Fast Discovery of Unexpected Patterns in Data, Relative to a Bayesian Network

Szymon Jaroszewicz
Technical University of Szczecin
Department of Computer Science
Zołnierska 49, 71-210 Szczecin, Poland
sj@cs.umb.edu

Tobias Scheffer
Humboldt-Universität zu Berlin
Deptartment of Computer Science
Unter den Linden 6, 10099 Berlin, Germany
scheffer@informatik.hu-berlin.de

ABSTRACT

We consider a model in which background knowledge on a given domain of interest is available in terms of a Bayesian network, in addition to a large database. The mining problem is to discover *unexpected patterns:* our goal is to find the strongest discrepancies between network and database. This problem is intrinsically difficult because it requires inference in a Bayesian network and processing the entire, potentially very large, database. A sampling-based method that we introduce is efficient and yet provably finds the approximately most interesting unexpected patterns. We give a rigorous proof of the method's correctness. Experiments shed light on its efficiency and practicality for large-scale Bayesian networks and databases.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Experimentation, Performance

Keywords

Bayesian Networks, Association Rules, Sampling

1. INTRODUCTION

The general task of knowledge discovery in databases (KDD) is the "automatic extraction of *novel*, useful, and valid knowledge from large sets of data" [5]. However, most data mining methods – such as the Apriori algorithm [1] – are bound to discover *any* knowledge that satisfies the chosen usefulness criterion, including (typically very many) rules that are *already known* to the user. Tuzhilin et al. [20, 15, 16] have studied the problem of finding unexpected rules relative to a set of rules that encode background knowledge.

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

Bayesian networks provide not only a graphical, easily interpretable alternative language for expressing background knowledge, but they also provide an inference mechanism; that is, the probability of arbitrary events can be calculated from the model. Intuitively, given a Bayesian network, the task of mining *interesting unexpected patterns* can be rephrased as discovering itemsets in the data which are much more – or much less – frequent than the background knowledge suggests.

An algorithm which performs this discovery task exactly has been discussed by Jaroszewicz and Simovici [10]. This approach, however, incurs two intrinsic problems. The first problem is the necessary exact inference in the Bayesian network – which is known to be a very hard problem. The second is the necessity to process the entire database in order to assess all patterns and identify the most interesting ones.

Inference can be approximated by sampling from the joint distribution of the network; evaluating the interestingness of patterns can be approximated by evaluating patterns on a random sample instead of the entire database. However, a discovery algorithm which follows this approach cannot come with the *guarantee* of finding the patterns which actually maximize the interestingness criterion. We devise an efficient *sequential sampling algorithm* which approximates the inference in the network and draws a sample from the database in such a way that the resulting interesting attribute sets are, with high probability $1 - \delta$, the most interesting ones up to some small difference in interestingness ε, where ε and δ are user-adjustable parameters.

Bayesian networks are widely used in practice; for instance, Volkswagen uses a Bayesian model in their production planning and scheduling system [13]; it contains 250 nodes with between 2 and 50 possible values. Finding discrepancies between a huge model and the (very large) transaction database is a difficult and relevant problem.

The rest of this paper is organized as follows. We discuss related work in Section 2 and introduce our framework and notation in Section 3. In Section 4, we present the sampling algorithm for finding frequent itemsets that are most interesting relative to background knowledge and our main result on its correctness. We study the behavior of this algorithm empirically in Section 5. Section 6 concludes.

2. RELATED WORK

The Apriori algorithm [1] finds frequent itemsets and, succeedingly, all sufficiently confident rules over these itemsets. Many measures of interestingness of rules and itemsets have

been discussed (*e.g.,* [12, 2, 9, 18]). A fundamental shortcoming of all of these interestingness measures is that neither of them takes into account whether discovered knowledge is entailed by previously available background knowledge.

Some algorithms take background knowledge expressed as rules into account during the discovery process [20, 15, 16]. These methods assess the unexpectedness of a new pattern given the background rules on a purely syntactic basis, without inference. That is, they discover rules that are unexpected given the background rules, rather than rules that are unexpected given *what can be inferred* from the background rules. For instance, if "$A \Rightarrow B$" and "$B \Rightarrow C$" were known, then "$A \Rightarrow C$" would still be considered an unexpected pattern. A more detailed discussion of data mining with background knowledge can be found in [10].

Bayesian networks are a powerful representational scheme for background knowledge; they are graphical models, easy to understand and modify. Bayesian networks encode the joint distribution over all attributes. Inference mechanisms are well understood; general inference in Bayesian networks is a hard problem that can be approximated by sampling (*e.g.,* [11]). The inference problem given values of some random variables is approximated by MCMC methods [6]. Jaroszewicz and Simovici [10] define interestingness of a pattern as difference between observed frequency and inferred probability. Their algorithm finds all itemsets whose interestingness is above ε but, unfortunately, relies on exact inference – which is tractable for small networks.

It is straightforward to replace the exact inference by sampling-based approximate inference, but we would like to do this in such a way that the algorithm's output maintains a well-defined optimality property. Sampling algorithms [21] estimate the interestingness of patterns by drawing examples from the database. The most elementary sampling schemes [21] calculate worst-case sample bounds, often based on Hoeffding's inequality. The sample size necessary for a desired ε/δ level of optimality can be reduced substantially by employing *data-dependent*, *sequential* sampling [3, 14]. Here, the data are processed incrementally; the necessary sample size is determined online, dependent on characteristics of the data that have already been processed.

Practical sequential sampling algorithms have been studied for interestingness functions which are averages over the data (such as accuracy) [7, 4, 8] as well as for more general interestingness functions [19]. These algorithms minimize the number of database accesses needed to find, with high probability, all approximately sufficiently interesting, or the n most interesting patterns. Unfortunately, all of these methods assume that the bottleneck in the assessment of candidate patterns lies only in the database access. By contrast, in our problem setting we need to manage uncertainty originating from limited database access as well as uncertainty originating from approximate inference in the Bayesian network.

3. PROBLEM SETTING

In this section, we introduce the necessary notation and define the problem that we will solve in the following.

We are given a database D with attributes $Z = \{A_1, \ldots, A_m, \}$; attributes are categorical with finite domains $\mathrm{Dom}(A)$. $P_I^D(\mathbf{i})$ denotes the probability that an attribute set $I \subseteq Z$ assumes a vector of values $\mathbf{i}$ in the database D.

A *Bayesian network* BN is a set of random variables (corresponding to our attributes) $Z = \{A_1, \ldots, A_m\}$ which constitute the vertices of a directed, acyclic graph, a set of edges $E \subseteq Z \times Z$, and, for each vertex A_i with direct ancestors $par(A_i)$, a conditional distribution $P_{A_i|par(A_i)}$. A Bayesian network defines a joint distribution $P_Z^{BN} = \prod_{i=1}^{m} P_{A_i|par(A_i)}$.

Given an attribute set I and values $\mathbf{i}$, we write $P_I^{BN}(\mathbf{i})$ to denote the probability of itemset $I = \mathbf{i}$ as determined by a Bayesian network BN. According to [10], we define the unexpectedness, or interestingness, of an event $I = \mathbf{i}$ as $\mathcal{I}(I, \mathbf{i}) = |P_I^D(\mathbf{i}) - P_I^{BN}(\mathbf{i})|$ — the absolute difference between an event's probability inferred from the network, and observed in the database. We will now leverage the definition of unexpectedness of events to *interestingness of attribute sets*: an attribute set I is interesting, if there is an event $I = \mathbf{i}$ for which inferred and observed probability diverge.

Definition 1. *Given a Bayesian network BN and data D, the interestingness of attribute set I is defined in Equation 1.*

$$\mathcal{I}(I) = \max_{\mathbf{i} \in \mathrm{Dom}(I)} \mathcal{I}(I, \mathbf{i}) \tag{1}$$

$$= \max_{\mathbf{i} \in \mathrm{Dom}(I)} |P_I^D(\mathbf{i}) - P_I^{BN}(\mathbf{i})| \tag{2}$$

The definition of interestingness refers to $P_I^{BN}(\mathbf{i})$, the exact probability of $I = \mathbf{i}$ inferred from the network, and $P_I^D(\mathbf{i})$, the probability of $I = \mathbf{i}$ in the (potentially very large) database. $P_I^{BN}(\mathbf{i})$ can be estimated by sampling from the network; $P_I^D(\mathbf{i})$ by sampling from the database. Since the network has no cycles we can always draw from the conditional distributions $P(A_i|par(A_i))$ of each vertex A_i after the values of all parents have been drawn. Thus we obtain a sample S^{BN} of independent assignments of values to the attributes according to P^{BN}. After additionally drawing records S^D from D independently under uniform distribution, we obtain an estimate $\hat{\mathcal{I}}(I, \mathbf{i})$ as in Equation 3, and of $\hat{\mathcal{I}}(I)$ in Equation 4. $\hat{P}_I^D(\mathbf{i})$ is the relative frequency of $I = \mathbf{i}$ in sample S^D, and $\hat{P}_I^{BN}(\mathbf{i})$ the relative frequency of $I = \mathbf{i}$ in S^{BN}.

$$\hat{\mathcal{I}}(I, \mathbf{i}) = \left| \hat{P}_I^D(\mathbf{i}) - \hat{P}_I^{BN}(\mathbf{i}) \right| \tag{3}$$

$$\hat{\mathcal{I}}(I) = \max_{\mathbf{i} \in \mathrm{Dom}(I)} \hat{\mathcal{I}}(I, \mathbf{i}) \tag{4}$$

A special case occurs when the Bayesian network is too large for exact inference but the database is compact and P_I^D can be determined exactly. In this case, only P_I^{BN} has to be approximated by $\hat{P}_I^{BN}$, but S^D can be the entire database D and therefore $\hat{P}_I^D = P_I^D$.

