Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information

Séverine Affeldt, Hervé Isambert
Institut Curie, Research Center, CNRS, UMR168, 26 rue d'Ulm, 75005, Paris France;
and Université Pierre et Marie Curie, 4 Place Jussieu, 75005, Paris, France
herve.isambert@curie.fr

Abstract

We report a novel network reconstruction method, which combines constraint-based and Bayesian frameworks to reliably reconstruct graphical models despite inherent sampling noise in finite observational datasets. The approach is based on an information theory result tracing back the existence of colliders in graphical models to negative conditional 3-point information between observed variables. In turn, this provides a confident assessment of structural independencies in causal graphs, based on the ranking of their most likely contributing nodes with (significantly) positive conditional 3-point information. Starting from a complete undirected graph, dispensible edges are progressively pruned by iteratively "taking off" the most likely positive conditional 3-point information from the 2-point (mutual) information between each pair of nodes. The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples. This "3off2" network reconstruction approach is shown to outperform constraint-based, search-and-score and earlier hybrid methods on a range of benchmark networks.

1 INTRODUCTION

The prospect of learning the direction of causal dependencies from mere correlations in observational data has long defied practical implementations (Reichenbach, 1956). The fact that causal relationships can, to some extent, be inferred from nontemporal statistical data is now known to hinge on the unique statistical imprint of colliders in causal graphical models, provided that certain assumptions are made about the underlying process of data generation, such as its faithfulness to a tree structure (Rebane and Pearl, 1988) or a directed acyclic graph model (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009).

These early findings led to the developments of two types of network reconstruction approaches; on the one hand, search and score methods (Cooper and Herskovits, 1992; Heckerman, Geiger, and Chickering, 1995; Chickering, 2002) need heuristic strategies, such as hill-climbing algorithms, to sample network space, on the other hand, constraint-based methods, such as the PC (Spirtes and Glymour, 1991) and IC (Pearl and Verma, 1991) algorithms, rely on the identification of structural independencies, that correspond to edges to be removed from the underlying network (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009). Yet, early errors in removing edges from the complete graph often lead to the accumulation of compensatory errors later on in the pruning process. Hence, despite recent, more stable implementations intending to overcome order-dependency in the pruning process (Colombo and Maathuis, 2014), constraint-based methods are not robust to sampling noise in finite datasets.

In this paper, we present a more robust constrained-based method and corresponding 3off2 algorithm. It is directly inspired by the PC anc IC algorithms but relies on a quantitative information theoretic framework to reliably uncover conditional independencies in finite datasets and subsequently orient and propagate edge directions between connected variables.

2 RESULTS

2.1 UNCOVERING CAUSALITY FROM A STABLE / FAITHFUL DISTRIBUTION

Consider a network $\mathcal{G} = (V, E)$ and a *stable* (or *faithful*) distribution $P(\mathbf{X})$ over V, implying that each structural independency (*i.e.* missing edge XY in $\mathcal{G}$)

corresponds to a vanishing conditional 2-point (mutual) information and reciprocally as,

$$(X \perp\!\!\!\perp Y|\{U_i\})_G \iff (X \perp\!\!\!\perp Y|\{U_i\})_P \quad (1)$$
$$\iff I(X;Y|\{U_i\}) = 0 \quad (2)$$

Eq.1 assumes, in particular, that $P(\mathbf{X})$ is a theoretical distribution, defined by a formal expression of its variables $\mathbf{X} = \{X, Y, U_1, U_2, \ldots\}$. Note, however, that no such expression is known *a priori*, in general, and $P(\mathbf{X})$ must typically be *estimated* from the available data. In principle, an infinite amount of data would be necessary to infer an 'exact' *stable* distribution $P(\mathbf{X})$ consistent with Eq.1. In the following, we will first assume that such an infinite amount of data is available and distributed as a stable $P(\mathbf{X})$ to establish how causality can be inferred statistically from conditional 2-point and 3-point information. We will then consider the more realistic situation for which $P(\mathbf{X})$ is not known exactly and must be estimated from a finite amount of data.

Let us first recall the *generic decomposition* of a conditional 2-point (or mutual) information $I(X;Y|\{U_i\})$ by the introduction of a third node Z and the conditional 3-point information $I(X;Y;Z|\{U_i\})$,

$$I(X;Y|\{U_i\}) = I(X;Y;Z|\{U_i\}) + I(X;Y|\{U_i\}, Z) \quad (3)$$

This relation can be taken as the definition of conditional 3-point information $I(X;Y;Z|\{U_i\})$ which is in fact symmetric in X, Y and Z,

$$\begin{aligned} I(X;Y;Z|\{U_i\}) &= I(X;Y|\{U_i\}) - I(X;Y|\{U_i\}, Z) \\ &= I(X;Z|\{U_i\}) - I(X;Z|\{U_i\}, Y) \\ &= I(Y;Z|\{U_i\}) - I(Y;Z|\{U_i\}, X) \end{aligned}$$

Note that Eq.3 is always valid, regardless of any assumption on the underlying graphical model and of the amount of data available to estimate conditional 2-point and 3-point information terms. Eq.3 will be used to prove the following lemmas and propositions, which trace back the origin of necessary causal relationships in a graphical model to the existence of a negative conditional 3-point information between *three* variables $\{X, Y, Z\}$, $I(X;Y;Z|\{U_i\}) < 0$, where $\{U_i\}$ accounts for a structural independency between two of them, *e.g.* $I(X;Y|\{U_i\}) = 0$ (see Theorem 4).

Lemma 1. *Given a stable distribution $P(\mathbf{X})$ on V, $\forall X, Y \in V$ not adjacent in $\mathcal{G}$, $\exists\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$ and $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y;Z|\{U_i\}) \leqslant 0$.*

Proof. If $X, Y \in V$ are not adjacent in $\mathcal{G}$, this corresponds to a structural independency, *i.e.* $\exists\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$. Then $\forall Z \neq X, Y, \{U_i\}$ Eq.3 implies $I(X;Y;Z|\{U_i\}) = -I(X;Y|\{U_i\}, Z) \leqslant 0$, as conditional mutual information is always positive. □

Corollary 2 (3-point contribution). *$\forall X, Y, Z \in V$ and $\forall\{U_i\} \subseteq V_{\backslash\{X,Y,Z\}}$ s.t. $I(X;Y;Z|\{U_i\}) > 0$, then $I(X;Y|\{U_i\}) > 0$ (as well as $I(X;Z|\{U_i\}) > 0$ and $I(Y;Z|\{U_i\}) > 0$ by symmetry of $I(X;Y;Z|\{U_i\})$).*

Corollary 2, which is a direct consequence of Eq.3 and the positivity of mutual information, will be the basis of the 3off2 causal network reconstruction algorithm, which iteratively "takes off" 3-point information from 2-point information, as $I(X;Y|\{U_i\}) - I(X;Y;Z|\{U_i\}) = I(X;Y|\{U_i\}, Z)$, and update $\{U_i\} \leftarrow \{U_i\} + Z$ as long as there remains some $Z \in V$ with (significantly) positive conditional 3-point information $I(X;Y;Z|\{U_i\}) > 0$.

Lemma 3 (vanishing conditional 2-point and 3-point information in undirected networks). *If $\mathcal{G}$ is an undirected (Markov) network, $\forall X, Y \in V$ and $\forall\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$, then $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y;Z|\{U_i\}) = 0$.*

Proof. If $\mathcal{G}$ is a Markov network, $\forall X, Y \in V$ and $\forall\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$, then $\forall Z \neq X, Y, \{U_i\}$, $I(X;Y|\{U_i\}, Z) = 0$ as conditioning observation cannot induce correlations in Markov networks (Koller and Friedman, 2009). This implies that $I(X;Y;Z|\{U_i\}) = 0$ through Eq.3. □

Note, however, that the converse of Lemma 3 is not true. Namely, (partially) directed networks can also have vanishing conditional 3-point information associated to all their structural independencies. In particular, tree-like bayesian networks without colliders (*i.e.* without v-structures, $X \rightarrow Z \leftarrow Y$) present only vanishing 3-point information associated to their structural independencies, *i.e.* $I(X;Y;Z|\{U_i\}) = 0$, $\forall X, Y, Z, \{U_i\} \in V$ s.t. $I(X;Y|\{U_i\}) = 0$. However, such a directed network must be Markov equivalent to an undirected network corresponding to the same structural independencies but lacking any trace of causal relationships (*i.e.* no directed edges). The probability distributions faithful to such directed networks do not contain evidence of obligate causality; *i.e.* no directed edges can be unambiguously oriented.

