AN ALGORITHM FOR COMPUTING THE NUCLEI OF TIME DEPENDENT DIRECTED GRAPHS

TECHNICAL INFORMATION

Quentin Vanhaelen (vanhaelen@insilicomedicine.com)

Abstract

This document is associated with the set of codes ECextractor and NUCLEIextractor which are designed to compute the set of nuclei in directed graphs used as a representation for dynamical systems. This document contains detailed description of the different output files generated by the three components of the pipeline as well as information about the different inputs required from the user. Guidelines to compile and run the three codes can be found in the README file.

1 ECEXTRACTOR

1.1 Input data

There are two input files required by ECEXTRACTOR:

1. The first one is the *in silico* data generated by TYPHON. The file generated must be opened in Excel and saved as txt file with tab as separators. The new file is then directly read by ECEXTRACTOR when launched. The format of the file is a two dimensional array with $N_{\text{total nodes}} + 1$ columns where the last column is used to store the time point.
2. The second file is a .txt tab delimited matrix containing the adjacency matrix of the directed graph. No header should be added in the file.

1.2 Output files

Output files generated by ECEXTRACTOR are as follows:

SUMMARY-SCC-CYCLES :

- column 1: number of disjoint parts in the graph: for the computation of the strongly connected components, if a graph has several completely disconnected parts then the Tarjan algorithm must be applied on each separated part. This variable is equal to the number of separated parts found for a given time point. It gives an indication on how the nodes interacts together.
- column 2: contains the number of strongly connected components obtained using the Tarjan algorithm.
- column 3: It contains the number of elementary cycles found using the Tiernan algorithm on each strongly connected components
- column 4: It contains the number of positive elementary cycles

- column 5: It contains the number of negative elementary cycles
- column 6: it is the label of the time point

SUMMARY-AUTO-REGULATIONS :

- column 1: it is the label of the time point
- column 2: it contains the number of positive auto regulations
- column 3: it contains the number of negative auto regulations

TIME-EVOLUTION-AUTO-REGULATIONS : This file contains a $M * N$ matrix where M is the number of time points and N is the number of nodes in the directed graph. A given entry is set to 1 if there is a positive diagonal term for the corresponding node, to -1 if the diagonal term is negative and set to 0 if there is no diagonal term.

RESULTS-LIST-ECS : this file contains the topology of all elementary cycles.

ECS-TEMPORAL-EVOLUTION This file contains a two dimensional array. The first column is the list of time points. The remaining columns are for all elementary cycles and auto-regulations. Each time a elementary cycle is present for a given time point the corresponding entry is set to 1.

simulation-general-information this file contains information about the duration of the simulation, number of time points used and date when the execution was performed.

LIST-SCREENSHOT-ID this file contains the ID of each screenshot taken during the simulation. More precisely it is a string with three numbers: first one is the time point, second is the total number of SCCs and last one is the number of active SCCs[1]. This file is used to generate the appropriate complete name of each screen shot so loading can be done automatically.

DIGRAPH-SCC it gives the detailed list of SCCs for each selected timepoint. Each row corresponds to one cycle. last row give the number of SCCs and the row just before gives the length of each SCC starting from the first one.

DIRECTED-GRAPH-EDGES-LIST list of the edges of the adjacency matrix. column A is the column label and column B is the row label of the adjacency matrix. third column is the sign of the interaction (-1 or 1) and last column is set to 1 if the interaction is effective and to 0 otherwise (it means that the node in the column A is actually not active when the picture is taken.)

2 R-script "R-SCRIPT-STEP-2"

This code takes various inputs and processes them. The code automatically selects the right directory where the files are located. The R-script can be executed by simple copy paste within the R working environment.

- Input Files:
  1. **TIME-SERIES-DATA.txt**
  2. **LIST-SCREENSHOT-ID**
  3. **RESULTS-LIST-ECS.csv**

[1] active SCCs are SCCs whose all involved nodes are presents in the system.

4. **ECS-TEMPORAL-EVOLUTION.csv**
5. **SUMMARY-SSC-CYCLES.csv**
6. **TIME-EVOLUTION-AUTO-REGULATIONS.csv**
7. **SUMMARY-SSC-CYCLES.csv**

- Output Files:

1. **FINAL-RESULTS-LIST-ECS.csv** It is the complete list of elementary cycles appearing across all time points. For each cycle, the first column contains the following information. the unique label of the cycle (the following columns at the same row contain the sign of each edge), the number of nodes included in the cycle, the sign of the cycle. The following columns contain the label of the nodes ordered according to the edge formed within the cycle. Each cycle is separated from the next one by a row of 0.
2. **FINAL-TEMPORAL-EVOLUTION-ECS.csv** The file gives the temporal changes for the presence of the elementary cycles over the time points. The number of rows is equal to the number of time points selected by the user for the analysis. The first column is the label of the time point and the remaining columns correspond to the elementary cycles with the exclusion of auto regulations. Each time a cycle appears, it is written 1 in the corresponding entry and 0 otherwise.
3. **Matrice-B**(.txt tab delimited) It is a square matrix whose dimension is equal to the total number of elementary cycles M found in the system during its complete temporal evolution. Each line corresponds to a elementary cycle and should be understood as follows. if an entry , say i, j is equal to one, it means that the elementary cycle i and j are disjoint. Thus all non zero entries in the first line give the list of elementary cycles which form a disjoint set with the elementary cycle number one. But this property is always checked two by two, thus if $1, 3$ and $1, 4$ are non zero it means that EC $1, 3$ and $1, 4$ form two disjoint sets but EC $3, 4$ are not necessarily disjoint.
4. **Tableau-C**](.txt tab delimited) This rectangular matrix of size $N+1, M+1$. The first column starts with a zero and the following N lines contain the label of a dynamical state. The M remaining columns corresponds to the elementary cycles. the first line is the length of the elementary cycle. then for each line corresponding to a state, the entry is set to one if this cycle appears in this state or set to zero otherwise.

3 NUCLEIEXTRACTOR

Assuming that the list of ECs is known, this code aims at computing the eventual nuclei of selected configuration of the associated dynamical system. The design of the searching method of this algorithm is based on the definition of a nucleus, i.e. a set of *disjoint* ECs containing all the active nodes of the directed graph.

3.1 Input data

The code requires two distinct input files which are built using the R script. This script also takes in charge the generation of two source files for NUCLEIextractor.

1. Matrice-B
2. Tableau-C

3.2 Output files

The code produces two distinct output files:

DETAILED-LIST-OF-NUCLEI This file contains the complete list of nuclei for the selected time points. The first column gives the time point, the second column gives the number of activated nodes for this time point. This number also corresponds to the sum of the elementary cycles lengths contained in the nucleus. The third column is the number n_{ec} of disjoint elementary cycles in the nucleus. the remaining n_{ec} occupied columns contain the label of each elementary cycle. The nuclei of all time points are listed without interruption. Note that when the auto regulations are included within the nucleus, the labeling is adapted as follows. if there are L elementary cycles of size 2 and greater, then the set of auto regulations are labeled starting with the number $L + 1$ which corresponds to the node 1.

TOTAL-NUMBER-OF-NUCLEI-PER-TIMEPOINTS This is a two dimensional array. The number of lines corresponds to the number of time points selected for the analysis. For each line, there are two columns: the first one is the label of the time point as saved in the output files produced by ECextractor. The second column gives the number of nuclei found for this time point.