A possible problem setting would be to find all attribute sets whose interestingness exceeds some ε. However, from the user's point of view it is often more natural to constrain the number of returned patterns, rather than the somewhat less intuitive interestingness value. We therefore define the n most interesting attribute sets problem as follows.

Definition 2. *Let D be a database over attributes Z and BN a Bayesian network. The n most interesting attribute sets problem is to find n attribute sets $H = \{I_1, \ldots, I_n\}$; $I_j \subseteq Z$, such that there is no other attribute set I' which is more interesting than any of H (Equation 5).*

$$\text{there is no } I' \subseteq Z : I' \notin H,\ \mathcal{I}(I') > \min_{I \in H} \mathcal{I}(I) \tag{5}$$

Any solution to the n most interesting attribute sets problem has to calculate the $\mathcal{I}(I)$ which requires exact inference in the Bayesian network and at least one pass over the entire database. We would like to find an alternative optimality property that can be guaranteed by an efficient algorithm. We therefore define the n approximately most interesting attribute sets problem as follows.

DEFINITION 3. *Let D be a database over itemsets Z and BN a Bayesian network. The n approximately most interesting attribute sets problem is to find n attribute sets $H = \{I_1, \ldots, I_n\}$; $I_j \subseteq Z$, such that, with high probability $1-\delta$, there is no other attribute set I' which is ε more interesting than any of H (Equation 6).*

$$\text{with confidence } 1-\delta, \text{ there is no } I' \subseteq Z : \\ I' \notin H \text{ and } \mathcal{I}(I') > \min_{I \in H} \mathcal{I}(I) + \varepsilon \tag{6}$$

4. FAST DISCOVERY OF INTERESTING ATTRIBUTE SETS

We are now ready to present our solution to the *n approximately most interesting attribute sets problem*. The AprioriBNS algorithm is presented in Table 1; it refers to confidence bounds provided in Table 2. We will now briefly sketch the algorithm, then state our main Theorem, and finally discuss some additional details and design choices.

AprioriBNS generates candidate attribute sets like the Apriori algorithm does: starting from all one-element sets in step 1, candidates with $i+1$ attributes are generated in step 2g by merging all sets which differ in only the last element, and pruning those with infrequent subsets.

In each iteration of the main loop, we draw a batch of database records and observations from the Bayesian network. Only one such batch is stored at a time and the sample size and frequency counts of all patterns under considerations are updated; the batch is deleted after an iteration of the loop and a new batch is drawn. The interestingness of each attribute set I is estimated based on $N^{BN}(I)$ observations from the network and $N^D(I)$ database records.

There are two mechanisms for eliminating patterns which are not among the n best ones. These rejection mechanisms are *data dependent*: if some attribute sets are very uninteresting, only few observations are needed to eliminate them from the search space and the algorithm requires few sampling operations. Step 2c is analogous to the pruning of low support itemsets in Apriori. $P_I^D(\mathbf{i})$ can be interpreted as the support of the itemset $I = \mathbf{i}$ with respect to the database, and $P_I^{BN}(\mathbf{i})$ as the support of $I = \mathbf{i}$ with respect to the network. The interestingness – which is the absolute difference – can be bounded from above by the maximum of these. No superset of I can be more frequent than I and therefore all supersets can be removed from the search space, if this upper bound is below the currently found n-th most interesting attribute set. Since only estimates $\hat{P}_I^{BN}(\mathbf{i})$ and $\hat{P}_I^D(\mathbf{i})$ are known, we add a confidence bounds $E_{\mathcal{I}}$ and E_s to account for possible misestimation.

The pruning step is powerful because it removes an entire branch, but it can only be executed when an attribute set is very infrequent. Therefore, in step 2d, we delete an attribute set I' if its interestingness (plus confidence bound) is below that of the currently n-th most interesting pattern (minus confidence bound). We can then delete I' but since interest-

Table 1: AprioriBNS: Fast Discovery of the Approximately Most Interesting Attribute Sets
Input: Bayesian network BN, database D over attributes Z, approximation and confidence parameters ε and δ, the desired number of interesting itemsets n.

1. **Let** $i \leftarrow 1$ (iteration); generate initial candidates $C_1 = \{\{A_i\} : A_i \in Z\}$; **let** $H_1 \leftarrow C_1$ (itemsets under consideration); **for all** $I \in H_1$, **initialize** $N^{BN}(I) = 0$ and $N^D(I) = 0$ (Bayesian network and database sample size for itemset I).
2. **Repeat** until break:
   (a) **Draw** batch of observations S_i^{BN} according to P^{BN} and a batch of database records S_i^D at random from D.
   (b) **For all** $I \in H_i$, **increment** $N^D(I)$ **by** $|S_i^D|$; **increment** $N^{BN}(I)$ **by** $|S_i^{BN}|$; update frequency counts $\hat{P}_I^D$, $\hat{P}_I^{BN}$, and consequently, $\mathcal{I}(I)$ (Equation 4). **Let** H_i^* be the n best itemsets in H_i, according to the current $\hat{\mathcal{I}}$.
   (c) **For all** $I' \in H_i \setminus H_i^*$: **if**
   $$\begin{aligned}&\max\{\max_{\mathbf{i}\in \mathrm{Dom}(I')} \hat{P}_{I'}^D(\mathbf{i}), \max_{\mathbf{i}\in \mathrm{Dom}(I')} \hat{P}_{I'}^{BN}(\mathbf{i})\} \\ &\quad + E_s\left(I', \tfrac{\delta}{3|H_i|i(i+1)}\right) < \\ &\min_{I\in H_i^*}\left\{\hat{\mathcal{I}}(I) - E_{\mathcal{I}}\left(I, \tfrac{\delta}{3|H_i|i(i+1)}\right)\right\}\end{aligned}$$
   then remove I' *and all its supersets* from H_i and C_i. (For $E_{\mathcal{I}}$ and E_s, refer to Table 2; neither I' nor any superset will ever become a champion.)
   (d) **For all** $I' \in H_i \setminus H_i^*$: **if**
   $$\begin{aligned}&\hat{\mathcal{I}}(I') + E_{\mathcal{I}}\left(I', \tfrac{\delta}{3|H_i|i(i+1)}\right) < \\ &\quad \min_{I\in H_i^*}\left\{\hat{\mathcal{I}}(I) - E_{\mathcal{I}}\left(I, \tfrac{\delta}{3|H_i|i(i+1)}\right)\right\}\end{aligned}$$
   then remove I' from H_i. (I' is most likely not a champion but its supersets might still.)
   (e) **If** $C_i = \emptyset$ and **for all** $I \in H_i^*$, $I' \in (H_i \setminus H_i^*)$:
   $$\begin{aligned}&\hat{\mathcal{I}}(I) - E_{\mathcal{I}}\left(I, \tfrac{\delta}{3|H_i|i(i+1)}\right) > \\ &\quad \hat{\mathcal{I}}(I') + E_{\mathcal{I}}\left(I', \tfrac{\delta}{3|H_i|i(i+1)}\right) - \varepsilon\end{aligned}$$
   then break.
   (f) **Let** $n^{BN} = \min_{I\in H_i} N^{BN}(I)$, $n^D = \min_{I\in H_i} N^D(I)$; **if** $C_i = \emptyset$ and
   $$E_d\left(n^{BN}, n^D, \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i} \frac{1}{j(j+1)}\right)}{\sum_{I\in H_i} |\,\mathrm{Dom}(I)|}\right) \leq \frac{\varepsilon}{2}$$
   then break.
   (g) $C_{i+1} \leftarrow$ generate new candidates from C_i.
   (h) **Let** $H_{i+1} \leftarrow H_i \cup C_{i+1}$; **let** $i \leftarrow i+1$.
3. **Return** the n best itemsets (according to $\hat{\mathcal{I}}$) from H_i.

