Uncertain Reasoning and Data Mining, year 2013-2014

Assignment 1

Suttipong Mungkala

---

1. Structure of the network

1.1 Objective

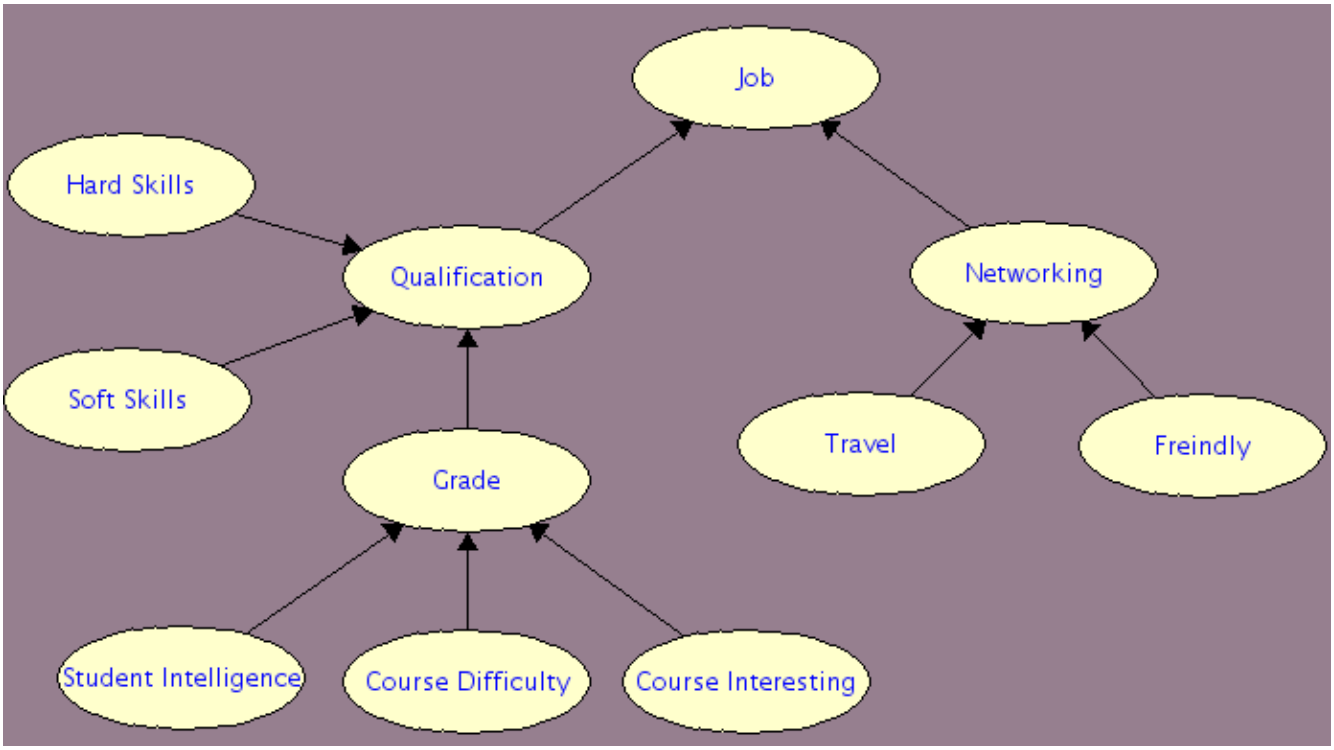
Why do recent graduate students have difficulty finding jobs? This has been the big question for ages. Students complain about finding jobs, employers equally complain about finding qualified applicants. And a significant of that what students believe employers want and what employers want turn out to be very different things.

Since I had four years working experiences and an opportunity to discuss with my former boss about what mattered to hiring managers. For instance, hard skills, soft skills and grades are extremely important to employers and personal connections are also important as people prefer to work with someone they know and like. The interesting point is the interaction between these factors, how they influence each other can be represented with the Bayesian Networks (BN). I hope this model can help students to estimate their chance in getting a job and what things they should improve in order to enhance their qualifications.

1.2 Variables

Before I constructed the dependencies in the BN for this problem, I started simplifying what factors that weighed to employers in a job application process for new graduate students. I defined 11 variables and for each one, how many states are defined and its arcs orientation is already self explanation.

The following BN is not the only model that represents the problem, there are other models. For instance, students who are brighter taking harder courses in which case, there might be potentially an edge between **Course Difficulty** and **Student Intelligence** but I do not use that model. Because of the fact that a model is not set in stone, it is a representation of how I believe the how the world works.



- **Grade** {c, b, a} and its influencing nodes **Student Intelligence** {low, high}, **Course Difficulty** {easy, hard} and **Course Interesting** {no, yes}

  The first variable is **Grade (G),** the grade of the student obviously depends on how difficult the course that he or she is taking, the interesting of the course and the intelligence of the student. So that gives us in addition to **G**, we have **Course Difficulty (CD)**, **Course Interesting (CI)** and **Student Intelligence (SI)** variable.

  The grade is considered as a grade point average (GPA) and is assigned as a range from *a* to *c,* where *a* is considered excellent, *b* is good and *c* is acceptable. The course is assumed to be the entire course of the student's study program. The difficulty of the course is clearly considered *easy* and *hard*, and the course interesting is considered *no* and *yes.*

- **Soft Skills** {no, yes} **Hard Skills** {no, yes}

  In addition to the student's grade, **Soft Skills (SS)** and **Hard Skills (HS)** are the significant factors that matter to hiring managers. The states of both nodes are considered as *no* and *yes* because I think the first sight that hiring managers can see is whether a candidate meets the minimum skills requirements needed for a particular job.

- **Qualification** {bad, average, good}

  It turns out that the overall qualification is influenced by the student's **Grade**, **Soft Skills** and **Hard Skills**. This is general agreement among hiring managers and students. The qualification

is basically relevant and tied to job duties, for instance, requires hard skills in JAVA programming, UNIX operating systems, Data Mining, and so on. If the student can fulfill all requirements he or she is considered as *good* qualification, otherwise *average* or *bad* qualification as well as soft skills (e.g. communication, teamwork, etc).

- **Networking** {no, yes} and its influencing nodes **Travel** {no, yes}, **Friendly** {no, yes}

  Many openning jobs filled by word of mouth. That's why networking is one of the best ways to find a job. But networking isn't about using other people or aggressively promoting yourself, it's about building relationships. The more you travel the more you meet new people, and if you are friendly you have better chance to gain your networking. As a result, these three variables are binary.

- **Job** {no, yes} - the decision node

  Finally, the chance of the student in getting a job obviously depends on the student's qualification and networking of the student. Given the knowledge of these two nodes (**Qualification** and **Networking**), it tells us the probability that the student will get the job.

