Estimating Min-Entropy For Large Output Spaces

Darryl Buller, Aaron Kaufer
Information Assurance Directorate
National Security Agency

Overview

- Background
- Our goal
- Using Bayesian Networks
- Optimizing with a Genetic Algorithm
- Computing Min-Entropy
- Examples

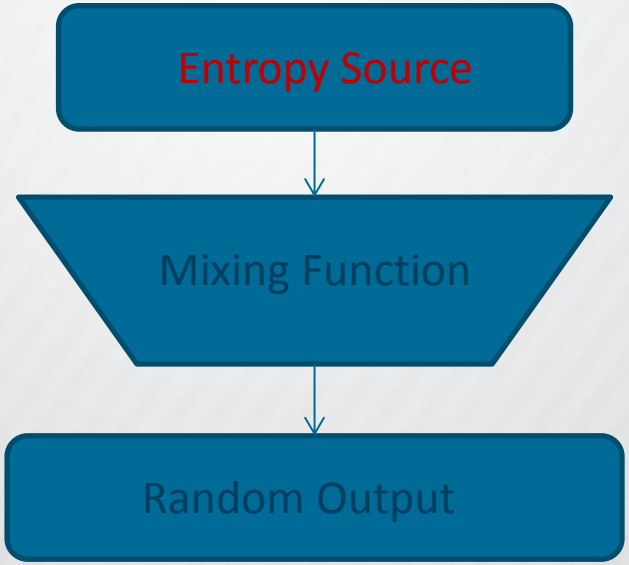
Background

Background

- Source of randomness can be quantified by its entropy
- **Min-entropy** measurement is useful for cryptographic applications
- $H_\infty = -log_2[Max\{\Pr[S = s_i]\}]$
- Corresponds to cost of optimal guessing attack
- Other measures of entropy can be misleading for these applications

Background

When data from entropy source is processed by a mixing function, we focus our analysis on the raw entropy source data



Background

- Suppose we have sample data from an entropy source
- We wish to find a statistical model and estimate the source's min-entropy
- Sample data is a sequence $\{s_1, s_2, \ldots, s_L\}$, each s_i an n-bit value sampled from an output space X

Background

- SP 800-90B has techniques for typical cases that satisfy the following two assumptions:
  - Output space *X* is reasonably small
  - Sample size *L* is large enough to detect non-IID properties (if they exist)

Background

- What if output space is very large; e.g., each s_i is dozens or hundreds of bits?
- Example: n = 50 bits, where 15^{th} bit tends to match 43^{rd} bit, or 17^{th} bit is influenced by 3^{rd}, 8^{th}, and 31^{st} bits, etc…
- Feasible sample sizes are far too small for us to fully understand the source and search for non-IID properties

Our Goal

Our Goal

Given *n* bit positions having an unknown joint distribution on 2^n possible values:

1. Compactly represent the essence of the joint distribution
2. Identify dependencies among bit positions
3. Estimate probability of most likely *n*-bit value; this lets us estimate min-entropy

Bayesian Networks

Bayesian Networks

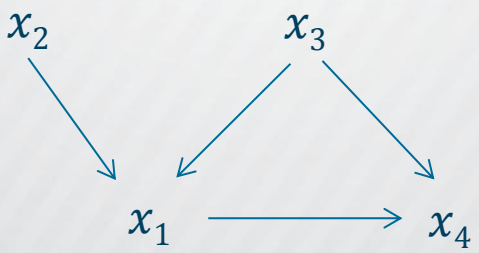
- Definition: Directed acyclic graph (DAG) whose nodes are random variables and edges indicate dependence
- Variables can depend on multiple other variables (in our case, each bit is a variable)

Bayesian Networks

Example:

- Suppose *X* consists of 4-bit outputs
- A possible BN would be:

$$\Pr[x_1, x_2, x_3, x_4] = \Pr[x_2]\Pr[x_3]\Pr[x_1|x_2, x_3]\Pr[x_4|x_1, x_3]$$



Bayesian Networks

- Given sample data, we want to find a BN that best explains the sample data
- Use resulting BN to estimate min-entropy
- *But how do we find the best BN given our data?*

Genetic Algorithms

Genetic Algorithms

- Optimization technique inspired by biology
- Represent a candidate solution as a “genome” (BN in our case)
- Maintain sequence of populations of candidate solutions
- Define fitness function that measures the quality of a particular genome

Genetic Algorithms

- The process:
  1. Randomly generate initial population of candidate solutions
  2. Repeatedly create new generation based on previous generation
- The goal is to eventually find the best-scoring candidate solution
- *How does this work?*

Genetic Algorithms

- In biology, crossover and mutation result in changes that affect fitness
- Increased fitness is rewarded by selection – population increasingly resembles optimal solution
- Decreased fitness is penalized – candidates are less likely to influence subsequent generations

Genetic Algorithms

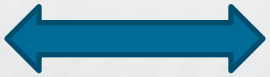
- Our implementation …

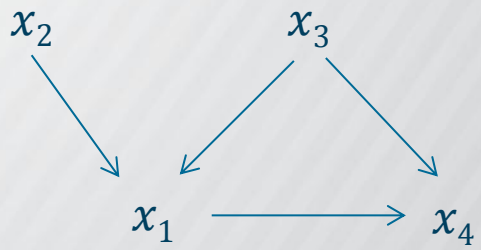
Genetic Algorithms

Genome: Encodes the details of a specific candidate solution

- Each candidate is a binary *nxn* adjacency matrix
- $A(i,j) = 1$ iff bit j is statistically dependent on bit i

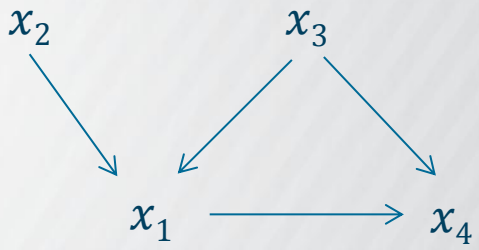
$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$





Genetic Algorithms

- Build conditional probability tables from the sample data as specified by the adjacency matrix
- For this example, we need
  - 1x2 table for $\Pr[x_2]$
  - 1x2 table for $\Pr[x_3]$
  - 4x2 table for $\Pr[x_1|x_2, x_3]$
  - 4x2 table for $\Pr[x_4|x_1, x_3]$

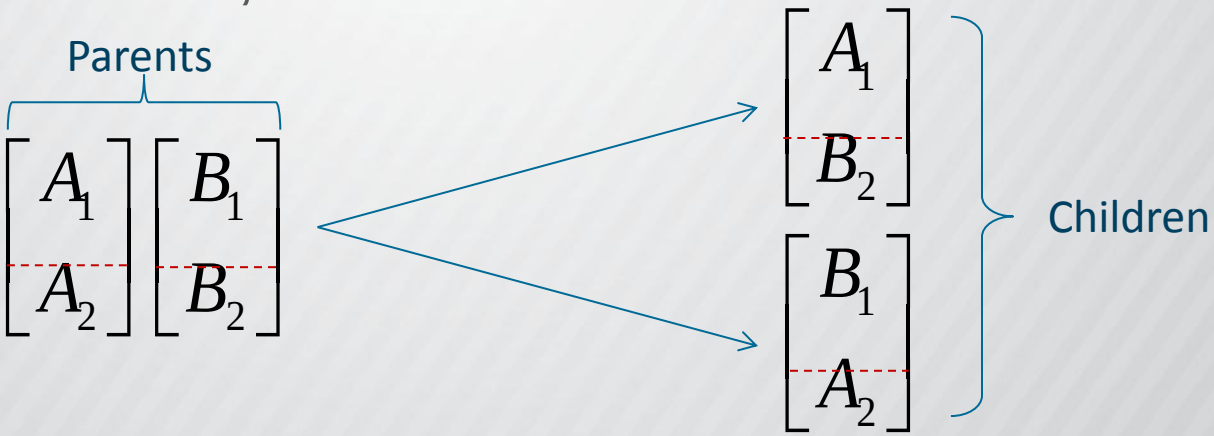


$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Genetic Algorithms

Crossover: produces two offspring by combining features of two parents

- Randomly pick a crossover point
- Join top part of one adjacency matrix and bottom part of the other, and vice-versa



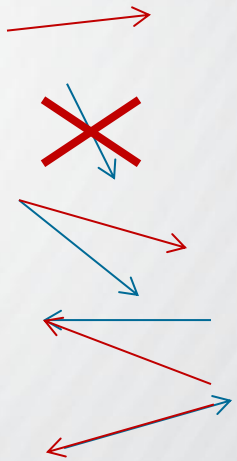
Genetic Algorithms

- Note that crossover often results in an invalid BN due to cycles
- Need a “de-cycling” step – children still contain characteristics of both parents

Genetic Algorithms

Mutation: A random change in a candidate's adjacency matrix

1. Add an edge
2. Remove an edge
3. Move an edge destination
4. Move an edge origin
5. Reverse an edge



Genetic Algorithms

Selection: Rewards high-fitness candidates by giving them a higher chance of selection to influence next generation:

1. Elitist selection: Directly copy most fit candidate to next generation
2. Fill remainder of next generation using rank selection to choose pairs of parents for crossover

Genetic Algorithms

- **Fitness function**: allows comparison of candidate solutions
- We use the Bayes Information Criterion (BIC)
- BIC rewards larger likelihood and simpler models; a smaller BIC is better (fitness-wise)

$$BIC = \mathbf{k} \ln \mathbf{N} - 2 \ln \mathbf{L}$$

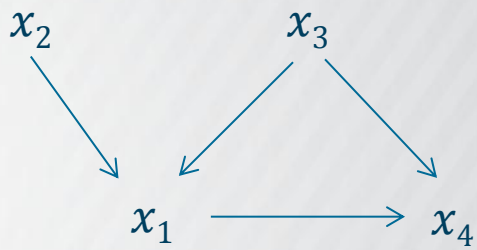
k: # of free parameters

N: # of sample outputs

L: likelihood of observed samples given the BN

Genetic Algorithms

- For the following BN:
  - 1x2 table for $\Pr[x_2]$
  - 1x2 table for $\Pr[x_3]$
  - 4x2 table for $\Pr[x_1|x_2, x_3]$
  - 4x2 table for $\Pr[x_4|x_1, x_3]$
- $\mathbf{k} = 1 + 1 + 4 + 4 = 10$



$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Min-Entropy

Min-Entropy

- Use *Max-Product Variable Elimination* algorithm to find the MAP of a BN
- Generalization of *Viterbi* algorithm

Examples

Example 1

- 32-bit blocks; sample size 15,000
- Bits 4-8, 11-15, 18-22, 25-29 follow biased joint distribution on 5-bit values
- All other bits unbiased and independent
- Actual min-entropy is 17.2877…

GENERATION #1 / 5000

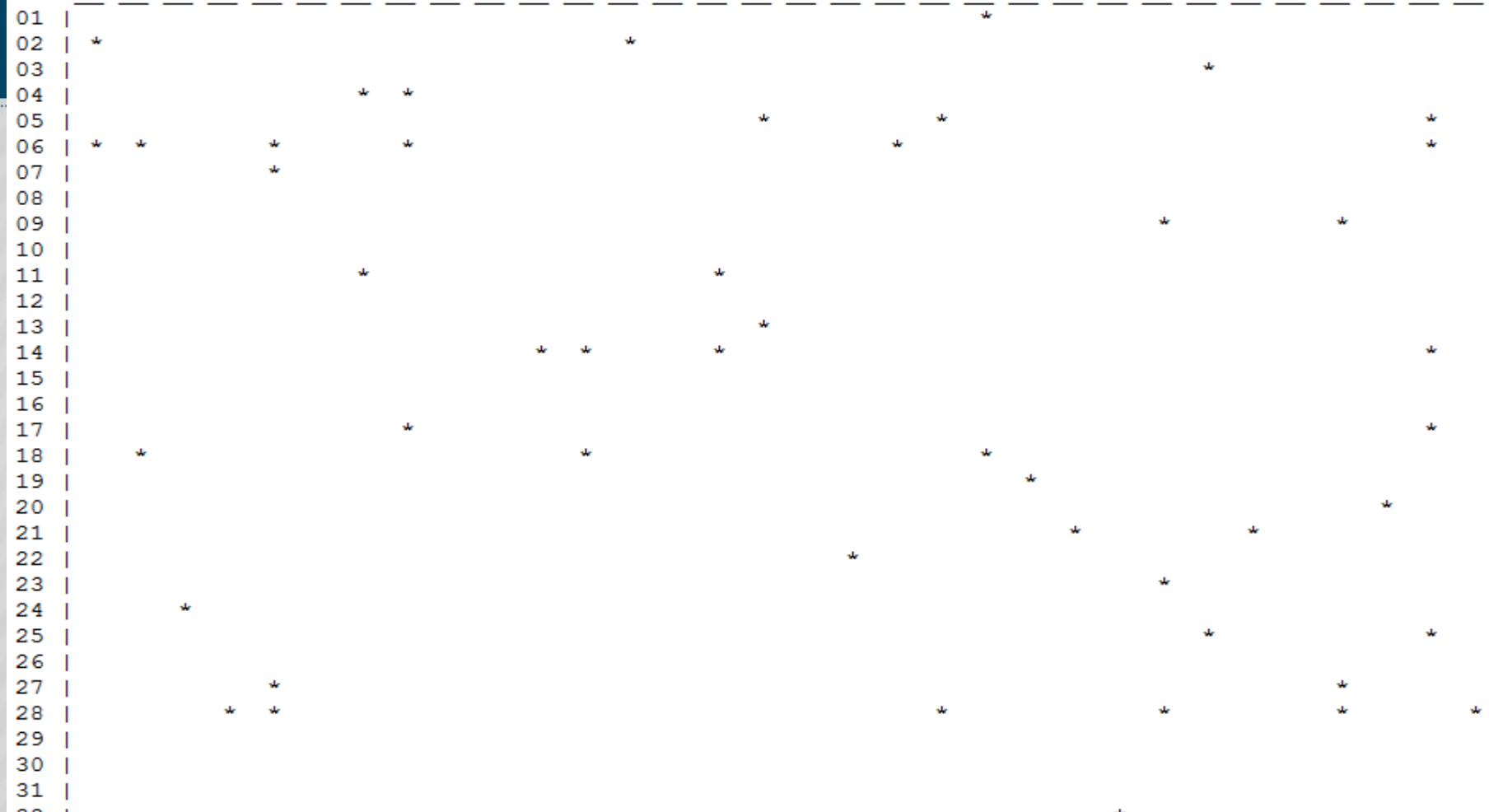
ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
01																					*											*
02																																
03								*								*														*		
04							*														*											
05																*				*					*						*	
06			*															*				*				*						
07					*									*			*													*	*	
08																																
09			*	*																								*		*		
10								*																								
11				*	*									*																		*
12					*		*		*							*		*														*
13				*			*				*	*																				
14																																
15													*																			
16										*															*							
17										*																						
18			*																													
19				*																					*	*						
20																						*			*							
21																																
22								*									*	*														
23																																
24		*						*							*																	
25																											*					
26				*																												
27														*									*						*			
28	*																									*			*			
29		*															*				*											
30								*																								
31																*																
32																	*		*													