Table 2: Confidence bounds used by AprioriBNS

Based on Hoeffding inequality, sampling from Bayesian network and data.
$E_{\mathcal{I}}(I,\delta) = \sqrt{\frac{1}{2}\frac{N^{BN}(I)+N^{D}(I)}{N^{BN}(I)N^{D}(I)}\log\frac{2\lvert\operatorname{Dom}(I)\rvert}{\delta}}, \quad E_s(I,\delta) = \sqrt{\log\frac{4\lvert\operatorname{Dom}(I)\rvert}{\delta}}\cdot\max\left\{\frac{1}{\sqrt{2N^{BN}(I)}},\frac{1}{\sqrt{2N^{D}(I)}}\right\}$ $E_d(n^{BN},n^{D},\delta) = \sqrt{\frac{1}{2}\frac{n^{BN}+n^{D}}{n^{BN}n^{D}}\log\frac{2}{\delta}}$
Based on Hoeffding inequality, all data used, sampling from Bayesian network only.
$E_s(I,\delta) = E_{\mathcal{I}}(I,\delta) = \sqrt{\frac{1}{2N^{BN}(I)}\log\frac{2\lvert\operatorname{Dom}(I)\rvert}{\delta}}, \quad E_d(n^{BN},\delta) = \sqrt{\frac{1}{2n^{BN}}\log\frac{2}{\delta}}$
Based on normal approximation, sampling from Bayesian network and data.
$E_{\mathcal{I}}(I,\delta) = z_{1-\frac{\delta}{2\lvert\operatorname{Dom}(I)\rvert}}\max_{\mathbf{i}\in\operatorname{Dom}(I)}\sqrt{\frac{\hat{P}_I^{BN}(\mathbf{i})\cdot(1-\hat{P}_I^{BN}(\mathbf{i}))}{N^{BN}(I)}+\frac{\hat{P}_I^{D}(\mathbf{i})\cdot(1-\hat{P}_I^{D}(\mathbf{i}))}{N^{D}(I)}\cdot\frac{\lvert D\rvert-N^{D}(I)}{\lvert D\rvert-1}}$ $E_s(I,\delta) = z_{1-\frac{\delta}{4\lvert\operatorname{Dom}(I)\rvert}}\max_{\mathbf{i}\in\operatorname{Dom}(I)}\max\left\{\sqrt{\frac{\hat{P}_I^{BN}(\mathbf{i})\cdot(1-\hat{P}_I^{BN}(\mathbf{i}))}{N^{BN}(I)}},\sqrt{\frac{\hat{P}_I^{D}(\mathbf{i})(1-\hat{P}_I^{D}(\mathbf{i}))}{N^{D}(I)}\cdot\frac{\lvert D\rvert-N^{D}(I)}{\lvert D\rvert-1}}\right\}$ $E_d(n^{BN},n^{D},\delta) = \frac{1}{2}z_{1-\frac{\delta}{2}}\sqrt{\frac{1}{n^{BN}}+\frac{1}{\lvert D\rvert}\cdot\frac{\lvert D\rvert-n^{D}}{N^{D}-1}}$
Based on normal approximation, all data used, sampling from Bayesian network only.
$E_s(I,\delta) = E_{\mathcal{I}}(I,\delta) = z_{1-\frac{\delta}{2\lvert\operatorname{Dom}(I)\rvert}}\max_{\mathbf{i}\in\operatorname{Dom}(I)}\sqrt{\frac{\hat{P}_I^{BN}(\mathbf{i})\cdot(1-\hat{P}_I^{BN}(\mathbf{i}))}{N^{BN}(I)}}, \quad E_d(n^{BN},n^{D},\delta) = \frac{1}{2}z_{1-\frac{\delta}{2}}\max_{\mathbf{i}\in\operatorname{Dom}(I)}\frac{1}{\sqrt{n^{BN}}}$

ingness does not decrease monotonically with the number of attributes, we cannot prune the entire branch.

There are two alternative stopping criteria. If every attribute set in the current set of "champions" H_i^* (minus an appropriate confidence bound) outperforms every attribute outside (plus confidence bound), then the current estimates are sufficiently accurate to end the search (step 2e). This stopping criterion is *data dependent*: If there are hypotheses which clearly set themselves apart from the rest of the hypothesis space, then the algorithm terminates early.

In addition, the algorithm may terminate if *all* estimates are tight up to $\frac{\varepsilon}{2}$. This worst-case criterion uses bounds which are independent of specific hypotheses (*data independent*) and a fixed amount of allowable error is set aside for it. Its purpose is to guarantee that the algorithm will always terminate. As long as the candidate set C_i is not empty, there are still hypotheses which have not yet been assessed at all. In this case, the search cannot yet terminate. After exiting the main loop, the n apparently most interesting attribute sets are returned.

AprioriBNS refers to error bounds which are detailed in Table 2. We provide both, exact but loose confidence bounds based on Hoeffding's inequality, and their practically more relevant normal approximation. Statistical folklore says normal approximations can be used for sample sizes from 30 onwards; in our experiments, we encounter sample sizes of 1000 or more. z denotes the inverse standard normal cumulative distribution function and n^{BN}, n^{D} the minimum sample size (from Bayesian network and database, respectively) for any $I \in H$. We furthermore distinguish the general case in which samples are drawn from both, the Bayesian network and database, from the special case in which the database is feasibly small and therefore $\hat{P}_I^D = P_I^D$, samples are drawn only from the network. We are now ready to state our main result on the optimality of the result returned by our discovery algorithm.

THEOREM 1. *Given a database D, a Bayesian network BN over nodes Z, and parameters n, ε, and δ, the AprioriBNS algorithm will output a set of the n* approximately most interesting attribute sets H^*. *That is, with probability $1-\delta$, there is no $I' \subseteq Z$ with $I \notin H^*$ and $\mathcal{I}(I') > \min_{I\in H^*}\mathcal{I}(I)+\varepsilon$. Furthermore, the algorithm will always terminate (even if the database is an infinite stream); the number of sampling operations from the database and from the Bayesian network is bounded by $O(|Z|\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$.*

The proof of Theorem 1 is given in the Appendix. We will conclude this section by providing additional design decisions of the algorithm. A copy of the source code is available from the authors for research purposes.

In step 2a, we are free to choose any size of the batch to draw from the network and database. As long as $C_i \neq \emptyset$, the greatest benefit is obtained by pruning attribute sets in step 2c (all supersets are removed from the search space). When $C_i = \emptyset$, then terminating early in step 2e becomes possible, and rejecting attribute sets in step 2d is as beneficial as pruning in step 2c, but easier to achieve. We select the batch size such that we can expect to be able to prune a substantial part of the search space ($C_i \neq \emptyset$), terminate early, or reject substantially many hypotheses ($C_i = \emptyset$).

We estimate the batch size required to prune 25% of the hypotheses by comparing the least interesting hypothesis in H_i^* to a hypothesis at the 75-th percentile of interestingness. We find the sample size that satisfies the precondition of step 2c for these two hypotheses (this is achieved easily by inverting $E_{\mathcal{I}}$ and E_s). If $C_i = \emptyset$, then we analogously find the batch size that would allow us to terminate early in step 2e and the batch size that would allow to reject 25% of the hypotheses in step 2d and take the minimum. In order to efficiently update the interestingness of many attribute sets simultaneously, we use a marginalization algorithm similar to the one described in [10].

5. EXPERIMENTS

Theorem 1 already guarantees that the attribute sets returned by the algorithm are, with high probability, nearly optimal with respect to the interestingness measure. But we still have to study the *practical usefulness* of the method for large-scale problems. In our experiments, we will first focus on problems that can be solved with the exact discovery method AprioriBN [10] and investigate whether the sampling approach speeds up the discovery process (while [10] call only the core part of their algorithm AprioriBN, we

use this term to refer to the entire exact discovery method). More importantly, we will then turn towards discovery problems with *large-scale* Bayesian networks that *cannot be handled by known exact methods*. We will investigate whether any of these problems can be solved using our sampling-based discovery method.

In order to study the performance of AprioriBN and AprioriBNS over a range of network sizes, we need a controlled environment with Bayesian networks of various sizes and corresponding datasets. We have to be able to control the divergence of background knowledge and data, and, in order to assure that our experiments are reproducible, we would like to restrict our experiments to publicly available data. We create an experimental setting which satisfies these requirements. For the first set of experiments, we use data sets from the UCI repository and learn networks from the data using the B-Course [17] website. These generated networks play the role of expert knowledge in our experimentation. In order to conduct experiments on a larger scale, we start from large Bayesian networks, generate databases by drawing from the network, and then learn a slightly distorted network from the data which again serves as expert knowledge (see below for a detailed description). For the small UCI datasets, the algorithm processes the entire database whereas, for the large-scale problems, AprioriBNS samples from both, the database and the network.

We first compare the performance of AprioriBN and AprioriBNS using the UCI data sets. For all experiments, we use $\varepsilon = 0.01$, $\delta = 0.05$, and $n = 5$. We constrain the cardinality of the attribute sets to $\max_k$. Here, the databases are small and therefore only the network is sampled and $\hat{P}_I^D = P_I^D$ for all I. Table 3 shows the performance results. The $|Z|$ column contains numbers of attributes in each dataset, $t[s]$ computation time, N^{BN} the number of samples drawn from the Bayesian network, $\max\hat{\mathcal{I}}$ and $\max\mathcal{I}$ are the estimated and actual interestingness of the most interesting attribute set found by AprioriBNS and AprioriBN, respectively.

