Finding Minimum Data Requirements Using Pseudo-Independence

Yoonheui Kim
Computer Science Dept.
University of Massachusetts,Amherst
Amherst, MA 01002

Victor Lesser
Computer Science Dept.
University of Massachusetts,Amherst
Amherst, MA 01002

Abstract

In situations where Bayesian networks (BN) inferencing approximation is allowable, we show how to reduce the amount of sensory observations necessary and in a multi-agent context the amount of agent communication. To achieve this, we introduce Pseudo-Independence, a relaxed independence relation that quantitatively differentiates the various degrees of independence among nodes in a BN. We combine Pseudo-Independence with Context-Specific Independence to obtain a measure, Context-Specific Pseudo-Independence (CSPI), that determines the amount of required data that needs to be used to infer within the error bound. We then use a Conditional Probability Table-based generation search process that utilize CSPI to determine the minimal observation set. We present empirical results to demonstrate that bounded approximate inference can be made with fewer observations.

1. Introduction

Bayesian networks (BN) provide a quantitative representation of relations among variables based on degrees of their dependence. However there are many independence relations such as context-specific independence [2] in Bayesian networks that are not captured in inference using traditional algorithms. Variables that are actually independent of each other are still involved in computing probabilities on each variable and add unnecessary complexity to inference tasks. There has been a long-term interest in approximating various inference tasks in BNs to reduce the complexity [4, 6, 5]. One of the formal ways of reducing complexity is to use independence relations not shown in the structure in Bayesian networks. Given this extra information on independence, a bigger set of variables can be ignored during inference without any loss of accuracy.

There has also been work on simplifying the reasoning in BN by reducing the size of the network itself as in many networks there are a lot of loose connections between variables (where disconnecting arcs do not significantly change the relation among variables in most cases)[4]. We were also interested in simplifying the network by disconnecting arcs among these variables where it does not significantly change the relation among variables. However, we have found that the degree of dependency is contingent on the context (i.e., values of observation variables) and many arcs cannot be permanently deleted from the network without a severe effect on the accuracy of inferencing.

Instead of altering the networks, we focus on reducing the amount of information necessary for computing posterior probabilities by exploiting weak dependence among nodes where they usually only play a role in a limited context of variables. We try to find a sufficient set of observations that make further observations unnecessary. Information that makes other information irrelevant is depicted as an independence relationship in BN. In particular, we look at Context-Specific Independence [2] that formalizes independence given posterior information. In addition, we trade off the cost of inferencing against accuracy by defining *Pseudo-Independence* which bounds the error in reasoning. We combine Context-Specific Independence and Pseudo-Independence and we model the minimum data requirements problem as a search for these weak dependence relations.

We formulate the problem of determining the minimum data requirements for an approximate reasoning task on BNs. Instead of focusing on saving computation, we focus instead on reducing the burden of data processing such as observation gatherings which arises during inference in particular domains of BN. We consider domains where a subset of hidden variables (about which information is not available) is inferred from observable variables and the property of observability does not change over time. Additionally, we assume collecting an observation has a communication cost, i.e. in a distributed environment.

Our algorithm can be directly applied to the sensor network domain by modeling the reasoning of the sensor network as a BN. Sensor readings are observational nodes in BNs. Since communicating or obtaining observations consume resources, this makes reducing the number of observations important. Additionally, because of the limitation of computation power and storage, it is reasonable to compute the probabilities off-line and then save them in a compact way. There is a large literature which uses Bayesian approaches for reasoning in sensor network domains e.g. [1, 3, 8]. We can think of two scenarios in the sensor network system for the use of our approximation inferencing algorithm: 1) a sensor network system with distributed reasoning and observations. Currently many systems perform heuristic reasoning without considering sensor nodes located geographically distant. Instead of ignoring those nodes, we can determine what to communicate and communicate only those observations relevant to reasoning; and, 2) a sensor network system where data can be obtained sequentially. Instead of performing all the measurements at the same time, we can make observations incrementally so as to minimize the number of observations and amount of communication.

