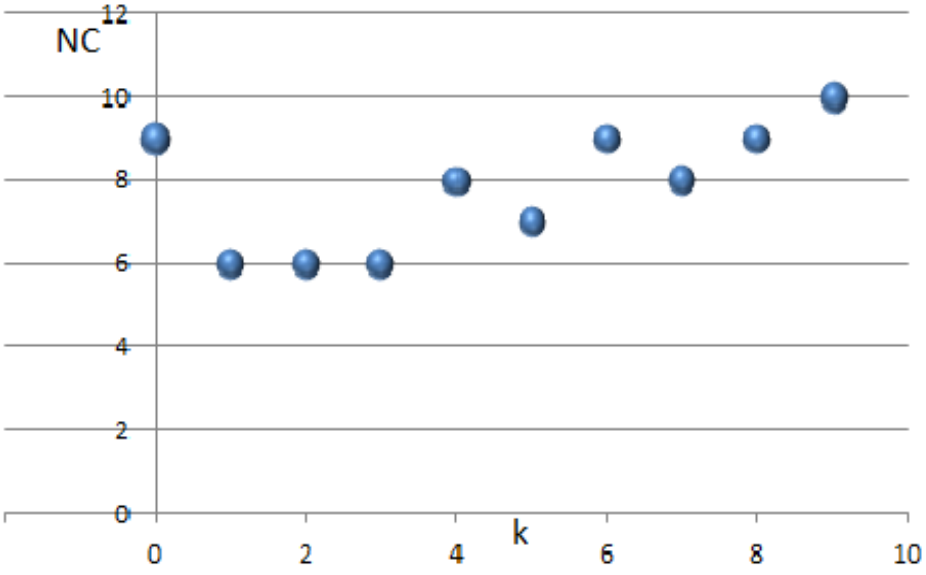

Figure 4.3 Community detection by Louvain clustering algorithm based on adjacent epistatic genes. NC denotes the number of communities found.

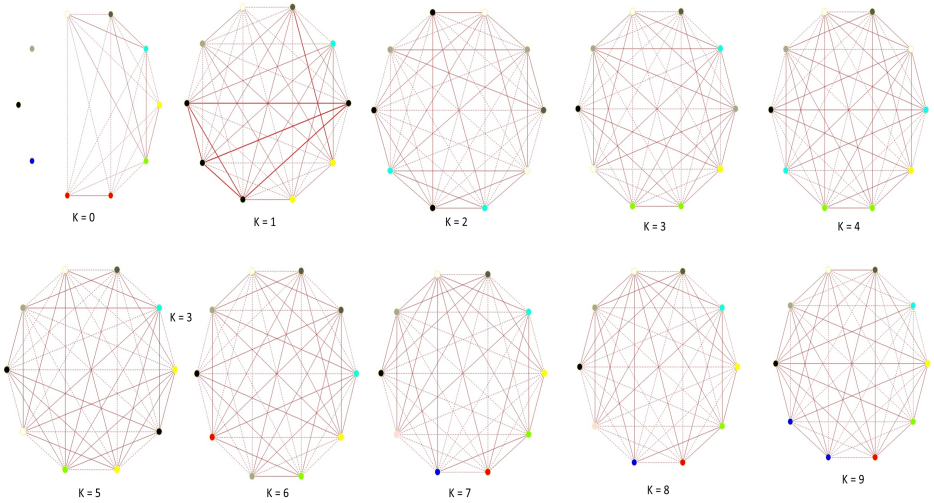

Figure 4.4 Community detection by Louvain clustering algorithm based on randomly assigned epistatic genes. NC denotes the number of communities found.

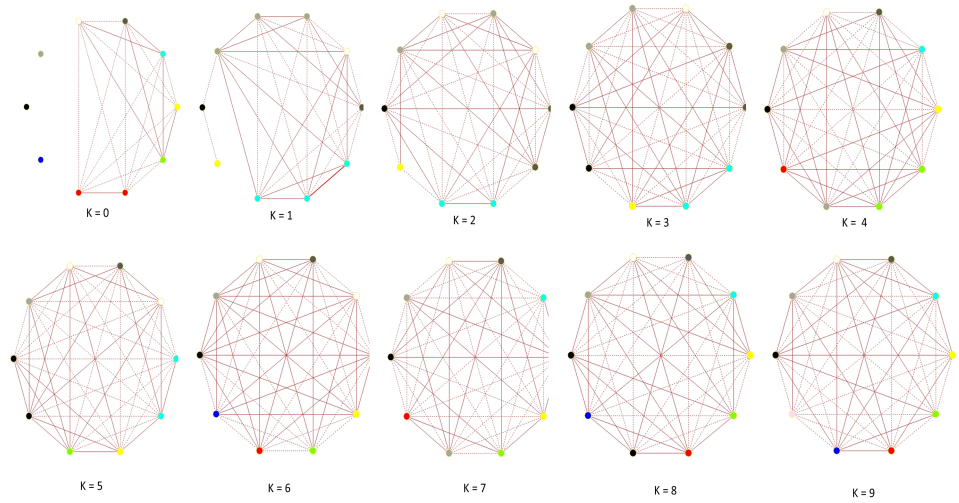

Figure 4.5 Community detection by Louvain clustering algorithm based on neighborhood selection with adjacent epistatic genes.

