University of St Andrews

Assignment 4

Bayesian Networks

Student ID: 120022067

University of St Andrews

CS5011 Artificial Intelligence Practice

Contents

1 Introduction

Bayesian Networks (BNs) are probabilistic graphical models representing a set of random variables and their conditional dependencies via directed acyclic graphs. Bayesian Networks have important applications in diagnostic expert systems particularly for medical applications, investigation of faulty systems, and legal domains.

The aim of this assignment is to learn how to model and use Bayesian networks to reason with uncertainty applied to diagnostic systems.

1.1 Parts Implemented

Part 1: All Bayesian Networks are saved as XML files (prob1.xml and prob2.xml) as stated in the requirements. The investigations mentioned in the requirement are also performed. Full details can be found in Section 3.

Part 2: For part 2, the Bayesian Networks representing the Blue Rainbow bridge scenario are created and saved as XML files (prob3.xml and prob4.xml). Full details can be found in Section 4.

Part 3: Implemented an expert agent assistant system in Java. The Bayesian Network used in the system is BN for medical diagnosis (prob2.xml). An in-depth discussion of part 3's implementation can be found in Section 5.

1.2 Software and Libraries

This section covers the third parties software and library used in this assignment.

1.2.1 AIspace Belief and Decision Networks Tool

AIspace Belief and Decision Networks tool is developed by AIspace (AIspace 2016). It was used to interact with Bayesian Networks.

1.2.2 Encog

Encog is a machine learning framework developed by Heaton Research (Research 2017). It was used to develop an application for an expert agent assistant for part 3.

2 Literature Review

2.1 Overview

Bayesian networks is a data structure used to represent the dependencies among variables (Russell & Norvig 2009). It can represent any full joint probability distribution (Russell & Norvig 2009). A network is a directed graph with each node containing quantitative

probability information. The full specification of Bayesian network is as follows (Russell & Norvig 2009):

1. Each node corresponds to a random variable, which may be discrete or continuous.
2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y. The graph has no directed cycles, and hence is a directed acyclic graph, or DAG.
3. Each node Xi has a conditional probability distribution P(Xi | P arents(Xi))that quantifies the effect of the parents on the node.

An example of Bayesian network is shown in Figure 1.

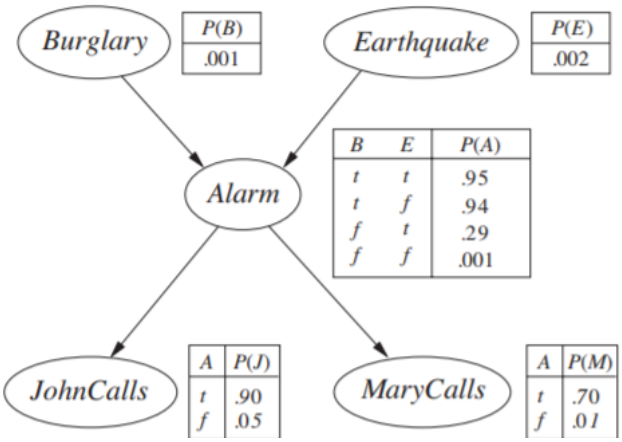

Figure 1: An example of Bayesian network (Russell & Norvig 2009).

2.2 Application of Bayesian Network

Bayesian network is used for modelling beliefs in (Data-Flair 2017):

- Gene Regulatory Networks
- Medicine
- Biomonitoring
- Document Classification
- Information Retrieval
- Image Processing
- Spam filter

3 Part 1

3.1 Design

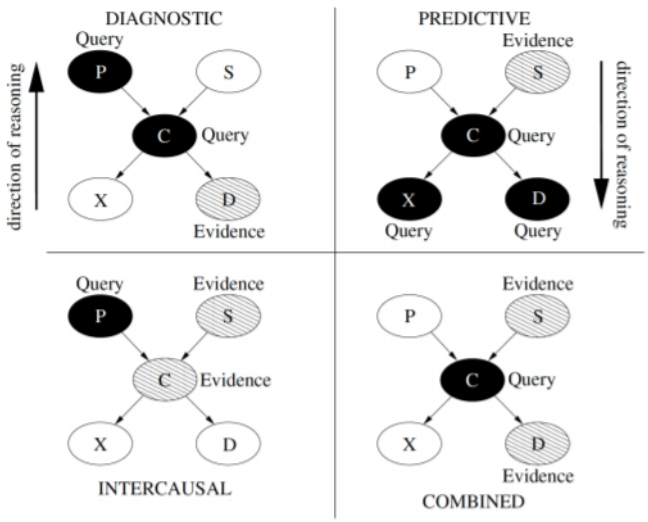

Figure 2: Types of inferential reasoning.

When investigating a query, the first step is to determine the posterior probability (e.g. P(Influenza | Smokes)). Then, the behavior of queried random variable can be estimated by observing the direction of the query (see Figure 2). Finally, the AISpace tool should be used to obtain the result.

3.2 Evaluation

3.2.1 BN for Car Faults Investigation

The queries performed are (Toniolo 2017):

1. Diagnostic: The agent observes that the car starter system is not OK, what other characteristics would the car present?

2. Diagnostic: The agent observes that the voltage at plug is low and the car stater system is not OK, what other characteristics would the car present?

3. Profiling: Determine the general characteristics of a car and what issues it might have depending on its battery age.

4. Predictive: If the starter system and the charging system are both not ok, what is the likely situation of the voltage at plug?

For the first query, the agent observes that the car starter system is not OK. Since the car starter system directly depend on battery voltage, the probability that battery voltage

is strong should decrease while the probability of having weak or dead battery voltage should increase.

Another variable that may be affected by the change in battery voltage's probability distribution is voltage at plug. Since the probability of having high voltage at plug is highly correlates with probability of string battery voltage, the probability of high voltage at plug is expected to decrease.

The two variables that have influence on the probability distribution of battery voltage are battery age and charging system. The probability distribution of charging system should have little to no changes at all because the chance that there is a problem with charging system is very small. However, the probability that battery age is new should decreases, while the probability of battery being old and very old should increases. The query result can be found in Figure 3.

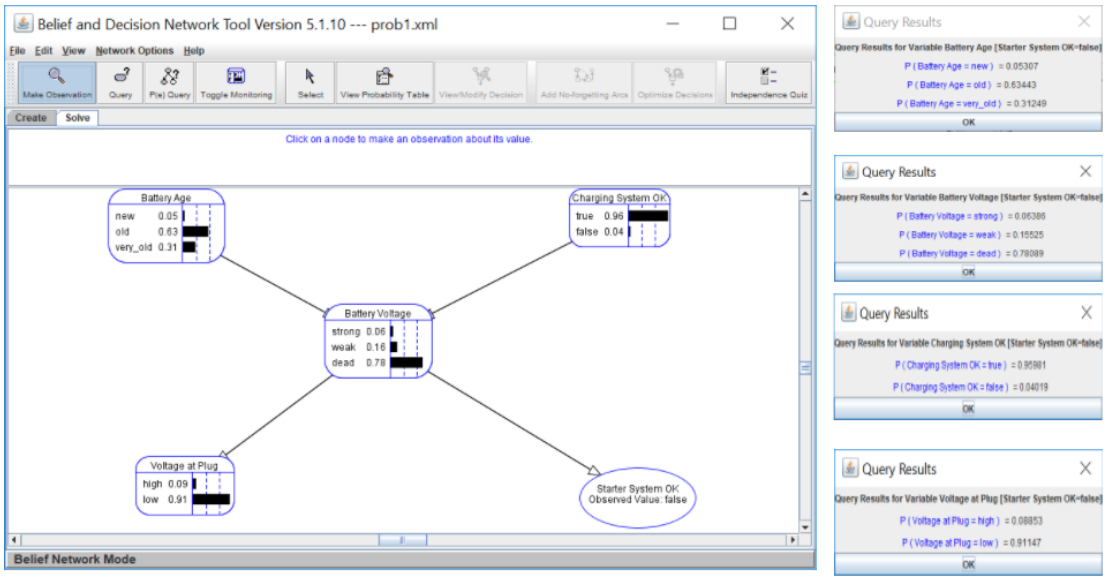

Figure 3: Result of query 1 (BN for Car Faults Investigation).

For the second query, the result should be similar to query 1 because the observations are made on variables that depend on battery voltage. The changes in battery voltage's probability distribution should also affect the probability distribution of battery age like in query 1. The query result is shown in Figure 3 and Figure 5.

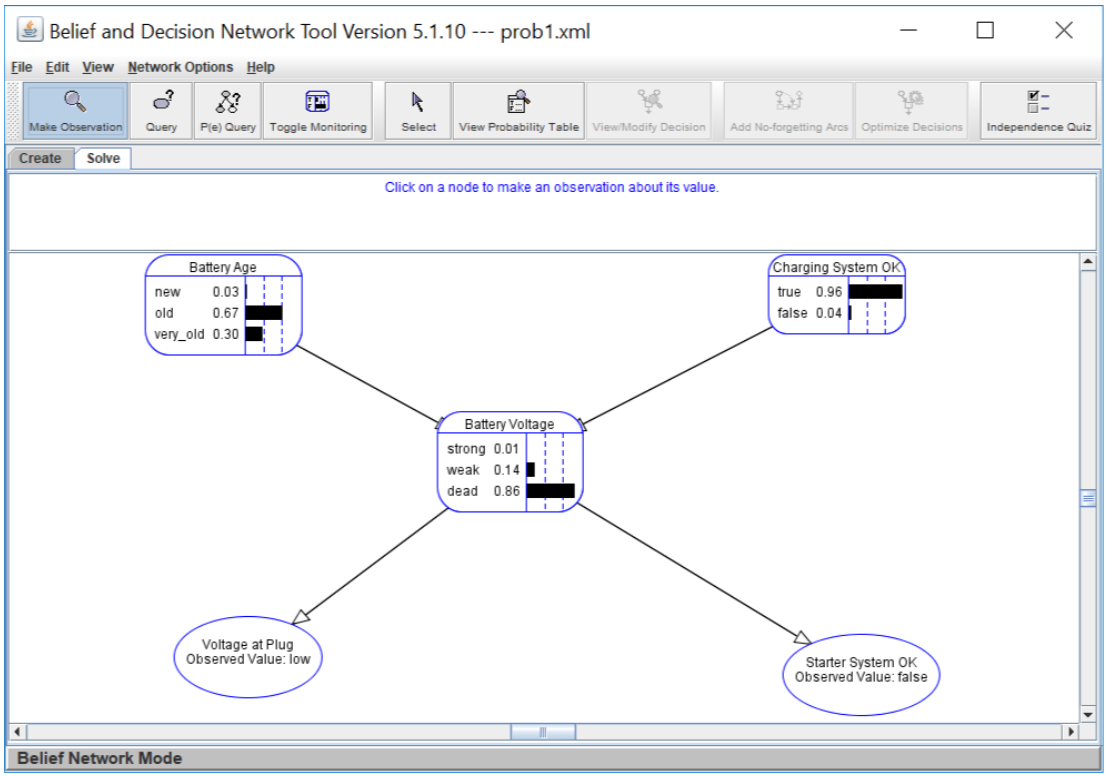

Figure 4: The agent observes that the voltage at plug is low and the car stater system is not OK.

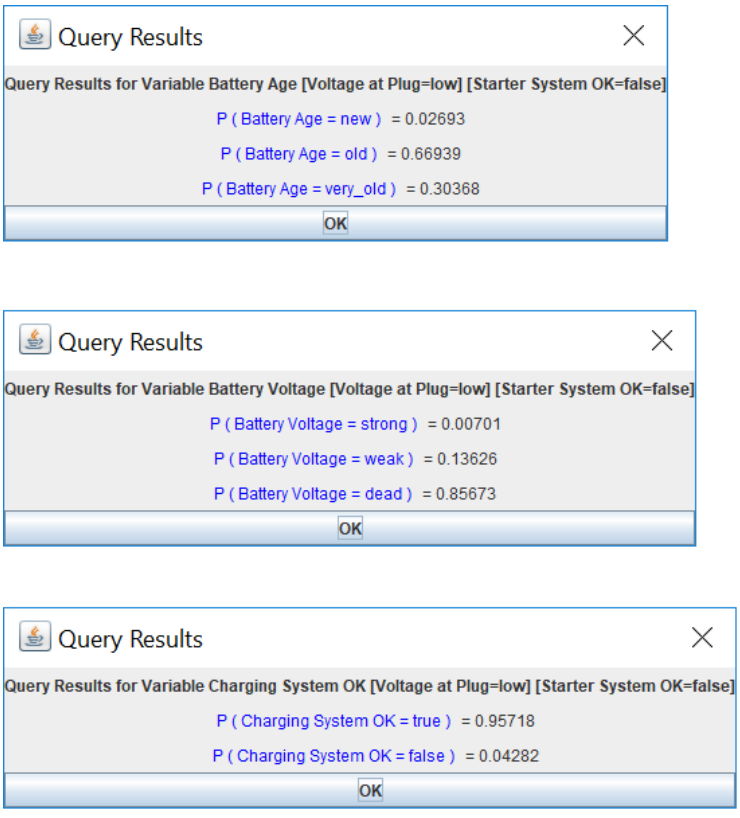

Figure 5: Result of query 2 (BN for Car Faults Investigation).

For the third query, the general characteristic of the car indicates that it is in good condition. As battery gets older, the probability of having high battery voltage should decreases and the probability that there are problems at voltage at plug and starter system should increases. The state of the car with different battery age can be found in Figure 6, Figure 7, and Figure 8.

