Estimation and Control of the False Discovery Rate of Bayesian Network Skeleton Identification

Angelos P. Armen and Ioannis Tsamardinos
Computer Science Department, University of Crete, and
Institute of Computer Science, Foundation for Research and Technology - Hellas
Heraklion, Crete, GR-714 09, Greece

Technical Report FORTH-ICS / TR-441 – January 2014

Abstract

An important problem in learning Bayesian networks is assessing confidence on the learnt structure. Prior work in constraint-based algorithms focuses on estimating or controlling the False Discovery Rate (FDR) when identifying the skeleton (set of edges without regard of direction) of a network. We present a unified approach to estimation and control of the FDR of Bayesian network skeleton identification and experimentally evaluate the performance of a standard FDR estimator in both tasks over several benchmark networks and sample sizes. We demonstrate that conservative estimation and strong control of FDR are not achieved in some cases due to insufficient sample size and/or unfaithfulness. We show that a permutation-based and a parametric-bootstrap-based FDR estimator achieve more accurate FDR estimation and strong control than the standard estimator. Finally, we present a relaxed definition of false positive that leads to more conservative estimation and control of FDR in relatively small sample sizes.

1 Introduction

Bayesian networks are graphical models that represent probabilistic relationships among variables. A Bayesian network consists of a directed acyclic graph (DAG) called the structure and the conditional probability distribution of each node given values of its parents in the DAG. Structure learning is concerned with learning structure from data, and constraint-based algorithms are a class of algorithms for this purpose. These algorithms work in two phases, skeleton identification and edge orientation. Skeleton identification is concerned with identifying the skeleton, that is, the set of links (edges without regard of direction) of the DAG. Edge orientation is concerned with orienting these links.

An important problem in structure learning is assessing confidence on the learnt structure. Listgarten and Heckerman (2007) propose an estimator of the False Discovery Rate (FDR) of structure learning, that is, the expected proportion of false edges in the output structure. Prior work in constraint-based algorithms uses the FDR of skeleton identification, that is, the expected proportion of false links in the output skeleton. Tsamardinos and Brown (2008) focus on FDR estimation when learning the parents and the children

of a target node (a case of *local* learning), while Li and Wang (2009) focus on FDR control in skeleton identification (global learning). In this work, we adapt the approach of Tsamardinos and Brown (2008) to skeleton identification and unify it with FDR control. We experimentally evaluate the performance of a standard estimator in both estimation and control over several benchmark networks and sample sizes. We demonstrate that conservative estimation and "strong control" of FDR are not achieved in some cases due to insufficient sample size and/or violations of the faithfulness condition. Then, we show that a permutation-based and a parametric-bootstrap-based FDR estimator achieve more accurate FDR estimation and strong control than the standard estimator. Finally, we present a relaxed definition of false positive that leads to more conservative estimation and control of FDR in relatively small sample sizes. Note that in this work we focus on categorical data, although our methods can be applied to continuous data as well.

The rest of this technical report is organized as follows. In Section 2, we review the necessary background on Bayesian network skeleton identification. In Section 3, we present our unified approach to estimation and control of the FDR of skeleton identification and present an experimental evaluation of the performance of a standard FDR estimator in both tasks. In Section 4, we identify and quantify the causes of the failure to achieve conservative estimation and strong control of FDR. In Section 5, we present and evaluate two non-p-value-based FDR estimators. In Section 6, we present an experimental evaluation of a relaxed definition of false positive. Finally, in Section 7, we discuss other approaches to assessing confidence in structure learning.

2 Bayesian network skeleton identification

In the first part of this section, we review basic Bayesian network theory including the concepts of d-separation, Markov equivalence and faithfulness. In the second part, we review Bayesian network skeleton identification and hypothesis tests of conditional independence.

2.1 Bayesian networks

Bayesian networks are graphical models that address the problems of encoding probabilistic relationships among a large set of variables and performing probabilistic inference with those variables (Neapolitan, 2004). A Bayesian network is a pair $(\mathbb{G}, P)$ of a directed acyclic graph (DAG) $\mathbb{G}$ and the joint probability distribution P of the set of nodes (variables) $\mathbf{V}$ of $\mathbb{G}$ that satisfies the *Markov condition*: Every node X in $\mathbf{V}$ is conditionally independent of the set $\mathbf{ND}_X$ of its non-descendants (excluding parents) given the set $\mathbf{PA}_X$ of its parents (Neapolitan, 2004):

$$X \perp\!\!\!\perp \mathbf{ND}_X \mid \mathbf{PA}_X$$

where $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ denotes the conditional independency of $\mathbf{A}$ and $\mathbf{B}$ given $\mathbf{C}$, each of which is a set of variables.[1] Owning to the Markov condition, P is equal to the product of the conditional distributions of all nodes given values of their parents:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \mathbf{pa}_{X_i})$$

[1]For simplicity, we denote both variable X and the set $\{X\}$ with X.

where $x_1, x_2, \ldots, x_n$ are values of all n nodes $X_1, X_2, \ldots, X_n$ in $\mathbf{V}$ and $\mathbf{pa}_X$ are the values of the parents of node X. $\mathbb{G}$ is called the *structure* and the values in the conditional probability distributions in P are called the *parameters*. Figure 1 shows a Bayesian network with four binary nodes.

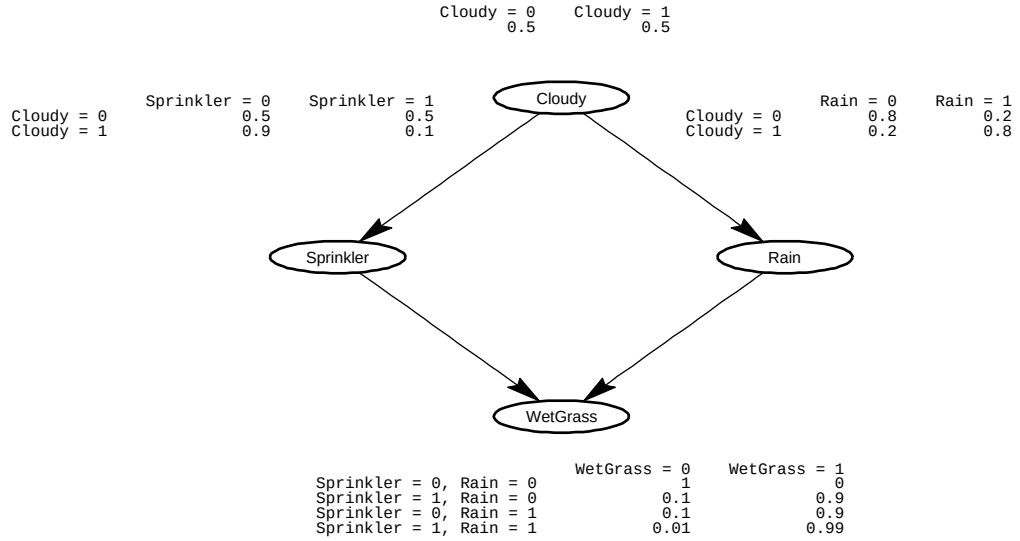


Figure 1: A Bayesian network with four binary nodes (Russell and Norvig, 2009)

.

2.1.1 d-separation

The conditional independencies in the Markov condition imply even more conditional independencies. We say that, based on the Markov condition, DAG $\mathbb{G}$ *entails* a conditional independency if the latter holds for every probability distribution P such that $(\mathbb{G}, P)$ satisfies the Markov condition (Neapolitan, 2004). Which conditional independencies are entailed is dictated by a graphical criterion called *d-separation*.

In a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$, a *chain* is the set of edges connecting the nodes of a set $\{X_1, X_2, \ldots, X_k\}$, where $k > 2$, such that $(X_{i-1}, X_i) \in \mathbf{E}$ or $(X_i, X_{i-1}) \in \mathbf{E}$ for $2 \leq i \leq k$ Neapolitan (2004). For chains $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \rightarrow Y$ and $X \rightarrow Z \leftarrow Y$, we say that edges meet *head-to-tail*, *tail-to-tail* and *head-to-head* at Z, respectively. We use – to denote a *link*, that is, an edge without regard of orientation. We call a chain $X - Z - Y$ such that X and Y are not adjacent an *uncoupled meeting* (Neapolitan, 2004). A chain ρ between distinct nodes X and Y in $\mathbf{V}$ is *blocked* by $\mathbf{A} \subseteq \mathbf{V} \setminus \{X, Y\}$ if one of the following holds (Neapolitan, 2004):

1. There is a node $Z \in \mathbf{A}$ on ρ, and the edges incident to Z on ρ meet head-to-tail at $\mathbf{Z}$.

2. There is a node $Z \in \mathbf{A}$ on ρ, and the edges incident to Z on ρ meet tail-to-tail at $\mathbf{Z}$.

3. There is a node Z, such that Z and all of Z's descendants are not in $\mathbf{A}$, on ρ, and the edges incident to Z on ρ meet head-to-head at $\mathbf{Z}$.

We say that distinct nodes X and Y in $\mathbf{V}$ are *d-separated* by $\mathbf{A}$ in $\mathbb{G}$ if every chain between X and Y is blocked by $\mathbf{A}$ (Neapolitan, 2004). For mutually disjoint subsets $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V}$, we say $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ in $\mathbb{G}$ (and denote it with $\mathbf{A} \perp \mathbf{B} \mid \mathbf{C}$) if for every $X \in \mathbf{A}$ and $X \in \mathbf{B}$, X and Y are d-separated by $\mathbf{C}$ (Neapolitan, 2004).

Theorem 1 *Let $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be mutually disjoint subsets of $\mathbf{V}$. Based on the Markov condition, a DAG $\mathbb{G}$ entails conditional independency $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ only if $\mathbf{A}$ and $\mathbf{B}$ are* d-separated *in $\mathbb{G}$ by $\mathbf{C}$:*

$$\mathbf{A} \perp \mathbf{B} \mid \mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

Proof The proof can be found in Neapolitan (2004, p. 79). ■

The following lemma, which relates d-separation to adjacency, is of great importance in learning the structure of a Bayesian network from data (see Section 2.2):

Lemma 2 *Let $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $X, Y \in \mathbf{V}$. Then X and Y are adjacent in $\mathbb{G}$ (denoted by $Adj(X, Y)$) if and only if they are not d-separated by some set in $\mathbb{G}$:*

$$Adj(X, Y) \iff \nexists\, \mathbf{S}_{XY} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ s.t. } X \perp Y \mid \mathbf{S}_{XY}$$

Proof The proof can be found in Neapolitan (2004, p. 89). ■

2.1.2 Markov equivalence

Many DAGs have the same d-separations and are said to belong to the same *Markov equivalence class*. All DAGs in such a class have the same links or, in other words, the same *skeleton*. Moreover, they have the same set of uncoupled head-to-head meetings. Thus, a Markov equivalence class can be represented with a single graph: a *DAG pattern* is the graph that has the same skeleton as the DAGs in the class and has oriented all and only the edges common to all of the DAGs in the class (Neapolitan, 2004).

2.1.3 Faithfulness

When $(\mathbb{G}, P)$ satisfies the *faithfulness* condition (see Neapolitan, 2004, p. 97, for definition), $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ entails conditional independency $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ for mutually disjoint subsets $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ of $\mathbf{V}$, if and only if $\mathbf{A}$ and $\mathbf{B}$ are d-separated in $\mathbb{G}$ by $\mathbf{C}$:

$$\mathbf{A} \perp \mathbf{B} \mid \mathbf{C} \iff \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

When $(\mathbb{G}, P)$ satisfies the faithfulness condition, we say that P and $\mathbb{G}$ are *faithful* to each other (Neapolitan, 2004). If $(\mathbb{G}, P)$ satisfies the faithfulness condition, then P satisfies the condition with all and only those DAGs in the Markov equivalence class gp of $\mathbb{G}$; we say

that *gp* and P are faithful to each other (Neapolitan, 2004). We say that P *admits a faithful DAG representation* if P is faithful to some DAG (and therefore to some DAG pattern) (Neapolitan, 2004). For a set $\mathbf{V}$ of categorical variables and a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$, if we randomly assign conditional distributions to the variables in $\mathbf{V}$, it is unlikely to come up with a joint distribution that is unfaithful to $\mathbb{G}$ (Meek, 1995).

2.2 Bayesian network skeleton identification

The goal of *structure learning* is to find a DAG $\mathbb{G}$ or DAG pattern *gp* faithful to a distribution P given a sample from P, assuming that P admits a faithful DAG representation. The *constraint-based* approach to structure learning involves two phases: first the d-separations in *gp* are identified and then they are used as *constraints* in generating *gp*. The first phase is called *skeleton identification* because it identifies the skeleton of *gp*; the second phase is called *edge orientation* because it orients the undirected edges of the identified skeleton.

Skeleton identification is based on Lemma 2. For each pair (X, Y) of nodes, a search for a subset of the rest nodes that renders X and Y conditionally independent takes place. Once such a subset (called a *sepset*) is found, (X, Y) is no longer considered; otherwise, the link $X - Y$ is discovered. Modern skeleton identification algorithms exploit the following corollary of Lemma 2 to speed up the search for a sepset:

Corollary 3 *Let $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $X, Y \in \mathbf{V}$. Then if X and Y are d-separated by some set, they are d-separated either by the set of the parents of X or the set of the parents of Y:*

$$\exists\, \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ s.t. } X \perp Y \mid \mathbf{Z} \implies X \perp Y \mid \mathbf{PA}_X \text{ or } X \perp Y \mid \mathbf{PA}_Y$$

Proof The proof can be found in Neapolitan (2004, p. 89). ■

Parents are, of course, unknown, so algorithms based on Corollary 3 ensure that they check the conditional independency of X and Y given all subsets of the set $\mathbf{ADJ}_X$ of neighbors of X and all subsets of the set $\mathbf{ADJ}_Y$ of neighbors of Y by checking subsets of current estimates $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$ of $\mathbf{ADJ}_X$ and $\mathbf{ADJ}_X$, respectively. These algorithms instantiate Algorithm Template 1. Examples of constraint-based structure-learning algorithms whose skeleton identification phase instantiates this template are *PC* (Spirtes et al., 2000) and algorithms belonging to the *Local to Global Learning (LGL)* class of constraint-based algorithms (Aliferis et al., 2010a,b), such as *MMHC* (Tsamardinos et al., 2006). LGL algorithms first learn the neighbors of each node, that is, they learn the links ending to that node; then they complete skeleton identification by combining the links to form the skeleton.

2.2.1 Testing conditional independence

Conditional independencies are identified from the sample from P by performing *hypothesis tests* of conditional independence. To determine whether conditional independency $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ holds, the test $\text{test}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ of the corresponding null hypothesis $\text{H}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ is performed. First, the value $t_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ of a test statistic $\text{T}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ is calculated and then the corresponding p-value $\text{p}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ is computed. If $\text{p}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}} > \alpha$, then $\text{H}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ is accepted. Subsequently,

Algorithm Template 1 Fast skeleton identification. Given a sample from the distribution P of the variables in some set $\mathbf{V}$, the algorithm identifies the skeleton common to every DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ that is faithful to P, assuming that P admits a faithful representation. $\widehat{\mathbf{ADJ}}_X$ is the current estimate of the set of neighbors of node (variable) X. $\mathbf{OPEN}_X$ is the set of nodes not yet considered for inclusion in $\widehat{\mathbf{ADJ}}_X$. $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$ denotes the conditional independency of variables X and Y given set of variables $\mathbf{S}_{XY}$. The termination criteria must be such that, at termination of the algorithm, $\mathbf{OPEN}_X$ is empty for each X. Some instantiations of the template also remove X from $\widehat{\mathbf{ADJ}}_Y$ and $\mathbf{OPEN}_Y$ when Y is removed from $\widehat{\mathbf{ADJ}}_X$, while others do not.

for each $X \in \mathbf{V}$ **do**
  $\widehat{\mathbf{ADJ}}_X \leftarrow \mathbf{U}$ s.t. $\mathbf{U} \subseteq \mathbf{V} \setminus X$
  $\mathbf{OPEN}_X \leftarrow \mathbf{V} \setminus (\widehat{\mathbf{ADJ}}_X \cup X)$
end for
repeat
  for each $X \in \mathbf{V}$ **do**
    repeat
      for each $Y \in \widehat{\mathbf{ADJ}}_X$ **do**
        search for $\mathbf{S}_{XY} \subseteq \widehat{\mathbf{ADJ}}_X \setminus Y$ s.t. $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$
        if such a set is found **then**
          $\widehat{\mathbf{ADJ}}_X \leftarrow \widehat{\mathbf{ADJ}}_X \setminus Y$
        end if
      end for
      $\widehat{\mathbf{ADJ}}_X \leftarrow \widehat{\mathbf{ADJ}}_X \cup \mathbf{W}$ s.t. $\mathbf{W} \subseteq \mathbf{OPEN}_X$
      $\mathbf{OPEN}_X \leftarrow \mathbf{OPEN}_X \setminus \mathbf{W}$
    until a termination criterion is met
  end for
until a termination criterion is met

(X, Y) is no longer considered and $X - Y$ is not discovered. If $\mathrm{p}_{X \perp\!\!\!\perp Y|\mathbf{Z}} \leq \alpha$, $\mathrm{H}_{X \perp\!\!\!\perp Y|\mathbf{Z}}$ is rejected and the search for a sepset continues.

Typically, the G test is employed when all variables in $\mathbf{V}$ are categorical (Tsamardinos et al., 2006; Aliferis et al., 2010a). The G test uses the G statistic. When $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ holds, the statistic asymptotically follows the χ^2 distribution with *df* degrees of freedom. Assuming no *structural zeros* (zeros in the contingency table where observations are impossible), *df* is given by the following equation:

$$df = (|\mathbf{D}_X| - 1)(|\mathbf{D}_Y| - 1) \prod_{Z \in \mathbf{Z}} |\mathbf{D}_Z|$$

where $\mathbf{D}_X$ is the domain of X. If every cell of the row (column) where $X = x$ ($Y = y$) in the conditional contingency table where $\mathbf{Z} = \mathbf{z}$ contains a structural zero (that is, $X \neq x$ ($Y \neq y$) when $\mathbf{Z} = \mathbf{z}$), the degrees of freedom corresponding to the conditional contingency table must be reduced by one. In their implementation of MMHC, Tsamardinos et al. (2006) reduce the degrees of freedom corresponding to each conditional contingency table by one for each row or column of the table whose each cell contains zero (whether a zero is

structural or not, is, of course, unknown to the test):

$$df = \sum_{\mathbf{z} \in \mathbf{D_Z}} \max\{|\mathbf{D}_X| - 1 - \sum_{x \in \mathbf{D}_X} [1 - I(N_{x\mathbf{z}})], 0\} \cdot \max\{|\mathbf{D}_Y| - 1 - \sum_{y \in \mathbf{D}_Y} [1 - I(N_{y\mathbf{z}})], 0\}$$

where $N_{x\mathbf{z}}$ ($N_{y\mathbf{z}}$) is the number of observations with $X = x$ ($Y = y$) and $\mathbf{Z} = \mathbf{z}$ in the sample, and function $I(\cdot)$ is 1 when its input is positive and 0 otherwise. We call this method the *degrees of freedom adjustment heuristic*. It is easy to see that

$$df = 0 \iff \forall \mathbf{z} \in \mathbf{D_Z} \text{ s.t. } N_{\mathbf{z}} > 0 : \{\forall x \in \mathbf{D}_X : \sum_{y \in \mathbf{D}_Y} I(N_{xy\mathbf{z}}) = 1\} \cup \{\forall y \in \mathbf{D}_Y : \sum_{x \in \mathbf{D}_X} I(N_{xy\mathbf{z}}) = 1\}$$

where $N_{xy\mathbf{z}}$ is the number of observations with $X = x$, $Y = y$ and $\mathbf{Z} = \mathbf{z}$, and $N_{\mathbf{z}}$ is the number of observations with $\mathbf{Z} = \mathbf{z}$ in the sample. That is, the degrees of freedom are zero if and only if for each instance of $\mathbf{Z}$ with positive count, either X or Y take only one of their values. This means that either X or Y are exactly determined by $\mathbf{Z}$ (*determinism*) or it seems so due to insufficient sample size (*close-to-determinism*). In their implementation of MMHC, Tsamardinos et al. (2006) ignore a test with $df = 0$. On the contrary, we set the p-value of such a test to one. We call this *determinism detection*. We have found that determinism detection results in greatly reduced execution times and more accurate FDR estimation and strong control in some cases (see Armen, 2011, for results without determinism detection).

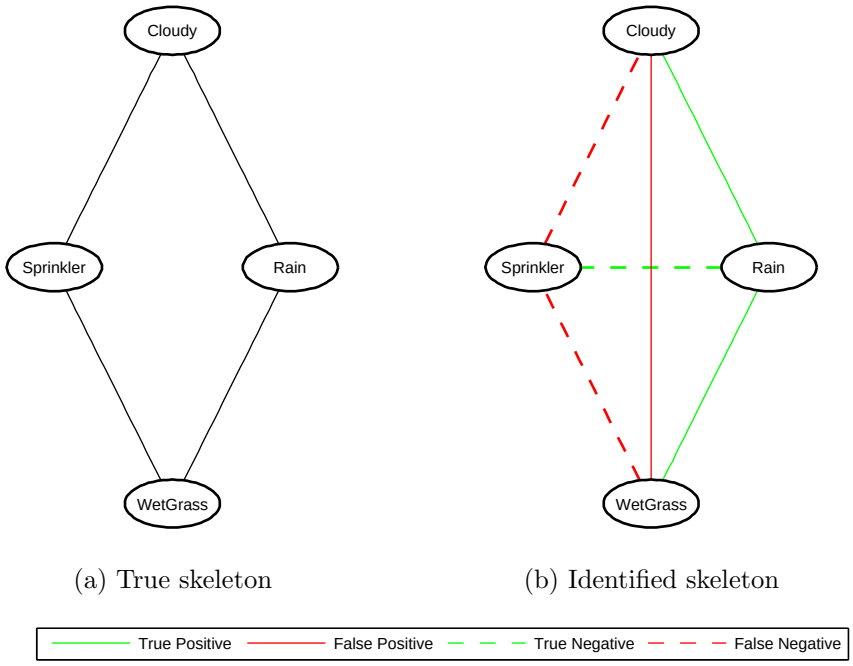
A test is typically performed only if it is reliable according to a *reliability criterion*. A reliable test both (a) meets the distributional assumptions of the statistic used and (b) has sufficient power (Fast, 2010). When a test is unreliable, a *default decision* is made. In MMHC, the default decision is independence when the conditioning set is empty and dependence otherwise (see Tsamardinos et al., 2006, Section 1.1, for a detailed justification).

For categorical variables, Fienberg (1977) recommends that there at least five observations per cell of the contingency table, on average, for the test to be reliable. We refer to the lower limit on the average number of observations per cell as the *heuristic power size* (denoted by *h-ps*) and to the corresponding reliability criterion as the *heuristic power rule*.

A *type I error* or *false positive* occurs when a test concludes dependence while independence holds. On the other hand, a *type II error* or *false negative* occurs when a test concludes independence while dependence holds. In this report, we use the terms false positive and false negative to refer to errors in the output skeleton and type I and type II error to refer to errors made by the tests. Since a link is not discovered once a sepset is found, a single type II error results in a false negative. In the rest of the report, we refer to the set that caused a pair not to no longer be considered as *the* sepset (regardless of whether it actually d-separates the nodes). Likelihood-ratio tests such as the G test have an asymptotic power of one, that is, they tend to not make type II errors as the sample size approaches infinity (see Li and Wang, 2009, Appendix B).

For a conditional-independence-test-based algorithm instantiating Algorithm Template 1, if P admits a faithful DAG representation and all tests considered by the algorithm are reliable according to the employed reliability criterion and yield the correct result, then the algorithm discovers all and only links in *gp* (see Armen, 2011, Theorem 2.10, p. 15).

Example 1 *We applied skeleton identification on a sample from the Bayesian network in Figure 1. The skeleton of the network is shown in Figure 2a, while the identified skeleton is shown in Figure 2b. 1/2 = 50% of the true links are correctly identified, while 1/3 ≈ 33% of the identified links are false.*



(a) True skeleton (b) Identified skeleton

Figure 2: Example of skeleton identification.

3 Estimation and control of the False Discovery Rate of skeleton identification

In addition to skeleton identification, we would like to assess confidence on the identified skeleton. To this end, we view skeleton identification as multiple hypothesis testing, each null hypothesis being the absence of a link. In multiple hypothesis testing, a reasonable error rate to use is the False Discovery Rate (FDR). We present a unified approach to estimation and control of the FDR of skeleton identification and an experimental evaluation of the performance of a standard FDR estimator in both tasks. Finally, we compare our approach to an existing approach to FDR control.

3.1 Multiple hypothesis testing and the False Discovery Rate

Multiple hypothesis testing is testing several hypotheses simultaneously (Shaffer, 1995). Suppose that m hypotheses are tested with corresponding p-values $p_1, p_2, \ldots, p_m$. Typically, a p-value threshold t is chosen and hypotheses with corresponding p-value $\leq t$ are rejected; t

can be either fixed beforehand or selected by a data-dependent thresholding procedure that controls some error rate, while maximizing power (Storey, 2010).

In single hypothesis testing, the False Positive Rate (FPR) is controlled at some level α, while maximizing power, by rejecting hypotheses with corresponding p-value $\leq \alpha$. In multiple hypothesis testing, a reasonable error rate to use is the *False Discovery Rate (FDR)* (Benjamini and Hochberg, 1995). FDR is loosely defined as the expected proportion of false positives among the rejected hypotheses ("discoveries") and it is useful when one is interested having mostly true positives among our discoveries. Precisely,

$$\mathrm{FDR} \triangleq \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R} \vee 1}\right] = \mathrm{E}\left[\frac{\mathrm{V}}{\mathrm{R}} \;\middle|\; \mathrm{R} > 0\right] \Pr(\mathrm{R} > 0)$$

where V is the number of rejected true null hypotheses, R is the number of rejections and $\mathrm{R} \vee 1$ corresponds to setting V / R to 0 when $\mathrm{R} = 0$. We refer to V / R as the *realized* FDR. There are two approaches to using FDR, namely *control* and *estimation*.

