Assignment 1

Network Analysis 2022

Recap

Question 1 (3 points)

Are the following statements true or false (0.5 point per statement)? Explain why.

1. To apply network models, one needs to assume that a network theory is true.
2. Within network models a link between nodes A and B represent a causal relationships between them.
3. In a PMRF, the absence of a link between nodes represents conditional independence between the corresponding variables and the presence of an edge represents conditional dependence.
4. An undirected network contains a total number of possible edges $m = n(n-1)/2$, with n being the number of nodes.
5. PMRFs may tell us something about the underlying causal structure between the variables.
6. The `minimum` argument in the `qgraph` function removes edges with a value under the indicated value.

Network Appraoches, Theory, and Models

Question 2 (2 points)

Write a short scientific essay on the network theory of mental disorders, answering the following questions. What is the external field of a mental disorder and how do you think this would impact it in the common cause framework and in the network framework? What is an implication that follows from the network perspective towards the diagnosis and treatment of mental disorders? Make sure to include a proper reference list, if you choose to use references (max 300 words).

Question 3 (2 points)

Choose a paper which makes use of network analysis. Summarize the paper and explain the authors' choice for performing network analysis (if you don't agree with the choice, explain why) (max 250 words).

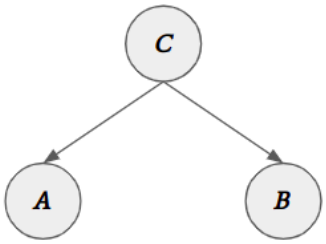
Question 4 (2 points)

Is it worth the hype? Critically reflect on the network approach to psychopathology. Think of one (conceptual or methodological) limitation of the network approach/network analysis in Psychology (max 250 words).

Causality, Conditional Independece & PMRFs

Question 5 (1 point)

For the indicated graph, explain the dependency structure at hand. What happens if variable C is observed?

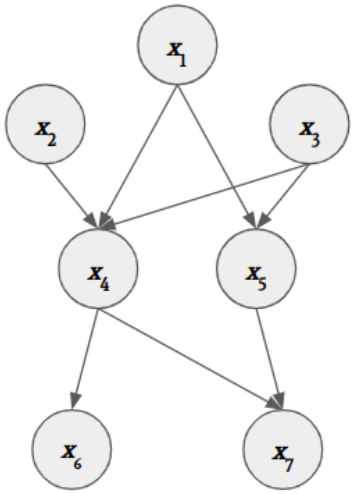


Question 6 (1 point)

For the indicated graph below first indicate for each node x_k its parent nodes pa_k. The joint distribution of a graph with K nodes can be expressed by:

$$p(x) = \prod_{k=1}^{K} p(x_k \mid pa_k)$$

where pa_k denotes the set of parents of x_k. Using this formula, write down the joint distribution $p(x_1, ..., x_7)$ for the graph below.



Question 7

Let's get started!

```
# Start up your R and enter the following commands
datafile <- read.table(file='http://borsboomdenny.googlepages.com/datafile.txt')
```

```
attach(datafile)
source('http://borsboomdenny.googlepages.com/program.txt')

# look at the datafile - should contain 5000 cases for six variables
# (genes, fingers, smoke, try, susceptible, cancer)
View(datafile)

# Compute the correlation between the variables genes and smoke - 0.1277
cor(genes,smoke)
```

```
## [1] 0.1277674
```

The dataset

The dataset consists of data from 5000 hypothetical subjects on six variables, each of which has two values:

1. `cancer` whether a person has cancer. 0=no, 1=yes.
2. `try` how often a person has tried cigarettes during adolescence. 0=not often, 1=often.
3. `genes` whether a person has good or bad genes. 0=good, 1=bad.
4. `smoke` whether a person is a smoker. 0=no, 1=yes.
5. `susceptibility` whether a person is susceptible to smoking addiction. 0=no, 1=yes.
6. `fingers` whether a person has yellow-stained fingers. 0= no, 1=yes.