  To sum up, we simplify this problem by basically binarizing every variables, except for the **Grade** and **Qualification.**

1.3 D-Separation

1.3.1 Course Difficulty and **Student Intelligence** are D-Separated if the **Grade** is not observed, and they are D-Connected when it is observed.

It is obviously that knowing the difficulty of the course {easy, hard} does not tell anything about the intelligence of the student as well as given the knowledge of student's intelligence does tell us how difficulty of the course he or she is taking. However, if we have information about the student's grade (**Grade** is observed), we are able to know how intelligence he/she is, since the grade is one of the evaluation methods that used in all education systems.

1.3.2 Grade and **Soft Skills** are D-Separated if the **Qualification** is not observed, and they are D-Connected when it is observed.

The grade tells how good student in studying and doing the exam, but it does not give the information about the student's soft skills (e.g. self-confidence, flexibility/adaptability). This sounds really make sense to me. From hiring managers perspective, one way to understand the student's soft skills is to have an interview, evaluate him/her and classify it as *{good, average, bad}* qualification. Thus, knowing the student's qualification (**Qualification** is observed) can tell us about his/her soft skills.

1.3.3 Travel and **Friendly** are D-Separated if the **Networking** is not observed, and they are D-Connected when it is observed.

People who like travelling does not mean they are friendly. This is so true based on my experience, while I was travelling there were a lot of people whom I met, they just loved travelling themselves and did not want to make connections with strangers. But once we have knowledge about their networking (**Networking** is observed), we can realize how friendly they are according to their networking information.

1.3.4 Hard Skills and **Soft Skills** are D-Separated if the **Qualification** is not observed, and they are D-Connected when it is observed.

This scenario can tell us the relations between hard skills and soft skills. Hard skills are teachable abilities or skill sets that are easy to quantify (e.g. computer programming skills). Soft skills, on the other hand, are subjective skills that are much harder to quantify (e.g teamwork). It is obviously seen that they have nothing in common, in which they are D-Separated in the network. During the job application and interview process, employers can check applicants on these two skill sets and categorise it. Once the **Qualification** is observed, you can see that the good qualification will have both hard skills and soft skills (almost the same %) while the bad qualification will have a big different number of probability between its states {yes, no} meaning they lack in one of these two skills.

2.) Conditional probability tables (CPTs)

- **Job** P(J | Q, N) conditional probability table - **Job | Qualification, Networking**

Qualification	bad		average		good	
Networking	no	yes	no	yes	no	yes
no	0.9	0.8	0.4	0.3	0.2	0.05
yes	0.1	0.2	0.6	0.7	0.8	0.95

The table was filled with the information retrieved from the internet, for example the students who have good qualification and networking will get a job but there is still a little chance to be denied by a company in case they are considered overqualified applicants, the students are possible to get a job even though they have bad qualification and no networking as it depends on what kind of job and which company they are applying for.

- **Friendly** P(F) and **Travel** P(T) prior probability table

P(F)

no	0.2
yes	0.8

P(T)

no	0.3
yes	0.7

The information here is based on my experience where most of people I have met, they are friendly and like travelling.

- **Networking** P(N| F, T) conditional probability table - **Networking | Friendly, Travel**

Freindly	no		yes	
Travel	no	yes	no	yes
no	0.9	0.3	0.8	0.3
yes	0.1	0.7	0.2	0.7

As mentioned earlier in the previous section, The more you travel the more you meet new people, and if you are friendly you have better chance to gain your networking. The probability table was filled in based on this sentence. I also put 0.1 for **Friendly**{=no} and **Travel**{=no} as you could still have networking from people/friends you already had, as well as **Friendly**{=yes} and **Travel**{=yes} there is possibility not to find or gain any networking as new people you meet just do not want to keep in touch with you.

- **Student Intelligence** P(SI), **Course Difficulty** P(CD) and **Course Interesting** (CI) prior probability table

P(SI)

low	0.7
high	0.3

P(CD)

easy	0.5
hard	0.5

P(CI)

no	0.5
yes	0.5

The information is based on what I have experienced during my studies, I have observed that about 30% of the total number of students are considered as high intelligence and the half of the course is difficult and interesting.

- **Grade** P(G | SI, CD, CI) conditional probability table - **Grade | Student Intelligence, Course Difficulty, Course Interesting**

Student Int...	low				high			
Course Dif...	easy		hard		easy		hard	
Course Int...	no	yes	no	yes	no	yes	no	yes
a	0.15	0.2	0.05	0.1	0.8	0.9	0.45	0.55
b	0.5	0.55	0.4	0.3	0.15	0.09	0.3	0.35
c	0.35	0.25	0.55	0.6	0.05	0.01	0.25	0.1

The grade obviously depends on how intelligence of the student, the difficulty of the course he or she is taking, and the interesting of the course. For example, if the student is intelligence, the course is easy and the student interested in that course, he or she will (90% chance) get grade *"a"*, or else 80% chance in case the course is not interesting because the student might pay less attention to that course, the probability of getting grade "*b*" and "*c*" is 9% and 1% respectively. I then applied the same concept to fill in the remaining probabilities of the table.

- **Hard Skills** P(HS) and **Soft Skills** P(SS) prior probability table

P(HS)

no	0.5
yes	0.5

P(SS)

no	0.6
yes	0.4

For the initial values of these two tables, I searched for about 10 different jobs on the internet and checked its requirements regarding hard skills and soft skills. It turned out that, my qualification matched about half of hard skills and 40% of soft skills required on the job page. So, I came up with the above probability tables.

