Bayesian Networks

CPSC 522 - AI (II)
January Assignment

By: Yaashaar Hadadian

Contents

1 Abstract

Bayesian Networks (also referred to as Belief Networks or Probabilistic Networks) is a Representation and Reasoning system (R&R) that is explicitly defined to exploit independence in probabilistic reasoning. The idea behind the Bayesian Network and its power is where, by exploiting the conditional independency of a variable X_i giving a subset of its predecessor nodes upon which the random variable directly depends, we can query as if we have the access over the full Joint Probability Distributions (JPD), while in fact, only conditional probability distributions were defined.
In practice, creating a Bayesian Network requires careful assessment of relationships among available random variables, to lead us to a compact representation of JPD that reduces the number of probabilities we need to specify. We would then be able to query the event of interest given the observations we made through some probabilistic inferring algorithms.

In this page, first, we introduce the Bayesian Network, then we discuss its internal structure and its constituents.Next, we present a step by step instruction for creating Bayesian Network from scratch, and ultimately we evaluate different possible structures of Bayesian Networks.Furthermore, we augmented our discussion in this page with an interesting case study as a practical example which we would work on it during different sections of this page to elaborate the concepts better.

2 Title

Bayesian Networks (also referred to as Belief Networks or Probabilistic Networks) as a Representation and Reasoning system (R&R), is a directed acyclic graphical representation for a joint probability distribution that is explicitly defined to exploit independence in probabilistic reasoning.

3 Builds on

Bayesian Network is one of the Graphical Models designed to depict a joint probability distribution for exploiting conditional (and marginal) independence in probabilistic reasoning.

4 Example Case Study: Suicide or death by natural cause !?

Let's say you are working for psychocriminology section of Vancouver police department.The department received a call and someone claimed that he found a student dead in his apartment. Autopsy results would not be released by next week, but the for some reason the police needs to find out the cause of death earlier.Therefore, the department puts you in charge of an investigation to find out the probability that he committed a suicide or he died of a natural cause. You are required to use your magical artificial intelligence skills to reason based on some pieces of information that you managed to extract from his diary that you found on his study table.

You decided to to represent the possible situations that lead to his death with a few indicator variables(only can be true or false):

- **Break Up (BU)**: Recent breaking up with somebody he was in relationship with
- **Relative Lost (RL)**: Recent lost of any close family member
- **Financial Problem (FNP)**: Living financially extremely tight for last few weeks
- **Work/Study Pressure (PRS)**: Being burdened with a lot of paperworks or assignments
- **No Friends & Family (NFF)**: Not having any family member or friend in the city
- **Insomnia (INS)**: Having recent insomnia
- **Serious illness (SIL)**: suffering from a serious illness
- **Depression(DEP)**: Suffering from depression
- **Emotional Shock (ES)**: Experiencing emotional shock in last few weeks
- **Chronic Stress (CST)**: Experiencing chronic
- **Prior Stroke (PS)**: Having a history of any type of stroke
- **Suicide (SU)**: Committing suicide

You soon realize a serious problem! Even if you would manage to find out the values of some of these [link]random variables from his diary, as there are 12 random variables, there would be 2^{12} different possible cases for each (boolean) variable; the [link] joint probability distributions for these 12 random variables would require you to define $2^{12} = 4096$ different numbers as probabilities to let you query them.

As you are banging your head against the wall, your colleague suggests you to create a Bayesian Network with these variables to substantially decrease the amount of probability values that you need to define based on his well written diary!

If you are interested in his suggestion, you need to be exposed to some definitions and formalism first.(I know...but, bear with me, we will get back to this.)

5 Definition

Bayesian Networks (also referred to as Belief Networks or Probabilistic Networks) is a Representation and Reasoning system (R&R) that is explicitly defined to exploit independence in probabilistic reasoning.It is considered a type of [Link] Markov Random Field in which, for each random variable, there is a factor that represents conditional probabilities and the directed graph is acyclic.