BIC: 813033.463068

GENERATION #25 / 5000

ADJACENCY MATRIX:



BIC: 726693.810980

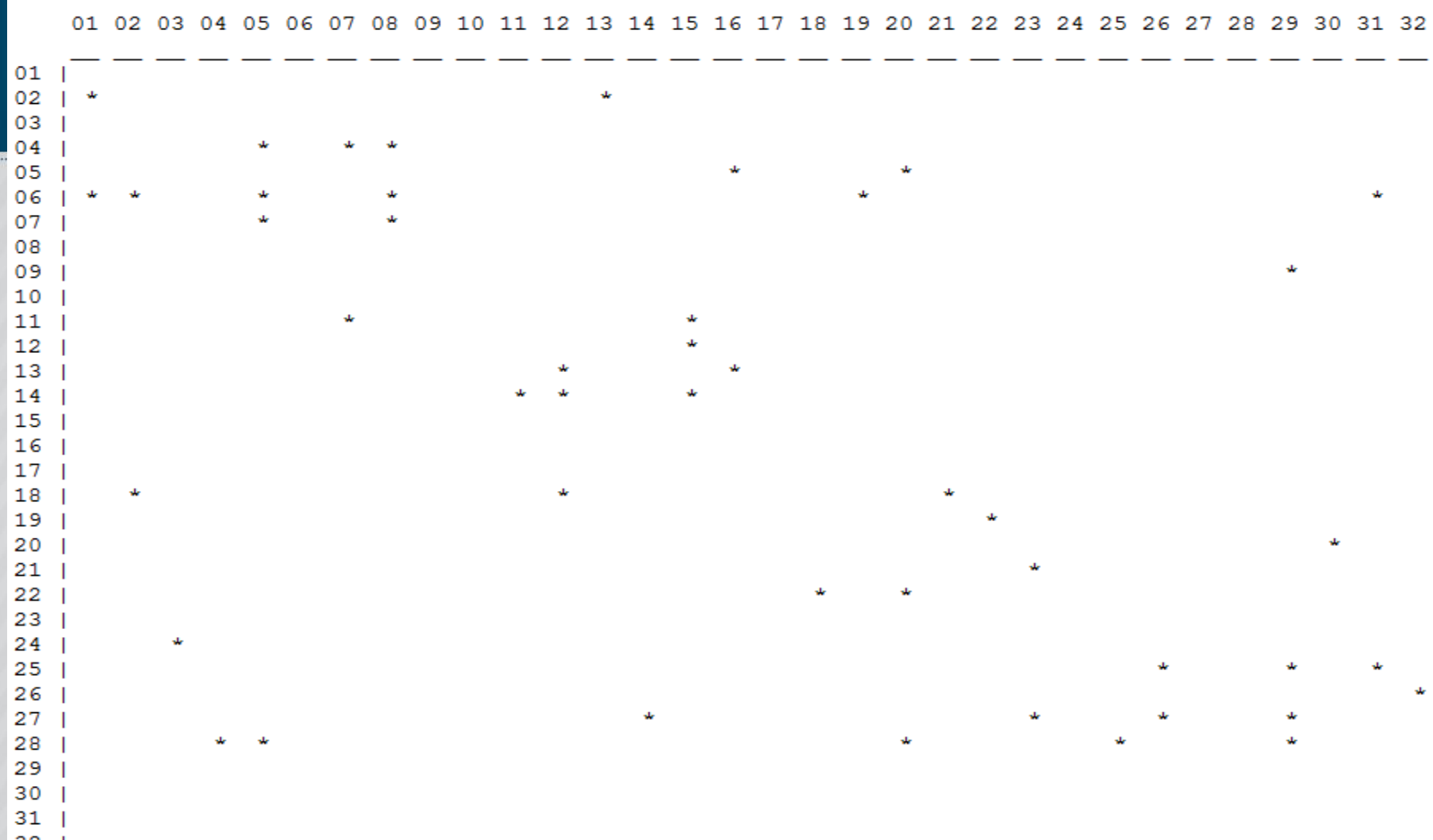
GENERATION #50 / 5000

ADJACENCY MATRIX:

BIC: 710074.763608

GENERATION #75 / 5000

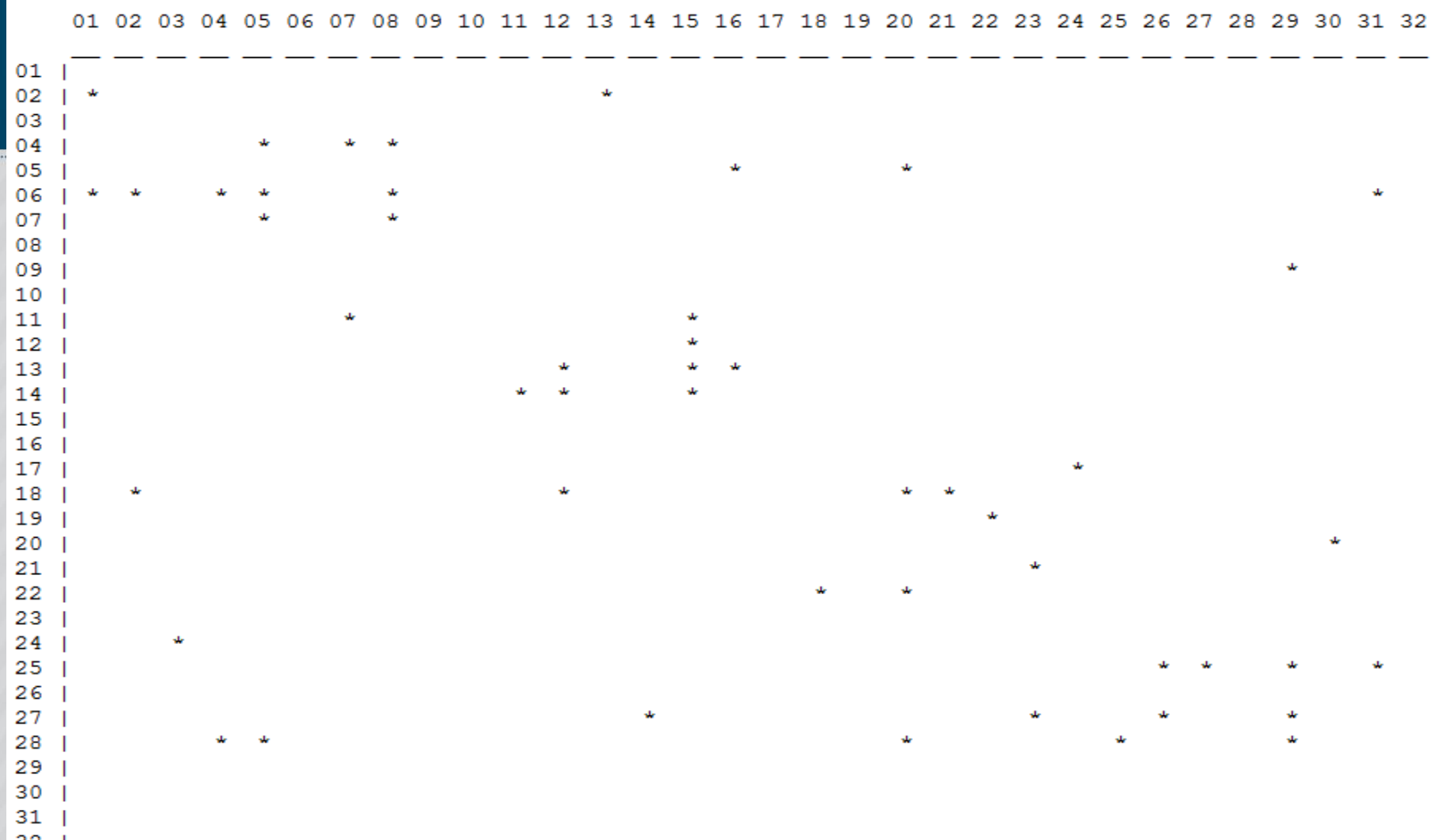
ADJACENCY MATRIX:



BIC: 699140.036300

GENERATION #100 / 5000

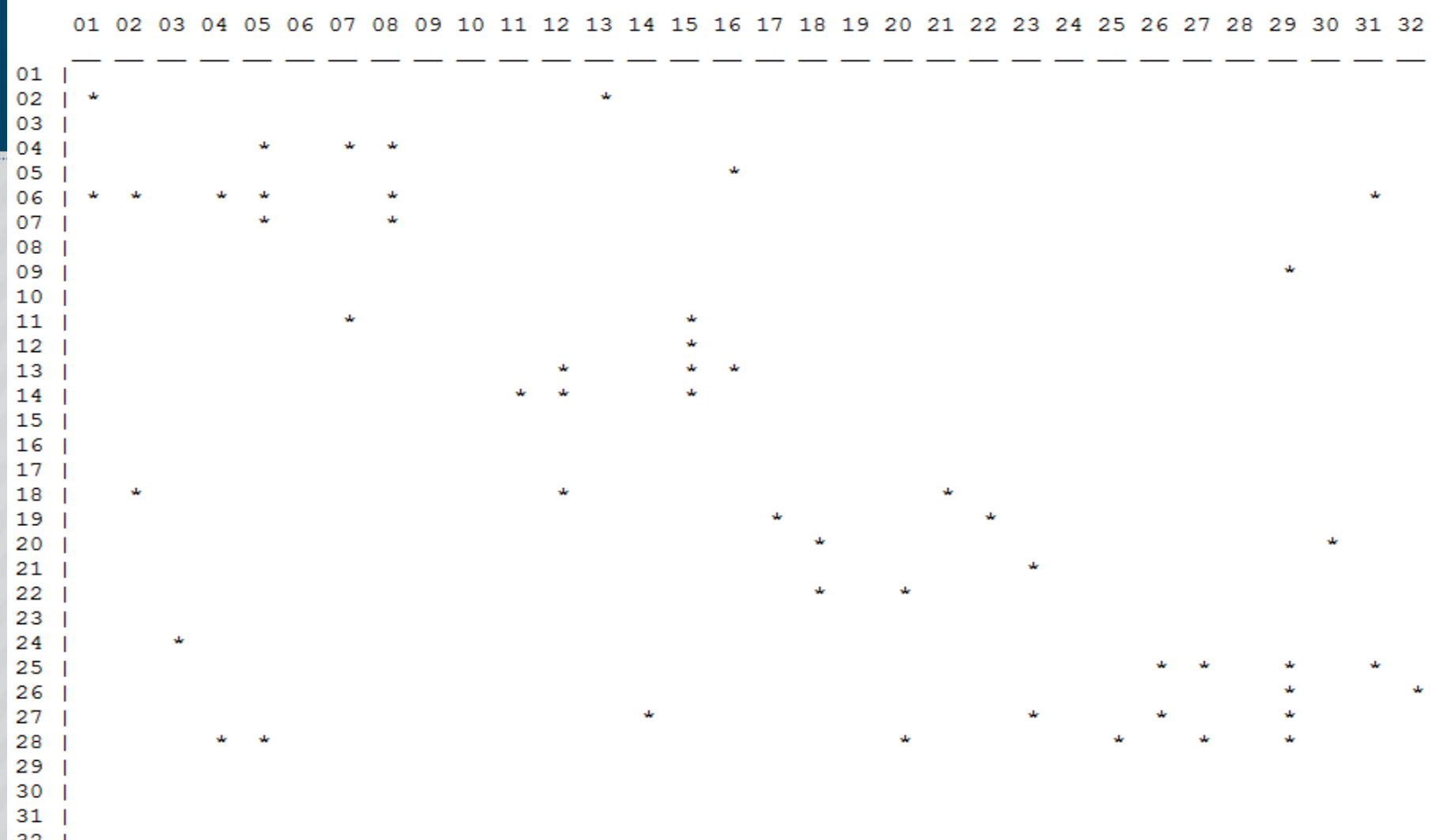
ADJACENCY MATRIX:



BIC: 683288.866511

GENERATION #125 / 5000

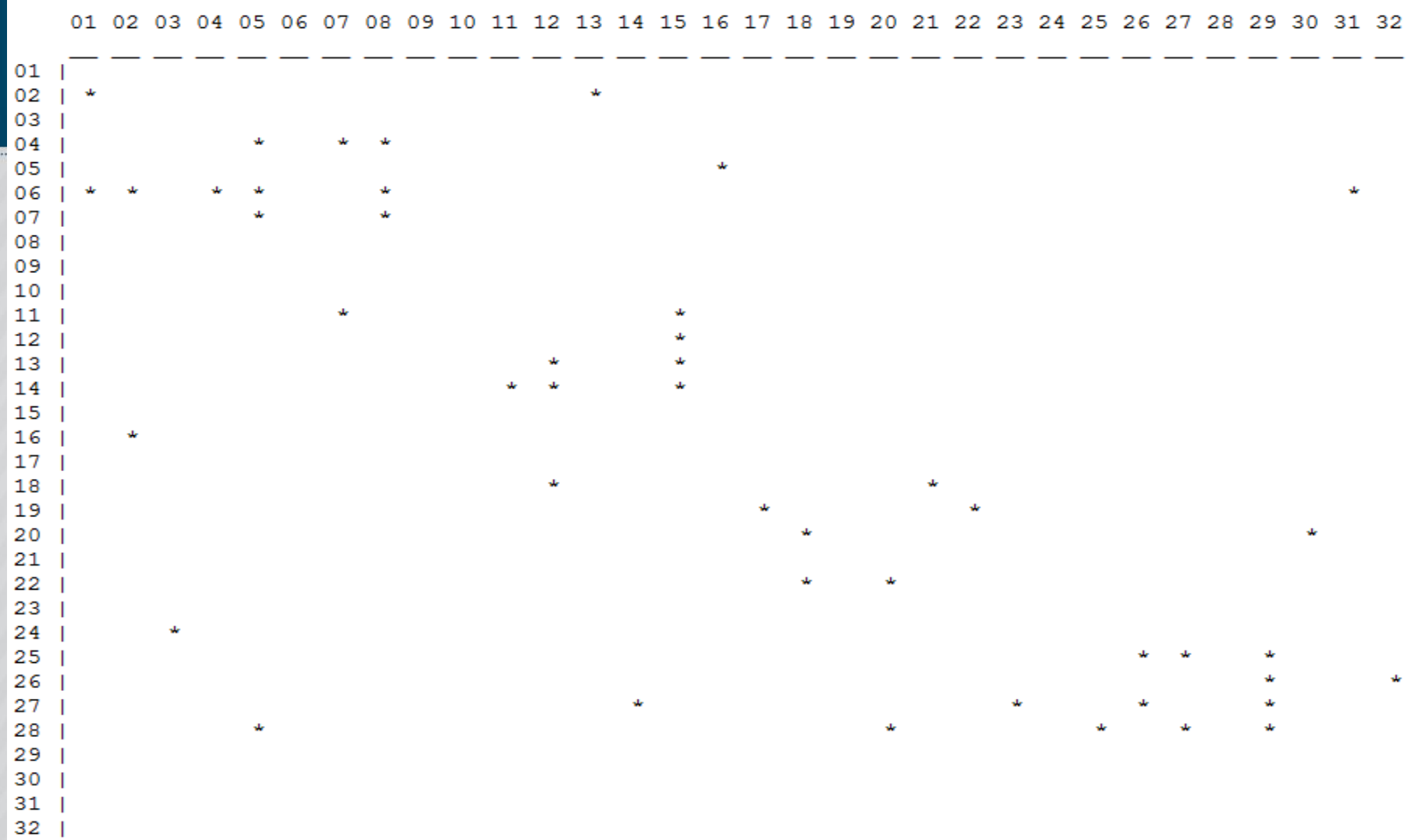
ADJACENCY MATRIX:



BIC: 682658.487793

GENERATION #150 / 5000

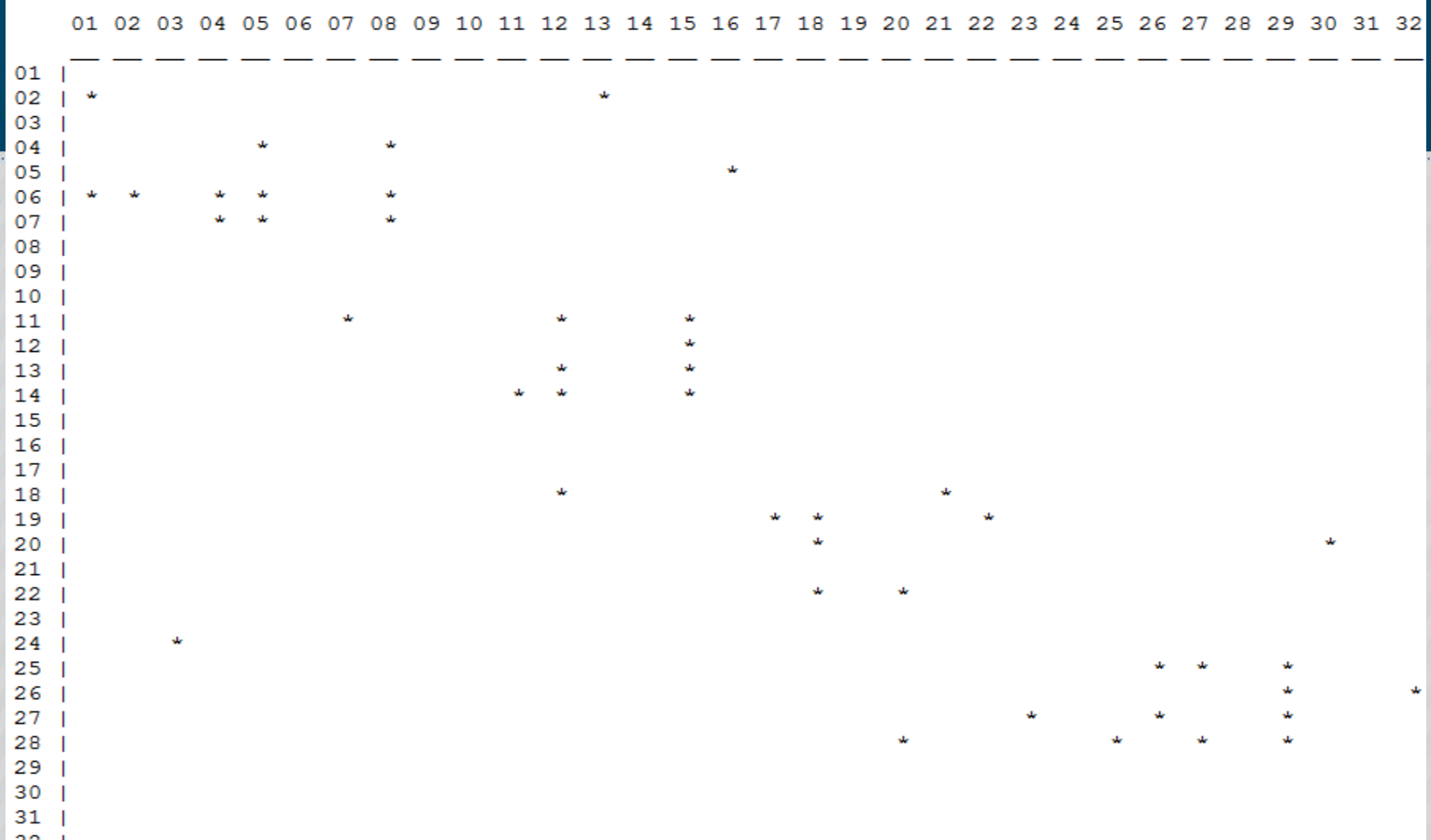
ADJACENCY MATRIX:



BIC: 682334.770794

GENERATION #175 / 5000

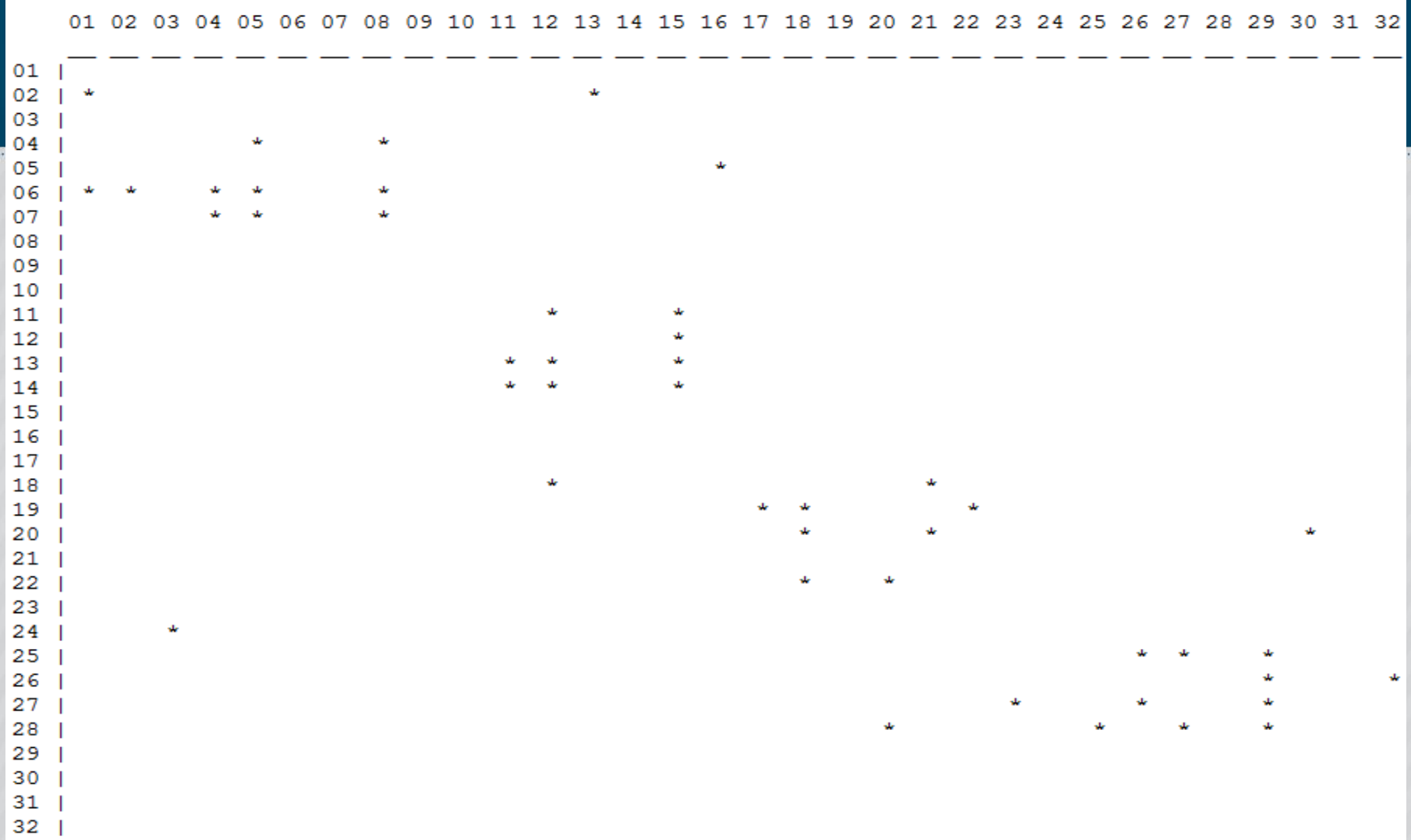
ADJACENCY MATRIX:



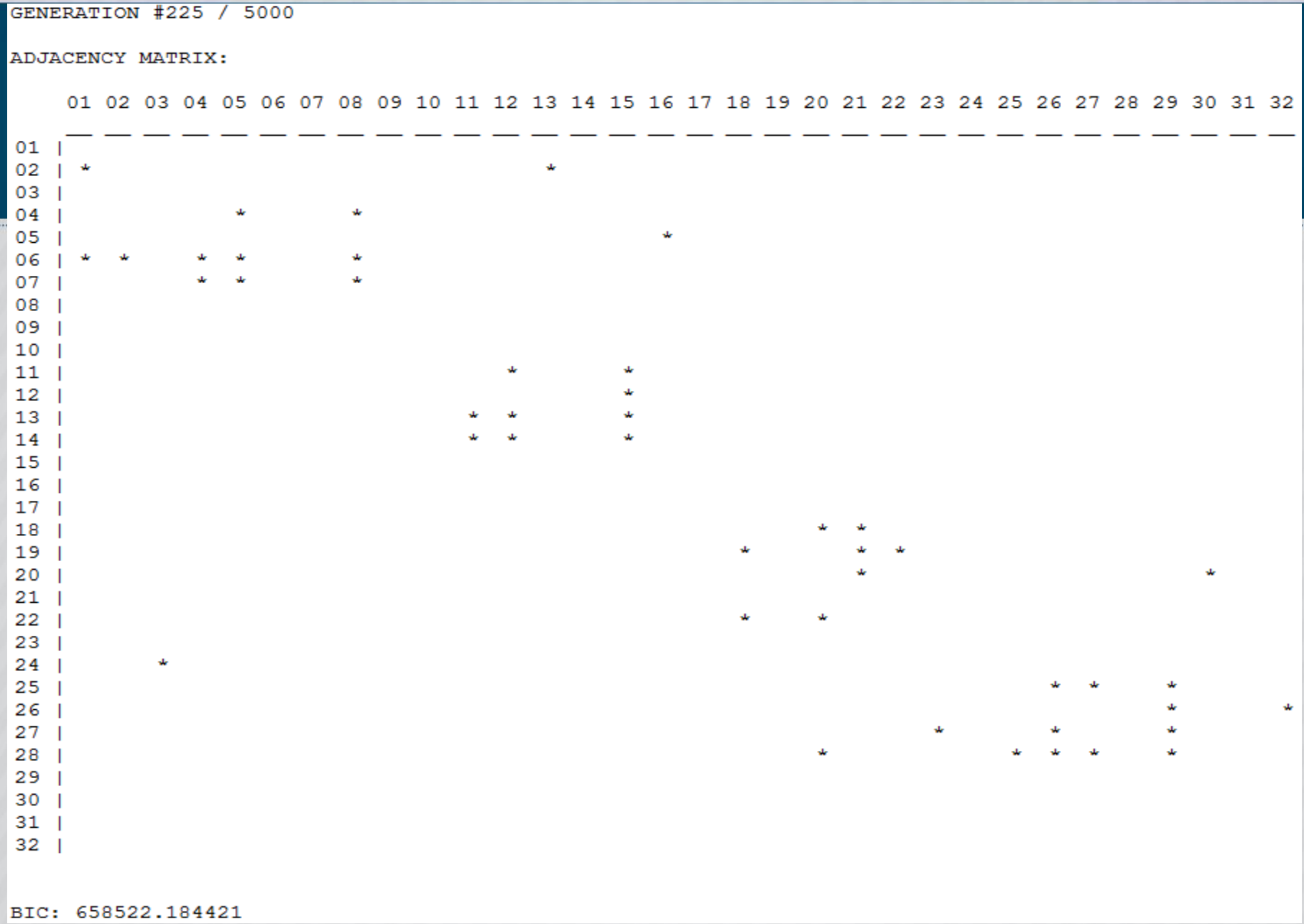
BIC: 675213.747863

GENERATION #200 / 5000

ADJACENCY MATRIX:



BIC: 674206.774448


GENERATION #225 / 5000
ADJACENCY MATRIX:
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
BIC: 658522.184421

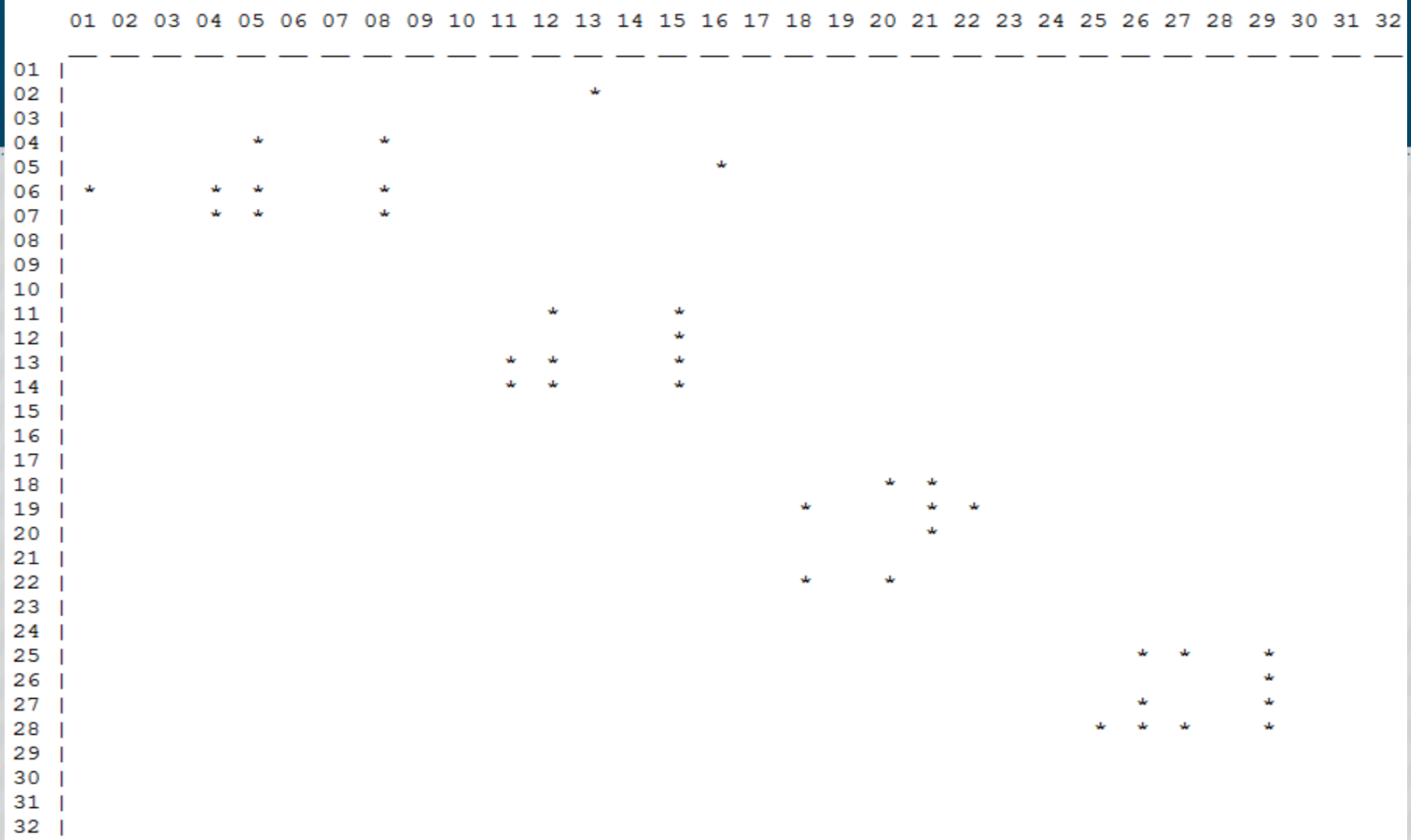
GENERATION #250 / 5000

ADJACENCY MATRIX:

BIC: 658468.903248

GENERATION #275 / 5000

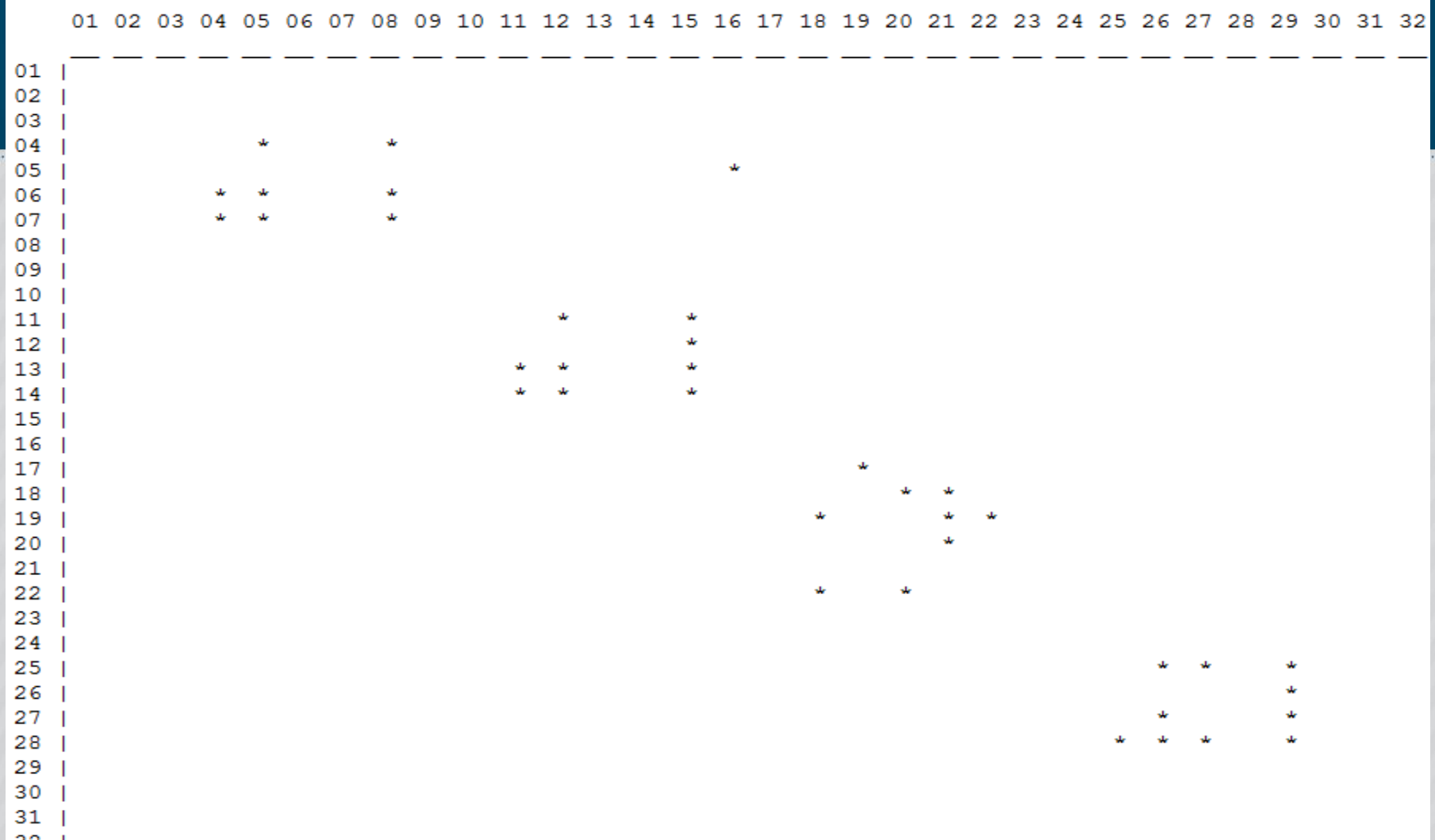
ADJACENCY MATRIX:



BIC: 658377.735340

GENERATION #300 / 5000

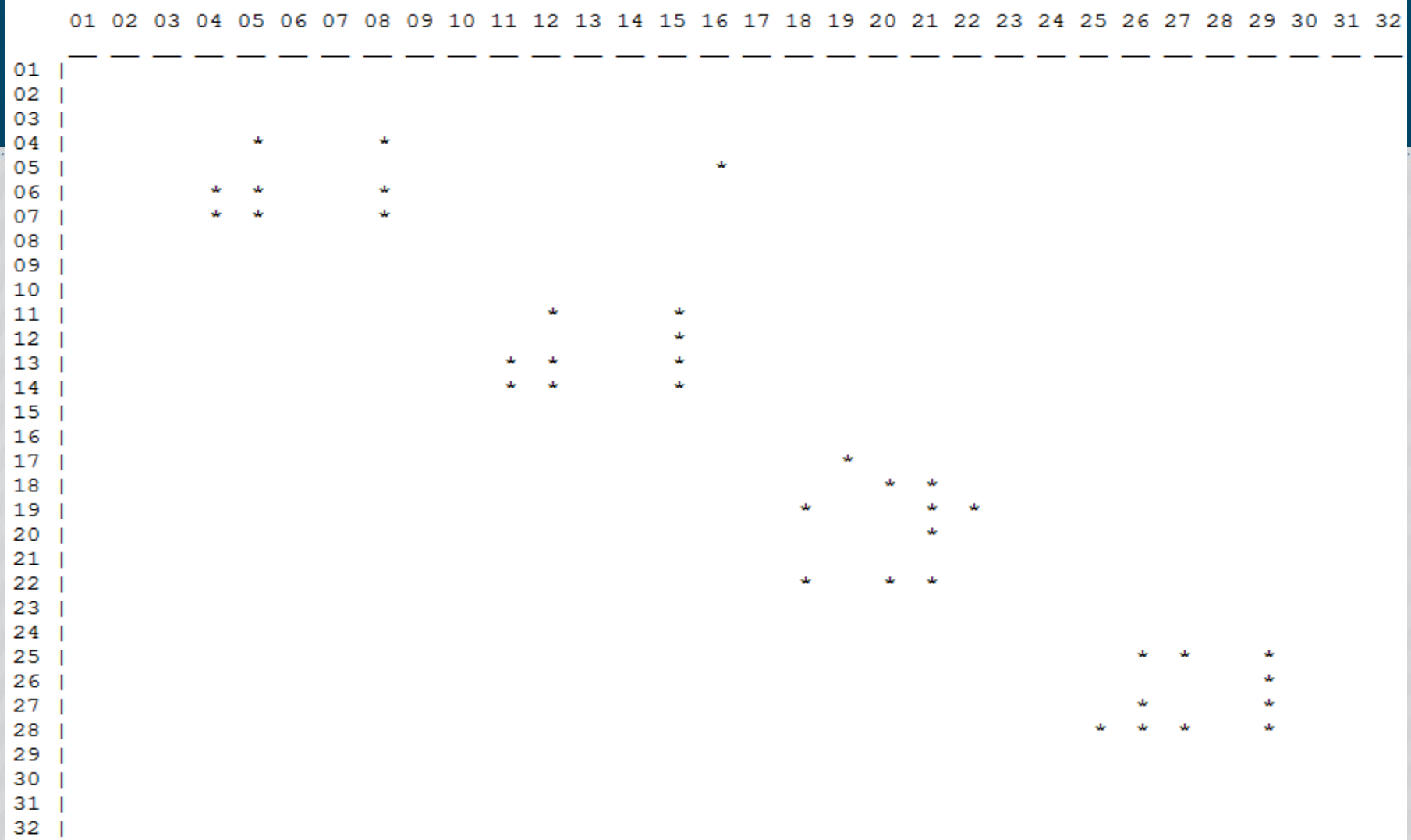
ADJACENCY MATRIX:



BIC: 658342.568077

GENERATION #325 / 5000

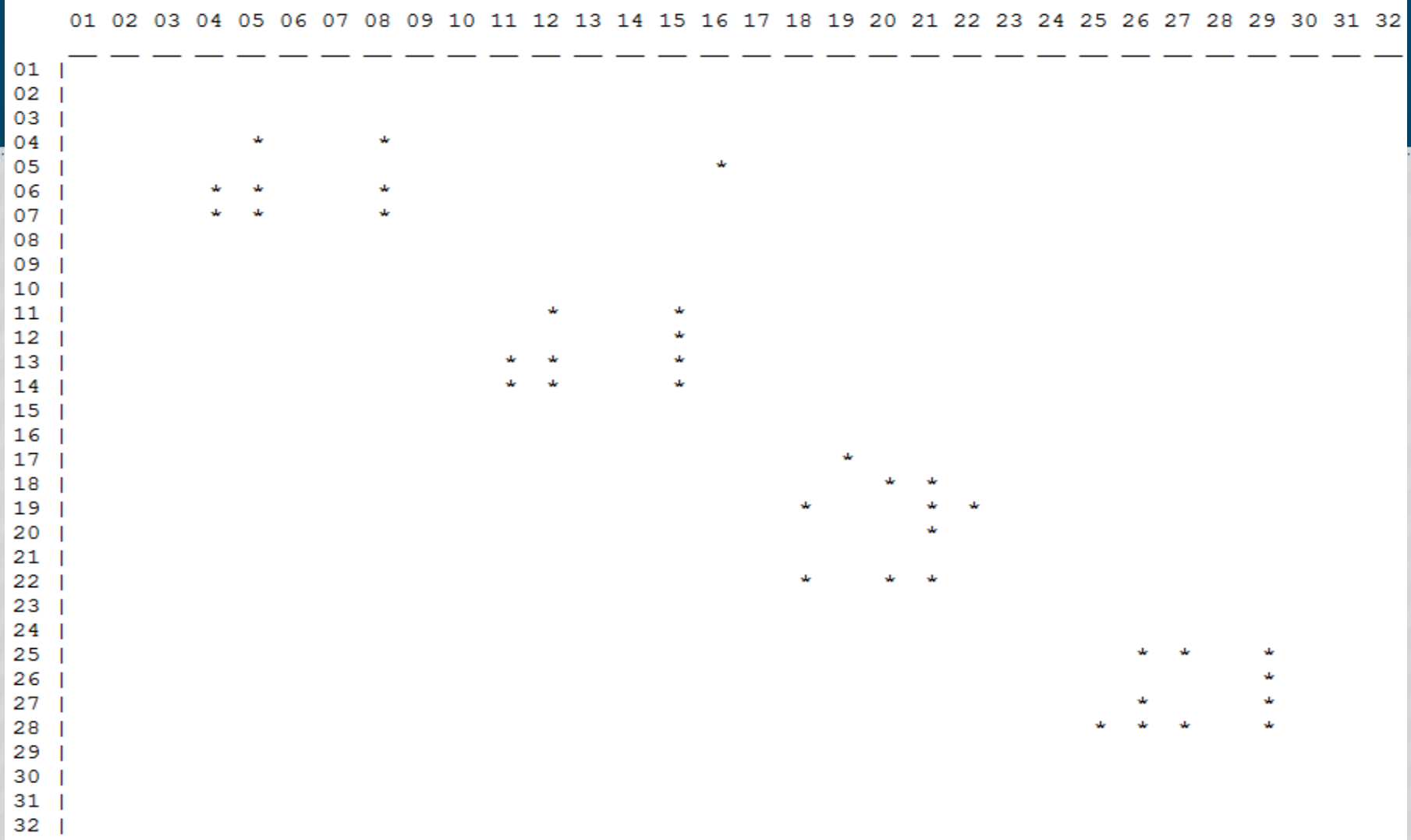
ADJACENCY MATRIX:



BIC: 657256.585419

GENERATION #350 / 5000

ADJACENCY MATRIX:



BIC: 657256.585419

GENERATION #375 / 5000

ADJACENCY MATRIX:

BIC: 657256.585419

GENERATION #400 / 5000

ADJACENCY MATRIX:

BIC: 657254.405126

GENERATION #425 / 5000

ADJACENCY MATRIX:

BIC: 651795.763259

GENERATION #450 / 5000

ADJACENCY MATRIX:

BIC: 651783.140737

GENERATION #475 / 5000

ADJACENCY MATRIX:

BIC: 651783.140737

GENERATION #500 / 5000

ADJACENCY MATRIX:

BIC: 651783.140737

BEST-FIT MODEL:

ADJACENCY MATRIX:

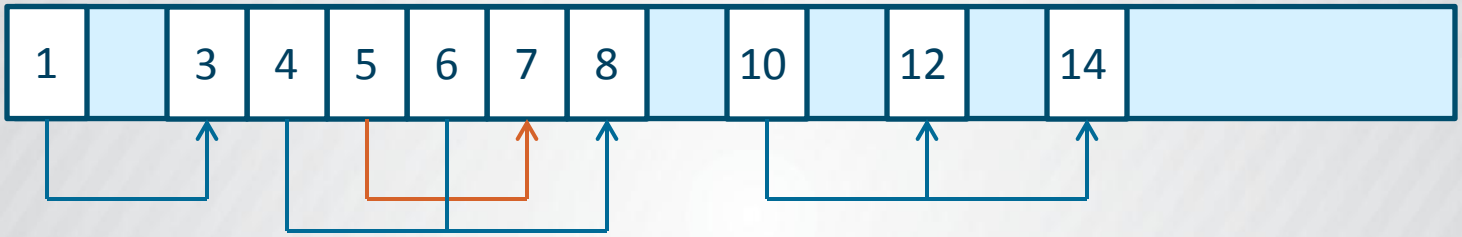
BIC: 651783.140737

Writing BEST-FIT MODEL to "edges.out" ...

Computing Min-Entropy Estimate ...

ESTIMATED MIN ENTROPY : 17.050647

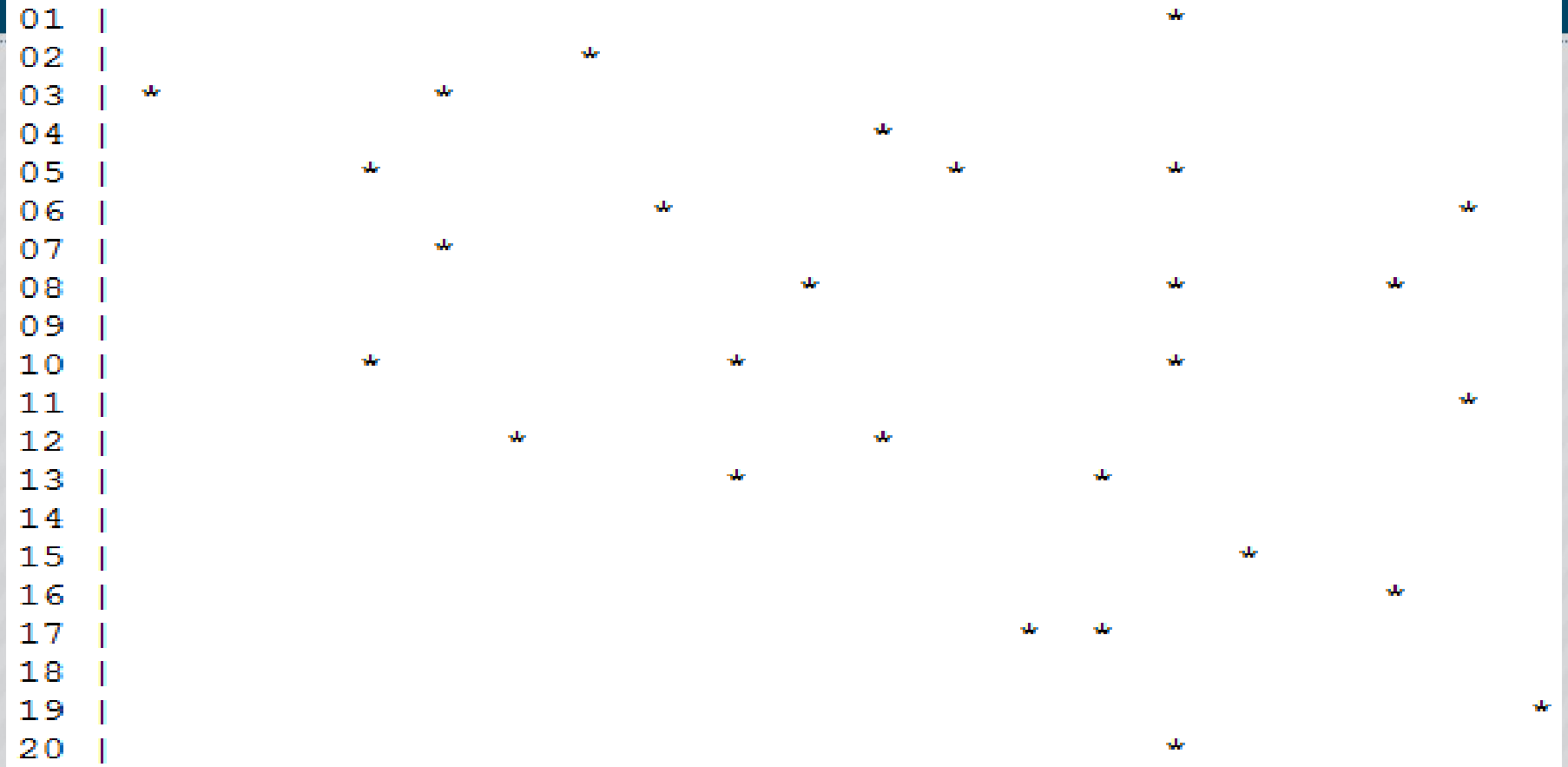
Example 2



- 20-bit blocks; sample size 15,000
- Bit 3 dependent on 1; 7 on 5; 8 on 4 and 6; 12 and 14 on 10
- All other bits unbiased and independent
- Actual min-entropy is 16.3104…

GENERATION #1 / 5000

ADJACENCY MATRIX:

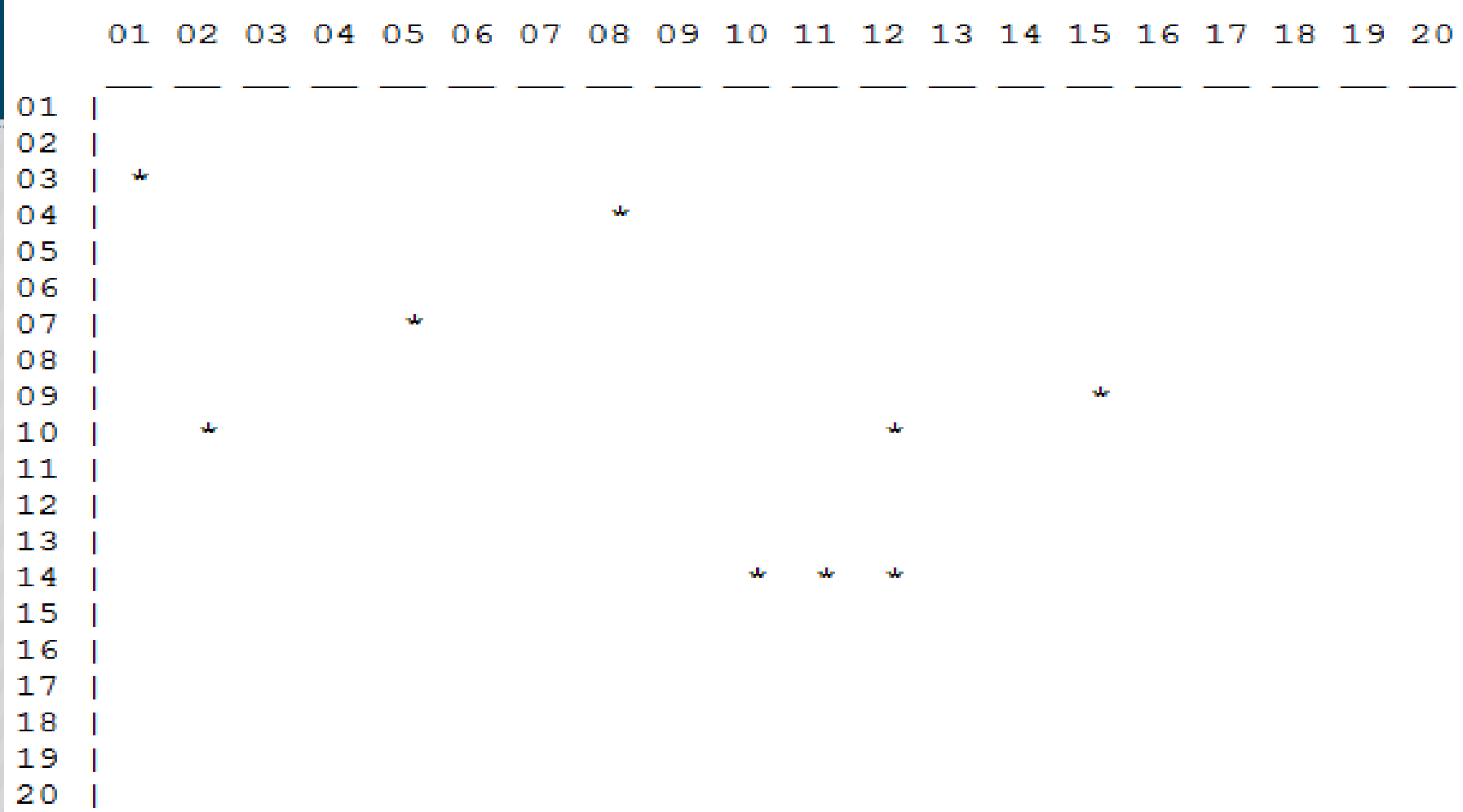


	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01							*								*					
02																				
03	*				*															
04											*									
05				*								*			*					
06								*											*	
07					*															
08										*					*			*		
09																				
10				*					*						*					
11																			*	
12						*					*									
13									*					*						
14																				
15																*				
16																		*		
17													*	*						
18																				
19																				*
20															*					