Your instruments

The model was generated using a DAG. Your job is to figure out how the arrows run. You can use two instruments for this purpose: first, you can check whether any two variables are independent, and second, you can check whether any two variables are conditionally independent, given a third. This works as follows:

```
# CHECKING INDEPENDENCE: To check if two variables X1 and X2 are independent,
# you type ind(X1,X2).
ind(smoke, cancer)
```

```
##
## *Contingency Table*
##
##     var2
## var1    0    1
##    0 3032  353
##    1 1315  300
##
## *Fitting Model*
##
## 2 iterations: deviation 5.684342e-14
## 2 iterations: deviation 0
##
## *Predicted frequencies under Independence*
##
##     var2
## var1        0        1
##    0 2942.919  442.081
```

```
##    1 1404.081  210.919
##
##
## *Results*
## Likelihood-Ratio ( df = 1 ) 60.96208
## p-value= 5.818484e-15
##
##
## *Conclusion*
## Independence of Var1 and Var2 does not hold.
## ************************************
##
##
```

```
# Interpreting the output: This returns the contingency table for smoke and
# cancer, the expected contingency table for smoke and cancer under independence,
# and a test of the null hypothesis that the variables are independent in the
# population. If p<.05, then the program concludes that the variables are
# dependent; otherwise that they are independent.
```

```
# CHECKING CONDITIONAL INDEPENDENCE: To check whether two variables X1 and X2 are
# conditionally independent, given X3, you type cind(var1=X1, var2=X2, blocker=X3).
# blocker is the variable you condition on.
cind(var1=smoke, var2=cancer, blocker=try)
```

```
##
## *Contingency Table*
##
## , , blocker = 0
##
##     var2
## var1    0    1
##    0 2517  289
##    1  573  124
##
## , , blocker = 1
##
##     var2
## var1    0    1
##    0  515   64
##    1  742  176
##
##
## *Fitting Model*
##
## 2 iterations: deviation 1.136868e-13
## 2 iterations: deviation 5.684342e-14
##
## *Predicted frequencies under CI*
##
## , , blocker = 0
##
##     var2
## var1          0          1
##    0 2475.17556  330.82444
```

```
##    1  614.82444   82.17556
##
## , , blocker = 1
##
##     var2
## var1          0          1
##    0  486.17435   92.82565
##    1  770.82565  147.17435
##
##
##
## *Results*
## Likelihood-Ratio ( df = 2 ) 45.65795
## p-value= 1.217591e-10
##
##
## *Conclusion*
## Conditional independence of var1 and var2, given blocker, does not hold.
```

```
# Interpret the output: This returns separate contingency tables for smoke and
# cancer for the two values of try,the expected contingency tables for smoke and
# cancer under conditional independence given try, and a test of the null hypothesis
# that smoke and cancer are conditionally independent of try in the population.
# If p < .05, then the program concludes that the variables are conditionally
# dependent given the blocker; otherwise that they are independent.
```

Question 7.1 (point)

Consider the three vairables `fingers`, `cancer` and `smoke`.

A) Are `fingers` and `cancer` independent?

B) Are `smoke` and `cancer` conditionally independent given `fingers`?

C) Are `fingers` and `cancer` conditionally independent given `smoke`?

D) Are `smoke` and `fingers` conditionally independent given `cancer`?

E) Do any additional checks you want. Which causal paths are consistent with the data for these three variables?

Question 7.2 (1 point)

Consider the three vairbales `smoke`, `try` and `susceptible`.

A) Are `susceptible` and `try` independent?

B) Are `susceptible` and `try` conditionally independent given `smoke`?

C) Do any additional checks you want. Which causal paths are consistent with the data for these three variables?

Question 7.3 (1 point)

Consider the three variables `try`, `smoke` and `cancer`.

A) Are `try` and `cancer` independent?

B) Are `try` and `cancer` conditionally independent given `smoke`?

C) Do any additional checks you want. Which causal paths are consistent with the data for these three variables?

Question 7.4 (1 point)

Consider the three variables `genes`, `susceptible` and `smoke`.

A) Are `smoke` and `genes` independent?

B) Are `smoke` and `susceptible` conditionally independent given `genes`?

C) Are `genes` and `smoke` conditionally independent given `susceptible`?

D) Do any additional checks you want. Which causal paths are consistent with the data for these three variables?

Question 7.5 (1 point)

Do any additional checks you want on the full dataset.Then draw your best guess of the DAG that created the data. Indicate the evidence you have for your DAG, as well as the evidence that speaks against it (if you have any). Also indicate for which parts of your DAG you have no conclusive evidence and explain why.

Remember that the data are simulated from the DAG, and therefore we have sampling errors: not all implied (in)dependencies in the generating DAG are necessarily detectable in this particular sample. Also, sometimes you may use sensibble theory to fix some arrows, but be sure to mention it. In general: use your brain as well as the statistics.