We refrain from drawing conclusions on the absolute running time of the algorithms because of a slight difference in the problems that AprioriBN and AprioriBNS solve (finding *all sufficiently* versus finding *the most* interesting rules). We do, however, conclude from Table 3 that the relative benefit of AprioriBNS over AprioriBN increases with growing network size. For 61 nodes, AprioriBNS is *many times* faster than AprioriBN. More importantly, AprioriBNS finds a solution for the `audiology` problem; AprioriBN exceeds time and memory resources for this problem.

The most interesting attribute set has always been picked correctly by the sampling algorithm and its estimated interestingness is close to the exact value. The remaining 4 most interesting sets were not always picked correctly, but remained within the bounds guaranteed by the algorithm.

We will now study how the execution time of AprioriBNS depends on the maximum attribute set size $\max_k$. Figure 1 shows the computation time for various values of $\max_k$ for the `lymphography` data set. Note that the search space size grows exponentially in $\max_k$ and this growth would be maximal for $\max_k = 10$ if no pruning was performed. By contrast, the runtime levels off after $\max_k = 7$, indicating that the pruning rule (step 2c of AprioriBNS) is effective and reduces the computation time substantially.

Let us now investigate whether AprioriBNS can solve discovery problems that involve *much larger* networks than

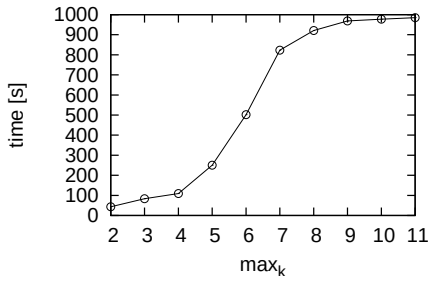

Figure 1: Computation time versus maximum attribute set size $\max_k$ for `lymphography` data.

AprioriBN can handle. We draw 1 million observations governed by the `Munin1` network from the Bayesian Network Repository. We then use a small part of the resulting dataset to learn a Bayesian network. Thus, the original network plays the role of a real world system (from which the dataset is obtained) and the network learned from a subset of the data plays the role of our imperfect knowledge about the system. By varying the sample size M used to build the network we can affect the quality of our 'background' knowledge. The `Munin1` network has 189 attributes. Exact inference from networks of this size is very hard in practice.

Table 4 shows the results for various values of M and $\max_k = 2 \ldots 3$. We sample at equal rates from the Bayesian network and from data; both numbers of examples are therefore equal and denoted by N in the table. We use the same setting for the next experiment with the `Munin2` network containing 1003 attributes. The problem is huge both in terms of the size of Bayesian network and the size of data: The file containing 1 million rows sampled from the original network is over 4GB large, and 239227 rows sampled by the algorithm amount to almost 1GB. The experiment took 4 hours and 50 minutes for $\max_k = 2$.

Figure 2 summarizes Tables 3 and 4, it details the relationship between the number of nodes in the network and the computation time of AprioriBN and AprioriBNS. We observe a roughly linear relationship between logarithmic network size and the logarithmic execution time, Figure 2 shows a model fitted to the data. From these experiments, we conclude that the AprioriBNS algorithm scales to very large Bayesian networks and databases, yet it is guaranteed to find a near-optimal solution to the most interesting attribute set problem with high confidence. We can apply the exact AprioriBN algorithm to networks of up to about 60 nodes. Using the same computer hardware, we can solve discovery problems over networks of more than 1000 nodes using the sampling-based AprioriBNS method.

6. CONCLUSION

We studied the problem of discovering unexpected patterns in a database. We formulated the approximately most interesting attribute sets problem and developed an algorithm which solves this problem. AprioriBNS uses sampling-based approximate inference in the Bayesian network and, when the database is large, also samples the data.

We proved that AprioriBNS always finds, with high confidence, the approximately most interesting attribute sets.

Table 3: Evaluation on networks learned from UCI datasets.

dataset	$\|Z\|$	$\max_k$	N^{BN}	AprioriBNS: $\max \hat{\mathcal{I}}$	AprioriBN: $\max \mathcal{I}$	AprioriBNS: $t[s]$	AprioriBN: $t[s]$
KSL	9	5	205582	0.03229	0.03201	55	**1**
lymphography	19	3	88333	0.09943	0.12308	43	**29**
lymphography	19	4	159524	0.12343	0.12631	**83**	106
soybean	36	3	282721	0.06388	0.06440	**409**	1292
soybean	36	4	292746	0.07185	0.07196	**1748**	7779
annealing	40	3	273948	0.04985	0.04892	**407**	1006
annealing	40	4	288331	0.06159	0.06118	**2246**	6762
splice	61	3	190164	0.03652	0.03643	**1795**	8456
audiology	70	3	211712	0.09723	–	**727**	–
audiology	70	4	228857	0.10478	–	**9727**	–

Table 4: Results for the `Munin` networks.

dataset	$\|Z\|$	M	$\max_k$	$t[s]$	N	$\max \hat{\mathcal{I}}$
Munin1	189	100	2	**874**	136972	0.4138
Munin1	189	150	2	**1754**	312139	0.2882
Munin1	189	200	2	**1004**	139500	0.2345
Munin1	189	250	2	**2292**	373191	0.1819
Munin1	189	500	2	**2769**	431269	0.1174
Munin1	189	1000	2	**3502**	480432	0.0674
Munin1	189	100	3	**14375**	375249	0.4603
Munin1	189	150	3	**16989**	450820	0.3272
Munin2	1003	100	2	**17424**	239227	0.3438

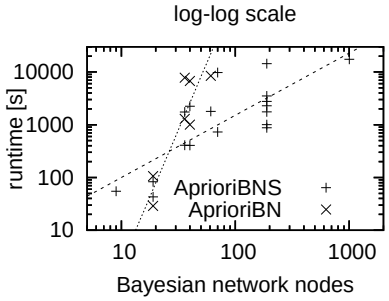

Figure 2: Network size and computation time.

We studied AprioriBNS empirically using moderately sized as well as large-scale Bayesian networks and databases. From our experiments, we can draw the following main conclusions. (1) The relative performance benefit of AprioriBNS over the corresponding exact method AprioriBN increases with the network size. For moderately sized networks, AprioriBNS can be many times faster than AprioriBN. (2) More importantly, while AprioriBN scales to networks with about 60 nodes, we can apply AprioriBNS to Bayesian networks of 1000 nodes and databases of several gigabytes using the same hardware.

Acknowledgments

T.S. is supported by the German Science Foundation under Grant SCHE 540/10-1.

7. REFERENCES

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, 1996.

[2] R. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining*, 1999.

[3] H. Dodge and H. Romig. A method of sampling inspection. *The Bell System Technical Journal*, 8:613–631, 1929.

[4] C. Domingo, R. Gavalda, and O. Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. *Data Mining and Knowledge Discovery*, 6(2):131–152, 2002.

[5] U. Fayyad, G. Piatetski-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *KDD-96*, 1996.

[6] W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1995.

[7] R. Greiner. PALO: A probabilistic hill-climbing algorithm. *Artificial Intelligence*, 83(1–2), July 1996.

[8] G. Hulten and P. Domingos. Mining complex models from aribtrarily large datasets in constant time. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining*, 2002.

[9] S. Jaroszewicz and D. Simovici. A general measure of rule interestingness. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery and Data Mining*, 2001.

[10] S. Jaroszewicz and D. Simovici. Interestingness of frequent itemsets using Bayesian networks as background knowledge. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[11] F. Jensen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2001.

[12] W. Klösgen. Assistant for knowledge discovery in data. In P. Hoschka, editor, *Assisting Computer: A New Generation of Support Systems*, 1995.

[13] R. Kruse. Knowledge-based operations on graphical models. In *Proceedings of the Dagstuhl Seminar on Probabilistic, Logical, and Relational Learning*, 2005. In print.

[14] O. Maron and A. Moore. Hoeffding races: Accelerating model selection search for classification and function approximating. In *Advances in Neural Information Processing Systems*, pages 59–66, 1994.

[15] B. Padmanabhan and A. Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999.

[16] B. Padmanabhan and A. Tuzhilin. Small is beautiful: discovering the minimal set of unexpected patterns. In *Proceedings of the Sixth SIGKDD Conference on Knowledge Discovery and Data Mining*, 2000.

[17] P.Myllymäki, T.Silander, H.Tirri, and P.Uronen. B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*, 11(3):369–387, 2002.

[18] T. Scheffer. Finding association rules that trade support optimally against confidence. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2001.

[19] T. Scheffer and S. Wrobel. Finding the most interesting patterns in a database quickly by using sequential

sampling. *Journal of Machine Learning Research*, 3:833–862, 2002.

[20] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the SIGKDD Conference on Knowledge Discovery and Data Mining*, 1995.

[21] H. Toivonen. Sampling large databases for association rules. In *Proceedings of the International Conference on Very Large Databases*, 1996.