A Bayesian Network consists of:

1. A [link] Directed Acyclic Graph, in which, each node represents a random variable X_i
2. A domain set for each variable X_i
3. A set of conditional probability distributions for each variable X_i given its parents $Pa(X_i)$: $P(X_i \mid Pa(X_i))$

Dependencies between nodes (random variables) are reflected by directed edges in the graph.
The parents of a variable X_i, refers to as $Pa(X_i)$ are those variables that are represented as a minimal set of X_i's predecessor nodes in the graph, upon which X_i directly depends and thereby, is conditionally independent from other variables:

$$P(X_i|X_i - 1, ..., X_1) = P(X_i|Pa(X_i)) \tag{1}$$

Also based on [link] Chain Rule, we can turn a conjunction of random variables into conditional probabilities:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_i - 1, ..., X_1) \tag{2}$$

Therefore, based on equation number 1 and 2, a Bayesian Network can be seen as a compact representation of a conjunction or Joint Probability Distributions (JPD) of a set of random variables (X_1,X_2,...,X_n):

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \tag{3}$$

6 Dependencies in Bayesian Networks

In Bayesian Networks we may encounter with one of the following three cases among random variables' dependencies. These case are from Death Diagnosis Bayesian Network:

6.1 Common Ancestors

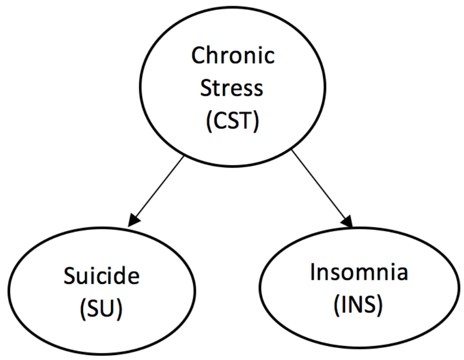


Figure 1: Dependencies : Common Ancestor

- Suicide and Insomnia are **dependant**; knowing the value of one, changes our belief about the probability of the other one.
- However, knowing the value of their common ancestor, Chronic Stress, makes Suicide and Insomnia to be **independent** of each other; by observing Chronic stress, knowing the value of one of Suicide and Insomnia doesn't change our belief about the other variable.

6.2 Chain

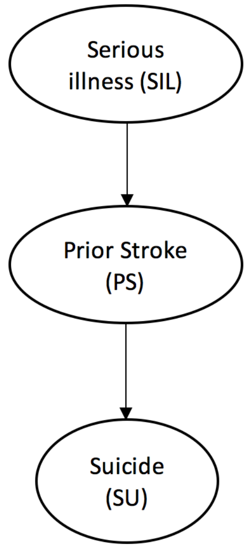


Figure 2: Dependencies : Chain

- Serious illness and Suicide are **dependent**; a change in value of one, affects on our belief about the other one.
- However, learning the value of Prior Stroke, makes them **independent** of each other; to know the effect of Serious illness, we no longer need to know Suicide's value, and to know the cause of suicide, we no longer need to know the value of Serious illness as its effect is **blocked** by Stroke.
- Intuitively, by knowing whether the subject had history of stroke, we can deduce that he had a serious illness either before having the stroke, or at least after having one, he was categorized as someone that had stroke which based on the assumptions of our simple example equals to having a serious illness.

6.3 Common Descendant

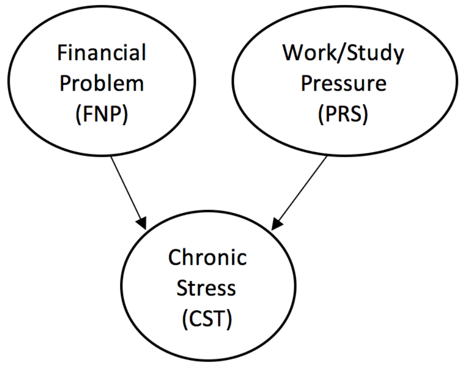


Figure 3: Dependencies : Common Descendant

- Financial Problem and Work/Study Pressure are **independent** of each other; knowing a value of one, doesn't change our belief about the other
- Learning Chronic Stress makes Financial Problem and Work/Study Pressure **dependent**; with observing Chronic Stress, any of the Financial Problem and Work/Study Pressure can be **explained away** by the other.