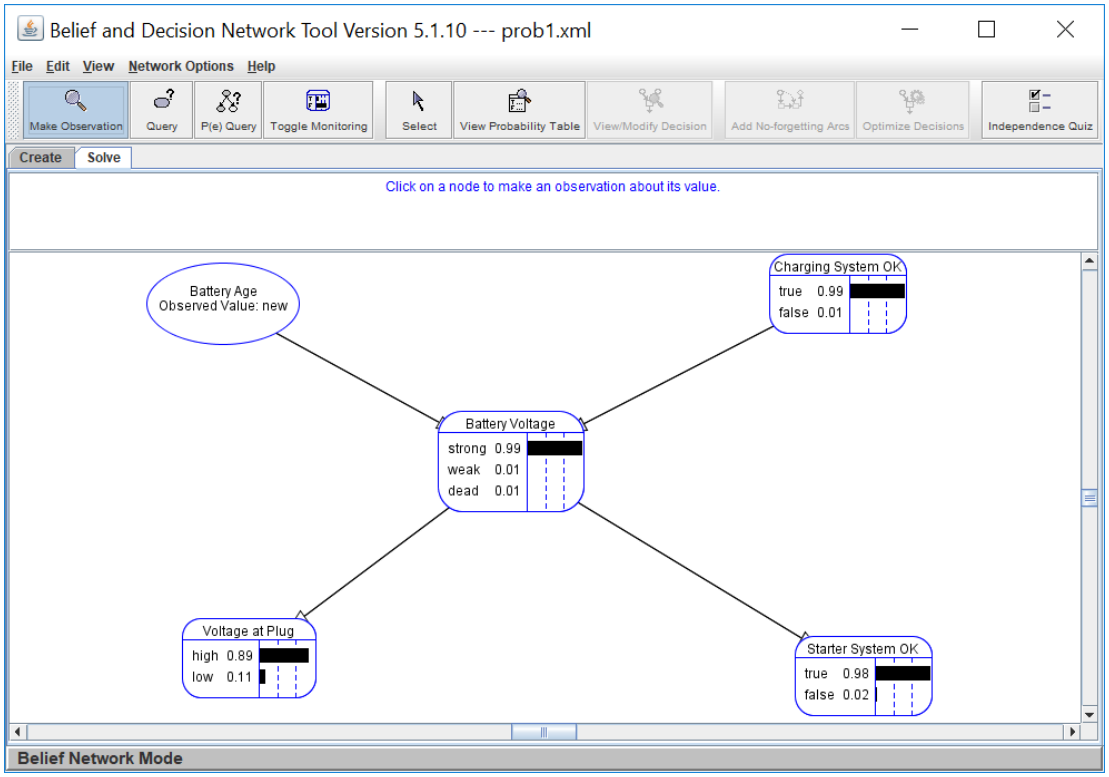

Figure 6: The state of a car with new battery.

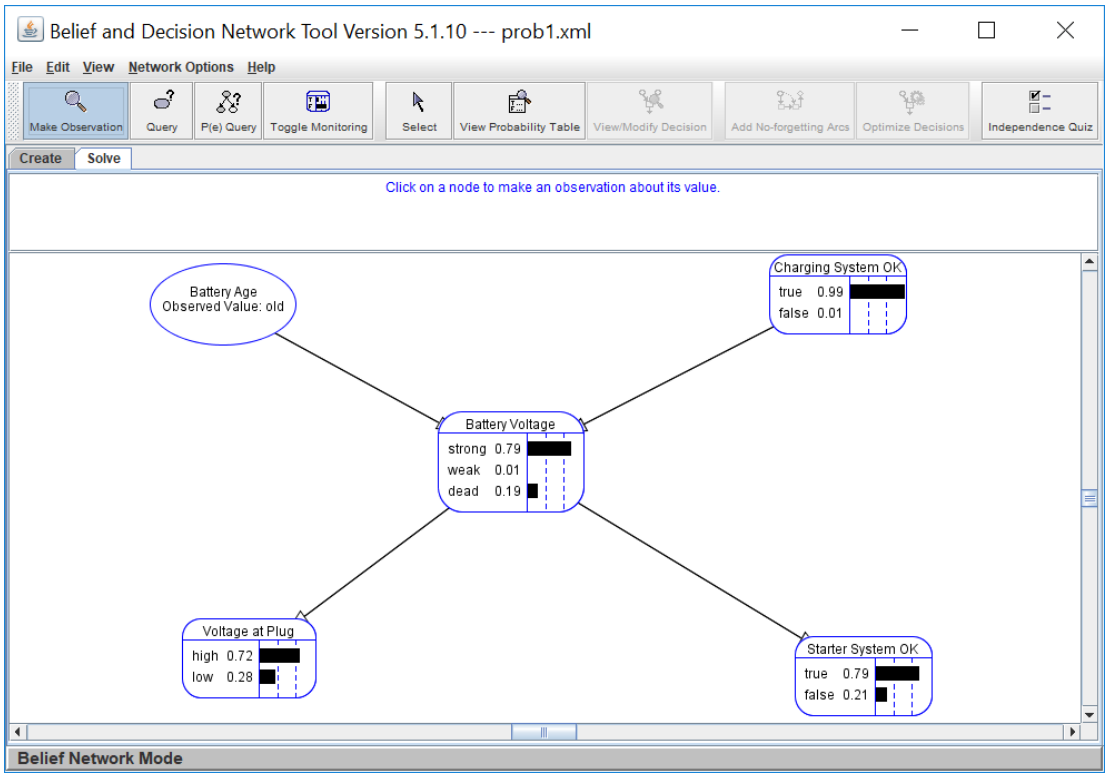

Figure 7: The state of a car with old battery.

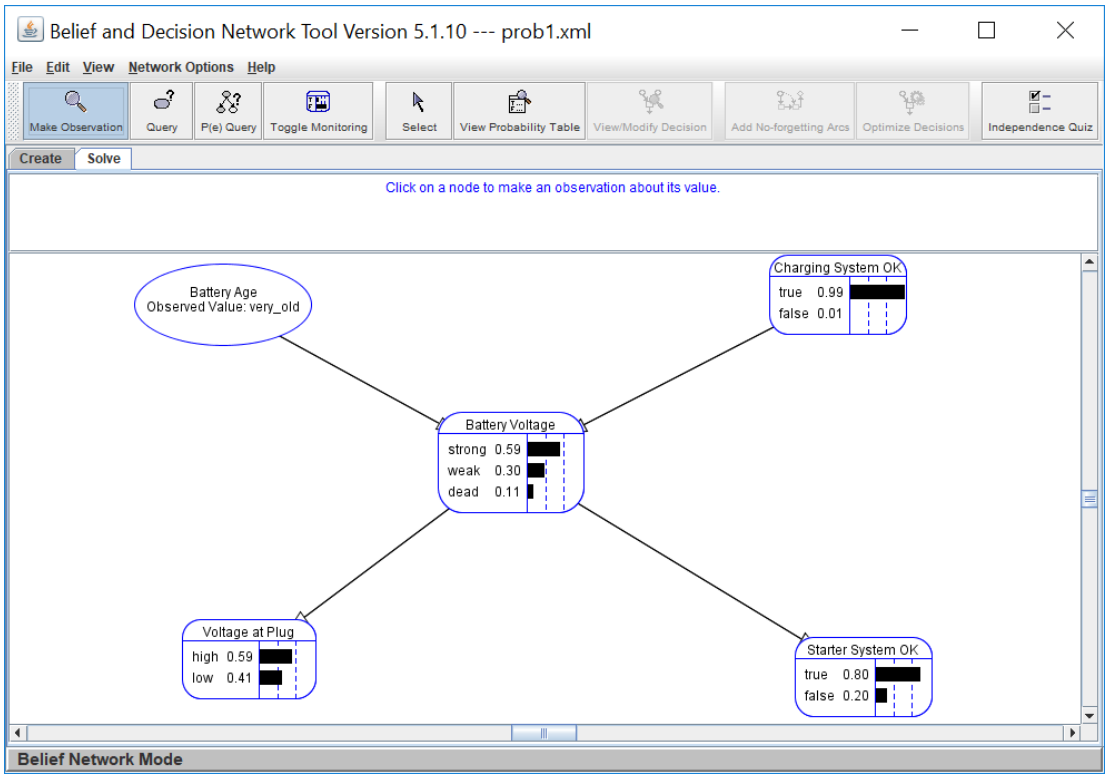

Figure 8: The state of a car with very old battery.

For the fourth query, the charging system and starter system are not OK. Since the two

variables affect the probability distribution of battery voltage, the probability of having low voltage at plug should be very high. The query result is shown is Figure 9 and Figure 10.

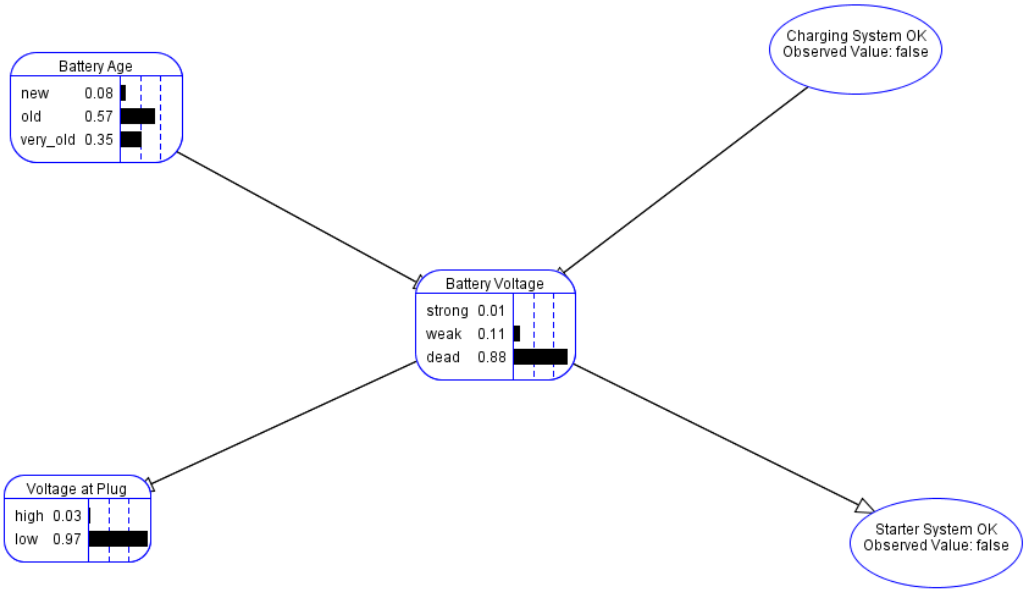

Figure 9: The agent observes that the charging system and starter system are not OK.

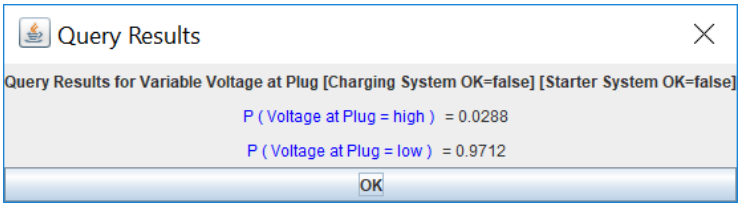

Figure 10: Result of query 4 (BN for Car Faults Investigation).

3.2.2 BN for Medical Diagnosis

The queries performed are (Toniolo 2017):

1. Diagnostic: The agent observes that the patient is wheezing, how likely is that the patient smokes?

2. Predictive: The agent observes that the patient is a smoker, how likely is that the patient will develop bronchitis? And that they will have sore throat?

3. Predictive: The agent observes that the patient is a smoker, and has influenza, how likely is that the patient will develop sore throat?

4. Predictive: The agent observes that the patient is a not smoker, and has influenza, how likely is that the patient will be coughing?

5. Intercausal: The agent observes that the patient is not a smoker and, has got bronchitis. How likely is that they have influenza?

6. Profiling: Determine the characteristics of patients that smoke and compare. those against patients that do not smoke

7. Predictive: If the patient has influenza, how likely is that the thermometer will indicate fever?

For the first query, if the patient is wheezing, the probability that the patient has bronchitis should increase because it is directly depend on wheezing. As a result, the probability that the patient smokes should also increase because it is one of the cause of bronchitis. The query result can be found in Figure 11 and Figure 12.

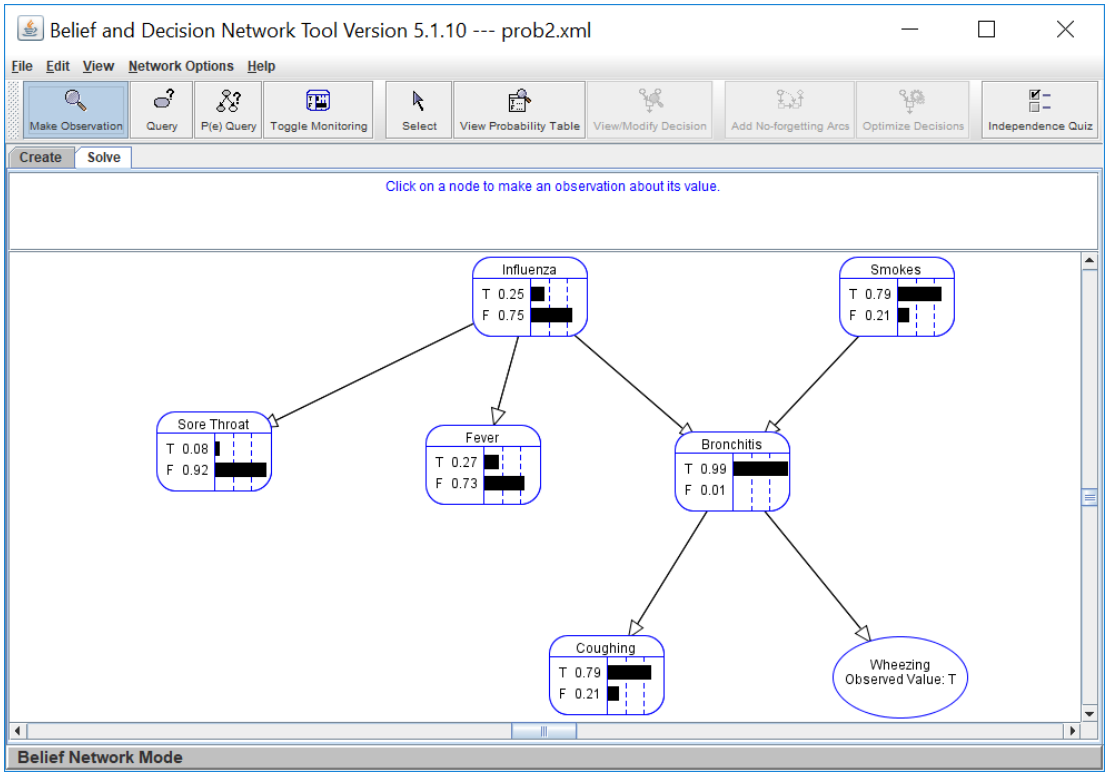

Figure 11: The agent observes that the patient is wheezing.