For a systematic and efficient search for Pseudo-Independence and Context-Specific Independence, we use a tree structure called CPT-tree (Conditional Probability Table-tree) [7] which represents conditional probability tables (CPT) in a compact way with related variables on internal nodes, values of those variables on arcs and the probabilities of inferred variables on leaf nodes (See Figure 2). Using a tree structure for a representation of simplified networks is not uncommon. Chow-Liu tree [5] is an approximate Bayesian network that is constructed by modifying the network into a tree structure that ignores some interdependencies among variables. In Friedman *et al.* [7], CPT-trees are used to illustrate the local relation of parent and child nodes depicted in conditonal probability tables given in the network. Figure 2(a) shows the CPT-tree on WetGrass as in the paper. Also, the probabilities are exact given values on parent nodes. For instance, given the example network (A modified WetGrass network) in Figure 1, CPT $P(WetGrass|Rain, Sprinkler)$ is given in the network, and the CPT-tree in Figure 2(a) is constructed. In our representation, instead of using CPTs given in the network we construct CPT-trees on the variables we infer given the observable nodes in the network and represent it as a CPT-tree.

In the example in Figure 1 and 2(b), we try to find a minimum set of observations to infer whether it had rained overnight (We infer node Rain). The CPT-tree shown in Figure 2(b) represents the minimum data requirement for an inference on the node Rain in Figure 1 where we allow up to a maximum 0.05 error in the probabilities. The nodes on each path are data O required to calculate $P(Rain = T|O)$ bounding the error to 0.05. The CPT-tree in Figure 2(b) shows that in no situation do we need the observation on node 6 in Figure 1. In addition, when the observation WetGrass is false, we do not need to collect any more information. Having paths with various lengths from root to leaf nodes, CPT-trees which we use in our framework straightforwardly represent the observation savings and the resulting probabilities.

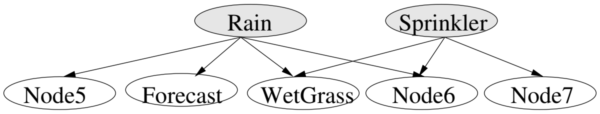


Figure 1. A Modified WetGrass network

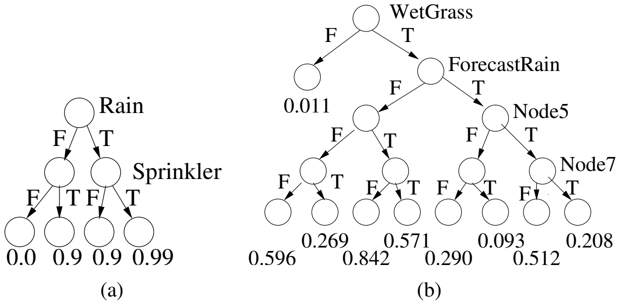


Figure 2. (a) The CPT-tree on WetGrass constructed as in Friendman *et al.* [7]. (b) The CPT-tree from the WetGrass network on $P(Rain|o)$ with an error bound of 0.05 where o is the values of observations on the path from root to leaf nodes in the tree.

Many inferencing problems in BNs can be viewed as a search where the search space is usually too large to fully explore. A similar problem like the Maximum A Posteriori problem that finds the best possible observations given a partial set is solved as a branch and bound search [10, 13]. In this problem, we search for the CPT-tree with the minimum number of leaf nodes necessary for accurate inferencing, which gives us the minimum data requirements. In building a CPT-tree, we add the most informative observation first to the tree until the current observation set satisfies one of two constraints that check the sufficiency of information for a correct reasoning.

In this paper, we first define the notion of Pseudo-Independence and extend Context-Specific Independence (CSI) based on Pseudo-Independence. We then define the problem of Minimum Data Requirement and construct the algorithm for finding a minimal set of observations for a given error bound. Finally in section 5, we analyze the efficiency and correctness of the algorithm.

2. Pseudo-Independence

There has been long-term interest in formalizing and finding independence relations not shown in structures of Bayesian networks (BN) [2, 11, 14]. These works are focused on precise independence with an exact equality where $I(X; Y|Z)$, independence of random variables X and Y given information Z is defined as $I(X; Y|Z) \Leftrightarrow P(X|Z) =$

$P(X|Y, Z)$.

When the interaction between variables is very weak, the probability of the variables changes minimally as values of weakly dependent variables change. By considering weak dependence as independent, we are able to save resources when inferencing in BN.

We introduce a new independence relation, Pseudo-Independence, which relaxes the strict conventional independence. Pseudo-Independence $I_{s(b)}$ as in Definition 2, 3 is an independence relation that allows some margin b over probabilities. To our best knowledge, our work is the first to formalize Pseudo-Independence in solving problems in BN. Pseudo-Independence is particularly advantageous when there is posterior information and the number of possible consequential probabilities explodes. This number reduces if we identify some of the observations are not relevant within the error bound. For instance, a network with 10 observational nodes with 2 possible values results in 2^{10} observational cases and the same number of consequential probabilities. If we identify that 3 of the observable nodes are independent given the error bound, we can reduce the number to 2^7. Both reasoning on each case and recording each result are expensive as observational cases are exponential in the number of observations.