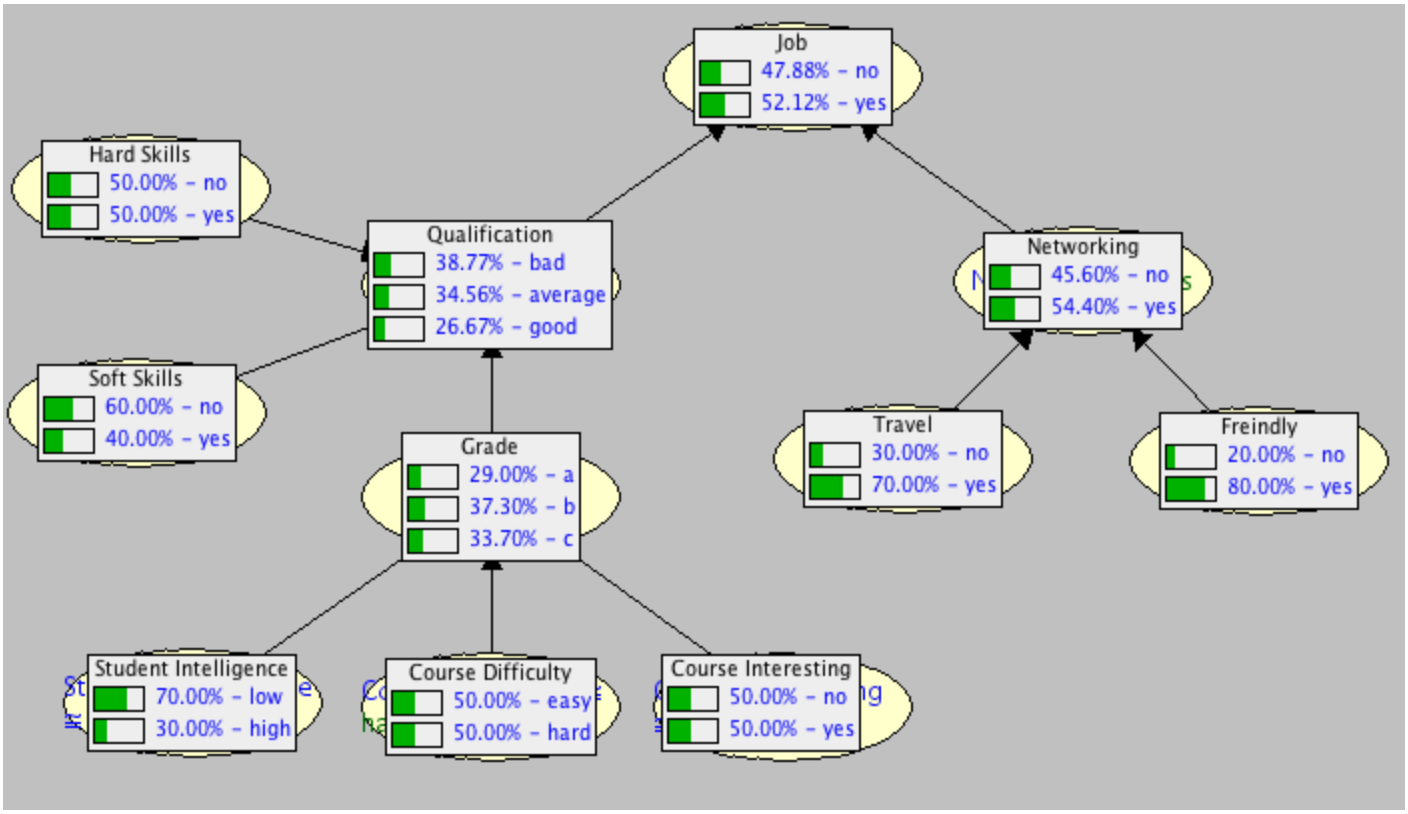
- **Qualification** P(Q | HS, SS, G) conditional probability table - **Qualification | Hard Skills, Soft Skills, Grade**

Grade	a				b				c			
Soft Skills	no		yes		no		yes		no		yes	
Hard Skills	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
bad	0.5	0.2	0.2	0.0	0.6	0.3	0.3	0.0	0.9	0.6	0.6	0.0
average	0.3	0.4	0.4	0.0	0.4	0.5	0.5	0.3	0.1	0.3	0.3	0.6
good	0.2	0.4	0.4	1.0	0.0	0.2	0.2	0.7	0.0	0.1	0.1	0.4

In order to explain which qualification is *good, average or bad,* I have searched information on the internet, what factors mattered to hiring managers?. Apart from student's academic background and hard skills (e.g. JAVA programming skills), soft skills such as communication, critical thinking, creativity and collaboration, as the area with the biggest gap. As a result, if the student lacks in soft skills and hard skills, his or her qualification **will not** be considered as "*good",* while lacking one of thse skills (either hard skills or soft skills) the qualification will likely be considered as *"average" or "bad"* depending on the grade as well.

3. Simulation

3.1 Comments on the simulation



Given the above BN without setting any evidence, I consider some major nodes and evaluate how realistic they are.

I first look at how **Student Intelligence, Course Difficulty** and **Course Interesting** influence **Grade**. They all make sense that the grade is generally distributed given half of the course difficulty is either *easy* or *hard* as well as course interesting is either *no* or *yes*, two third of students are not so intelligence.

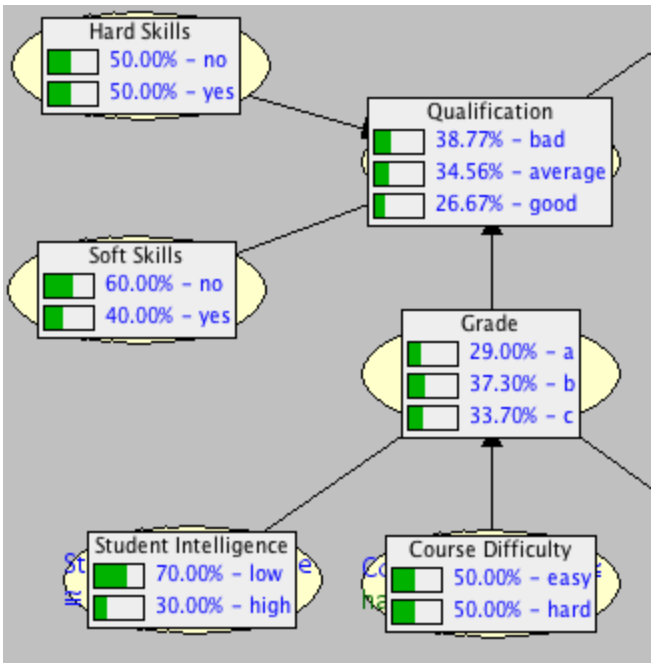
Qualification given the the knowledge of **Hard Skills**, **Soft Skills** and **Grade.** This is also realistic because once I had a chance to screen candidates with my former boss and we found out that the applications consist of good, average and bad qualifications, with a less number of good candidates. Moreover, a survey by staffing company Adecco mentioned that about 40% of respondents cited soft skills, such as communication, critical thinking, creativity and collaboration, as the area with the biggest gap.

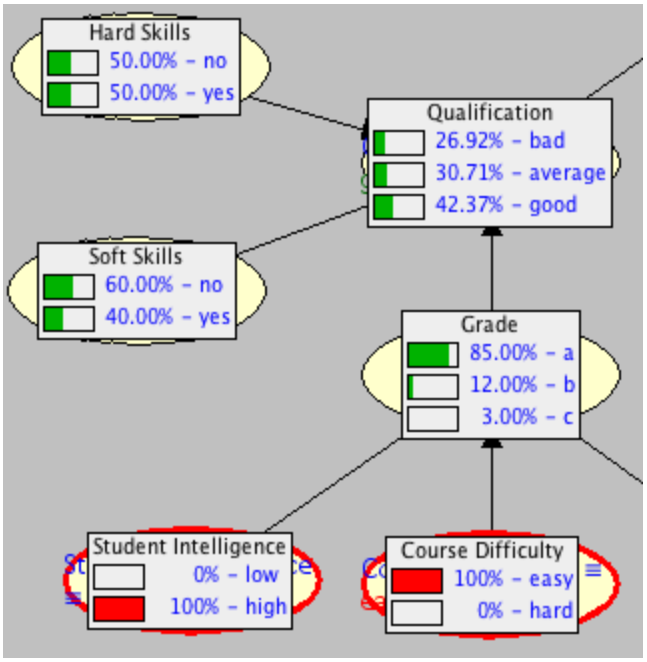
Finally, whether a student will get a **Job** obviously depends on how good of his or her **Qualification** and the **Networking** of the student. Without considering any other factors (e.g. economic situation), without setting the evidence, it is likely that the student is going to succeed in a job hunting with 52% of the probability.