The following Theorem 4 establishes the existence of negative conditional 3-point information as statistical evidence of obligate causality in graphical models. For the purpose of generality in this section, we do not exclude the possibility that unobserved 'latent' variables might mediate the causal relationships among observed variables. However, this requires dissociating the labelling of the two endpoints of each edges. Let us first introduce three different endpoint marks associated to such edges in mixed graphs: they are the tail ($-$), the head ($>$) and the unspecified ($\circ$) endpoint marks. In addition, we will use the asterisk symbol ($*$) as a wild card denoting any of the three marks.

Theorem 4 (negative conditional 3-point information as statistical evidence of causality). *If $\exists X, Y, Z \in V$ and $\{U_i\} \subseteq V_{\backslash\{X,Y,Z\}}$ s.t. $I(X;Y|\{U_i\}) = 0$ and $I(X;Y;Z|\{U_i\}) < 0$ then, $\mathcal{G}$ is (partially) directed, i.e. some variables in $\mathcal{G}$ are causally linked, either directly or indirectly through other variables, including possibly unknown, 'latent' variables unobserved in $\mathcal{G}$.*

Proof. Theorem 4 is the contrapositive of Lemma 3, with the additional use of Lemma 1. □

Proposition 5 (origin of causality at unshielded triples with negative conditional 3-point information). *for all unshielded triple, $X \ast\!\!-\!\!\circ\, Z \circ\!\!-\!\!\ast\, Y$, $\exists\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$, if $Z \notin \{U_i\}$ then $I(X;Y;Z|\{U_i\}) < 0$ and the unshielded triple should be oriented as $X \ast\!\!\rightarrow Z \leftarrow\!\!\ast\, Y$.*

Proof. if $I(X;Y|\{U_i\})=0$ with $Z\notin\{U_i\}$, the unshielded triple has to be a collider and $I(X;Y|\{U_i\},Z)>0$, by faithfulness, hence, $I(X;Y;Z|\{U_i\})<0$ by Eq.3. □

Hence, the origin of causality manifests itself in the form of colliders or v-structures in graphical models which reveal 'genuine' causations ($X\!\rightarrow Z$ or $Y\!\rightarrow Z$) or, alternatively, 'possible' causations ($X\circ\!\!\rightarrow Z$ or $Y\circ\!\!\rightarrow Z$), provided that the corresponding correlations are not due to unobserved 'latent' variables L or L' as, $X \dashleftarrow L \dashrightarrow Z$ or $Y \dashleftarrow L' \dashrightarrow Z$.

Following the rationale of constraint-based approaches, it is then possible to 'propagate' further the orientations downstream of colliders, through positive (conditional) 3-point information, if one assumes that the underlying distribution $P(\mathbf{X})$ is faithful to an *ancestral graph* $\mathcal{G}$ on V. An *ancestral graph* is a mixed graph, that is, with three types of edges, undirected ($-$), directed ($\leftarrow$ or $\rightarrow$) or bidirectional ($\leftrightarrow$), but with *i.)* no directed cycle, *ii.)* no almost directed cycle (including one bidirectional edge) and *iii.)* no undirected edge with incoming arrowhead (such as $X \ast\!\!\rightarrow Z\!-\!Y$). In particular, Directed Acyclic Graphs (DAG) are subclasses of ancestral graphs (*i.e.* without undirected nor bidirectional edges).

Proposition 6 ('propagation' of causality at unshielded triples with positive conditional 3-pt information). *Given a distribution $P(\mathbf{X})$ faithful to an ancestral graph $\mathcal{G}$ on V, for all unshielded triple with already one converging orientation, $X \ast\!\!\rightarrow Z \circ\!\!-\!\!\ast\, Y$, $\exists\{U_i\} \subseteq V_{\backslash\{X,Y\}}$ s.t. $I(X;Y|\{U_i\}) = 0$, if $Z \in \{U_i\}$ then $I(X;Y;Z|\{U_i\}_{\backslash Z}) > 0$ and the first orientation should be 'propagated' to the second edge as $X \ast\!\!\rightarrow Z \rightarrow Y$.*

Proof. if $I(X;Y|\{U_i\}) = 0$ with $Z \in \{U_i\}$, the unshielded triple cannot be a collider and, since $\mathcal{G}$ is assumed to be an ancestral graph, the edge $Z\!-\!Y$ cannot be an undirected edge either. Hence, it has to be a directed edge, $Z\!\rightarrow Y$ and $I(X;Y;Z|\{U_i\}_{\backslash Z}) > 0$ by faithfulness and Eq.3. □

Note that the propagation rule of Proposition 6 can be applied iteratively to successive unshielded triples corresponding to positive conditional 3-point information. Yet, all arrowhead orientations can be ultimately traced back to a negative conditional 3-point information, Theorem 4 and Proposition 5.

2.2 ROBUST RECONSTRUCTION OF CAUSAL GRAPHS FROM FINITE DATASETS

We now turn to the more practically relevant situation of finite datasets consisting of N independent data points. The associated sampling noise will instrinsically limit the accuracy of causal network reconstruction. In particular, conditional independencies cannot be exactly achieved ($I(X;Y|\{U_i\}) = 0$) but can be reliably established using statistical criteria that depend on the number of data points N.

Given N independent datapoints from the available data $\mathcal{D}$, let us introduce the maximum likelihood, $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$, that they might have been generated by the graphical model $\mathcal{G}$ (Sanov, 1957),

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = \frac{e^{N\sum_{\{x_i\}} p(\{x_i\})\log(q(\{x_i\}))}}{Z(\mathcal{G},\mathcal{D})} \quad (4)$$

where $H(\mathcal{G},\mathcal{D}) = -\sum_{\{x_i\}} p(\{x_i\})\log(q(\{x_i\}))$ is the cross entropy between the "true" probability distribution $p(\{x_i\})$ of the data $\mathcal{D}$ and the theoretical probability distribution $q(\{x_i\})$ of the model $\mathcal{G}$ and $Z(\mathcal{G},\mathcal{D})$ is a data- and model-dependent factor ensuring proper normalization condition. The structural constraints of the model $\mathcal{G}$ can be included *a priori* in the factorization form of the theoretical probability distribution, $q(\{x_i\})$. In particular, if we assume a Bayesian network as underlying graphical model, $q(\{x_i\})$ factorizes as $q(\{x_i\}) = \prod_i p(x_i|\{\mathrm{pa}_{x_i}\})$, where $\{\mathrm{pa}_{x_i}\}$ denote the values of the parents of node X_i, $\{\mathrm{Pa}_{X_i}\}$, and leads to the following maximum likelihood expression,

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-N\sum_i H(X_i|\{\mathrm{Pa}_{X_i}\})}}{Z(\mathcal{G},\mathcal{D})} \quad (5)$$