6.4 Parents and Markov Blanket

Among all variables in the variable set, there are some of them that directly affects variable X_i, which are called X_i's **Markove Blanket**, thereby, X_i is conditionally independent of other variables.

In other words, a node(random variable) is conditionally independent from other nodes(random variables) given its parents, children, and its children's parents.

Markov Blanket of Chronic Stress is indicated with grey nodes in the figure 4.

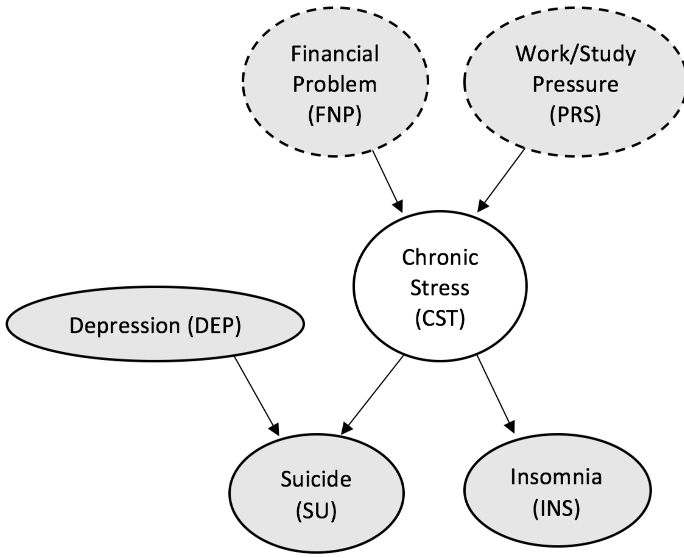


Figure 4: Markov Blanket (Grey Nodes) and Parents (Grey and Dashed Nodes) of Chronic Stress

Parents of a random variable X_i, $Pa(X_i)$, as small subset of its predecessor nodes, is a subset of its Markov Blanket, $MB(X_i)$:

$$Pa(X_i) \subseteq MB(X_i) \tag{4}$$

Parents of Chronic Stress are indicated with grey and dashed nodes in the figure 4.

7 Constructing Bayesian Networks

There is a straight forward recipe for creating a Bayesian Network out of a some random variables, which is organized and is presented in this section as simple ToDo List.

7.1 ToDo List

1. Defining a total order for our random variables (X_1,...,X_n)

2. Applying [Link] **Chain Rule**:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_i - 1, ..., X_1)$$

3. Identify the parents of each random variable X_i, so that, X_i is conditionally independent from all of its predecessors given $Pa(X_i)$:

$$P(X_i | X_i - 1, ..., X_1) = P(X_i | Pa(X_i))$$

4. With respect to the step 2 and 3, rewriting the equation to get a **compact representation** of the initial conjunction(joint probability distributions) of random variables as result of exploiting conditional independencies among them:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$$

5. Construct the network:

   - Draw **nodes** for random variables
   - Draw **directed arcs** from each random variable $Pa(X_i)$ to X_i
   - Define a **conditional probability table (CPT)** for each random variable X_i given its parents:

$$P(X_i | Pa(X_i))$$

7.2 Example

Now that you know the recipe, it is time to get back to our case study and following the above recipe to create a Bayesian Network:

1. **Total Ordering**
   In the introductory section of the case study, we chose our indicator(boolean) random variables for our domains, listed and clearly described them. Now we need to define an ordering that follows the natural sequence of events. One strategy(although not the best) could be to group similar variables together,define an ordering that makes sense within each group and then define a total ordering between groups and put their variables right after each other:

   Break Up (BU),Relative Lost (RL),Emotional Shock (ES),No Friends & Family (NFF),Depression(DEP),
   Financial Problem (FNP),Work/Study Pressure (PRS),Chronic Stress (CST),Insomnia (INS),
   Serious illness (SIL), Prior Stroke (PS),
   Suicide (SU)

