RC22270 (W0112-008) December 3, 2001
Computer Science

IBM Research Report

Inference complexity as a model-selection criterion for learning Bayesian networks

Alina Beygelzimer
Department of Computer Science
University of Rochester, Rochester, NY 14627

Irina Rish
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Inference complexity as a model-selection criterion for learning Bayesian networks

Alina Beygelzimer
Department of Computer Science
University of Rochester, Rochester, NY 14627
beygel@cs.rochester.edu

Irina Rish
IBM T. J. Watson Research Center
30 Saw Mill River Road, Hawthorne, NY 10532
rish@us.ibm.com

Abstract

Learning Bayesian networks from data is commonly viewed as a (constrained) optimization problem, where a particular scoring metric is used for evaluating the quality of a candidate model. Commonly used scoring metrics, such as BIC/MDL, prefer models that fit the data well and have low representation complexity (i.e. the number of parameters). In this paper, we argue that a model selection criteria must also include another important property of Bayesian networks, namely their *inference complexity*, which is exponential in *treewidth* of the underlying graph. We demonstrate that traditional metrics may not distinguish between networks with drastically different treewidths. In particular, we show that it is quite likely to have networks that represent close distributions (in terms of KL-divergence) and have similar representation complexity, but fairly different treewidths. We also investigate the relationships between the treewidth and other structural properties of graphs that suggest efficient probabilistic tests for a quick estimation of the treewidth during learning.

Keywords: *Bayesian networks*, *learning*, *efficient inference*, *treewidth*.

1 Introduction

This paper is motivated by the goal of learning Bayesian networks that allow *efficient inference*, as opposed to learning the networks first without consideration of their inference complexity and *then* dealing with this complexity by means of approximation. As we demonstrate in this paper, traditional model selection criteria that aim at fitting the data and minimizing the *representation* complexity are unable to capture the *inference* complexity, and thus must be extended appropriately.

A Bayesian network (BN) is a graphical probabilistic model that exploits conditional independencies among a set of random variables $\mathbf{X} = \{X_1, ..., X_n\}$ to provide a compact product-form representation of their joint probability distribution $P(\mathbf{X})$. The *structure* of a BN is a directed acyclic graph (dag), denoted $G = (V, E)$, where the nodes in V correspond (by one-to-one mapping) to the variables[1] and the edges encode probabilistic dependencies. Every node X_i is associated with a local conditional probability distribution (CPD) $\mathbf{P}(X_i \mid \Gamma(X_i))$, where $\Gamma(X_i)$ denotes the set of *parents* of node X_i (i.e., nodes pointing to X_i) in the graph; a node without parents is associated with its prior distribution $\mathbf{P}(X_i)$. Thus, a Bayesian network is a pair (G, Θ) where G is its dag, and Θ is the set of all CPD parameters. The joint distribution encoded by the network is given by the product of the CPDs: $\mathbf{P}(X_1, ..., X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid \Gamma(X_i))$. By the *representation complexity* of a particular Bayesian network structure we mean the number of *independent parameters* (i.e., the minimal number of conditional and prior probabilities) needed to specify the network. For example, to specify the CPD for a node X with k parents we need 2^k probabilities $P(X = 0 \mid \gamma(X))$, where $\gamma(X)$ is a particular assignment to the parents of X. Given a network structure G, let $\mu(G)$ denote the representation complexity of G.

Bayesian networks are used for *probabilistic inference*, i.e. for answering probabilistic queries, such as *belief updating*, or finding the posterior probability $P(\mathbf{Y}|\mathbf{Z} = \mathbf{z})$ of an unknown set of variables $Y \subseteq \mathbf{X}$ given *evidence* $\mathbf{Z} = \mathbf{z}$, or finding the most-likely variable assignment $\mathbf{y}' = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{Z} = \mathbf{z})$, called the *maximum aposteriory probability (MAP)* assignment. However, (exact) probabilistic inference in Bayesian networks is NP-hard [4]. Commonly used inference algorithms, such as the *junction tree* (or *join-*

[1] A common assumption in Bayesian network literature is that random variables have a finite set of values. In this paper, we particularly focus on networks with binary variables; extending the results to general finite domains is straightforward.

tree) algorithm [10] and closely related to it *variable-elimination* techniques [5, 18, 6], are exponential in the size (the dimensionality) of largest dependence recorded by the algorithm, which is equivalent to the size of largest clique (minus 1) "induced" in the *moral graph* of BN[2]. This graph parameter is called the *induced width*, or the *treewidth* [7], and is formally defined as follows. Given a (directed or undirected) graph G, the *width* of X_i along ordering o is the number of X_i's neighbors preceding X_i in o. The *width of the graph* along o, denoted w_o, is the maximum width along o. The *induced graph* of G along o is obtained by connecting the preceding neighbors of each X_i, going from $i = n$ to $i = 1$. The induced width along o, denoted w_o^*, is the width of the induced graph along o, while the induced width w^* is the minimum induced width along any ordering.