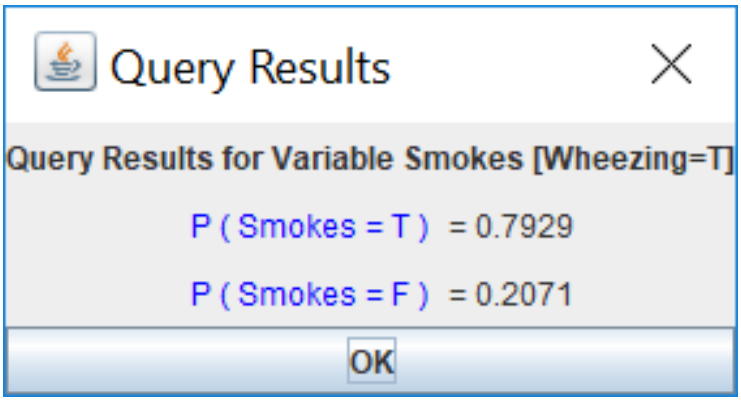

Figure 12: Probability that the patient smokes.

The second query states that the patient is a smoker. This observation suggests that the probability that the patient has bronchitis should increase because smoking directly affects bronchitis. However, the probability distribution of variable sore throat should be unaffected because there is no connection between smoking and sore throat. The query result can be found in Figure 13 and Figure 14.

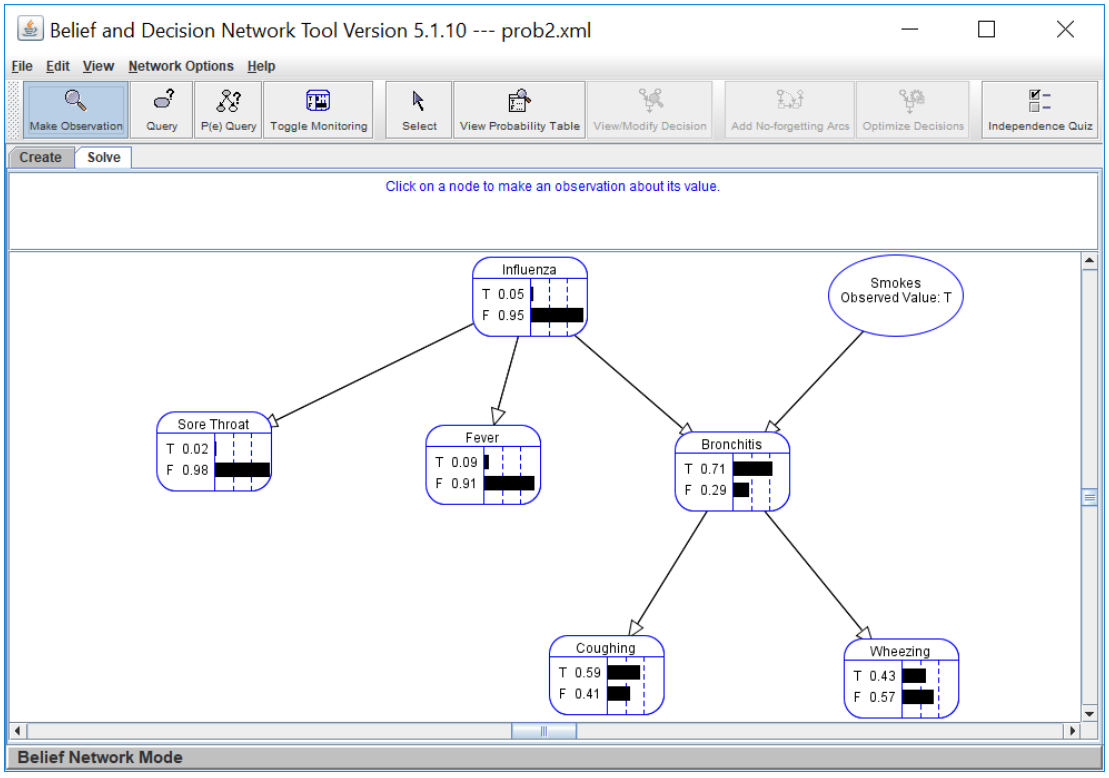

Figure 13: The agent observes that the patient is a smoker.

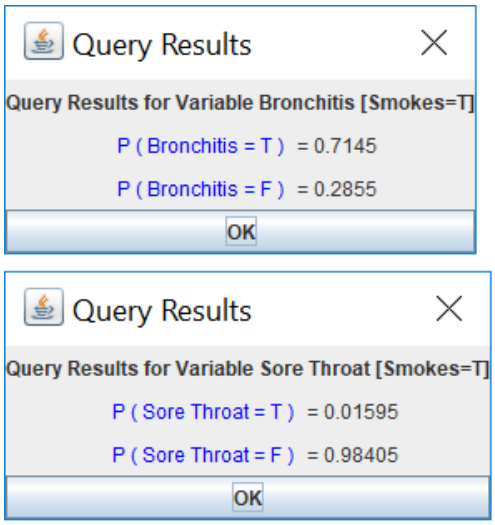

Figure 14: Probability that the patient has bronchitis and sore throat.

The third query observes that the patient is a smoker, and has influenza. Since the patient has influenza, the probability of having sore throat should increase. The observation on smokes variable should have no effect on the probability distribution of sore throat because there is no dependency between the two variables. The query result can be found in Figure 15 and Figure 16.

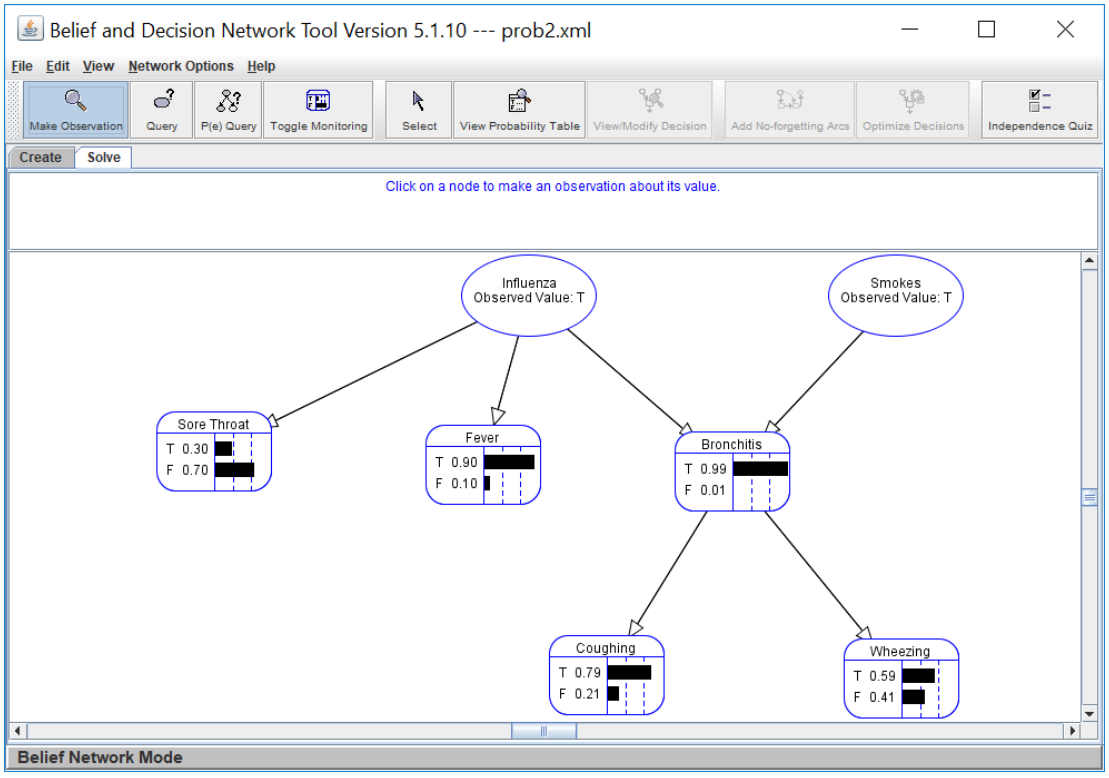

Figure 15: The agent observes that the patient is a smoker, and has influenza.

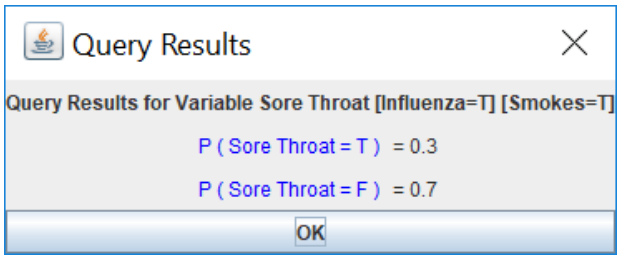

Figure 16: Probability that the patient has sore throat.

The fourth query observes that the patient is a not smoker, and has influenza. In this case, the probability of bronchitis should increase because influenza causes bronchitis. As a result, the probability of coughing should also increase because it is one of the effect of bronchitis. The query result can be found in Figure 17 and Figure 18.

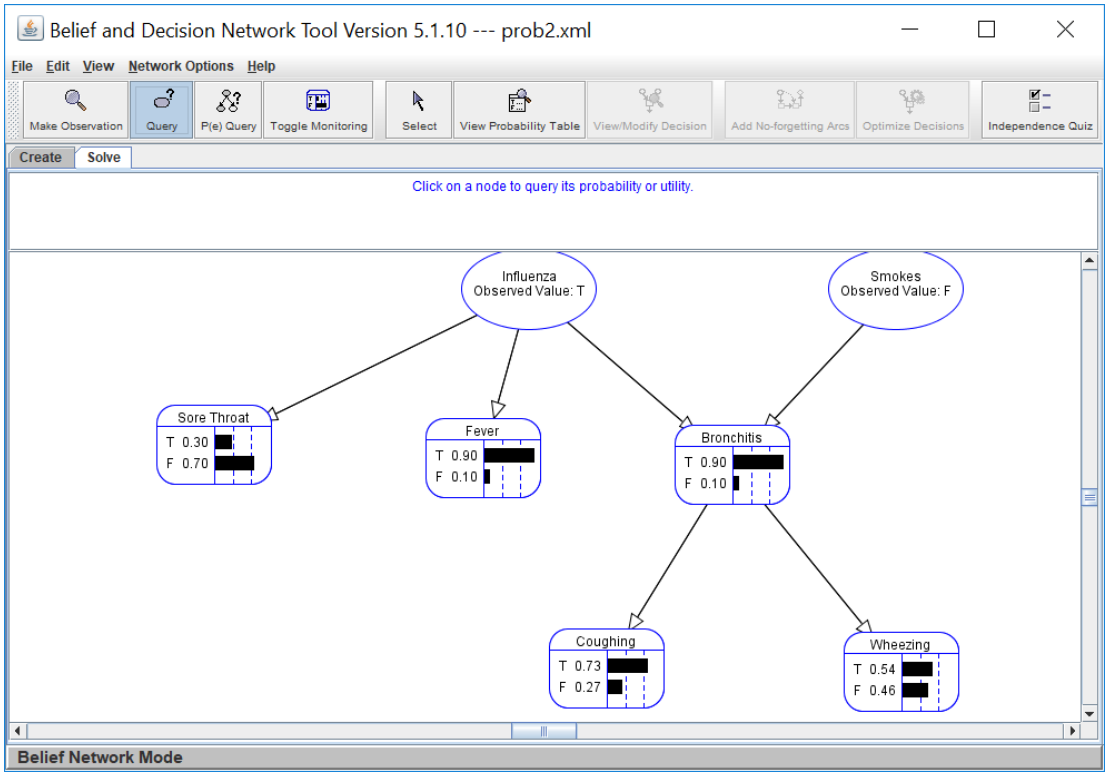

Figure 17: The agent observes that the patient is a not smoker, and has influenza.

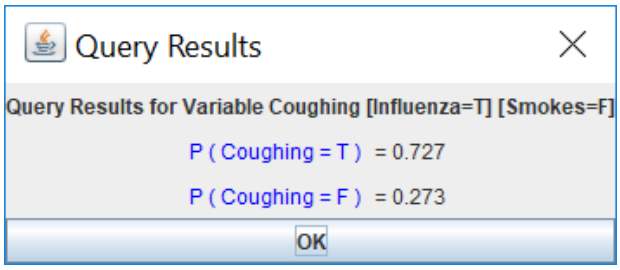

Figure 18: Probability that the patient is coughing.