2. **Applying Chain Rule**
   P(BU,RL,ES,NFF,DEP,FNP,PRS,CST,INS,SIL,PS,SU) = $P(BU) \times P(RL|BU) \times P(ES|BU, RL) \times$
   $P(NFF|BU, RL, ES) \times P(DEP|BU, RL, ES, NFF) \times$
   $P(FNP|BU, RL, ES, NFF, DEP) \times P(PRS|BU, RL, ES, NFF, DEP, FNP) \times$
   $P(CST|BU, RL, ES, NFF, DEP, FNP, PRS) \times P(INS|BU, RL, ES, NFF, DEP, FNP, PRS, CST) \times$
   $P(SIL|BU, RL, ES, NFF, DEP, FNP, PRS, CST, INS) \times$
   $P(PS|BU, RL, ES, NFF, DEP, FNP, PRS, CST, INS, SIL) \times$
   $P(SU, BU, RL, ES, NFF, DEP, FNP, PRS, CST, INS, SIL, PS)$

3. **Identifying Parents**
   P(BU,RL,ES,NFF,DEP,FNP,PRS,CST,INS,SIL,PS,SU) =

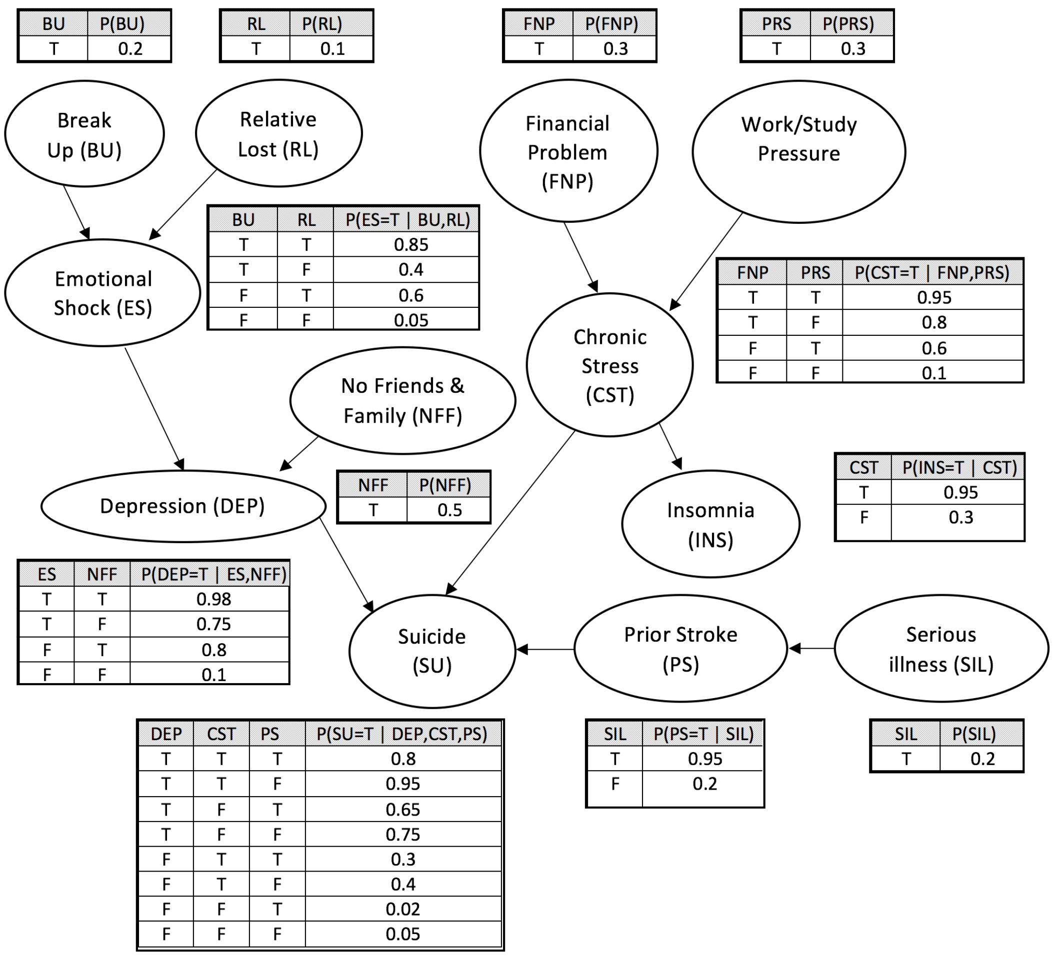
   - **P(BU)**: has only one variable and nothing to play with
   - × **P(RL|BU)** : Does relative Lost (RL) depends on Break Up (BU) or in other words, is BU parent of RL? the answer is No, therefore we get rid of BU → P(RL)
   - × **P(ES|BU,RL)**: Break Up (BU) and Relative Lost (RL) both could be reasonable parents for Emotional Shock, therefore, we would keep this conditional probability as it is
   - × **P(NFF|BU,RL,ES)**: No Friends and Family (NFF) doesn't get influenced (directly) by BU,RL or ES, therefore we remove these variables from this conditional dependency → P(NFF)
   - × **P(DEP|BU,RL,ES,NFF)**: Depression (DEP) could be a direct result of having No Friends and Family (NFF) and Emotional Shock (ES); NFF and ES could be parents of DEP, but BU and RL doesn't directly affect depression → P(DEP|ES,NFF)
   - ... *Repeating this process for the remaining conditional probabilities as well ...*