The model $\mathcal{G}$ can then be compared to the alternative model $\mathcal{G}_{\backslash X\rightarrow Y}$ with one additional missing edge $X \rightarrow Y$ using the maximum likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\backslash X\rightarrow Y}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI(X;Y|\{\mathrm{Pa}_Y\}_{\backslash X})}\frac{Z(\mathcal{G},\mathcal{D})}{Z(\mathcal{G}_{\backslash X\rightarrow Y},\mathcal{D})} \quad (6)$$

where $I(X;Y|\{\mathrm{Pa}_Y\}_{\backslash X}) = H(Y|\{\mathrm{Pa}_Y\}_{\backslash X}) - H(Y|\{\mathrm{Pa}_Y\})$. However, Eq.6 cannot be used as such to learn the underlying graphical model, as it assumes that the order between the nodes and their parents is already known (see however (de Campos, 2006)). Yet, following the rationale of constraint-based approaches, Eq.6 can be reformulated by replacing the parent nodes with an unknown separation set $\{U_i\}$ to be learnt simultaneously with the missing edge candidate XY,

$$\frac{\mathcal{L}_{\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(X;Y|\{U_i\})+k_{X;Y|\{U_i\}}} \tag{7}$$
$$k_{X;Y|\{U_i\}} = \log\left(Z(\mathcal{G},\mathcal{D})/Z(\mathcal{G}_{\setminus XY|\{U_i\}},\mathcal{D})\right)$$

where the factor $k_{X;Y|\{U_i\}} > 0$ tends to limit the complexity of the models by favoring fewer edges. Namely, the condition, $I(X;Y|\{U_i\}) < k_{X;Y|\{U_i\}}/N$, implies that simpler models compatible with the structural independency, $X \perp\!\!\!\perp Y|\{U_i\}$, are more likely than model $\mathcal{G}$, given the finite available dataset. This replaces the 'perfect' conditional independency condition, $I(X;Y|\{U_i\}) = 0$, valid in the limit of an infinite dataset, $N \to \infty$. A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria (Rissanen, 1978; Hansen and Yu, 2001),

$$k^{\mathsf{MDL}}_{X;Y|\{U_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i} \log N \tag{8}$$

where r_x, r_y and r_{u_i} are the number of levels of the corresponding variables. The MDL complexity, Eq.8, is simply related to the normalisation constant of the normal distribution reached in the asymptotic limit of a large dataset $N \to \infty$ (Central Limit Theorem). However, such a central limit distribution is only reached for very large datasets in practice. Alternatively, the normalisation of the maximum likelihood can also be done over all possible datasets including the same number of data points to yield a (universal) Normalized Maximum Likelihood (NML) criteria (Shtarkov, 1987; Rissanen and Tabus, 2005) and its decomposable (Kontkanen and Myllymäki, 2007; Roos et al., 2008) and XY-symmetric version, $k^{\mathsf{NML}}_{X;Y|\{U_i\}}$, defined in the Supplementary Methods.

Then, instead of exploring the combinatorics of sepset composition $\{U_i\}$ for each missing edge candidate XY as in traditional constraint-based approaches, we propose that Eq.7 can be used to iteratively extend a *likely* sepset using the maximum likelihood ratios between two successive sepset candidates, *i.e.* between the already ascertained $\{U_i\}$ and the possible extended $\{U_i\} + Z$, as,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}} = e^{NI(X;Y;Z|\{U_i\})+k_{X;Y;Z|\{U_i\}}} \tag{9}$$

using Eq.3 for $I(X;Y;Z|\{U_i\})$ and introducing a similar 3-point complexity conditioned on $\{U_i\}$ as,

$$k_{X;Y;Z|\{U_i\}} = k_{X;Y|\{U_i\},Z} - k_{X;Y|\{U_i\}} \tag{10}$$

where $k_{X;Y;Z|\{U_i\}} \geqslant 0$, unlike 3-point information, $I(X;Y;Z|\{U_i\})$ which can be positive or negative.

Introducing also the shifted 2-point and 3-point information for finite datasets as,

$$I'(X;Y|\{U_i\}) = I(X;Y|\{U_i\}) - \frac{k_{X;Y|\{U_i\}}}{N}$$
$$I'(X;Y;Z|\{U_i\}) = I(X;Y;Z|\{U_i\}) + \frac{k_{X;Y;Z|\{U_i\}}}{N}$$

leads to maximum likelihood ratios equivalent to Eqs.7 and 9,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X;Y|\{U_i\})} \tag{11}$$
$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}} = e^{NI'(X;Y;Z|\{U_i\})} \tag{12}$$

As will become apparent in the following discussion, learning, iteratively, the most likely edge to be removed XY and its corresponding separation set $\{U_i\}$ will imply to simultaneously minimize 2-point information (Eq.11) while maximizing 3-point information (Eq.12).

We start the discussion with 3-point information, Eq.12. The sign and magnitude of shifted conditional 3-point information $I'(X;Y;Z|\{U_i\})$ determine the probability that Z should be included in or excluded from the sepset candidate $\{U_i\}$,

• If $I'(X;Y;Z|\{U_i\}) > 0$, Z is more likely to be included in $\{U_i\}$ with probability,

$$P_{\mathsf{nv}}(X;Y;Z|\{U_i\}) = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}$$
$$= \frac{1}{1 + e^{-NI'(X;Y;Z|\{U_i\})}} \tag{13}$$

• If $I'(X;Y;Z|\{U_i\}) < 0$, Z is more likely to be excluded from $\{U_i\}$, suggesting obligatory causal relationships in the form of a v-structure or collider between X, Y, Z with probability,

$$P_{\mathsf{v}}(X;Y;Z|\{U_i\}) = 1 - P_{\mathsf{nv}}(X;Y;Z|\{U_i\})$$
$$= \frac{1}{1 + e^{NI'(X;Y;Z|\{U_i\})}} \tag{14}$$

But, in the case $I'(X;Y;Z|\{U_i\}) > 0$, Eq.12 can also be interpreted as quantifying the likelihood increase that the edge XY should be removed from the model by extending the candidate sepset from $\{U_i\}$ to $\{U_i\} + Z$, *i.e.* $\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}} = \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}} \times \exp(NI'(X;Y;Z|\{U_i\}))$, with $\exp(NI'(X;Y;Z|\{U_i\})) > 1$. Yet, as the 3-point information, $I'(X;Y;Z|\{U_i\})$, is actually symmetric with respect to the variables, X, Y and Z, the factor $\exp(NI'(X;Y;Z|\{U_i\})) > 1$ provides in fact the same likelihood increase for the removal of the three edges XY, XZ and ZY, conditioned on the same initial set of nodes $\{U_i\}$, namely,

$$\begin{aligned}\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}} &= \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{U_i\},y}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{U_i\}}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{U_i\},x}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{U_i\}}}}\\ &= e^{NI'(X;Y;Z|\{U_i\})}\end{aligned}$$

However, despite this symmetry of 3-point information, $I'(X;Y;Z|\{U_i\})$, the likelihoods that the edges XY, XZ and ZY should be removed are not the same, as they depend on different 2-point information, $I'(X;Y|\{U_i\})$, $I'(X;Z|\{U_i\})$ and $I'(Z;Y|\{U_i\})$, Eq.11. In particular, the likelihood ratio between the removals of the alternative edges XY and XZ is given by,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{U_i\},Y}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{U_i\}}}} = \frac{e^{-NI'(X;Y|\{U_i\})}}{e^{-NI'(X;Z|\{U_i\})}} \tag{15}$$

and similarly between edges XY and ZY.

