# Dublin R Workshop on Probabilistic Graphical Models: Bayesian Networks

Mick Cooney
michael.cooney@applied.ai

Nov 19, 2015

`https://bitbucket.org/kaybenleroll/dublin_r_workshops`.

Code is available in the `wspgm201511` directory.

Content in this workshop is based on the book Graphical Models with R by Søren Højsgaard.

Also look at the vignettes for the packages `gRain` and `gRbase`

Remember that this topic is massive. I could easily give a full semester course on this stuff to really do it justice, so most of this workshop is just me working through the material as I learn it.

As a result, it is highly likely this worksheet and code contains typos, errors, logical flaws and other mistakes in need of correction in this workshop, so if you note any, please let me know so I can try to fix them!

If you want to look into this topic more, there is an old Coursera course by Daphne Koller (tough going but excellent):

`https://www.coursera.org/course/pgm`

This course was based on her textbook Probabalistic Graphical Models: Principles and Techniques

# 1. Introduction

A graph is a mathematical object that can be defined as a pair $\mathcal{G} = (V, E)$, where $V$ is a set of *vertices* or *nodes*, and $E$ is a set of *edges* that joins two vertices. Edges in general may be directed, undirected or bidirected. They are typically visualised by using shapes or points for the nodes and lines for the edges.

The concept of *conditional independence* is related to that of *statistical independence*. Suppose we have three random variables $A$, $B$ and $C$, then $A$ and $B$ are *conditionally independent* given $C$, writen $A \perp B \,|\, C$, iff, for every given value $c$ in $C$, $A$ and $B$ are independent in the conditional distribution given $C = c$.

Another way of saying this is that for some $f$ a generic density or probability mass function, then one characteristic of $A \perp B \,|\, C$ is that

$$f(a, b \,|\, c) = f(a \,|\, c) f(b \,|\, c).$$

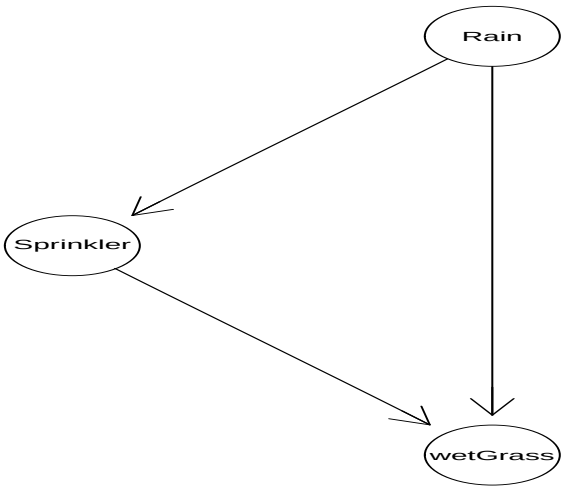
An equivalent characterisation is that the joint density of $A$, $B$ and $C$ factorises as

$$f(a, b, c) = g(a, c)\, h(b, c).$$

Finally, we will also make heavy use of Bayes' Rule, the standard formula for relating conditional probabilities:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

## 2. The Sprinkler Network




Two events can cause grass to be wet: Either the sprinkler is on or it is raining. Rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on).

This can be modeled with a Bayesian network. The variables (R)ain, (S)prinkler, Wet(G)rass have two possible values: (y)es and (n)o.

We can factorise the joint probability mass function as

$$p_{GSR}(g, s, r) = p_{G|SR}(g|s, r)p_{S|R}(s|r)p_R(r)$$

or overloading the notation a little:

$$P(G, S, R) = P(G|S, R)P(S, R) = P(G|S, R)P(S|R)P(R)$$

This means we can construct the joint probability table by starting with the *conditional probability tables* (CPTs).

**Exercise 2.1** Create the 3 CPTs using the `parray` function and the following conditional probabilities:

$$P(R) = 0.2$$
$$P(S|R) = 0.01 \qquad P(S|\neg R) = 0.4$$
$$P(G|S, R) = 0.99 \qquad P(G|S, \neg R) = 0.9 \qquad P(G|\neg S, R) = 0.8 \qquad P(G|\neg S, \neg R) = 0$$

**Exercise 2.2** Calculate the full joint probability function $P(G, S, R)$. *HINT:* The function `tabListMult()` might be of use.

**Exercise 2.3** Calculate the probability that it is raining given that the grass is wet. *HINT:* The functions `tabMarg()` and `tabDiv()` may be of use.