3.1.1 Control of the False Discovery Rate

The FDR control approach (Benjamini and Hochberg, 1995) is to set an FDR threshold q and find a thresholding procedure that achieves *strong control*[2] of FDR below q: $\mathrm{FDR} \leq q$. Procedure 2, referred to as the BH procedure, is proven to achieve strong control, assuming independent p-values (Benjamini and Hochberg, 1995) or what is called *positive regression dependence* of the p-values on each of the p-values corresponding to the true null hypotheses (Benjamini and Yekutieli, 2001):

Procedure 2 BH procedure. Controls, at level q, the False Discovery Rate in testing m hypotheses, when the corresponding p-values are independent.

Let $\mathrm{p}_{(1)} \leq \mathrm{p}_{(2)} \cdots \leq \mathrm{p}_{(m)}$ be the ordered p-values
Let $k = \arg\max_i \{\mathrm{p}_{(i)} \leq \frac{i}{m} q\}$
Reject hypotheses corresponding to $\mathrm{p}_{(i)} : i = 1 \ldots k$ if k exists, otherwise none

3.1.2 Estimation of the False Discovery Rate

The FDR estimation approach (Storey, 2002) is to set a p-value threshold t and estimate $\mathrm{FDR}(t)$, the FDR resulting from rejecting all hypotheses with corresponding p-value $\leq t$, in a conservative manner:

$$\mathrm{E}[\widehat{\mathrm{FDR}}(t)] \geq \mathrm{FDR}(t)$$

where $\widehat{\mathrm{FDR}}(t)$ is the estimator of $\mathrm{FDR}(t)$. We denote $\mathrm{E}[\widehat{\mathrm{FDR}}(t)] - \mathrm{FDR}(t)$, the *bias* of $\widehat{\mathrm{FDR}}(t)$, with $\mathrm{bias}(\widehat{\mathrm{FDR}}, t)$. Storey (2002) introduces the following estimator, which is

[2] In the statistical literature, the term "strong control" refers to the case that an error rate is controlled under any configuration of true and false null hypotheses, while the term "weak control" refers to the case that an error rate is controlled when all null hypothesis are true.

proven to be conservative when the p-values are independent:[3]

$$\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t) \triangleq \frac{m \cdot t}{\mathrm{R}(t) \vee 1}$$

FDR estimators can be used to define valid FDR controlling procedures (Storey et al., 2004). Rejecting hypotheses with corresponding p-value $\leq \max\{t \text{ s.t. } \widehat{\mathrm{FDR}}_{\mathrm{BH}}(t) \leq q\}$ is the same as applying the BH procedure (Storey, 2002). We refer to rejecting hypotheses with corresponding p-value less than or equal to

$$\mathrm{t_q}(\widehat{\mathrm{FDR}}) \triangleq \max\{t \text{ s.t. } \widehat{\mathrm{FDR}}(t) \leq q\}$$

as the *FDR controlling procedure* with FDR estimator $\widehat{\mathrm{FDR}}$ and FDR threshold q. It is not hard to see that the procedure strongly controls FDR at level q if and only if $\widehat{\mathrm{FDR}}[\mathrm{t_q}(\widehat{\mathrm{FDR}}]$ is conservative. We refer to $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}]$ as the bias of the procedure and denote it by $\mathrm{procbias}(\widehat{\mathrm{FDR}}, q)$. We denote the power resulting from rejecting all hypotheses with corresponding p-value $\leq t$ by $\mathrm{power}(t)$ and the power $\mathrm{power}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ of the procedure by $\mathrm{procpower}(\widehat{\mathrm{FDR}}, q)$.

Benjamini and Yekutieli (2001) prove that the BH procedure with $q/\left(\sum_{i=1}^{m} \frac{1}{i}\right)$ in place of q achieves strong control for any form of dependence of the p-values. We refer to this modified procedure as the *BY procedure*. The BY procedure is the FDR controlling procedure with the following estimator:

$$\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t) \triangleq \frac{m \cdot t \cdot \left(\sum_{i=1}^{m} \frac{1}{i}\right)}{\mathrm{R}(t) \vee 1}$$

Therefore, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is conservative for any form of dependence.

3.2 Using the False Discovery Rate in skeleton identification

In order to use FDR, skeleton identification is viewed as multiple hypothesis testing, each null hypothesis being the absence of a link. It is not hard to see that the p-value $\mathrm{p}_{\neg \mathrm{Adj}(X,Y)}$ corresponding to the hypothesis $\mathrm{H}_{\neg \mathrm{Adj}(X,Y)}$ of absence of link $X - Y$ is the probability that, for every performed test $\mathrm{test}_{X \perp\!\!\!\perp Y|\mathbf{B}}$ of the hypothesis $\mathrm{H}_{X \perp\!\!\!\perp Y|\mathbf{B}}$, the corresponding statistic $\mathrm{T}_{X \perp\!\!\!\perp Y|\mathbf{B}}$ is greater than or equal to the observed value $\mathrm{t}_{X \perp\!\!\!\perp Y|\mathbf{B}}$, when $\mathrm{H}_{\neg \mathrm{Adj}(X,Y)}$ is true:

$$\mathrm{p}_{\neg \mathrm{Adj}(X,Y)} = \Pr\left(\bigcap_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} \mathrm{T}_{X \perp\!\!\!\perp Y|\mathbf{B}} \geq \mathrm{t}_{X \perp\!\!\!\perp Y|\mathbf{B}} \,\middle|\, \neg \mathrm{Adj}(X, Y) \right)$$

where $\underline{\mathbf{B}}_{XY}$ is the set of subsets $\mathbf{B}$ of $\mathbf{V} \backslash \{X, Y\}$ such that $\mathrm{test}_{X \perp\!\!\!\perp Y|\mathbf{B}}$ is performed. Although unavailable, $\mathrm{p}_{\neg \mathrm{Adj}(X,Y)}$ can be upper-bounded thanks to the following theorem, whose proof can be found in Appendix A:

[3] The estimators of Storey (2002) also include a $\hat{\pi}_0(\lambda)$ term, an estimator of the proportion of true null hypotheses. Because $\hat{\pi}_0(\lambda)$ is not applicable in this context, we use $\hat{\pi}_0(\lambda) = 1$ instead.

Theorem 4 *Let $(\mathbb{G}, P)$ be a Bayesian network of the variables in some set $\mathbf{V}$ and X and Y be distinct nodes in $\mathbf{V}$. Suppose that an algorithm instantiating Algorithm Template 1 is applied on a sample from P. If*

1. *$\mathbb{G}$ and P are faithful to each other,*
2. *all tests considered by the algorithm are reliable, and*
3. *performed tests never make a type II error,*

then $p_{\neg Adj(X,Y)}$ is upper-bounded by the maximal among $p_{X \perp\!\!\!\perp Y|\mathbf{B}}$ for $\mathbf{B} \in \underline{\mathbf{B}}_{XY}$:

$$p_{\neg Adj(X,Y)} \leq \max_{\mathbf{Z} \in \underline{\mathbf{B}}_{XY}} p_{X \perp\!\!\!\perp Y|\mathbf{B}}$$

Skeleton identification can be viewed as a multiple testing procedure that rejects link absence hypotheses with corresponding p-value $\leq \alpha$: $\mathrm{H}_{\neg \mathrm{Adj}(X,Y)}$ is accepted once some $\mathrm{p}_{X \perp\!\!\!\perp Y|\mathbf{B}} \geq \alpha$, or equivalently, once the up-to-date upper bound $\max_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} \mathrm{p}_{X \perp\!\!\!\perp Y|\mathbf{B}}$ of $\mathrm{p}_{\neg \mathrm{Adj}(X,Y)}$ exceeds α. Thus, under the assumptions of Theorem 4 plus the assumption that the link absence p-values are independent, skeleton identification controls FPR at level α.

3.2.1 Existing approaches to using FDR in structure learning

Listgarten and Heckerman (2007) present a Bayesian and a frequentist approach to estimating the number of false *edges* (in contrast to links) in a learned structure. We discuss their former approach in Section 7.1. Their latter one essentially uses a simulation-based FDR estimator (Storey and Tibshirani, 2001) to estimate FDR. We discuss this approach in Section 5.1, where we show that their simulation method is incorrect and we propose an alternative one.

Tsamardinos and Brown (2008) focus on FDR estimation when learning the neighbors of a target node X. The learning task is viewed as multiple hypothesis testing, each null hypothesis being the absence of a node from $\mathbf{ADJ}_X$. Since hypotheses with corresponding p-value $\leq \alpha$ are rejected, the FDR of learning neighbors is $\mathrm{FDR}(\alpha)$. Neighbors of different nodes are learned from samples of size $n \in \{1000, 5000, 10000\}$ from benchmark networks. Then, $\mathrm{FDR}(\alpha)$ is estimated by $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$. For $n \in \{5000, 10000\}$, $\mathrm{FDR}(\alpha)$ is conservatively estimated, in general. For $n = 1000$, however, this is not the case; the situation is improved using a relaxed definition of false positive (see Section 6).

Li and Wang (2009) focus on FDR control in skeleton identification (global learning). The skeleton identification phase of PC, called *PC-skeleton*, is modified to obtain *PC_{FDR}-skeleton*. Instead of thresholding the up-to-date upper bounds on the link absence p-values at some FPR threshold α, $\mathrm{PC}_{\mathrm{FDR}}$-skeleton applies an FDR controlling procedure with some FDR threshold q. It is not hard to see that Theorem 4 holds for $\mathrm{PC}_{\mathrm{FDR}}$-skeleton too, when the last assumption is replaced by the assumption that no application of the FDR controlling procedure yields a false negative. Under the additional assumption that the procedure supports the dependence between the link absence p-values, $\mathrm{PC}_{\mathrm{FDR}}$-skeleton controls FDR at level q. $\mathrm{PC}_{\mathrm{FDR}}$-skeleton employing the BH procedure with $q = 0.05$ is applied to samples of size $n = 500$ from randomly generated networks with varying number of nodes and characteristics. $\mathrm{PC}_{\mathrm{FDR}}$-skeleton achieves FDR noticeably lower than q, while a

heuristic modification of the algorithm achieves FDR around q. However, this modification is not theoretically proven to strongly control FDR.

3.2.2 A unified approach to estimation and control

In this work, we apply the FDR estimation approach of Tsamardinos and Brown (2008) to skeleton identification, extend it to any p-value threshold $t \leq \alpha$, and unify it with FDR control at any FDR threshold q. Our unified approach to estimation and control of FDR of skeleton identification is simply as follows: first perform skeleton identification with significance level α; then either estimate FDR at a p-value threshold $t \leq \alpha$ or apply an FDR controlling procedure with FDR threshold q.

There is no point in estimating FDR(t) for $t > \alpha$, since skeleton identification, viewed as a multiple testing procedure, has already accepted link absence hypotheses with corresponding p-value $> \alpha$. For the same reason, when applying the FDR controlling procedure with $\widehat{\text{FDR}}$ and q, there is no point in retaining links with corresponding p-value in the interval $(\alpha \ \text{t}_\text{q}(\widehat{\text{FDR}})]$ if $\text{t}_\text{q}(\widehat{\text{FDR}}) > \alpha$.

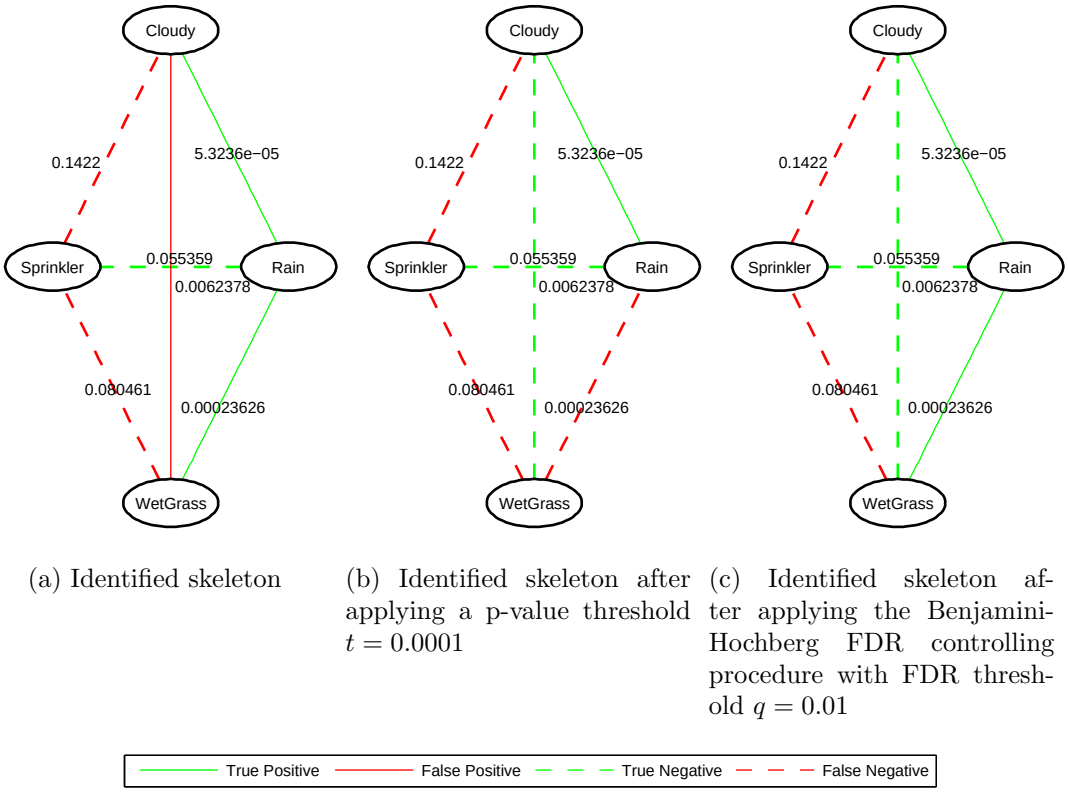
Example 2 *We applied skeleton identification with $\alpha = 0.05$ on a sample of the Bayesian network of Figure 1. The identified skeleton is shown in 3a. The realized FDR for $t = \alpha$ is $1/3$. Table 1 lists the values of FDR-related quantities corresponding to each pair of nodes in the identified skeleton, in ascending order of p-values. By applying a p-value threshold $t = 0.0001$ we get the skeleton of Figure 3b and zero realized FDR. We estimate $FDR(0.0001)$ by $\widehat{FDR}_{BH}(0.0001) = 6 \cdot 0.0001/1 = 0.0006$. By applying the BH procedure with $q = 0.01$ we get the skeleton of Figure 3c and zero realized FDR.*

Pair of nodes	p-value ↓	$\widehat{\text{FDR}}_{\text{BH}}$	realized FDR
$(Rain, Cloudy)$	5.324e-05	0.00031941	0
$(WetGrass, Rain)$	0.0002363	0.00070878	0
$(WetGrass, Cloudy)$	0.006238	0.012476	0.33333
$(Rain, Sprinkler)$	0.05536		
$(WetGrass, Sprinkler)$	0.08046		
$(Sprinkler, Cloudy)$	0.1422		

Table 1: Values of False Discovery Rate (FDR) - related quantities corresponding to each pair of nodes (link absence hypothesis) in the identified skeleton of Example 2, in ascending order of p-values. Only values corresponding to p-values $\leq \alpha = 0.05$ are listed, where α is the significance level of the underlying hypothesis tests. $\widehat{\text{FDR}}$ is the value of the FDR estimator of Storey (2002) at the corresponding p-value. The realized FDR corresponding to a p-value is the realized FDR when rejecting link absence hypotheses with less or equal corresponding p-value.

3.3 Experiment

In order to find out what the FDR of skeleton identification is and whether conservative estimation and strong control of FDR are achieved in practice, we applied skeleton iden-



(a) Identified skeleton

(b) Identified skeleton after applying a p-value threshold $t = 0.0001$

(c) Identified skeleton after the Benjamini-Hochberg FDR controlling procedure with FDR threshold $q = 0.01$

Figure 3: Example of the unified approach to estimation and control of the False Discovery Rate (FDR) in skeleton identification. The number by each line is the p-value corresponding to the pair of nodes the line connects.

tification to samples of several sizes from several benchmark networks. For each network and sample size, we evaluated $\widehat{\text{FDR}}_{\text{BY}}$[4] in both tasks of estimation and control over several p-value and FDR thresholds, respectively.

First, we obtained the *Alarm*, *Andes*, *Barley*, *Hailfinder*, *Hepar II*, *Insurance*, *Mildew*, *Power Plant*, *Water* and *Win95pts* Bayesian networks with categorical variables from online repositories (see Appendix D) and generated 100 random samples of size 10000 from each network. Then, we applied the skeleton identification phase of MMHC, which phase we call *MMPC-skeleton*[5], to each sample from each network, each time using the first $n \in \{100, 1000, 10000\}$ observations. We used the G test with $\alpha = 0.05$, the heuristic power rule with *h-ps* $= 5$, and the degrees of freedom adjustment heuristic with determinism detection. All expectations reported in the rest of the report were approximated by the respective means over the 100 samples.

[4]We did not use $\widehat{\text{FDR}}_{\text{BH}}$ because the link absence p-values are, of course, not independent, and proving whether they have positive regression dependence on the null link absence p-values is beyond the scope of this work.

[5]We call it this way because it is based on the MMPC local learning algorithm (Tsamardinos et al., 2006).

For each network and sample size, we approximated $\mathrm{FDR}(t)$[6] for 50 logarithmically spaced in $[10^{-8}, 10^{-1}]$ values of t (Figure 6). $\mathrm{FDR}(t)$ varies greatly among networks and increases as t increases or n decreases, in general.

In order to evaluate $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ in FDR estimation, we approximated $\mathrm{bias}(\widehat{\mathrm{FDR}}_{\mathrm{BY}}, t)$ for each network and sample size (Figure 4). $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ gets more conservative as t increases, in general. For large enough sample size, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is conservative (or almost conservative) on all networks except Barley and Mildew. For large t, however, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is *overly* conservative in the aforementioned cases.

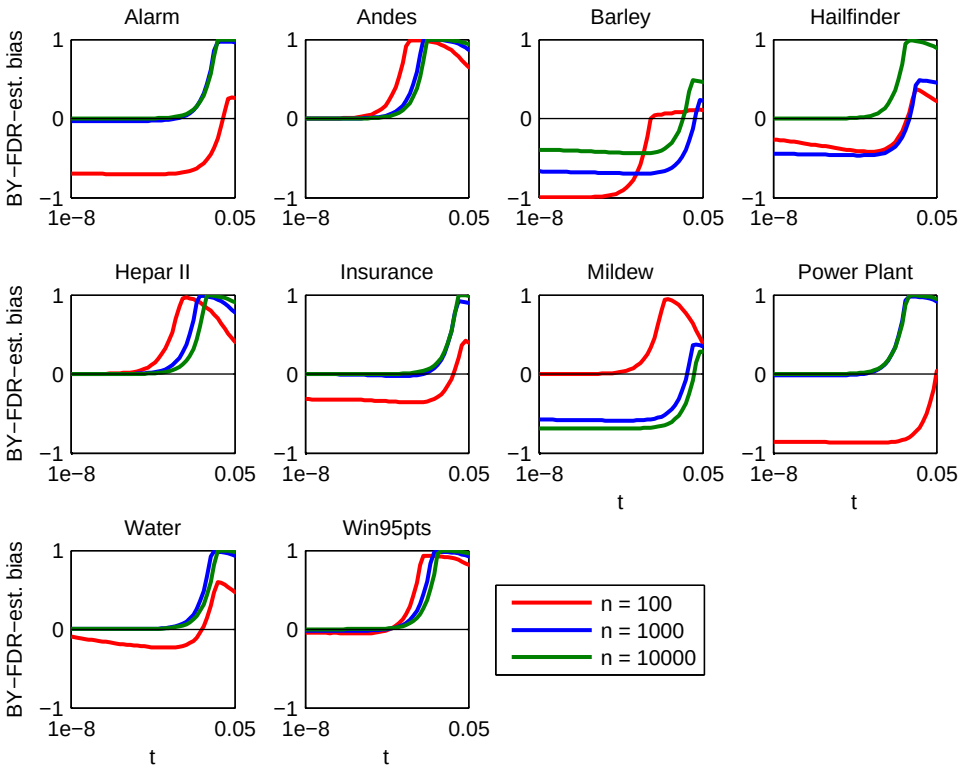


Figure 4: $\mathrm{bias}(\widehat{\mathrm{FDR}}_{\mathrm{BY}}, t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ of Benjamini and Yekutieli (2001) at each p-value threshold t, for each network and sample size n (see text for definition). X-axes are in logarithmic-10 scale. Bias increases as t increases, in general. Bias is nonnegative or slightly negative for $n \in \{1000, 10000\}$ on Alarm, for all n on Andes, for $n = 10000$ on Hailfinder, for all n on Hepar II, for $n \in \{1000, 10000\}$ on Insurance, for $n = 100$ on Mildew, for $n \in \{1000, 10000\}$ on Power Plant and Water, and for all n on Win95pts. However, bias at large t is very large in the aforementioned cases.

[6] We also approximated pFDR, but found that $\mathrm{FDR}(t) = \mathrm{pFDR}(t)$ for all networks and samples sizes so we did not use it any further.

In order to evaluate $\widehat{\text{FDR}}_{\text{BY}}$ in FDR control, we applied the BY procedure with 50 logarithmically spaced in $[10^{-3}, 10^{-1}]$ values of q to the skeletons identified from each sample of each network and approximated procbias$(\widehat{\text{FDR}}_{\text{BY}}, q)$. The BY procedure achieves (or almost achieves) tight strong control of FDR for large enough sample size on all networks except Barley and Mildew.

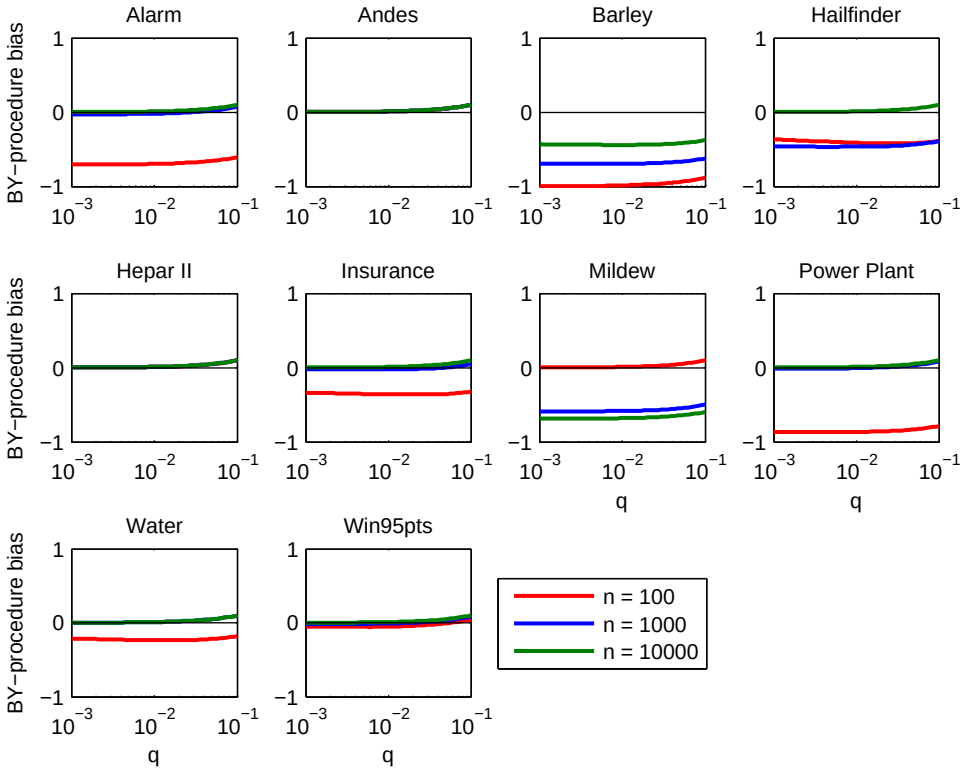


Figure 5: procbias$(\widehat{\text{FDR}}_{\text{BY}}, q)$: bias of the False Discovery Rate (FDR) controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold q for each network and sample size n (see text for definition). X-axes are in logarithmic-10 scale. Bias is close to zero for $n \in \{1000, 10000\}$ on Alarm, for all n on Andes, for $n = 10000$ on Hailfinder, for all n on Hepar II, for $n \in \{1000, 10000\}$ on Insurance, for $n = 100$ on Mildew, for $n \in \{1000, 10000\}$ on Power Plant and Water, and for all n on Win95pts.

In summary, FDR varies greatly among networks; estimation is conservative and tight strong control is achieved on some networks for large enough sample size; in these cases, however, estimation at large p-value thresholds is overly conservative. In section 4, we identify and quantify the causes of the failure to achieve conservative estimation and strong control.