Hence, for XY to be the most likely edge to be removed conditioned on the sepset $\{U_i\} + Z$, not only Z should contribute through $I'(X;Y;Z|\{U_i\}) > 0$ with probability $P_{\mathsf{nv}}(X;Y;Z|\{U_i\})$ (Eq.13), but XY must also correspond to the 'weakest' edge of XY, XZ and ZY conditioned on $\{U_i\}$, as given by the lowest conditioned 2-point information, Eq.15. Note that removing the edge XY with the lowest conditional 2-point information is consistent, as expected, with the Data Processing Inequality, $I(X;Y|\{U_i\}) \leqslant \min(I(X;Z|\{U_i\}), I(Z;Y|\{U_i\}))$, in the limit of large datasets. However, quite frequently, XZ or ZY might also have low conditional 2-point information, so that the edge removal associated with the symmetric contribution $I(X;Y;Z|\{U_i\})$ will only be consistent with the Data Processing Inequality (DPI) with probability,

$$\begin{aligned}&P_{\mathsf{dpi}}(XY;Z|\{U_i\}) =\\ &\quad= \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{U_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{U_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{U_i\}}}}\\ &\quad= \frac{1}{1 + \frac{e^{-NI'(X;Z|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}} + \frac{e^{-NI'(Z;Y|\{U_i\})}}{e^{-NI'(X;Y|\{U_i\})}}}\end{aligned} \tag{16}$$

In practice, taking into account this DPI-consistency probability $P_{\mathsf{dpi}}(XY;Z|\{U_i\})$, as detailed below, significantly improves the results obtained by relying solely on the 'non-v-structure' probability $P_{\mathsf{nv}}(X;Y;Z|\{U_i\})$. Conversely, the DPI-consistency probability $P_{\mathsf{dpi}}(XY;Z|\{U_i\})$ is not sufficient on its own to uncover causal relationships between variables, which require to compute 3-point information $I(X;Y;Z|\{U_i\})$ and the probability $P_{\mathsf{nv}}(X;Y;Z|\{U_i\})$ (see Proposition 7 and Proposition 8, below).

To optimize the likelihood that the edge XY can be accounted for by the additional contribution of Z conditioned on previously selected $\{U_i\}$, we propose to combine the maximum of 3-point information (Eq.13) and the minimum of 2-point information (Eq.16) by defining the score $S_{\mathsf{lb}}(Z;XY|\{U_i\})$ as the lower bound of $P_{\mathsf{nv}}(X;Y;Z|\{U_i\})$ and $P_{\mathsf{dpi}}(XY;Z|\{U_i\})$, since both conditions need to be fulfilled to warrant that edge XY is likely to be absent from the model $\mathcal{G}$,

$$\begin{aligned}&S_{\mathsf{lb}}(Z;XY|\{U_i\}) =\\ &\qquad= \min\Big[P_{\mathsf{nv}}(X;Y;Z|\{U_i\}), P_{\mathsf{dpi}}(XY;Z|\{U_i\})\Big]\end{aligned}$$

Hence, the pair of nodes XY with the most likely contribution from a third node Z and likely to be absent from the model can be ordered according to their rank $R(XY;Z|\{U_i\})$ defined as,

$$R(XY;Z|\{U_i\}) = \max_Z \big(S_{\mathsf{lb}}(Z;XY|\{U_i\})\big) \tag{17}$$

Then, Z can be iteratively added to the set of contributing nodes (*i.e.* $\{U_i\} \leftarrow \{U_i\} + Z$) of the top edge $XY = \mathrm{argmax}_{XY} R(XY;Z|\{U_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph.

Implementing this local optimization scheme, the 3off2 algorithm eventually learns the network skeleton by collecting the nodes of the separation sets one-by-one, instead of exploring the full combinatorics of sepset composition without any likelihood guidance. Indeed, the 3off2 scheme amounts to identify $\{U_i\}$ by "taking off" iteratively the "most likely" conditional 3-point information from each 2-point information as,

$$\begin{aligned}I(X;Y|\{U_i\}_n) = I(X;Y) &- I(X;Y;U_1)\\ &- I(X;Y;U_2|U_1) - \cdots\\ &- I(X;Y;U_n|\{U_i\}_{n-1})\end{aligned}$$

or equivalently between the shifted 2-point and 3-point information terms,

$$\begin{aligned}I'(X;Y|\{U_i\}_n) = I'(X;Y) &- I'(X;Y;U_1)\\ &- I'(X;Y;U_2|U_1) - \cdots\\ &- I'(X;Y;U_n|\{U_i\}_{n-1})\end{aligned}$$

This leads to the following Algorithm 1 for the reconstruction of the graph skeleton using the 3off2 scheme.

Note, in particular, that the 3off2 scheme to reconstruct graph skeleton is solely based on identifying structural independencies, which can also be applied to graphical models for undirected Markov networks.

Algorithm 1: 3off2 Skeleton Reconstruction

```
In:   observational data of finite size $N$
Out:  skeleton of causal graph $\mathcal{G}$

Initiation
Start with complete undirected graph
forall edges $XY$ do
    if $I'(X;Y)<0$ then
        $XY$ edge is non-essential and removed
        separation set of $XY$: $\mathrm{Sep}_{XY} = \emptyset$
    else
        find the most contributing node $Z$
        neighbor of $X$ or $Y$ and compute 3off2 rank,
        $R(XY;Z|\emptyset)$
    end
end

Iteration
while $\exists$ $XY$ edge with $R(XY;Z|\{U_i\}) > 1/2$ do
    for edge $XY$ with highest rank $R(XY;Z|\{U_i\})$ do
        expand contributing set $\{U_i\} \leftarrow \{U_i\} + Z$
        if $I'(X;Y|\{U_i\})<0$ then
            $XY$ edge is non-essential and removed
            separation set of $XY$: $\mathrm{Sep}_{XY} = \{U_i\}$
        else
            find next most contributing node $Z$
            neighbor of $X$ or $Y$ and compute new
            3off2 rank: $R(XY;Z|\{U_i\})$
        end
        sort the 3off2 rank list $R(XY;Z|\{U_i\})$
    end
end
```

Then, given the skeleton obtained from Algorithm 1, Eqs.13 and 14 lead to the following Proposition 7 and Proposition 8 for the orientation and propagation rules of unshielded triples, which are equivalent to Proposition 5 and Proposition 6 but for underlying DAG models (assuming no latent variables) and for finite datasets with the corresponding probabilities for the initiation/propagation of orientations.

Proposition 7 (Significantly negative conditional 3-point information as robust statistical evidence of causality in finite datasets). *Assuming that the underlying graphical model is a DAG $\mathcal{G}$ on V, $\forall X,Y,Z \in V$ and $\forall \{U_i\} \subseteq V_{\setminus\{X,Y,Z\}}$ s.t. $I'(X;Y|\{U_i\}) < 0$ (i.e. no XY edge) and $I'(X;Y;Z|\{U_i\}) < 0$ then,*

i. if X,Y,Z form an unshielded triple, $X \circ\!\!-\!\!\circ Z \circ\!\!-\!\!\circ Y$, then it should be oriented as $X \rightarrow Z \leftarrow Y$, with probabilities,

$$P^{\circ}_{X\rightarrow Z} = P^{\circ}_{Y\rightarrow Z} = \frac{1+e^{NI'(X;Y;Z|\{U_i\})}}{1+3e^{NI'(X;Y;Z|\{U_i\})}}$$

ii. similarly, if X,Y,Z form an unshielded triple, with one already known converging arrow, $X \rightarrow Z \circ\!\!-\!\!\circ Y$, with probability $P_{X\rightarrow Z} > P^{\circ}_{X\rightarrow Z}$, then the second edge should be oriented to form a v-structure, $X \rightarrow Z \leftarrow Y$, with probability,

$$P_{Y\rightarrow Z} = P_{X\rightarrow Z}\left(\frac{1}{1+e^{NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2}$$

Proof. The implications $(i.)$ and $(ii.)$ rely on Eq.14 to estimate the probability that the two edges form a collider. We start proving $(ii.)$ using the probability decomposition formula:

$$\begin{aligned} P_{Y\rightarrow Z} &= P_{X\rightarrow Z}\frac{P_{X\rightarrow Z\leftarrow Y}}{P_{X\rightarrow Z\leftarrow Y}+P_{X\rightarrow Z\rightarrow Y}} \\ &\quad + (1-P_{X\rightarrow Z})\frac{P_{X\leftarrow Z\leftarrow Y}}{P_{X\leftarrow Z\leftarrow Y}+P_{X\leftarrow Z\rightarrow Y}} \\ &= P_{X\rightarrow Z}\left(\frac{1}{1+e^{NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2} \end{aligned}$$

which also leads to $(i.)$ if one assumes $P_{X\rightarrow Z} = P_{Y\rightarrow Z}$ by symmetry in absence of prior information on these orientations. □

Following the rationale of constraint-based approaches, it is then possible to 'propagate' further the orientations downstream of colliders, using Eq.13 for positive (conditional) 3-point information. For simplicity and consistency, we only implement the propagation of orientation based on likelihood ratios, which can be quantified for finite datasets as proposed in the following Proposition 8. In particular, we do not extend the propagation rules (Meek, 1995) to inforce acyclic constraints that are necessary to have a complete reconstruction of the Markov equivalent class of the underlying DAG model.