For the fifth query, the patient is not a smoker and, has got bronchitis. In this case, the probability that the patient has influenza should be 1.00. This is because bronchitis is caused by influenza and smoking. Since the agent observes that the patient is not a smoker, the only thing that can cause bronchitis is influenza. The query result is shown in Figure 19 and Figure 20.

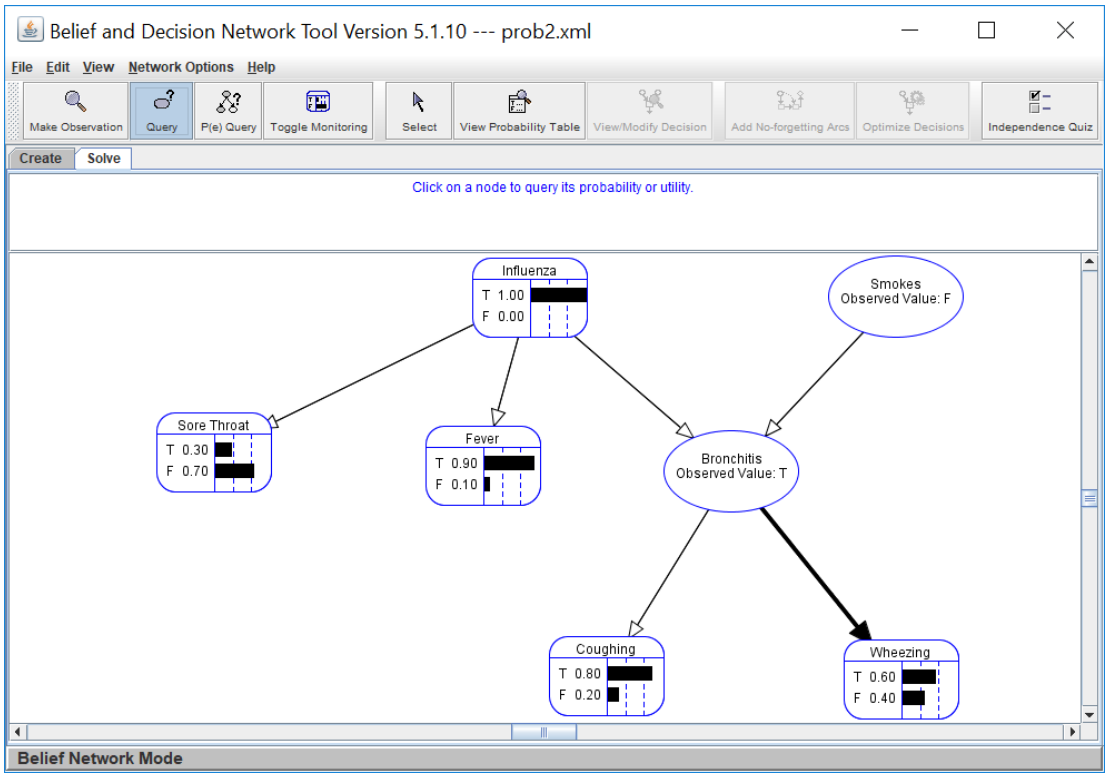

Figure 19: The agent observes that the patient is a not smoker, and has bronchitis.

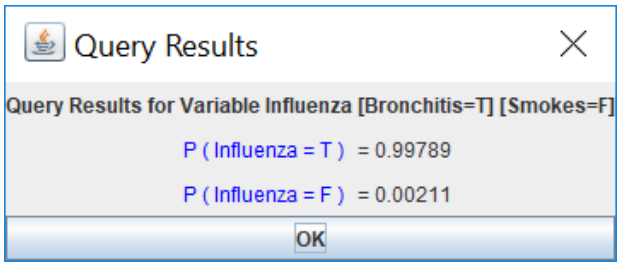

Figure 20: Probability that the patient has influenza.

The sixth query determines the characteristics of patients that smoke and compare those against patients that do not smoke. Generally, smoker should have higher chance of being ill than non smoker because smoking causes bronchitis, which can cause coughing and wheezing. The query result can be found in Figure 21 and Figure 22.

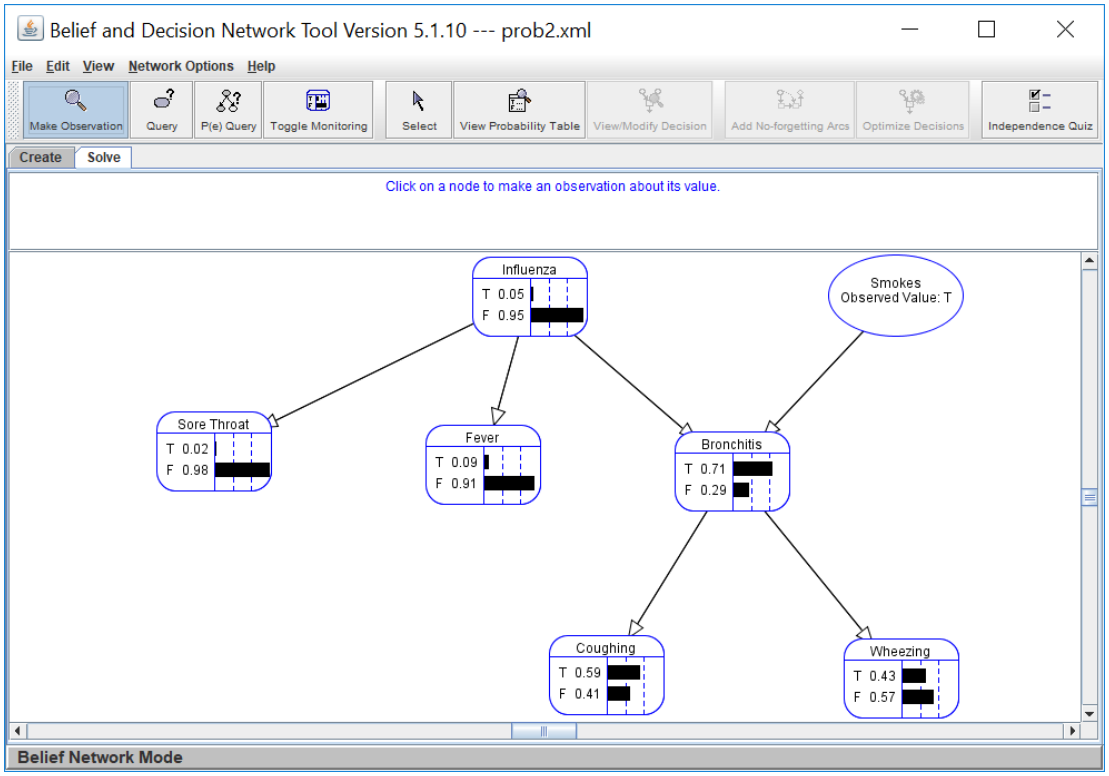

Figure 21: Characteristics of smokers.

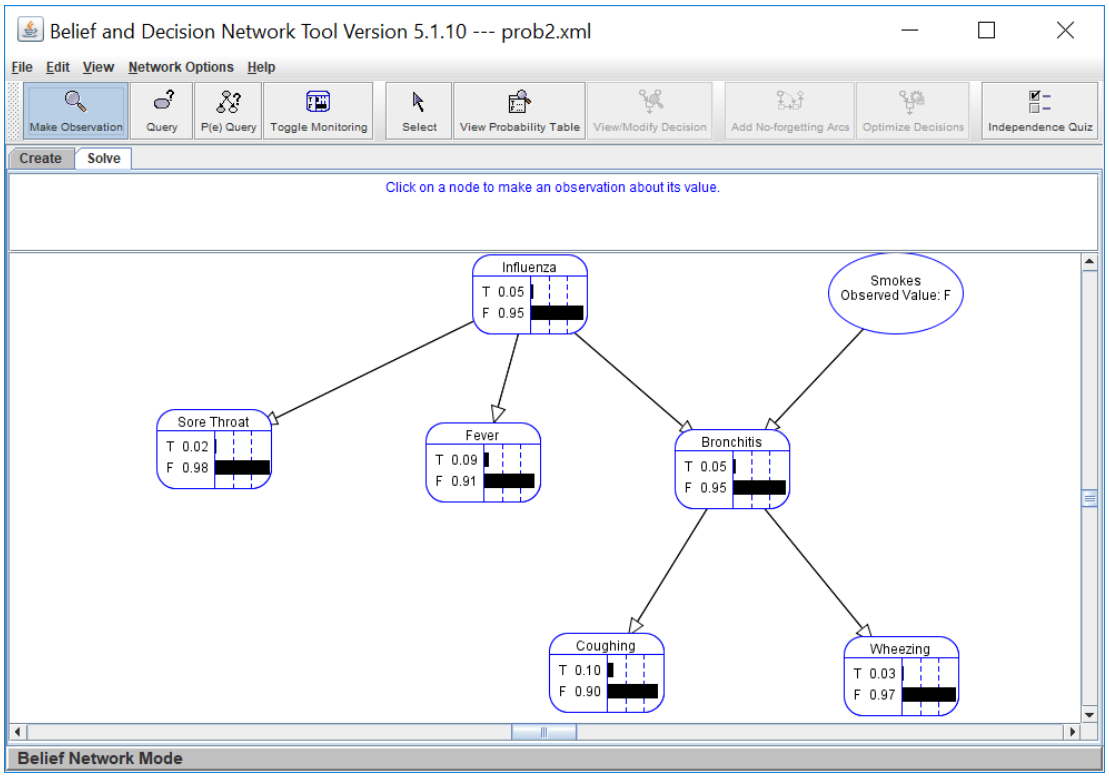

Figure 22: Characteristics of non-smokers.

The seventh query states that the patient has influenza. This means that the probability of fever should increases. Since the therm is highly correlates with fever, the probability that the thermometer indicate fever should be high. The query result can be found in Figure 23 and Figure 24.

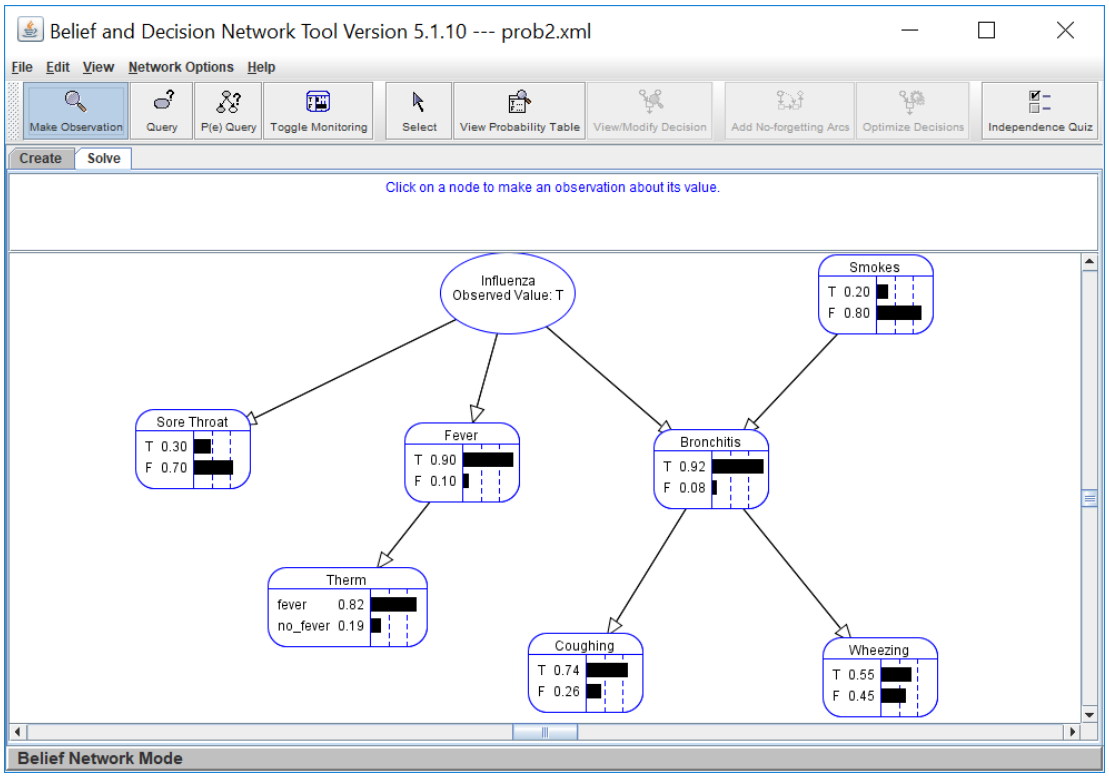

Figure 23: The agent observes that the patient is has influenza.

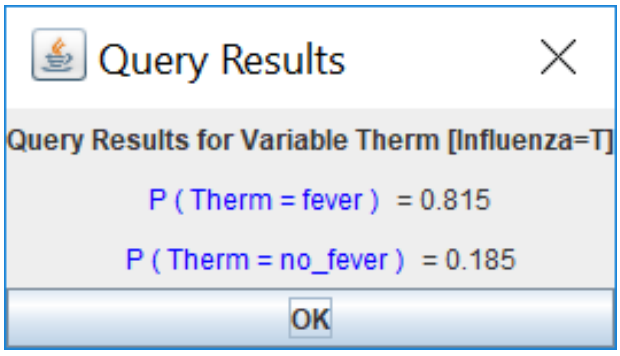

Figure 24: Probability that the thermometer will indicate fever.