BIC: 578019.435300

GENERATION #25 / 5000

ADJACENCY MATRIX:



BIC: 570684.444407

GENERATION #50 / 5000

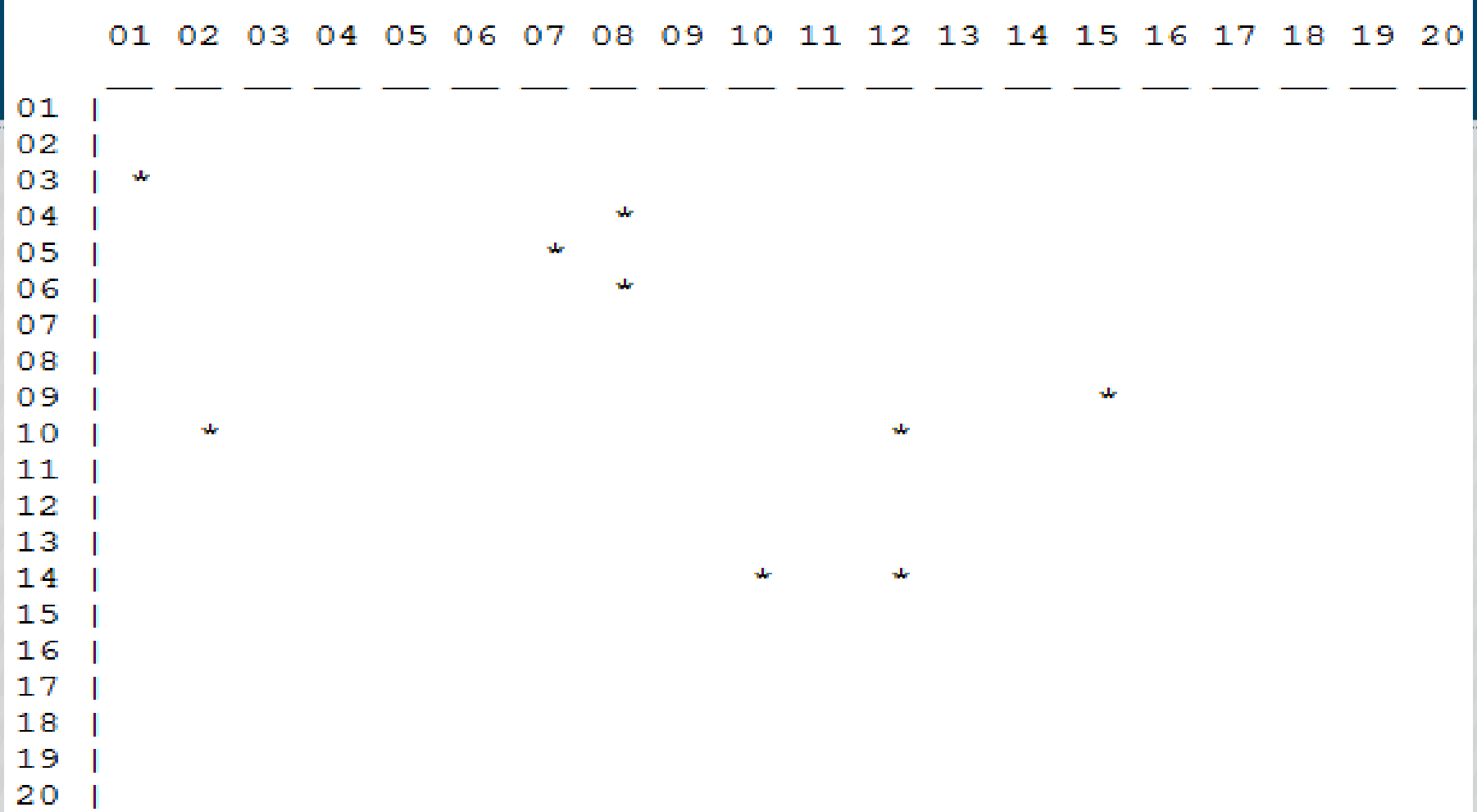
ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01																				
02																				
03	*																			
04								*												
05																				
06								*												
07					*															
08																				
09															*					
10		*										*								
11																				
12																				
13																				
14										*		*								
15																				
16																				
17																				
18																				
19																				
20																				

BIC: 570455.316561

GENERATION #75 / 5000

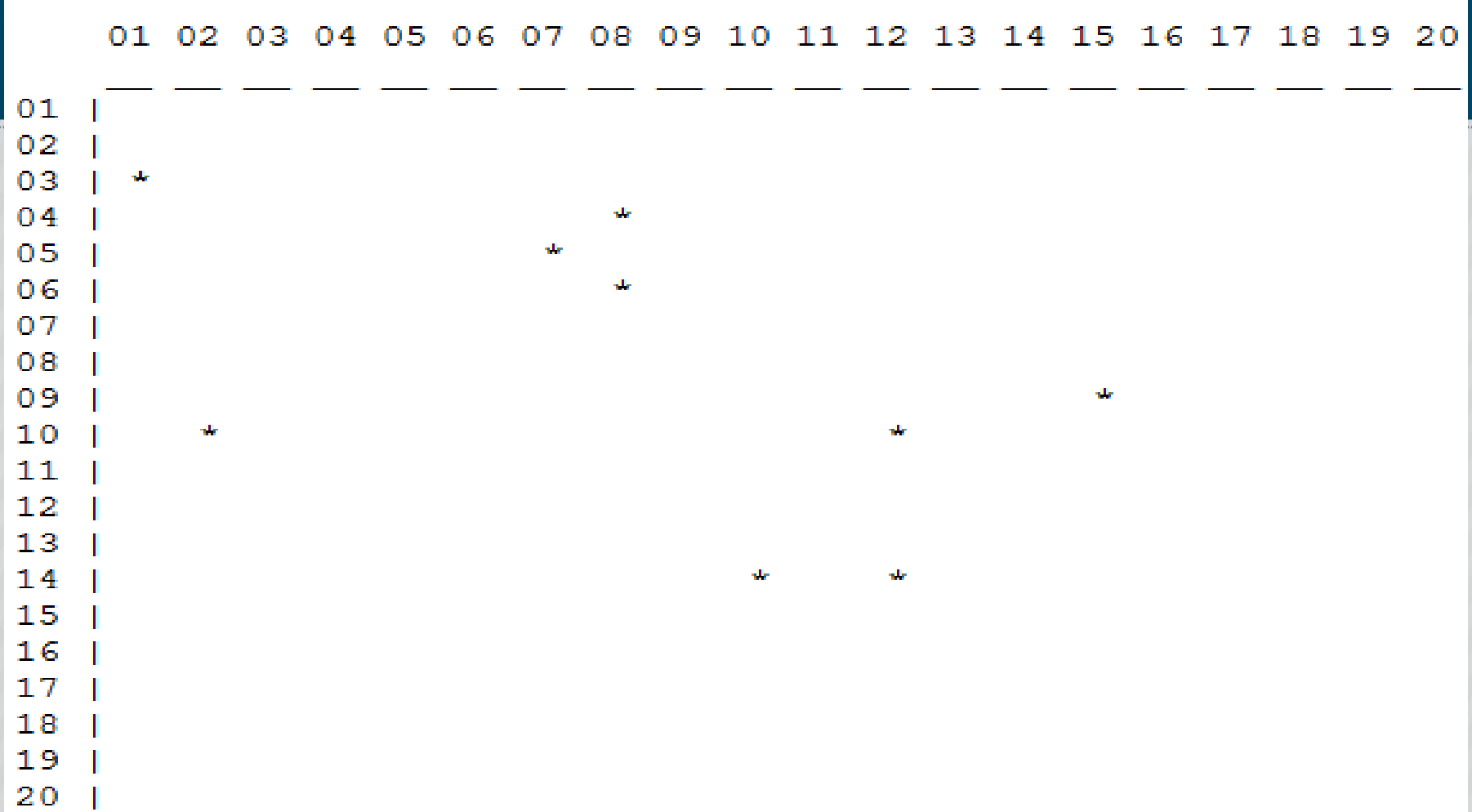
ADJACENCY MATRIX:



BIC: 570455.316561

GENERATION #100 / 5000

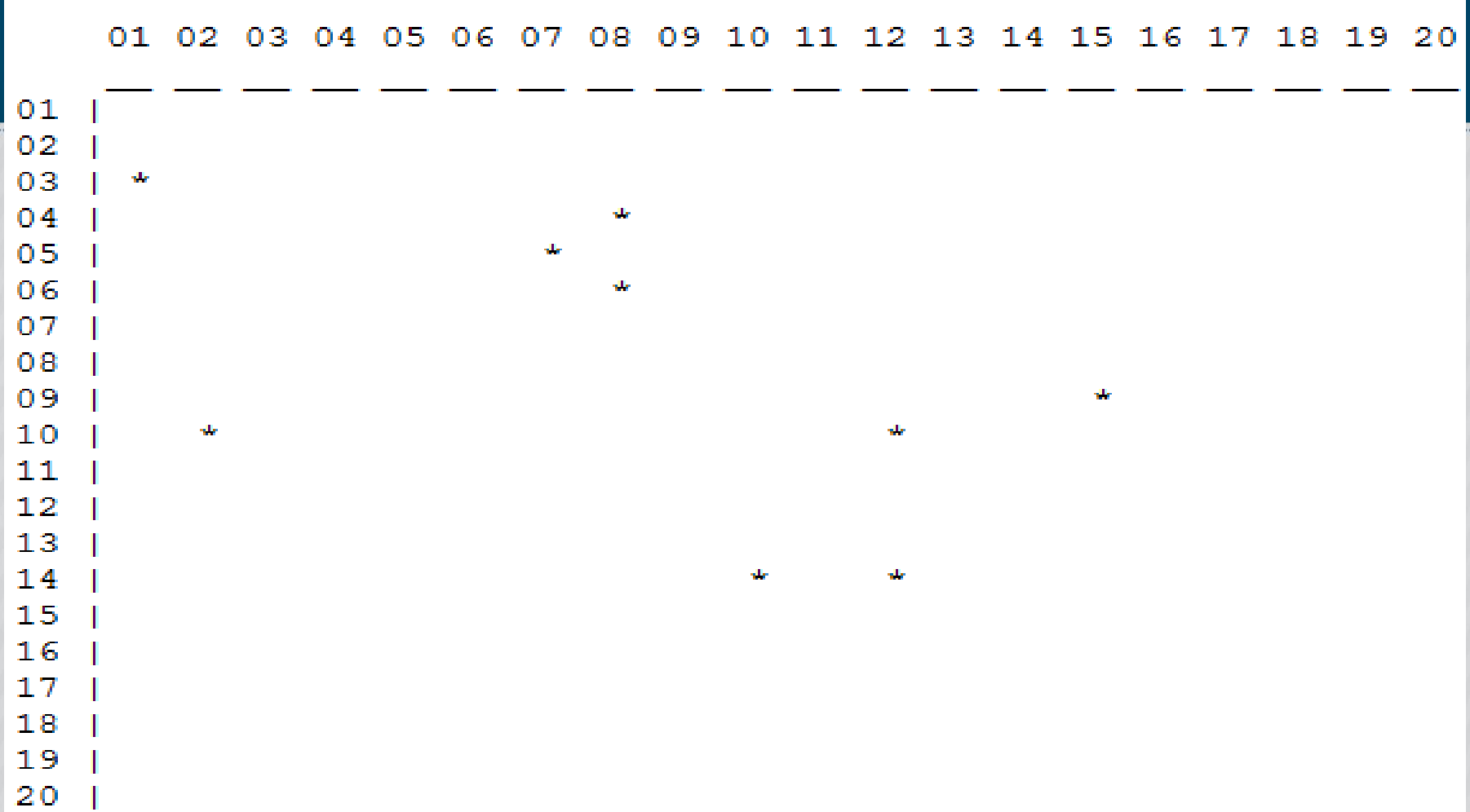
ADJACENCY MATRIX:



BIC: 570455.316561

GENERATION #125 / 5000

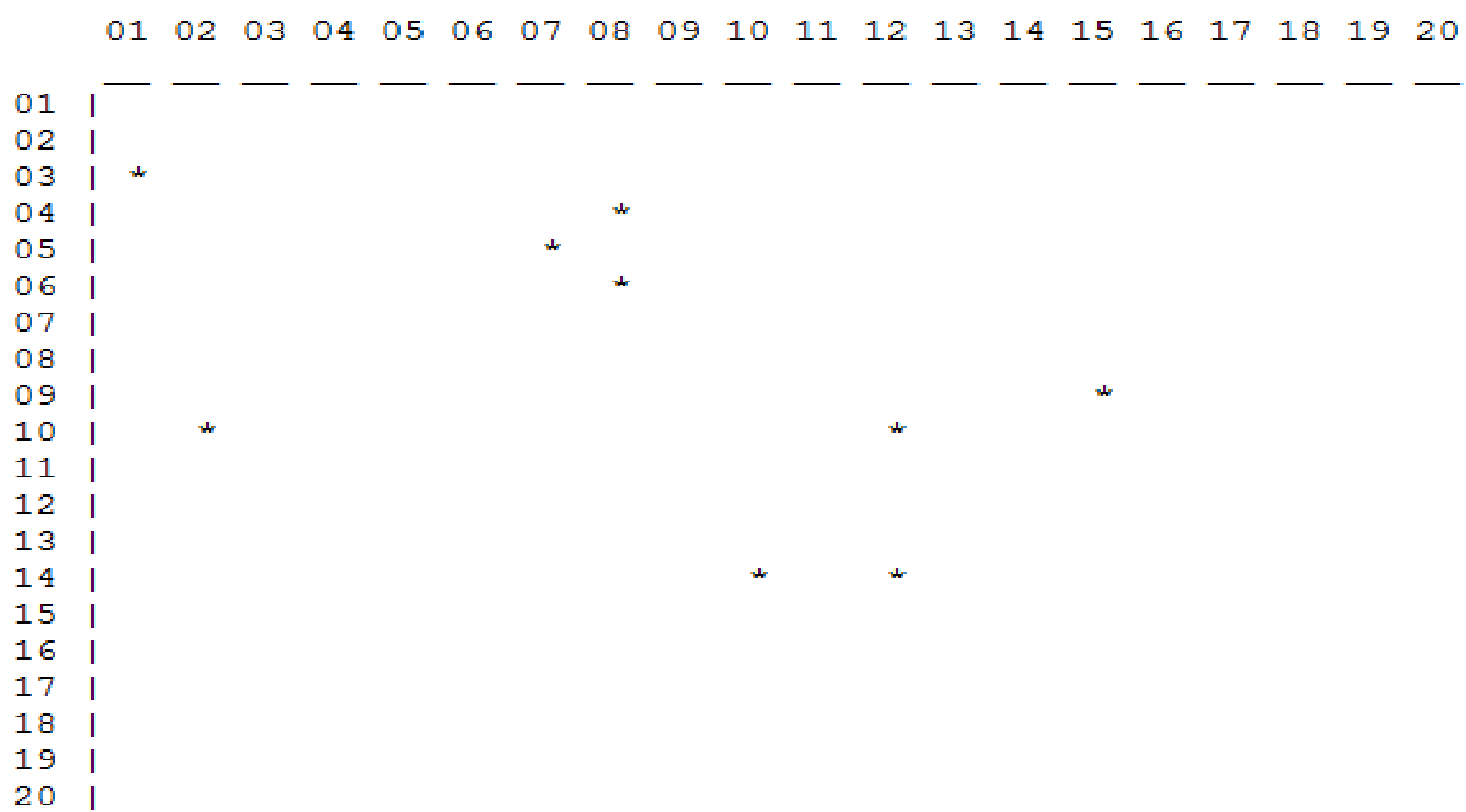
ADJACENCY MATRIX:



BIC: 570455.316561

GENERATION #150 / 5000

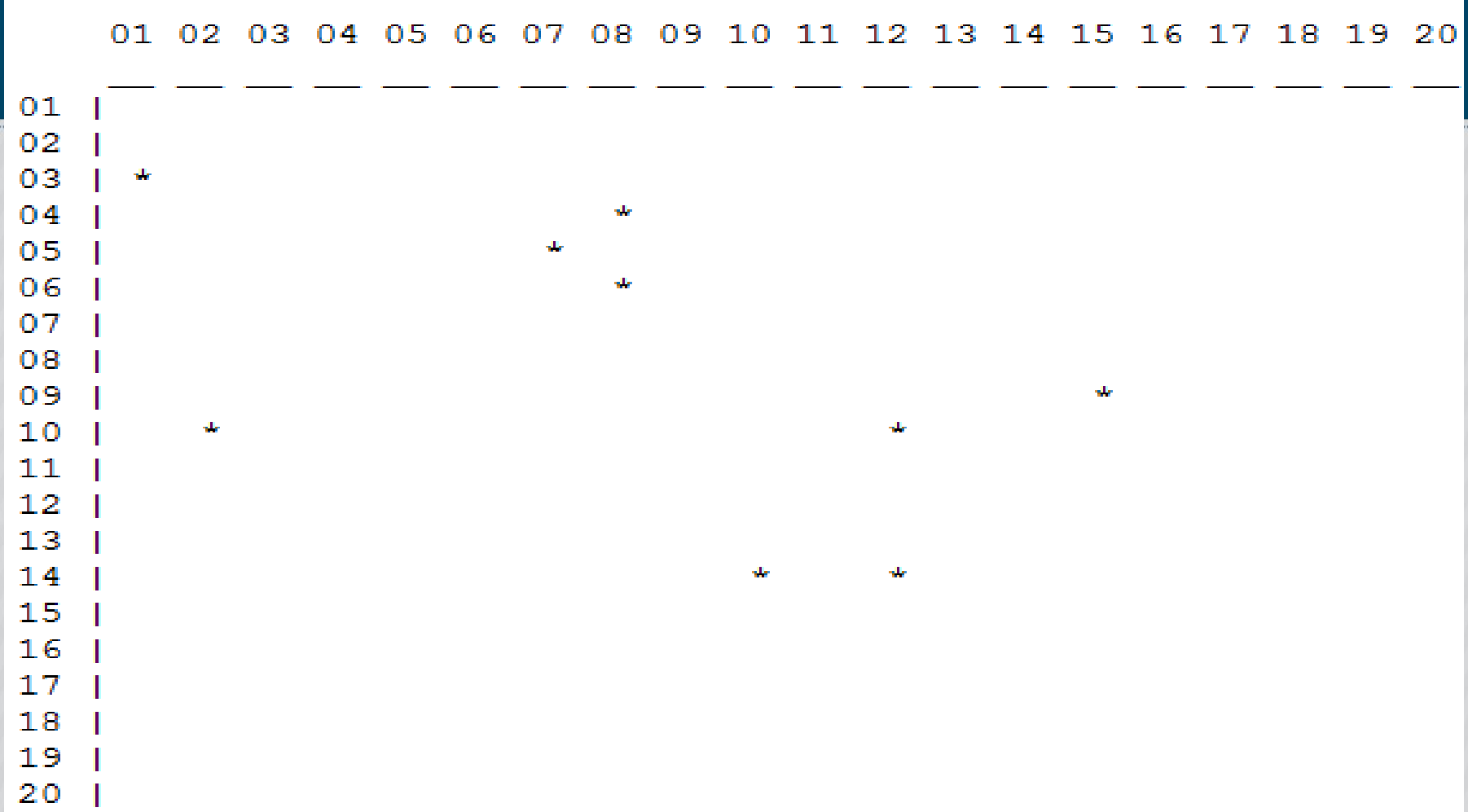
ADJACENCY MATRIX:



BIC: 570455.316561

GENERATION #175 / 5000

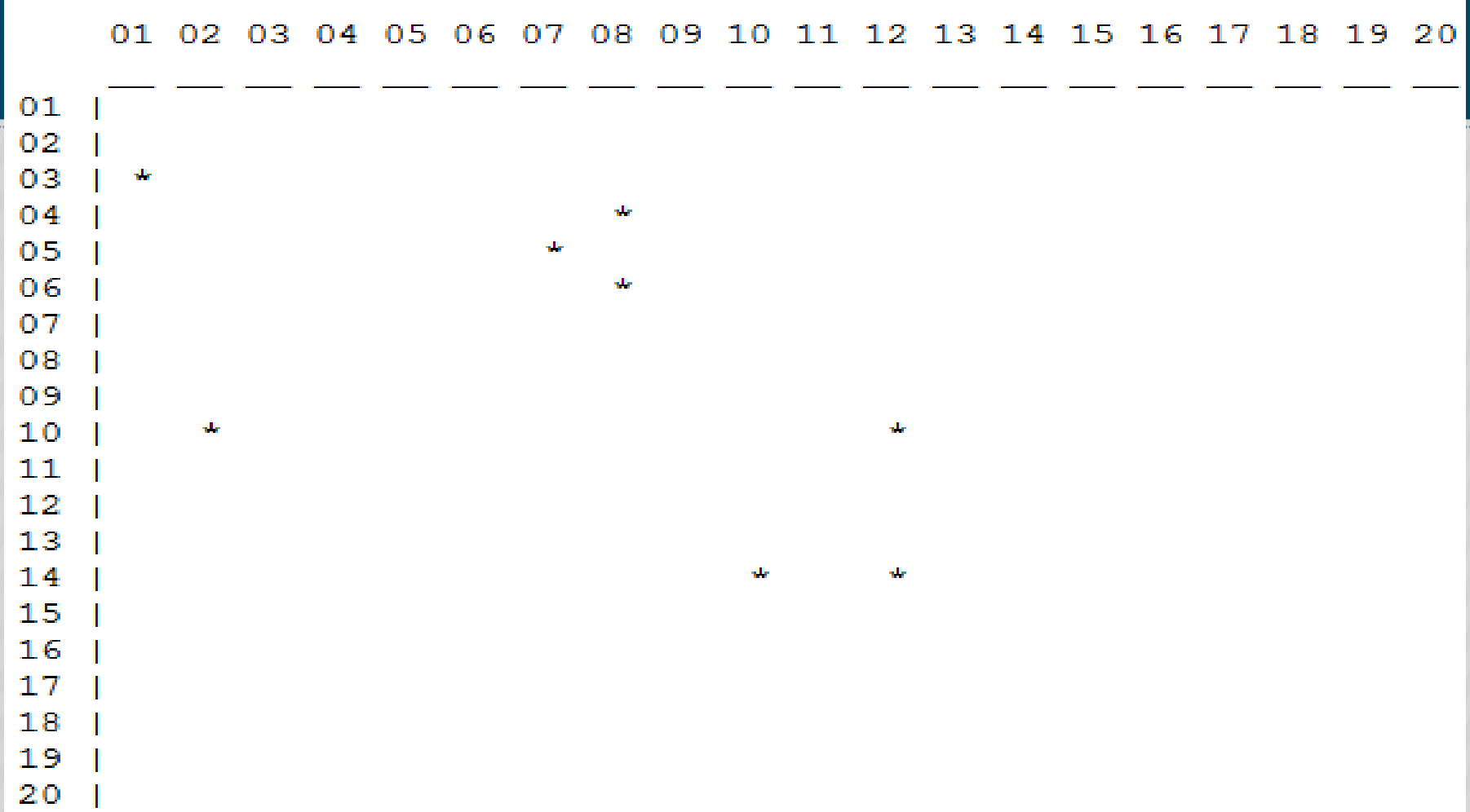
ADJACENCY MATRIX:



BIC: 570455.316561

GENERATION #200 / 5000

ADJACENCY MATRIX:



BIC: 570443.807480

GENERATION #225 / 5000

ADJACENCY MATRIX:

```
    01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __
01 |
02 |
03 |  *
04 |                       *
05 |                    *
06 |                       *
07 |
08 |
09 |
10 |     *                                 *
11 |
12 |
13 |
14 |                                   *   *
15 |
16 |
17 |
18 |
19 |
20 |
```