Proposition 8 (robust 'propagation' of causality at unshielded triples with significantly positive conditional 3-pt information). *Assuming that the underlying graphical model is a DAG $\mathcal{G}$ on V, $\forall X,Y,Z \in V$ and $\forall \{U_i\} \subseteq V_{\setminus\{X,Y,Z\}}$*

s.t. $I'(X;Y|\{U_i\},Z) < 0$ *(i.e. no* XY *edge) and* $I'(X;Y;Z|\{U_i\}) > 0$*, then if* X,Y,Z *form an unshielded triple with one already known converging orientation,* $X \rightarrow Z \circ\!\!\!-\!\!\!* \, Y$*, with probability* $P_{X\rightarrow Z} > 1/2$*, this orientation should be 'propagated' to the second edge as* $X \rightarrow Z \rightarrow Y$*, with probability,*

$$P_{Z\rightarrow Y} = P_{X\rightarrow Z}\left(\frac{1}{1+e^{-NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2}$$

Proof. This results is shown using the probability decomposition formula,

$$\begin{aligned} P_{Z\rightarrow Y} &= P_{X\rightarrow Z}\frac{P_{X\rightarrow Z\rightarrow Y}}{P_{X\rightarrow Z\leftarrow Y} + P_{X\rightarrow Z\rightarrow Y}} \\ &\quad + (1 - P_{X\rightarrow Z})\frac{P_{X\leftarrow Z\rightarrow Y}}{P_{X\leftarrow Z\leftarrow Y} + P_{X\leftarrow Z\rightarrow Y}} \\ &= P_{X\rightarrow Z}\left(\frac{1}{1+e^{-NI'(X;Y;Z|\{U_i\})}} - \frac{1}{2}\right) + \frac{1}{2} \end{aligned}$$

□

Proposition 7 and Proposition 8 lead to the following Algorithm 2 for the orientation of unshielded triples of the graph skeleton obtained from Algorithm 1.

Algorithm 2: 3off2 Orientation / Propagation Step

In: Graph skeleton from Algorithm 1 and corresponding conditional 3-point information $I'(X;Y;Z|\{U_i\})$.

Out: Partially oriented causal graph $\mathcal{G}$ with edge orientation probabilities.

```
3off2 Orientation / Propagation Step

sort list of unshielded triples, $\mathcal{L}_c = \{\langle X,Z,Y\rangle_{X\not- Y}\}$,
in decreasing order of their orientation/propagation
probability initialized at 1/2 and computed from:
    - (i.) Proposition 7, if $I'(X;Y;Z|\{U_i\})<0$, or
    - (ii.) Proposition 8, if $I'(X;Y;Z|\{U_i\})>0$
repeat
    Take $\langle X,Z,Y\rangle_{X\not- Y} \in \mathcal{L}_c$ with highest orientation
    / propagation probability $> 1/2$.
    if $I'(X;Y;Z|\{U_i\}) < 0$ then
        Orient/propagate edge direction(s) to form a
        v-structure $X \rightarrow Z \leftarrow Y$ with probabilities
        $P_{X\rightarrow Z}$ and $P_{Y\rightarrow Z}$ given by Proposition 7.
    else
        Propagate second edge direction to form a
        non-v-structure $X \rightarrow Z \rightarrow Y$ assigning
        probability $P_{Z\rightarrow Y}$ from Proposition 8.
    end
    Apply new orientation(s) and sort remaining list
    of unshielded triples $\mathcal{L}_c \leftarrow \mathcal{L}_c \backslash \langle X,Z,Y\rangle_{X\not- Y}$ after
    updating propagation probabilities.
until no additional orient./propa. probability $>1/2$ ;
```

2.3 APPLICATIONS TO CAUSAL GRAPH BENCHMARKS

We have tested the 3off2 method on a range of benchmark networks of 50 nodes with up to 160 edges generated with the causal modeling tool Tetrad IV (http://www.phil.cmu.edu/tetrad). The average connectivity $\langle k\rangle$ of these benchmark networks ranges between 1.6 to 6.4, and the average maximal in/out-degree between 3.2 to 8.8 (see Table S1 for a detailed description). The evaluation metrics are the Precision, $Prec = TP/(TP + FP)$, the Recall, $Rec = TP/(TP + FN)$ and the $F\!-\!score = 2Prec.Rec/(Prec + Rec)$. However, in order to take into account the orientation/non-orientation of edges in the predicted networks and compare them with the CPDAG of the benchmark graphs, we define orientation-dependent counts as, $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$, where $TP_{\text{misorient}}$ corresponds to all true positive edges of the skeleton with different orientation/non-orientation status as in the CPDAG reference.

The first methods used for comparison with 3off2 are the PC-stable algorithm (Colombo and Maathuis, 2014) with conservative (Ramsey et al, 2006) or majority orientation rules, implemented in the `pcalg` package (Kalisch et al., 2012; Kalisch and Bühlmann, 2008) and the hybrid method MMHC combining constraint-based skeleton and Bayesian orientation (Tsamardinos, Brown, and Aliferis, 2006), implemented in the `bnlearn` package (Scutari, 2010). Figs. 1-5 give the average CPDAG comparison results over 100 dataset replicates from 5 different benchmark networks (Table S1). The causal graphical models predicted by the 3off2 method are obtained using either the MDL/BIC or the NML complexities (see Supplementary Methods). Figs. S1-S6 provide additional results on the prediction of the network skeletons and execution times. The PC and MMHC results are shown, Figs. 1-5, for an independence test parameter $\alpha = 0.1$, as reducing α tends to worsen the CPDAG F-score for benchmark networks with $\langle k\rangle \geqslant 1.6$ (Figs. S7-S18). All in all, we found that the 3off2 method outperforms both PC-stable and MMHC methods on all tested datasets, Figs. 1-5.

Additional comparisons were obtained with Bayesian inference implemented in the `bnlearn` package (Scutari, 2010), using AIC, BDe and BIC/MDL scores and hill-climbing heuristics with 30 to 100 random restarts, Figs. S19-S30. 3off2 reaches equivalent or significantly better F-scores than Bayesian hill-climbing for all dataset sizes on benchmark networks up to 120 edges ($\langle k\rangle \leqslant 4.8$). In particular, 3off2 with MDL scores reaches excellent F-scores on sparse networks (Figs. S19 & S20) and keeps one of the best F-scores over all sample sizes for less sparse networks when combined to NML complexity (Figs. S21 & S22). For somewhat denser networks ($\langle k\rangle \simeq 5$), the 3off2 F-score appears slightly lower than for Bayesian inference methods, Fig. S23, although it eventually becomes equivalent for large datasets ($N \geqslant 1000$).