4 Part 2

4.1 Design

4.1.1 Domain Variables and Relationship

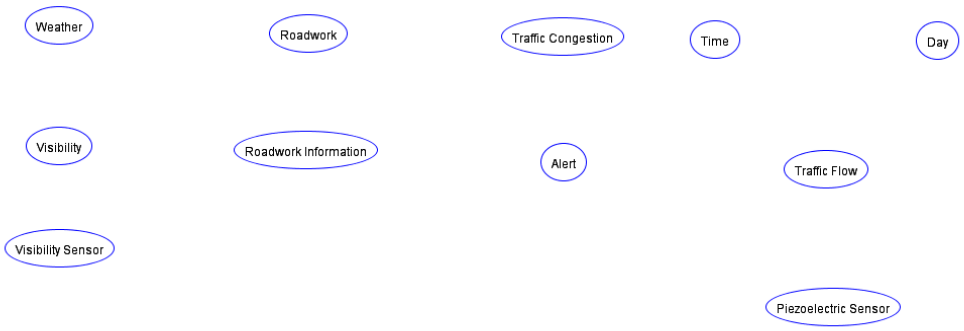

Figure 25: Domain variables of BRB Bayesian Network (prob3.xml).

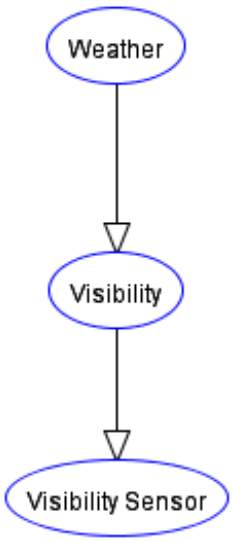

Figure 26: Relationship between weather and visibility.

The first step is to decide what domain variables are and what relationships are between the domain variables. The domain variables for the network are shown in Figure 25. The next step is to determine the relationships between the variables. The requirement states that visibility depends on weather, and the visibility sensor's reading depends on visibility (Toniolo 2017). The relationship between the three variables is shown in Figure 26.

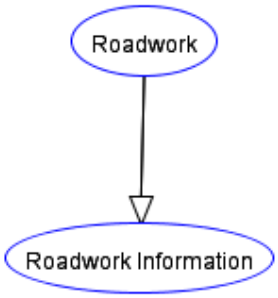

Figure 27: Relationship between roadwork related variables.

According to Toniolo (2017), roadwork information system directly depends on whether there is planned roadwork or not. The relationship of roadwork related variables is illustrated in Figure 27.

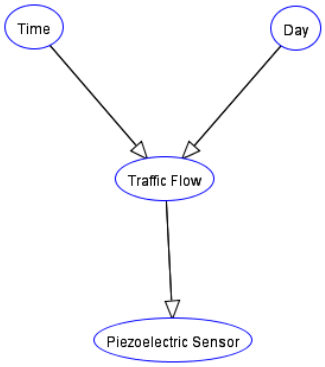

Figure 28: Relationship between traffic flow related variables.

The requirement suggests that time of day and type of day affects number of cars, and number of cars affects piezoelectric sensor's reading (Toniolo 2017). The relationship of traffic flow related variables is shown in Figure 28.

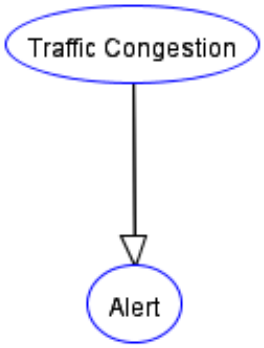

Figure 29: Relationship between traffic congestion and alert.

Toniolo (2017) states that alert will trigger 80% of the time if the system predicts traffic congestion. The relationship of traffic congestion and alert is shown in Figure 29.

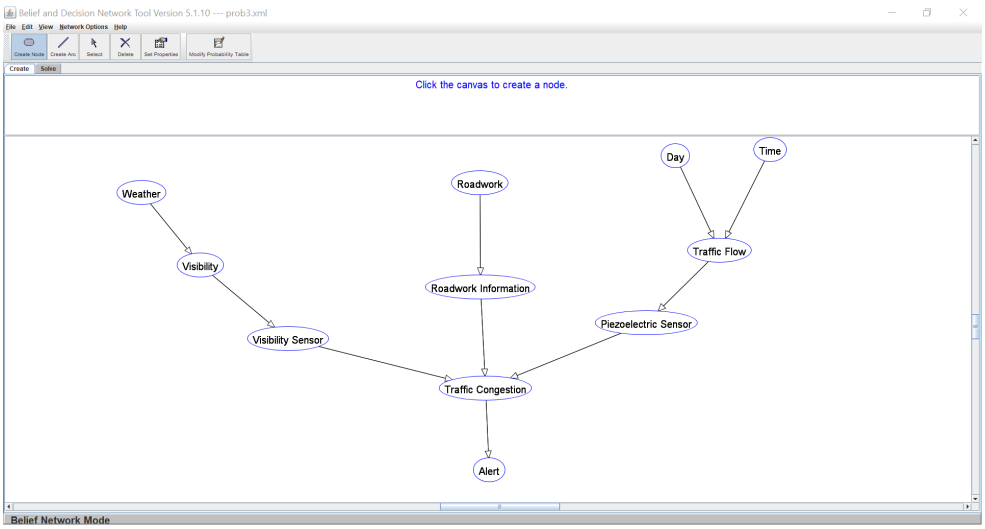

Figure 30: Fully connected BRB Bayesian Network (prob3.xml).

After that, all variables need to be connected. The fully connected network is shown in Figure 32.

4.1.2 Probability Distribution

For this part, we assume that BRB is located in a place where it is sunny most of the time. The factors that affect traffic congestion are visibility, roadwork, and traffic flow. The probability distribution of BRB Bayesian Network can be found in Figure 31.

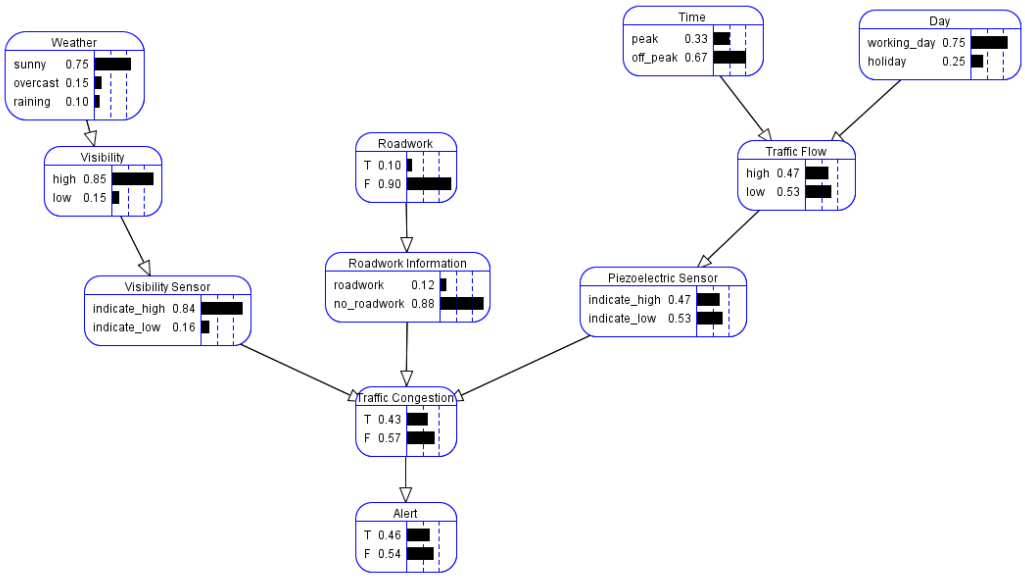

Figure 31: Probability distribution of BRB Bayesian Network (prob3.xml).

4.1.3 Adding Correlations Between Variables

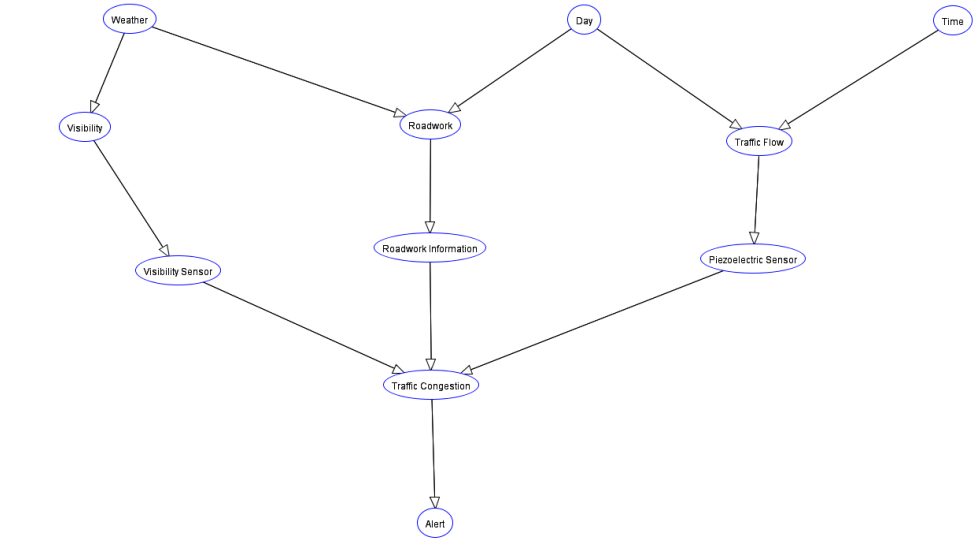

Figure 32: Fully connected BRB Bayesian Network (prob4.xml).

For this part, the following relationships are added to the network:

1. Workers are less likely to do the roadwork when the weather is bad i.e. raining and overcast.

2. There are not many workers available to do roadwork during holiday period.

The modified network is shown in Figure 33.

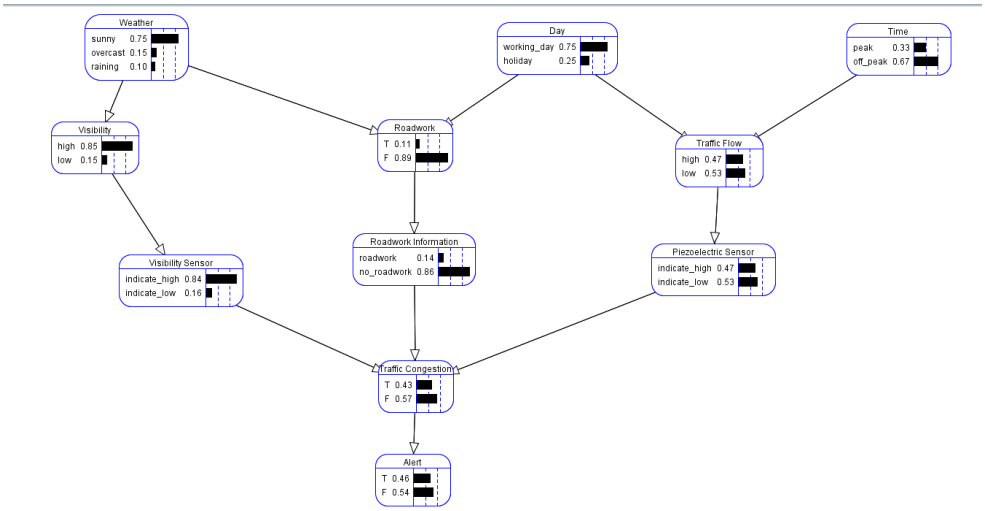

Figure 33: Probability distribution of BRB Bayesian Network (prob4.xml).

4.2 Examples and Testing

Simple queries are also made to test whether the probability constraints stated in the requirement are satisfied. The result can be found in Table 1.

Query	Expected outcome	Actual outcome
Probability of visibility sensor indicates high visibility given that visibility is low	0.01	0.01
Probability of roadwork information system indicate that there is roadwork taking place given that there is no planned roadwork	0.03	0.03
Probability of triggering alert given that there is traffic congestion	0.8	0.8
Probability of piezoelectric sensor indicate high traffic flow given that there is low traffic flow	0.1	0.1

Table 1: The probability constraints test results

4.3 Evaluation

4.4 BRB Bayesian Network

The queries performed are:

1. Predictive: The system observes that today is working day, how likely is that traffic flow is high?
2. Predictive: The system observes that it is currently peak time, how likely is that traffic flow is high?
3. Predictive: The system observes that piezoelectric sensor indicates high traffic flow, how likely is that traffic congestion is true?
4. Diagnostic: The system observes that traffic congestion is true, how likely is that visibility sensor indicate high visibility?
5. Diagnostic: The system observes that traffic congestion is true, how likely is that roadwork information system indicate planned roadwork?

The first query makes an observation that today is working day. The probability of traffic flow being high should increase because more people take their cars. However, another cause of high traffic flow is peak time. Since the probability of peak time is quite low (8 out of 24 hours), the probability of high traffic flow should be less than 60 %. The query result can be found in Figure 34 and Figure 35.