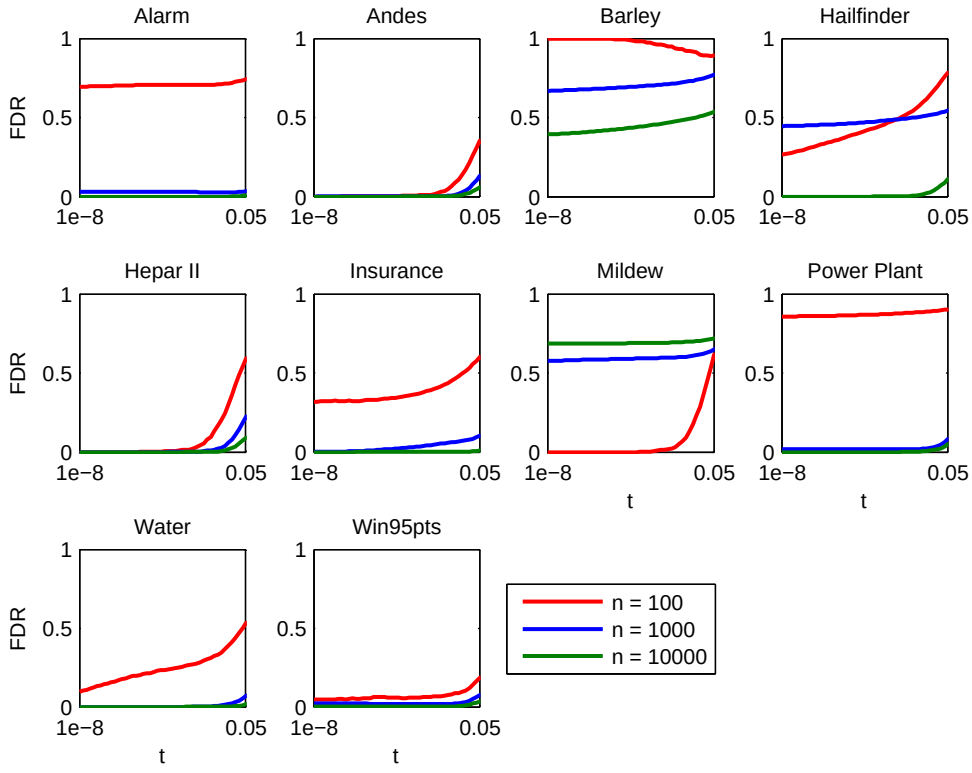


Figure 6: FDR(t): False Discovery Rate (FDR) at each p-value threshold t for each network and sample size n. X-axes are in logarithmic-10 scale. FDR varies greatly among networks. FDR increases as t increases in all cases except for $n = 100$ and large t on Barley. FDR also increases as n decreases on all networks except Hailfinder and Mildew.

3.4 Comparison of approaches to FDR control

It is easy to see that both our approach and that of Li and Wang (2009) are special cases of the following more general approach to FDR estimation and control in skeleton identification: first, perform skeleton identification while applying a thresholding rule to the up-to-date upper bounds on the link absence p-values every time one is updated; then, estimate FDR at a p-value threshold $t \leq \mathrm{t}(\widehat{\mathrm{FDR}})$, where $\mathrm{t}(\widehat{\mathrm{FDR}})$ is the threshold selected in the final application of the rule, or control FDR below an FDR threshold q. During skeleton identification, our approach involves thresholding at a fixed threshold α, while the approach of Li and Wang (2009) involves the application of an FDR controlling procedure.

Although the approach of Li and Wang (2009) is an FDR control approach, it still allows for FDR estimation, as discussed above. In the task of FDR control, their approach does not require the specification of the FPR threshold α nor an FDR control postprocessing step, in contrast to ours. On the other hand, their approach requires the specification of the FDR threshold q in advance; however, an FDR controlling procedure with another threshold $q' \leq q$ may be applied as a postprocessing step.

We experimentally compared the two approaches to FDR control. We applied PC_{FDR}-skeleton and our approach with PC-skeleton ($\alpha = 0.05$), both with FDR threshold 0.05, to the samples from each network, each time using the first $n \in \{100, 1000, 10000\}$ observations. We used our own implementation of $\text{PC}_{(\text{FDR})}$-skeleton, because we found it to be faster than that of Li and Wang (2009). Moreover, we used the default decision of MMHC and not that of PC. Finally, we used the G test, in contrast to Li and Wang (2009), who use the Cochran-Mantel-Haenszel (CMH) test. Fast (2010) experimentally demonstrates that, compared to the G test, the CMH test decreases power and increases FPR.

We approximated $\text{procbias}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$, $\text{procpower}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$, and the expected skeleton identification time (Tables B.1 – B.3, Appendix B) for each network, sample size, and FDR control approach. For PC_{FDR}-skeleton, $\text{procbias}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$ and $\text{procpower}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$ correspond to the final application of the procedure. To aid our comparison, we define the *undercontrol*

$$\text{uc}(\widehat{\text{FDR}}, q) \triangleq \max\{\text{FDR}[\text{t}_{\text{q}}(\widehat{\text{FDR}})] - q, 0\}$$

and the *overcontrol*

$$\text{oc}(\widehat{\text{FDR}}, q) \triangleq \max\{q - \text{FDR}[\text{t}_{\text{q}}(\widehat{\text{FDR}})], 0\}$$

of the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ and FDR threshold q. Strong control is achieved in a collection of cases if and only if the mean undercontrol across the cases is zero. The less the mean overcontrol across the cases, the more tight the strong control. Table 2 summarizes the results of the experimental comparison, which are similar for both FDR control approaches.

	1-stage	2-stage
$\text{uc}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$	0.2774	0.272
$\text{oc}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$	0.01426	0.01849
$\text{procpower}(\widehat{\text{FDR}}_{\text{BY}}, 0.05)$	0.5802	0.5656
E[time] (s)	33.28	31.31

Table 2: Summary of the results of the experimental comparison of False Discovery Rate (FDR) control approaches. *1-stage* and *2-stage* refer to the single-stage approach of Li and Wang (2009) and our two-stage approach, respectively, using the PC-skeleton algorithm. $\text{uc}(\widehat{\text{FDR}}_{\text{BH}}, 0.05)$ and $\text{oc}(\widehat{\text{FDR}}_{\text{BH}}, 0.05)$ denote undercontrol and overcontrol, respectively, of the FDR by the FDR controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold 0.05; $\text{procpower}(\widehat{\text{FDR}}_{\text{BH}}, 0.05)$ denotes the power of the procedure; E[time] denotes the expected execution time (see text for definitions). Each cell of the table contains the mean value of the corresponding quantity across various networks and sample sizes. All values are similar for both FDR control approaches.

4 Causes of failure to achieve conservative estimation and strong control

The experimental results in Section 3.3 imply that some of the assumptions of Theorem 4 are violated in some cases. In this section, we analyze the causes of these violations. To facilitate our analysis, we present the theorem below, which establishes a single sufficient criterion for conditional-independence-test-based skeleton identification algorithms to upper-bound $\mathrm{p}_{\neg \mathrm{Adj}(X,Y)}$. The proof can be found in Appendix A.

Theorem 5 *Let $(\mathbb{G}, P)$ be a Bayesian network of the variables in some set $\mathbf{V}$ and $X, Y \in \mathbf{V}$. Suppose that an algorithm instantiating Algorithm Template 1 is applied on a sample from P. If*

1. *there is a set $\mathbf{F} \in \underline{\mathbf{B}}_{XY}$ such that $H_{X \perp Y|\mathbf{F}}$ would be true if $H_{\neg Adj(X,Y)}$ was true, and*
2. *$X \perp\!\!\!\perp Y \mid \mathbf{F} \implies X \perp Y \mid \mathbf{F}$ holds,*

then $p_{\neg Adj(X,Y)}$ is upper-bounded by the maximal among $p_{X \perp\!\!\!\perp Y|\mathbf{B}}$ for $\mathbf{B} \in \underline{\mathbf{B}}_{XY}$:

$$p_{\neg Adj(X,Y)} \leq \max_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} p_{X \perp\!\!\!\perp Y|\mathbf{B}}$$

In fact, it is not hard to see that only *false positive* p-values need to be upper-bounded for $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ to be conservative. The following corollary of Theorem 5 establishes a single sufficient criterion for such a p-value to be upper-bounded:

Corollary 6 *Let $(\mathbb{G}, P)$ be a Bayesian network of the variables in some set $\mathbf{V}$ and X and Y be distinct variables in $\mathbf{V}$ that are not adjacent in $\mathbb{G}$. Suppose that an algorithm instantiating Algorithm Template 1 is applied on a sample from P. If there is a set $\mathbf{F} \in \underline{\mathbf{B}}_{XY}$ such that $X \perp Y \mid \mathbf{F}$, then $p_{\neg Adj(X,Y)}$ is upper-bounded by the maximal among $p_{X \perp\!\!\!\perp Y|\mathbf{B}}$ for $\mathbf{B} \in \underline{\mathbf{B}}_{XY}$:*

$$p_{\neg Adj(X,Y)} \leq \max_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} p_{X \perp\!\!\!\perp Y|\mathbf{B}}$$

Proof Owning to the assumption, $\mathrm{H}_{X \perp Y|\mathbf{F}}$ is true. Since X and Y are not adjacent in $\mathbb{G}$, $\mathrm{H}_{\neg \mathrm{Adj}(X,Y)}$ is also true. Furthermore, because $X \perp\!\!\!\perp Y \mid \mathbf{F}$ holds due to Theorem 1, $X \perp\!\!\!\perp Y \mid \mathbf{F} \implies X \perp Y \mid \mathbf{F}$ holds. The proof concludes due to Theorem 5. ■

In the rest of the report, we refer to a test with a conditioning set that is a sepset as a *septest*, and to the test with the conditioning set that is *the* sepset as *the* septest. According to the Corollary above, a performed septest implies an upper-bounded p-value for a false positive.

We now categorize false positives according to their direct cause. Only false positives due to type I errors have an upper-bounded p-value guaranteed by Corollary 6. The existence of false positives of the other types imply violations of one or more assumptions of Theorem 4. By counting the false positives of each type, we can identify the direct causes of these violations. We distinguish three types of false positives:

- *Type-I-error false positive* (t1FP): a false positive with one or more septests performed. All performed septests made a type I error, that is, they concluded dependence.

- *Default-decision false positive* (ddFP): a false positive with one or more septests considered but none of them performed, each time triggering the default (incorrect) decision (dependence). The tests were not performed because they were all deemed unreliable according to the employed reliability criterion. In the case of the heuristic power rule, unreliability is declared due to insufficient sample size. The existence of ddFPs implies violations of assumption 2 (all tests considered by the algorithm are reliable) of Theorem 4.

- *False-negative false positive* (fnFP): a false positive with no septest considered. No septest is considered due to the (incorrect) removal of one or more sepset members from $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$. As we will shortly see, the existence of fnFPs implies violations of one or more of the assumptions of Theorem 4.

The type of a false positive can be determined using the flowchart of Figure 7.

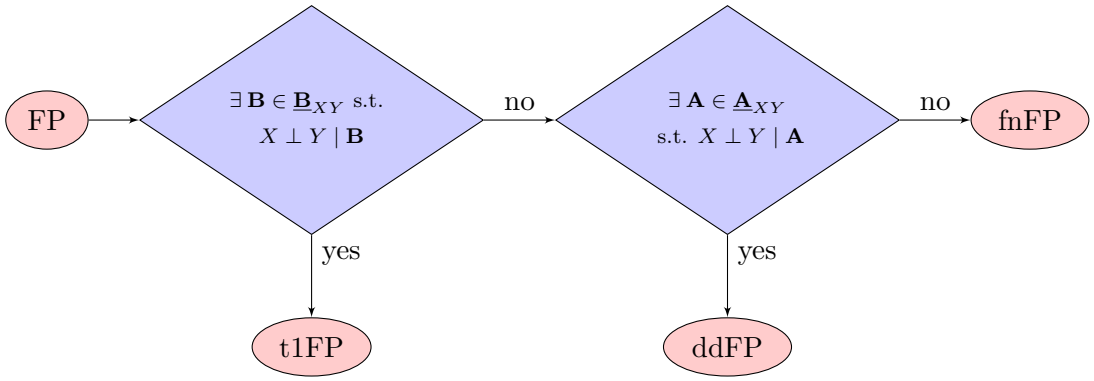


Figure 7: Flowchart used to determine the type of a false positive (FP) $X - Y$. *t1FP*, *ddFP* and *fnFP* denote type I error, default-decision and false-negative false positive, respectively (see text for definitions). $\underline{\mathbf{A}}_{XY}$ and $\underline{\mathbf{B}}_{XY}$ is the set of subsets $\mathbf{Z}$ of $\mathbf{V}$ such that the test of conditional independence of X and Y given $\mathbf{Z}$ is considered but not performed and performed, respectively.

We approximated the expected number of false positives of each type for each network and sample size (Figure 8). As n increases, the expected number of false positives decreases, in general. Most false positives are expected to be either t1FPs or fnFPs; a few ddFPs are expected in some cases. As n increases, the expected number of t1FPs and fnFPs decreases, in general.

False negatives are the direct cause of fnFPs. We distinguish three types of false negatives according to their direct cause.

- *Default decision false negative* (ddFN): a false negative with non-performed septest. This only happens when the test given the empty set is not performed, triggering the default (incorrect) decision (independence). The existence of ddFNs implies violations of assumption 2 (all tests considered by the algorithm are reliable) of Theorem 4.

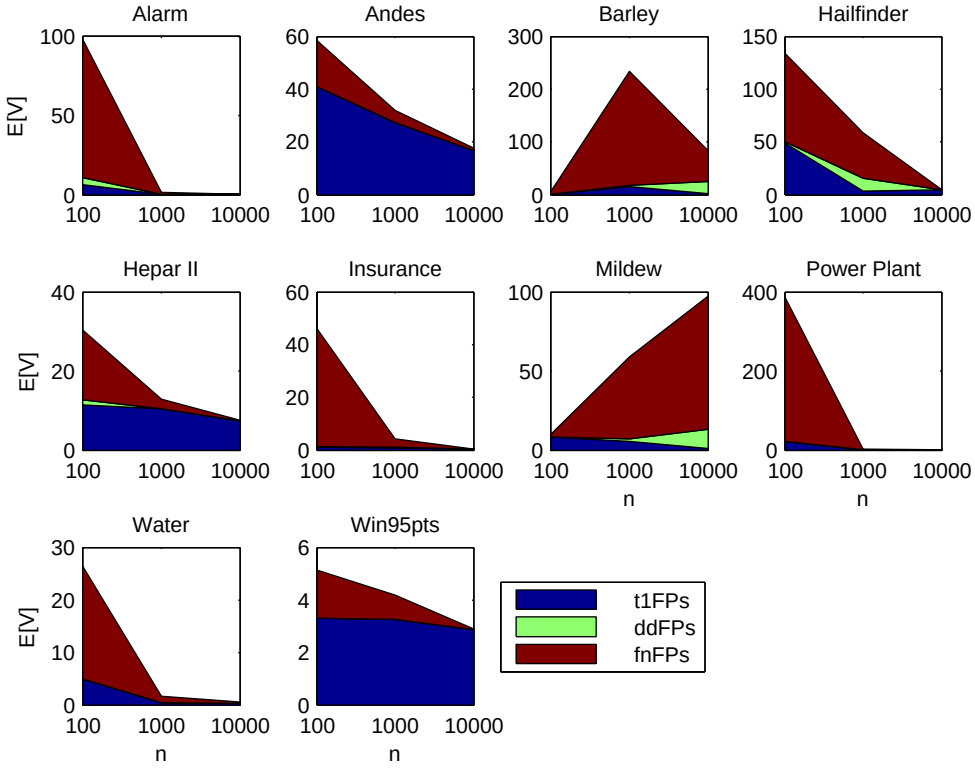


Figure 8: E[V]: expected number of false positives (FP) of each type for each network and sample size n. t1FPs, ddFPs, and fnFPs denote type-I-error, default-decision and false-negative false positives, respectively (see text for definitions). X-axes are in logarithmic-10 scale. As n increases, the expected number of false positives decreases in all networks except Barley. Most or all false positives are expected to be t1FPs for all n on Andes, Hepar II, and Win95pts. t1FPs are also expected for $n = 100$ on Alarm, $n = 1000$ on Barley, for all n on Hailfinder, Insurance, Mildew, $n = 100$ on Power Plant, and all n on Water. As n increases, the expected number of t1FPs decreases on all networks except Barley and Hailfinder. Most false positives are expected to be fnFPs for $n = 100$ on Alarm, for all n on Barley, for $n \in \{100, 1000\}$ on Hailfinder, for $n = 100$ on Hepar II, for $n \in \{100, 1000\}$ on Insurance, for $n \in \{1000, 10000\}$ on Mildew, for $n = 100$ on Power Plant and for $n \in \{100, 1000\}$ on Water. fnFPs are also expected for $n = 1000$ on Alarm, for all n on Andes, for $n = 1000$ on Hepar II, for $n = 100$ on Mildew, for $n = 10000$ on Water and for $n \in \{100, 1000\}$ on Win95pts. As n increases, the expected number of fnFPs decreases on all networks except Barley and Mildew. A few ddFPs are expected for $n = 100$ on Alarm, $n = 10000$ on Barley, $n = 1000$ on Hailfinder, $n = 100$ on Hepar II, and $n \in \{1000, 10000\}$ on Mildew.

- *Unfaithfulness false negative* (ufFN): a false negative with performed septest and the conditional independency given the sepset actually holding (unfaithfulness). The

existence of ufFNs implies violations of assumption 1 ($\mathbb{G}$ and P are faithful to each other) of Theorem 4.

- *Type-II-error false negative* (t2FN): a false negative with performed septest and the conditional independency given the sepset not holding. The septest made a type II error. Type II errors are due to *close-to-unfaithfulness*, that is, the situation when independence seems to hold due to insufficient sample size (Zhang and Spirtes, 2008). The existence of ufFNs implies violations of assumption 3 (performed tests never make a type II error) of Theorem 4.

The type of a false negative can be determined using the flowchart of Figure 9.

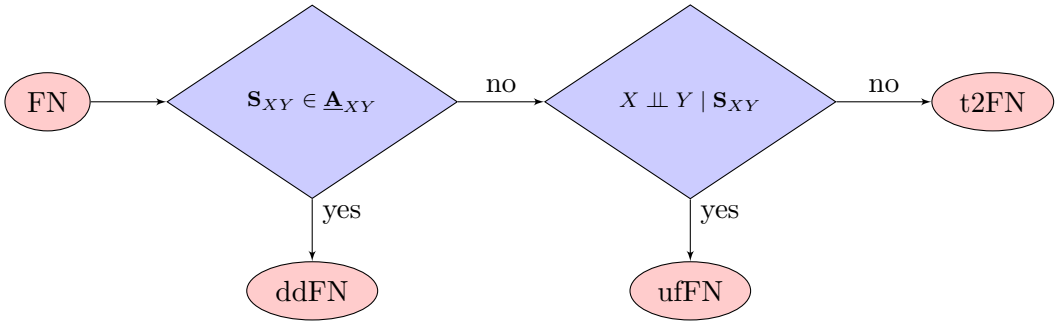


Figure 9: Flowchart used to determine the type of a false negative (FN) $X - Y$. *ddFN*, *ufFP* and *t2FN* denote default decision, unfaithfulness and type-II-error false negative (see text for definitions). $\mathbf{S}_{XY}$ denotes the sepset of $X - Y$. $\underline{\mathbf{A}}_{XY}$ is the set of subsets $\mathbf{Z}$ of $\mathbf{V}$ such that the test of conditional independence of X and Y given $\mathbf{Z}$ is considered but not performed.

We approximated the expected number of false negatives of each type for each network and sample size (Figure 10). $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$ holds if and only if the *conditional mutual information* of X and Y given $\mathbf{S}_{XY}$:

$$\mathrm{I}(X, Y \mid \mathbf{S}_{XY}) = \sum_{x \in \mathbf{D}_X, y \in \mathbf{D}_Y, \mathbf{z} \in \mathbf{D}_\mathbf{Z}} P(x, y, \mathbf{z}) \log \frac{P(x, y \mid \mathbf{z})}{P(x \mid \mathbf{z}) P(y \mid \mathbf{z})}$$

is zero (Cheng et al., 1997). For Alarm, Hepar II, Insurance, and Power Plant the above probabilities were exactly computed by the *Variable Elimination* algorithm (Russell and Norvig, 2009). For the rest networks, the probabilities were approximated by the respective relative frequencies in a sample of size 100000 created by merging the first 10 samples from the network. Most false negatives are t2FNs in most cases. As n increases, the expected number of t2FPs decreases, in general. Most or all false negatives are ddFNs for $n = 100$ on Barley and Hailfinder and for all n on Mildew. Indeed, these networks have some variables with a very large domain (e.g., one variable in Mildew takes 100 values). Even pairwise tests of independence involving these variables are deemed unreliable, leading to the default decision to discard the corresponding links. As n increases, the expected number of ddFNs decreases. This is expected because more and more pairwise tests of independence are deemed reliable. Although the expected number of false negatives decreases, the expected

number of fnFPs increases for $n \in \{1000, 10000\}$ on Mildew and for $n = 1000$ on Barley. This may be due to tests conditioning on previously ddFNs resulting in new t2FNs, which in turn cause the fnFPs. The existence of ufFNs on Andes, Hailfinder, Water and Win95pts shows that these networks do not satisfy the faithfulness condition. As n increases, the expected number of ufFNs increases. This is expected because the more tests are performed, the greater the chance to conclude independence due to unfaithfulness.

Note that not all false negatives are necessarily responsible for fnFPs; for example, there are a lot of false negatives but no fnFPs for $n = 10000$ on Hailfinder. However, we do not distinguish between subtypes of fnFPs in order to avoid overcomplicating our analysis. Therefore, we cannot deem any single cause of false negatives responsible for the failure to achieve conservative estimation and strong control.

To sum up, the failure to achieve conservative estimation and strong control in all cases of the experiment of Section 3.3 are traced back to insufficient sample size and/or unfaithfulness. We identified several approaches to eliminating false positives with non-upper-bounded p-value and experimentally evaluated two of them; the work can be found in Appendix C. The results, however, do not improve over the ones in Section 3.3. We conclude that increasing the sample size may be the only way to achieve conservative p-value-based FDR estimation and control under faithfulness, at the expense of overly conservative FDR estimation at large p-value thresholds (see Section 3.3). An alternative is to use non-p-value-based FDR estimators, as we do in the next section.

5 Non-p-value-based estimation and control

We now present non-p-value-based FDR estimators that aim at circumventing the problem of non-upper-bounded p-values.

5.1 Estimating the False Discovery Rate via simulation of null statistics

Storey and Tibshirani (2001) propose the following estimator of $\mathrm{FDR}(C)$, the FDR when rejecting hypotheses with corresponding statistic in critical region C, for any kind of dependence between the hypotheses:[7]

$$\widehat{\mathrm{FDR}}_{\mathrm{ST}}(C) \triangleq \frac{\hat{\mathrm{E}}[\mathrm{R}^0(C)]}{\mathrm{R}(C) \vee 1}$$

where $\mathrm{R}(C)$ is the number of rejected hypotheses and $\hat{\mathrm{E}}[\mathrm{R}^0(C)]$ is an estimator of $\mathrm{E}[\mathrm{R}^0(C)]$, the expected number of rejected true null hypotheses if all null hypotheses were true. Storey and Tibshirani (2001) also propose a method for estimating $\mathrm{E}[\mathrm{R}^0(C)]$ via simulation of null statistics. Under general dependence, $\widehat{\mathrm{FDR}}_{\mathrm{ST}}(C)$ is proven to be conservative under some conditions. For independent p-values, we set $\hat{\mathrm{E}}[\mathrm{R}^0(t)] = m \cdot t$ to obtain $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)$. If we set $\hat{\mathrm{E}}[\mathrm{R}^0(t)] = m \cdot t \cdot \left(\sum_{i=1}^{m} \frac{1}{i}\right)$, we end up with $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$.

[7] Actually, they propose it as an estimator of an alternative definition of FDR, called the *positive* FDR (pFDR) (Storey, 2002, 2003) (See Section 7.4). However, it can be used as a conservative estimator of FDR too, since $\mathrm{FDR} \leq \mathrm{pFDR}$. The original estimator also includes a $\hat{\pi}_0(\lambda)$ term, which is an estimator of the proportion of true null hypotheses π_0 based on a tuning parameter λ. Because $\hat{\pi}_0(\lambda)$ is not applicable in the context of Bayesian network skeleton identification, we use $\hat{\pi}_0(\lambda) = 1$ instead.

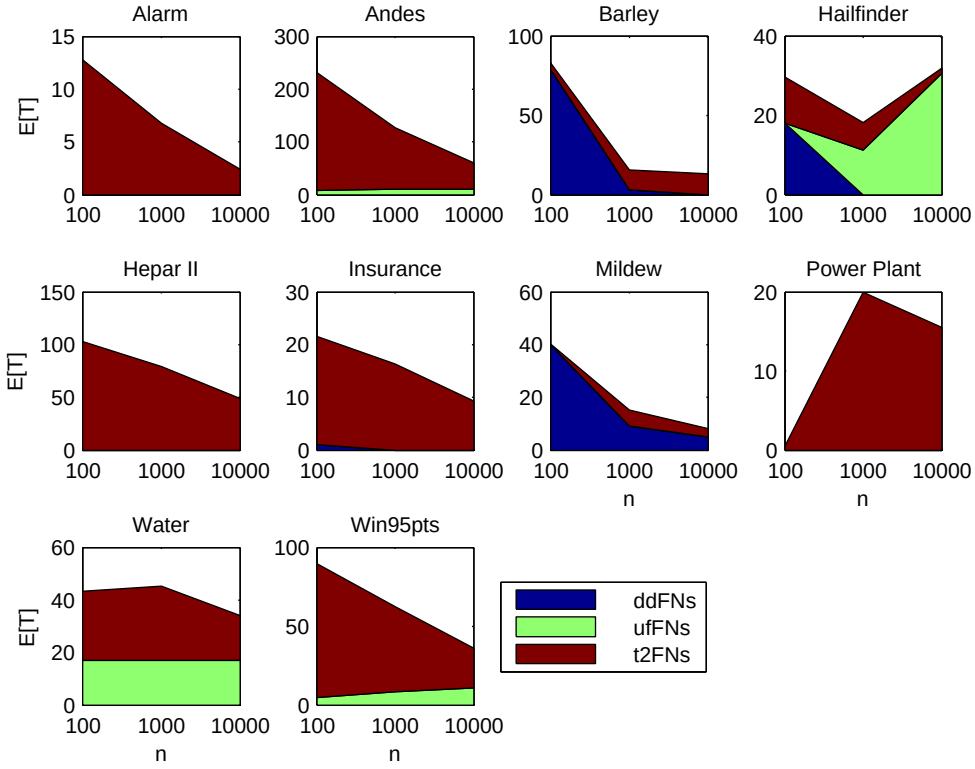


Figure 10: E[T]: expected number of false negatives (FN) of each type for each network and sample size n. ddFNs, ufFNs and t2FNs denote default-decision, unfaithfulness and type-II-error false negatives (see text for definitions). X-axes are in logarithmic-10 scale. Most or all false negatives are t2FNs for all n on Alarm and Andes, for $n \in \{1000, 10000\}$ on Barley, and for all n on Hepar II, Insurance, Power Plant, Water and Win95pts. There are also t2FNs for $n = 100$ on Barley, for all n on Hailfinder and for $n \in \{1000, 10000\}$ on Mildew. As n increases, the expected number of t2FPs decreases on all networks except Barley, Power Plant and Water. Most false negatives are ufFNs for $n \in \{1000, 10000\}$ on Hailfinder. There are also ufFNs for all n on Andes, Water and Win95pts. As n increases, the expected number of ufFNs increases. Most or all false negatives are ddFNs for $n = 100$ on Barley and Hailfinder and for all n on Mildew. There are also a few ddFNs for $n = 100$ on Insurance. As n increases, the expected number of ddFNs decreases.