# 3. Genetic Inheritance

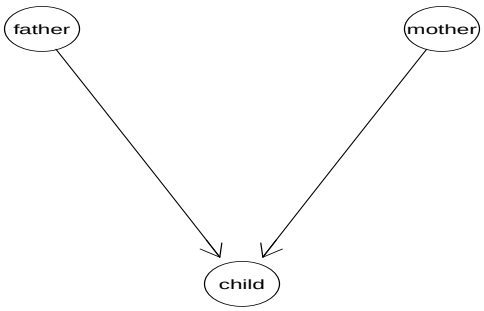
We now turn our attention to analysing genetic inheritance on the chromosomes for a given DNA sequence.

An *allele* is the DNA sequence at a marker and can take two values marked $A$ or $B$ (in practice there can be 10 or 20 different values).

A *genotype* is an unordered pair of alleles: $AA$, $AB$, or $BB$.

The genotype of a person at a specific marker is a random variable with state space $\{AA, AB, BA\}$.

We are interested in the joint distribution of genotypes for a group of people.




A child inherits one allele from each parent independently.

The parents two alleles have equal probability of being passed on to the child.

Each combination has probability 0.25; some lead to the same genotype for the child.

```
##    child AA_AA AA_AB AA_BB AB_AA AB_AB AB_BB BB_AA BB_AB BB_BB
## 1:    AA     1   0.5     0   0.5  0.25   0.0     0   0.0     0
## 2:    AB     0   0.5     1   0.5  0.50   0.5     1   0.5     0
## 3:    BB     0   0.0     0   0.0  0.25   0.5     0   0.5     1
```

So in this case we have the the joint probability distribution as being

$$p(m, f, c) = p(m)\, p(f)\, p(c \mid m, f)$$

**Exercise 3.1** Assuming the population frequency of alleles $A$ and $B$ is 0.3 and 0.7 respectively, calculate the distribution of the genotypes. *HINT:* You probably want to work with the binomial distribution for this, `dbinom()`.

**Exercise 3.2** Construct the probability tables for the three nodes in the chart. *HINT:* Look at the function `cptable()` for this.

**Exercise 3.3** Build the Bayesian network and plot it out.

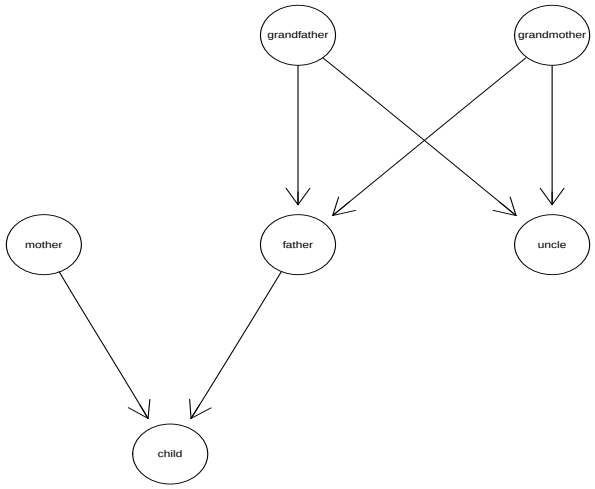
**Exercise 3.4** What is the marginal distribution of the father's genotype?

**Exercise 3.5** What is the joint distribution of mother and child?

**Exercise 3.6** What is the conditional joint distribution of the father, given values for mother and child?

**Exercise 3.7** A mother with a genotype $BB$ has a child with genotype $AB$. Given that a man has genotype $AB$, how can we determine if the man is likely to be the child's father?

**Exercise 3.8** Construct the Bayesian network as seen below.




**Exercise 3.9** Suppose the 'father' is not willing to give a sample but his brother is, and tests $AA$. What is the probability of observing this evidence if the man is in fact the father?

# 4. The Chest Clinic Example

We now move up to a bigger example, the Chest Clinic example as discussed in Lauritzen and Spiegelhalter (1988):

> Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

**Exercise 4.1** Breaking down the above paragraph into discrete facts, construct a network graph that captures the relationships described.

**Exercise 4.2** Load the dataset `chestSim500` and create the Bayesian network using this data. What is the unconditional probability of an individual having lung cancer according to this network?

**Exercise 4.3** Given that we know the individual has visited Asia and has dyspnoea, what is the conditional probability now that the person has lung cancer?

**Exercise 4.4** Repeat the above process using the dataset `chestSim1000`. How much do the probabilities change?

**Exercise 4.5** Repeat the above process using the datasets for all the simulated data in the `gRbase` package. Does this have much of an effect on the outputted probabilities? *HINT:* Use the R command `data()` to find all available datasets from a package.