3.1 Scenarios of Evidence

3.1.1. Course Difficulty {easy} and Student Intelligence {high}

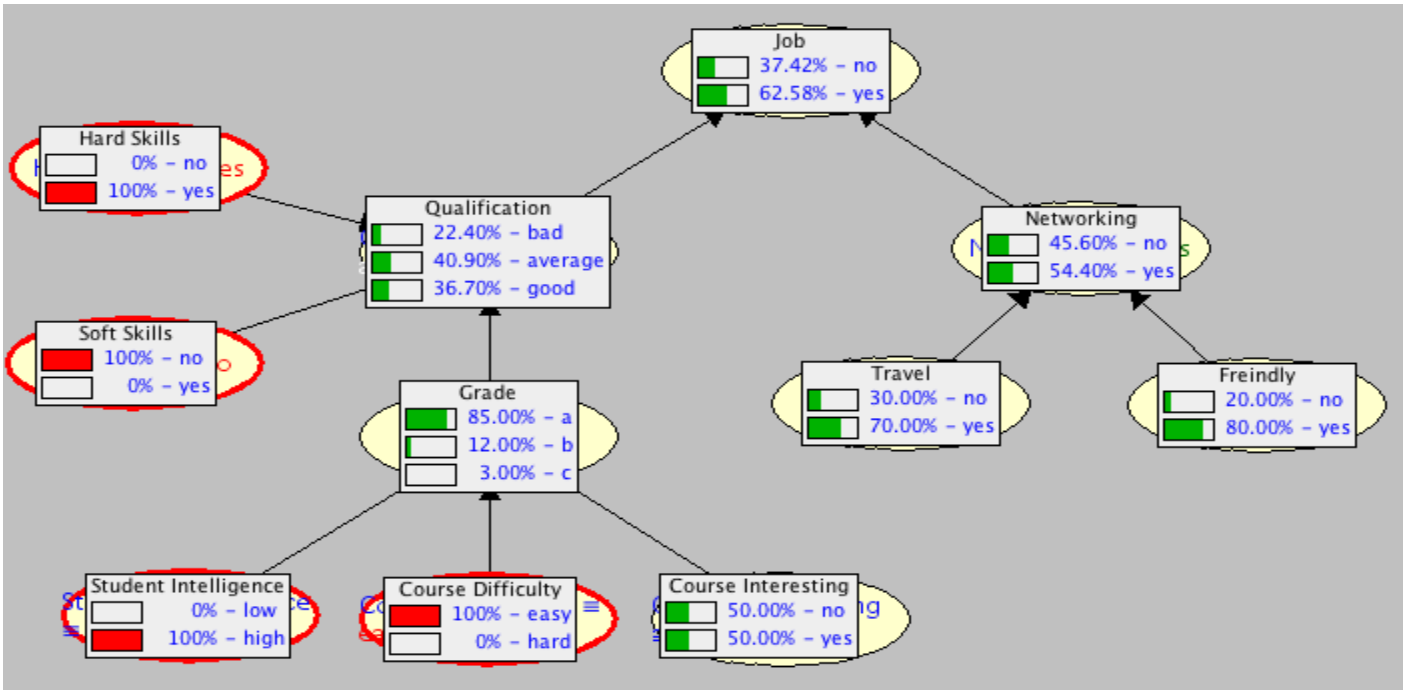
It is noticeable that the student will have 85% of probability to get the grade *"a"*, while it is possible that the student will get "b" and "c" as the outcome. This evidence is true in most cases. However, if we look at the model precisely the number of good **Qualification** is also changed, which could be difficult for hiring managers to distinguish whether the student's qualification is really good or just the student is taking the easy course. Student intelligence does not mean he or she can meet hard skills or soft skills that required for a workplace. For this reason, the other factors (**Hard Skills and Soft Skills**) need to be taken into account as well. You can see more detail in the next scenario.





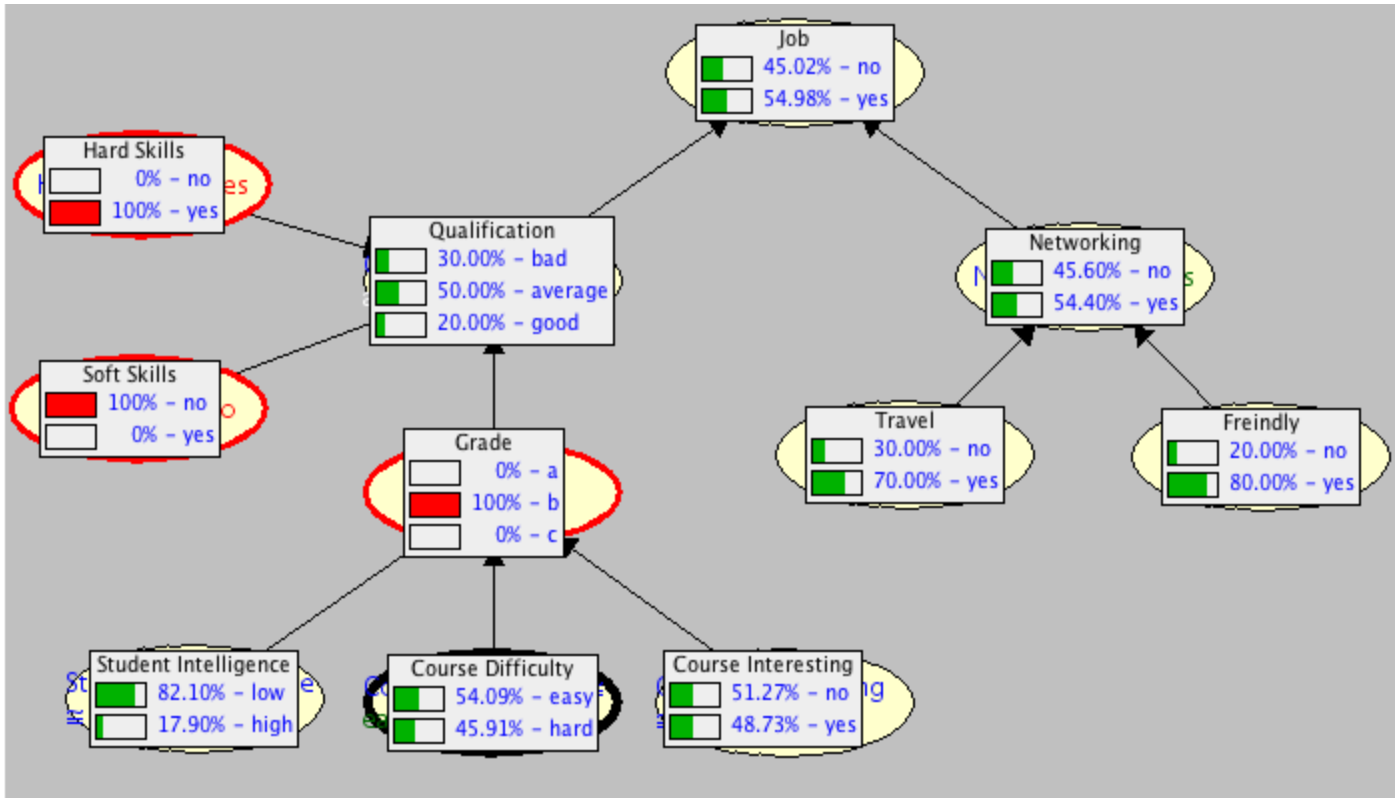
3.1.2. Course Difficulty {easy}, Student Intelligence {high}, Hard Skills {yes} and Soft Skills {no}