The frequentist approach of Listgarten and Heckerman (2007) essentially uses $\widehat{\mathrm{FDR}}_{\mathrm{ST}}$ to estimate the FDR of structure learning. It is assumed that the structure learning algorithm can be decomposed into independent searches for the parents of each node. Then, the distribution under the hypothesis that node X_j is a not parent of X_i is simulated by randomly permuting the values x_i of X_i in the sample from P. However, this permutation method is not theoretically correct. The probability distribution under the hypothesis that

node X_j is a not parent of X_i is

$$P^{ji0}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} P^{ji0}(x_k \mid \mathbf{pa}_{X_k}^{ji0})$$

where $\mathbf{pa}_{X_i}^{ji0} = \mathbf{pa}_{X_i} \setminus \{x_j\}$ and $\mathbf{pa}_{X_k}^{ji0} = \mathbf{pa}_{X_k}$ for $k \neq i$. To sample from this distribution one needs to know the parents of each node, that is, the real DAG $\mathbb{G}$. Then one would randomly permute x_i's while keeping $\mathbf{pa}_{X_i}^{ji0}$'s fixed. However, this results in incorrect values for the descendants of X_i. One should, therefore, randomly permute both x_i's and the values of X_i's descendants while keeping both $\mathbf{pa}_{X_i}^{ji0}$'s and the values of the parents (that are not descendants of X_i) of each descendant of X_i fixed. $\mathbb{G}$ is, of course, unknown. However, we can use $\hat{\mathbb{G}}$ as an approximation. We expect that the closer $\hat{\mathbb{G}}$ to $\mathbb{G}$, the more accurate the estimation.

The permutation method just described can also be applied to the estimation of FDR of skeleton identification using LGL algorithms. The method is as follows: For each X and Y not adjacent in $\hat{\mathbb{G}}$, randomly permute either x's or y's as described above and then learn the neighbors of either X or Y to obtain a null maximal conditional independence p-value. Repeat this B times and then estimate $\mathrm{E}[\mathrm{R}^0(t)]$ by

$$\hat{\mathrm{E}}[\mathrm{R}^0(t)] = \frac{1}{\hat{\pi}_0} \cdot \frac{1}{B} \sum_{i=1}^{B} \#\{\max_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} \mathrm{p}_{X \perp\!\!\!\perp Y \mid \mathbf{B}} \leq t\}$$

where $\hat{\pi}_0$ is the proportion of non-edges among edges and non-edges in $\hat{\mathbb{G}}$. Since neighbor learning is $\mathrm{O}(|\mathbf{V}| \cdot 2^{|\mathbf{V}|})$ in the number of tests (Tsamardinos et al., 2006), this permutation method is $\mathrm{O}(B \cdot |\mathbf{V}|^2 \cdot 2^{|\mathbf{V}|})$. It is easy to see that the frequentist approach of Listgarten and Heckerman (2007), applied to skeleton identification, is a special case of this method for $\hat{G}$ being the empty graph.

5.1.1 Estimating the False Discovery Rate using the parametric bootstrap

Another possibility is to estimate FDR using the *parametric bootstrap* (Friedman et al., 1999). We denote the resulting estimator with $\widehat{\mathrm{FDR}}_{\mathrm{PB}}$. The computation of $\widehat{\mathrm{FDR}}_{\mathrm{PB}}$ is straightforward: Learn a Bayesian network $(\hat{\mathbb{G}}, \hat{P})$, generate B samples from it and apply skeleton identification to each sample. Then estimate FDR by the mean, across the samples, realized FDR, replacing the latter with zero when it not defined:

$$\widehat{\mathrm{FDR}}_{\mathrm{PB}}(t) \triangleq \frac{1}{B} \sum_{i=1}^{B} \frac{\mathrm{V}_i(t)}{\mathrm{R}_i(t) \vee 1}$$

where $\mathrm{V}_i(t)$ and $\mathrm{R}_i(t)$ are the number of rejected true null hypotheses and the number of rejections, respectively, corresponding to the i-th sample. Again, we expect that the closer $\hat{\mathbb{G}}$ to $\mathbb{G}$, the more accurate the estimation. Since MMPC-skeleton is $\mathrm{O}(|\mathbf{V}|^2 \cdot 2^{|\mathbf{V}|})$ in the number of tests (Tsamardinos et al., 2006), parametric bootstrap is $\mathrm{O}(B \cdot |\mathbf{V}|^2 \cdot 2^{|\mathbf{V}|})$.

5.2 Experiment

In order to compare $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ in the task of FDR estimation, we define the *underestimation*

$$\text{ue}(\widehat{\text{FDR}}, t) \triangleq \max\{\text{FDR}(t) - \text{E}[\widehat{\text{FDR}}(t)], 0\}$$

and the *overestimation*

$$\text{oe}(\widehat{\text{FDR}}, t) \triangleq \max\{\text{E}[\widehat{\text{FDR}}(t)] - \text{FDR}(t), 0\}$$

of $\text{FDR}(t)$ by $\widehat{\text{FDR}}(t)$. We approximated the mean $\text{ue}(\widehat{\text{FDR}}, t)$ and $\text{oe}(\widehat{\text{FDR}}, t)$ (Tables B.4 and B.5, respectively, Appendix B) for $t \in [0, \alpha]$ via trapezoidal numerical integration with step $\alpha/100$ for each network, sample size and FDR estimator. We used $B = 10$ in both $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{ST}}$. $\text{FDR}(t)$ and $\text{power}(t)$ are, of course, independent of the FDR estimator. Estimation is conservative in a p-value-threshold region if and only if the mean underestimation is zero in the region. The greater the mean overestimation, the more conservative the estimation. We say that estimation is *accurate* in a region when both mean underestimation and mean overestimation in that region are close to zero.

In order to compare the estimators in the task of FDR control, we approximated the mean $\text{uc}(\widehat{\text{FDR}}, q)$, $\text{oc}(\widehat{\text{FDR}}, q)$ and $\text{procpower}(\widehat{\text{FDR}}, q)$ (Tables B.6–B.8, Appendix B) for $q \in [0.001, 0.1]$ via trapezoidal numerical integration with step 0.001 for each network, sample size and $\widehat{\text{FDR}}$. Strong control is achieved in an FDR-threshold region if and only if the mean undercontrol is zero in the region. The less the mean overcontrol, the more tight the strong control. We say that strong control is *accurate* in a region when both mean undercontrol and mean overcontrol in that region are close to zero.

Finally, we approximated the expected *extra estimation time* for each network, sample size and FDR estimator (Table B.9, Appendix B). By extra estimation time we refer to the time spent in computing the quantities in $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{PB}}$ that are not present in $\widehat{\text{FDR}}_{\text{BY}}$. By definition, the extra estimation time of $\widehat{\text{FDR}}_{\text{BY}}$ is zero.

Table 3 summarizes the results of the experiment. $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{PB}}$ achieve, on average, more accurate estimation and control than $\widehat{\text{FDR}}_{\text{BY}}$; $\widehat{\text{FDR}}_{\text{PB}}$ does better than $\widehat{\text{FDR}}_{\text{ST}}$ and it is faster to compute. Thus, we recommend the use of $\widehat{\text{FDR}}_{\text{PB}}$ in both estimation and control if the extra computation time is not a problem. Note that a smaller B in $\widehat{\text{FDR}}_{\text{PB}}$ may be sufficient for accurate FDR estimation and control; however, we do investigate the effect of B in this report.

6 Relaxing the definition of false positive

In this section, we consider relaxed definitions of false positive, such that false positives according to these definitions have an upper-bounded p-value guaranteed under more realistic assumptions, thus increasing the likelihood that FDR is conservatively estimated and strongly controlled. We present the experimental evaluation of one such definition.

	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$	0.0331	0.04446	0.04423
Mean oe($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$	0.6204	0.009822	0.01228
Mean uc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.2022	0.0677	0.05473
Mean oc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.02687	0.01028	0.01241
Mean procpower($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.4763	0.3499	0.3404
E[time] (s)	0	420.5	737.1

Table 3: Summary of the results of the comparison of False Discovery Rate (FDR) estimators. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. ue($\widehat{\text{FDR}}, t$) and oe($\widehat{\text{FDR}}, t$) denote underestimation and overestimation, respectively, of FDR(t) by FDR estimator $\widehat{\text{FDR}}(t)$; α is the significance level of the underlying hypothesis tests; uc($\widehat{\text{FDR}}, q$) and oc($\widehat{\text{FDR}}, q$) denote undercontrol and overcontrol, respectively, of the FDR by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ and FDR threshold q; procpower($\widehat{\text{FDR}}, q$) denotes the power of the procedure; E[time] denotes the expected extra estimation time (see text for definitions). Each cell of the table contains the mean value of the corresponding quantity across various networks and sample sizes. $\widehat{\text{FDR}}_{\text{BY}}$ achieves, on average, the greatest mean power and requires no extra estimation time; $\widehat{\text{FDR}}_{\text{PB}}$ achieves, on average, the lowest mean FDR understimation and undercontrol; $\widehat{\text{FDR}}_{\text{ST}}$ achieves, on average, the lowest mean FDR overestimation and overcontrol.

6.1 Relaxed definitions of false positive

In order to facilitate the presentation of our first consideration for a relaxed definition of false positive, we present the following theorem that relaxes assumption 2 of Theorem 4. The proof of the theorem can be found in Appendix A.

Theorem 7 *Let $(\mathbb{G}, P)$ be a Bayesian network of the variables in some set $\mathbf{V}$ and $X, Y \in \mathbf{V}$. Suppose that an algorithm instantiating Algorithm Template 1 is applied on a sample from P. If*

1. *$\mathbb{G}$ and P are faithful to each other*

2. *there is a sepset $\mathbf{S}_{XY}$ that is a subset of $\mathbf{ADJ}_X$ or $\mathbf{ADJ}_Y$ such that $H_{X \perp\!\!\!\perp Y | \mathbf{S}_{XY}}$ would be true if $H_{\neg Adj(X,Y)}$ was true and $test_{X \perp\!\!\!\perp Y | \mathbf{S}_{XY}}$ is reliable according to the employed reliability criterion,*

3. *performed tests never make a type II error,*

then $p_{\neg Adj(X,Y)}$ *is upper-bounded by the maximal among* $p_{X \perp\!\!\!\perp Y|\mathbf{B}}$ *for* $\mathbf{B} \in \underline{\mathbf{B}}_{XY}$*:*

$$p_{\neg Adj(X,Y)} \leq \max_{\mathbf{Z} \in \underline{\mathbf{B}}_{XY}} p_{X \perp\!\!\!\perp Y|\mathbf{B}}$$

The following corollary of Theorem 7 specializes in false positive p-values:

Corollary 8 *Let* $(\mathbb{G}, P)$ *be a Bayesian network of the variables in some set* $\mathbf{V}$ *and* X *and* Y *be distinct nodes in* $\mathbf{V}$ *that are not adjacent in* $\mathbb{G}$*. Suppose that an algorithm instantiating Algorithm Template 1 is applied on a sample from* P*. If*

1. $\mathbb{G}$ *and* P *are faithful to each other*
2. *there is a sepset* $\mathbf{S}_{XY}$ *that is a subset of* $\mathbf{ADJ}_X$ *or* $\mathbf{ADJ}_Y$ *such that* $test_{X \perp\!\!\!\perp Y|\mathbf{S}_{XY}}$ *is reliable according to the employed reliability criterion,*
3. *performed tests never make a type II error,*

then $p_{\neg Adj(X,Y)}$ *is upper-bounded by the maximal among* $p_{X \perp\!\!\!\perp Y|\mathbf{B}}$ *for* $\mathbf{B} \in \underline{\mathbf{B}}_{XY}$*:*

$$p_{\neg Adj(X,Y)} \leq \max_{\mathbf{B} \in \underline{\mathbf{B}}_{XY}} p_{X \perp\!\!\!\perp Y|\mathbf{B}}$$

Proof Owning to assumption 2, $\mathrm{H}_{X \perp\!\!\!\perp Y|\mathbf{S}_{XY}}$ is true. Since X and Y are not adjacent in $\mathbb{G}$, $\mathrm{H}_{\neg \mathrm{Adj}(X,Y)}$ is also true. The proof concludes due to Theorem 7. ■

If we consider as false positives only the ones for which assumption 2 of Corollary 8 actually holds, we end up with the following relaxed definition of false positive, originally by Tsamardinos and Brown (2008):

Definition 9 *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* P *of the variables in some set* $\mathbf{V}$ *satisfies the faithfulness condition and that a conditional-independence-test-based algorithm instantiating Algorithm Template 1 is applied on a sample from* P*. A* ***false positive*** *is a falsely discovered link* $X - Y$ *for which there is sepset–subset of the neighbors of* X *or* Y *such that the corresponding test is reliable according to the employed reliability criterion:*

$$\exists\, \mathbf{S}_{XY} \subseteq \mathbf{ADJ}_X \text{ or } \mathbf{ADJ}_Y \text{ s.t. } X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY} \text{ and } test_{X \perp\!\!\!\perp Y|\mathbf{S}_{XY}} \text{ is reliable}$$

Definition 9 is intuitive only if the reliability criterion is intuitive. In general, as conditioning set cardinality $|\mathbf{Z}|$ increases, reliability decreases (Tsamardinos et al., 2006). Thus, an intuitive criterion is to consider a test reliable only if $|\mathbf{Z}|$ is below some upper limit *max-k*. We refer to this criterion as the *maximal conditioning set cardinality rule*. For categorical variables, this rule is usually used in conjunction with the heuristic power rule as a way to reduce the execution time of the algorithm (Aliferis et al., 2010a). When employing the maximal conditioning set cardinality rule, a false positive according to Definition 9 is a falsely discovered link $X - Y$ for which there is a sepset–subset of the neighbors of X or Y with cardinality not greater than *max–k*. Note that, when employing the heuristic power

rule, using Definition 9 is essentially an approach to dealing with insufficient sample size to declare reliability.

We now present a relaxed definition of false positive, originally by Tsamardinos et al. (2008), that is intuitive regardless of the employed reliability criterion. The idea behind this definition is to "translate" the employed reliability criterion to the maximal conditioning set cardinality rule and use Definition 9 with the latter criterion. The translation corresponds to finding the *worst-case maximal conditioning set cardinality worst-max-k* of the employed reliability criterion:

Definition 10 *Let P be a joint probability distribution of the variables in some set $\mathbf{V}$. The **worst-case maximal conditioning set cardinality** worst-max-k of a reliability criterion for samples of size n from P is the maximal max-k such that, the test of conditional independence of every variable X with every other variable Y in $\mathbf{V}$ given a subset $\mathbf{Z}$ of $\mathbf{V} \setminus \{X, Y\}$ with cardinality not greater than worst-max-k is reliable according to the criterion:*

$$worst\text{-}max\text{-}k \triangleq \max max\text{-}k \text{ s.t.}$$
$$\forall\, X, Y \in \mathbf{V},\ X \neq Y,\ \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ and } |\mathbf{Z}| \leq max\text{-}k,\ test_{X \perp\!\!\!\perp Y | \mathbf{Z}} \text{ is reliable}$$

It is not hard to see that the following algorithm calculates the *worst-max-k* of the heuristic power rule:

Algorithm 3 Calculate worst-max-k (see text for definition) for the heuristic power rule.

Let $\mathrm{d}_{(1)} \geq \mathrm{d}_{(2)} \cdots \geq \mathrm{d}_{(m)}$ be the ordered domain sizes of the variables in $\mathbf{V}$
$worst\text{-}max\text{-}k \leftarrow 0$
while $worst\text{-}max\text{-}k \leq |\mathbf{V}| - 2$ and $n / \prod_{i=1}^{2+worst\text{-}max\text{-}k} \mathrm{d}_{(i)} \geq h\text{-}ps$ **do**
  $worst\text{-}max\text{-}k \leftarrow worst\text{-}max\text{-}k + 1$
end while
$worst\text{-}max\text{-}k \leftarrow worst\text{-}max\text{-}k - 1$
if $worst\text{-}max\text{-}k < 0$ **then**
  $worst\text{-}max\text{-}k \leftarrow nonexistent$
end if

Using Definition 9 with the maximal conditioning set cardinality rule and $max\text{-}k = worst\text{-}max\text{-}k$ corresponds to using the following definition of false positive:

Definition 11 *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution P of the variables in some set $\mathbf{V}$ satisfies the faithfulness condition and that a conditional-independence-test-based algorithm instantiating Algorithm Template 1 is applied on a sample from P. A **false positive** is a falsely discovered link $X - Y$ such that there is a sepset–subset of the neighbors of X or Y with cardinality not greater than worst-max-k:*

$$\exists\, \mathbf{S}_{XY} \subseteq \mathbf{ADJ}_X \text{ or } \mathbf{ADJ}_Y \text{ s.t. } |\mathbf{S}_{XY}| \leq worst\text{-}max\text{-}k \text{ and } X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$$

where worst-max-k is the worst-case maximal conditioning set cardinality of the employed reliability criterion.

From now on, we refer to Definition 11 as the relaxed definition of false positive. An alternative way to use the definition is to first set a desired conditioning set cardinality *max-k* and then find the *worst-case minimal sample size worst-min-n* of the employed reliability criterion:

Definition 12 *Let P be a joint probability distribution of the variables in some set $\mathbf{V}$. The* ***worst-case minimal sample size*** *worst-min-n of a reliability criterion for conditioning set cardinality max-k is the minimal size n for a sample from P such that, the test of conditional independence of every variable X with every other variable Y in $\mathbf{V}$ given a subset $\mathbf{Z}$ of $\mathbf{V} \setminus \{X, Y\}$ with cardinality not greater than max-k is reliable according to the criterion:*

$$\textit{worst-min-n} \triangleq \min n \text{ s.t.}$$
$$\forall\, X, Y \in \mathbf{V},\ X \neq Y,\ \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ and } |\mathbf{Z}| \leq \textit{max-k},\ \textit{test}_{X \perp\!\!\!\perp Y | \mathbf{Z}} \text{ is reliable}$$

It is easy to see that $\textit{worst-min-n} = \lceil \textit{h-ps} \cdot \prod_{i=1}^{2+\textit{max-k}} \mathrm{d}_{(i)} \rceil$ for the heuristic power rule. It is also easy to see that, for samples of size *worst-min-n*, $\textit{worst-max-k} = \textit{max-k}$.

6.2 Experiment

First, we calculated *worst-max-k* for each network and sample size (Table B.10, Appendix B). *worst-max-k* does not exist for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance and for all n on Mildew. Then, we approximated mean $\mathrm{ue}(\widehat{\mathrm{FDR}}, t)$ and $\mathrm{oe}(\widehat{\mathrm{FDR}}, t)$ for $t \in [0, \alpha]$, and mean $\mathrm{uc}(\widehat{\mathrm{FDR}}, q)$ and $\mathrm{oc}(\widehat{\mathrm{FDR}}, q)$ for $q \in [0.001, 0.1]$ (Tables B.11–B.14, Appendix B) for each network, sample size, and definition of false positive (regular and relaxed). In the results, we include the performance of $\widehat{\mathrm{FDR}}_{\mathrm{PB}}$ using the regular definition of false positive. $\mathrm{power}(t)$ is, of course, independent of both the FDR estimator and the definition of false positive, while $\mathrm{procpower}(\widehat{\mathrm{FDR}}, q)$ is only independent of the definition of false positive.

Table 4 summarizes the results of the experiment. $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ always achieves conservative estimation of the FDR according to the relaxed definition of false positive, but the estimation is overly conservative. For the task of FDR estimation, therefore, we recommend the use of the regular definition with $\widehat{\mathrm{FDR}}_{\mathrm{PB}}$ if the extra estimation time is not a problem and with $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ otherwise. On average, the combination of the relaxed definition with $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ leads to much less FDR undercontrol but also larger overcontrol than the combination of the regular definition with either FDR estimator, and is more powerful than the combination of the regular definition with $\widehat{\mathrm{FDR}}_{\mathrm{PB}}$. For the task of FDR control, therefore, we recommend the use of the relaxed definition of false positive with $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ over the use of the regular definition with either FDR estimator.

7 Other approaches to assessing confidence in structure learning

Confidence can be assessed on entire structures or certain structural features (for example, links). Skeleton identification, viewed as multiple hypothesis testing, assesses confidence on

	Reg	Reg (PB)	Rel
Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$	0.02462	0.004486	0
Mean oe($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$	0.7366	0.03624	0.8648
Mean uc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.1111	0.004645	0.0001246
Mean oc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.03293	0.02277	0.04803
Mean procpower($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$	0.4763	0.3018	0.4763
E[time] (s)	0	420.5	0

Table 4: Summary of the results of the comparison of False Positive definitions. *Reg* and *Rel* refer to the combination of the regular and the relaxed, respectively, definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001), while *Reg (PB)* refers to the combination of the regular definition of false positive with the parametric-bootstrap-based FDR estimator. ue($\widehat{\text{FDR}}, t$) and oe($\widehat{\text{FDR}}, t$) denote underestimation and overestimation, respectively, of FDR(t) by FDR estimator $\widehat{\text{FDR}}(t)$; α is the significance level of the underlying hypothesis tests; uc($\widehat{\text{FDR}}, q$) and oc($\widehat{\text{FDR}}, q$) denote undercontrol and overcontrol, respectively, of the FDR by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ and FDR threshold q; procpower($\widehat{\text{FDR}}, q$) denotes the power of the procedure; E[time] denotes the expected extra estimation time (see text for definitions). Each cell of the table contains the mean value of the corresponding quantity across various networks and sample sizes. Rel achieves zero FDR underestimation while Reg and Reg (PB) do not. Also, mean undercontrol is much less for Rel than Reg and Reg (PB) and mean power for Rel (same as for Reg) is larger than for Reg (PB), on average. However, mean overestimation for Rel is larger than for Reg and much larger than for Reg, on average, and mean overcontrol is larger for Rel than for Reg or Reg (PB), on average.

links (through the link absence p-values) and entire skeletons (through FPR or FDR). We now discuss other approaches to assessing confidence in structure learning, namely search-and-score methods, Bayesian model averaging, and the use of the bootstrap in assessing confidence on structural features.

7.1 Search-and-score methods and Bayesian model averaging

Search-and-score structure-learning methods search for the DAG or DAG pattern $\mathbb{G}$ that maximizes some score function (Tsamardinos et al., 2006). Thus, confidence is assessed on entire structures (through the score). A typical score function is the posterior probability $\Pr(\mathbb{G} \mid d)$ of $\mathbb{G}$ given the data d. $\Pr(\mathbb{G} \mid d)$ is called the *Bayesian score* of $\mathbb{G}$. Exhaustive consideration of possible structures is computationally infeasible when the number of nodes is large, because the number of possible DAGs is super-exponential to the number of nodes (Robinson, 1977). The number of DAG patterns is smaller but it is still large (Steinsky,

2003). For this reason, algorithms that perform a heuristic search over the space of DAGs or DAG patterns have been devised (Neapolitan, 2004).

When the number of variables is small and the sample size is large, a single structure can be orders of magnitude more probable than the rest structures (Heckerman et al., 1999). However, when the sample size is small relative to the number of variables, there are often many structures that are equally probable and choosing one would be arbitrary (Neapolitan, 2004). In such cases, one would perform *Bayesian model averaging* to compute the probability $\Pr(\text{Present}(f) \mid d)$ of the event $\text{Present}(f)$ that a certain feature f of the DAG or DAG pattern $\mathbb{G}$ is present, given the data d:

$$\Pr(\text{Present}(f) \mid d) = \sum_{\mathbb{G}} 1\{\text{Present}_{\mathbb{G}}(f)\} \cdot P(\mathbb{G} \mid d)$$

where $1\{\cdot\}$ is the indicator function which is one when its input occurs and 0 otherwise and $\text{Present}_{\mathbb{G}}(f)$ denotes the event that f is present in $\mathbb{G}$.

The exact computation of $\Pr(\text{Present}(f) \mid d)$ by averaging over all possible structures is computationally infeasible when the the number of nodes is not small for the reasons discussed above. In these cases, $\Pr(\text{Present}(f) \mid d)$ can be approximated by searching for highly probable structures and then average over them (Neapolitan, 2004). This is usually done using *Monte Carlo Markov Chain* (MCMC) based methods (Madigan et al., 1995; Friedman and Koller, 2003; Eaton and Murphy, 2007; Grzegorczyk and Husmeier, 2008), which are computationally very expensive.

$\Pr(\text{Present}(f) \mid d)$, where f is a subnetwork (for example, a single edge) can be computed exactly for moderately sized networks (about 25 nodes or less) using dynamic programming (Koivisto and Sood, 2004; Koivisto, 2006; Tian and He, 2009). Application to larger networks is currently prohibited due to space requirements. Further work by Parviainen and Koivisto (2009, 2010) is targeted on reducing these requirements.

