Appendix B

Example of correcting individuals

This appendix gives an example of how an individual is generated using either the LTM and ATM methods as explained in Section 4.5.1. Just as an example, we will assume that we have a case where the model and data graphs contain respectively 8 and 10 vertices ($|V_M| = 8$ and $|V_D| = 10$). Therefore, the length of an individual will be of 10 variables.

Both LTM and ATM are methods that act directly on the simulation step of EDAs. Therefore, we need to perform first the learning step following the EDA that we have chosen. The learning step will return a Bayesian network (in this case with a size of 10 nodes) as well as the estimation of the distribution $p_l(\boldsymbol{x})$ obtained from the N selected individuals, being the latter in the form of the different conditional probabilities θ_{ijk} as defined in Equation 4.3.

We will assume that the learned Bayesian network structure is the one shown in Figure B.1. As explained before, the Bayesian network shows interdependencies between the variables (e.g. in this case this structure is showing that the value taken by variable 1 is dependent on the value of variables 2 and 9, that is, the matching assigned to vertex 1 of the data graph G_D is dependent on the matching assigned to vertices 2 and 9 of the same graph G_D, while the latter vertices can be matched to any vertex of G_M independently of the rest of matches). Following this Bayesian network, it is important to have a look at the different combination of values of the parent-variables: in the case of nodes 2, 5 and 9, they do not have parents, so they are considered as independent. Nodes 3, 4, 6, 7 and 10 have a single parent, and therefore the possible combination of values for the parents is $|V_M|$=8 (i.e. the number of values that the only parent can take). Finally, nodes 1 and 8 have two parents each, and therefore the number of possible combinations of values of the two parents is $|V_M|^2 = 64$. Having all this into account, the probabilities that we will have to compute are just the following: θ_{2-k}, θ_{9-k}, $\theta_{31k} \ldots \theta_{38k}$, $\theta_{11k} \ldots \theta_{1(64)k}$, $\theta_{41k} \ldots \theta_{48k}$, $\theta_{61k} \ldots \theta_{68k}$, $\theta_{71k} \ldots \theta_{78k}$, $\theta_{(10)1k} \ldots \theta_{(10)8k}$, θ_{5-k}, and $\theta_{81k} \ldots \theta_{8(64)k}$, where $k = 1 \ldots |V_M|$ in all the cases.

Just as an example for our purposes, we will assume that the value of these probabilities is uniform (this is not normally the case, we just do it for simplicity).

At this stage, we will start the simulation step in order to create the new R individuals of the next generation. Each individual $\boldsymbol{x}= (x_1, \ldots, x_{|V_D|})$ has to be generated by instantiating each of the variables one after another. For this we will use the PLS method in which an ancestral ordering $\boldsymbol{\pi}$ of the nodes in the Bayesian network is followed as explained in Section 4.2.2. An ancestral ordering is any ordering in which any variable is placed after all its parent variables on the Bayesian network. A possible ancestral ordering for the Bayesian network in Figure B.1 is $(2, 9, 3, 1, 6, 4, 7, 10, 5, 8)$, but others such as $(2, 9, 5, 3, 4, 6, 1, 7, 10, 8)$ or $(9, 1, 7, 10, 5, 8, 2, 3, 4, 6)$ could also be considered. Any of these could be used, but we will

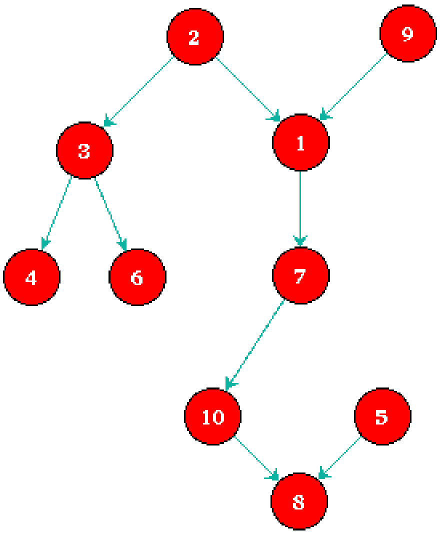

Figure B.1: Example of a Bayesian network structure.

select the first one: $\boldsymbol{\pi} = (2, 9, 3, 1, 6, 4, 7, 10, 5, 8)$.

