Statistical Data Analysis 2 – Final Project

November 2025

Project description

The purpose of this project is to investigate how the type and amount of data describing network dynamics influence the accuracy of inferring network structure within the framework of Bayesian networks.

The first part of the project focuses on constructing various Boolean networks, simulating them under asynchronous and synchronous update modes to generate data, using the BNFinder2 software tool (https://bioputer.mimuw.edu.pl/software/bnf/) to build dynamic Bayesian networks from the simulated data, and finally assessing the quality of the reconstructed network structures.

In the second part of the project, the insights gained from the first part will be applied to reconstruct the structure of a network model representing a real-life biological process.

Part I

The detailed tasks of the first part of the project are as follows.

1. Construct several Boolean networks with sizes (measured by the number of nodes or variables) ranging from 5 to 16.[†] Each node should have no more than three parent nodes, and the Boolean functions governing individual nodes should be generated at random.

2. Simulate trajectories of the generated networks in both synchronous and asynchronous modes to create datasets. The datasets should vary in (i) the proportion of transient and attractor states, (ii) the trajectory sampling frequency (i.e. the number of time steps between consecutive sampled states), and (iii) their overall size (i.e. the number and length of trajectories used to construct each dataset); see Example 1 below for further explanations.

3. Use the datasets to infer **dynamic Bayesian networks** with the BNFinder2 software tool. Unlike classical Bayesian networks, dynamic Bayesian networks can contain cycles, making them more suitable for this task. Consider two scoring functions (referred to as scoring criteria in the BNFinder2 terminology): Minimal Description Length (MDL) and BDe (Bayesian–Dirichlet equivalence).

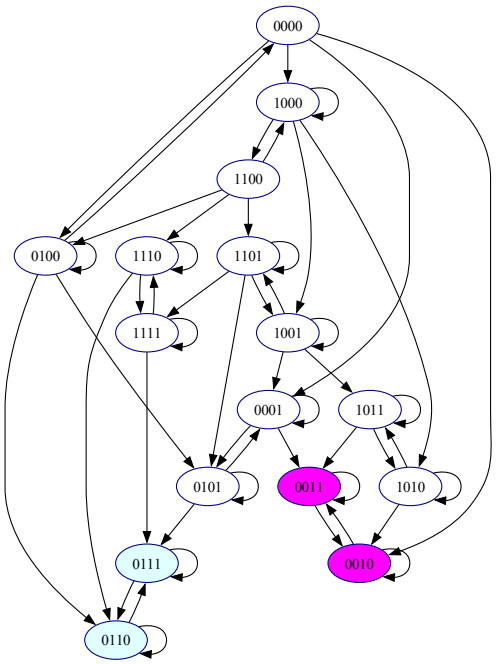
4. Evaluate the accuracy of the reconstructed network structures with respect to the characteristics of the datasets and the scoring functions used. Consider the original

[†]Not all sizes within this range need to be considered; it suffices to generate networks of sizes 5, 16, and a few intermediate sizes. The maximum size is suggested in view of the computational challenges associated with attractor detection and Bayesian network inference, but you may consider larger systems if your computational resources and attractor detection implementation allow it.

Boolean network (i.e. the one used to generate the dataset) as the ground truth. The evaluation should employ at least two structure-based graph distance measures of your choice, with a brief justification for the selected measures.

Example 1. *Consider a Boolean network with four nodes* x_1, x_2, x_3, *and* x_4 *and the following associated Boolean functions:* $f_1 = \neg x_2 \wedge \neg x_3 \wedge \neg x_4$, $f_2 = (x_2 \wedge x_3) \vee (\neg x_2 \wedge \neg x_3)$, $f_3 = \text{TRUE}$, *and* $f_4 = \neg x_3 \vee \neg x_4$. *The asynchronous state transition system of the BN with two highlighted complex attractors is shown below. Starting from state* 0000, *a trajectory of length 15 is simulated:* 0000 → 0100 → 0100 → 0000 → 0001 → 0001 → 0001 → 0101 → 0001 → 0001 → 0011 → 0010 → 0010 → 0011 → 0011 → 0010. *This entire trajectory can be considered one dataset. Some examples of other possible datasets are as follows:*

- *A dataset obtained by sampling the trajectory with frequency 3:* 0000 → 0000 → 0001 → 0001 → 0010 → 0010.
- *A dataset with a large proportion of transient states:* 0000 → 0100 → 0100 → 0000 → 0001 → 0001 → 0001 → 0101 → 0001 → 0001 → 0011.
- *A dataset with a large proportion of attractor states:* 0101 → 0001 → 0001 → 0011 → 0010 → 0010 → 0011 → 0011 → 0010
- *A dataset consisting of two trajectories, i.e. the one above* 0000 → 0100 → 0100 → 0000 → 0001 → 0001 → 0001 → 0101 → 0001 → 0001 → 0011 → 0010 → 0010 → 0011 → 0011 → 0010 *and another starting from the state* 1000*:* 1000 → 1000 → 1100 → 1110 → 0110 → 0110 → 0110 → 0110 → 0110 → 0110 → 0111 → 0111 → 0111 → 0111.



Part II

In the second part of the project, your task is to consider a validated Boolean network model of a real-life biological mechanism. To this end, select a Boolean network

model of your choice from the 'models' subfolder of the Biodivine repository, available at https://github.com/sybila/biodivine-boolean-models, with the number of nodes (variables) not exceeding 16.[‡]. Using the insights gained from the first part of the project, generate an appropriate dataset for the network inference task and reconstruct the network structure with BNFinder2, applying a scoring function chosen based on your previous experience. Evaluate the accuracy of the reconstruction.

Technical remarks

- A tutorial on using BNFinder2 is available at:

  https://bioputer.mimuw.edu.pl/software/bnf/bnfinder_tutorial.pdf.

- Examples of inferring dynamic Bayesian networks from time-series data can be found in Section 1.1 of the tutorial.

- Installation instructions for BNFinder2 are provided in the User's Manual, available at https://bioputer.mimuw.edu.pl/software/bnf/bnfinder_manual.pdf. Note that BNFinder2 requires Python 2 and is not compatible with Python 3.

Project report

You should prepare and submit a comprehensive report describing your experimental setup, the datasets generated, and the results obtained. The report should include the specifications of all Boolean networks considered in your study, along with the corresponding input files containing the simulated datasets provided to BNFinder2. It should justify all methodological choices made and present the results both in written form and through appropriate graphical representations. Finally, the conclusions drawn from your experiments should be clearly articulated.

Project realisation

The project should be carried out in groups of three-four Students. You are free to create teams across lab groups. The submitted project report should contain a clear statement of the contributions of individual team members.

Good luck! :)

[‡]As before, the maximum size is suggested with the computational challenges in mind, but you may consider larger systems.