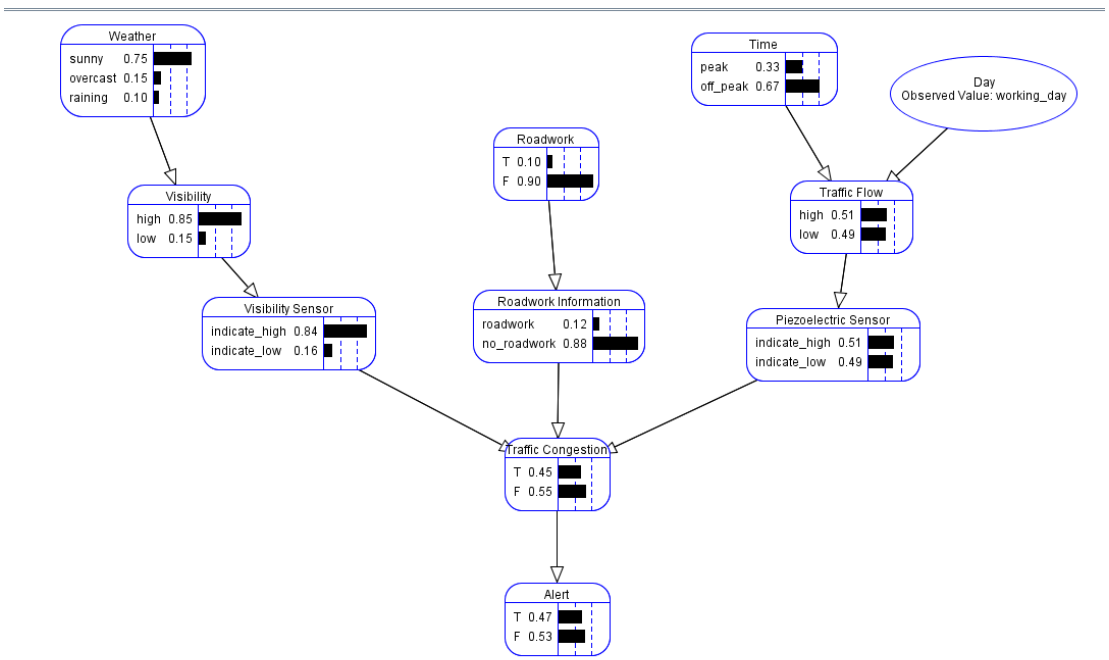

Figure 34: The system observes that today is working day.

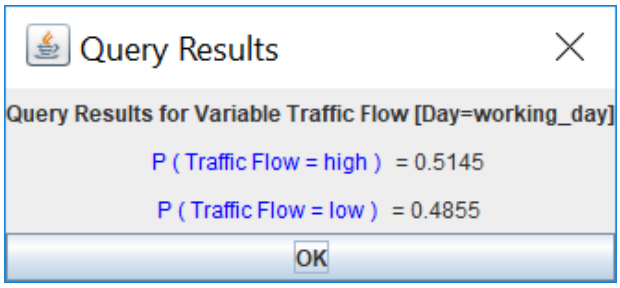

Figure 35: Probability that traffic flow is high.

The second query makes an observation that it is currently peak time. Since peak time is one of the causes of high traffic flow, it is likely that high traffic flow probability will increase. The query result can be found in Figure 36 and Figure 37.

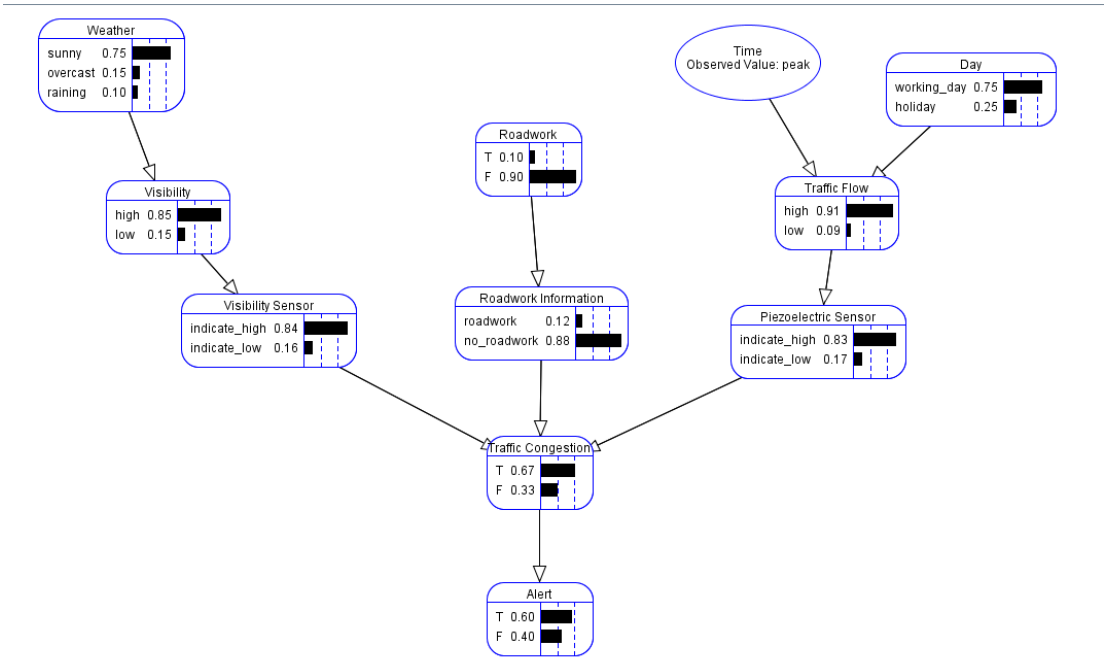

Figure 36: The system observes that it is currently peak time.

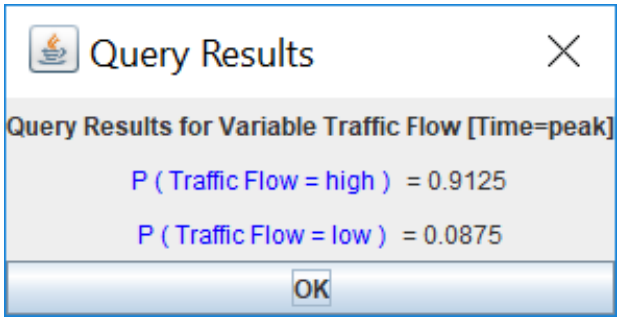

Figure 37: Probability that traffic flow is high.

The third query makes an observation that piezoelectric sensor indicates high traffic flow. Since high traffic flow is one of the causes of traffic congestion, it is likely that traffic congestion probability is high. The query result can be found in Figure 38 and Figure 39.

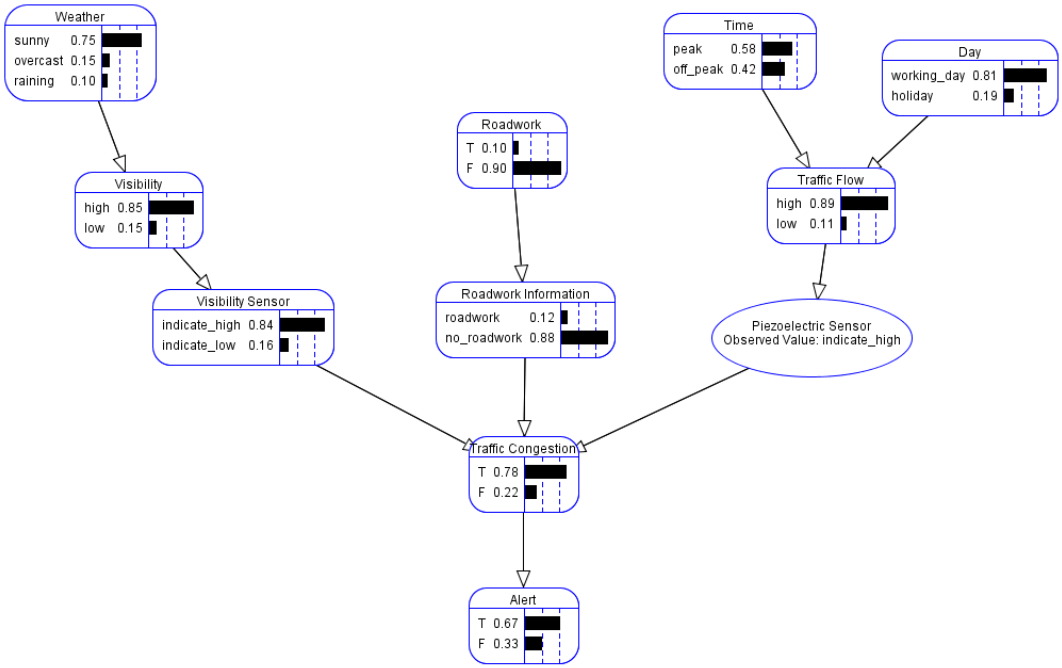

Figure 38: The system observes that piezoelectric sensor indicates high traffic flow.

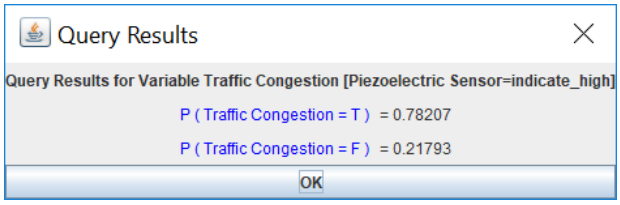

Figure 39: Probability that there is traffic congestion.

The fourth query states that there is traffic congestion. In this case, the probability of visibility sensor indicate high visibility should slightly decrease because low visibility is one of the causes of traffic congestion. However, it is very likely that visibility is still high because the it is usually sunny at BRB. The query result can be found in Figure 40 and Figure 41.

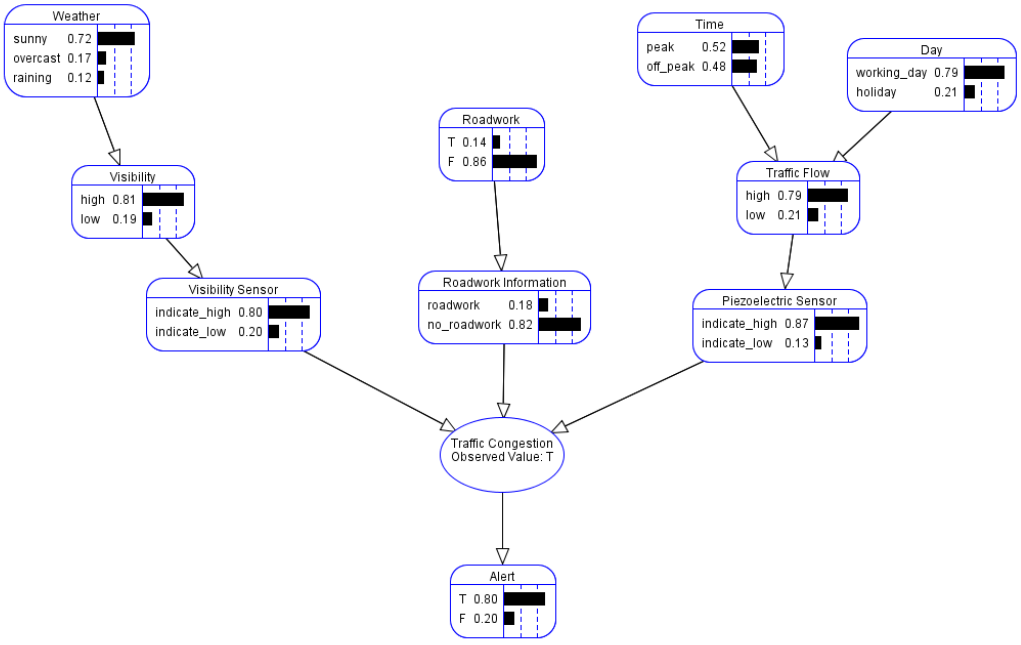

Figure 40: The agent observes that there is traffic congestion.

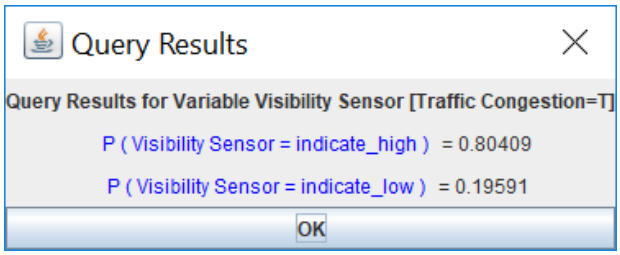

Figure 41: Probability that visibility sensor indicate high visibility.

The fifth query states that there is traffic congestion. In this case, the probability that roadwork information system indicate planned roadwork should slightly increase because roadwork is one of the causes of traffic congestion. However, it is very likely that probability of roadwork taking place is still low. The query result can be found in Figure 42 and Figure 43.

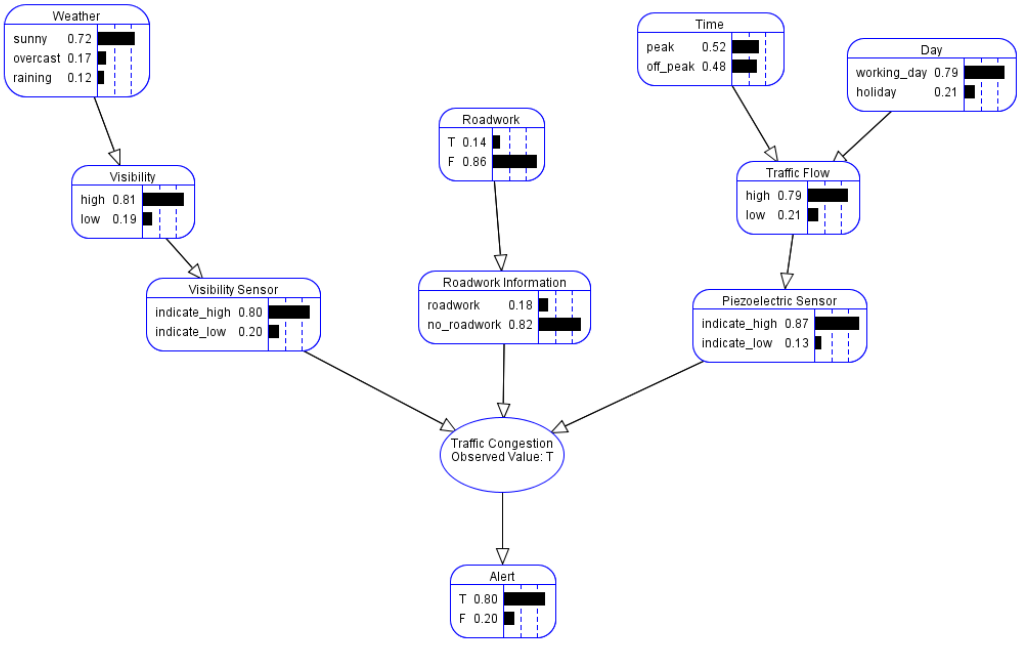

Figure 42: The agent observes that there is traffic congestion.