Once the ancestral ordering has been found, we will start generating individuals. We will instantiate the variables or each individual following the ancestral ordering, $\boldsymbol{\pi}$. We will proceed similarly at the beginning either for LTM and ATM, where initially $VNO(V_M)^1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $vns^1 = |V_D| = 10$.

B.1 Simulation with LTM

Following $\boldsymbol{\pi}$, we will start with variable X_2 ($\pi(1) = 2$). Variable X_2 is independent from the rest, and the probability to take any of its possible values is the same ($\forall k = 1 \ldots |V_M|, \theta_{2-k} = \frac{1}{|V_M|}$). As the condition $|VNO(V_M)^1| = vns^1$ is not satisfied no modifications are to be done on the probabilities, so we will select a value at random and we will assign it to variable X_2. Let us imagine that this value is 1. Variable X_2 in the individual is set with this value, which in other words means that in the solution represented by this individual we are matching vertex 2 of G_D with vertex 1 of G_M.

The next variable to instantiate is X_9 ($\pi(2) = 9$) which is also independent from the rest, and its situation is the same ($\forall k = 1 \ldots |V_M|, \theta_{9-k} = \frac{1}{|V_M|}$). This time we have that $VNO(V_M)^2 = \{2, 3, 4, 5, 6, 7, 8\}$ and $vns^2 = 9$, and therefore $|VNO(V_M)^2| = vns^2$ is not satisfied. Therefore we will select a value at random. Let us again imagine that this value is 1, then we will assign the value 1 to variable X_9.

So far, after finishing this second step we have the following individual:

	1							1	

The following variable to work with is X_3 ($\pi(3) = 3$). Following the Bayesian network, this variable is dependent of variable X_2, which has already been instantiated. We said before that its probabilities are equal, and therefore we have that $\forall j, k = 1 \ldots |V_M|, \theta_{3jk} = \frac{1}{|V_M|}$. As $VNO(V_M)^3 = \{1, 3, 4, 5, 6, 7, 8, 10\}$ and $vns^3 = 8$, we have again that $|VNO(V_M)^3| \neq vns^3$, following its distribution we select a value at random. Let us now imagine that this value is 4, thus so we assign the value 4 to variable X_3.

In a similar way, we will now consider next variable X_1 $(\pi(4) = 1)$, which is dependent on variables X_2 and X_9. As it depends in two parent-variables instead on in one as before, in this case $\forall k = 1 \dots |V_M| \;\; j = 1 \dots |V_M|^2, \theta_{12k} = \frac{1}{|V_M|}$ and $\theta_{1jk} = \frac{1}{|V_M|}$. We also have that $VNO(V_M)^4 = \{1, 2, 3, 5, 6, 7, 8\}$ and $vns^4 = 7$, $|VNO(V_M)^4| \neq vns^4$. Therefore following the distribution of this value, we select a value at random, and let us imagine that we obtain one more time the value 1. So far we have the following individual:

1	1	4						1	

Next, the variable X_4 is treated $(\pi(3) = 4)$. Variable X_4 is dependent only on variable X_3, which has already been instantiated. In this case we have a single parent, and therefore $\forall j, k = 1 \dots |V_M|, \theta_{6jk} = \frac{1}{|V_M|}$. If we were in an ordinary PLS simulation approach, a value would have been chosen at random. However, this time we have that $VNO(V_M)^5 = \{2, 3, 5, 6, 7, 8\}$ and $vns^5 = 6$, that is, the condition $|VNO(V_M)^5| = vns^5$ is satisfied. As a result, following the LTM approach, we have to modify the θ_{6jk} probabilities before instantiation so that values already appeared in previous steps do not appear again. We do this because the number of variables to instantiate equals the number of values still not appeared in the individual. Following Equation 4.42 we modify the probabilities as follows:

$$\theta_{6j1} = 0, \theta_{6j4} = 0, \theta_{6jk} = \frac{1}{|V_M| - 2} \qquad \forall j = 1 \dots |V_M|, \quad k = 2, 3, 5, 6, 7, 8, 9, 10 \tag{B.1}$$

In other words, we set the probabilities for the values already appeared to 0, avoiding them to appear for this variable, and we normalize the rest of probabilities. Doing it so, we make sure that the next value that will be instantiated will not be neither 1 nor 4.

