TASK 1: Set up and reading

a) Visit http://bayesian-ai.eecs.qmul.ac.uk/bayesys/
b) Download the Bayesys user manual.
c) Set up the NetBeans project by following the steps in Section 1 of the manual.
d) Read Sections 2, 3, 4 and 5 of the manual.
e) *Skip* Section 6.
f) Read Section 7 and repeat the example.
    i. *Skip* subsections 7.3 and 7.4.
g) Read Section 8 and repeat the example.
h) *Skip* Sections 9, 10, 11 and 12.
i) Read Section 13.
    i. *Skip* subsection 13.6.

TASK 2: Determine research area and prepare data set

You are free to choose or collate your own data set. As with Coursework 1, we recommend that you address a problem you are interested in or related to your professional field. If you are motivated by the subject matter, the project will be more fun for you, and you will likely perform better.

Data requirements:

- **Size of data:** The data set must contain at least 8 variables (yes, penalty applies for using <8 variables). There is no upper-bound restriction on the number of the variables. However, we recommend using <50 variables for the purposes of the coursework to make it much easier for you to visualise the causal graph, and to save computational runtime. While the vast majority of submissions typically rely on relatively small data sets that take a few seconds to 'learn', keep in mind some algorithms might take hours to complete learning when given more than 100 variables!
    i. You do not need to use a special technique for feature selection – it is up to you to decide which variables to keep. We will not be assessing feature selection decisions.
    ii. There is no sample-size restriction and you are free to use a part of the samples. For example, your data set may contain millions of rows and you may want to use fewer to speed-up learning.

- **Re-use data from CW1:** You are allowed to reuse the data set you have prepared for Coursework 1, as long as: a) you consider that data set to be suitable for causal structure learning (refer to Q1 in Section 3), and b) it contains at least 8 variables.

- **Bayesys repository:** You are not allowed to use any of the data sets available in the Bayesys repository for this coursework.

- **Categorical data:** Bayesys assumes the input data are categorical or discrete; e.g., {"low","medium","high"}, {"yellow","blue","green"}, {"<10","10-20","$20+$"} etc, rather than a continuous range of numbers. If your data set contains continuous variables, Bayesys will consider each value of a continuous variable as a different category. This will cause problems with model dimensionality, leading to poor accuracy and high runtime (if this is not clear why, refer to the Conditional Probability Tables (CPTs) covered in the lectures).

  To address this issue, you should discretise all continuous variables to reduce the number of states to reasonable levels. For example, a variable with continuous values ranging from 1 to 100 (e.g., {"14.34","78.56","89.23"}) can be discretised into categories such as {"1to20","21to40","41to60","61to80","81to100"}. Because Coursework 2 is not concerned with data pre-processing, you are free to follow any approach you wish to discretise continuous variables. You could discretise the variables manually as discussed in the above example, or even use k-means which we covered in previous lectures, or any other data discretisation approach. We will not be assessing data discretisation decisions.

- **Missing data values:** The input data set must not contain missing values/empty cells. If it does, the easiest solution would be to replace ALL empty cells with a new category value called *missing* (or use a different relevant name). This will force the algorithms to consider missing values as an additional state. Alternatively, you could use any data imputation approach, such as MissForest. We will not be assessing data imputation decisions.

Once you ensure your data set is consistent with what has been stated above, rename your data set to *trainingData.csv* and place it in folder *Input*.

TASK 3: Draw out your knowledge-based graph

1. Use your own knowledge to produce a knowledge-based causal graph based on the variables you decide to keep in your data set. Remember that this graph is based on *your* knowledge, and it is not necessarily correct or incorrect. You will compare the graphs learnt by the different algorithms with reference to your knowledge graph.

   You may find it easier if you start drawing the graph by hand, and then record the directed relationships in the *DAGtrue.csv* file. In creating your *DAGtrue.csv* file, we recommend that you edit one of the sample files that come with Bayesys; e.g., create a copy of the *DAGtrue_ASIA.csv* file available in the directory *Sample input files/Structure learning*, then rename the file to *DAGtrue.csv*, and then replace the directed relationships with those present in your knowledge graph.