APPENDIX

A. PROOF OF THEOREM 1

The proof of Theorem 1 has two parts: we will first prove the guaranteed sample bound of $O(|Z|\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$. We will then show that AprioriBNS in fact solves the approximately most interesting attribute sets problem.

A.1 AprioriBNS Samples Only Polynomially Many Observations

THEOREM 2. *The number of sampling operations of AprioriBNS from the database and from the Bayesian network is bounded by* $O(|Z|\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$.

PROOF. We can disregard the possibility of early stopping and show that the stopping criterion in step 2f applies after polynomially many sampling operations.

Let $r = \max_{A\in Z}|\mathrm{Dom}(A)|$. First note that $\sum_{I\in H_{i_{max}}}|\mathrm{Dom}(I)| \le (r+1)^{|I|}$. For clarity of the presentation, let $n^{BN} = n^D = N$. The stopping condition becomes Equation 7.

$$\sqrt{\frac{1}{N}\log\frac{2\sum_{I\in H_{i_{max}}}|\mathrm{Dom}(I)|}{\delta\left(1-\frac{2}{3}\sum_{j=1}^{i_{max}}\frac{1}{j(j+1)}\right)}} \le \frac{\varepsilon}{2} \tag{7}$$

From the Hoeffding bound, it follows that Equation 7 is satisfied for N given in Equation 8.

$$N \ge \frac{4}{\varepsilon^2}\log\frac{2\sum_{I\in H_{i_{max}}}|\mathrm{Dom}(I)|}{\delta\left(1-\frac{2}{3}\sum_{j=1}^{i_{max}}\frac{1}{j(j+1)}\right)} \tag{8}$$

$$\le \frac{4}{\varepsilon^2}\log\frac{2(r+1)^{|Z|}}{\frac{1}{3}\delta} = \frac{4}{\varepsilon^2}|Z|\log\frac{6(r+1)}{\delta} \tag{9}$$

This proves Theorem 2 ∎

A.2 AprioriBNS Solves Approximately Most Interesting Attribute Sets Problem

Throughout the proof, $\sum_{\mathbf{i}}$ and $\max_{\mathbf{i}}$ are abbreviations for $\sum_{\mathbf{i}\in\mathrm{Dom}(I)}$ and $\max_{\mathbf{i}\in\mathrm{Dom}(I)}$ respectively. We will first define a helpful concept which we call the *support* of an attribute set. The support of an attribute set I is the maximum support of any itemset $I=\mathbf{i}$ with respect to the Bayesian network or the database, whichever is greater. Using this definition, it is easy to see that support upper-bounds interestingness which is helpful to understand the pruning mechanism of step 2c.

DEFINITION 4. *The* support *of an attribute set I is defined in Equation 10.*

$$\mathrm{supp}(I) = \max\left\{\max_{\mathbf{i}} P_I^{BN}(\mathbf{i}), \max_{\mathbf{i}} P_I^{D}(\mathbf{i})\right\} \tag{10}$$

We write the estimated support as $\widehat{\mathrm{supp}}(I) = \max\left\{\max_{\mathbf{i}}\hat{P}_I^{BN}(\mathbf{i}), \max_{\mathbf{i}}\hat{P}_I^{D}(\mathbf{i})\right\}$.

Table 5: Notation used in the proof.

G	"Good" hypotheses output by AprioriBNS in step 3.
R_i	Attribute sets rejected before iteration i. Note that if an attribute set is pruned (step 2c) then R will contain that set and all of its supersets.
H_i	Collection of attribute sets still under consideration in iteration i.
H_i^*	n most interesting attribute sets in H_i.
C_i	Collection of candidate attribute sets in iteration i.
U_i	Collection of unseen attribute sets: $U_i = 2^Z \setminus (\{\emptyset\}\cup H_i \cup R_i)$.
i_{max}	Value of i after the main loop terminates.
$\hat{\mathcal{I}}(I)$	Estimate of interestingness of attribute set I during i-th iteration.
$\widehat{\mathrm{supp}}(I)$	Estimate of support of attribute set I during current iteration.
n^{BN}, n^D	Minimum sample size (from Bayesian network and database, respectively) for any $I\in H_i$.

LEMMA 1. *The support of an attribute set I upper-bounds its interestingness:* $\mathrm{supp}(I) \ge \mathcal{I}(I)$.

PROOF. The proof of Lemma 1 follows directly from Definition 1: the absolute difference $\max_{\mathbf{i}}|P_I^{BN}(\mathbf{i}) - P_I^D(\mathbf{i})|$ is greatest if either $P_I^{BN}(\mathbf{i})$ or $P_I^D(\mathbf{i})$ is zero. ∎

Table 5 defines additional notation that we use during the proof. U_i is the set of unseen attribute sets in iteration i. It is important to note that no hypotheses remain unseen when the candidate set C_i is empty.

LEMMA 2. *$C_i = \emptyset$ implies $U_i = \emptyset$ for all $1 \le i \le i_{max}$.*

PROOF. Lemma 2 follows primarily from the completeness of Apriori's candidate generation procedure invoked in step 2g: if no attribute set was ever pruned, then $\bigcup_i C_i = 2^Z \setminus \{\emptyset\}$. In step 2h, the candidates C_i are accumulated in H_{i+1}. In step 2d, one or more hypotheses I' can be removed from H_i. By the definition of R_i, each removed I' is then an element of R_i. In step 2c, hypotheses I' and all their supersets are removed from C_i and H_i. In this case, supersets of C_i will not be generated in step 2g but, by the definition of R_i all of them become members of R_i. This implies that $U_i = 2^Z \setminus \{\emptyset\} \setminus H_i \setminus R_i = \emptyset$. ∎

The proof heavily relies on confidence intervals for estimates of the interestingness, support, and the difference of interestingness values. We have to show that the confidence bounds given in Table 2 are in fact valid.

LEMMA 3. *$E_{\mathcal{I}}$ as defined in Table 2 is a valid confidence bound: $Pr\left[|\mathcal{I}(I) - \hat{\mathcal{I}}(I)| > E_{\mathcal{I}}(I,\delta)\right] \le \delta$. Table 2 details different versions of $E_{\mathcal{I}}$ based on the (exact but loose) Hoeffding bound, and an approximate (but practically useful) bound based on the normal approximation. For the special case $\hat{P}_I^D = P_I^D$ and samples are drawn only from the Bayesian network (but not from the database), additional Hoeffding and normal bounds are given.*

PROOF. Let us begin by giving a bound on the difference of estimated probabilities. Let $X_1,\ldots,X_n$ be independent random variables and let $X_i \in [a_i, b_i]$. Let $S_n = \sum_{i=1}^n X_i$.

Hoeffding's inequality states that

$$Pr\big[|S_n - E(S_n)| \geq \varepsilon\big] \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right),$$

where $E(S_n)$ denotes the expected value of S_n. Since $\hat{P}_I^{BN}(\mathbf{i}) - \hat{P}_I^{D}(\mathbf{i})$ is a sum of $N^{BN}(I)$ random variables taking values in $\{0, \frac{1}{N^{BN}(I)}\}$, and $N^D(I)$ random variables taking values in $\{0, -\frac{1}{N^D(I)}\}$, Equation 11 follows.

$$\begin{aligned} Pr\left[\left|(\hat{P}_I^{BN}(\mathbf{i}) - \hat{P}_I^{D}(\mathbf{i})) - (P_I^{BN}(\mathbf{i}) - P_I^{D}(\mathbf{i}))\right| \geq \varepsilon\right] \\ \leq \quad 2\exp\left(-2\varepsilon^2 \frac{N^{BN}(I)N^D(I)}{N^{BN}(I)+N^D(I)}\right). \end{aligned} \tag{11}$$