Learning Bayesian networks from data has recently become an active area of research [9]. A typical approach is based on defining (1) a scoring metric for evaluating the quality of a given structure, and (2) a search strategy for traversing the space of candidate models using some local transformations (e.g. adding, deleting, or reversing an edge). A learning algorithm then searches for a model that optimizes the scoring metric. Related approaches may also use some prior knowledge about the domain (e.g. marginal independence between certain variables) to constrain the search.

Commonly used information-theoretic scoring metrics such as BIC or MDL, AIC, MML and others, provide a tradeoff between the accuracy and the complexity of a model by minimizing the sum of the encoding length of the model (i.e. its representation complexity) and the encoding length of the data given the model, which is measured by the negative log-likelihood and thus reflects the model error. Formally, the *minimum description length* (MDL) [15] score of a Bayesian network $B = (G, \Theta)$ (which is equivalent to Bayesian information criterion, or BIC), is defined as $\mathrm{MDL}(G, \Theta \mid D) = \frac{\log N}{2}\mu(G) - \log \mathbf{P}(D \mid G, \Theta)$, where D is a set of N observations (training data), and $\mu(G)$ is the number of independent parameters needed to specify the CPDs. The first term of the MDL score is the description length of the model, while the second term, the negative log-likelihood of the model given data, is the number of bits needed to describe the data given the model. Thus

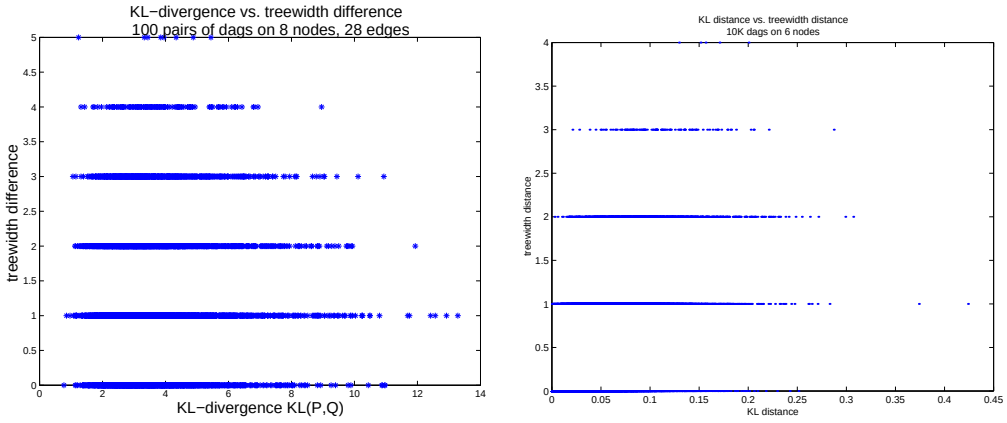
[2]The *moral* graph of G results from connecting ("marrying") all the parents of every node in G by an undirected clique, then dropping the directionality of other edges. Moral graphs basically allow to represent the set of dependencies encoded by the original directed graph without the use of edge directionality. (Notice that every pair of nodes participating in the same CPD is connected in the moral graph).

MDL favors models that predict data better (have higher log-likelihood) and have lower representation complexity. However, it is independent of other important properties of the learned model, in particular, its inference complexity. For example, we may learn two networks that fit the data equally well, and have similar representation complexity, yet their treewidth (and therefore, their inference complexity) is quite different. In this paper, we provide asymptotic (with increasing networks size) theoretical analysis and some empirical evidence suggesting that such situations are frequent enough to argue for using the treewidth test during learning. We demonstrate that two networks having drastically different treewidths may quite often represent close distributions (in terms of KL-divergence) and have similar number of parameters (and thus similar BIC/MDL scores). Moreover, we describe a construction of two large natural classes of BNs that have this property.

In this paper, we argue that learning algorithms should incorporate treewidth as a part of their model selection criteria. Clearly, testing the treewidth of a candidate model during learning must be very efficient. Computing the exact treewidth of a given graph is known to be NP-hard, although there are polynomial-time heuristics for upper bounding the treewidth [7, 1]. However, we do not need to compute the treewidth exactly, nor we need a good approximation of it; it suffices to have a quick probabilistic test that asserts whether the treewidth is "small enough" or very large (i.e. far from being "small enough"). We want the complexity of such test to be independent of n, depending only on the quality of this assertion and the parameter quantifying "small enough". Furthermore, since the search is usually based on local transformations of the structure (e.g. adding, deleting, or reversing a single edge), the treewidth test can be incremental. We initiate an investigation of the relationships between structural properties of graphs and the complexity of inference in the graphs. In particular, we explore the relationships between the number of parameters needed to specify the network and the treewidth of the network. These investigations suggest efficient probabilistic tests for predicting the treewidth from other (easy-to-compute) graph invariants without actually computing it directly.