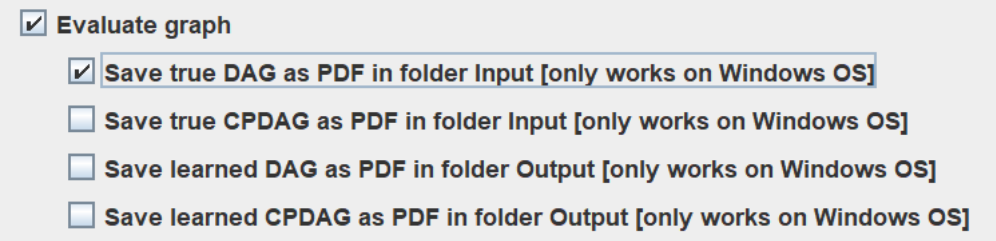
   NOTE: Your knowledge graph should have a maximum node in-degree of 11; i.e., no node in the graph should have more than 11 parents (this is a library/package restriction).

2. Once you are happy with the graph you have prepared, ensure the file is called *DAGtrue.csv* and placed in folder *Input*.

   NOTE: If your OS is not showing the file extensions (e.g., .CSV or .PDF), name your file *DAGtrue* and not *DAGtrue.csv*; otherwise, the file might end up being called *DAGtrue.csv.csv* unintentionally (when the file extension is not visible). If this happens, Bayesys will be unable to locate the file.

3. Make a copy of the *DAGtrue.csv* file, and rename this copy into *DAGlearned.csv* and place it in folder *Output*. You can discard the copied file once you complete **Task 3**.

4. Ensure that your *DAGtrue.csv* and *trainingData.csv* (from **Task 2**) files are in folder *Input*, and the *DAGlearned.csv* file is in folder *Output*. Run Bayesys in NetBeans. Under tab *Main*, select *Evaluate graph* and then click on the first subprocess as shown below. Then hit the *Run* button found at the bottom of tab *Main*.



   The above process will generate output information in the terminal window of NetBeans. Save the last three lines, as highlighted in the Fig below; you will need this information later when answering some of the questions in Section 3.

Additionally, the above process should have generated one PDF files in folder *Input* called *DAGtrue.pdf*. Save this file as you will need it for later.

```
SHD score [CPDAG]: 0.000
DDM score [CPDAG]: 1.000
BSF score [CPDAG]: 1.000
# of independent graphical fragments: 1 (includ
_____                Inference-based evaluation
LL for graph [log2]: -32348.864
BIC score [log2] -32468.454
# of free parameters 18
BUILD SUCCESSFUL (total time: 6 seconds)
```

This only concerns MAC/Linux users: The above process might return an error while creating the PDF file, due to compatibility issues. Even if the system completes the process without errors, the PDF files generated may be corrupted and not open on MAC/Linux. If this happens, you should use the online GraphViz editor to produce your graphs, available here: https://edotor.net/ , which converts text into a visual drawing. As an example, copy the code shown below in the web editor:

```
digraph {
    Earthquake -> Alarm
    Burglar -> Alarm
    Alarm -> Call
}
```

If you are drawing a CPDAG containing undirected edges, then consider:

```
digraph {
      Earthquake -> Alarm
      Burglar -> Alarm
      Alarm -> Call [arrowhead=none];
}
```

You can then edit the above code to be consistent with your *DAGtrue.csv*. You could copy-and-paste the variable relationships (e.g., Earthquake → Alarm) directly from *DAGtrue.csv* into the code editor, taking care to remove commas and quote any variable names containing spaces.

TASK 4: Perform structure learning

1. Run Bayesys. Under tab *Main*, select *Structure learning* and algorithm *HC* (default selection). Select *Evaluate graph* and then click on the last two (out of four) options so that you also generate the learned DAG and CPDAG in PDF files, in addition to the *DAGlearned.csv* file which is generated by default. Then, hit the *Run* button.

2. Once the above process completes, you should see:

   i. Relevant text generated in the terminal window of NetBeans.

   ii. The files *DAGlearned.csv*, *DAGlearned.pdf* and *CPDAGlearned.pdf* should be generated in folder *Output*. As stated in **Task 3**, the PDF files may be corrupted on MAC/Linux, and you will have to use the online GraphViz editor to produce the graph corresponding to *DAGlearned.csv* (simply copy the relationships from the CSV file into the editor as discussed in **Task 3**).

3. Repeat the above process for the other four algorithms; i.e., TABU, SaiyanH, MAHC and GES. Save the same output information and files that each algorithm produces (ensure you first read the NOTE below).

   NOTE: As stated in the manual, Bayesys overwrites the output files every time it runs. You need to remember to either rename or move the output files to another folder before running the next algorithm.

   Similarly, if you happen to have one of the output files open – for example, viewing the *DAGlearned.pdf* in Adobe Reader while running structure learning - Bayesys will fail to replace the PDF file, and the output file will not reflect the latest iteration. Ensure you close all output files before running structure learning.