In Equation 12 we expand the definition of $\mathcal{I}$. We remove the absolute value in Equation 13 by summing over the two possible ways in which the absolute value can exceed the bound $E_{\mathcal{I}}$. Since $\max_i\{a_i - b_i\} \geq \max_i\{a_i\} - \max_i\{b_i\}$, Equation 14 follows. We apply the union bound in Equation 15, replace the two symmetric differences by the absolute value in Equation 16. Since $|a-b| \geq ||a|-|b||$, Equation 17 follows; we expand $E_{\mathcal{I}}$, apply (11) and arrive in Equation 18

$$Pr\big[|\mathcal{I}(I) - \hat{\mathcal{I}}(I)| \geq E_{\mathcal{I}}(I,\delta)\big] \tag{12}$$

$$= Pr\big[|\max_{\mathbf{i}}|\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| - \max_{\mathbf{i}}|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})|| \geq E_{\mathcal{I}}\big]$$

$$= Pr\big[\max_{\mathbf{i}}|\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| - \max_{\mathbf{i}}|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| \geq E_{\mathcal{I}}\big] \tag{13}$$

$$+ Pr\big[\max_{\mathbf{i}}|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| - \max_{\mathbf{i}}|\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| \geq E_{\mathcal{I}}\big]$$

$$\leq Pr\big[\max_{\mathbf{i}}\Big(|\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| - |P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})|\Big) \geq E_{\mathcal{I}}\big] \tag{14}$$

$$+ Pr\big[\max_{\mathbf{i}}\Big(|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| - |\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})|\Big) \geq E_{\mathcal{I}}\big]$$

$$\leq \sum_{\mathbf{i}}\Big(Pr\big[|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| - |\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| \geq E_{\mathcal{I}}\big] \tag{15}$$

$$+ Pr\big[|P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| - |\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})| \geq E_{\mathcal{I}}\big]\Big)$$

$$= \sum_{\mathbf{i}} Pr\big[||P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})| - |\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})|| \geq E_{\mathcal{I}}\big] \tag{16}$$

$$\leq \sum_{\mathbf{i}} Pr\left[\left|(P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i})) - (\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i}))\right| \geq \sqrt{\frac{1}{2}\frac{N^{BN}(I)+N^D(I)}{N^{BN}(I)N^D(I)}\log\frac{2|\operatorname{Dom}(I)|}{\delta}}\right] \tag{17}$$

$$= \sum_{\mathbf{i}} \frac{\delta}{|\operatorname{Dom}(I)|} = \delta \tag{18}$$

To prove the bounds based on normal approximation notice that $\hat{P}_I^{BN}(\mathbf{i})$ follows the binomial distribution, which can be approximated by the normal distribution with mean $P_I^{BN}(\mathbf{i})$ and standard deviation $\sqrt{(N^{BN}(I))^{-1}P_I^{BN}(\mathbf{i})\cdot(1-P_I^{BN}(\mathbf{i}))}$. When sampling from data, $\hat{P}_I^{D}(\mathbf{i})$ follows the hypergeometric distribution which can be approximated by the normal distribution with mean $P_I^{D}(\mathbf{i})$, and standard deviation

$$\sqrt{\frac{P_I^D(\mathbf{i})(1-P_I^D(\mathbf{i}))}{N^D(I)}\cdot\frac{|D|-N^D(I)}{|D|-1}}. \tag{19}$$

Combining the two we get

$$Pr\left[|(\hat{P}^{BN}(\mathbf{i}) - \hat{P}^{D}(\mathbf{i})) - (P^{BN}(\mathbf{i}) - P^{D}(\mathbf{i}))| \geq z_{1-\frac{\delta}{2}} \sqrt{\frac{\hat{P}_I^{BN}(\mathbf{i})(1-\hat{P}_I^{BN}(\mathbf{i}))}{N^{BN}(I)} + \frac{\hat{P}_I^{D}(\mathbf{i})(1-\hat{P}_I^{D}(\mathbf{i}))}{N^{D}(I)}\cdot\frac{|D|-N^D(I)}{|D|-1}}\right] \leq \delta. \tag{20}$$

Since we use the estimates of probabilities to compute the standard deviation, Student's t distribution governs the exact distribution, but for large sample sizes used in the algorithm the t distribution is very close to normal.

The proof is identical to the Hoeffding case until Equation 16, where the Hoeffding bound needs to be replaced by the above expression. The special case of sampling only from the Bayesian network ($\hat{P}_I^D = P_I^D$) follows immediately from the more general case discussed in detail. ∎

LEMMA 4. *E_s as defined in Table 2 is a valid confidence bound for the support: $Pr\big[|\operatorname{supp}(I) - \widehat{\operatorname{supp}}(I)| > E_s(I,\delta)\big] \leq \delta$. Table 2 details a Hoeffding bound and an approximate normal bound. For the special case that $\hat{P}_I^D = P_I^D$ and samples are drawn only from the Bayesian network, Hoeffding and normal bounds are given, too.*

PROOF. In Equation 21, we expand the support defined in Equation 10. To replace the absolute value, we sum over both ways in which the absolute difference can exceed E_s in Equation 22. In Equation 23, we exploit $\max_i\{a_i - b_i\} \geq \max_i\{a_i\} - \max_i\{b_i\}$; we then use the union bound and introduce the absolute value again in Equation 24. Equation 25 expands the definition of E_s; the Chernoff bound (Equation 26) proves that the the confidence is in fact δ.

$$Pr\big[|\widehat{\operatorname{supp}}(I) - \operatorname{supp}(I)| \geq E_s(I,\delta)\big] \tag{21}$$

$$= Pr\big[|\max_{\mathbf{i}}\max\{\hat{P}_I^{BN}(\mathbf{i}), \hat{P}_I^{D}(\mathbf{i})\} - \max_{\mathbf{i}}\max\{P_I^{BN}(\mathbf{i}), P_I^{D}(\mathbf{i})\}| \geq E_s\big]$$

$$= Pr\big[\max_{\mathbf{i}}\max\{\hat{P}_I^{BN}(\mathbf{i}), \hat{P}_I^{D}(\mathbf{i})\} - \max_{\mathbf{i}}\max\{P_I^{BN}(\mathbf{i}), P_I^{D}(\mathbf{i})\} \geq E_s\big] \tag{22}$$

$$+ Pr\big[\max_{\mathbf{i}}\max\{P_I^{BN}(\mathbf{i}), P_I^{D}(\mathbf{i})\} - \max_{\mathbf{i}}\max\{\hat{P}_I^{BN}(\mathbf{i}), \hat{P}_I^{D}(\mathbf{i})\} \geq E_s\big]$$

$$\leq Pr\big[\max_{\mathbf{i}}\max\{\hat{P}_I^{BN}(\mathbf{i}) - P_I^{BN}(\mathbf{i}), \hat{P}_I^{D}(\mathbf{i}) - P_I^{D}(\mathbf{i})\} \geq E_s\big] \tag{23}$$

$$+ Pr\big[\max_{\mathbf{i}}\max\{P_I^{BN}(\mathbf{i}) - \hat{P}_I^{BN}(\mathbf{i}), P_I^{D}(\mathbf{i}) - \hat{P}_I^{D}(\mathbf{i})\} \geq E_s\big]$$

$$\leq \sum_{\mathbf{i}} Pr\big[|\hat{P}_I^{BN}(\mathbf{i}) - P_I^{BN}(\mathbf{i})| \geq E_s\big] + Pr\big[|P_I^{D}(\mathbf{i}) - P_I^{D}(\mathbf{i})| \geq E_s\big] \tag{24}$$

$$\leq \sum_{\mathbf{i}} Pr\left[|\hat{P}_I^{BN}(\mathbf{i}) - P_I^{BN}(\mathbf{i})| \geq \sqrt{\frac{1}{2N^{BN}(I)}\log\frac{4|\operatorname{Dom}(I)|}{\delta}}\right] \tag{25}$$

$$+ Pr\left[|\hat{P}_I^{D}(\mathbf{i}) - P_I^{D}(\mathbf{i})| \geq \sqrt{\frac{1}{2N^{D}(I)}\log\frac{4|\operatorname{Dom}(I)|}{\delta}}\right]$$

$$= \sum_{\mathbf{i}}\left[2\frac{\delta}{2|\operatorname{Dom}(I)|}\right] = \delta \tag{26}$$

For the normal approximation based bounds we start with Equation 25 above which becomes

$$\sum_{\mathbf{i}} Pr\left[|\hat{P}_I^{BN}(\mathbf{i}) - P_I^{BN}(\mathbf{i})| \geq z_{1-\frac{\delta}{4|\operatorname{Dom}(I)|}}\sqrt{\frac{\hat{P}_I^{BN}(\mathbf{i})\cdot(1-\hat{P}_I^{BN}(\mathbf{i}))}{N^{BN}(I)}}\cdot\right]$$

$$+ Pr\left[|\hat{P}_I^{D}(\mathbf{i}) - P_I^{D}(\mathbf{i})| \geq z_{1-\frac{\delta}{4|\operatorname{Dom}(I)|}}\sqrt{\frac{\hat{P}_I^{D}(\mathbf{i})(1-\hat{P}_I^{D}(\mathbf{i}))}{N^{D}(I)}\cdot\frac{|D|-N^D(I)}{|D|-1}}\right]$$

$$= \sum_{\mathbf{i}}\left[2\frac{\delta}{2|\operatorname{Dom}(I)|}\right] = \delta.$$

The special case of $\hat{P}_I^D = P_I^D$ follows immediately from the general case. ∎

LEMMA 5. *E_d as defined in Table 2 is a valid, data independent confidence bound for the interestingness value of an attribute set values: $Pr\big[|\mathcal{I}(I,\mathbf{i}) - \hat{\mathcal{I}}(I,\mathbf{i})| >$*

$E_d(n^{BN}, n^D, \delta)\big] \leq \delta$. *A Hoeffding bound for the special case that* $\hat{P}_I^D = P_I^D$ *and samples are drawn only from the Bayesian network is given.*

PROOF. The proof for the Hoeffding inequality based bound follows directly from (11) in the proof of Lemma 3. For the normal case, it follows from (20) by substituting $\hat{P}_I^{BN}(\mathbf{i}) = \hat{P}_I^D(\mathbf{i}) = \frac{1}{2}$ which corresponds to the maximum possible standard deviation. ∎

THEOREM 3. *Let G be the collection of attribute sets output by the algorithm. After the algorithm terminates the following condition holds with the probability of $1-\delta$:*

$$\begin{aligned} &\textit{there is no } I' \in 2^Z \setminus \{\emptyset\} \textit{ such that } I' \notin G \\ &\textit{and } \mathcal{I}(I') > \min_{I \in G} \mathcal{I}(I) + \varepsilon \end{aligned} \tag{27}$$

PROOF. We will first assume that, throughout the course of the algorithm, the estimates of all quantities lie within their confidence intervals (assumptions A1a, A1b, and A2). We will show that under this assumption the assertion in Equation 27 is always satisfied when the algorithm terminates. We will then quantify the risk that over the entire execution of the algorithm at least one estimate lies outside of its confidence interval; we will bound this risk to at most δ. These two parts prove Theorem 3.