Context-Specific Independence (CSI) [2], independence relation arising from posterior information, describes the hidden independence relation where the values of some observable nodes are determined and results in some variables becoming independent. However, their relationships are contingent on parameter values and are not readily available as the ones that are posed on the static structure of the BN. No clear rules exists such as D-separation, which describes independence relations based on the structure of Bayesian network that are helpful for finding these relations. Moreover, although CSI is useful, the number of possible sets in CSI relations is potentially exponential in the number of variables because the sets Y and Z in $I(X; Y|Z)$ are not determined beforehand. Some researchers have tried to find CSI relationships which were strictly defined to a small set of comparisons such as parent and child nodes [11], but this greatly limits the possible use of CSI. As an alternative approach, we develop an efficient search for these relations using KL-divergence, and target a specific set of independence relations that are relatively useful for our inference tasks. In addition to the use of this search for CSI relations, the explosion of these relations can be reduced by Pseudo-Independence, which allows loose connections to be viewed as independent. This enables us to ignore small differences in the probabilities caused by some observations and many observation cases can be summarized by a smaller set.

Context-Specific Independence [2] is defined given context, i.e., value-assigned variables. Consider a finite set $U = X_1, \ldots, X_n$ of discrete random variables where each variable $X_i \in U$ may take on values from a finite domain. We use capital letters, such as X, Y, Z for variable names and lower case letters x, y, z to denote specific values taken by those variables. A set of variables is denoted by boldface capital letters **X, Y, Z**. The set of all values of variable X is denoted val(X).

Definition 1 (Context-Specific Independence) *: Let **X,Y,Z,C** be a pairwise disjoint set of variables. X and Y are contextually independent given Z and the context c* $\in$ *val(**C**), denoted* $I_c(X; Y|Z, c)$, *if* $P(X|Z, c, Y) = P(X|Z, c)$ *whenever* $P(Y, Z, c) > 0$.

Definition 2 (Pseudo-Independence 1) *: Let P be a joint probability distribution over the variables in U, and let **X,Y,Z** be subsets of U. **X** and **Y** are pseudo-independent given **Z**, denoted* $I_{s(b)}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$ *if for all* $x \in val(X), y \in val(Y), z \in val(Z)$, *the following relationship holds:*
$|P(x|z, y) - P(x|z)| \leq b$ *whenever* $P(y, z) \geq 0$ *where* $P(x|y, z)$ *is defined.*

Definition 3 (Pseudo-Independence 2) *: Using the same notations, **X** and **Y** are pseudo-independent given **Z**, denoted* $I_{s(b)}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$ *if for all* $x \in val(X), y \in val(Y), z \in val(Z)$, *the following relationship holds:*
$|P(x|z, y)/P(x|z)| \leq b$ *whenever* $P(y, z) \geq 0$.

The first definition of Pseudo-Independence is intuitive. It bounds the difference between $P(\mathbf{X}|\mathbf{Z})$ and $P(\mathbf{X}|\mathbf{Y}, \mathbf{Z})$, which is required to be 0 for strict independence. In the second definition, the proportion of two probabilities should remain within the bound, which in turn allow us to bound KL-divergence [9] to a constant value. KL-divergence is a metric used in information theory to determine how similar two probability distributions are. KL-divergence $D_{KL}(P \parallel Q)$ of two probability distributions P,Q is defined as $\sum_i P(i) \log \frac{P(i)}{Q(i)}$. For any additional observation value set y, KL-divergence between $P(\mathbf{X}|\mathbf{z}, \mathbf{y})$ and $P(\mathbf{X}|z)$ is bounded by $\sum_{\mathbf{X}} \log b$. Although the second definition has the nice property of having the bound of KL-divergence, we focus on the first definition of Pseudo-Independence in this paper.

We can now extend Context-Specific Independence to Context-Specific Pseudo-Independence which is used as a termination criteria in our algorithm for finding minimum data requirements.

Definition 4 (Context-Specific Pseudo-Independence) *: Let **X,Y,Z,C** be pairwise disjoint sets of variables. X and Y are contextually pseudo-independent given Z and the context c* $\in$ *val(**C**), denoted* $I_{sc(b)}(\mathbf{X}; \mathbf{Y}|\mathbf{Z}, \mathbf{c})$, *if* $|P(X|Z, c, Y) - P(X|Z, c)| \leq b$ *whenever* $P(Y, Z, c) > 0$.