The Bayesian approach of Listgarten and Heckerman (2007) is to estimate the expected number $\text{E}[\text{S} \mid d]$ of true edges given data d by averaging over all possible structures:

$$\text{E}[\text{S} \mid d] = \sum_{\mathbb{G}} \text{S}(\mathbb{G}, \hat{\mathbb{G}}) \cdot P(\mathbb{G} \mid d)$$

where $\text{S}(\mathbb{G}, \hat{\mathbb{G}})$ is the number of edges that are present in both $\hat{\mathbb{G}}$ and $\mathbb{G}$.

7.2 The bootstrap

Friedman et al. (1999) use the bootstrap to estimate the probability $\Pr(\text{Present}_{\hat{\mathbb{G}}}(f) \mid |D| = n)$ of the event $\text{Present}_{\hat{\mathbb{G}}}(f)$ that a certain feature f of the DAG or DAG pattern $\hat{\mathbb{G}}$ learned from a sample of size n is present. If the structure learning algorithm is *consistent*, then we can expect that, as n increases, $\Pr(\text{Present}_{\hat{\mathbb{G}}}(f) \mid |D| = n) \to 1$ if $\text{Present}_{\mathbb{G}}(f)$ occurs and $\Pr(\text{Present}_{\hat{\mathbb{G}}}(f) \mid |D| = n) \to 0$ if $\text{Present}_{\mathbb{G}}(f)$ does not occur (Friedman et al., 1999).

Non-parametric bootstrap works as follows: First, B bootstrap samples (that is, samples with replacement) are generated from the original sample d. Then structure learning is

applied to each bootstrap sample. Finally, $\Pr(\text{Present}_{\hat{\mathbb{G}}}(f) \mid |D| = n)$ is estimated by

$$\frac{1}{B}\sum_{i=1}^{B} 1\{\text{Present}_{\hat{\mathbb{G}}_i}(f)\}$$

where $\hat{\mathbb{G}}_i$ denotes the structure learned from the i-th bootstrap sample. The learned structures from the non-parametric bootstrap can be also used as the highly probable structures in the estimation of $\Pr(\text{Present}(f) \mid d)$ in Bayesian model averaging. *Parametric* bootstrap is similar to the non-parametric except that we generate B samples from the network learned from the original sample.

7.3 Classification of pairs of nodes

The scores in the search-and-score approach are not directly comparable to the error rates in the constraint-based approach. However, methods estimating or exactly computing $\text{E}[\text{S} \mid d]$ can be compared to FDR estimators and methods estimating or exactly computing $\Pr(\text{Present}(X - Y) \mid d)$ or $\Pr(\text{Present}_{\hat{\mathbb{G}}}(X - Y) \mid |D| = n)$ for each pair (X, Y) of nodes in $\mathbf{V}$ can be compared to skeleton identification in the context of classification.

Skeleton identification, Bayesian model averaging for estimating or exactly computing $\Pr(\text{Present}(X - Y) \mid d)$ and the bootstrap for estimating $\Pr(\text{Present}_{\hat{\mathbb{G}}}(X - Y) \mid |D| = n)$ can be viewed as scoring classification of pairs of nodes as *links* or *non-links*. The score is $\Pr(\text{Present}(X - Y) \mid d)$, $\Pr(\text{Present}_{\hat{\mathbb{G}}}(X - Y) \mid |D| = n)$ and $-\,\text{p}_{\neg\text{Adj}(X,Y)}$ (the negative link absence p-value) for each method respectively.

Receiver Operating Characteristic (ROC) analysis (Fawcett, 2003) has been extensively used for comparing methods for estimating or exactly computing $\Pr(\text{Present}(X \to Y) \mid d)$ or $\Pr(\text{Present}_{\hat{\mathbb{G}}}(X \to Y) \mid |D| = n)$ to each other (Friedman and Koller, 2003; Koivisto, 2006; Eaton and Murphy, 2007; Grzegorczyk and Husmeier, 2008). An experimental comparison of skeleton identification to methods for estimating or exactly computing $\Pr(\text{Present}(X - Y) \mid d)$ or $\Pr(\text{Present}_{\hat{\mathbb{G}}}(X - Y) \mid |D| = n)$ for each pair (X, Y) of nodes in $\mathbf{V}$ is yet to be conducted.

The *Positive Predictive Value* (PPV) of a classifier on a dataset is defined as the ratio S/R of the number S of true positives to the number R of predicted positives. It is easy to see that the PPV of multiple hypothesis testing, viewed as binary classification, is equal to 1 minus realized FDR. Thus, $1 - \widehat{\text{FDR}}$ can be used an estimator of E[PPV]. $\text{E}[\text{S} \mid d]/R$, where R is the number of links in $\hat{\mathbb{G}}$, can be also used as an estimator of E[PPV]. Listgarten and Heckerman (2007) compare their two approaches in terms of PPV-estimation error in the context of classification of possible edges as edges or non-edges. An experimental comparison of the Bayesian approach of Listgarten and Heckerman (2007) to FDR estimation is yet to be conducted in the context of classification of pairs of nodes as links or non-links.

7.3.1 Skeleton identification as classification of pairs of nodes

We generated the ROC curve (Figure 11) of skeleton identification and calculated the area under the curve (AUC) (Table 5) for each network and sample size. We set link absence p-values $> \alpha$ to 1 because we are not interested in the ability of skeleton identification in scoring discarded links (see Section 3.2.2). Classification performance is generally improved

as n increases. Classification performance is above average in some cases where conservative estimation and strong control are not achieved by $\widehat{\text{FDR}}_{\text{BY}}$; however, classification performance is below average in some cases where conservative estimation and strong control *are* achieved. Thus, it seems that there is no correspondence between classification performance and FDR estimation and control performance.

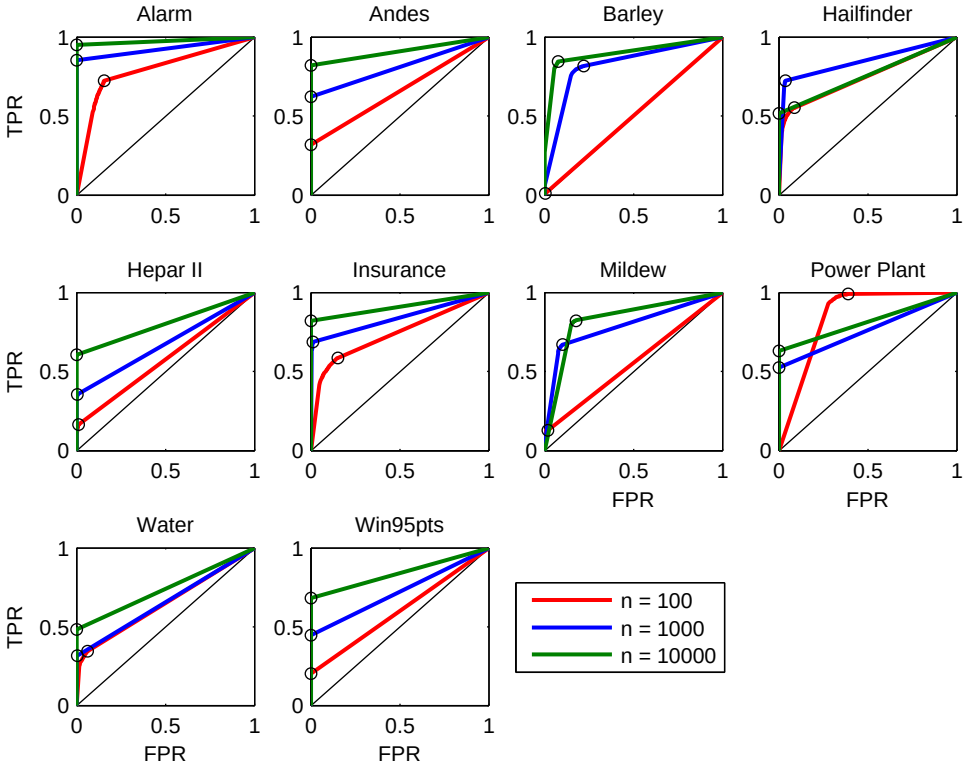


Figure 11: Receiver Operating Characteristic (ROC) curve for each network and sample size n. The point of each curve that corresponds to p-value threshold $\alpha = 0.05$ is surrounded by a circle. Curves tend to $(1, 1)$ as n increases in all networks except Hailfinder and Power Plant.

7.4 A Bayesian interpretation of the False Discovery Rate in skeleton identification

Storey (2003) introduces a modified version of FDR called the *positive False Discovery Rate* (pFDR), and discusses its advantages over FDR. pFDR and FDR are asymptotically ($m \to \infty$) equivalent for a fixed p-value threshold (Storey, 2002); we confirmed this for the experiment of Section 3.3. The definition of pFDR is the following:

$$\text{pFDR} \triangleq \text{E}\left[\frac{\text{V}}{\text{R}} \,\middle|\, \text{R} > 0\right]$$

Network	n = 100	n = 1000	n = 10000
Alarm	0.791	0.9254	0.9741
Andes	0.6573	0.8114	0.9108
Barley	0.5015	0.8251	0.8967
Hailfinder	0.7456	0.8455	0.757
Hepar II	0.5761	0.6758	0.8013
Insurance	0.7363	0.839	0.9101
Mildew	0.5562	0.7916	0.8306
Power Plant	0.8431	0.7617	0.8147
Water	0.6458	0.6553	0.7412
Win95pts	0.5995	0.7214	0.8397
Mean:	0.6652	0.7852	0.8476

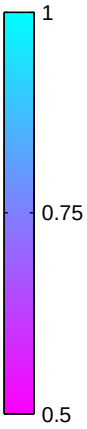


Table 5: Area under the Receiver Operating Characteristic curve (AUC) for each network and sample size n. AUC increases as n increases on all networks except Hailfinder and Power Plant. Even if the FDR estimator by Benjamini and Yekutieli (2001) does not achieve FDR conservative estimation (Figure 4) and strong control (Figure 5) for $n = 100$ on Alarm, for $n \in \{1000, 10000\}$ on Barley, $n = 1000$ on Hailfinder, $n \in \{1000, 10000\}$ on Mildew, and for $n = 100$ on Power Plant, AUC is ≥ 0.75 in those cases. On the other hand, while conservative estimation and strong control are achieved for $n \in \{100, 1000\}$ on Hepar II, for $n = 100$ on Mildew, for $n \in \{1000, 10000\}$ on Water, and for $n \in \{100, 1000\}$ on Win95pts, AUC is < 0.75 in those cases.

An interesting property of pFDR is that, if we assume that the test statistics come from a mixture distribution, then pFDR can be expressed as a Bayesian posterior probability (Storey, 2003):

Theorem 13 *Suppose m identical hypothesis tests are performed with the statistics T_1, ..., T_m and critical region C. Assume that (T_i, H_i) are independent, identically distributed random variables, $T_i \mid H_i$ $(1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and H_i Bernoulli(π_1) for $i = 1, \ldots, m$. Then,*

$$pFDR(C) = Pr(H = 0 \mid T \in C) \tag{1}$$

where $pFDR(C)$ is the pFDR obtained when rejecting hypotheses with corresponding statistic $\in C$ and $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

Proof The proof can be found in Storey (2003). ■

Note that Eq. 1 does not depend on m and $\Pr(H_i = 0 \mid T_i \in C)$ is the same for $i = 1, \ldots, m$.

As mentioned in Section 7.1, the Bayesian score assumes a prior probability distribution of DAGs or DAG patterns. If we assume for any pair of nodes (X, Y) that $\Pr(\neg \text{Adj}(X, Y)) =$

π_0 and $\Pr(X - Y) = \pi_1$ such that $\pi_0 = 1 - \pi_1$, $\Pr(P_{\neg\text{Adj}(X,Y)} \leq t \mid \neg\text{Adj}(X,Y)) = F_0$ and $\Pr(P_{\neg\text{Adj}(X,Y)} \leq t \mid X - Y) = F_1$ instead, then:

$$\text{pFDR}(t) = \Pr(\neg\text{Adj}(X,Y) \mid P_{\neg\text{Adj}(X,Y)} \leq t)$$

That is, pFDR is the probability of a link in the output skeleton being false.

8 Summary and future work

This report focuses on the problem of estimating and controlling FDR of identifying the skeleton of a Bayesian network. The main contributions are the following:

- A unified approach to estimation and control of the FDR of Bayesian network skeleton identification
- An experimental evaluation of a standard FDR estimator in both tasks over several benchmark networks and sample sizes showing that conservative estimation and strong control of FDR are not achieved in some cases
- A thorough analysis of the causes of the failure to achieve conservative estimation and strong control of FDR in those cases, tracing that failure back to insufficient sample size and/or the violation of the faithfulness condition
- An experimental evaluation of several approaches to improving estimation and control of FDR, including a permutation-based and a parametric-bootstrap-based FDR estimator demonstrated to achieve more accurate estimation and strong control
- An experimental evaluation of a relaxed definition of false positive, showing that the latter leads to more conservative estimation and strong control
- A discussion about the relationship of the FDR of skeleton identification with other approaches to assessing confidence in structure learning

We note the following as future work in using FDR when learning Bayesian network structure and when estimating structural uncertainty in general:

- Unfaithfulness and close-to-unfaithfulness are causes of failure to achieve conservative estimation and strong control of FDR yet to be dealt with.
- An experimental comparison, in the context of classification of pairs of nodes as links or non-links, of skeleton identification to Bayesian model averaging and the bootstrap, is yet to be conducted.
- In this work we have focused on learning the structure of the whole network. However, the FDR of learning the neighbors of a single node, may be of interest. The use of FDR of neighbor learning has some special issues (see Tsamardinos et al., 2008) and has to be studied separately.

Acknowledgements

This work was partially funded by the Institute of Computer Science, Foundation for Research and Technology - Hellas. We would like to thank Sofia Triantafillou for her feedback.

References

B. Abramson, J. Brown, W. Edwards, A. Murphy, and R.L. Winkler. Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71, 1996.

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research, Special Topic on Causality*, 11:171–234, 2010a.

C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification. part ii: Analysis and extensions. *Journal of Machine Learning Research, Special Topic on Causality*, 11:235–284, 2010b.

A.P. Armen. Estimation and control of the false discovery rate in bayesian network skeleton identification, with application to biological data. Master's thesis, University of Crete, Heraklion, March 2011.

IA Beinlich, HJ Suermondt, RM Chavez, and GF Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *AI in Medicine in Europe*.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995. ISSN 00359246. URL `http://www.jstor.org/stable/2346101`.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001. ISSN 0090-5364.

J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, 1997. ISSN 0885-6125.

J. Cheng, D.A. Bell, and W. Liu. Learning belief networks from data: An information theory based approach. In *Proceedings of the sixth international conference on Information and knowledge management*, pages 325–331. ACM, 1997.

C. Conati, A. S. Gertner, K. VanLehn, and M. J. Druzdzel. On-line student modeling for coached problem solving using Bayesian networks. In *Proceedings of the Sixth International Conference on User Modeling (UM–96)*, pages 231–242, Vienna, New York, 1997. Springer Verlag.

D. Eaton and K. Murphy. Bayesian structure learning using dynamic programming and mcmc. *Proceedings of the Twenty-Third Confererence on Uncertainty in Artificial Intelligence (UAI 2007)*, 2007.

A. Fast, M. Hay, and D. Jensen. Improving Accuracy of Constraint-Based Structure Learning. 2009.

A.S. Fast. *Learning the Structure of Bayesian Networks with Constraint Satisfaction*. PhD thesis, University of Massachusetts Amherst, 2010.

T. Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. *HP Laboratories technical report*, 2003.

S.E. Fienberg. The analysis of cross-classified categorical data. 1977.

N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1):95–125, 2003. ISSN 0885-6125.

N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks. *AI&STAT VII*, 1999.

M. Grzegorczyk and D. Husmeier. Improving the structure mcmc sampler for bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71(2):265–305, 2008.

D. Heckerman, C. Meek, and G. Cooper. A Bayesian Approach to Causal Discovery. *Computation, causation, and discovery*, page 141, 1999.

A.L. Jensen and F.V. Jensen. Midasan influence diagram for management of mildew in winter wheat. In *Proceedings of the 12th annual conference on uncertainty in artificial intelligence (UAI-96)*, pages 349–356, 1996.

FV Jensen. U. kj rul, kg olesen, and j. pedersen (1989), et forprojekt til et ekspertsystem for drift af spildevandsrensning (an expert system for control of waste water treatment—a pilot project). Technical report, Technical Report, Judex Datasystemer A/S, Aalborg, Denmark (in Danish), 1989.

M. Koivisto. Advances in exact bayesian structure discovery in bayesian networks. In *Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 241–248, 2006.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004. ISSN 1532-4435.

K. Kristensen and I.A. Rasmussen. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33(3):197–217, 2002. ISSN 0168-1699.

J. Lemeire, S. Meganck, F. Cartella, and T. Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *International Journal of Approximate Reasoning*, 2012.

J. Li and Z. J. Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *J. Mach. Learn. Res.*, 10:475–514, 2009. ISSN 1532-4435.

J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned dag model: Investigation of a bayesian and a frequentist approach. In *23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995. ISSN 0306-7734.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the eleventh international conference on uncertainty in artificial intelligence*, 1995.

R.E. Neapolitan. *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004. ISBN 0130125342.

T.D. Nielsen and F.V. Jensen. On-line alert systems for production plants: A conflict based approach. *International journal of approximate reasoning*, 45(2):255–270, 2007.

A. Onisko. *Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders*. PhD thesis, Ph. D. Dissertation, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw, 2003.

P. Parviainen and M. Koivisto. Exact structure discovery in bayesian networks with less space. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 436–443. AUAI Press, 2009.

P. Parviainen and M. Koivisto. Bayesian structure discovery in bayesian networks with less space. pages 589–596, 2010.

R. Robinson. Counting unlabeled acyclic digraphs. *Combinatorial mathematics V*, pages 28–43, 1977.

S. Russell and P. Norvig. Artificial intelligence: A modern approach. 2009.

J.P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46, 1995.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. 2000. ISBN 0262194406.

B. Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete mathematics*, 270(1-3):267–278, 2003. ISSN 0012-365X.

J. D. Storey. False Discovery Rates. *International Encyclopedia of Statistical Science, Lovric M (editor)*, 2010.

J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002. ISSN 1467-9868.

J.D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. ISSN 0090-5364.

J.D. Storey and R. Tibshirani. *Estimating the positive false discovery rate under dependence, with applications to DNA microarrays.* 2001.

J.D. Storey, J.E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004. ISSN 1467-9868.

J. Tian and R. He. Computing posterior probabilities of structural features in bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 538–547. AUAI Press, 2009.

I. Tsamardinos and G. Borboudakis. Permutation Testing Improves Bayesian Network Learning. *Machine Learning and Knowledge Discovery in Databases*, pages 322–337, 2010.

I. Tsamardinos and L. E. Brown. Bounding the false discovery rate in local bayesian network learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, volume 2, pages 1100–1105. AAAI Press, 2008.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning Journal*, 65:31–78, 2006.

I. Tsamardinos, L.E. Brown, and S. Triantafylloy. Controlling the False Discovery Rate in Bayesian Network Structure Learning. 2008.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.

Appendix A

Here we prove Theorems 4, 5 and 7. Their proofs use the following three lemmas:

Lemma 14 *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* P *of the variables in some set* $\mathbf{V}$ *is a Bayesian network. Suppose further that a conditional-independence-test-based algorithm instantiating Algorithm Template 1 is applied on a sample from* P*. If*

1. $\mathbb{G}$ *and* P *are faithful to each other and*
2. *performed tests never make a type II error*

then the algorithm discovers all links in $\mathbb{G}$*.*

Proof For each $X \in \mathbf{V}$ and $Y \in \mathbf{V} \setminus X$, the algorithm eventually inserts Y in $\widehat{\mathbf{ADJ}}_X$. The algorithm removes Y from $\widehat{\mathbf{ADJ}}_X$ if and only if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ or $\widehat{\mathbf{ADJ}}_Y$ such that $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$ is concluded. Owing to assumption 2, the algorithm concludes $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$ if and only if $\text{test}_{X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}}$ is performed and $X \perp\!\!\!\perp Y \mid \mathbf{S}_{XY}$ holds. The latter is not the case if $Y \in \mathbf{PA}_X$, due to Lemma 2. Thus, $\mathbf{PA}_X$ is never removed from $\widehat{\mathbf{ADJ}}_X$. That is, the algorithm discovers all links in $\mathbb{G}$. ■

Lemma 15 *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* P *of the variables in some set* $\mathbf{V}$ *is a Bayesian network. Suppose further that a conditional-independence-test-based algorithm instantiating Algorithm Template 1 is applied on a sample from* P *and discovers a link* $X - Y$*. If*

1. $\mathbb{G}$ *and* P *are faithful to each other and*
2. *performed tests never make a type II error*

then for each subset $\mathbf{W}$ *of* $\mathbf{ADJ}_X$ *or* $\mathbf{ADJ}_Y$ *the algorithm considers* $\text{test}_{X \perp\!\!\!\perp Y \mid \mathbf{W}}$*.*

Proof The algorithm eventually inserts Y in $\widehat{\mathbf{ADJ}}_X$ and X in $\widehat{\mathbf{ADJ}}_Y$ and considers $\text{test}_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$ for each subset $\mathbf{Z}$ of $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$. $\mathbf{W}$ is never removed from $\widehat{\mathbf{ADJ}}_X$ if $\mathbf{W} \subseteq \mathbf{ADJ}_X$ or from $\widehat{\mathbf{ADJ}}_Y$ if $\mathbf{W} \subseteq \mathbf{ADJ}_Y$ due to Lemma 14. Thus, the algorithm eventually considers $\text{test}_{X \perp\!\!\!\perp Y \mid \mathbf{W}}$. ■

Lemma 16 *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* P *of the variables in some set* $\mathbf{V}$ *is a Bayesian network. Suppose further that* $\underline{\mathbf{C}}$ *is a set of sets* $\mathbf{C}$ *such that* $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ *for* $X, Y \in \mathbf{V}$*. Let* $H_{\exists\, \mathbf{S} \in \underline{\mathbf{C}} \text{ s.t. } X \perp Y \mid \mathbf{S}}$ *be the hypothesis that there is a set* $\mathbf{S}$ *in* $\underline{\mathbf{C}}$ *such that* $X \perp Y \mid \mathbf{S}$*. Let further* $\underline{\mathbf{S}}$ *be the subset of* $\underline{\mathbf{C}}$ *such that for each* $\mathbf{S} \in \underline{\mathbf{S}}$*,* $H_{X \perp Y \mid \mathbf{S}}$ *would be true if* $H_{\exists\, \mathbf{S} \in \underline{\mathbf{C}} \text{ s.t. } X \perp Y \mid \mathbf{S}}$ *was true. If there is a set* $\mathbf{W} \in \underline{\mathbf{S}}$ *for which* $X \perp\!\!\!\perp Y \mid \mathbf{W} \implies X \perp Y \mid \mathbf{W}$ *holds, then the p-value*

$p_{\exists\, \mathbf{S}\in\underline{\mathbf{C}}\ s.t.\ X\perp Y|\mathbf{S}}$ *of* $H_{\exists\, \mathbf{S}\in\underline{\mathbf{C}}\ s.t.\ X\perp Y|\mathbf{S}}$ *is upper-bounded by the maximal among the p-values* $p_{X\perp\!\!\!\perp Y|\mathbf{C}}$ *for* $\mathbf{C}\in\underline{\mathbf{C}}$*:*

$$p_{\exists\, \mathbf{S}\in\underline{\mathbf{C}}\ s.t.\ X\perp Y|\mathbf{S}} \leq \max_{\mathbf{C}\in\underline{\mathbf{C}}} p_{X\perp\!\!\!\perp Y|\mathbf{C}}$$

Proof Owning to the assumption and Theorem 1, there is a set $\mathbf{W}\in\underline{\mathbf{S}}$ for which $X\perp\!\!\!\perp Y \mid \mathbf{W} \iff X\perp Y \mid \mathbf{W}$ holds; let $\underline{\mathbf{F}}$ be the set of such sets. Let $E_{\mathbb{G}0}$, $E^i_{\mathbb{G}0}$, E^i_{P0} and $E^i_{\mathbb{G}1}$ denote the event that $\mathrm{H}_{\exists\, \mathbf{S}\in\underline{\mathbf{C}}\ \mathrm{s.t.}\ X\perp Y|\mathbf{S}}$, $\mathrm{H}_{X\perp Y|\mathbf{C}}$, $\mathrm{H}_{X\perp\!\!\!\perp Y|\mathbf{C}}$ and $\mathrm{H}_{\neg X\perp Y|\mathbf{C}}$ is true, respectively. Without loss of generality, assume that for $i\leq k_2$, $\mathbf{C}\in\underline{\mathbf{S}}$ and for $i\leq k_1\leq k_2$, $\mathbf{C}\in\underline{\mathbf{F}}$. Let $p=\mathrm{p}_{\exists\, \mathbf{S}\in\mathbf{C}\ \mathrm{s.t.}\ X\perp\!\!\!\perp Y|\mathbf{S}}$, $T_i=\mathrm{T}_{X\perp\!\!\!\perp Y|\mathbf{C}}$, $t_i=\mathrm{t}_{X\perp\!\!\!\perp Y|\mathbf{C}}$, $p_i=\mathrm{p}_{X\perp\!\!\!\perp Y|\mathbf{C}}$.

$$\begin{aligned}
p &= \Pr\Big(\bigcap_i T_i\geq t_i \Big| E_{\mathbb{G}0}\Big) = \prod_i \Pr\Big(T_i\geq t_i \Big| \Big\{\bigcap_{j<i} T_j\geq t_j\Big\}\cap E_{\mathbb{G}0}\Big)\\
&= \prod_i \Big[\Pr\Big(T_i\geq t_i \Big| \Big\{\bigcap_{j<i} T_j\geq t_j\Big\}\cap E_{\mathbb{G}0}\cap E^i_{\mathbb{G}0}\Big)+\\
&\qquad \Pr\Big(T_i\geq t_i \Big| \Big\{\bigcap_{j<i} T_j\geq t_j\Big\}\cap E_{\mathbb{G}0}\cap E^i_{\mathbb{G}1}\Big)\Big]\\
&\leq \prod_{i\leq k_2} \Pr\Big(T_i\geq t_i \Big| \Big\{\bigcap_{j<i} T_j\geq t_j\Big\}\cap E^i_{\mathbb{G}0}\Big) \leq \prod_{i\leq k_1} \Pr\Big(T_i\geq t_i \Big| \Big\{\bigcap_{j<i} T_j\geq t_j\Big\}\cap E^i_{P0}\Big)\\
&\leq \Pr\Big(T_1\geq t_1 \Big| E^1_{P0}\Big) = p_1 \leq \max_i p_i
\end{aligned}$$

The second equality is due to the chain rule for random variables. The third equality is due to the law of total probability. The first inequality is due to $E^i_{\mathbb{G}0}\subseteq E_{\mathbb{G}0}$ for $i\leq k$. ■

A version of Lemma 16, with a similar proof, is found in Tsamardinos and Brown (2008) and concerns the upper-bounding of the p-value corresponding to a hypothesis related to $\mathrm{H}_{\neg\mathrm{Adj}(X,Y)}$. Now we give the proofs of Theorem 5, 7 and 4, in this order:

Proof of Theorem 5. Owing to assumption 1, $\mathrm{H}_{\neg\mathrm{Adj}(X,Y)}$ and $\mathrm{H}_{\exists\, \mathbf{S}\in\underline{\mathbf{B}}_{XY}\ \mathrm{s.t.}\ X\perp Y|\mathbf{S}}$ are equivalent. The proof concludes due to assumption 2 and Lemma 16. ■

Proof of Theorem 7. The algorithm considers $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{S}_{XY}}$ due to Lemma 15. Owing to assumption 2, $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{S}_{XY}}$ is performed. The proof concludes due to Theorem 5. ■

Proof of Theorem 4. Owing to Lemma 15, the algorithm considers $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{PA}^0_X}$ and $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{PA}^0_Y)}$, where:

- $\mathbf{PA}^0_X=\mathbf{PA}_X\setminus Y$ and $\mathbf{PA}^0_Y=\mathbf{PA}_Y$, if $Y\in\mathbf{PA}_X$
- $\mathbf{PA}^0_X=\mathbf{PA}_X$ and $\mathbf{PA}^0_Y=\mathbf{PA}_Y\setminus X$, if $X\in\mathbf{PA}_Y$
- $\mathbf{PA}^0_X=\mathbf{PA}_X$ and $\mathbf{PA}^0_Y=\mathbf{PA}_Y$, if X and Y are not adjacent in $\mathbb{G}$

Owing to Corollary 3, either $\mathrm{H}_{X\perp Y|\mathbf{PA}^0_X}$ or $\mathrm{H}_{X\perp Y|\mathbf{PA}^0_Y}$ would be true if $\mathrm{H}_{\neg\mathrm{Adj}(X,Y)}$ was true. Owing to assumption 2, both $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{PA}_X}$ and $\mathrm{test}_{X\perp\!\!\!\perp Y|\mathbf{PA}_Y}$ are performed. The proof

concludes due to Theorem 5. ■

Appendix B

B.1 Comparison of False Discovery Rate control approaches

In this section, we present the results of the experimental comparison of our approach to FDR control to that of Li and Wang (2009). We refer to the single-stage approach of Li and Wang (2009) and our two-stage approach, both using the PC-skeleton algorithm, as *1-stage* and *2-stage*, respectively.