In order to reduce ambiguous information about the qualification, hard skills and soft skills also need to be observed. Having good grade and hard skills fulfillment do not mean the student's qualification will be considered as *good* , it could be considered as *average* and *bad* as shown below. This is due to the fact that the student is missing soft skills which are important for hiring managers. As a result, the student will still have about 60% of chance in getting a job which less probability in comparison to the previous scenario.



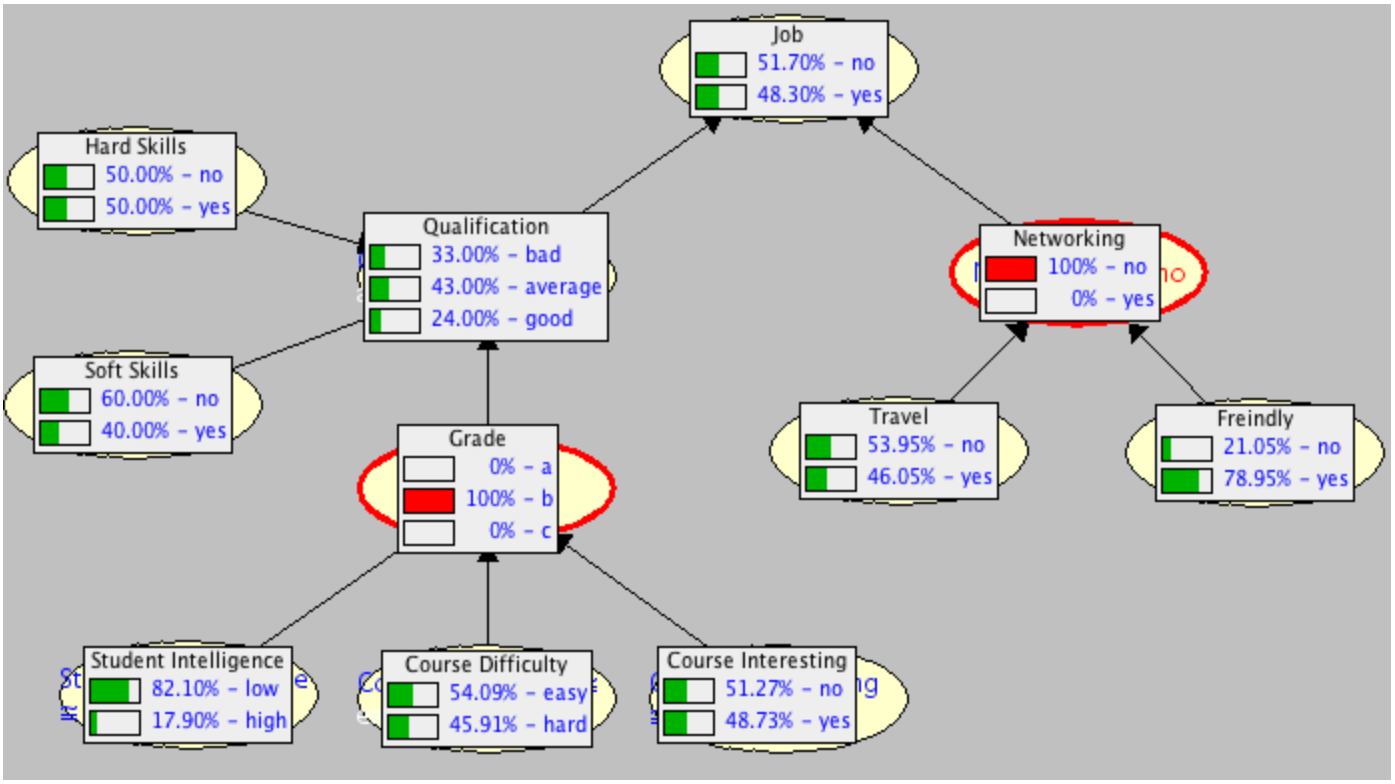
3.1.3 Grade {b}, Hard Skills {yes} and Soft Skills {no}

This scenario can also be difficult in taking decisions whether the student will get a job since his or her qualification is considered average and can only meet hard skills requirements. As a result, the chance in getting a job is about 50/50 as shown below.