2 Treewidth difference of similar distributions

In this section, we outline our preliminary results suggesting that it is quite common to encounter two Bayesian network models that represent similar



(a) (b)

Figure 1: (a) treewidth difference (y-axis) versus KL-divergence (x-axis) for 100 pairs of randomly generated Bayesian networks (8 nodes, 28 edges). (b) treewidth difference (y-axis) versus symmetric KL-divergence (x-axis) on 10,000 Bayesian networks with 6 nodes constructed using the "strong-core" technique.

distributions and yet have very different treewidths[3]. We show that there are many networks that encode close distributions (measured by their *KL-divergence*), but have a relatively large difference in the treewidth. The *KL-divergence* between two probability distributions p and q is defined as $KL(p,q) = \mathbf{E}[\log p/q]$, where the expectation is taken with respect to p. We will use the "symmetrical" KL-divergence defined as $d(p,q) = KL(p,q) + KL(q,p)$.

Figure 1(a) shows KL-divergence versus treewidth difference for 100 pairs of randomly generated Bayesian networks on $n = 8$ nodes with $\binom{n}{2} = 28$ edges. We observe no correlation between the treewidth difference and the KL-divergence: close distribution may have quite different treewidth, and vice versa.

Furthemore, we provide a way of constructing BNs that represent close

[3]Note that our analysis is asymptotic, i.e. we assume sufficiently large data sets, so that we can compare the distributions represented by competing models directly, instead of comparing the corresponding data likelihoods used in model-selection criteria.

distributions whose treewidths can be very different. Given an "initial" BN $B = (G, \Theta)$ representing a joint probability distribution $P(\mathbf{X})$, we will construct a family of BNs by significantly modifying G, and at the same time keeping a small distance between $P(\mathbf{X})$ and its projection (in terms of symmetric KL-divergence) onto the set of distributions consistent with a modified G. The main idea here is that removal (or addition) of "weak" edges, i.e. the edges corresponding to "weak" probabilistic dependencies will not change much the distribution but may significantly change the underlying graph G. The strength of an edge between the variable X_i and its parent Y, denoted as $I_i(Y)$, can be quantified, for example, by the mutual information between the two variables, as proposed in [13] (see [3] for details).

Definition 1 *Given a Bayesian network $B = (G, \Theta)$ containing a node X_i and its parent Y, an edge $Y \rightarrow X_i$ is said to be* δ-strong *(for $0 < \delta < 1$) if $I_i(Y) \geq \delta$, where $I_i(Y)$ is a measure of Y's influence on X_i. The* δ-core *of a B, denoted G_δ, is the subgraph of G containing all δ-strong edges and all vertices of G.*

The cores suggest a way of constructing BNs with the desired property. Figure 1(b) shows empirical results from generating networks with strong cores (containing most of the probability mass of the corresponding distributions) and weak symmetric differences. This allows to keep the distributions close and, at the same time, to have a large difference in the treewidth. Clearly, the properties of the network we starts with, determine the limits of the family this network spawns. For example, if the distribution of arc weights (defined in terms of mutual information [13]) is highly non-uniform (i.e. only a few terms make large contributions to the multi-information encoded by the network), then the network must have many "weak" edges that can be removed without losing much accuracy. In our construction (see Figure 1(b)), we started with random dags; this was good enough to give networks with small KL-distance, yet largely different graph structures. Of course, our analysis can only be considered as preliminary and requiring further experiments on larger graphs; we believe that with increasing graph size, we will often observe the increasing gap in the treewidth among BNs representing similar distributions.

3 Representation complexity versus inference complexity

In this section, we focus on the representation complexity part of the model selection criteria. We demonstrate its inability to capture the inference complexity by constructing two parametric families of Bayesian networks that have (asymptotically) similar representation complexity (the number of parameters), but drastically different inference complexity (treewidth).

Notice that a small representation complexity does not imply a small treewidth. Consider a simple, well-known example of the $n \times n$ *grid*, where the direction of all vertical (horizontal) edges is from left to right (from top to bottom respectively). If the nodes are binary, the grid requires less than $4n^2$ parameters. However, it is easy to see that the treewidth of its moral graph is n (notice that the moral graph is already triangulated); hence the inference complexity is exponential in n. The grid was used in [2] to prove the following proposition.