(A1a) $\forall i \in \{1, \ldots, i_{max}\} \forall I \in H_i : |\hat{\mathcal{I}}(I) - \mathcal{I}(I)| \leq E_{\mathcal{I}}\left(I, \frac{\delta}{3|H_i|i(i+1)}\right)$

(A1b) $\forall i \in \{1, \ldots, i_{max}\} \forall I \in H_i : |\widehat{\text{supp}}(I) - \text{supp}(I)| \leq E_s\left(I, \frac{\delta}{3|H_i|i(i+1)}\right)$

(A2) If $E_d\Big(n_{i_{max}}^{BN}, n_{i_{max}}^D, \frac{\delta\left(1-\frac{2}{3}\sum_{j=1}^{i}\frac{1}{j(j+1)}\right)}{\sum_{I \in H_{i_{max}}} |\text{Dom}(I)|}\Big) \leq \frac{\varepsilon}{2}$ then $\forall I \in H_{i_{max}}^* \forall I' \in (H_{i_{max}} \setminus H_{i_{max}}^*) : \mathcal{I}(I) \geq \mathcal{I}(I') - \varepsilon$

Equation (Inv1) shows the main loop invariant which, as we will now show, is satisfied after every iteration of the main loop as well as when the loop is exited.

(Inv1) $\forall K \in R_i$ there exist distinct $I_1, \ldots, I_n \in H_i : \forall j \in \{1, \ldots, n\} \mathcal{I}(I_j) \geq \mathcal{I}(K)$

We will prove the loop invariant (Inv1) by induction. For the base case ($R_i = \emptyset$), (Inv1) is trivially true. For the inductive step, let us assume that (Inv1) is satisfied for R_i and H_i before the loop is entered and show that it will hold for R_{i+1} and H_{i+1} after the iteration. (Inv1) refers to R and H, so we have to study steps 2c, 2d, and 2h, which alter these sets. Note that, by the definition of R, R_{i+1} is always a superset of R_i; it contains all elements of R_i in addition to those that are added in steps 2c and 2d.

Step 2c
Let K be an attribute set pruned in this step. The pruning condition together with our definition of support (Equation 10) implies Equation 28; we omit the confidence parameter of E_s for brevity. Equation 28 is equivalent to Equation 29. Assumption (A1a) says that $\hat{\mathcal{I}}(I'') - E_{\mathcal{I}}(I'') \leq \mathcal{I}(I'')$; from assumption (A1b) we can conclude that $\widehat{\text{supp}}(K) + E_s(K) \geq \text{supp}(K)$ which leads to Equation 30. From the definition of support, it follows that all supersets J of K must have a smaller or equal support (Equation 31); Lemma 1 now implies that if the support of K is lower than that of J, so must be the interestingness (Equation 32).

$$\widehat{\text{supp}}(K) \leq \min_{I \in H_i^*} \left\{\hat{\mathcal{I}}(I) - E_{\mathcal{I}}(I)\right\} - E_s(K) \tag{28}$$

$$\Leftrightarrow \forall I'' \in H_i^* : \widehat{\text{supp}}(K) + E_s(K) \leq \hat{\mathcal{I}}(I'') - E_{\mathcal{I}}(I'') \tag{29}$$

$$\Rightarrow \forall I'' \in H_i^* : \text{supp}(K) \leq \mathcal{I}(I'') \tag{30}$$

$$\Rightarrow \forall I'' \in H_i^* \forall J \supseteq K : \text{supp}(J) \leq \mathcal{I}(I'') \tag{31}$$

$$\Rightarrow \forall I'' \in H_i^* \forall J \supseteq K : \mathcal{I}(J) \leq \mathcal{I}(I'') \tag{32}$$

K cannot be an element of H_i^* because, in order to satisfy Equation 28, the error bound E_s would have to be zero or negative which can never be the case. Since $K \notin H_i^*$, and $|H_i^*| = n$, we can choose $I_1, \ldots, I_n$ to lie in H_i^*. AprioriBNS now prunes K and all supersets $J \supseteq K$, but Equation 32 implies that for any $J \supseteq K$: $\mathcal{I}(J) \leq \mathcal{I}(I_1), \ldots, \mathcal{I}(I_n)$. Therefore, (Inv1) is satisfied for $R_{i+1} = R_i \cup$ (supersets of K) and the "new" H_i ($H_i \setminus$ rejected hypotheses).

Step 2d
Let K be one of the attribute sets rejected in this step. The condition of rejection implies Equation 33; we omit the confidence parameter of $E_{\mathcal{I}}$ for brevity. Let I'' be any attribute set in H_i^*. Equation 33 implies Equation 34. Together with assumption (A1a), this leads to Equation 35.

$$\hat{\mathcal{I}}(K) \leq \min_{I \in H_i^*} \left\{\hat{\mathcal{I}}(I) - E_{\mathcal{I}}(I)\right\} - E_{\mathcal{I}}(K) \tag{33}$$

$$\Leftrightarrow \quad \forall I'' \in H_i^* : \hat{\mathcal{I}}(K) + E_{\mathcal{I}}(K) \leq \hat{\mathcal{I}}(I'') - E_{\mathcal{I}}(I'') \tag{34}$$

$$\Rightarrow \quad \forall I'' \in H_i^* : \mathcal{I}(K) \leq \mathcal{I}(I'') \tag{35}$$

Note also that a rejected hypothesis K cannot be an element of H_i^* because otherwise the error bounds $E_{\mathcal{I}}$ and E_s would have to be zero or negative which can never be the case. Since $K \notin H_i^*$, and $|H_i^*| = n$, we can choose $I_1, \ldots, I_n$ to lie in H_i^* and Equation 35 implies 36. Since furthermore $R_{i+1} = R_i \cup \{K\}$, Equation 36 implies (Inv1) for R_{i+1} and the "new" H_i ($H_i \setminus$ rejected hypotheses); below "$\exists^*$" abbreviates "there exist distinct".

$$\exists^* I_1, \ldots, I_n \in H_i \setminus \{K\} : \forall j \in \{1, \ldots, n\} \mathcal{I}(I_j) \geq \mathcal{I}(K) \tag{36}$$

This implies that (Inv1) holds for R_{i+1} and the current state of H_i after step 2d.

Step 2h
R_{i+1} is not altered, H_{i+1} is assigned a superset of H_i. (Inv1) requires the existence of n elements in H. If it is satisfied for R_{i+1} and H_i (which we have shown in the previous paragraph), it also has to be satisfied for any superset $H_{i+1} \supseteq H_i$. This proves that the loop invariant (Inv1) is satisfied after each loop iteration.

Final Step (immediately before Step 3)
The main loop terminates only when $C_i = \emptyset$, from Lemma 2 we know that $U_{i_{max}} = \emptyset$. Since $U_{i_{max}} = \emptyset$, and $G = H_{i_{max}}^*$ we have $2^Z \setminus (\{\emptyset\} \cup G) = R_{i_{max}} \cup (H_{i_{max}} \setminus H_{i_{max}}^*)$ and it suffices to show that all attribute sets in G are better than all sets in $R_{i_{max}}$ and in $H_{i_{max}} \setminus H_{i_{max}}^*$. We distinguish between the two possible termination criteria of the main loop.

Case (a): Early stopping in Step 2e
The stopping criterion, we are assured the Equation 37 is satisfied. By assumption (A1a), this implies Equation 38.

$$\begin{aligned} &\forall I \in H_i^*, I' \in H_i \setminus H_i^* : \\ &\qquad \hat{\mathcal{I}}(I) + E_{\mathcal{I}}(I) > \hat{\mathcal{I}}(I') - E_{\mathcal{I}}(I') - \varepsilon \qquad (37) \\ &\Rightarrow \quad \forall I \in H_i^*, I' \in H_i \setminus H_i^* : \mathcal{I}(I) > \mathcal{I}(I') - \varepsilon \qquad (38) \end{aligned}$$

From the invariant (Inv1) we know that $\forall K \in R_{i_{max}} \exists^* I_1, \ldots, I_n \in H_{i_{max}} : \forall j \in \{1, \ldots, n\} \mathcal{I}(I_j) \geq \mathcal{I}(K)$; that is, for every rejected hypothesis there are n hypotheses in H_i which are at least as good. Take any such $S = \{I'_1, \ldots, I'_n\}$. For every $I' \in S$ either $I' \in H^*_{i_{max}}$ or $I' \notin H^*_{i_{max}}$. In the former case it follows immediately that $I' \in G$; that is, I' is better than the rejected K and I' is in the returned set G. If $I' \notin H^*_{i_{max}}$, then Equation 38 guarantees that every hypothesis $I \in H^*_{i_{max}}$ is "almost as good as I'": $\forall I \in H^*_{i_{max}} : \mathcal{I}(I) \geq \mathcal{I}(I') - \varepsilon$. This proves case (a) of Theorem 3.