3. Minimum Data Requirements

In an environment where obtaining information is costly, it is useful to trade off the amount of information used in reasoning and accuracy. We define the problem of finding minimal data requirements of a Bayesian network to decide a compact observation set sufficient for approximate inference with a well-defined bound on error. The minimal data requirements problem can be viewed as finding a compact conditional probability table using additional independence relations in the Bayesian network by making a trade-off between accuracy and compactness. Shen *et al.* [12], have tried to achieve a similar reduction of data to communicate within the Bayesian networks by formulating a approximate decision problem using DEC-POMDP for determining which information to send; however they use the complete independence assumption between multiple agents' networks that makes the problem much simpler. Yang *et al.* [15] tries to obtain a message passing/communication policy in multiply-sectioned Bayesian networks for achiving exact inference when each agent does not reveal the local structure. The communication policy in this framework is not designed for communication savings and the number of messages for reaching the solution remains the same for any observation. We search for Pseudo-Independence among variables such that partial observation set makes inference unncessary for other remaining observations. The output of the search are the sets of observations and resulting probabilities that summarize possible observation cases similar to a compact conditional probability table (see Figure 2). The input and output of the problem are summarized as follows.

Input : Bayesian network, Hypothesis variable set (The variables of reasoning), The error bound d.
Output : The CPT-tree with minimal observations and the resulting probabilities.

Definition 5 (Minimum Data Requirements) *MDR(BN, b, X): Let a Bayesian network BN have nodes* $< \mathbf{O}, \mathbf{H} >$ *where* $\mathbf{O}$ *is the set of observational nodes and H is a set of hidden nodes. Let* $\mathbf{o_p} \in val(\mathbf{O_p})$ *where* $\mathbf{O_p} \subset \mathbf{O}$ *and* $\mathbf{X} \in \mathbf{H}$ *a set of nodes to be reasoned about.* $MDR(BN, b, \mathbf{X})$ *is a set S of tuples,* $< \mathbf{o_p}, P(\mathbf{X}|\mathbf{o_p}) >$ *where S is constructed by considering all values of observation set* $\mathbf{O}$ *factoring out observable variables whose values affect* $P(\mathbf{X}|\mathbf{o_p})$ *only within a small bound b. The goal is to find a set* S_{min} *with minimum number of tuples* $S_{min} = argmin_S|S|$ *where* $|S|$ *corresponds to the number of leaf nodes in a CPT-tree.*

A naive form of solution S is the conditional probability table where $\mathbf{o_p} = \mathbf{o}$ for all tuples $< \mathbf{o_p}, P(\mathbf{X}|\mathbf{o_p}) >$ in S. A tuple $< \mathbf{o_p}, P(\mathbf{X}|\mathbf{o_p}) >$ describes the probability $P(\mathbf{X}|\mathbf{o_p})$ where observation o has a value $\mathbf{o_p}$ on a subset of observation $\mathbf{O_p}$. For instance, suppose a network with two observations {WeatherForecast, WetGrass} and we have a tuple $< WetGrass = F, 0.01 >$ with a given error bound 0.05. The tuple means the probability $P(\mathbf{X}|\mathbf{o})$ satisfies $0.01 - b \leq P(\mathbf{X}|\mathbf{o}) \leq 0.01 + b$ for any $\mathbf{o} \in val(\mathbf{O})$ where $WetGrass = F$, regardless of the value of $WeatherForecast$. The change in the value of observation WeatherForecast only leads to a difference of probability $|P(X|WetGrass = F) - P(X|WetGrass = F, WeatherForecast)|$ within the bound of 0.05.

4. Solving Data Requirement as Search

We solve the Minimum Data Requirement problem as a search for a CPT-tree with a minimum number of leaf nodes. We formulate a forward search with incremental pruning that adds one observation to the tree at a time as the size of fully expanded tree can be too large to handle and explore. Two problems arise in building up the tree. One is to determine which evidence to add to the tree. The other is to determine when to stop expanding branches of the tree. When the search terminates, the probability of $\mathbf{X}$ given observations on the path from root to leaf is computed and associated with the corresponding leaf node.