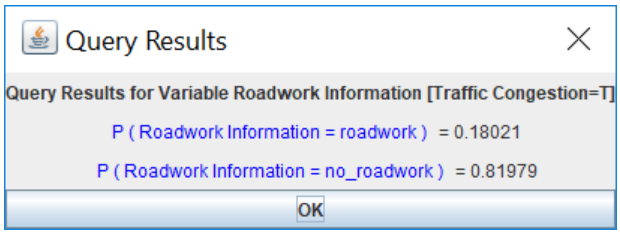

Figure 43: Probability that roadwork information system indicate planned roadwork.

4.5 BRB Bayesian Network with Correlation

The queries performed are:

1. Predictive: If there is roadwork, how likely is that alert will be triggered?
2. Predictive: The system observes that it is raining, how likely is that roadwork is taking place?
3. Predictive: The system observes overcast on a working day during peak time, how likely is that alert will be triggered?
4. Diagnostic: The system observes that there is no traffic congestion, how likely is that it is currently peak time?
5. Diagnostic: The system observes that there is roadwork, how likely is that today is working day?

The first query states that there is roadwork. In this case, the probability traffic congestion should increase. As a result, the probability of triggering alert should increase as well because traffic congestion directly affects alert. The query result can be found in Figure 44 and Figure 45.

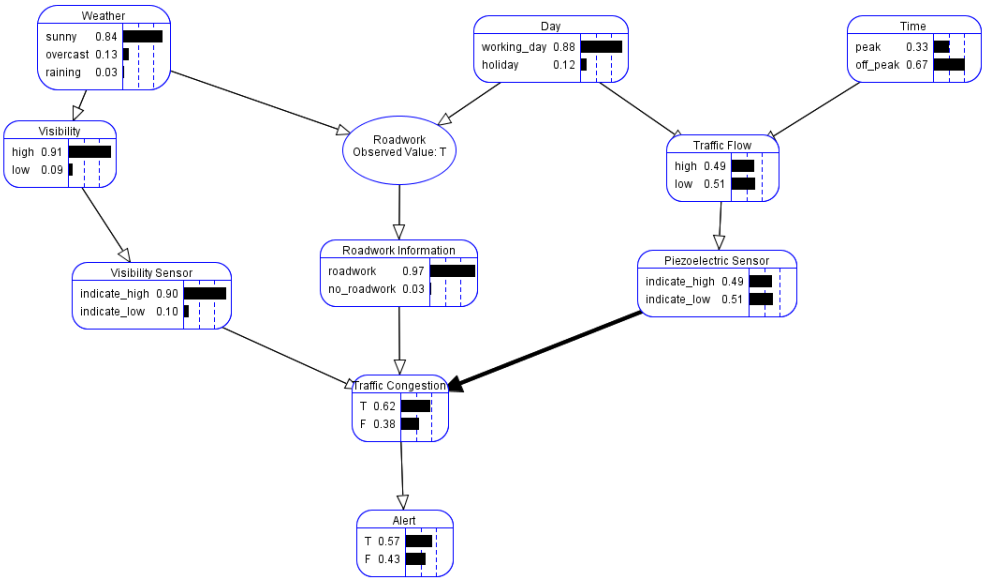

Figure 44: The agent observes that there is roadwork.

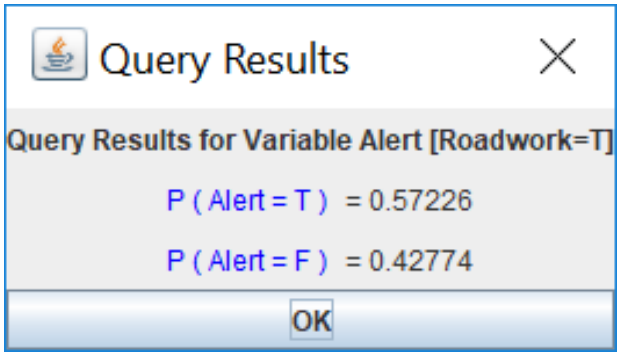

Figure 45: Probability of triggering alert.

The second query states that it is raining. In this case, the probability of roadwork should decrease because roadwork directly depend on weather. The query result can be found in Figure 46 and Figure 47.

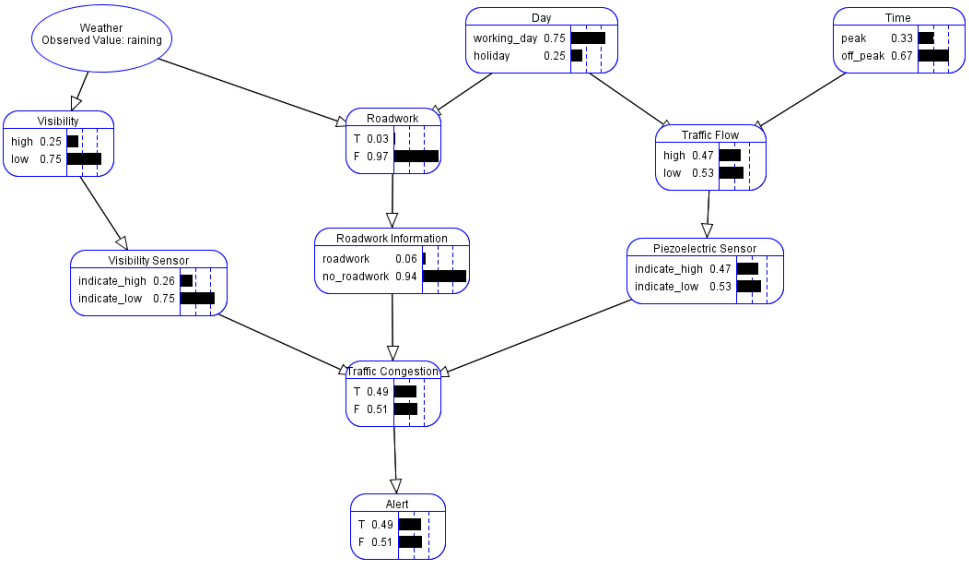

Figure 46: The agent observes that it is raining.

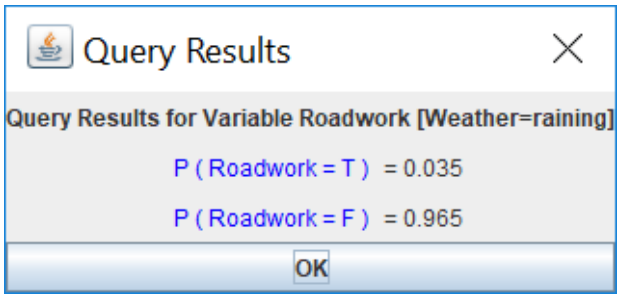

Figure 47: Probability of roadwork.

The third query observes overcast on a working day during peak time. In this case, the probability of alert should increase because all observed variables affects traffic congestion, and alert directly depend on traffic congestion. The query result can be found in Figure 48 and Figure 49.

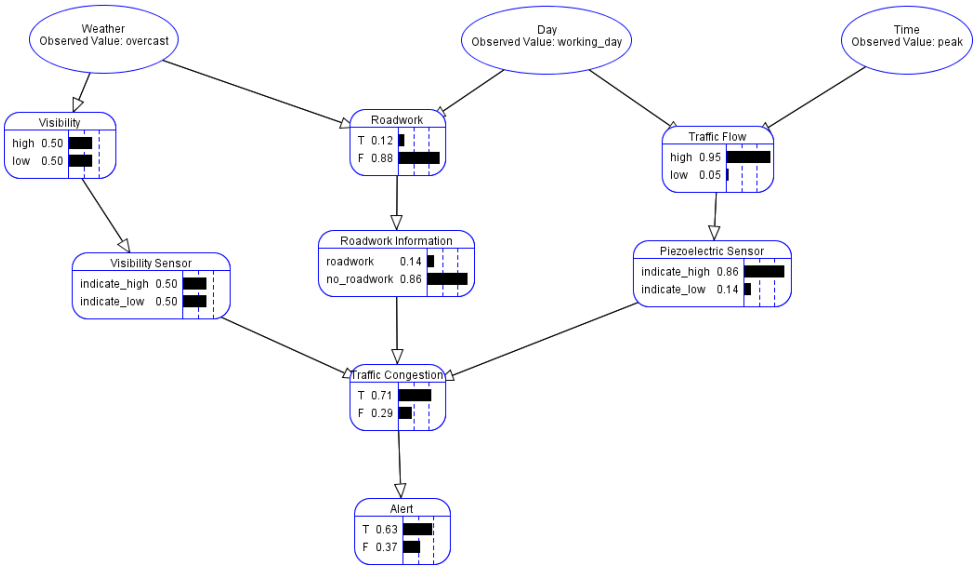

Figure 48: The agent observes that it is raining.

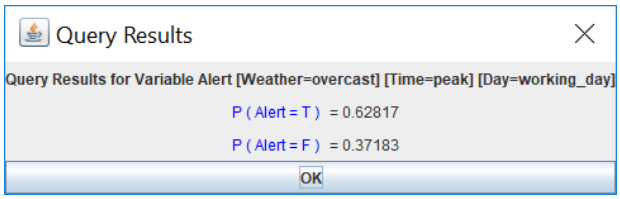

Figure 49: Probability of roadwork given that it is raining.

The fourth query states that there is no traffic congestion. In this case, the probability that this is currently peak time should decrease because high traffic flow causes traffic congestion, and peak time causes high traffic flow. The query result can be found in Figure 50 and Figure 51.

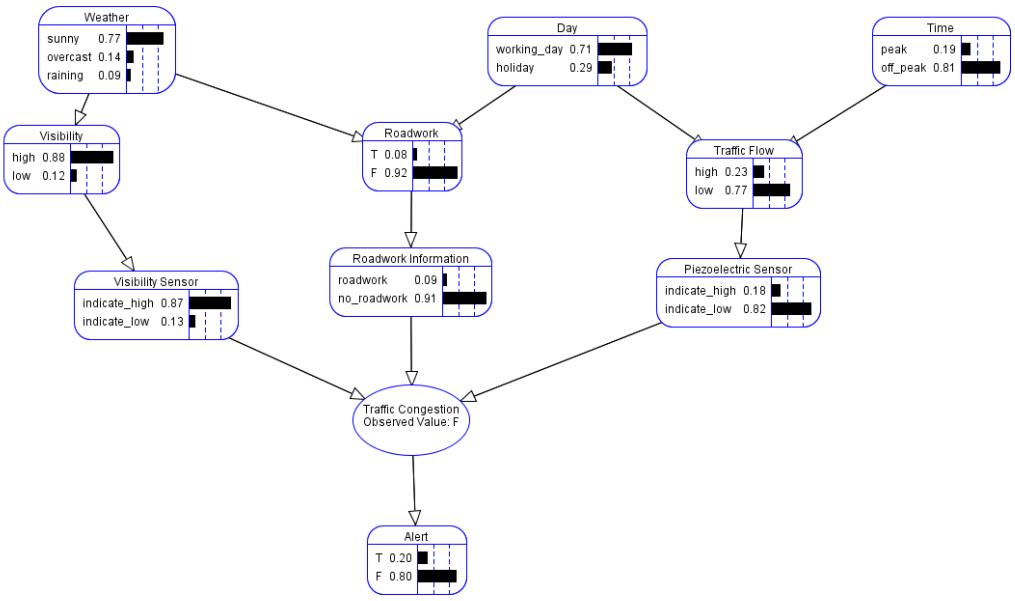

Figure 50: The agent observes that there is no traffic congestion.

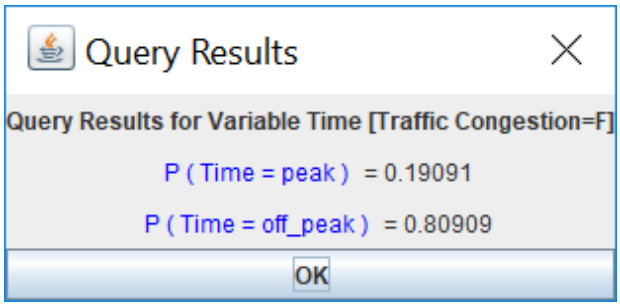

Figure 51: Probability of peak time given that there is no traffic congestion.

The fifth query states that there is roadwork. In this case, the probability that today is working day should be very high because there is high correlation between roadwork planning and working day. The query result can be found in Figure 52 and Figure 53.

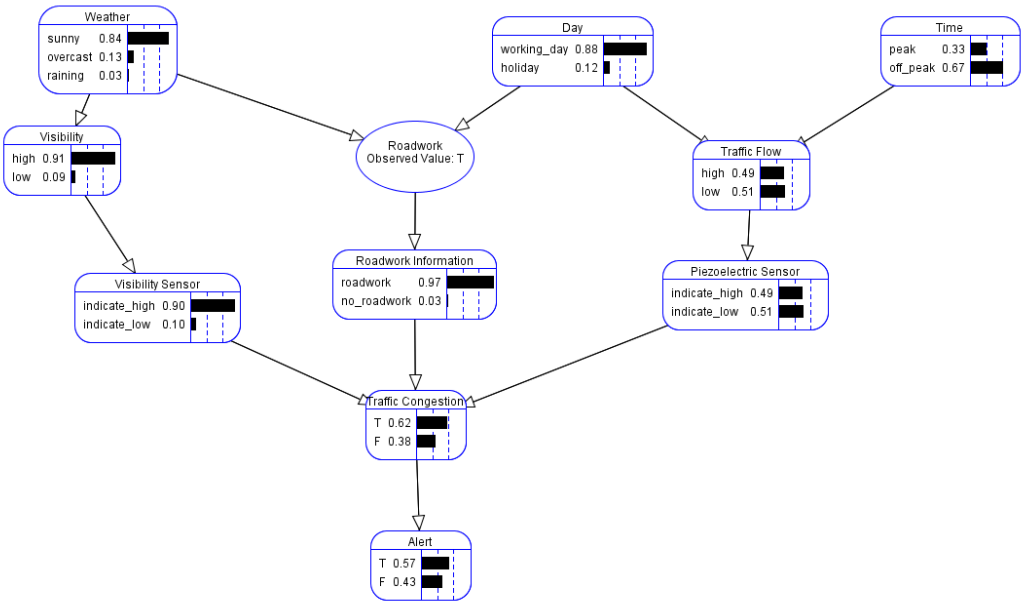

Figure 52: The agent observes that there is roadwork.