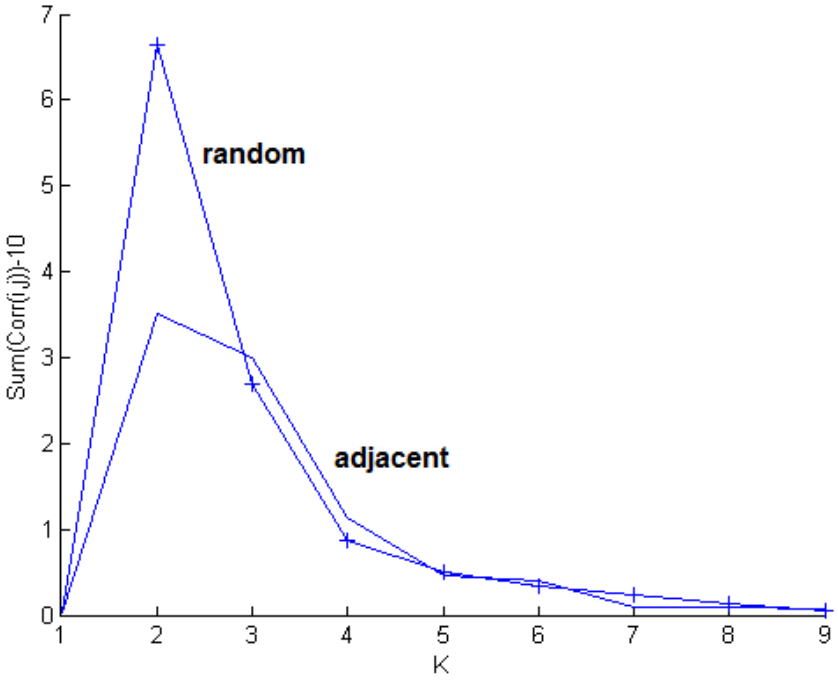

Figure 4.6 Community detection by Louvain clustering algorithm based on randomly assigned epistatic genes.

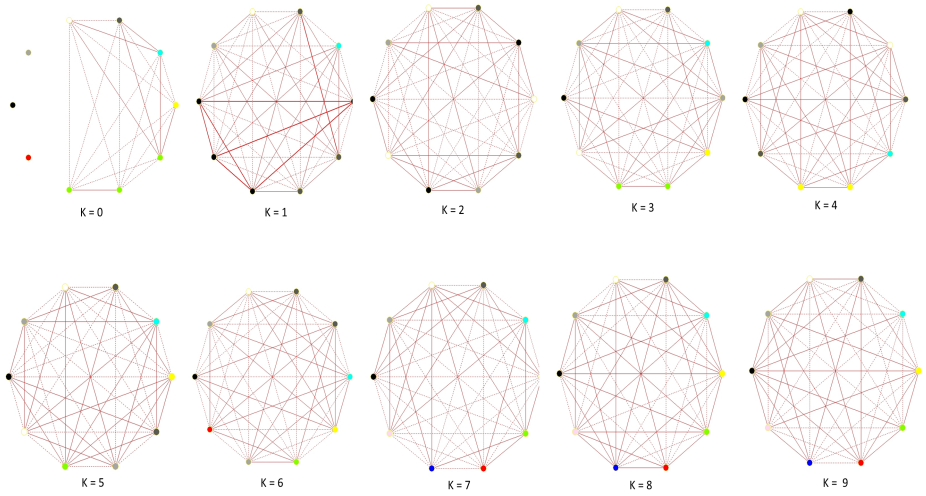

Figure 4.7 Community detection by VOS clustering algorithm based on neighborhood selection with randomly assigned epistatic genes.