BIC: 570443.807480

GENERATION #250 / 5000

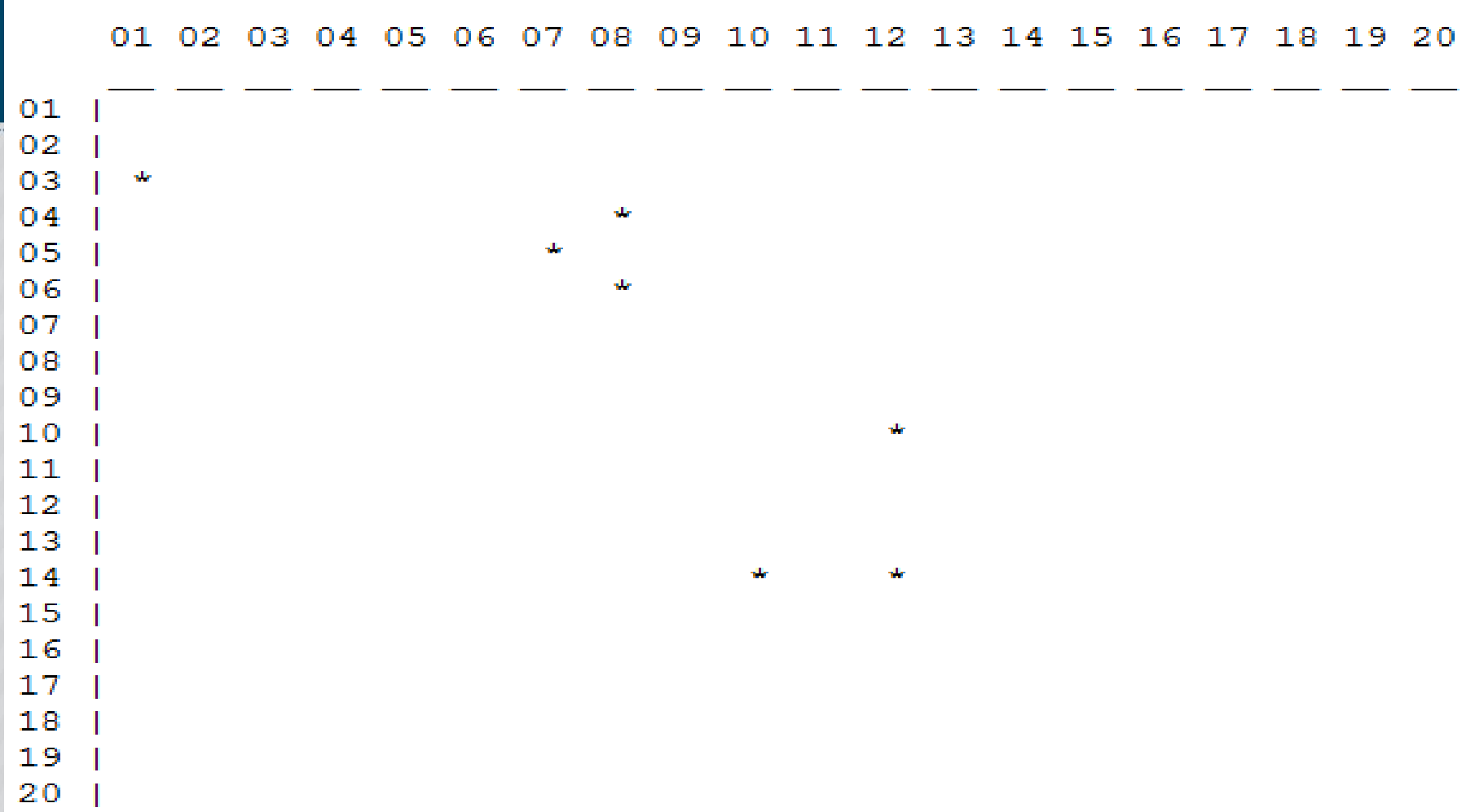
ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01																				
02																				
03	*																			
04								*												
05							*													
06								*												
07																				
08																				
09																				
10												*								
11																				
12																				
13																				
14										*		*								
15																				
16																				
17																				
18																				
19																				
20																				

BIC: 570432.721112

GENERATION #275 / 5000

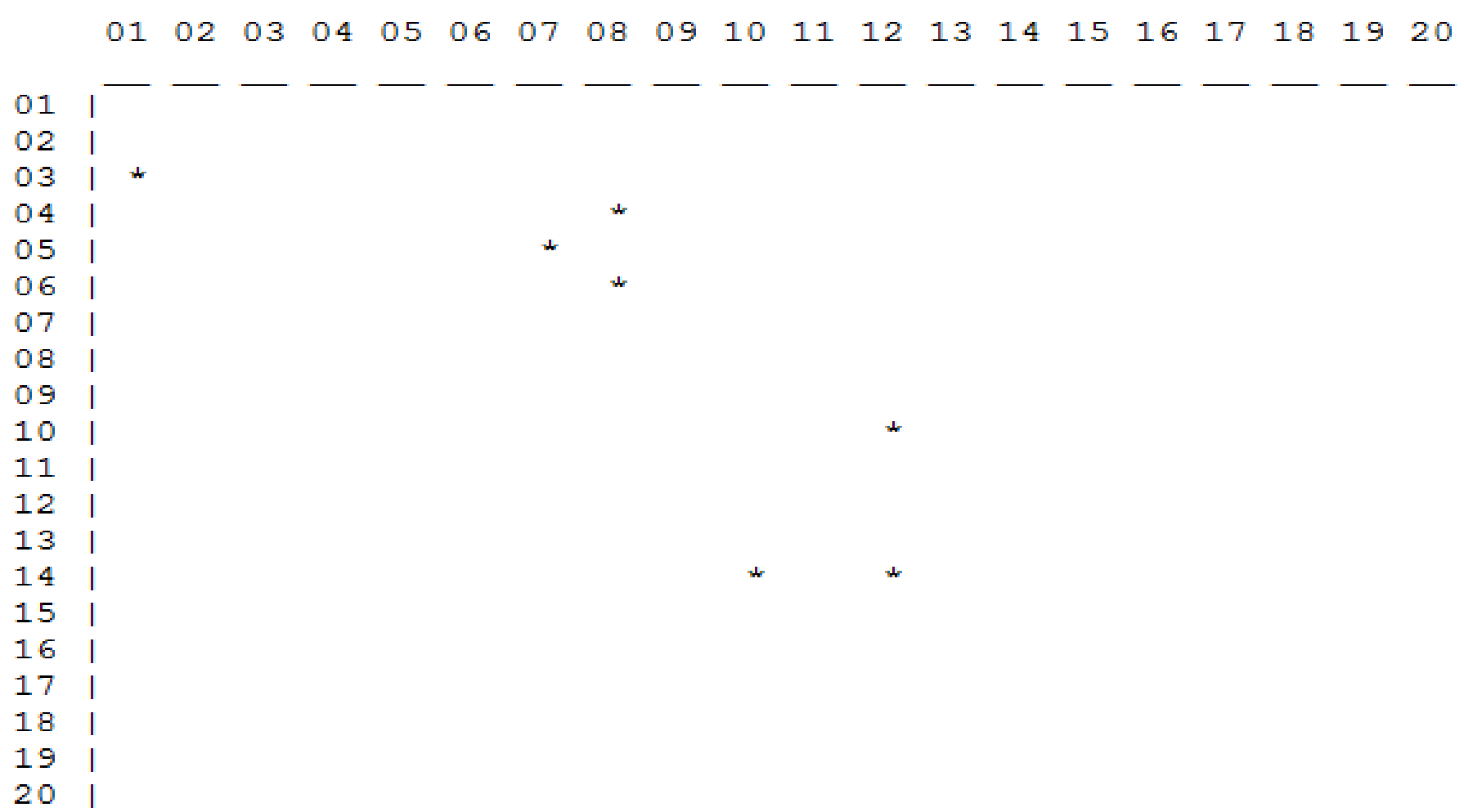
ADJACENCY MATRIX:



BIC: 570432.721112

GENERATION #300 / 5000

ADJACENCY MATRIX:



BIC: 570432.721112

GENERATION #325 / 5000

ADJACENCY MATRIX:

```
    01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
01 |
02 |
03 |  *
04 |                       *
05 |                    *
06 |                       *
07 |
08 |
09 |
10 |                                      *
11 |
12 |
13 |
14 |                                   *     *
15 |
16 |
17 |
18 |
19 |
20 |
```

BIC: 570432.721112

GENERATION #350 / 5000

ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01																				
02																				
03	*																			
04								*												
05							*													
06								*												
07																				
08																				
09																				
10												*		*						
11																				
12																				
13																				
14												*								
15																				
16																				
17																				
18																				
19																				
20																				

BIC: 570432.721112

GENERATION #375 / 5000

ADJACENCY MATRIX:

```
    01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __
01 |
02 |
03 |  *
04 |                       *
05 |                    *
06 |                       *
07 |
08 |
09 |
10 |                                      *        *
11 |
12 |
13 |
14 |                                      *
15 |
16 |
17 |
18 |
19 |
20 |
```

BIC: 570432.721112

GENERATION #400 / 5000

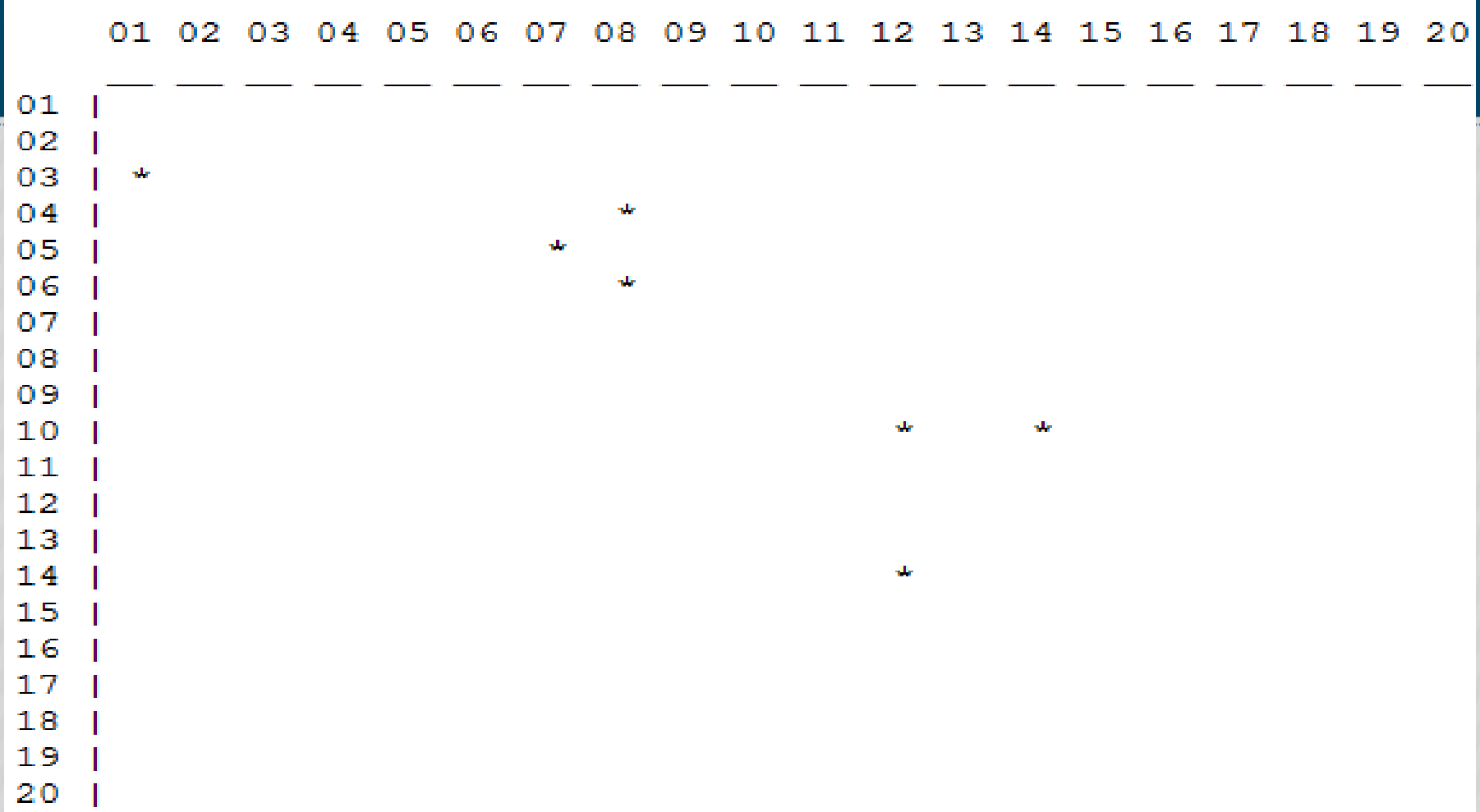
ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01																				
02																				
03	*																			
04								*												
05							*													
06								*												
07																				
08																				
09																				
10												*		*						
11																				
12																				
13																				
14												*								
15																				
16																				
17																				
18																				
19																				
20																				

BIC: 570432.721112

GENERATION #425 / 5000

ADJACENCY MATRIX:



BIC: 570432.721112

GENERATION #450 / 5000

ADJACENCY MATRIX:

```
    01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __
01 |
02 |
03 |  *
04 |                       *
05 |                    *
06 |                       *
07 |
08 |
09 |
10 |                                      *        *
11 |
12 |
13 |
14 |                                      *
15 |
16 |
17 |
18 |
19 |
20 |
```

BIC: 570432.721112

GENERATION #475 / 5000

ADJACENCY MATRIX:

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01																				
02																				
03	*																			
04								*												
05							*													
06								*												
07																				
08																				
09																				
10												*		*						
11																				
12																				
13																				
14												*								
15																				
16																				
17																				
18																				
19																				
20																				

BIC: 570432.721112

GENERATION #500 / 5000

ADJACENCY MATRIX:

```
    01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- --
01 |
02 |
03 |  *
04 |                       *
05 |                    *
06 |                       *
07 |
08 |
09 |
10 |                                      *        *
11 |
12 |
13 |
14 |                                      *
15 |
16 |
17 |
18 |
19 |
20 |
```

BIC: 570432.721112

```
BEST-FIT MODEL:

ADJACENCY MATRIX:

     01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
    __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __ __
01  |
02  |
03  |  *
04  |                       *
05  |                    *
06  |                       *
07  |
08  |
09  |
10  |                                      *        *
11  |
12  |
13  |
14  |                                      *
15  |
16  |
17  |
18  |
19  |
20  |


BIC: 570432.721112

Writing BEST-FIT MODEL to "edges.out" ...

Computing Min-Entropy Estimate ...

ESTIMATED MIN ENTROPY : 16.251261
```

References

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models, Principles and Techniques*. Cambridge, Massachusetts: The MIT Press.