On denser networks ($\langle k\rangle \geqslant 5-6$), Bayesian inference exhibits better F-scores than 3off2, in particular with AIC score, Fig. S24. However, the good performance with AIC strongly relies on its high Recall (but low Precision), due to its very small penalty term on large datasets, which makes it favor more complex networks (Figs. S24) but perform very poorly on sparse graphs (Figs. S19-S21). By contrast, the reconstruction of dense networks is impeded with the 3off2 scheme, as it is not always possible to uncover structural independencies, $I(X;Y|\{U_i\}_n)\simeq 0$, in dense graphs through an ordered set $\{U_i\}_n$ with only positive conditional 3-point information, $I'(X;Y;U_k|\{U_i\}_{k-1}) > 0$. Indeed in complex graphs, there are typically many indirect paths $X \rightarrow U_j \rightarrow Y$ between unconnected node pairs (X, Y). At the beginning of the pruning process, this is prone to suggest likely v-structures $X \rightarrow Y \leftarrow U_j$, instead of the correct non-v-structures, $X \rightarrow U_j \rightarrow Y$ (for instance if $I(X;U_j) \ll I(X;Y)$, $I(X;U_j) \ll I(U_j;Y)$ and $I(X;U_j) - I(X;U_j|Y) = I(X;Y;U_j) < 0$, for all j). Such elimination of FN edge $X \rightarrow U_j$ and conservation of FP $X \rightarrow Y$ tend to decrease both Precision and Recall, although 3off2 remains significantly more robust than PC and MMHC, Fig. 5. Besides, for most practical applications on real life data, interpretable causal models should remain relatively sparse and avoid to display multiple indirected paths between unconnected nodes.

Finally, 3off2 running times on these benchmark networks are similar to MMHC and Bayesian hill-climbing heuristic methods (with 100 restarts) and 10 to 100 times faster than PC for large datasets, Figs. S1-S30.

3 DISCUSSION

In this paper, we propose to combine constraint-based and score-based frameworks to improve network reconstruction. Earlier hybrid methods, including MMHC, have also attempted to exploit the best of these two types of inference approaches by combining the robustness of Bayesian scores with the attractive conceptual features of constraint-based approaches (Dash

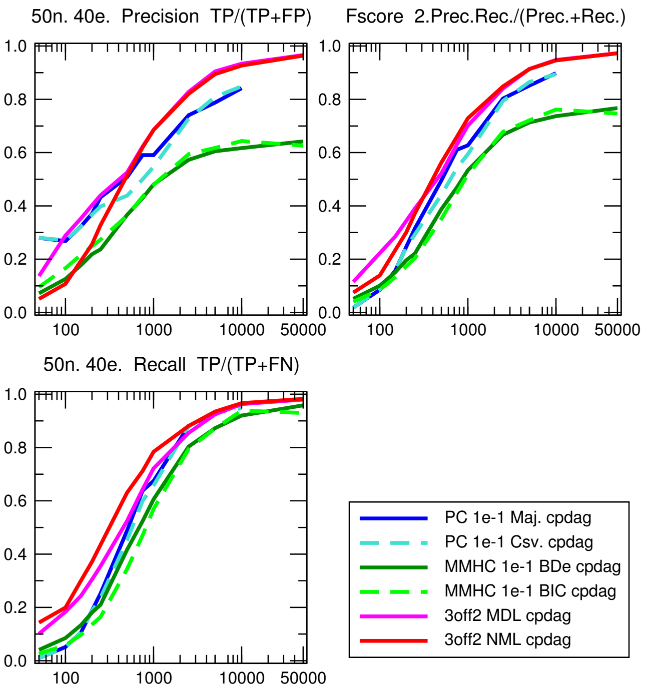

Figure 1: **CPDAG comparison between 3off2, PC-stable and MMHC.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k\rangle = 1.6$, $\langle k^{in}_{\max}\rangle = 3.2$, $\langle k^{out}_{\max}\rangle = 3.6$. PC-stable benchmarks were tested up to N=10,000 due to their sharp increase in execution time, see Figs. S7-S12.

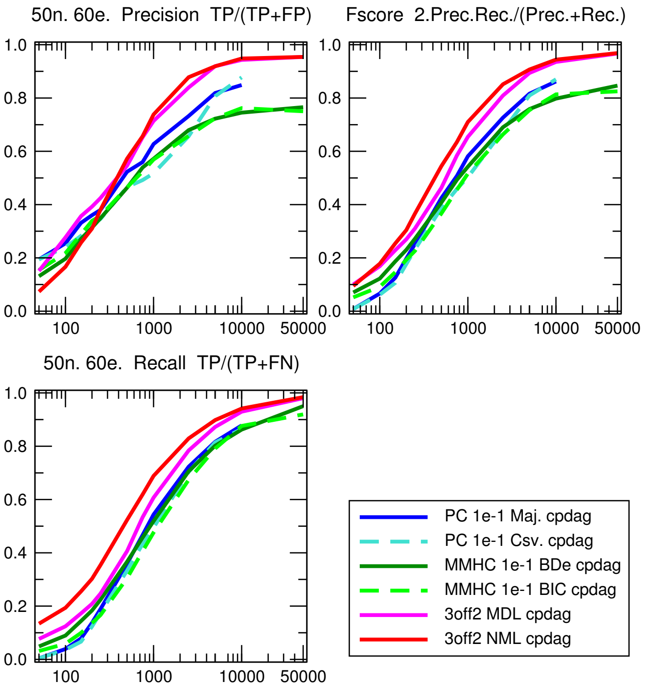

Figure 2: **CPDAG comparison between 3off2, PC-stable and MMHC.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k\rangle = 2.4$, $\langle k^{in}_{\max}\rangle = 4.6$, $\langle k^{out}_{\max}\rangle = 3.6$. PC-stable benchmarks were tested up to N=10,000 due to their sharp increase in execution time, see Figs. S7-S12.

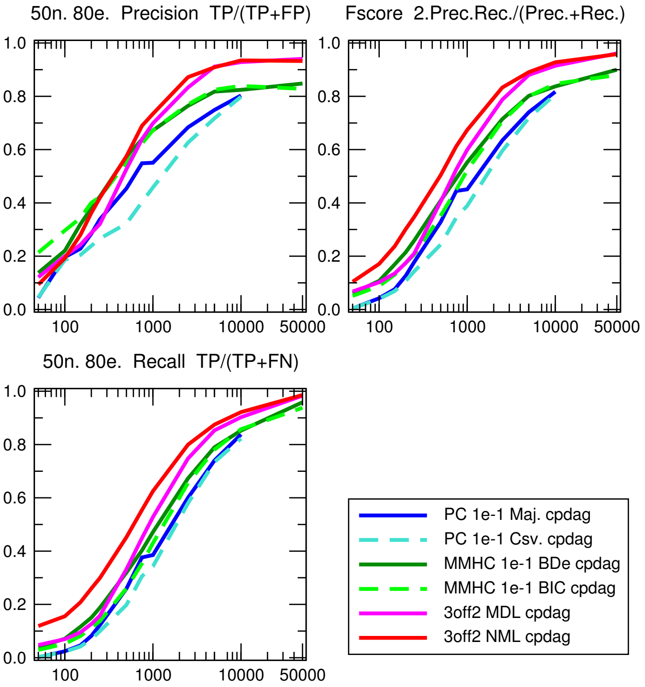

Figure 3: **CPDAG comparison between 3off2, PC-stable and MMHC.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k_{\max}^{in} \rangle = 4.8$, $\langle k_{\max}^{out} \rangle = 5.6$. PC-stable benchmarks were tested up to N=10,000 due to their sharp increase in execution time, see Figs. S7-S12.

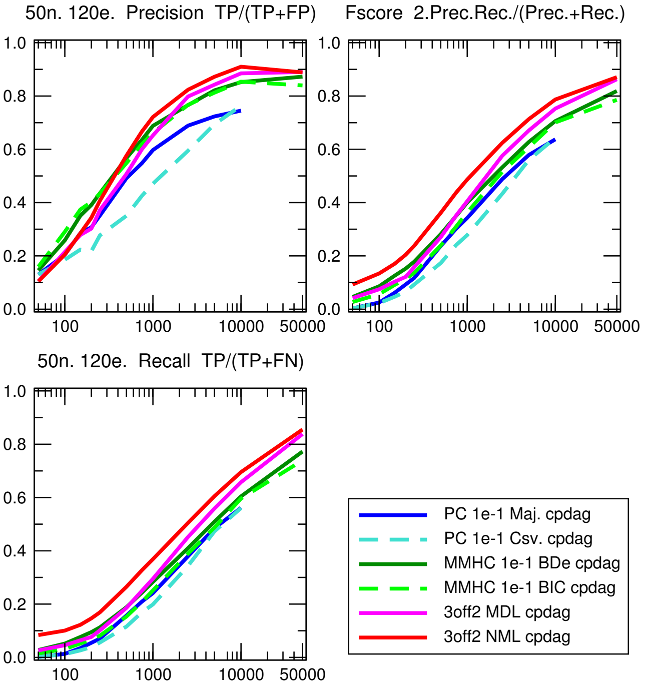

Figure 4: **CPDAG comparison between 3off2, PC-stable and MMHC.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k_{\max}^{in} \rangle = 8.8$, $\langle k_{\max}^{out} \rangle = 7.2$. PC-stable benchmarks were tested up to N=10,000 due to their sharp increase in execution time, see Figs. S7-S12.