For the first problem of choosing the most informative observation, we compute KL-divergence [9] of two probability distributions, one with the given observation set $\mathbf{O}_p$ and the other with an additional observation O_n. The KL-divergence of two probability distributions on a variable set $\mathbf{X}$ indicates information gain due to the additional observation O_n.

$$D_{KL}(P(\mathbf{X}|\mathbf{o_p} \cup \{O_n\}||P(\mathbf{X}|\mathbf{o_p})) = \sum_{o_n \in val(O_n)} \sum_{\mathbf{x} \in val(\mathbf{X})} P(\mathbf{x}|\mathbf{o_p}, o_n) \log(\frac{P(\mathbf{x}|\mathbf{o_p}, o_n)}{P(\mathbf{x}|\mathbf{o_p})}) \tag{1}$$

This corresponds to Line 21-24 of Algorithm 1. Using KL-divergence as a criteria to reduce the search space is relatively common approach in various domains. For example, in Figure 1, when $\mathbf{X} = \{Rain\}, \mathbf{O}_p = \emptyset$, KL-divergence $D_{KL}(P(Rain|\{O_n\}||P(Rain)) = \Sigma_{o_n} \Sigma_{Rain=T,F} P(Rain|\{o_n\}) \log(\frac{P(Rain|\{o_n\})}{P(Rain)})$. For each observation o_n, we compute D_{KL} in Equation 1. $O_n = WetGrass$ is selected to be added to the tree because it has the largest D_{KL} value.

We solve the second problem of determining when to stop, using two stopping criteria. The search continues to add the most informative observation to the tree until every observation set satisfies one of the following two stopping criteria. The first stopping criterion uses Context-Specific Pseudo-Independence and D-separation rules to find set $\mathbf{Y}$ that contains all observation values already obtained and satisfies Context-Specific Pseudo-Independence $I_{sc(b)}(\mathbf{X}; \mathbf{O} - \mathbf{Y} \mid \mathbf{Y} - \mathbf{Y}', \mathbf{y}')$. The first stopping criterion uses independence relations in Bayesian networks as follows: Let $\mathbf{X}, \mathbf{O}$, and $\mathbf{Y}$ be sets of variables where $\mathbf{O}$

is a set which contains all observation nodes, $\mathbf{Y}$ a subset of $\mathbf{O}$, $\mathbf{y} \in val(\mathbf{Y}), \mathbf{o} \in val(\mathbf{O})$. If $I(\mathbf{X};\mathbf{O}-\mathbf{Y} \mid \mathbf{Y})$, then $P(\mathbf{X} \mid \mathbf{Y})$ bounds the probabilities of set $\mathbf{X}$ on all observation cases. Because of the independence $I(\mathbf{X};\mathbf{O}-\mathbf{Y} \mid \mathbf{Y})$, $P(\mathbf{X} \mid \mathbf{O}) = \mathbf{P}(\mathbf{X} \mid \mathbf{Y})$. For each $\mathbf{o} \in val(\mathbf{O})$, there exists a value set $y \in val(\mathbf{Y})$ that satisfies $P(\mathbf{X} \mid \mathbf{o}) = P(\mathbf{X} \mid \mathbf{y})$. Therefore for all $\mathbf{o}$, $minP(\mathbf{X} \mid \mathbf{Y}) \leq P(\mathbf{X} \mid \mathbf{o}) \leq maxP(\mathbf{X} \mid \mathbf{Y})$. Likewise, for value $\mathbf{y'} \in val(\mathbf{Y'})$, subset of $\mathbf{Y}$, $P(\mathbf{X} \mid \mathbf{Y}_{\mathbf{Y'}=\mathbf{y'}})$, the probability where the value of subset $\mathbf{Y'}$ is fixed to $\mathbf{y'}$ bounds the probability of $P(\mathbf{X} \mid \mathbf{o}_{\mathbf{Y'}=\mathbf{y'}})$. This criteria is not always applicable for the following two reasons; First there might not be a set Y that satisfies the independence constraint $I(\mathbf{X};\mathbf{O}-\mathbf{Y} \mid \mathbf{Y})$. Second, we limit the size of the set $\mathbf{Y}-\mathbf{Y'}$ whose values are unknown in $\mathbf{Y}$ to 3 (line 7) because we do not want to construct $\mathbf{Y}$ with too many members which would result in a computational burden in checking the constraints.

For situations where there does not exist $\mathbf{Y}$ which satisfies the conditions, we introduce the second criterion shown in Line 10-20 in Algorithm 1. This criterion is heuristic, based on the reasoning that if adding any remaining observation does not individually result in changes in the probabilities, it is likely that adding all the remaining observations would not dramatically change the probabilities. This is a heuristic criteria which does not guarantee the exact solution, but it is reasonable and experimentally shown to be an effective one. The criteria is satisfied if any single additional observation does not change the resulting probabilities more than the bound divided by the number of remaining evidences n. The second contraint is given in Equation 2.