**Exercise 4.6** Given the above datasets, what is the marginal probabilities of the three diseases mentioned above? (Lung cancer, Tuberculosis and Dyspnoea)

# 5. Scaling the Networks

All of the above approaches are example of the 'Brute Force' approach which is done by calculating the full joint distribution for the network $p(V)$ as a multiple of the CPTs that comprise it.

$$p(V) = p(a)\,p(t|a)\,p(s)\,p(l|s)\,p(b|s)\,p(e|t,l)\,p(d|e,b)\,p(x|e)$$

This gives $p(V)$ represented by a table with $2^8 = 256$ entries.

We can then marginalise and condition as desired to calculate whatever probabilities we need.

This scales appalling badly. A network with 80 variables, each with 10 values has a joint probability space of $10^{80}$, approximately the count of atoms in the universe.

We are going to need a bigger boat...

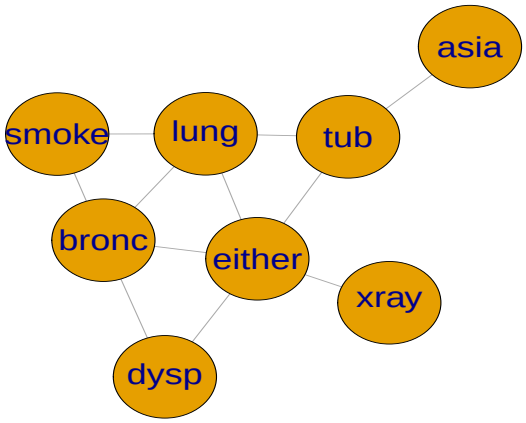
So, we need a way to not need the full joint distribution, instead focusing on the the low dimensional CPTs and send 'messages' between them.

To use a network it first needs to be *compiled* and then *propagated*. Compilation of a network based on CPTs is first *moralised* — edges are added between the parents of each node, and then directed edges are replaced with undirected ones. It is then *triangulated* to form a triangulated graph.

The CPTs are transformed into *clique potentials* defined on the cliques of the chordal graph.

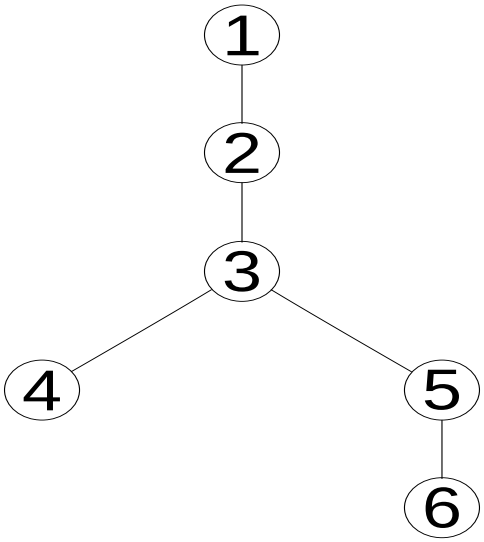
We can see this process below:

```
chestclinic.dag <- dag(list(
    "asia"
   ,c("tub", "asia")
   ,"smoke"
   ,c("lung", "smoke")
   ,c("bronc", "smoke")
   ,c("either", "lung", "tub")
   ,c("xray", "either")
   ,c("dysp", "bronc", "either")
    ));


chestclinic.moralized    <- moralize(chestclinic.dag);
chestclinic.triangulated <- triangulate(chestclinic.moralized);

iplot(chestclinic.triangulated);
```




Once we have this DAG created, we also want to see how the triangulated data looks as a junction tree. The messages passed between connected nodes on the graph involve the common variables in the nodes, and propagating the information just involves a double pass down and up the tree.

```
## cliques
##   1 : asia tub
##   2 : either lung tub
##   3 : either lung bronc
##   4 : smoke lung bronc
##   5 : either dysp bronc
##   6 : either xray
## separators
##   1 :
##   2 : tub
##   3 : either lung
##   4 : lung bronc
##   5 : either bronc
##   6 : either
## parents
##   1 : 0
##   2 : 1
##   3 : 2
##   4 : 3
##   5 : 3
##   6 : 5
```




**Exercise 5.1** Create the `grain` object from the `chestSim500` data.

**Exercise 5.2** Find the marginal, joint and conditional probability for lung cancer and bronchitis given that the person recently visited Asia and displayed symptoms of dyspnoea

**Exercise 5.3** Suppress the automatic propagation of the data and redo the code, but this time adding the evidence one piece at a time.