Proposition 1 [2] *For every $n > 3$, there exist dags $G = G(n)$ and $G' = G'(n)$ such that $\mu(G) > \mu(G')$, whereas $w^*(G) = 2$, while $w^*(G') = \sqrt{n}$.*

Proof: See Appendix. ∎

Now we give a construction promised above. In fact, it will be sufficient to restrict ourselves to bipartite dags, an important family of structures that often arise in various diagnosis and prediction problems. For example, one can think of a bipartite graph as a diagnosis network with one component representing independent unknown variables to be diagnosed (e.g., diseases or computer faults) and the other component representing the observed symptoms caused by those unknown variables (see, for example, the QMR-DT network for medical diagnosis [16], or the BN for computer network fault diagnosis [14]).

3.1 Construction

We need to exhibit two families of dags with the same parameter complexity: an *easy*-family $\mathcal{G}' = \{\mathcal{G}'_n\}$ with small treewidth, and a *hard*-family $\mathcal{G}'' = \{\mathcal{G}''_n\}$ with large treewidth, where $\mathcal{G}'_n$ and $\mathcal{G}''_n$ are the families restricted to n-node

dags. Both families will be non-explicit (i.e. given by a probabilistic construction). In fact, both families will be instances of the same parameterized construction, which we define next.

Let the (l, r, d)-*diffuser* be a directed bipartite graph D with l "left" nodes, each of degree d, and r "right" nodes, such as the direction of all edges is from right to left. Define a *random* (l, r, d)-diffuser, constructed by connecting every left node to a randomly chosen subset of d right nodes [4].

Consider such a random diffuser D with parameters $(l = n - 2k, r = 2k, d = k)$. Let $\mathcal{G}'_n$ be the probability space of D, i.e. the space of all bipartite graphs constructed as above. Clearly, the treewidth of the moralized D is at most $2k - 1$. We also have $\mu(D) = (n - 2k)2^k + 2k$. The number of graphs in the family is $\binom{2k}{k}^{n-2k} \sim (\frac{2^{2k}}{\sqrt{2k}})^{n-2k}$. For $k = \log n$, we have $|\mathcal{G}'_n| = O(n^{n/2})$, $\mu = O(n^2)$, and $w^* = O(\log n)$.

Now let the hard family $\mathcal{G}''_n$ be the probability space of a random $(l = n - n^{2/3}, r = n^{2/3}, d = \log n)$-diffuser. We have $|\mathcal{G}''_n| = r^{dl} = O(n^{n \log n})$, which gives a large family of n-node dags with at most $\mu = n^{2/3} + n(n - n^{2/3}) = O(n^2)$ parameters. We will show that the treewidth of the moral graph of such a diffuser is at least $n^{2/3} - 1$.

Lemma 1 *With probability at least* $1 - exp(-\frac{ld^2}{r^2} + 2 \ln r)$, *a random* (l, r, d)-*diffuser has treewidth* $r - 1$.

Proof: See Appendix. ∎

In particular, for $l = n - n^{2/3}$, $r = n^{2/3}$, and $d = \log n$, we have $w^* \geq n^{2/3} - 1$ with probability at least $1 - exp(-n^{1/2} \log^2 n)$. We have defined two large families of bipartite dags such that the graphs in both families have $O(n^2)$ parameters, yet the graphs in one family have treewidth $O(\log n)$ while the graphs in the other family have treewidth $O(n^{2/3})$. Notice that the treewidth ratio in this case is $O(n^{2/3 - \frac{\log \log n}{\log n}})$, which is asymptotically larger than the ratio $O(n^{1/2})$ obtained in Proposition 1.

[4] Random diffusers are equivalent to the graph structure of *low-density parity-check codes (Gallager codes)* [12], currently considered to be the state-of-the-art codes when combined with the appropriate probabilistic decoder [8]. Gallager codes transform a block of r input bits into l additional parity-check bits, each parity-check bit is computed for a subset of d randomly selected input bits.

3.2 Bounds on treewidth

A natural question that arises is that of the exact relation between $\mu(\cdot)$ and $w^*(\cdot)$. In this section we examine the distribution of n-node dags according to their number of parameters. Clearly, the maximum complexity is that of the clique (when all variables are pairwise dependent), in which case $\mu = 2^n - 1$.

Figure 2 shows how $w^*(\cdot)$ is distributed as a function of $\mu(\cdot)$, i.e. it shows all possible combinations of (μ, w^*) pairs over all non-isomorphic dags on a fixed number of nodes.

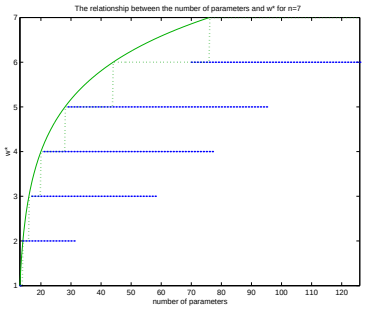


Figure 2: Relationship between $\mu(\cdot)$ and $w^*(\cdot)$. The solid curve corresponds to the upper bound; the corresponding integer upper bound is shown as a step-wise dotted line).