Given bound b,

$$\begin{aligned} &\forall O_n \in \mathbf{O}-\mathbf{Y} \text{ and } o_n \in val(O_n), \\ &|P(\mathbf{X} \mid \mathbf{Y}, o_n) - P(\mathbf{X} \mid \mathbf{Y})| \leq \frac{b}{n} \end{aligned} \tag{2}$$

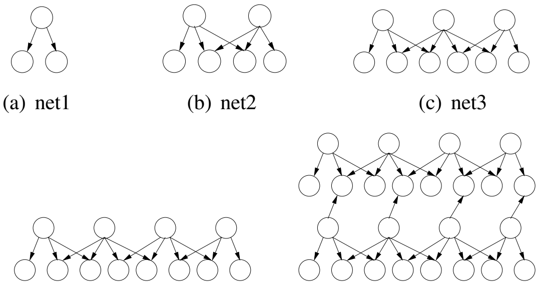

(a) net1 (b) net2 (c) net3 (d) net4 (e) net8

Figure 3. Example 2-level networks with upper reasonings and lower observable nodes

Algorithm 1 Minimum Data Requirements(BN, X, b) where X is nodes of inference in the Bayesian network BN and b is an error bound.

```
1:  To Construct a CPT-tree
2:  $S$ is the set of all observable variables
3:  $EX \longleftarrow \{\{\}\}$ : the set of evidence to examine
4:  $CPT \longleftarrow \phi$ : the set of enough evidences
5:  for $\forall A \in EX$ do
6:    $maxdiff = 0$
7:    $EX = EX - A$
8:    Exhaustively search for set Y of size $\leq 3$, $I(X; S - Y \cup A | Y \cup A)$
9:    if $\max P(X|Y \cup A) - b \leq P(X|A) \leq minP(X|Y \cup A) + b$ then
10:     include A in CPT and continue
11:   else {First constraint not satisfied, check the second constraint}
12:     for each value of evidence $e \in S$ do
13:       Compute $maxdiff = max(|P(X|A,e) - P(X|A)|, maxdiff)$,
14:       if $maxdiff > b/|S - A|$ then
15:         break
16:       end if
17:     end for
18:     if $maxdiff \leq b/|S - A|$ then
19:       include A in CPT and continue
20:     end if
21:   end if
22:   choose e that maximize KL-divergence $KL(P(X|A,e)||P(X|A))$
23:   for each value k of $e$ do
24:     $EX = EX + \{\forall a \in A, e = k\}$
25:   end for
26: end for
27: return CPT
```

5. Experimental Results

5.1. Reasoning Results on Sample Networks

In this section, we experimentally evaluate our algorithm. We construct example networks to measure the performance in terms of amount of observation savings and how close the approximate inference is to the actual probabilities given full observations.

The algorithm was evaluated on two-level networks as shown in Figure 3 with randomly generated parameter values. The upper-level nodes are hidden nodes and the lower-level nodes are observable nodes. Figure 3(e) is a grid network where the set of hidden nodes are placed in a grid, a common structure from the sensor network domain. The nodes in the second row of the grid are dependent on observations of the nodes in the first row. We evaluated the

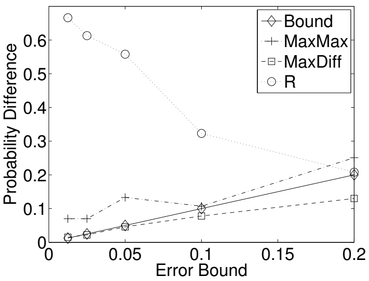


Figure 4. Maximum reasoning errors with error bound change in net4

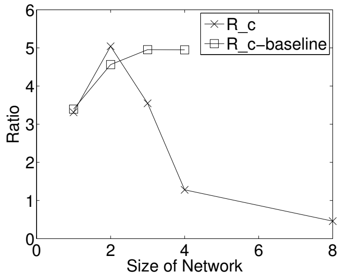


Figure 7. R_c, the ratio of total number of computations required by the algorithm to the number of computations required for original CPT. Baselines are generated by the algorithm adding variables in a random order to the CPT-tree