The bias (Table B.1) and power (Table B.2) for 1-stage and 2-stage are similar in all cases. Expected execution time (Table B.3) varies among networks, sample sizes and FDR control approaches, but the mean expected execution time of 1-stage and 2-stage is similar.

Network	n	1-stage	2-stage
Alarm	100	-0.6645	-0.6547
	1000	0.02194	0.02134
	1e+04	0.05	0.05
Andes	100	-0.01512	0.006977
	1000	0.04177	0.04354
	1e+04	0.02619	0.03172
Barley	100	-0.874	-0.874
	1000	-0.7027	-0.7012
	1e+04	-0.4245	-0.4182
Hailfinder	100	-0.8155	-0.8156
	1000	-0.4324	-0.4329
	1e+04	-0.5802	-0.5787
Hepar II	100	0.04148	0.05
	1000	0.04858	0.05
	1e+04	0.04969	0.05
Insurance	100	-0.3514	-0.3435
	1000	-0.002402	-0.001155
	1e+04	0.04356	0.04903
Mildew	100	-0.8485	-0.8485
	1000	-0.7864	-0.7865
	1e+04	-0.6386	-0.6375
Power Plant	100	-0.8322	-0.8322
	1000	-0.04867	-0.007802
	1e+04	-0.004295	0.04312
Water	100	-0.2183	-0.2183
	1000	0.04947	0.05
	1e+04	0.01335	0.04773
Win95pts	100	-0.06148	-0.01006
	1000	-0.02198	0.01303
	1e+04	0.04191	0.04816
Mean FDR undercontrol		0.2774	0.272
Mean FDR overcontrol		0.01426	0.01849

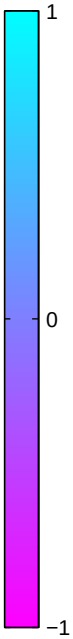


Table B.1: procbias($\widehat{\text{FDR}}_{\text{BY}}, 0.05$): bias of the False Discovery Rate (FDR) controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold 0.05 (see main text for definition) for each network, sample size n and FDR control approach. *1-stage* and *2-stage* refer to the single-stage approach of Li and Wang (2009) and our two-stage approach, respectively, using the PC-skeleton algorithm. For 1-stage, the bias corresponds to the final application of the procedure. The bias for 1-stage and 2-stage is similar in all cases.

Network	n	1-stage	2-stage
Alarm	100	0.5574	0.5374
	1000	0.7622	0.7387
	1e+04	0.9335	0.9254
Andes	100	0.1214	0.06361
	1000	0.4961	0.4596
	1e+04	0.761	0.7541
Barley	100	0.9451	0.9451
	1000	0.8039	0.8001
	1e+04	0.8045	0.8055
Hailfinder	100	0.6912	0.6882
	1000	0.6927	0.6874
	1e+04	0.5152	0.5148
Hepar II	100	0.04203	0.02423
	1000	0.2254	0.1881
	1e+04	0.5062	0.4787
Insurance	100	0.4542	0.4442
	1000	0.6233	0.6173
	1e+04	0.79	0.779
Mildew	100	0.9896	0.9896
	1000	0.8287	0.8283
	1e+04	0.913	0.9115
Power Plant	100	0.9505	0.9505
	1000	0.4719	0.47
	1e+04	0.5845	0.5321
Water	100	0.242	0.242
	1000	0.2668	0.2664
	1e+04	0.3992	0.3917
Win95pts	100	0.1059	0.08277
	1000	0.3313	0.2788
	1e+04	0.5985	0.5726
	Mean:	0.5802	0.5656

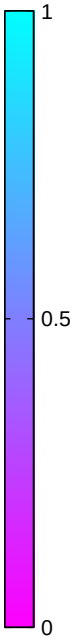


Table B.2: procpower($\widehat{\text{FDR}}_{\text{BY}}, 0.05$): power of the False Discovery Rate (FDR) controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold 0.05 (see main text for definition) for each network, sample size n and FDR control approach. *1-stage* and *2-stage* refer to the single-stage approach of Li and Wang (2009) and our two-stage approach, respectively, using the PC-skeleton algorithm. For 1-stage, the power corresponds to the final application of the procedure. The power of 1-stage and 2-stage is similar in all cases.

Network	n	1-stage	2-stage
Alarm	100	1.40	2.13
	1000	2.83	4.29
	1e+04	4.44	7.43
Andes	100	161.09	96.18
	1000	217.78	150.93
	1e+04	301.60	252.80
Barley	100	2.54	2.47
	1000	6.61	6.99
	1e+04	23.32	29.68
Hailfinder	100	5.16	4.43
	1000	8.76	9.03
	1e+04	56.79	73.56
Hepar II	100	8.76	6.36
	1000	10.79	11.21
	1e+04	69.36	156.28
Insurance	100	1.41	1.58
	1000	4.36	5.41
	1e+04	14.91	18.13
Mildew	100	0.96	0.87
	1000	2.52	2.51
	1e+04	7.53	9.43
Power Plant	100	4.73	4.02
	1000	8.90	7.93
	1e+04	11.14	11.67
Water	100	1.63	1.32
	1000	2.10	2.17
	1e+04	4.06	4.57
Win95pts	100	11.12	9.82
	1000	15.13	15.41
	1e+04	26.54	30.56
	Mean:	33.28	31.31

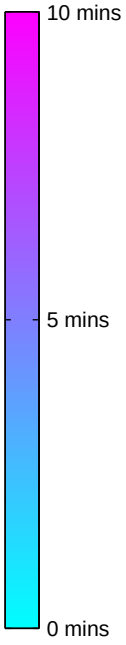


Table B.3: Expected execution time (seconds) for each network, sample size n and FDR control approach. *1-stage* and *2-stage* refer to the single-stage approach of Li and Wang (2009) and our two-stage approach, respectively, using the PC-skeleton algorithm. Execution time varies among networks, sample sizes, and FDR control approaches. The longest times are observed on Andes, which is the largest among the networks. The mean expected execution time of 1-stage and 2-stage is similar.

B.2 Comparison of False Discovery Rate estimators

In this section, we present the results of the experimental comparison of $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ FDR estimators (see main text for definitions).

Mean underestimation (Table B.4) by $\widehat{\text{FDR}}_{\text{BY}}$ and $\widehat{\text{FDR}}_{\text{ST}}$ is large in one case each, while mean underestimation by $\widehat{\text{FDR}}_{\text{PB}}$ is small. Mean overestimation (Table B.5) by $\widehat{\text{FDR}}_{\text{BY}}$ is pretty large to very large in all but one case, while mean underestimation by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{ST}}$ is small. On average, $\widehat{\text{FDR}}_{\text{PB}}$ achieves the lowest mean underestimation, followed by $\widehat{\text{FDR}}_{\text{BY}}$ and $\widehat{\text{FDR}}_{\text{ST}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ achieves the lowest mean overestimation, followed by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{BY}}$.

Mean undercontrol (Table B.6) by $\widehat{\text{FDR}}_{\text{BY}}$ is very large in many cases, while mean undercontrol by $\widehat{\text{FDR}}_{\text{PB}}$ or $\widehat{\text{FDR}}_{\text{ST}}$ is much lower in these cases. However, the mean power (Table B.8) of the FDR controlling procedure with $\widehat{\text{FDR}}_{\text{PB}}$ or $\widehat{\text{FDR}}_{\text{ST}}$ is much lower than by $\widehat{\text{FDR}}_{\text{BY}}$. Mean overcontrol (Table B.7) by all estimators is small in all cases. On average, $\widehat{\text{FDR}}_{\text{PB}}$ achieves the lowest mean undercontrol, followed by $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{ST}}$ achieves the lowest mean overcontrol, followed by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{BY}}$, and $\widehat{\text{FDR}}_{\text{BY}}$ achieves the highest mean power, followed by $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{PB}}$.

As n increases, $\widehat{\text{FDR}}_{\text{PB}}$ estimation time increases on all networks and $\widehat{\text{FDR}}_{\text{ST}}$ estimation time increases on all but four networks (Table B.9). The longest estimation times are observed on Andes, which is the largest among the networks. On average, $\widehat{\text{FDR}}_{\text{PB}}$ is faster to compute than $\widehat{\text{FDR}}_{\text{ST}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0.1052	0.09543	0.06849
	1000	0	0.01467	0.001801
	1e+04	0	8.652e-07	0
Andes	100	0	0	0.03902
	1000	0	0	0.003225
	1e+04	0	7.919e-05	0
Barley	100	0	0.4312	0.4629
	1000	0.1608	0.0676	0.02181
	1e+04	0.03443	0.09145	0.09755
Hailfinder	100	0.01124	0.09217	0.1518
	1000	0.0173	0.1402	0.1032
	1e+04	0	0.02185	0
Hepar II	100	0	0	0.02212
	1000	0	5.235e-07	5.057e-06
	1e+04	0	4.331e-05	0
Insurance	100	0.04299	0.03324	0.02559
	1000	0.0001046	0.03026	0
	1e+04	0	0.0008426	0
Mildew	100	0	6.202e-05	0
	1000	0.06317	0.0005779	0
	1e+04	0.1485	0.07372	0.05302
Power Plant	100	0.4044	0.09247	0.08811
	1000	0	0.001845	0.009285
	1e+04	0	0.007761	0.0005261
Water	100	0.004804	0.09371	0.07805
	1000	0	0.01918	0.01776
	1e+04	0	0.003009	0.0001807
Win95pts	100	0	0.008278	0.06997
	1000	0	0.01213	0.01238
	1e+04	0	0.002153	0
	Mean:	0.0331	0.04446	0.04423

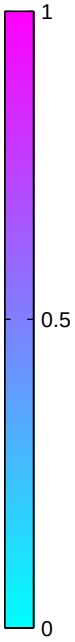


Table B.4: Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$: mean underestimation of the False Discovery Rate (FDR) by FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0, \alpha]$ for each network, sample size n and FDR estimator $\widehat{\text{FDR}}$. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. Mean underestimation by $\widehat{\text{FDR}}_{\text{BY}}$ and $\widehat{\text{FDR}}_{\text{ST}}$ is large while mean underestimation by $\widehat{\text{FDR}}_{\text{PB}}$ is small for $n = 100$ on Power Plant and Barley, respectively. On average, $\widehat{\text{FDR}}_{\text{PB}}$ achieves the lowest mean underestimation, followed by $\widehat{\text{FDR}}_{\text{BY}}$ and $\widehat{\text{FDR}}_{\text{ST}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0.152	0	0
	1000	0.8816	0	0
	1e+04	0.8968	0.0006242	0.005155
Andes	100	0.7588	0.07483	0
	1000	0.914	0.01442	0.0007701
	1e+04	0.95	0.001123	0.006759
Barley	100	0.09949	0	0
	1000	0.09872	0	0
	1e+04	0.3559	0	0
Hailfinder	100	0.2522	0	0
	1000	0.4116	0	0
	1e+04	0.9019	0	0.01729
Hepar II	100	0.5626	0.1041	9.432e-05
	1000	0.8487	0.04026	0.01012
	1e+04	0.9104	0.001581	0.004515
Insurance	100	0.2463	0	0
	1000	0.7479	0	0.03385
	1e+04	0.8035	2.41e-06	0.0613
Mildew	100	0.5928	0.01547	0.05951
	1000	0.2505	0.03127	0.1355
	1e+04	0.1306	0	0
Power Plant	100	7.806e-05	0	0
	1000	0.916	2.39e-05	0
	1e+04	0.9362	7.203e-07	0.0003907
Water	100	0.471	0	0
	1000	0.8984	0	0
	1e+04	0.8973	0	0.0008545
Win95pts	100	0.854	0.01087	0
	1000	0.9278	0	0
	1e+04	0.9447	3.338e-05	0.03231
	Mean:	0.6204	0.009822	0.01228

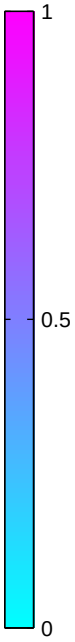


Table B.5: Mean oe($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$: mean overestimation of the False Discovery Rate (FDR) by FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0, \alpha]$ for each network, sample size n and FDR estimator $\widehat{\text{FDR}}$. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. Mean overestimation by $\widehat{\text{FDR}}_{\text{BY}}$ is pretty large to very large in all cases except on Power Plant for $n = 100$, where mean underestimation is large (Table B.4). On the contrary, mean underestimation by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{ST}}$ is small. On average, $\widehat{\text{FDR}}_{\text{ST}}$ achieves the lowest mean overestimation, followed by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{BY}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0.6488	0.1365	0.1142
	1000	0.00421	0.003392	2.773e-05
	1e+04	0	4.329e-06	1.819e-06
Andes	100	0	0.000365	0.02236
	1000	0	0	0.001816
	1e+04	0	5.227e-05	0
Barley	100	0.9281	0.9267	0.9247
	1000	0.6575	0.2416	0.01392
	1e+04	0.4106	0.06723	0.06227
Hailfinder	100	0.4039	0.1091	0.08242
	1000	0.4294	0.2566	0.2376
	1e+04	0	0.0241	0
Hepar II	100	0	0.001555	0.01368
	1000	0	2.181e-05	0.0004021
	1e+04	0	8.331e-05	0
Insurance	100	0.3448	0.04508	0.003066
	1000	0.007042	0.02125	4.09e-05
	1e+04	0	0.000111	0
Mildew	100	0	0.001872	0.0005017
	1000	0.5408	0.0007617	0.001855
	1e+04	0.6366	0	0
Power Plant	100	0.8224	0.0682	0.006263
	1000	0.001262	0.005997	0.008787
	1e+04	0	0.004275	0.001304
Water	100	0.213	0.06272	0.05849
	1000	0	0.01396	0.01004
	1e+04	0	0.0006721	0.0001841
Win95pts	100	0.01675	0.02908	0.06938
	1000	0.0016	0.008797	0.008602
	1e+04	0	0.000983	0
	Mean:	0.2022	0.0677	0.05473

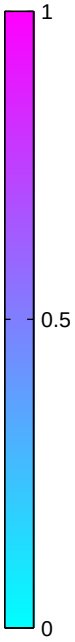


Table B.6: Mean uc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean "undercontrol" of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0.001, 0.1]$ for each network, sample size n and $\widehat{\text{FDR}}$. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. Mean undercontrol by $\widehat{\text{FDR}}_{\text{BY}}$ is very large on Alarm for $n = 100$, on Barley for all n, on Hailfinder for $n \in \{100, 1000\}$, on Insurance for $n = 100$, on Mildew for $n \in \{1000, 10000\}$, and on Power Plant and Water for $n = 100$, while mean undercontrol by $\widehat{\text{FDR}}_{\text{PB}}$ or $\widehat{\text{FDR}}_{\text{ST}}$ (except on Barley for $n = 100$) is much lower in these cases. On average, $\widehat{\text{FDR}}_{\text{PB}}$ achieves the lowest mean undercontrol, followed by $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{BY}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0	0	0
	1000	0.02539	0.02232	0.02448
	1e+04	0.04999	0.04337	0.044
Andes	100	0.04857	0.009284	0
	1000	0.04979	0.009178	0.0008158
	1e+04	0.04997	0.008754	0.01265
Barley	100	0	0	0
	1000	0	0	0.003748
	1e+04	0	0	0
Hailfinder	100	0	0	0
	1000	0	0	0
	1e+04	0.04939	0	0.01548
Hepar II	100	0.04749	0.008358	5.584e-06
	1000	0.04963	0.01279	0.001191
	1e+04	0.04994	0.001529	0.003958
Insurance	100	0	0	0.01725
	1000	0.01177	0	0.02052
	1e+04	0.04953	0.04197	0.04649
Mildew	100	0.04875	0.003035	0.003817
	1000	0	0.02072	0.0153
	1e+04	0	0.04999	0.04999
Power Plant	100	0	0	0.01984
	1000	0.03415	0.002614	0.002043
	1e+04	0.04999	0.01271	0.01409
Water	100	0	0	0
	1000	0.0494	0.004102	0.005178
	1e+04	0.04954	0.03437	0.03598
Win95pts	100	0.009272	0	0
	1000	0.03383	0.002758	0.003168
	1e+04	0.04983	0.02052	0.0323
	Mean:	0.02687	0.01028	0.01241

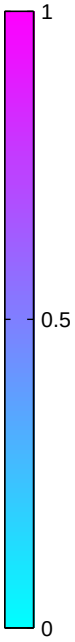


Table B.7: Mean $\text{oc}(\widehat{\text{FDR}}, q)$ for $q \in [0.001, 0.1]$: mean overcontrol of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0.001, 0.1]$ for each network, sample size n and FDR estimator $\widehat{\text{FDR}}$. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. Mean overcontrol by all estimators is small in all cases. On average, $\widehat{\text{FDR}}_{\text{ST}}$ achieves the lowest mean overcontrol, followed by $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{BY}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0.5299	0.07428	0.06773
	1000	0.7251	0.8061	0.7645
	1e+04	0.917	0.9385	0.9359
Andes	100	0.07387	0.1796	0.2036
	1000	0.449	0.5729	0.5766
	1e+04	0.7433	0.8064	0.7884
Barley	100	0.000569	0.0006917	0.000731
	1000	0.7559	0.09998	0.07354
	1e+04	0.7933	0.3323	0.3414
Hailfinder	100	0.3604	0.1851	0.1822
	1000	0.6711	0.1289	0.1177
	1e+04	0.5095	0.5101	0.4882
Hepar II	100	0.02715	0.05817	0.06185
	1000	0.1878	0.2773	0.2843
	1e+04	0.4771	0.5751	0.5702
Insurance	100	0.4167	0.0631	0.08576
	1000	0.6081	0.6391	0.5555
	1e+04	0.7733	0.8117	0.6908
Mildew	100	0.08429	0.108	0.1064
	1000	0.6115	0.04304	0.04304
	1e+04	0.785	0	0
Power Plant	100	0.9377	0.042	0.002829
	1000	0.4862	0.4534	0.5026
	1e+04	0.5301	0.6181	0.6092
Water	100	0.2373	0.1578	0.1752
	1000	0.2633	0.3045	0.3046
	1e+04	0.3963	0.4775	0.4753
Win95pts	100	0.1026	0.1487	0.1731
	1000	0.2911	0.4191	0.4178
	1e+04	0.5433	0.6657	0.614
	Mean:	0.4763	0.3499	0.3404

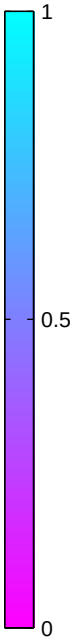


Table B.8: Mean procpower($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean power of the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0.001, 0.1]$ for each network, sample size n and FDR estimator $\widehat{\text{FDR}}$. $\widehat{\text{FDR}}_{\text{BY}}$, $\widehat{\text{FDR}}_{\text{PB}}$, and $\widehat{\text{FDR}}_{\text{ST}}$ denote the FDR estimator by Benjamini and Yekutieli (2001), the parametric-bootstrap-based FDR estimator, and the FDR estimator by Storey and Tibshirani (2001), respectively. Mean power by $\widehat{\text{FDR}}_{\text{PB}}$ or $\widehat{\text{FDR}}_{\text{ST}}$ is much lower than by $\widehat{\text{FDR}}_{\text{BY}}$ on Alarm for $n = 100$, on Barley for $n \in \{1000, 10000\}$, on Hailfinder for $n \in \{100, 1000\}$, on Insurance for $n = 100$, on Mildew for $n \in \{1000, 10000\}$, and on Power Plant and Water for $n = 100$. On average, $\widehat{\text{FDR}}_{\text{BY}}$ achieves the highest mean power, followed by $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\text{FDR}}_{\text{PB}}$.

Network	n	$\widehat{\text{FDR}}_{\text{BY}}$	$\widehat{\text{FDR}}_{\text{PB}}$	$\widehat{\text{FDR}}_{\text{ST}}$
Alarm	100	0.00	45.56	138.89
	1000	0.00	70.18	71.31
	1e+04	0.00	158.30	140.96
Andes	100	0.00	947.14	2302.50
	1000	0.00	1590.61	2639.65
	1e+04	0.00	5633.94	8810.07
Barley	100	0.00	35.31	41.99
	1000	0.00	85.84	704.62
	1e+04	0.00	316.07	1476.47
Hailfinder	100	0.00	78.12	251.07
	1000	0.00	127.98	298.52
	1e+04	0.00	241.35	257.19
Hepar II	100	0.00	128.92	217.12
	1000	0.00	171.23	234.68
	1e+04	0.00	1081.95	1134.27
Insurance	100	0.00	24.08	68.08
	1000	0.00	52.95	71.82
	1e+04	0.00	182.94	221.92
Mildew	100	0.00	22.37	31.37
	1000	0.00	40.37	122.75
	1e+04	0.00	162.52	520.27
Power Plant	100	0.00	93.51	867.09
	1000	0.00	98.36	110.30
	1e+04	0.00	210.91	155.85
Water	100	0.00	27.59	47.20
	1000	0.00	37.96	39.00
	1e+04	0.00	111.86	73.06
Win95pts	100	0.00	150.07	186.35
	1000	0.00	199.95	268.27
	1e+04	0.00	487.09	610.07
	Mean:	0	420.5	737.1

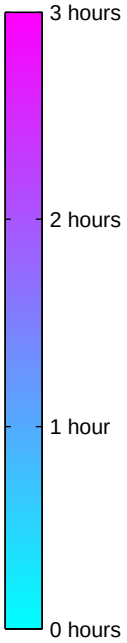


Table B.9: Expected False Discovery Rate (FDR) extra estimation time (seconds) for each network, sample size n and FDR estimator (see main text for definition). $\widehat{\text{FDR}}_{\text{PB}}$ and $\widehat{\text{FDR}}_{\text{ST}}$ denote the parametric-bootstrap-based FDR estimator and the FDR estimator by Storey and Tibshirani (2001), respectively. As n increases, $\widehat{\text{FDR}}_{\text{PB}}$ estimation time increases on all networks and $\widehat{\text{FDR}}_{\text{ST}}$ estimation time increases on all networks except Alarm, Hailfinder, Power Plant and Water. The longest estimation times are observed on Andes, which is the largest among the networks. On average, $\widehat{\text{FDR}}_{\text{PB}}$ estimation time is shorter than $\widehat{\text{FDR}}_{\text{ST}}$ estimation time.