We derive bounds on w^*, which explain the bounding curves in Figure 3.2. Moreover, we show that both the lower and the upper bound below may be attained.

Proposition 2 *(Lower bound) For any dag G of order n,*

$$w^*(G) \geq \log \mu(G) - \log n.$$

Proof: Fix a dag G of order n. Let μ be the representation complexity of G. By the averaging argument, there must exist at least one vertex with $\log(\mu/n)$ parents; therefore, the moralized G has a clique of size at least $\log(\mu/n) + 1$, which proves the proposition. ∎

We show that the above bound is tight for small μ in the sense that there exist dags that come arbitrarily close to it.

Proposition 3 *For each integer $c \geq 1$, there exists a graph G_c with $w^*(G_c) = \frac{1}{c}\log n$ such that $\log \frac{\mu(G_c)}{n} = \log(n^{1/c} - O(\frac{\log n}{n^{1-1/c}}))$.*

Proof: Fix $c \geq 1$. Let G_c be the complete bipartite graph with the smaller component of size $\frac{1}{c}\log n$ such that all the edges are directed from the smaller component to the larger. It is easy to see that $w^*(G_c) = \frac{1}{c}\log n$. On the other hand, $\mu(G_c) = \frac{1}{c}\log n + n^{1/c}(n - \frac{1}{c}\log n)$. The claim easily follows. ∎

Notice that the bound is not tight up to a $\log n$ additive term as $\mu(K_n) = 2^n - 1$, and $w^*(K_n) = n - 1$ (where K_n denotes clique on n nodes), while the bound gives $w^*(K_n) > n - \log n$. The claim follows from the convexity of the bound.

Proposition 4 *(Upper bound) For any dag G of order n, $w^*(G) \leq \lfloor \log(\mu(G) - 2(n+1)) \rfloor + 1$.*

Proof: Omitted (see [3]). ∎

4 Testing the treewidth

In this section we present some observations on how to quickly bound the treewidth during learning. We do not attempt to approximate the treewidth (let alone compute it exactly); this would be an overkill in this setting. We consider a more relaxed task of testing whether the treewidth is "small enough". The tests will be allowed to err with small probability, provided they are very fast (sublinear in the number of nodes in the network). This would allow us to choose between alternative networks without actually incurring the expense of running good (but much more expensive) approximation algorithms for determining the treewidth.

Let $\bar{G} = (V, \bar{E})$ denote the complement (to a complete graph) of an undirected graph $G = (V, E)$, i.e. $\bar{E} = S$
E where S is the set of all possible edges among the nodes in V. Let us also use the notation $w^*(G, \bar{G}) = w^*(G) + w^*(\bar{G})$. First, we show that for any graph G, either G or its complement $\bar{G}$ must have large treewidth, i.e. that $w^*(G, \bar{G})$ is $\Theta(n)$, where $n = |V|$.

Proposition 5 *For any undirected graph G, $w^*(G, \bar{G}) \geq (2 - \sqrt{3})n$.*

Proof: See Appendix. ∎

The proof of the proposition (see Appendix) suggests that any two-sided error probabilistic algorithm for testing whether a given graph G has treewidth at most some fixed integer k, first approximates the density of G by randomly selecting a small set of vertex pairs and outputting the fraction of them that are adjacent in the graph. By the Chernoff bound, a sample of size m is sufficient to get within an additive error of δ with probability at least $\exp(-2\delta^2 m)$; thus setting $m = \delta^{-2}$ upper bounds the probability of error by $e^{-2} < 1/7$. If the estimate is less than $nr - \binom{r+1}{2}$ (as follows from the proof), we reject. Thus we may assume that G has $O(nr)$ edges, i.e. that G is sparse.