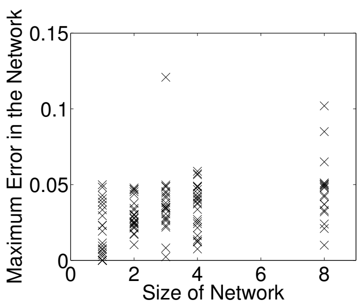


Figure 5. Distribution of maximum errors on 24 networks with different parameter values of the networks

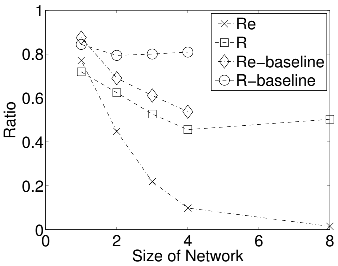
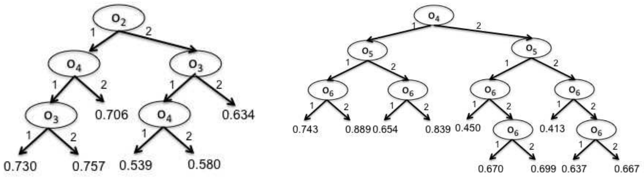


Figure 6. R_e, the ratio of entries in resulting observation sets to the entries in the original CPT and R, the ratio of required evidence to all evidence. Baselines are generated by the algorithm adding variables in a random order to the CPT-tree



(a) Agent 1

(b) Agent 2

Figure 8. the CPT-tree constructed from the network in Figure 3(c)

Table 1. Data for networks in figure 3. Results of 24 runs with 0.05 error bound

size of network	net1	net2	net3	net4	net8
maxdiff	0.012	0.046	0.047	0.059	0.0565
R	0.719	0.624	0.527	0.456	0.503
numcheck	13.250	80.667	227.000	327.417	30029.667
treesize	2.958	9.000	18.500	25.292	1013

algorithm on 24 networks with randomly generated parameters for each structure.

For each instance, we compute 4 measures to verify the performance of the algorithm. R is the ratio of the number of required evidences/observations to the number of total evidence averaged on every observation case. In Table 1, R=0.456 for net4, which indicates that approximately 45% of total number of evidences/observations are required to reach a conclusion within an error bound 0.05. *maxdiff* is a measure of maximum discrepancy between exact probabilities given all observations and approximated probabilities given partial observations on a network. *numcheck* is the number of probability computations during construction of a CPT-tree. *treesize* is the number of leaf nodes in a CPT-tree. As *treesize* gets smaller, we get smaller CPT-trees and less data is required. R increases on net8 because the network is more densely connected than net4 or net3 and it needs more observations from neighboring nodes.

Figure 4 shows value change to error bound on net4 (Figure 3(d)). *maxdiff*, remains smaller than the error bound. *maxmax*, maximum difference observed in our entire experiments for the network also remains close to the given error bound. R decreases as the error bound increases showing that we need less evidence when we can tolerate more error.

The result in Figure 5 shows that the maximum error, the maximum difference between exact and approximate probabilities remains smaller or close to the error bound for most cases.

Figure 6 and 7 show the savings of observations and storage/computation. We compare the result with a baseline algorithm that randomly selects a variable to add to CPT-trees. The baseline algorithm results in a bigger CPT-tree, more observations are needed, and more computations are required to build a CPT-tree. R_e in Figure 6 is the ratio of number of leaf nodes to the number of entries in the full CPTs. The ratio decreases dramatically as the network grows since the number of entries explodes exponentially in relation to the linear increase in the number of observable nodes. The consequence of the ratio decrease is savings both in the number of observations and in the storage necessary to hold the CPTs. In Figure 7, the ratio R_c is the ratio of the number of probability computations required for a CPT-tree to the number of computation for full CPT. The result shows the ratio is significantly reduced, and it goes below 1 for net8 cases, suggesting that the number of computations at least does not exceed the cost of actually computing the full conditional probability table.