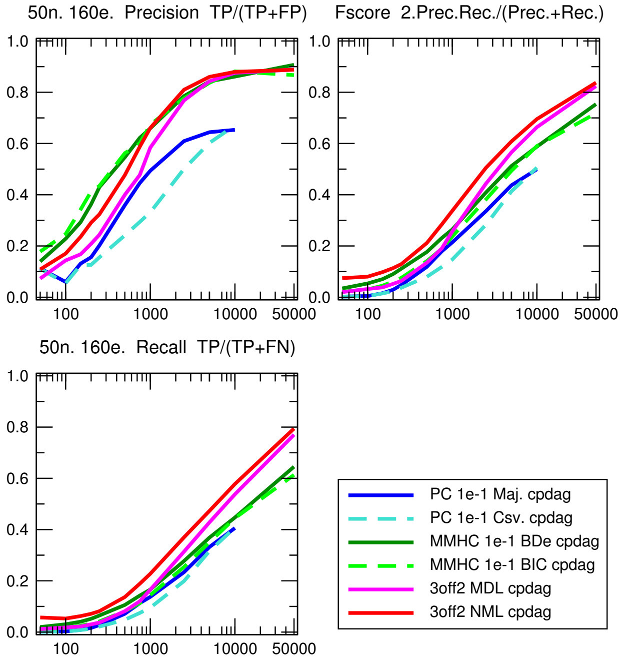

Figure 5: **CPDAG comparison between 3off2, PC-stable and MMHC.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k_{\max}^{in} \rangle = 8.6$, $\langle k_{\max}^{out} \rangle = 8.6$.

and Druzdzel, 1999; Tsamardinos, Brown, and Aliferis, 2006; Cano, Gomez-Olmedo, and Moral, 2008; Claassen and Heskes, 2012). In particular, (Dash and Druzdzel, 1999) have proposed to exploit an intrinsic weakness of the PC algorithm, its sensitivity to the order in which conditional independencies are tested on finite data, to rank these different order-dependent PC predictions with Bayesian scores. More recently, (Claassen and Heskes, 2012) have also combined constraint-based and Bayesian approaches to improve the reliability of causal inference. They proposed to use Bayesian scores to directly assess the reliability of conditional independencies by *summing* the likelihoods over compatible graphs. By contrast, we propose to use Bayesian scores to progressively uncover the best supported conditional independencies, by iteratively "taking off" the most likely indirect contributions of conditional 3-point information from every 2-point (mutual) information of the causal graph. In addition, using likelihood ratios (Eqs.11 & 12) instead of likelihood sums (Claassen and Heskes, 2012) circumvents the need to score conditional independencies over a potentially intractable number of compatible graphs.

All in all, we found that 3off2 outperforms constraint-based, search-and-score and earlier hybrid methods on a range of benchmark networks, while displaying similar running times as hill-climbing heuristic methods.

Acknowledgements

S.A. acknowledges support from Ministry of Research and Association pour la Recherche contre le Cancer.

References

Cano, A.; Gomez-Olmedo, M.; and Moral, S. 2008. A score based ranking of the edges for the pc algorithm. In *Proceedings of the European Workshop on Probabilistic Graphical Models (PGM)*, 41–48.

Chickering, D. M. 2002. Learning equivalence classes of bayesian-network structures. *Journal of Machine Learning Research* 2:445–498.

Claassen, T., and Heskes, T. 2012. A bayesian approach to constraint based causal inference. In *In Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 207–216. Morgan Kaufmann.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15:3741–3782.

Cooper, G. F., and Herskovits, E. 1992. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* 9(4):309–347.

Dash, D., and Druzdzel, M. J. 1999. A hybrid anytime algorithm for the construction of causal models from sparse data. In *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence*, 142–149. Morgan Kaufmann.

de Campos, L. M. 2006. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research* 7:2149–2187.

Hansen, M. H., and Yu, B. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96:746–774.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20(3):197–243. Available as Technical Report MSR-TR-94-09.

Kalisch, M., and Bühlmann, P. 2008. Robustification of the pc-algorithm for directed acyclic graphs. *Journal Of Computational And Graphical Statistics* 17(4):773–789.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; and Bühlmann, P. 2012. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software* 47(11):1–26.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press.

Kontkanen, P., and Myllymäki, P. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* 103(6):227–233.

Meek, C. 1995. Causal inference and causal explanation with background knowledge. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU. Morgan Kaufmann. 403–418.

Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.* 441–452.

Pearl, J. 2009. *Causality: models, reasoning and inference.* Cambridge University Press, 2nd edition.

Ramsey, J.; Spirtes, P.; and Zhang, J. 2006. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, UAI2015, 401–408. Oregon, USA: AUAI Press.

Rebane, G., and Pearl, J. 1988. The recovery of causal poly-trees from statistical data. *Int. J. Approx. Reasoning* 2(3):341.

Reichenbach, H. 1956. *The Direction of Time.* California library reprint series. University of California Press.

Rissanen, J., and Tabus, I. 2005. Kolmogorovs structure function in mdl theory and lossy data compression. In *Adv. Min. Descrip. Length Theory Appl.* MIT Press. Chap. 10.

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

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymäki, P. 2008. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008).* IEEE Press. invited paper.

Sanov, I. 1957. On the probability of large deviations of random variables. *Mat. Sbornik* 42:11–44.

Scutari, M. 2010. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35(3):1–22.

Shtarkov, Y. M. 1987. Universal sequential coding of single messages. *Problems of Information Transmission (Translated from)* 23(3):3–17.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62–72.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search.* The MIT Press, Cambridge, Massachusetts, 2nd edition.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* 65(1):31–78.

SUPPLEMENTARY INFORMATION

for the manuscript "Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information"

Séverine Affeldt, Hervé Isambert
Institut Curie, Research Center, UMR168, 26 rue d'Ulm, 75005, Paris France;
and Université Pierre et Marie Curie, 4 Place Jussieu, 75005, Paris, France
herve.isambert@curie.fr

SUPPLEMENTARY METHODS

Complexity of graphical models

The complexity $k_{\mathcal{G},\mathcal{D}}$ of a graphical model is related to the normalization constant $Z(\mathcal{G},\mathcal{D})$ of its maximum likelihood as $k_{\mathcal{G},\mathcal{D}} = \log Z(\mathcal{G},\mathcal{D})$,

$$\mathcal{L}_\mathcal{G} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = e^{-NH(\mathcal{G},\mathcal{D})-k_{\mathcal{G},\mathcal{D}}} \tag{1}$$