As with this last case, in the successive variables to simulate, the condition $|VNO(V_M)^m| = vns^m \;\; m = 6, 7, 8, 9, 10$ will be satisfied, and therefore for each of these variable to instantiate a value not yet appeared will be assigned. Therefore LTM will ensure that all the vertices in G_M will have at least a vertex from G_D to which are matched.

Note that the procedure followed with LTM is basically the same as PLS until the condition $|VNO(V_M)^m| = vns^m$ is satisfied. If random values would be different, the latter condition was never satisfied, and the LTM procedure will behave as an ordinary PLS approach.

B.2 Simulation with ATM

ATM is somehow more complex than LTM in the sense that all the θ_{ijk} probabilities are manipulated even before the condition $|VNO(V_M)^m| = vns^m$ is satisfied.

In ATM the probabilities change in relation to a value $K = \left\lceil \frac{N - vns^m}{vns^m - |VNO(V_M)^m|} \right\rceil$. This will be used for adapting the probabilities when the condition $|VNO(V_M)^m| = vns^m$ is satisfied. The finality is to give more probability to values not appeared yet and to lower the rest.

Following our example, we will assume that the value for N is 1000 and that the ancestral ordering of choice is the same as in the LTM example. Following it, variable X_2 will be treated first $(\pi(1) = 2)$. As the condition $|VNO(V_M)^1| = vns^1$ is not satisfied, following the ATM approach we compute the values K and P_{Indiv}:

$$K = \left\lceil \frac{N - vns^1}{vns^1 - |VNO(V_M)^1|} \right\rceil = \left\lceil \frac{1000 - 10}{10 - 8} \right\rceil = 495$$

$$P^1_{Indiv} = \sum_{k \ | \ u^k_M \in V_M \setminus VNO(V_M)^1} \theta_{2-k} = 0.$$

As we have at the beginning that all vertices of G_M are in $VNO(V_M)^1$, then all the probabilities will be changed following Equation 4.43 by multiplying then by the following factor:

$$\frac{K - P^m_{Indiv}}{K \cdot \left(1 - P^m_{Indiv}\right)} = \frac{495 - 0}{495 \cdot 1}.$$

This means that all the probabilities will not be changed. As before a value will be obtained for variable X_2 at random following its distribution ($\forall k = 1 \ldots |V_M|, \theta_{2-k} = \frac{1}{|V_M|}$). Let us imagine that this value is 1 as in the LTM example .

The second variable to treat is X_9 ($\pi(2) = 9$). We know that $|VNO(V_M)^2| = vns^2$ is not satisfied, so following the ATM approach we have that:

$$K = \left\lceil \frac{N - vns^2}{vns^2 - |VNO(V_M)^2|} \right\rceil = \left\lceil \frac{1000 - 9}{9 - 7} \right\rceil = 496$$

$$P^1_{Indiv} = \sum_{k \ | \ u^k_M \in V_M \setminus VNO(V_M)^1} \theta_{2-k} = \frac{1}{|V_M|} = \frac{1}{8} = 0.125.$$

Following Equation 4.43 we will now have a slight change on the probabilities:

$$\theta^*_{9-k} = \begin{cases} \theta_{9-k} \cdot \frac{496 - 0.125}{496 \cdot (1 - 0.125 = \theta_{9-k} \cdot 1.14)} & \text{if } k = 2, 3, 4, 5, 6, 7, 8 \\ \frac{\theta_{9-k}}{496} & \text{if } k = 1 \ . \end{cases}$$

This example shows that as $|VNO(V_M)^m|$ and vns^m are more similar the effect on the probabilities will be stronger. This change in the probabilities is applied to all the variables to instantiate until the condition $|VNO(V_M)^m| = vns^m$ is satisfied, and then ATM behaves like LTM. The effect on this method is that learned probabilities are manipulated even more than with LTM, and therefore values not yet appeared are more possible to appear.