4. **Rewriting**
   Finally from previous step we would get:

$$\begin{aligned} P(BU, RL, ES, NFF, DEP, FNP, PRS, CST, INS, SIL, PS, SU) = \\ P(BU) \times P(RL) \times P(ES|BU, RL) \times P(DEP|ES, NFF)\times \\ P(FNP) \times P(PRS) \times P(CST|FNP, PRS) \times P(INS|CST)\times \\ P(SIL) \times P(PS|SIL) \times P(SU|DEP, CST, PS) \end{aligned} \tag{5}$$

5. **Constructing The Network**
In the final step we draw nodes, their parents and arcs from parents to nodes based on the propositions in the last step. Ultimately we will draw a conditional probability table (CPT) for each node and fill each cell of each CPT according to information we already extracted from diary of the student. Figure 5 shows our final Bayesian Networks for the case study.
Note that, we only define probability of the event to be equal to True, for each row of our CPTs and the probability of the event to be equal to False would be computed based on the probability axiom:

$$P(\neg a) = 1 - P(a)$$
$$\rightarrow P(X_i = F) = 1 - P(X_i - T)$$



BU	P(BU)
T	0.2

RL	P(RL)
T	0.1

FNP	P(FNP)
T	0.3

PRS	P(PRS)
T	0.3

BU	RL	P(ES=T | BU,RL)
T	T	0.85
T	F	0.4
F	T	0.6
F	F	0.05

FNP	PRS	P(CST=T | FNP,PRS)
T	T	0.95
T	F	0.8
F	T	0.6
F	F	0.1

NFF	P(NFF)
T	0.5

CST	P(INS=T | CST)
T	0.95
F	0.3

ES	NFF	P(DEP=T | ES,NFF)
T	T	0.98
T	F	0.75
F	T	0.8
F	F	0.1

DEP	CST	PS	P(SU=T | DEP,CST,PS)
T	T	T	0.8
T	T	F	0.95
T	F	T	0.65
T	F	F	0.75
F	T	T	0.3
F	T	F	0.4
F	F	T	0.02
F	F	F	0.05

SIL	P(PS=T | SIL)
T	0.95
F	0.2

SIL	P(SIL)
T	0.2

Figure 5: Student's Death Diagnosis Bayesian Network Model

In reality, CPT's numbers are provided by **Domain Experts**, or from **Previous Observations/Records**, or though **Machine Learning** Techniques.