B.3 Comparison of False Positive definitions

In this section, we present the results of the experimental comparison of False Positive definitions. We refer to the combination of the regular and the relaxed definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001) as *Reg* and *Rel*, respectively, and to the combination of he regular definition of false positive and the parametric-bootstrap-based FDR estimator as *Reg (PB)*.

Mean underestimation (Table B.11) is zero for Rel in all cases, while it is not for Reg or Reg (PB). However, mean overestimation (Table B.12) for Rel is larger than for Reg and much larger than for Reg (PB), on average. Mean undercontrol (Table B.13) for Rel is much smaller than for Reg or Reg (PB) and mean power (Table B.8) for Rel (same for Reg) is larger than for Reg (PB), on average. However, mean overcontrol (Table B.14) is larger for Rel than for Reg or Reg (PB), on average.

Network	n = 100	n = 1000	n = 10000
Alarm	0	1	3
Andes	2	5	8
Barley			0
Hailfinder		0	1
Hepar II	0	1	3
Insurance		1	3
Mildew			
Power Plant	0	2	4
Water	0	1	3
Win95pts	2	5	8
Mean:	0.6667	2.5	4.833

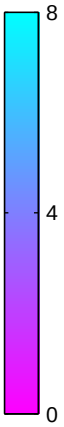


Table B.10: *worst-max-k*: worst-case maximal conditioning set cardinality (see main text for definition) for each network and sample size n. *worst-max-k* does not exist for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance, and for all n on Mildew.

Network	n	Reg	Reg (PB)	Rel
Alarm	100	0.1052	0.008478	0
	1000	0	0.007707	0
	1e+04	0	0	0
Andes	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Barley	100	0	0	
	1000	0.1608	0.0429	
	1e+04	0.03443	0	0
Hailfinder	100	0.01124	0.01446	
	1000	0.0173	0.06522	0
	1e+04	0	0	0
Hepar II	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Insurance	100	0.04299	0.008305	
	1000	0.0001046	0	0
	1e+04	0	0	0
Mildew	100	0	0.0001249	
	1000	0.06317	0	
	1e+04	0.1485	0.006219	
Power Plant	100	0.4044	0.002255	0
	1000	0	0	0
	1e+04	0	0	0
Water	100	0.004804	0.01949	0
	1000	0	2.583e-05	0
	1e+04	0	0	0
Win95pts	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
	Mean:	0.02462	0.004486	0

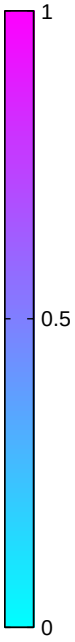


Table B.11: Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$: mean underestimation of the False Discovery Rate (FDR) by FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0, \alpha]$ for each network, sample size n and definition of false positive / FDR estimator combination. The relaxed definition of FDR (see main text) is not applicable for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance and for all n on Mildew (see Table B.10). *Reg* and *Rel* refer to the combination of the regular and the relaxed, respectively, definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001), while *Reg (PB)* refers to the combination of the regular definition of false positive with the parametric-bootstrap-based FDR estimator. Mean underestimation is zero for Rel in all cases, while it is not for Reg or Reg (PB).

Network	n	Reg	Reg (PB)	Rel
Alarm	100	0.152	0	0.7416
	1000	0.8816	0	0.9075
	1e+04	0.8968	0.01597	0.8968
Andes	100	0.7588	0.08494	0.7872
	1000	0.914	0.01882	0.914
	1e+04	0.95	0.0338	0.95
Barley	100	0.09949	0.0641	
	1000	0.09872	8.251e-07	
	1e+04	0.3559	0.03552	0.8255
Hailfinder	100	0.2522	0.001168	
	1000	0.4116	0	0.9115
	1e+04	0.9019	0.05595	0.9021
Hepar II	100	0.5626	0.09595	0.9003
	1000	0.8487	0.04246	0.8633
	1e+04	0.9104	0.01616	0.9106
Insurance	100	0.2463	1.53e-05	
	1000	0.7479	0.0161	0.828
	1e+04	0.8035	0.0313	0.8035
Mildew	100	0.5928	0.01448	
	1000	0.2505	0.03949	
	1e+04	0.1306	0	
Power Plant	100	7.806e-05	1.415e-05	0.4504
	1000	0.916	0.06564	0.916
	1e+04	0.9362	0.08939	0.9362
Water	100	0.471	0	0.8491
	1000	0.8984	0.005785	0.9285
	1e+04	0.8973	0.02808	0.898
Win95pts	100	0.854	0.07995	0.8986
	1000	0.9278	0.04739	0.9278
	1e+04	0.9447	0.07019	0.9447
	Mean:	0.7366	0.03624	0.8648

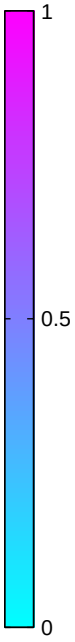


Table B.12: Mean $\mathrm{oe}(\widehat{\mathrm{FDR}}, t)$ for $t \in [0, \alpha]$: mean overestimation of the False Discovery Rate (FDR) by FDR estimator $\widehat{\mathrm{FDR}}$ (see main text for definition) in $[0, \alpha]$ for each network, sample size n and definition of false positive / FDR estimator combination. The relaxed definition of FDR (see main text) is not applicable for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance and for all n on Mildew (see Table B.10). *Reg* and *Rel* refer to the combination of the regular and the relaxed, respectively, definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001), while *Reg (PB)* refers to the combination of the regular definition of false positive with the parametric-bootstrap-based FDR estimator. Mean overestimation for Rel is noticeably larger than for Reg for $n = 100$ on Alarm, for $n = 10000$ on Barley, for $n = 1000$ on Hailfinder, for $n = 100$ on Hepar II, and for $n = 100$ on Power Plant and Water and much larger than for Reg (PB) in all cases. On average (without taking the cases where the relaxed definition of FDR is not applicable into account), mean overestimation for Rel is larger than for Reg and much larger than for Reg (PB).

Network	n	Reg	Reg (PB)	Rel
Alarm	100	0.6488	0.04162	0
	1000	0.00421	0.00252	0
	1e+04	0	0	0
Andes	100	0	2.422e-06	0
	1000	0	0	0
	1e+04	0	0	0
Barley	100	0.9281	0.02401	
	1000	0.6575	0.01957	
	1e+04	0.4106	0.0006011	0
Hailfinder	100	0.4039	0.0001545	
	1000	0.4294	0	0
	1e+04	0	0	0
Hepar II	100	0	0.001069	0
	1000	0	1.404e-05	0
	1e+04	0	0	0
Insurance	100	0.3448	0.0338	
	1000	0.007042	2.236e-06	0
	1e+04	0	0	0
Mildew	100	0	0.001872	
	1000	0.5408	0.0007617	
	1e+04	0.6366	0	
Power Plant	100	0.8224	0.04823	0
	1000	0.001262	0.003341	0.001262
	1e+04	0	0	0
Water	100	0.213	0.008094	0
	1000	0	3.404e-05	0
	1e+04	0	0	0
Win95pts	100	0.01675	0.0004328	4.091e-06
	1000	0.0016	0.0008659	0.0016
	1e+04	0	0	0
	Mean:	0.1111	0.004645	0.0001246

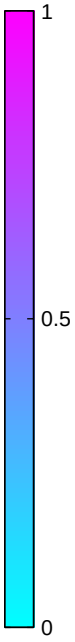


Table B.13: Mean uc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean "undercontrol" of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0.001, 0.1]$ for each network, sample size n and definition of false positive / FDR estimator combination. The relaxed definition of FDR (see main text) is not applicable for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance and for all n on Mildew (see Table B.10). *Reg* and *Rel* refer to the combination of the regular and the relaxed, respectively, definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001), while *Reg (PB)* refers to the combination of the regular definition of false positive with the parametric-bootstrap-based FDR estimator. Mean undercontrol for Rel is noticeably smaller than for Reg for $n = 100$ on Alarm, for $n = 10000$ on Barley, for $n = 1000$ on Hailfinder, for $n = 100$ on Power Plant and Water and smaller than or equal to mean undercontrol for Reg (PB) in all cases except for $n = 1000$ on Win95pts. On average (without taking the cases where the relaxed definition of FDR is not applicable into account), mean undercontrol for Rel is much smaller than for Reg or Reg (PB).

Network	n	Reg	Reg (PB)	Rel
Alarm	100	0	0	0.0485
	1000	0.02539	0.02245	0.04999
	1e+04	0.04999	0.04433	0.04999
Andes	100	0.04857	0.01687	0.04857
	1000	0.04979	0.01263	0.04979
	1e+04	0.04997	0.02855	0.04997
Barley	100	0	0.00435	
	1000	0	0	
	1e+04	0	0.01813	0.04989
Hailfinder	100	0	0.005907	
	1000	0	0.04999	0.04999
	1e+04	0.04939	0.0405	0.04939
Hepar II	100	0.04749	0.008904	0.04999
	1000	0.04963	0.01481	0.04963
	1e+04	0.04994	0.01458	0.04994
Insurance	100	0	0	
	1000	0.01177	0.01092	0.04999
	1e+04	0.04953	0.04461	0.04953
Mildew	100	0.04875	0.003082	
	1000	0	0.02072	
	1e+04	0	0.04999	
Power Plant	100	0	0	0.04639
	1000	0.03415	0.02473	0.03415
	1e+04	0.04999	0.04568	0.04999
Water	100	0	0	0.04725
	1000	0.0494	0.007923	0.04999
	1e+04	0.04954	0.04015	0.04954
Win95pts	100	0.009272	0.006673	0.04853
	1000	0.03383	0.0271	0.03383
	1e+04	0.04983	0.04428	0.04983
	Mean:	0.03293	0.02277	0.04803

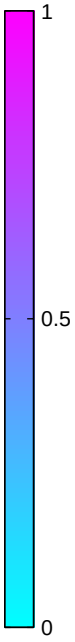


Table B.14: Mean oc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean overcontrol of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}$ (see main text for definition) in $[0.001, 0.1]$ for each network, sample size n and definition of false positive / FDR estimator combination. The relaxed definition of FDR (see main text) is not applicable for $n \in \{100, 1000\}$ on Barley, for $n = 100$ on Hailfinder and Insurance and for all n on Mildew (see Table B.10). *Reg* and *Rel* refer to the combination of the regular and the relaxed, respectively, definition of false positive with the FDR estimator by Benjamini and Yekutieli (2001), while *Reg (PB)* refers to the combination of the regular definition of false positive with the parametric-bootstrap-based FDR estimator. On average (without taking the cases where the relaxed definition of FDR is not applicable into account), mean overcontrol is larger for Rel than for Reg or Reg (PB).

Appendix C

In Section 4 of the main text, we identified and quantified the direct and indirect causes of false positives with non-upper-bounded p-value, namely the default decision, the employed reliability criterion, insufficient sample size to declare reliability, unfaithfulness, type II errors and close-to-unfaithfulness. Here we identify the following approaches to dealing with them:

- Default decision: reverse default decision
- Reliability criterion: replace reliability criterion with a less stringent one, decrease *h-ps* if the heuristic power rule is employed, or increase the significance level of the tests (if the reliability criterion depends on it) so more tests are attempted
- Insufficient sample size to declare reliability: relax the definition of false positive such that a falsely discovered link $X - Y$ is not considered a false positive if there is no sepset–subset of the neighbors of X or Y such that the corresponding test is reliable according to the employed reliability criterion (see Section 6 of the main text)
- Type II errors: increase the significance level or replace test with a more powerful one, so less type II errors are made, increase *h-ps* when using the heuristic power rule, replace reliability criterion with a more stringent one, or decrease the significance level of the tests (if the reliability criterion depends on it), so less tests are attempted

Obviously, insufficient sample size to declare reliability and close-to-unfaithfulness can be dealt with by increasing the sample size. However, we are interested in improving estimation and control of FDR when the sample size is fixed. Relaxing the definition of false positive such that false positives according to the relaxed definition have an upper-bounded p-value guaranteed under more realistic assumptions is a general approach to dealing with causes of false positives with non-upper-bounded p-value including insufficient sample size to declare reliability. This approach is presented in Section 6 of the main text. Finally, dealing with unfaithfulness and close-to-unfaithfulness for a fixed sample size is beyond the scope of this work (see Lemeire et al. (2012) for an algorithm for structure learning under unfaithfulness).

We evaluated varying the reliability criterion (see Section C.1) and increasing the significance level of the hypothesis tests (see Section C.2). We did not try reversing the default decision for unattempted tests given a nonempty set (that is, concluding independence instead of dependence), because this should increase false negatives, as also pointed out by Fast (2010). We also did not try reversing the default decision for unattempted tests given the empty set (that is, concluding dependence instead of independence) either due to the reason discussed by Tsamardinos et al. (2006, Section 11.1). Finally, we did not evaluate other tests (see Section C.3).

C.1 Varying the reliability criterion

The POWER correction (Fast et al., 2009) is an approach for controlling the power of the tests above a specified threshold $1 - \beta$, where β is the False Negative Rate threshold of the tests. The power of a conditional-independence test using a statistic that follows the χ^2

distribution is a function of the sample size n, the significance level α, the degrees of freedom *df* and the *effect size* w in the data. POWER is the first approach to address all four factors (Fast, 2010). w can either be specified in advance or estimated via cross-validation.

Fast (2010) first chooses values for w for sample size $n \in \{500, 1000, 2000, 5000\}$ using cross-validation with the *Diabetes*, *Hailfinder*, *Barley* and *Insurance* networks, randomly selected from the Bayesian Network Repository (see Appendix D). Then, using these values of w and power threshold $1 - \beta = 0.95$, he evaluates POWER along with three other reliability criteria (including a baseline one) with the *Alarm*, *Mildew*, *Pathfinder*, *Water* and *Win95pts* networks. He demonstrates that POWER is the only one to result in significance decreases (compared to the others, minus the baseline one) in false negatives, accompanied, however, with a significance increase in false positives.

Among the reliability criteria evaluated by Fast (2010) is what he calls the "rule of thumb", supposedly used by many structure learning algorithms. According to this criterion, a test is considered reliable if there are at least five observations per *degree of freedom*, on average. This is, however, different from the criterion actually used in most constraint-based algorithms, which is the heuristic power rule. According to the latter, a test is considered reliable if there are at least *h-ps* (usually 5) observations per *cell of the contingency table*, on average. In contrast to POWER, the rule of thumb and the heuristic power rule do not maintain a constant level of power for all sample sizes (Fast, 2010).

We applied MMPC–skeleton to the samples from Alarm, Andes, Hepar II, Mildew, Power Plant, Water, and Win95pts, each time using the first $n \in \{500, 1000, 5000\}$ observations and the heuristic power rule, rule of thumb and POWER reliability criteria. We used $\beta = 0.05$ with POWER. For the rest parameters we used the same values as in Section 3.3 of the main text. We did not use Hailfinder, Barley and Insurance because they were used in the calculation of w values by Fast (2010), as mentioned above. We used $n = 500$ and $n = 5000$ this time instead of $n = 100$ and $n = 10000$ because Fast (2010) did not calculate w for the latter sample sizes. Skeleton identification from the samples of Mildew using the rule of thumb was aborted because it was taking too much time.

For each network and sample size, we approximated mean FDR, power, and FDR under- and overestimation in $[0, 0.05]$ (Tables C.1 – C.4). We also approximated mean FDR under- and overcontrol, and FDR-controlling-procedure power in $[0.001, 0.1]$ (Tables C.5 – C.7). Finally, we approximated the expected skeleton identification time (Table C.8). Table C.9 summarizes the results. Compared to Heuristic, POWER achieves, on average, greater mean power and less mean overestimation and overcontrol, but greater mean FDR, underestimation and undercontrol and Thumb achieves, on average, less mean FDR, underestimation and undercontrol, but less mean power and greater mean overestimation and overcontrol. Mean expected skeleton identification time is similar for all criteria. Since POWER achieves greater mean FDR, underestimation and undercontrol than the heuristic power rule and use of the rule of thumb may severely slow down skeleton identification, as on Mildew, we do not recommend POWER or the rule of thumb over the heuristic power rule.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.107	0.7761	0.007126
	1000	0.02877	0.7807	0.004259
	5000	0.002493	0.5056	0.002526
Andes	500	0.0951	0.09443	0.09544
	1000	0.06904	0.06805	0.06935
	5000	0.03708	0.03714	0.03708
Hepar II	500	0.1893	0.3045	0.1938
	1000	0.1267	0.2746	0.1282
	5000	0.0593	0.1209	0.05981
Mildew	500	0.5756	0.2162	
	1000	0.6177	0.2308	
	5000	0.7526	0.5921	
Power Plant	500	0.05202	0.8997	0.05223
	1000	0.04623	0.9026	0.04557
	5000	0.03048	0.1653	0.03048
Water	500	0.03519	0.5381	0.01569
	1000	0.03722	0.5569	0.01635
	5000	0.01018	0.4641	0.01017
Win95pts	500	0.07417	0.07418	0.07417
	1000	0.04597	0.0458	0.04597
	5000	0.02468	0.02434	0.02469
	Mean:	0.0595	0.3685	0.05072

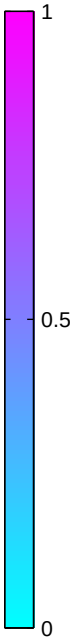


Table C.1: Mean FDR(t) for $t \in [0, \alpha]$: mean False Discovery Rate (FDR) in $[0, \alpha]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably less FDR for all n on Mildew but noticeably greater FDR for all n on Alarm, Hepar II, Power Plant, and Water and Thumb achieves noticeably less FDR for $n \in \{500, 1000\}$ on Alarm and Water. On average (without taking Mildew into account), Thumb achieves the least FDR, followed by Heuristic and POWER.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.7732	0.7574	0.6891
	1000	0.8205	0.8056	0.7708
	5000	0.9118	0.9632	0.8959
Andes	500	0.5024	0.5108	0.5001
	1000	0.5851	0.6006	0.5818
	5000	0.7543	0.755	0.7542
Hepar II	500	0.2453	0.3319	0.2377
	1000	0.3171	0.4371	0.3117
	5000	0.4984	0.6128	0.4937
Mildew	500	0.5136	0.06425	
	1000	0.6464	0.06425	
	5000	0.7269	0.1667	
Power Plant	500	0.4832	0.9849	0.4685
	1000	0.512	0.9851	0.5075
	5000	0.5565	0.7295	0.5565
Water	500	0.2766	0.211	0.2711
	1000	0.2934	0.2464	0.2903
	5000	0.398	0.5853	0.3971
Win95pts	500	0.3349	0.3348	0.3349
	1000	0.4087	0.4117	0.4087
	5000	0.5747	0.5847	0.5747
	Mean:	0.5137	0.6027	0.5025

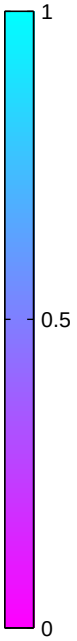


Table C.2: Mean power(t) for $t \in [0, \alpha]$: mean power in $[0, \alpha]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably greater power for $n = 5000$ on Alarm, for all n on Hepar II, for $n \in \{500, 1000\}$ on Power Plant, and for $n = 5000$ on Water, but noticeably less power for all n on Mildew and for $n \in \{500, 1000\}$ on Water. Thumb and Heuristic achieve similar power. On average (without taking Mildew into account), POWER achieves the greatest power, followed by Heuristic and Thumb.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.000254	0.2086	0
	1000	0	0.2324	0
	5000	0	0.04851	0
Andes	500	0	0	0
	1000	0	0	0
	5000	0	0	0
Hepar II	500	0	0	0
	1000	0	0	0
	5000	0	0	0
Mildew	500	0.03039	0	
	1000	0.06317	0	
	5000	0.1898	0.003201	
Power Plant	500	0	0.508	0
	1000	0	0.523	0
	5000	9.076e-05	2.97e-05	9.076e-05
Water	500	0	0.009609	0
	1000	0	0.01951	0
	5000	0	0.03846	0
Win95pts	500	0	0	0
	1000	0	0	0
	5000	0	0	0
	Mean:	1.915e-05	0.08823	5.042e-06

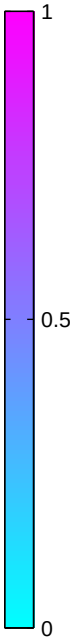


Table C.3: Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, \alpha]$: mean underestimation (see main text for definition) of the False Discovery Rate (FDR) by FDR estimator $\widehat{\text{FDR}}_{\text{BY}}$ by Benjamini and Yekutieli (2001) in $[0, \alpha]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably greater mean underestimation for all n on Alarm, for $n \in \{500, 1000\}$ on Power Plant, and for all n on Water. Thumb and Heuristic achieve the same mean underestimation in all cases except $n = 500$ on Alarm. On average (without taking Mildew into account), Thumb achieves the least mean underestimation, followed by Heuristic and POWER.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.8031	0.07064	0.9215
	1000	0.8816	0.05859	0.9139
	5000	0.8989	0.3424	0.9004
Andes	500	0.8895	0.89	0.8892
	1000	0.914	0.9146	0.9137
	5000	0.9436	0.9435	0.9436
Hepar II	500	0.7905	0.6686	0.7865
	1000	0.8487	0.6897	0.8478
	5000	0.9037	0.8325	0.9035
Mildew	500	0.321	0.7658	
	1000	0.2505	0.7512	
	5000	0.08953	0.3674	
Power Plant	500	0.9132	0	0.9144
	1000	0.916	0	0.9172
	5000	0.9304	0.7826	0.9304
Water	500	0.9033	0.3898	0.9234
	1000	0.8984	0.356	0.9197
	5000	0.9072	0.3525	0.9072
Win95pts	500	0.9031	0.9031	0.9031
	1000	0.9278	0.9279	0.9278
	5000	0.9418	0.9415	0.9418
	Mean:	0.8953	0.5591	0.9058

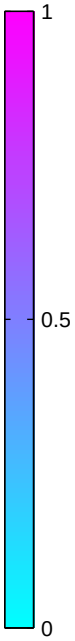


Table C.4: Mean $\mathrm{oe}(\widehat{\mathrm{FDR}}, t)$ for $t \in [0, \alpha]$: mean overestimation (see main text for definition) of the False Discovery Rate (FDR) by FDR estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ by Benjamini and Yekutieli (2001) in $[0, \alpha]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably less mean overestimation for all n on Alarm, for $n \in \{500, 1000\}$ on Hepar II and Power Plant, and for all n on Water. Thumb and Heuristic achieve similar mean overestimation in all cases. On average (without taking Mildew into account), POWER achieves the least mean overestimation, followed by Heuristic and Thumb.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.04273	0.7175	0
	1000	0.00421	0.7261	0
	5000	0	0.4449	0
Andes	500	0	0	0
	1000	0	0	0
	5000	0	0	0
Hepar II	500	0	0.02849	0
	1000	0	0.08576	0
	5000	0	0.01395	0
Mildew	500	0.4574	0	
	1000	0.5408	0	
	5000	0.6891	0.2889	
Power Plant	500	0.001977	0.8488	0.002232
	1000	0.001262	0.8537	0.001428
	5000	4.11e-06	0.07415	4.11e-06
Water	500	0	0.3435	0
	1000	0	0.4155	0
	5000	0	0.3808	0
Win95pts	500	0.009453	0.009464	0.009444
	1000	0.0016	0.001592	0.0016
	5000	2.37e-07	6.618e-08	2.37e-07
	Mean:	0.003402	0.2747	0.0008171

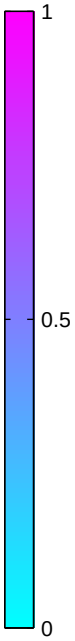


Table C.5: Mean uc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean undercontrol (see main text for definition) of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\text{FDR}}_{\text{BY}}$ by Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably less mean undercontrol for all n on Mildew but noticeably greater mean undercontrol for all n on Alarm, Hepar II, Power Plant, and Water and Thumb achieves noticeably less mean undercontrol for $n \in \{500, 1000\}$. On average (without taking Mildew into account), Thumb achieves the least mean undercontrol, followed by Heuristic and POWER.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.0001424	0	0.04973
	1000	0.02539	0	0.04974
	5000	0.04999	0	0.04999
Andes	500	0.04997	0.04974	0.04998
	1000	0.04979	0.0499	0.04989
	5000	0.04998	0.04998	0.04998
Hepar II	500	0.04999	0.0001871	0.04999
	1000	0.04963	0	0.04964
	5000	0.04999	0.009673	0.04999
Mildew	500	0	0.04873	
	1000	0	0.04863	
	5000	0	0	
Power Plant	500	0.03233	0	0.03146
	1000	0.03415	0	0.03358
	5000	0.0482	0	0.0482
Water	500	0.04981	0	0.04999
	1000	0.0494	0	0.04999
	5000	0.04845	0	0.04846
Win95pts	500	0.01618	0.0162	0.01619
	1000	0.03383	0.03388	0.03383
	5000	0.04796	0.048	0.04796
	Mean:	0.04084	0.01431	0.04492

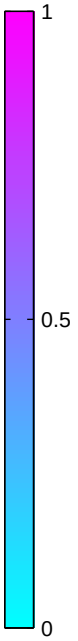


Table C.6: Mean $\text{oc}(\widehat{\text{FDR}}, q)$ for $q \in [0.001, 0.1]$: mean overcontrol (see main text for definition) of the False Discovery Rate (FDR) by the FDR controlling procedure of Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably less mean overcontrol for all n on Alarm, for $n = 1000$ on Andes, and for all n on Hepar II, Power Plant and Water, but noticeably greater mean overcontrol for $n \in \{500, 1000\}$ on Mildew. Thumb and Heuristic achieve similar mean overcontrol in all cases except $n \in \{500, 1000\}$ on Alarm. On average (without taking Mildew into account), POWER achieves the least mean overcontrol, followed by Heuristic and Thumb.