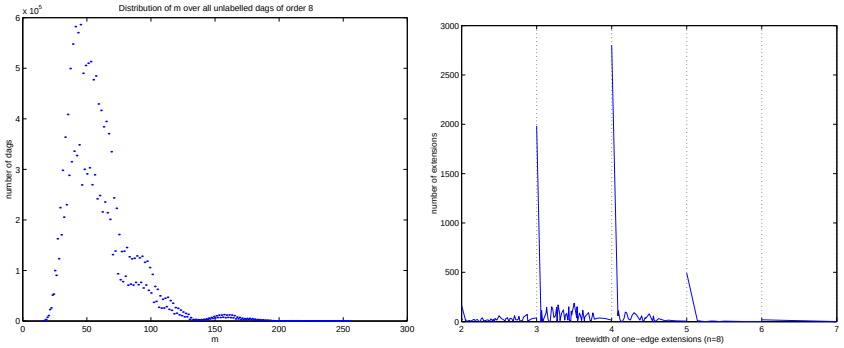
Since most learning algorithms are incremental, i.e. based on local edge transformations, we only need to provide incremental treewidth checks; in other words, we only need to make sure that such local modifications do not inflict large increases in the treewidth. Clearly, no edges should be added to a node, which already has more than r parents (since moralization will induced a clique of size more than $r+1$). Notice that this degree bound is necessary, but not sufficient (recall the grid example). It is not hard to see that adding a single directed edge cannot increase the treewidth of the resulting moral graph by more than one (see Appendix). Furthermore, it turns out that *most* one-edge extensions do not increase the treewidth for undirected graphs. Figure 3(b) shows the statistics for the average increase in treewidth over all one-edge extensions, where x-axis corresponds to the treewidth of the average extension, and y-axis gives the number of undirected graphs whose average extension has the corresponding treewidth.

5 A counting argument

Due to page limit, we give a sketch of the argument. The main observation is that for certain values of μ there exist a huge number of dags with μ parameters, while there are not that many dags with sufficiently small treewidth. Therefore, not all dags with μ parameters can have small treewidth (although there exist such dags); that is, there must exist dags with the same parameter complexity and a whole spectrum of treewidths.

First, we examine how $\mu(\cdot)$ is distributed across all dags on a given number of nodes (see Figure 3). Our purpose here is to explain the behavior of the

distribution and to give a feeling for the asymptotics, rather than to give rigorous estimates; very crude bounds are sufficient for the argument.



(a) $n = 8$ (b) $n = 8$

Figure 3: (a) The distribution of dags according to $\mu(\cdot)$. (b) Average increase in treewidth over all one-edge extensions.

It is prominant that most dags are concentrated on a small number of μ values. Consider a labeled n-node dag G. We can associate G with a linear ordering of nodes $1, \cdots, n$ in a natural way (so that there is no edge from a higher- to a lower-numbered node). Let d_i denote the in-degree of node i; notice that for all i, $d_i \leq i - 1$. We have $\mu(G) = \sum_{i=1}^{n} 2^{d_i}$. What values of μ are shared by many dags? Clearly the multiplicity of μ depends on the number of different partitions of μ, i.e. ordered sequences of in-degrees $(d_1, \ldots, d_n)$ satisfying $\sum_{i=1}^{n} 2^{d_i} = \mu$ and $d_i \leq i - 1$ for all i. The multiplicity of μ is the sum over all such partitions of the number of labeled dags with this in-degree sequence. Clearly, the number of ordered labelled dags with a given sequence $(d_1, \ldots, d_n)$ is bounded by $\binom{n-1}{d_n}\binom{n-2}{d_{n-1}} \cdots \binom{1}{d_2}$, which is maximized[5] when $d_i = \left\lceil \frac{i-1}{2} \right\rceil$ for all i. (This results in dags with $\lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil$ edges.) So what values of μ correspond to this choice of d_i's? We have $\mu = \sum_{i=1}^{n} 2^{\lceil \frac{n-i}{2} \rceil} \leq 3 \cdot 2^{n/2}$. Indeed, the values around $3 \cdot 2^{\frac{n}{2}}$ correspond to the sharp peak in the distribution of μ. This is already true for small values of n. For example, for $n = 8$ the peak is around $3 \cdot 2^{\frac{n}{2}} = 48$ (see Figure 3a).

[5]This number is approximately $\frac{\prod_{i=1}^{n-1} 2^i}{\sqrt{(n-1)!}} \sim 2^{\frac{n^2 - n \log n}{2}}$.

Also notice that there are two self-similar curves in Figure 3a, one above the other, corresponding to odd and even values of $\mu(\cdot)$ respectively. This can be easily explained by observing that μ of a dag is odd exactly when the dag has an odd number of sources (i.e. vertices with in-degree 0). Let $p(n,k)$ be the probability that a random n-node dag has k sources. The limits of $p(n,k)$ were determined by Liskovec [11]. Roughly, $\frac{p(n,k-1)}{p(n,k)} \sim \frac{k2^{k-1}}{\rho A(k)}$, where $\rho \approx 1.488$ and $A(k)$ is a decreasing function in k with values in $(1,2)$. Clearly the ratio grows exponentially with k. We have $A(2) \approx 1.68$, hence the ratio of the number of dags with a single source to the number of dags with two sources is $\frac{4}{\rho A(2)} \approx 1.6$, which explains why there are more dags with odd number of parameters.