8 Compactness of Bayesian Networks

As we saw earlier in the introductory section of case study, to represent the joint probability distributions for N random variable(s) with domain size of D each with K parent(s), we would have required to specify:

$$D^K$$

numbers. In the worst case, the number of parents of each node(K), could be approximately equal to the number of variables (N), thus, the complexity would be:

$$K \approx N \rightarrow O(D^N)$$

In this case the network is **fully connected**.Bayesian Networks are highly beneficial when, each of its variables, only interacts with a small fraction of other variables($K \ll N$). To reach such a structure, sometimes, some simplification assumptions should be made to reduce the dependencies among variables in the network and to **relax** the problem.

In our example we would have required to specify 2^{12} numbers for presenting the full joint probability distributions in the worst case scenario (fully connected network),however, by exploiting the conditional independences among variables through the Bayesian Network, we only defined:

$$N \times D^K$$

8.1 Trade off Between Structures

Variables ordering is the first and probably the most important step in constructing Bayesian Network, since different ordering of variables, leads to different structure of networks.For example, we could define different variables ordering in the first step:

$$SIL, PS, SU, DEP, ES, BU, RL, NFF, FNP, CST, PRS, INS$$

and then apply the chain rule in the second step and identify different set of parents for each variable in the third step:

$$\begin{aligned}
&P(SIL, PS, SU, DEP, ES, BU, RL, NFF, FNP, CST, PRS, INS) = \\
&\quad P(SIL) \times P(PS|SIL) \times P(SU|PS) \times P(DEP|SIL, SU) \times \\
&\quad P(ES|SIL, SU, DEP) \times P(BU|ES) \times P(RL|ES) \times P(NFF|SU, DEP) \times \\
&\quad P(FNP|SU, DEP, NFF) \times P(CST|SIL, SU, DEP, FNP) \times \\
&\quad P(PRS|SU, DEP, CST) \times P(INS|SU, DEP, CST, PRS)
\end{aligned} \tag{6}$$

Equation 6 and the equation 5, that we used before in the original solution, would be equivalent if the Conditional Probability Tables (CPT) are carefully specified to satisfy the conditional dependencies among variables and to result in the same probability; if CPTs make equation 6 and 5 to be equal to each other.

Moreover, it would be much efficient if pay an extra attention in step 3, where we identify random variable's parent nodes and if we make that our different structure doesn't result in having extra dependencies among variables than what are considered to be the direct ones, otherwise, we would require to specify more numbers to fill in the CPTs. Figure emphasize the significance of this matter by illustrating the alternative Bayesian Network (bottom) that could be yield as a result of different variables ordering which corresponds to equation 6, in comparison with the former one (top), which we initially created based on equation 5.