Case (b): Stopping in Step 2f
Assumption (A2) assures Equation 39.

$$\forall I \in H^*_{i_{max}} \forall I' \in (H_{i_{max}} \setminus H^*_{i_{max}}) \mathcal{I}(I) \geq \mathcal{I}(I') - \varepsilon \quad (39)$$

Analogously to case (a), we can argue that (Inv1) guarantees that $\forall K \in R_{i_{max}} \exists^* I_1, \ldots, I_n \in H_{i_{max}} : \forall j \in \{1, \ldots, n\} \mathcal{I}(I_j) \geq \mathcal{I}(K)$. Identically to case (a), this implies Theorem 3.

We have shown that if the main loop terminates, the output will be correct. It is easy to see that the loop will in fact terminate after finitely many iterations: Since Z is finite, the candidate generation has to stop at some point i with $C_i = \emptyset$. When the sample size becomes large enough, the loop will be exited in step 2f. This is guaranteed because a guaranteed fraction $\frac{\delta}{3}$ is reserved for the error bound of step 2f and the error bound (Table 2) vanishes for large sample sizes.

Risk of violation of (A1a), (A1b), and (A2)
We have proven Theorem 3 under assumptions (A1a), (A1b), and (A2). We will now bound the risk of a violation of any of these assumptions during the execution of AprioriBNS. We first focus on the risk of a violation of (A1a). A violation of $|\mathcal{I}(I) - \hat{\mathcal{I}}(I)| \leq E_{\mathcal{I}}$ can occur in any iteration of the main loop and for any $I \in H_i$ (Equation 40). We use the union bound to take all of these possibilities into account (Equation 41). Lemma 3 implies Equation 42.

$$\begin{aligned} & Pr[(A1a) \text{ is violated for some } I \text{ in some iteration}] \\ = \; & Pr\left[\bigvee_{i=1}^{i_{max}} \bigvee_{I \in H_i} \left|\hat{\mathcal{I}}(I) - \mathcal{I}(I)\right| > E_{\mathcal{I}}(I)\right] && (40) \\ \leq \; & \sum_{i=1}^{i_{max}} \sum_{I \in H_i} Pr\left[\left|\hat{\mathcal{I}}(I) - \mathcal{I}(I)\right| > E_{\mathcal{I}}\left(I, \frac{\delta}{3|H_i|i(i+1)}\right)\right] && (41) \\ \leq \; & \sum_{i=1}^{i_{max}} \sum_{I \in H_i} \frac{\delta}{3|H_i|i(i+1)} = \frac{\delta}{3} \sum_{i=1}^{i_{max}} \frac{1}{i(i+1)} && (42) \end{aligned}$$

The risk of violating assumption (A1b) can be bounded similarly in Equations 43 and 44.

$$\begin{aligned} & Pr[(A1b) \text{ is violated for some } I \text{ in some iteration}] \\ = \; & Pr\left[\bigvee_{i=1}^{i_{max}} \bigvee_{I \in H_i} |\widehat{\text{supp}}(I) - \text{supp}(I)| > E_s(I)\right] && (43) \\ \leq \; & \sum_{i=1}^{i_{max}} \sum_{I \in H_i} Pr\left[|\widehat{\text{supp}}(I) - \text{supp}(I)| > \right. && (44) \\ & \left. E_s\left(I, \frac{\delta}{3|H_i|i(i+1)}\right)\right] = \frac{\delta}{3} \sum_{i=1}^{i_{max}} \frac{1}{i(i+1)} \end{aligned}$$

We now address the risk of a violation of (A2). In step 2b, H^*_i is assigned the hypotheses with highest values of $\hat{\mathcal{I}}(I)$; *i.e.,* for all $I \in H^*_i$ and $I' \notin H^*_i : \hat{\mathcal{I}}(I) \geq \hat{\mathcal{I}}(I')$. For (A2) to be violated, there has to be an $I \in H^*_{i_{max}}$ and an $I' \in H_{i_{max}} \setminus H^*_{i_{max}}$ such that $\mathcal{I}(I) < \mathcal{I}(I') - \varepsilon$ but Equation 45 is satisfied in spite. This is only possible if there is at least one hypothesis $I \in H_{i_{max}}$ with $|\mathcal{I}(I) - \hat{\mathcal{I}}(I)| > \frac{\varepsilon}{2}$. Intuitively, Equation 45 assures that all elements of $H_{i_{max}}$ have been estimated to within a two-sided confidence interval of $\frac{\varepsilon}{2}$; since all $I \in H^*_{i_{max}}$ appear at least as good as $I' \notin H^*_{i_{max}}$, I' can be at most ε better than I.

$$E_d\left(n^{BN}_{i_{max}}, n^D_{i_{max}}, \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i} \frac{1}{j(j+1)}\right)}{\sum_{I \in H_i} |\text{Dom}(I)|}\right) \leq \frac{\varepsilon}{2} \quad (45)$$

In Equation 46 we substitute Equation 45 into this condition. We expand the definition of interestingness in Equation 47, use the union bound in Equation 48 and refer to Lemma 5 in Equation 49.

$$\begin{aligned} & Pr\left[\exists I \in H_{i_{max}} : |\hat{\mathcal{I}}(I) - \mathcal{I}(I)| > \frac{\varepsilon}{2}\right] \\ \leq \; & Pr\left[\exists I \in H_{i_{max}} : |\hat{\mathcal{I}}(I) - \mathcal{I}(I)| > \right. \\ & \left. E_d\left(n^{BN}_{i_{max}}, n^D_{i_{max}}, \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i_{max}} \frac{1}{j(j+1)}\right)}{\sum_{I \in H_i} |\text{Dom}(I)|}\right)\right] && (46) \\ \leq \; & Pr\left[\exists I \in H_{i_{max}}, \mathbf{i} \in \text{Dom}(I) : \right. \\ & \left. \left||\hat{P}^{BN}_I(\mathbf{i}) - \hat{P}^D_I(\mathbf{i})| - |P^{BN}_I(\mathbf{i}) - P^D_I(\mathbf{i})|\right| > \right. \\ & \left. E_d\left(n^{BN}_{i_{max}}, n^D_{i_{max}}, \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i_{max}} \frac{1}{j(j+1)}\right)}{\sum_{I \in H_i} |\text{Dom}(I)|}\right)\right] && (47) \\ \leq \; & \sum_{\substack{I \in H_{i_{max}}, \\ \mathbf{i} \in \text{Dom}(I)}} Pr\left[\left||\hat{P}^{BN}_I(\mathbf{i}) - \hat{P}^D_I(\mathbf{i})| - |P^{BN}_I(\mathbf{i}) - P^D_I(\mathbf{i})|\right| > \right. \\ & \left. E_d\left(n^{BN}_{i_{max}}, n^D_{i_{max}}, \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i_{max}} \frac{1}{j(j+1)}\right)}{\sum_{I \in H_i} |\text{Dom}(I)|}\right)\right] && (48) \\ \leq \; & \sum_{\substack{I \in H_{i_{max}}, \\ \mathbf{i} \in \text{Dom}(I)}} \frac{\delta\left(1 - \frac{2}{3}\sum_{j=1}^{i_{max}} \frac{1}{j(j+1)}\right)}{\sum_{I \in H_i} |\text{Dom}(I)|} && (49) \\ = \; & \delta\left(1 - \frac{2}{3}\sum_{j=1}^{i_{max}} \frac{1}{j(j+1)}\right) && (50) \end{aligned}$$

We can now calculate the combined risk of any violation of (A1a), (A1b), or (A2) using the union bound in Equation 51; this risk can be bounded to at most δ in Equation 52 (note that $\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = 1$).

$$\begin{aligned} & Pr[\text{(A1a), (A1b), or (A2) violated during execution}] \\ \leq \; & \frac{2\delta}{3}\sum_{i=1}^{i_{max}} \frac{1}{i(i+1)} + \delta\left(1 - \frac{2}{3}\sum_{i=1}^{i_{max}} \frac{1}{i(i+1)}\right) && (51) \\ = \; & \delta \sum_{i=1}^{i_{max}} \frac{1}{i(i+1)} < \delta && (52) \end{aligned}$$

This completes the proof of Theorem 3. ∎
Together, Theorems 3 and 2 prove Theorem 1. ∎