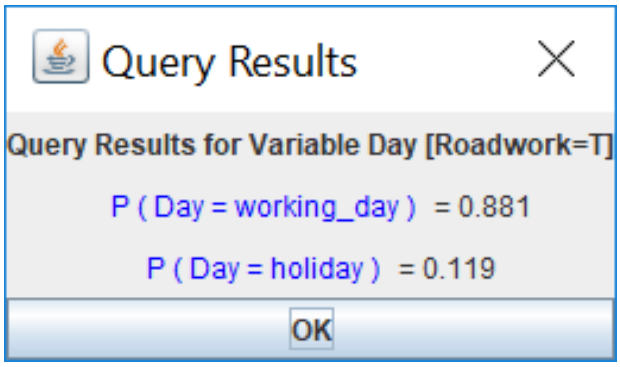

Figure 53: Probability of working day given that there is roadwork.

5 Part 3

5.1 Design

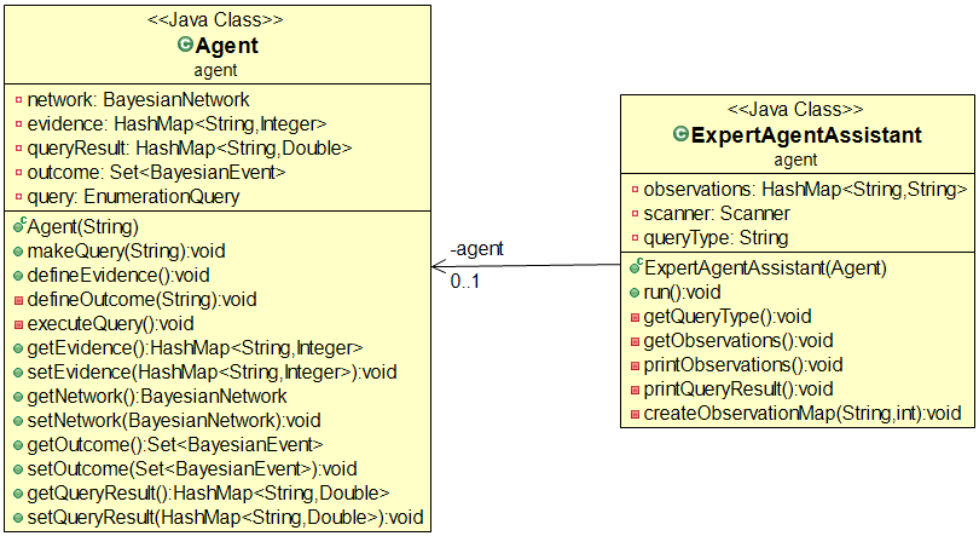

Figure 54: Class diagram of expert agent assistant application.

The class diagram of the system is shown in Figure 54. ***Agent*** class is responsible for using observations about the patient to make diagnostic or predictive query. ***ExpertAgentAssistant*** class interacts with the user. It gathers observations from the user and display the agent's query in a user friendly way. The flowchart showing execution of the expert agent assistant application is shown in Figure 55.

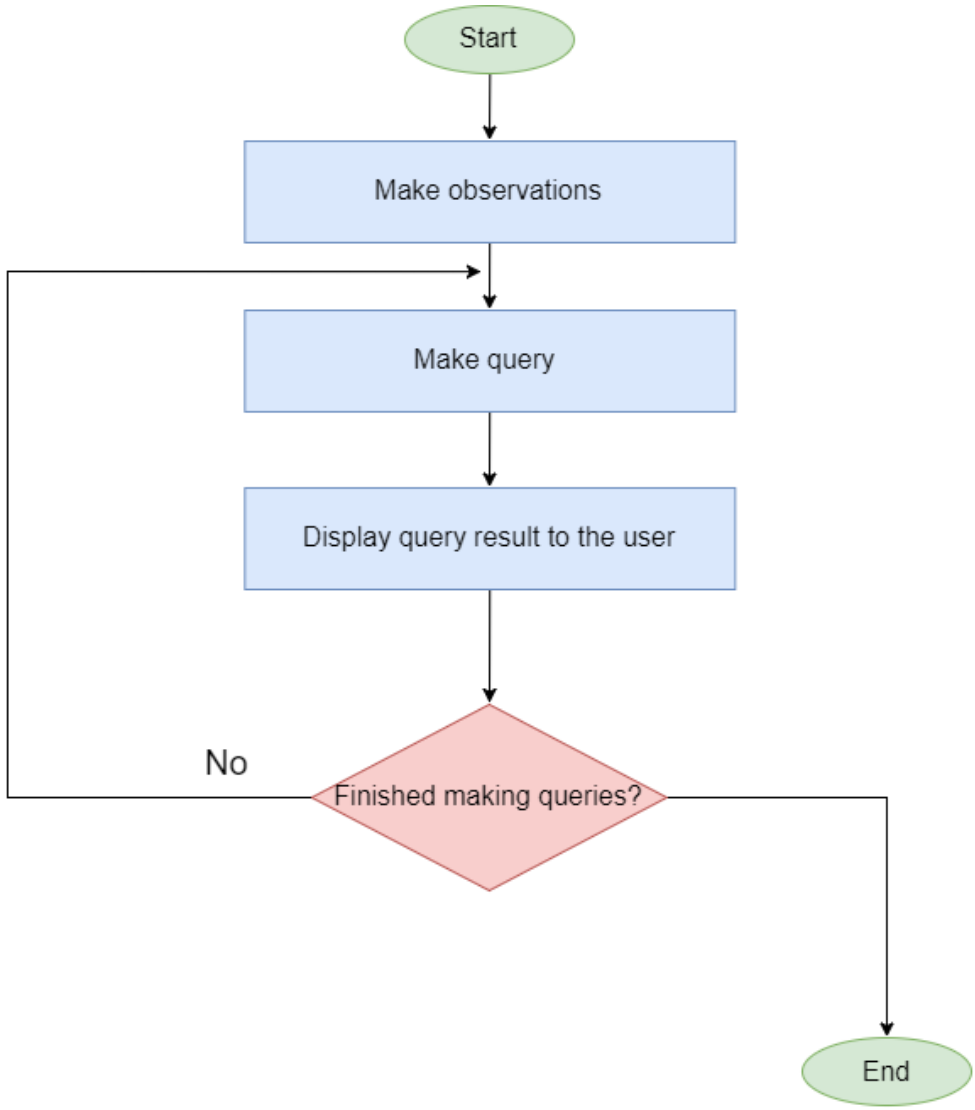

Figure 55: flowchart showing execution of the expert agent assistant application.

5.2 Examples and Testing

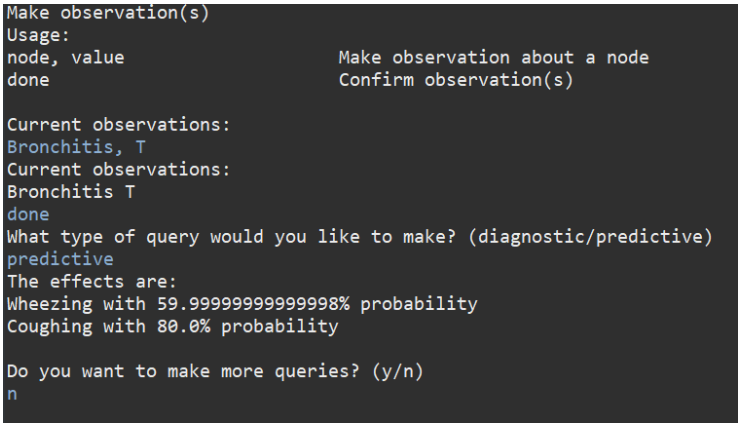

Figure 56: An example of program run. In this case, the agent observes that the patient has bronchitis.

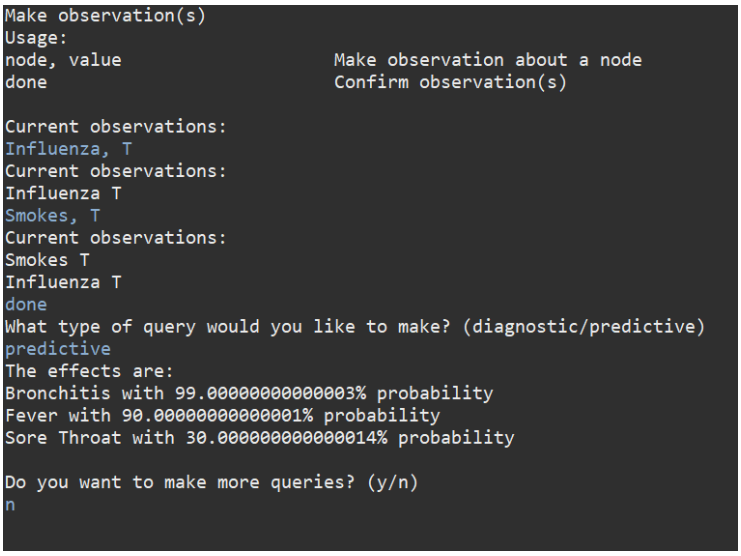

Figure 57: An example of program run. In this case, the agent observes that the patient is a smoker, and has influenza.

Examples of program run can be found in Figure 56 and Figure 57. The program let the user query the network using text-based interface.

5.3 Evaluation

The application developed has all basic features stated in the requirements. It can load the medical diagnosis network using Encog library. The expert agent assistant allows the user to interact with the network using text-based interface . The system supports diagnostic and predictive query. Due to time constraints, it is not possible to implement support for

other types of query. Another improvement is to allow the user to query specific node instead of displaying all possible causes and effects of the symptoms.

5.4 Running

prob.jar can be executed with the command:

```
java -jar prob.jar
```

Listing 1: The command used to run prob.jar.

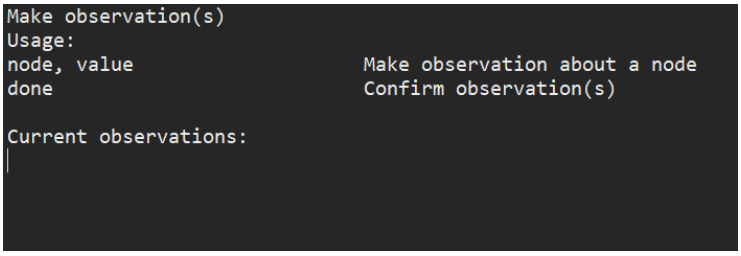

Figure 58: The application asks the user to make observations about the patient.

The system will ask the user to make observations about the patient (see Figure 58). If the user wants to make an observation that the patient has bronchitis, the command in Listing 2 should be used.

```
Bronchitis, T
```

Listing 2: The command used to make an observation.

Once the user has finished making observations, the command in Listing should be used to proceed to making queries.

```
done
```

Listing 3: The command used to confirm observations.

The next step is making queries. The user can choose whether they want to make diagnostic or predictive query (see Figure 59).

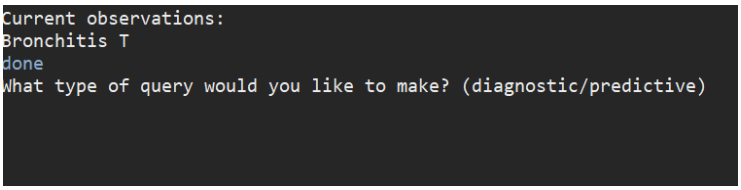

Figure 59: The application asks the user to make queries.

The system will display query result to the user (see Figure 60).

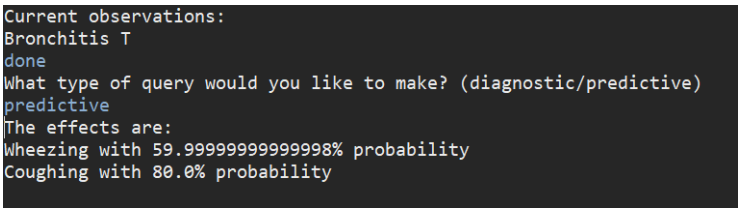

Figure 60: The application displays query result to the user.

Bibliography

AIspace (2016), 'Belief and decision networks', http://www.aispace.org/bayes/index.shtml. Accessed: 2017-12-17.

Data-Flair (2017), 'Top 10 real-world bayesian network applications', https://data-flair.training/blogs/bayesian-network-applications/. Accessed: 2017-12-21.

Research, H. (2017), 'Encog machine learning framework', http://www.heatonresearch.com/encog/. Accessed: 2017-12-17.

Russell, S. J. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach*, 3 edn, Pearson Education.

Toniolo, A. (2017), 'Cs5011: Assignment 4 - reasoning with uncertainty - bayesian networks', https://studres.cs.st-andrews.ac.uk/CS5011/Practicals/A4/A4.pdf. Accessed: 2017-12-17.