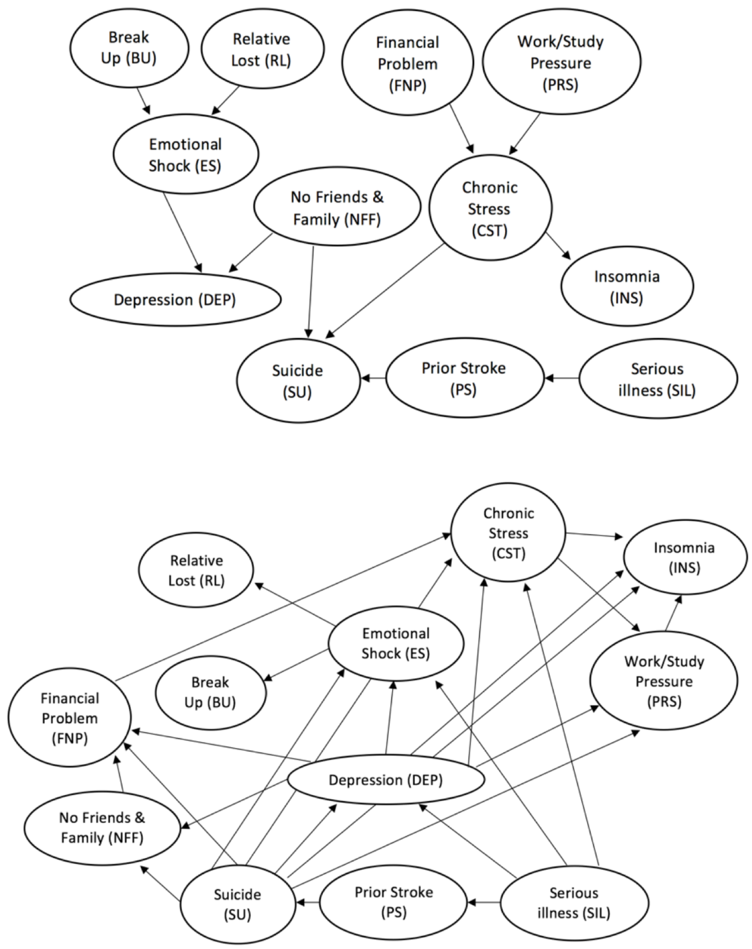


Figure 6: Student's Death Diagnosis Bayesian Networks, original solution (top), and alternative one (bottom)

It is extremely important to carefully assign the variables ordering as in the realistic models, the network can get easily complicated because of the number of random variables involved. Figure 7 is related to the Bayesian Network designed for diagnosis of liver disorders [1], which can illustrate the importance of only keeping the direct dependencies and perhaps even incorporating some relaxing assumptions.

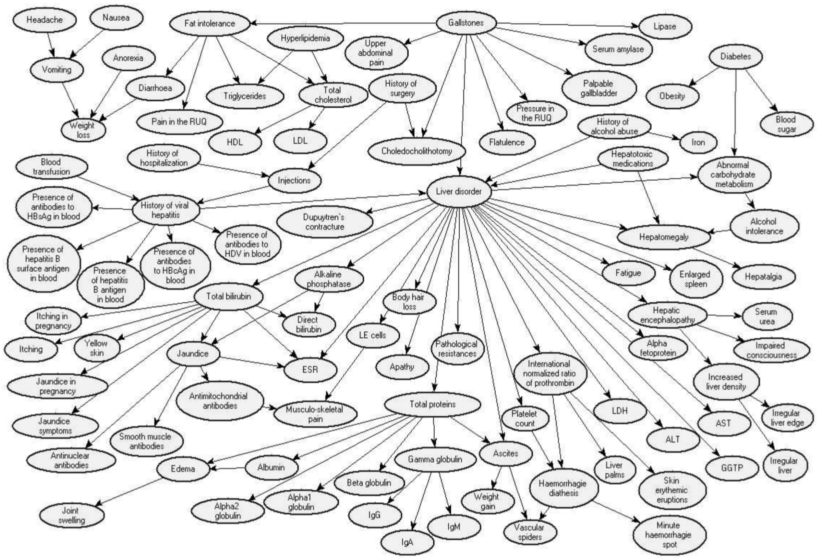


Figure 7: Bayesian network model for liver diagnosis [1]

8.2 The Power of Bayesian Network

The idea behind the Bayesian Network and its power is where, by exploiting the conditional independency of a variable X_i giving a subset of its predecessor nodes, we can query as if we have the access over the full joint probability distributions, while in fact, only conditional probability distributions were defined.
In practice, we use some algorithms to query the Bayesian Network and to compute the probability of the event of interest giving the observations we made.This procedure is called **probabilistic Inference**.
Based on the scale of the network we can perform either:

- **Exatct Inference** through [Link]Variable Elimination technique and algorithms like [Link]Viterabi
- **Approximate Inference** through [Link]Forward sampling, Rejection Sampling, Importance Sampling, or Particle Sampling

These techniques and algorithms are described and discussed in area of [Link]Probabilistic Inference.

References

[1] A. Onisko, M. J. Druzdzel, and H. Wasyluk, "A bayesian network model for diagnosis of liver disorders," in *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering*, vol. 2. Citeseer, 1999, pp. 842–846.