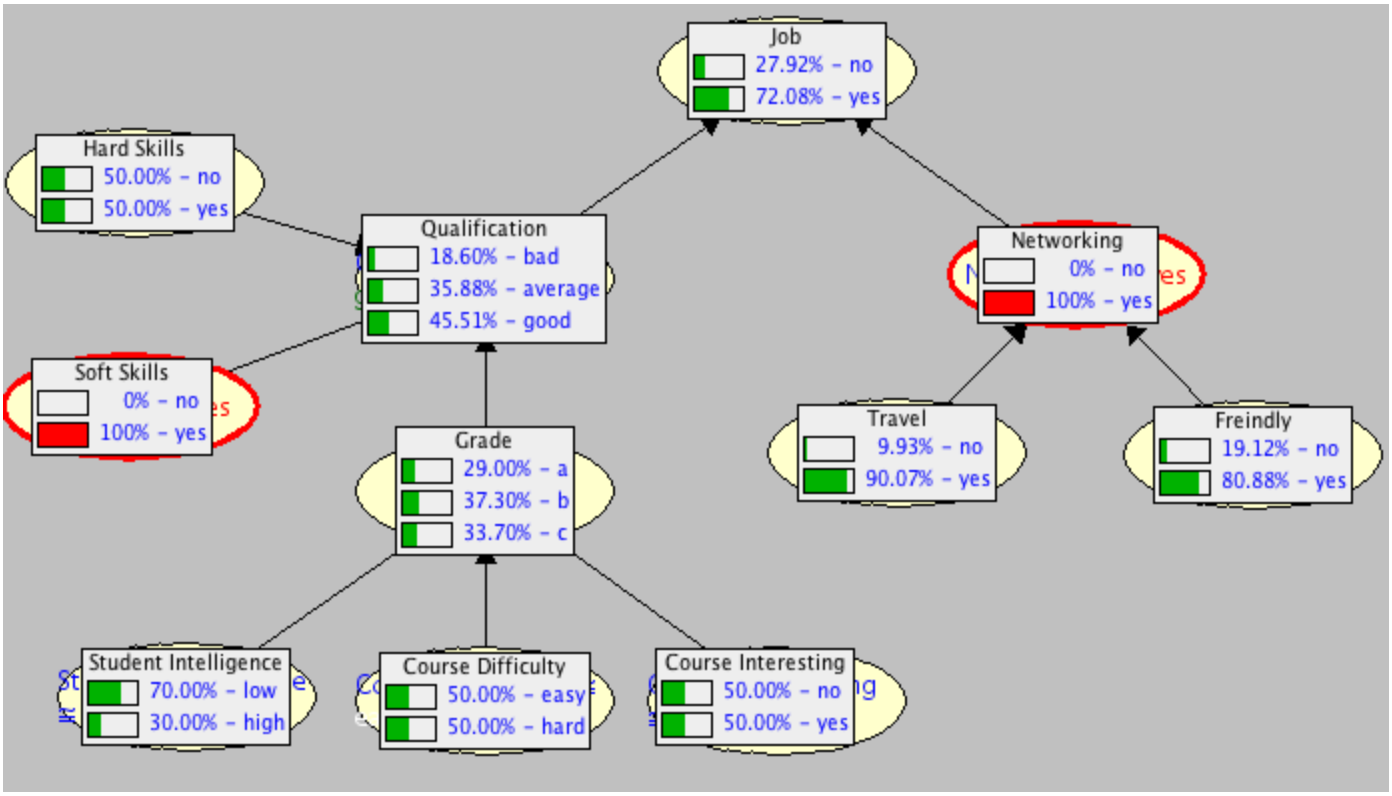
3.1.4. Grade {b} and Networking {no}

For this scenario, I tried setting the evidence in the networking {=no} node and the grade {=b} to see how it affects the posterior belief in the decision node. The job node represented about 50 of probability in getting a job, which is considered a difficult decision making scenario.



3.1.5. Soft Skills {yes} and Networking {yes}

I read many internet articles about employers saying that applications lack soft skills (e.g. communication and interpersonal skills). It is clear that the student who fulfills this requirement, has a better chance to get a job. Also, the networking is one of the best ways to find a job because people do business primarily with people they know and like, the job that matches your profile and you want may not be advertised at all. Networking leads to information and helps you convince employers to hire you. The following evidences demonstrate this fact. There was only 50% chance of getting a job before the **Soft Skills** and **Networking** were set. The network now shows that the student will have about 70% of chance in getting a job.



3.2 Conclusion on the scenarios

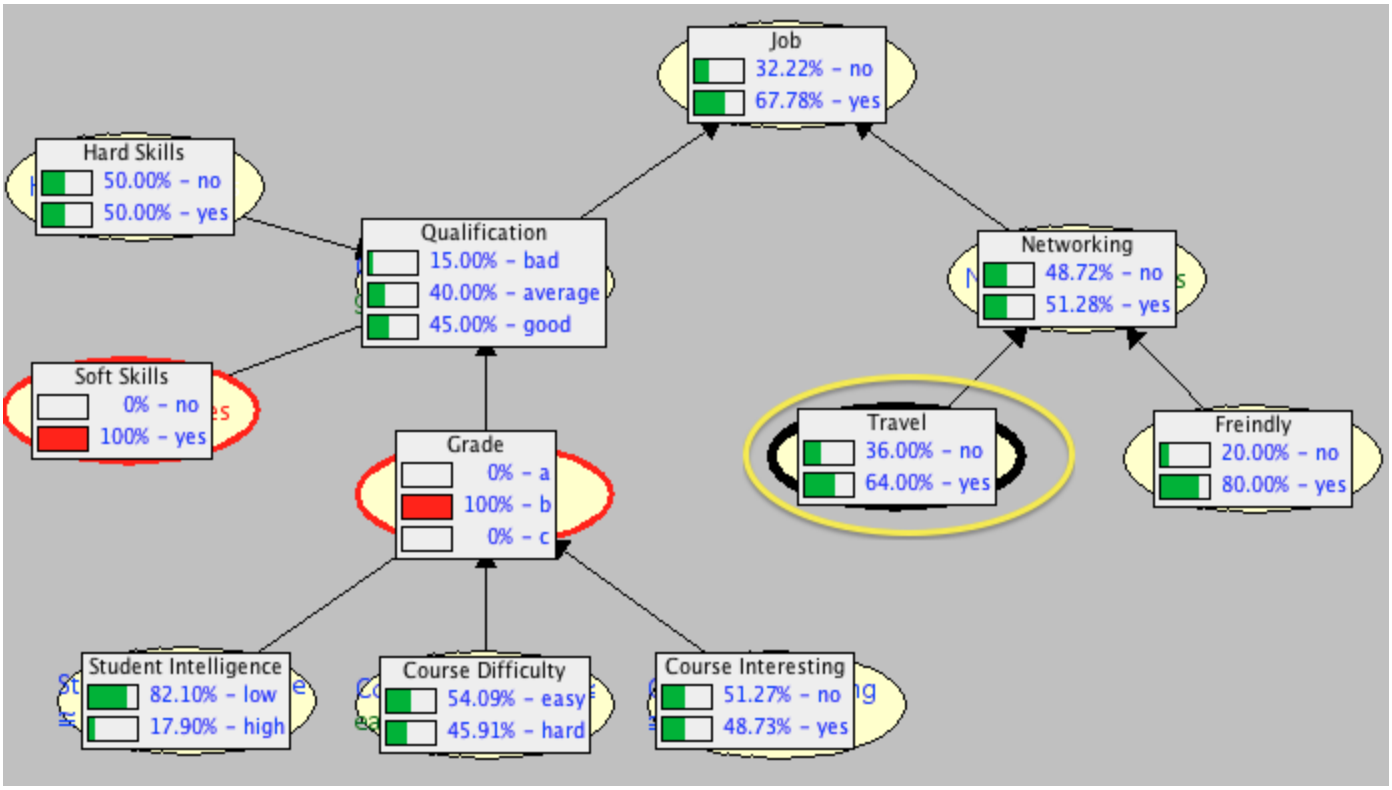
So far you have seen 5 different scenarios of evidence that related to different situations in which it could be hard in making a decision. The model works well in representing how the prior belief influences the posterior belief given some variables are observed.