Notice a sharp decrease of the number of dags after μ becomes larger than 2^{n-1}. This is due to the fact that all dags with at least 2^{n-1} parameters must have a node of in-degree at least $(n-2)$ (since there can be at most three nodes with in-degree $(n-3)$, and even if the other $(n-3)$ nodes form a clique, the number of parameters can be at most $3 \cdot 2^{n-3} + 2^{n-3} - 1 = 2^{n-1} - 1$). It is not hard to see that the number of such dags is of the same order as the number of dags on $(n-1)$ nodes. Stanley [17] gave an asymptotic formula for the number of labeled n-node dags, yielding the ratio $\sim \frac{n2^{n-1}}{\rho}$ for the number of dags on n nodes with at least 2^{n-1} parameters to the total number of n-node dags; hence the small bump to the right of the main peak is just an exponentially scaled-down version of this peak for $(n-1)$-node dags.

6 Conclusions

Traditional model-selection criteria used for learning Bayesian networks favor models that fit the data and have low representation complexity (i.e. the number of parameters needed to specify the network). However, those criteria do not take into account the complexity of inference in the resulting network, which is exponential in the network's treewidth. Herein, we demonstrate the importance of using the treewidth as a part of model selection criteria. We show that the cases when two networks have drastically different treewidths, and yet represent close distributions (in terms of KL-divergence) and have similar representation complexities (and thus similar BIC/MDL scores), are sufficiently frequent. We also suggest some efficient probabilistic tests for a quick estimation of the treewidth during learning.

References

[1] E. Amir. Efficient approximation for triangulation of minimum treewidth. In *17th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2001.

[2] A. Beygelzimer and I. Rish. On the importance of using treewidth as a model-selection criterion for learning bayesian networks. In *Submitted to the 7th Valencia International Meeting On Bayesian Statistics*, 2001.

[3] Alina Beygelzimer and Irina Rish. Learning Bayesian networks that are efficient for inference. Technical report, IBM T.J. Watson Research Center, 2001.

[4] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, 1990.

[5] B. D'Ambrosio. Symbolic probabilistic inference in large BN2O networks. In *Proc. Tenth Conf. on Uncertainty in Artificial Intelligence*, pages 128–135, 1994.

[6] R. Dechter. Bucket elimination: A unifying framework for probabilistic inference. In *Proc. Twelfth Conf. on Uncertainty in Artificial Intelligence*, pages 211–219, 1996.

[7] R. Dechter. *Bucket elimination: A unifying framework for probabilistic reasoning*. In M. I. Jordan (Ed.), Learning in Graphical Models, Kluwer Academic Press, 1998.

[8] B.J. Frey and D.J.C. MacKay. A revolution: Belief propagation in graphs with cycles. *Advances in Neural Information Processing Systems*, 10, 1998.

[9] D. Heckerman. A tutorial on learning Bayesian networks, technical report msr-tr-95-06. Technical report, Microsoft Research, 1995.

[10] S.L. Lauritzen and D.J. Spiegelhalter. Local computation with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[11] V. A. Liskovec. The number of maximal vertices of a random acyclic digraph. *Theory Prob. Applic.*, 20:401–409, 1975.

[12] D.J.C. MacKay and R.M. Neal. Near Shannon limit performance of low density parity check codes. *Electronic Letters*, 33:457–458, 1996.

[13] Ann E. Nicholson and Nathalie Jitnah. Using mutual information to determine relevance in bayesian networks. In *Pacific Rim International Conference on Artificial Intelligence*, pages 399–410, 1998.

[14] Irina Rish, Mark Brodie, Haiqin Wang, and Sheng Ma. Intelligent probing: a cost-efficient approach to fault diagnosis in computer networks. *Submitted to IBM Systems Journal*, 2001.

[15] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[16] M. Shwe, B.F. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the Internist-1/QMR knowledge base: I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241 – 255, 1991.

[17] R. Stanley. Acyclic orientations of graphs. *Discrete Mathematics*, 5:171–178, 1973.

[18] N.L. Zhang and D. Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.

Appendix

Proposition 1 [3, 2] *For every $n > 3$, there exist dags $G = G(n)$ and $G' = G'(n)$ such that $\mu(G) > \mu(G')$, whereas $w^*(G) = 2$, while $w^*(G') = \sqrt{n}$.*