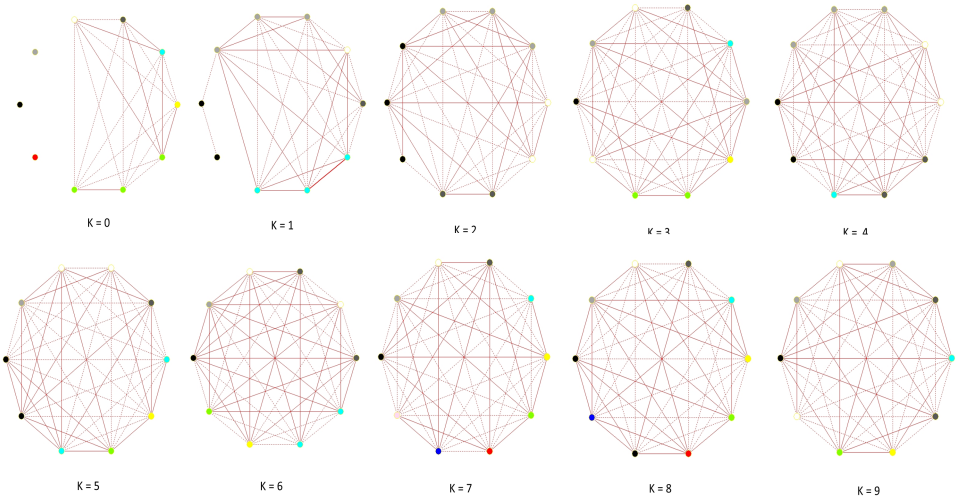

Figure 4.8 Community detection by VOS clustering algorithm based on neighborhood selection with adjacent epistatic genes.

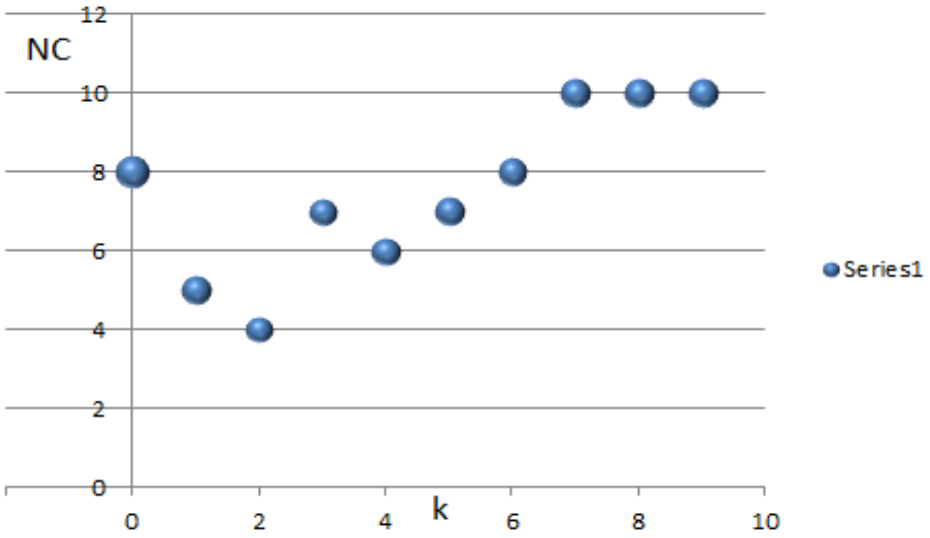

Figure 4.9 Community detection by VOS clustering algorithm based on randomly assigned epistatic genes.

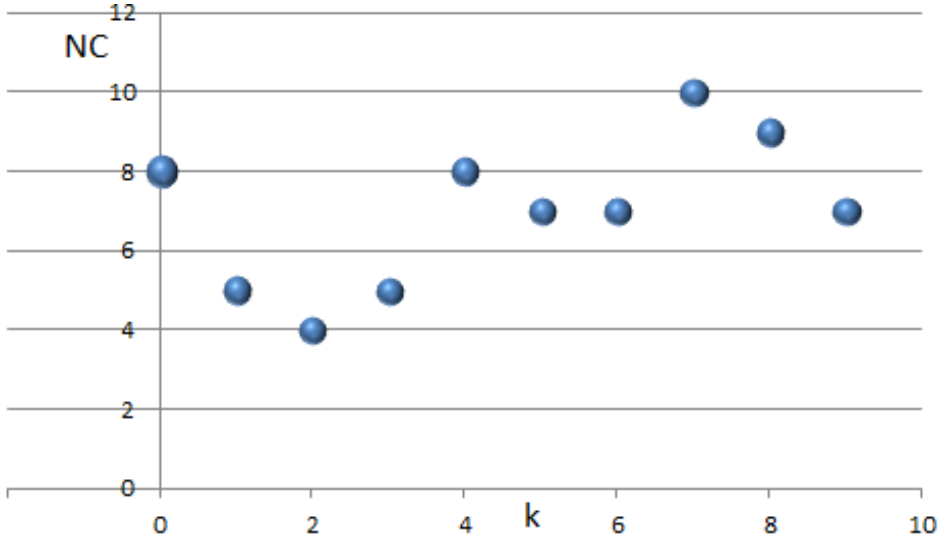

Figure 4.10 Community detection by VOS clustering algorithm based on adjacent epistatic genes.