3.3 Analysis of Sensitivity

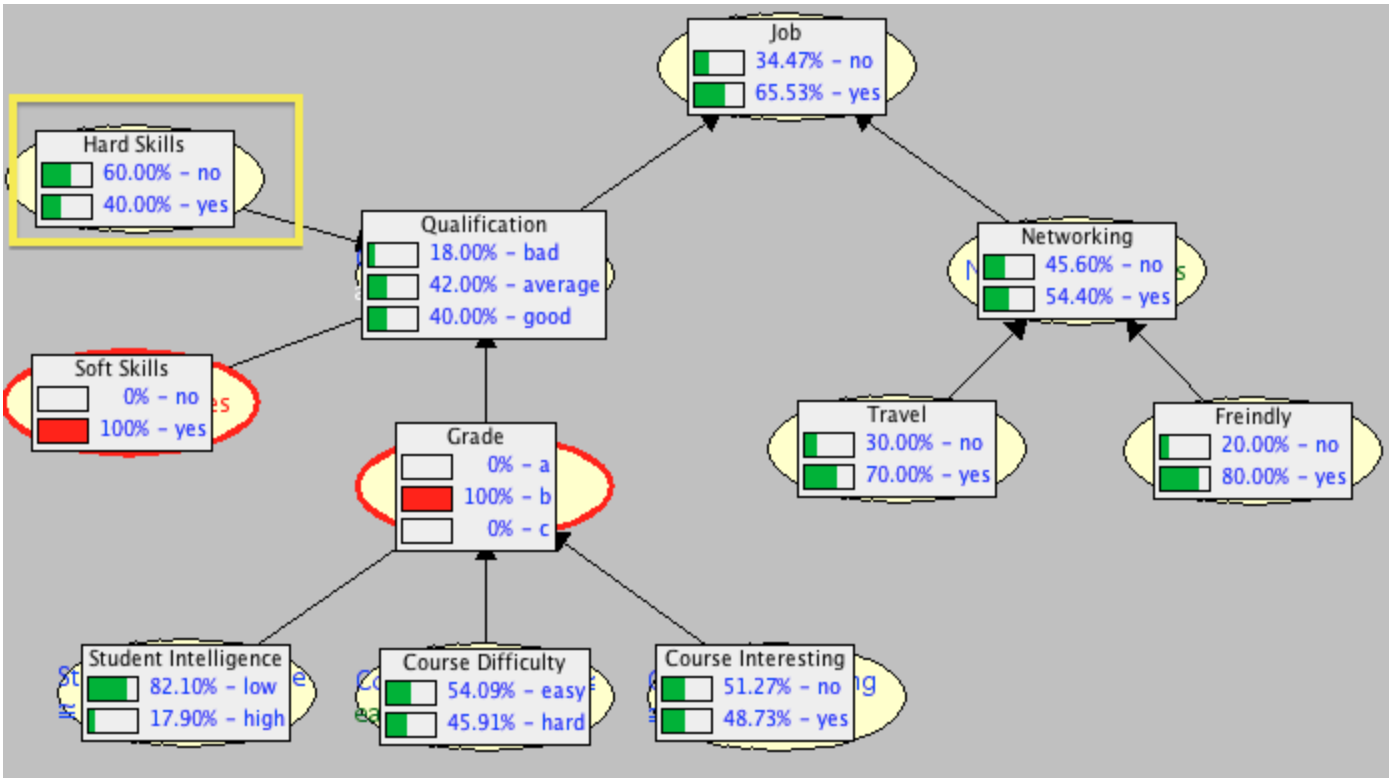
I set the following evidences to perform an analysis of sensitivity:
Soft Skills = {yes}
Grade = {b}

I did not only focus on the probability distribution of the decision node (**Job)** but also other nodes (e.g. Networking and Qualification). I estimated that the uncertainty about this distribution is not that much and Job's state would still have more probability to be "yes". I modified the following distributions (Travel and Hard Skills) by increasing of 20% its initial value and mainly observed the change occurred on all nodes

1.) Increase 20% of Travel {=no}



2.) Increase 20% of Hard Skills {=no}

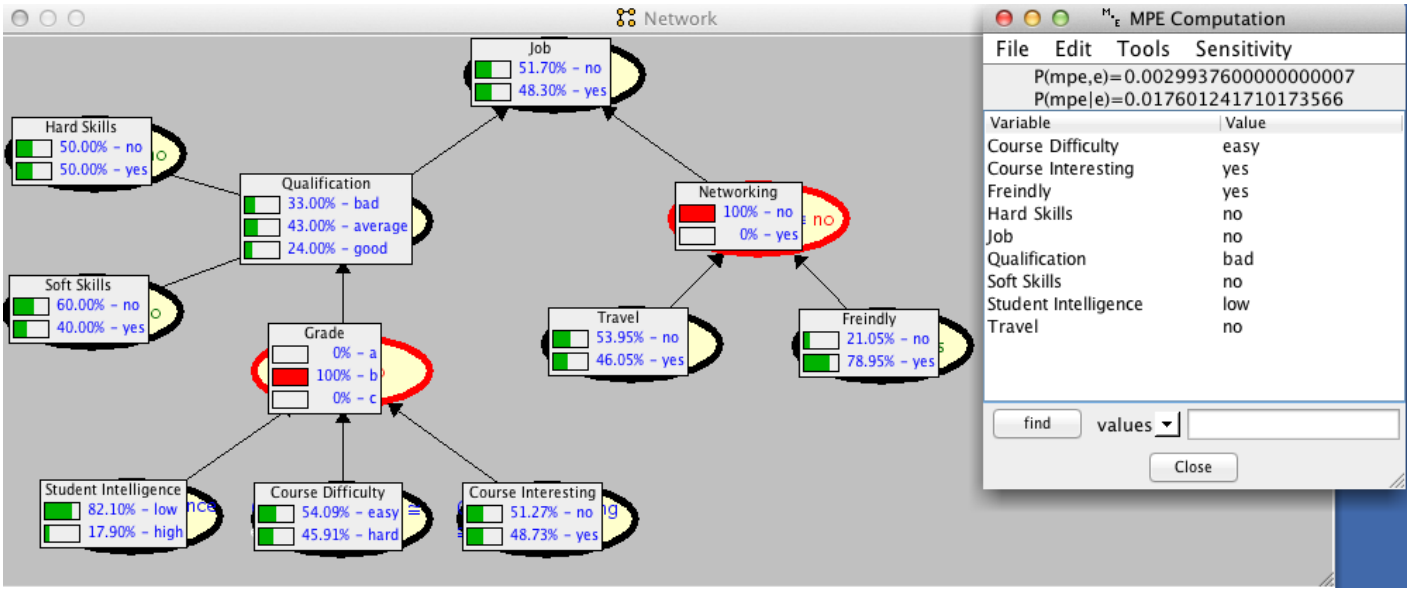


As a result, the new values of the distribution in the new networks above are just slightly changes from its initial values. I also tried running the sensitivity analysis through Samlam which gave me the positive result. I can therefore say that the network is robust with respect to these deviations.