Proof: Let G be the complete bipartite graph with the "top" component of size 2 such that the direction of all edges is from this component to the "bottom" component of size $n - 2$. Let G' be the $\sqrt{n} \times \sqrt{n}$ rectangular grid. (Since the proposition is of asymptotic nature, we ignore the issue of divisibility.) Let the direction of all vertical (resp. horizontal) edges in G'

be from left to right (resp. from top to bottom). Clearly, the moral graph of G has treewidth 2 (note that the moral graph is already triangulated), while the treewidth of the moralized G' is $\sqrt{n}$. We only need to show that $\mu(G) \sim \mu(G')$. Indeed, $\mu(G) = 4(n-2)+2$ (4 parameters for each of the $(n-2)$ nodes in the bottom level, and one prior for each top-level node), while $\mu(G') = 4(\sqrt{n}-1)^2 + 4(\sqrt{n}-1) + 1 = 4(n-\sqrt{n})+1$. Thus in fact $\mu(G') < \mu(G)$. Clearly the two quantities can be made equal by adding additional edges to G', which can only increase its treewidth. ∎

Lemma 1 *With probability at least* $1 - exp(-\frac{ld^2}{r^2} + 2\ln r)$, *a random* (l, r, d)-*diffuser has treewidth* $r-1$.

Proof: It is easy to show that the treewidth does not exceed $r-1$, i.e. one less than the size of the clique on the parent (i.e., right) component. Therefore, it sufficies to show that the moralization can indeed induce such a clique with the required probability. Let L and R denote the left and the right components respectively. What is the probability that a pair of parents $p, q \in R$ is not connected in the moralized graph?

$$\begin{aligned} &\mathbf{Pr}[p \text{ and } q \text{ are not connected}] \\ &\quad = \mathbf{Pr}[\forall v \in L: \; p \notin \Gamma(v) \;\vee\; q \notin \Gamma(v)] \\ &\quad = (1 - d^2/r^2)^l \\ &\quad \leq e^{-ld^2/r^2}. \end{aligned}$$

Let P be the expected number of pairs in R that are not induced. Then $\mathbf{E}[P] = \binom{r}{2} e^{-ld^2/r^2}$, and by Markov's inequality, the probability that there are no more than $k\mathbf{E}[P]$ such edges is at least $1 - 1/k$. Thus we want the largest k such that $k\mathbf{E}[P] < 1$, so $k \sim exp(\frac{ld^2}{r^2} - 2\ln r)$, proving the lemma. ∎

Proposition 5 *For any undirected graph* G, $w^*(G, \bar{G}) \geq (2 - \sqrt{3})n$.

Proof: Let $e(G)$ and $r = w^*(G)$ denote the number of edges and the treewidth of a graph $G = (V, E)$, respectively. It is easy to see that $e(G) \leq nr - \frac{1}{2}r(r+1)$. Indeed, consider the ordering of V minimizing the treewidth. Every vertex in V is adjacent to at most r higher ordered vertices except for the last r vertices, which, in the worst case, form a clique. Thus we have $e(G) \leq \sum_{i=1}^{n-r} r + \binom{r}{2} = nr - \frac{1}{2}r(r+1)$, proving the claim. We need to show that

$w^*(\bar{G}) \geq (2-\sqrt{3})n-r$. Assume the contrary; then a similar argument shows that $e(\bar{G}) \leq n((2-\sqrt{3})n-r-1)-\frac{1}{2}((2-\sqrt{3})n-r)((2-\sqrt{3})n-r-1)$. This, however, contradicts $e(\bar{G}) = n(n-1)/2-e(G) \geq n(n-1)/2-nr+\frac{1}{2}r(r+1)$. Indeed, we merely need to verify that this lower bound is greater than the upper bound given by the assumption on $w^*(\bar{G})$, which reduces to showing that $2r(r+1)+c^2+c > 2r(n-c)$ holds for all $0 < r < n-c$, where $c = (\sqrt{3}-1)n$. We have $r \leq n-c-1$; otherwise the proposition trivially holds. Therefore it sufficies to show that $(\sqrt{3}-1)^2n^2 > 2(2-\sqrt{3})^2n^2$, which is certainly true. ∎

Proposition 6 *For any dag G, adding a single directed edge to G cannot increase $w^*(G)$ by more than one.*

Proof: Indeed, if G has treewidth w, then there must exist an ordering of vertices $o = v_1, \ldots, v_n$ such that for each i, the width of v_i is at most w. Assume that we add a directed edge $v_i \rightarrow v_j$ (clearly it should not create a directed cycle in G). The moralization of the augmented graph induces a clique on $\Gamma(v_j) \cup \{v_i\}$. We will construct an ordering of this moral graph with width at most $w+1$. In fact, we only need to move v_i to the end of o. Moving a node downward in the ordering can only decrease its width. The width of v_j and the width of every node in $\Gamma(v_j)$ will increase by one. Clearly, the width of any other vertex with index smaller than i in o, will not change. The only worry are the neighbors of v_i with index greater than i that have now percolated up, and thus will be eliminated before v_i during the inference. However, all that can happen in the worst case, is that every such node will get connected to v_i, increasing its width by one. Thus the width of every node can increase by at most one. ∎