Visualizing Networks

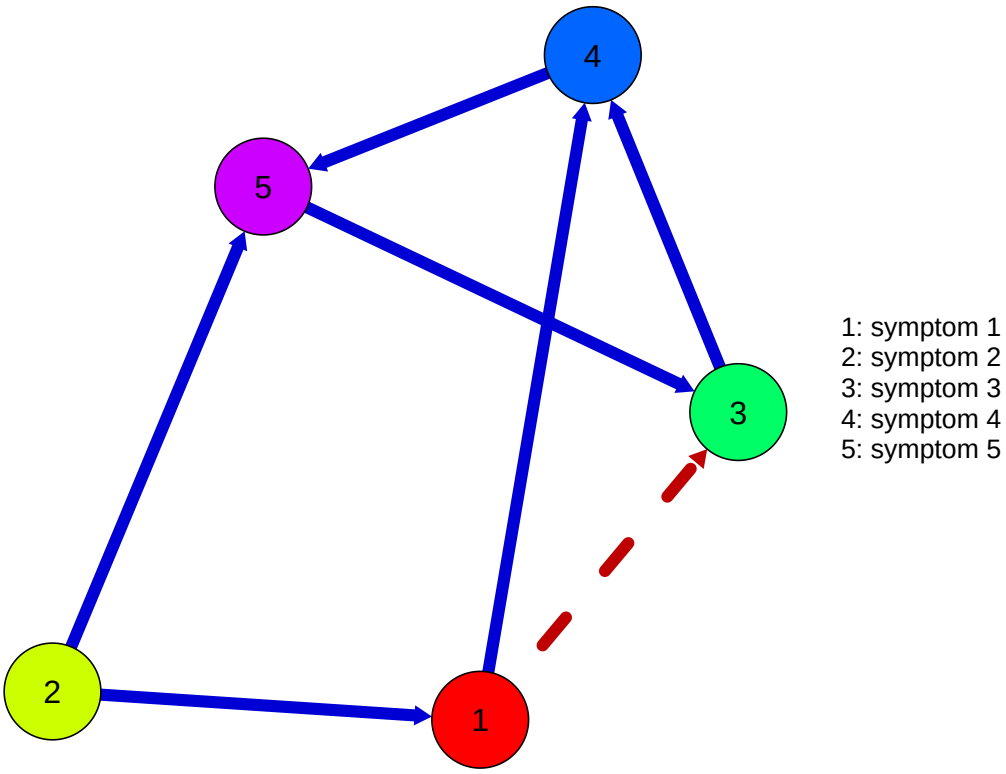
```
# Make sure to install qgraph
install.packages("qgraph")
```

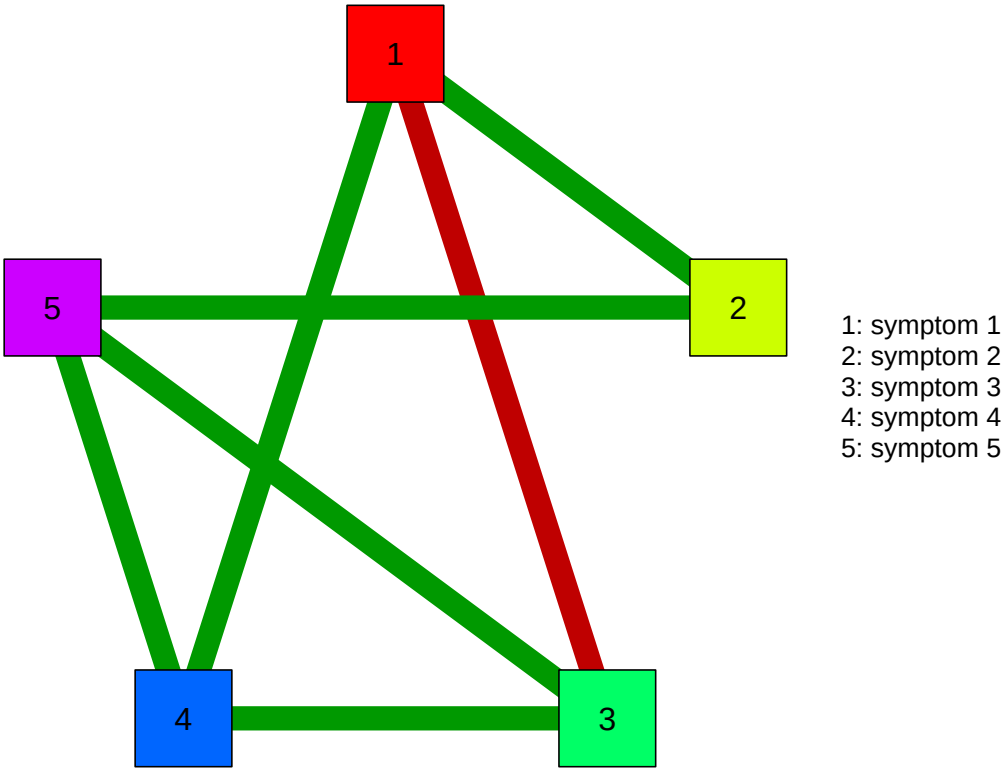
```
## Installing package into '/home/emily/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(qgraph)
```

Question 8 (2 points)

Recreate the following networks as close as possible.





Question 9 (2 points)

Time to be creative – Find a data set online to construct a (social) network (*see* e.g., http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/). **Important** – the data must already be in the right format for the `qgraph` function, e.g., weight matrix or edge list. Visualize the network in R and write a short report (maximum 300 words) describing what the network represents and how it can be interpreted. Most importantly, explain why you visualized it the way you did – which is the focus of this assignment, so really think about how important information can best be visualized within a network.