Network	n	Heuristic	POWER	Thumb
Alarm	500	0.6492	0.7091	0.5188
	1000	0.7251	0.781	0.6714
	5000	0.8936	0.961	0.8751
Andes	500	0.349	0.3552	0.3461
	1000	0.449	0.4728	0.4461
	5000	0.6741	0.6798	0.6726
Hepar II	500	0.1218	0.1844	0.1148
	1000	0.1878	0.2919	0.1777
	5000	0.3863	0.512	0.382
Mildew	500	0.4765	0.06457	
	1000	0.6115	0.06457	
	5000	0.7086	0.1559	
Power Plant	500	0.401	0.9897	0.3748
	1000	0.4862	0.99	0.4651
	5000	0.5282	0.5762	0.5282
Water	500	0.2518	0.1532	0.2507
	1000	0.2633	0.2098	0.2633
	5000	0.3264	0.5328	0.3266
Win95pts	500	0.2117	0.2116	0.2119
	1000	0.2911	0.292	0.2911
	5000	0.4569	0.471	0.4568
	Mean:	0.4251	0.5208	0.4096

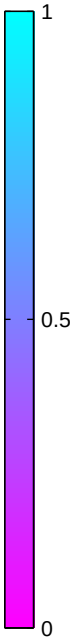


Table C.7: Mean procpower($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean power of the False Discovery Rate (FDR) controlling procedure of Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and reliability criterion. $\alpha = 0.05$ is the significance level of the underlying hypothesis tests. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. Compared to Heuristic, POWER achieves noticeably greater power for all n on Alarm and Hepar II, for $n \in \{500, 1000\}$ on Power Plant, and for $n = 5000$ on Water, but noticeably less power for all n on Mildew and for $n \in \{500, 1000\}$ on Water. Thumb and Heuristic achieve similar power in all cases except $n \in \{500, 1000\}$ on Alarm. On average (without taking Mildew into account), POWER achieves the greatest power, followed by Heuristic and Thumb.

Network	n	Heuristic	POWER	Thumb
Alarm	500	3.36	4.74	3.42
	1000	3.83	5.12	3.83
	5000	4.81	6.74	5.35
Andes	500	84.98	126.34	87.36
	1000	146.84	127.70	145.55
	5000	253.89	205.63	234.72
Hepar II	500	12.45	11.62	11.62
	1000	14.41	12.45	12.48
	5000	34.86	21.86	38.42
Mildew	500	2.43	1.50	
	1000	2.94	1.50	
	5000	6.63	1.87	
Power Plant	500	7.34	8.89	5.13
	1000	7.91	9.19	4.91
	5000	9.05	8.56	6.55
Water	500	2.51	1.76	1.68
	1000	2.83	1.80	1.85
	5000	3.89	2.92	2.89
Win95pts	500	15.77	14.94	10.28
	1000	17.33	16.73	11.60
	5000	27.01	25.83	18.74
	Mean:	36.28	34.05	33.69

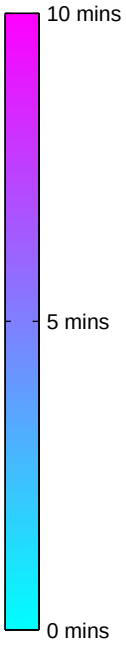


Table C.8: Expected skeleton identification time (seconds) for each network, sample size n and reliability criterion. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. On average (without taking Mildew into account), Thumb achieves the least time, followed by POWER and Heuristic.

	Heuristic	POWER	Thumb
Mean FDR(t) for $t \in [0, \alpha]$	0.0595	0.3685	0.05072
Mean power(t) for $t \in [0, \alpha]$	0.5137	0.6027	0.5025
Mean ue($\widehat{\text{FDR}}_{\text{BY}}, t$) for $t \in [0, \alpha]$	1.915e-05	0.08823	5.042e-06
Mean oe($\widehat{\text{FDR}}_{\text{BY}}, t$) for $t \in [0, \alpha]$	0.8953	0.5591	0.9058
Mean uc($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.003402	0.2747	0.0008171
Mean oc($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.04084	0.01431	0.04492
Mean procpower($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.4251	0.5208	0.4096
E[time] (s)	36.28	34.05	33.69

Table C.9: Summary of the results of the comparison of reliability criteria. Heuristic, POWER, and Thumb denote the heuristic power rule, the POWER correction, and the rule of thumb respectively. There are no results for Thumb on Mildew. FDR(t) and power(t) denote the False Discovery Rate (FDR) and the power, respectively, at p-value threshold t; ue($\widehat{\text{FDR}}_{\text{BY}}, t$) and oe($\widehat{\text{FDR}}_{\text{BY}}, t$) denote underestimation and overestimation, respectively, of FDR(t) by FDR estimator $\widehat{\text{FDR}}_{\text{BY}}(t)$ of Benjamini and Yekutieli (2001); α is the significance level of the underlying hypothesis tests; uc($\widehat{\text{FDR}}_{\text{BY}}, q$) and oc($\widehat{\text{FDR}}_{\text{BY}}, q$) denote undercontrol and overcontrol, respectively, of the FDR by the FDR controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold q; procpower($\widehat{\text{FDR}}, q$) denotes the power of the procedure; E[time] denotes the expected skeleton identification time (see main text for definitions). Each cell of the table contains the mean value of the corresponding quantity across various networks and sample sizes. Compared to Heuristic, POWER achieves, on average, greater mean power and less mean overestimation and overcontrol, but greater mean FDR, underestimation and undercontrol and Thumb achieves, on average, less mean FDR, underestimation and undercontrol, but less mean power and greater mean overestimation and overcontrol. Mean expected skeleton identification time is similar for all criteria.

C.2 Increasing the significance level

We applied MMPC-skeleton to the samples from each network, each time using the first $n \in \{100, 1000, 10000\}$ observations and significance level $\alpha \in \{0.05, 0.1, 0.2\}$. For the rest parameters we used the same values as in Section 3.3 of the main text.

We approximated mean FDR, power, and FDR under– and overestimation (Tables C.10 – C.13) in $[0, 0.05]$, in order for the results to be comparable. We also approximated mean FDR under– and overcontrol, and FDR-controlling-procedure power (Tables C.14 – C.16) in $[0.001, 0.1]$, as usually. Finally, we approximated the expected skeleton identification time (Table C.17). Table C.18 summarizes the results. On average, when α increases, mean FDR, underestimation, and undercontrol slightly decrease, but mean power and FDR-controlling-procedure power slightly decrease, mean overestimation and overcontrol slightly increase, and expected skeleton identification time greatly increases. Therefore, we do not

recommend using $\alpha > 0.05$, because the slight decrease in underestimation and undercontrol is overshadowed by the great increase in skeleton identification time.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.7164	0.7122	0.7073
	1000	0.02877	0.02866	0.02725
	1e+04	0.002623	0.002282	0.001406
Andes	100	0.2264	0.1694	0.09735
	1000	0.06904	0.04819	0.02179
	1e+04	0.03021	0.02107	0.009522
Barley	100	0.8856	0.8856	0.8856
	1000	0.7404	0.7397	0.7382
	1e+04	0.5086	0.5073	0.504
Hailfinder	100	0.6915	0.6907	0.6895
	1000	0.5203	0.5194	0.5186
	1e+04	0.06023	0.04406	0.02112
Hepar II	100	0.4225	0.3704	0.3026
	1000	0.1267	0.09773	0.05944
	1e+04	0.04764	0.03626	0.01922
Insurance	100	0.5302	0.5285	0.526
	1000	0.0817	0.08068	0.07705
	1e+04	0.003949	0.003554	0.001912
Mildew	100	0.3851	0.3851	0.3851
	1000	0.6177	0.6169	0.6155
	1e+04	0.7019	0.701	0.6998
Power Plant	100	0.8812	0.8812	0.8812
	1000	0.04623	0.03943	0.03258
	1e+04	0.0236	0.01977	0.01382
Water	100	0.4417	0.4417	0.4417
	1000	0.03722	0.03443	0.02843
	1e+04	0.00658	0.00554	0.004564
Win95pts	100	0.1297	0.1154	0.09879
	1000	0.04597	0.03629	0.02359
	1e+04	0.01779	0.01427	0.009196
	Mean:	0.3009	0.2926	0.2814

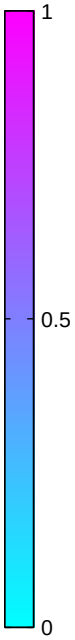


Table C.10: Mean FDR(t) for $t \in [0, 0.05]$: mean False Discovery Rate (FDR) in $[0, 0.05]$ for each network, sample size n and significance level α. On average, mean FDR slightly decreases when α increases.

bayesNetLabel	nObs	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.6551	0.6525	0.6482
	1000	0.8205	0.8191	0.8163
	1e+04	0.9295	0.9289	0.9273
Andes	100	0.2713	0.2539	0.2305
	1000	0.5851	0.5768	0.561
	1e+04	0.7979	0.7946	0.7883
Barley	100	0.00651	0.00651	0.00651
	1000	0.7909	0.7894	0.7867
	1e+04	0.8197	0.8184	0.8166
Hailfinder	100	0.5059	0.5053	0.5045
	1000	0.7003	0.6993	0.6967
	1e+04	0.5086	0.5086	0.5085
Hepar II	100	0.1278	0.1216	0.1136
	1000	0.3171	0.3055	0.289
	1e+04	0.5715	0.5661	0.5559
Insurance	100	0.5316	0.5301	0.5274
	1000	0.6583	0.6568	0.6533
	1e+04	0.7966	0.7939	0.7899
Mildew	100	0.1232	0.1232	0.1232
	1000	0.6464	0.6464	0.6461
	1e+04	0.8039	0.8032	0.8012
Power Plant	100	0.9646	0.9646	0.9646
	1000	0.512	0.511	0.5096
	1e+04	0.6015	0.5912	0.5887
Water	100	0.3015	0.3015	0.3015
	1000	0.2934	0.2918	0.2876
	1e+04	0.4597	0.4577	0.4542
Win95pts	100	0.1805	0.1771	0.1713
	1000	0.4087	0.4009	0.3923
	1e+04	0.6463	0.6413	0.6339
	Mean:	0.5445	0.5412	0.5365

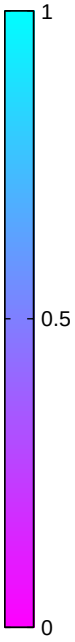


Table C.11: Mean power(t) for $t \in [0, 0.05]$: mean power in $[0, 0.05]$ for each network, sample size n and significance level α. On average, mean power slightly decreases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.1052	0.1018	0.09763
	1000	0	0	0
	1e+04	0	0	0
Andes	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Barley	100	0	0	0
	1000	0.1608	0.1595	0.157
	1e+04	0.03443	0.03402	0.03307
Hailfinder	100	0.01124	0.01112	0.01091
	1000	0.0173	0.01726	0.01722
	1e+04	0	0	0
Hepar II	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Insurance	100	0.04299	0.04204	0.0409
	1000	0.0001046	8.43e-05	3.604e-05
	1e+04	0	0	0
Mildew	100	0	0	0
	1000	0.06317	0.06295	0.06271
	1e+04	0.1485	0.1475	0.1462
Power Plant	100	0.4044	0.4044	0.4044
	1000	0	0	0
	1e+04	0	0	0
Water	100	0.004804	0.004804	0.004804
	1000	0	0	0
	1e+04	0	0	0
Win95pts	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
	Mean:	0.0331	0.03285	0.03249

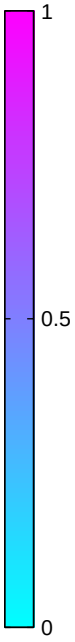


Table C.12: Mean ue($\widehat{\text{FDR}}, t$) for $t \in [0, 0.05]$: mean underestimation (see main text for definition) of the False Discovery Rate (FDR) by FDR estimator $\widehat{\text{FDR}}_{\text{BY}}$ by Benjamini and Yekutieli (2001) in $[0, 0.05]$ for each network, sample size n and significance level α. On average, mean underestimation slightly decreases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.152	0.1576	0.1644
	1000	0.8816	0.8819	0.8835
	1e+04	0.8968	0.8973	0.8984
Andes	100	0.7588	0.8157	0.8878
	1000	0.914	0.935	0.9616
	1e+04	0.95	0.9592	0.9708
Barley	100	0.09949	0.09949	0.09949
	1000	0.09872	0.09985	0.1021
	1e+04	0.3559	0.3576	0.3613
Hailfinder	100	0.2522	0.2531	0.2545
	1000	0.4116	0.4125	0.4136
	1e+04	0.9019	0.9182	0.9412
Hepar II	100	0.5626	0.6147	0.6826
	1000	0.8487	0.8783	0.9173
	1e+04	0.9104	0.9222	0.9399
Insurance	100	0.2463	0.2493	0.2536
	1000	0.7479	0.7494	0.7545
	1e+04	0.8035	0.8046	0.8074
Mildew	100	0.5928	0.5928	0.5928
	1000	0.2505	0.2515	0.2532
	1e+04	0.1306	0.1319	0.1337
Power Plant	100	7.806e-05	7.806e-05	7.806e-05
	1000	0.916	0.9229	0.9298
	1e+04	0.9362	0.9408	0.9469
Water	100	0.471	0.471	0.471
	1000	0.8984	0.9014	0.9078
	1e+04	0.8973	0.8988	0.9006
Win95pts	100	0.854	0.8686	0.8855
	1000	0.9278	0.9381	0.9514
	1e+04	0.9447	0.9485	0.9539
	Mean:	0.6204	0.6291	0.6407

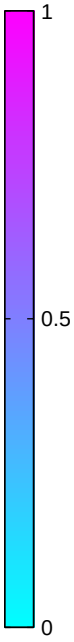


Table C.13: Mean $\mathrm{oe}(\widehat{\mathrm{FDR}}, t)$ for $t \in [0, 0.05]$: mean overestimation (see main text for definition) of the False Discovery Rate (FDR) by FDR estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ by Benjamini and Yekutieli (2001) in $[0, 0.05]$ for each network, sample size n and significance level α. On average, mean overestimation slightly increases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.6488	0.6463	0.6424
	1000	0.00421	0.004241	0.004129
	1e+04	0	0	0
Andes	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Barley	100	0.9281	0.9281	0.9281
	1000	0.6575	0.6566	0.6549
	1e+04	0.4106	0.4088	0.4047
Hailfinder	100	0.4039	0.4037	0.4025
	1000	0.4294	0.4295	0.4289
	1e+04	0	0	0
Hepar II	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Insurance	100	0.3448	0.3432	0.3417
	1000	0.007042	0.006098	0.004259
	1e+04	0	0	0
Mildew	100	0	0	0
	1000	0.5408	0.5408	0.5405
	1e+04	0.6366	0.636	0.6354
Power Plant	100	0.8224	0.8224	0.8224
	1000	0.001262	0.001253	0.001256
	1e+04	0	0	0
Water	100	0.213	0.213	0.213
	1000	0	0	0
	1e+04	0	0	0
Win95pts	100	0.01675	0.01341	0.01192
	1000	0.0016	0.0008591	0.0002403
	1e+04	0	0	0
	Mean:	0.2022	0.2018	0.2012

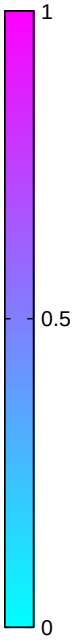


Table C.14: Mean $\mathrm{uc}(\widehat{\mathrm{FDR}}, q)$ for $q \in [0.001, 0.1]$: mean undercontrol (see main text for definition) of the False Discovery Rate (FDR) by the FDR controlling procedure with FDR estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ by Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and significance level α. On average, mean undercontrol slightly decreases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0	0	0
	1000	0.02539	0.02535	0.02563
	1e+04	0.04999	0.04999	0.04999
Andes	100	0.04857	0.04946	0.04999
	1000	0.04979	0.04993	0.04999
	1e+04	0.04997	0.04999	0.04999
Barley	100	0	0	0
	1000	0	0	0
	1e+04	0	0	0
Hailfinder	100	0	0	0
	1000	0	0	0
	1e+04	0.04939	0.04974	0.04999
Hepar II	100	0.04749	0.04888	0.04999
	1000	0.04963	0.04967	0.04999
	1e+04	0.04994	0.04999	0.04999
Insurance	100	0	0	0
	1000	0.01177	0.01222	0.01468
	1e+04	0.04953	0.04952	0.04975
Mildew	100	0.04875	0.04875	0.04875
	1000	0	0	0
	1e+04	0	0	0
Power Plant	100	0	0	0
	1000	0.03415	0.03413	0.03411
	1e+04	0.04999	0.04999	0.04999
Water	100	0	0	0
	1000	0.0494	0.04949	0.04974
	1e+04	0.04954	0.04954	0.0497
Win95pts	100	0.009272	0.01162	0.01362
	1000	0.03383	0.03757	0.04265
	1e+04	0.04983	0.04983	0.04983
	Mean:	0.02687	0.02719	0.02761

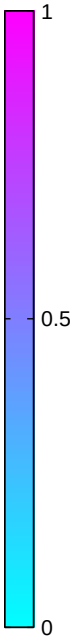


Table C.15: Mean oc($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean overcontrol (see main text for definition) of the False Discovery Rate (FDR) by the FDR controlling procedure of Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and significance level α. On average, mean overcontrol slightly increases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	0.5299	0.5253	0.5207
	1000	0.7251	0.723	0.7184
	1e+04	0.917	0.9155	0.9122
Andes	100	0.07387	0.06589	0.05845
	1000	0.449	0.4413	0.4286
	1e+04	0.7433	0.7401	0.7338
Barley	100	0.000569	0.000569	0.000569
	1000	0.7559	0.7539	0.7498
	1e+04	0.7933	0.7928	0.792
Hailfinder	100	0.3604	0.3595	0.3585
	1000	0.6711	0.6697	0.6684
	1e+04	0.5095	0.5094	0.509
Hepar II	100	0.02715	0.0246	0.02217
	1000	0.1878	0.1799	0.1696
	1e+04	0.4771	0.4698	0.4582
Insurance	100	0.4167	0.4152	0.4123
	1000	0.6081	0.6059	0.6017
	1e+04	0.7733	0.771	0.7675
Mildew	100	0.08429	0.08429	0.08429
	1000	0.6115	0.6114	0.6113
	1e+04	0.785	0.784	0.7816
Power Plant	100	0.9377	0.9377	0.9377
	1000	0.4862	0.4847	0.4837
	1e+04	0.5301	0.5176	0.5157
Water	100	0.2373	0.2373	0.2373
	1000	0.2633	0.2629	0.2615
	1e+04	0.3963	0.394	0.3888
Win95pts	100	0.1026	0.09882	0.0952
	1000	0.2911	0.2821	0.2717
	1e+04	0.5433	0.5371	0.529
	Mean:	0.4763	0.4732	0.4693

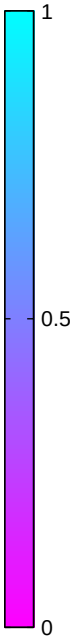


Table C.16: Mean procpower($\widehat{\text{FDR}}, q$) for $q \in [0.001, 0.1]$: mean power of the False Discovery Rate (FDR) controlling procedure of Benjamini and Yekutieli (2001) in $[0.001, 0.1]$ for each network, sample size n and significance level α. On average, mean overcontrol slightly increases when α increases.

Network	n	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Alarm	100	2.83	4.82	6.03
	1000	3.83	6.30	7.76
	1e+04	6.09	8.67	11.12
Andes	100	77.22	139.95	209.07
	1000	146.84	174.84	222.19
	1e+04	370.92	764.06	878.31
Barley	100	2.92	3.01	3.02
	1000	8.67	9.88	12.32
	1e+04	31.85	38.41	49.97
Hailfinder	100	6.83	7.53	9.36
	1000	12.59	14.66	18.94
	1e+04	13.29	14.97	19.15
Hepar II	100	11.34	13.29	20.07
	1000	14.41	18.61	23.74
	1e+04	91.86	121.81	191.74
Insurance	100	2.09	2.45	3.04
	1000	4.76	5.77	7.78
	1e+04	14.58	16.31	20.34
Mildew	100	1.78	1.82	1.99
	1000	2.94	3.33	4.08
	1e+04	10.29	12.09	15.21
Power Plant	100	8.33	9.43	10.67
	1000	7.91	8.85	10.08
	1e+04	10.37	11.64	13.08
Water	100	2.22	2.46	2.64
	1000	2.83	3.29	4.04
	1e+04	5.09	5.85	7.07
Win95pts	100	13.51	15.32	17.78
	1000	17.33	20.73	28.22
	1e+04	36.89	44.08	56.76
	Mean:	31.41	50.14	62.85

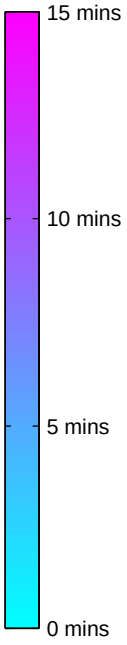


Table C.17: Expected skeleton identification time (seconds) for each network, sample size n and significance level α. Expected time increases when α increases.

	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
Mean FDR(t) for $t \in [0, 0.05]$	0.3009	0.2926	0.2814
Mean power(t) for $t \in [0, 0.05]$	0.5445	0.5412	0.5365
Mean ue($\widehat{\text{FDR}}_{\text{BY}}, t$) for $t \in [0, 0.05]$	0.0331	0.03285	0.03249
Mean oe($\widehat{\text{FDR}}_{\text{BY}}, t$) for $t \in [0, 0.05]$	0.6204	0.6291	0.6407
Mean uc($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.2022	0.2018	0.2012
Mean oc($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.02687	0.02719	0.02761
Mean procpower($\widehat{\text{FDR}}_{\text{BY}}, q$) for $q \in [0.001, 0.1]$	0.4763	0.4732	0.4693
E[time] (s)	31.41	50.14	62.85

Table C.18: Summary of the results of increasing the significance level α. FDR(t) and power(t) denote the False Discovery Rate (FDR) and the power, respectively, at p-value threshold t; ue($\widehat{\text{FDR}}_{\text{BY}}, t$) and oe($\widehat{\text{FDR}}_{\text{BY}}, t$) denote underestimation and overestimation, respectively, of FDR(t) by FDR estimator $\widehat{\text{FDR}}_{\text{BY}}(t)$ of Benjamini and Yekutieli (2001); α is the significance level of the underlying hypothesis tests; uc($\widehat{\text{FDR}}_{\text{BY}}, q$) and oc($\widehat{\text{FDR}}_{\text{BY}}, q$) denote undercontrol and overcontrol, respectively, of the FDR by the FDR controlling procedure of Benjamini and Yekutieli (2001) with FDR threshold q; procpower($\widehat{\text{FDR}}, q$) denotes the power of the procedure; E[time] denotes the expected skeleton identification time (see main text for definitions). Each cell of the table contains the mean value of the corresponding quantity across various networks and sample sizes. On average, when α increases, mean FDR, underestimation, and undercontrol slightly decrease, but mean power and FDR-controlling-procedure power slightly decrease, mean overestimation and overcontrol slightly increase, and expected skeleton identification time greatly increases.

C.3 Varying the test statistic

Fast (2010) evaluates the Cochran-Mantel-Haenszel (CMH) test of independence as a means to increase power compared to the usual G test. He demonstrates that the CMH test actually decreases power and increases the FPR on all three networks (Alarm, Insurance and Win95pts) used in the evaluation. Therefore, we do not consider the CMH test.

The p-value of the G test is calculated assuming that the test statistic follows the χ^2 distribution. However, the latter is only asymptotically true. Thus, the calculated p-value is only asymptotically correct. Tsamardinos and Borboudakis (2010) demonstrate that the G test with degrees of freedom calculated according to Tsamardinos et al. (2006) underestimates the p-value when the sample size is small. Tsamardinos and Borboudakis (2010) devise a permutation test that is well-calibrated, i.e., its FPR matches the significance level α. We do not consider this test as a means to increase the power of the tests either, because not underestimating the p-value would only decrease the power of the tests.

Appendix D

D.1 Data sources

Summary statistics of the Bayesian networks used throughout this thesis are given in Table D.1. The Win95pts network was developed at Microsoft Research and contributed to the community by Jack Breese. The networks were downloaded from the following three online repositories:

Bayesian Networks and Decision Graphs (BNDG)
`http://bndg.cs.aau.dk/html/bayesian_networks.html`

Bayesian Network Repository (BNR)
`http://www.cs.huji.ac.il/site//labs/compbio/Repository/`

GeNIe & SMILE Network Repository (GS)
`http://genie.sis.pitt.edu/networks.html`

Name	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	$\overline{\|\mathbf{PA}_X\|}$	$\overline{\|\mathbf{D}_X\|}$	Repository	Reference
Alarm	37	46	1.24	2.84	GS	Beinlich et al.
Andes	223	338	1.52	2	GS	Conati et al. (1997)
Barley	48	84	1.75	8.77	GS	Kristensen and Rasmussen (2002)
Hailfinder	56	66	1.18	3.98	GS	Abramson et al. (1996)
Hepar II	70	123	1.76	2.31	GS	Onisko (2003)
Insurance	27	52	1.93	3.30	BNR	Binder et al. (1997)
Mildew	35	46	1.31	17.6	BNR	Jensen and Jensen (1996)
Water	32	66	2.06	3.63	BNR	Jensen (1989)
Power Plant	45	42	13	0.93	BNDG	Nielsen and Jensen (2007)
Win95pts	76	112	1.47	2.00	GS	

Table D.1: Summary statistics of the Bayesian networks used throughout this thesis. $|\mathbf{V}|$ denotes the number of nodes, $|\mathbf{E}|$ the number of edges, $\overline{|\mathbf{PA}_X|}$ the mean number of parents and $\overline{|\mathbf{D}_X|}$ the mean number of levels.

D.2 Software

- The *Probabilistic Graphical Model Toolbox* (PGM Toolbox) for MATLAB® is developed by Angelos P. Armen and was used in the experiments of this work. PGM Toolbox can be downloaded from `http://www.mensxmachina.org/`.
- The *POWER* correction, a reliability criterion evaluated in Section C.1 of Appendix C, is available within the *PowerBayes* software package by Andrew Fast. PowerBayes can be downloaded from `http://kdl.cs.umass.edu/powerbayes/`.