5.2. Example of Multi-Agent Communication Policy

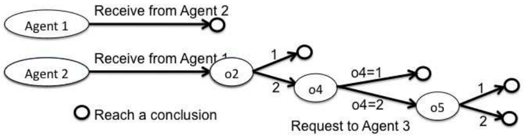


Figure 9. Example communication policy constructed based on the simple scheme of requesting required data using CPT-trees in Figure 8

A multi-agent communication policy for each agent can be generated with local observations and CPT-trees. One simple scheme is to send information necessary to other agents which is included in every path to the leaf nodes in CPT-trees, and in the next step based on acquired information, agents request more information if necessary. This scheme leads to communication savings, because an agent usually needs only a subset of information available in the network. Sample CPT-trees in Figure 8 are constructed from the network of size 3 as in Figure 3(c). The communication policy in Figure 9 is generated from CPT-trees in Figure 8 on above scheme. Consider the following scenario with these agent observations [Agent 1, (o1,1) (o2,2)] [Agent 2, (o3,2), (o4,2)] [Agent3, (o5,2), (o6,2)]. Agent 1 transmits its observations to agent 2, and agent 2 to agent 1 because $o2$ is necessary for reasoning for agent 2 and $o4$ or $o3$ for agent 1. Additionally in this observation scenario, Agent 2 requests information to Agent 3 after receiving the information that $o2 = 2$ from agent 1. The communication ends in 1 and 3 steps for agent 1 and 2 respectively as in Figure 9. Although the communication policy is only a local solution using only locally available information and the CPT-tree, each agent does not have to transfer all information to achieve the required accuracy. Note that agent 1 does not need information from agent 3 in any case. We ran the experiments on communication savings with 100 sample scenarios generated based on the probabilities on the network for 24 networks. We assume each agent sends all local observations at once which is 2 pieces here. On average, each agent in the network require 1.766, 2.320, 6.340 messages for the network of net3, net4, net8 respectively, whereas without using the CPT-tree each agent sends 2,3,7 messages.

6. Conclusion

We have defined Minimum Data Requirement problem which is concerned with minimizing the number of observations needed for approximate inference with a specified error bound in Bayesian networks (BN). We introduce Pseudo-Independence relation to quantitatively describe weak dependencies among variables in a Bayesian network and extend Context-Specific Independence (CSI) relation to Context-Specific Pseudo-Independen ce (CSPI). We provide an algorithm which searches the sets of required observations and probabilities given a BN using Context-Specific Pseudo-Independence as a termination criteria and build a CPT-tree. The limitation of the algorithm is in the assumption that the contribution of variables to inference is additive which is not always true. Although the algorithm does not guarantee the approximated probabilities to be within the error bound, the experimental results are promising showing the accuracy and efficiency of the algorithm. Further, even with a simple multi-agent communication policy which is based on progressively acquiring observations starting with the most informative observations first in the CPT tree, we get between 10 to 20

Acknowledgements

This material is based upon work supported by the National Science Foundation Award No. IIS-0414711. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of National Science Foundation.

References

[1] R. Biswas, S. Thrun, and L. J. Guibas. A probabilistic approach to inference with limited information in sensor networks. *IPSN '04: Proceedings of the Third International symposium on Information Processing in Sensor Networks*, pages 269–276, 2004.

[2] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[3] N. Carver and V. Lesser. Domain monotonicity and the performance of local solutions strategies for CDPS-based distributed sensor interpretation and distributed diagnosis. *Autonomous Agents and Multi-Agent Systems*, 6(1):35–76, 2003.

[4] A. Choi and A. Darwiche. A variational approach for approximating Bayesian networks by edge deletion. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 80–89, 2006.

[5] C. Chow and N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[6] R. Dechter and I. Rish. A scheme for approximating probabilistic inference. *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 132–14, 1997.

[7] N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. *UAI*, pages 252–262, 1996.

[8] A. Krause and C. Guestrin. Optimal nonmyopic value of information in graphical models - efficient algorithms and theoretical limits. *In Proc. of IJCAI*, pages 1339–1345, 2005.

[9] S. Kullback and A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[10] J. D. Park and A. Darwiche. Approximating map using local search. *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 403–410, 2001.

[11] D. Poole and N. L. Zhang. Exploiting contextual independence in probabilistic inference. *J. Artif. Intell. Res. (JAIR)*, 18:263–313, 2003.

[12] J. Shen, V. Lesser, and N. Carver. Minimizing communication cost in a distributed Bayesian network using a decentralized MDP. *Proceedings of Second International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2003)*, pages 678–685, July 2003.

[13] X. Sun, M. J. Druzdzel, and C. Yuan. Dynamic weighting A^* search-based MAP algorithm for Bayesian networks. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 2385–2390, 2007.

[14] S. K. M. Wong and C. J. Butz. Contextual weak independence in Bayesian networks. *UAI*, pages 670–679, 1999.

[15] Y. Xiang. A probabilistic framework for cooperative multiagent distributed interpretation and optimization of communication. *Artificial Intelligence*, 87(1-2):295–342, 1996.