3.4 Most Probable Explanation (MPE)

I selected Scenario 4th, where Grade{=b} and Networking{=no} and compute the MPE: The remaining nodes have the following MPEs:

Course Difficulty: easy
Course Interesting: yes
Friendly: yes
Hard Skills: no
Job: no
Qualification: bad
Soft Skills: no
Student Intelligence: low
Travel: no



Before computed the MPE, I doubted especially the states on the following nodes: Course Interesting, Qualification and Job. I did not expect that the status of Qualification would be *"bad"* as I thought it's supposed to be *"average"* according to its conditional probabilities shown from the network. But once I studied about how MPE works, everything is more clear as the MPE is the joint assignment of values of the non-evidence variables that maximize the posterior probabilities conditional on the evidence. This assignment provides the posterior probability of 0.0176 given the evidence according to Samlam.

3.4 Joint Distribution

I select the following variables: Job, Qualification, Networking, Travel and Friendly.
I set an evidence on these variables: Job = {yes}, Qualification = {good}

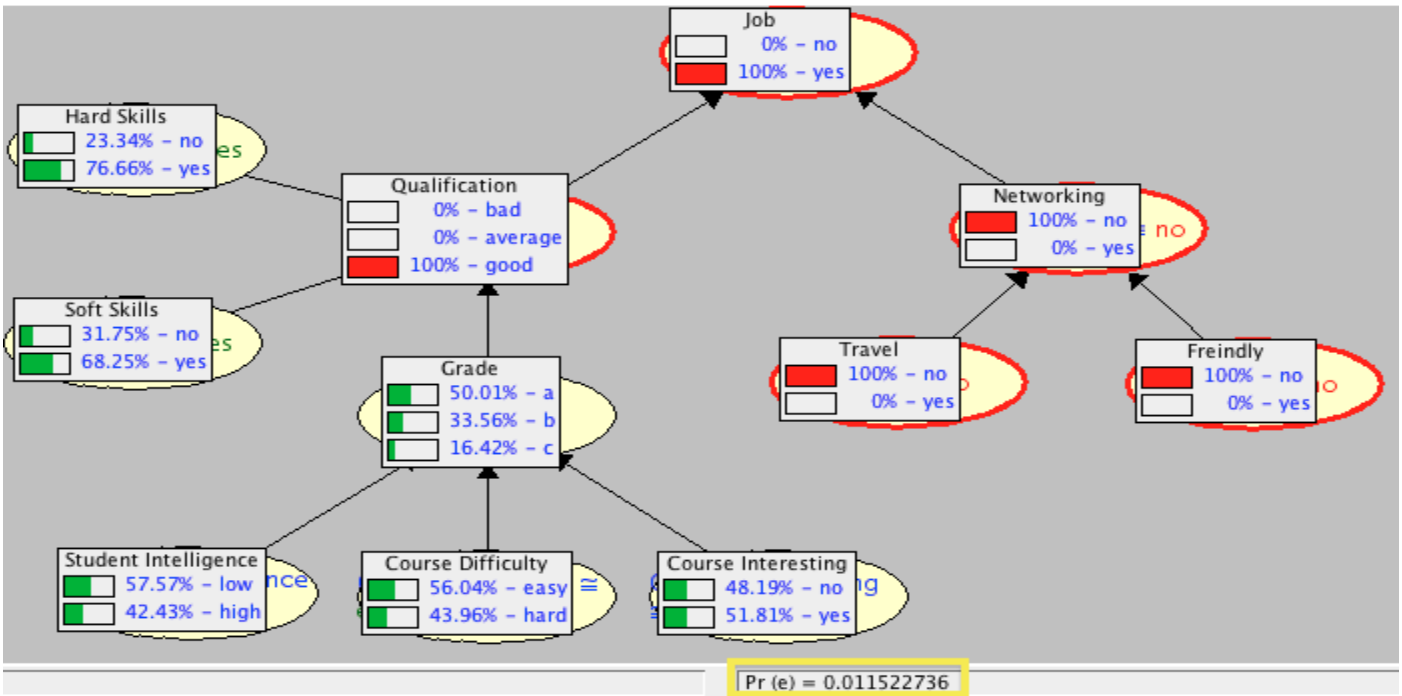
Computing the joint probability distribution P(Networking, Travel, Friendly | Job=yes, Qualification=yes) can be done by following steps:

= P(N, T, F | J=yes, Q=yes)
= P(N, T, F | e) *e* is the evidence (J=yes, Q=yes)
= $\frac{P(N,T,F,e)}{P(e)}$ P(e) = 0.2351 according to Samlam,
P (N, T, F, e) for each combination is also provided by Samlam

P(Networking, Travel, Friendly | Job=yes, Qualification=yes)

			P(e) =	0.2351
Network	**Travel**	**Friendly**	**P (N,T,F,e)**	**P (N,T,F,e) / P(e)**
no	no	no	0.0115	**0.0489**
no	no	yes	0.0409	**0.1740**
no	yes	no	0.0089	**0.0379**
no	yes	yes	0.0358	**0.1523**
yes	no	no	0.0015	**0.0064**
yes	no	yes	0.0122	**0.0519**
yes	yes	no	0.0248	**0.1055**
yes	yes	yes	0.0993	**0.4224**

For example P(N=no, T=no, F=no, e) = 0.0115 according to Samlam below. The remaining probabilities can be obtained with the same approach.



3.5 Conclusion

Overall, the network met my initial goal by representing the probabilistic relationships among various variables of my interest. For example, perhaps I should take an easy course to get a good result but it does not mean having a good grade will secure my job, thus I should also think about other factors that influence my qualification (e.g. hard skills and soft skills) as well as making networking. It is absolutely able to do sensible inference about the variables of interest as the outcome of the probability distributions were realistic. The model can also predict which factor acts the most influencing role in making decision.

The current model could be improved by setting the prior probabilities and conditional probabilities with the more accurate information, but this would take time in researching and collecting such data. It is also possible to add more nodes (e.g. multi-language skills and work permit) to make the model more sensible or introduce an edge between the Course Difficulty and Student Intelligence, in case students who are brighter taking harder courses. But as mentioned, a model is not set in stone, it is a representation of how I believe the how the world works.