For Bayesian networks with decomposable entropy, *i.e.* $H(\mathcal{G},\mathcal{D}) = \sum_i H(X_i|\{\text{Pa}_{X_i}\})$, it is convenient to use decomposable complexities, $k_{\mathcal{G},\mathcal{D}} = \sum_i k_{X_i|\{\text{Pa}_{X_i}\}}$,

$$\mathcal{L}_\mathcal{G} = e^{-N\sum_i H(X_i|\{\text{Pa}_{X_i}\})-\sum_i k_{X_i|\{\text{Pa}_{X_i}\}}} \tag{2}$$

such that the comparison between alternative models $\mathcal{G}$ and $\mathcal{G}_{\backslash X\to Y}$ (*i.e.* $\mathcal{G}$ with one missing edge $X \to Y$) leads to a simple local increment of the score,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash X\to Y}}}{\mathcal{L}_\mathcal{G}} = e^{-NI(X;Y|\{\text{Pa}_Y\}_{\backslash X})+\Delta k_{Y|\{\text{Pa}_Y\}_{\backslash X}}} \tag{3}$$

$$I(X;Y|\{\text{Pa}_Y\}_{\backslash X}) = H(Y|\{\text{Pa}_Y\}_{\backslash X}) - H(Y|\{\text{Pa}_Y\}) \geqslant 0$$
$$\Delta k_{Y|\{\text{Pa}_Y\}_{\backslash X}} = k_{Y|\{\text{Pa}_Y\}} - k_{Y|\{\text{Pa}_Y\}_{\backslash X}} \geqslant 0$$

A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria (Rissanen, 1978; Hansen and Yu, 2001),

$$k^{\text{MDL}}_{Y|\{\text{Pa}_Y\}} = \frac{1}{2}(r_y - 1)\prod_j^{\text{Pa}_Y} r_j \,\log N \tag{4}$$

$$\Delta k^{\text{MDL}}_{Y|\{\text{Pa}_Y\}_{\backslash X}} = \frac{1}{2}(r_x-1)(r_y - 1)\prod_j^{\text{Pa}_{y\backslash X}} r_j \,\log N \tag{5}$$

where r_x, r_y and r_j are the number of levels of each variable, x, y and j. The MDL complexity, Eq.4, is simply related to the normalisation constant of the normal distribution reached in the asymptotic limit of a large dataset $N \to \infty$ (Central Limit Theorem). The MDL complexity can also be derived from the Stirling approximation on the Bayesian measure (Schwarz, 1978; Bouckaert, 1993). Yet, in practice, this central limit distribution is only reached for very large datasets, as some of the least-likely $(r_y-1)\prod_j r_j$ combinations of states of variables are in fact rarely (if ever) sampled in typical finite datasets. As a result, the MDL complexity criteria tends to underestimate the relevance of edges connecting variables with many levels, r_i, leading to the removal of false negative edges.

To avoid such biases with finite datasets, the normalisation of the maximum likelihood can be done over all possible datasets with the same number N of data points. This corresponds to the (universal) Normalized Maximum Likelihood (NML) criteria (Shtarkov, 1987; Rissanen and Tabus, 2005; Kontkanen and Myllymäki, 2007; Roos et al., 2008),

$$\mathcal{L}_\mathcal{G} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(\mathcal{G},\mathcal{D}')}} = e^{-NH(\mathcal{G},\mathcal{D})-k^{\text{NML}}_{\mathcal{G},\mathcal{D}}} \tag{6}$$

We introduce here the factorized version of the NML criteria (Kontkanen and Myllymäki, 2007; Roos et al., 2008) which corresponds to a decomposable NML score, $k^{\text{NML}}_{\mathcal{G},\mathcal{D}} = \sum_{X_i} k^{\text{NML}}_{X_i|\{\text{Pa}_{X_i}\}}$, defined as,

$$k^{\text{NML}}_{Y|\{\text{Pa}_Y\}} = \sum_j^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} \tag{7}$$

$$\Delta k^{\text{NML}}_{Y|\{\text{Pa}_Y\}_{\backslash X}} = \sum_j^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} - \sum_{j'}^{q_y/r_x} \log \mathcal{C}^{r_y}_{N_{yj'}} \tag{8}$$

where N_{yj} is the number of data points corresponding to the jth state of the parents of Y, $\{\text{Pa}_Y\}$, and $N_{yj'}$ the number of data points corresponding to the j'th state of the parents of Y, excluding X, $\{\text{Pa}_Y\}_{\backslash X}$. Hence, the factorized NML score for each node X_i corresponds to a separate normalisation for each state

$j = 1, ..., q_i$ of its parents and involving exactly N_{ij} data points of the finite dataset,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(X_i|\{\mathrm{Pa}_{X_i}\}) - \sum_j^{q_i} \log \mathcal{C}^{r_i}_{N_{ij}}} \tag{9}$$

$$= e^{N\sum_i \sum_j^{q_i} \sum_k^{r_i} \frac{N_{ijk}}{N} \log\left(\frac{N_{ijk}}{N_{ij}}\right) - \sum_j^{q_i} \log \mathcal{C}^{r_i}_{N_{ij}}} \tag{10}$$

$$= \prod_i \prod_j^{q_i} \frac{\prod_k^{r_i} \left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}}}{\mathcal{C}^{r_i}_{N_{ij}}} \tag{11}$$

where N_{ijk} corresponds to the number of data points for which the ith node is in its kth state and its parents in their jth state, with $N_{ij} = \sum_k^{r_i} N_{ijk}$. The universal normalization constant $\mathcal{C}^r_n$ is then obtained by averaging over all possible partitions of the n data points into a maximum of r subsets, $\ell_1 + \ell_2 + \cdots + \ell_r = n$ with $\ell_k \geqslant 0$,

$$\mathcal{C}^r_n = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n} \frac{n!}{\ell_1!\ell_2!\cdots\ell_r!} \prod_{k=1}^{r} \left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{12}$$

which can in fact be computed in linear-time using the following recursion (Kontkanen and Myllymäki, 2007),

$$\mathcal{C}^r_n = \mathcal{C}^{r-1}_n + \frac{n}{r-2}\mathcal{C}^{r-2}_n \tag{13}$$

with $\mathcal{C}^r_0 = 1$ for all r, $\mathcal{C}^1_n = 1$ for all n and applying the general formula Eq.12 for $r = 2$,

$$\mathcal{C}^2_n = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{14}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) (Szpankowski, 2001; Kontkanen et al., 2003; Kontkanen, 2009),

$$\mathcal{C}^2_n = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{15}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{16}$$

Then, following the rationale of constraint-based approaches, we can reformulate the likelihood ratio of Eq. 3 by replacing the parent nodes $\{\mathrm{Pa}_Y\}_{\backslash X}$ in the conditional mutual information, $I(X;Y|\{\mathrm{Pa}_Y\}_{\backslash X})$, with an unknown separation set $\{U_i\}$ to be learnt simultaneously with the missing edge candidate XY,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash XY|\{U_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(X;Y|\{U_i\}) + k_{X;Y|\{U_i\}}} \tag{17}$$

where we have also transformed the asymmetric parent-dependent complexity difference, $\Delta k_{Y|\{\mathrm{Pa}_Y\}_{\backslash X}}$, into a $\{U_i\}$-dependent complexity term, $k_{X;Y|\{U_i\}}$, with the same XY-symmetry as $I(X;Y|\{U_i\})$,

$$k^{\mathsf{MDL}}_{X;Y|\{U_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i} \log N \tag{18}$$

$$k^{\mathsf{NML}}_{X;Y|\{U_i\}} = \frac{1}{2}\sum_{j'}^{\{U_i\}} \Bigg(\sum_{k_x}^{r_x} \log \mathcal{C}^{r_y}_{N_{k_x j'}} - \log \mathcal{C}^{r_y}_{N_{j'}} + \sum_{k_y}^{r_y} \log \mathcal{C}^{r_x}_{N_{k_y j'}} - \log \mathcal{C}^{r_x}_{N_{j'}} \Bigg) \tag{19}$$

Note, in particular, that the MDL complexity term in Eq.18 is readily obtained from Eq.5 due to the Markov equivalence of the MDL score, corresponding to its XY-symmetry whenever $\{\mathrm{Pa}_Y\}_{\backslash X} = \{\mathrm{Pa}_X\}_{\backslash Y}$. By contrast, the factorized NML score, Eq.7, is not a Markov-equivalent score (although its non-factorized version, Eq.6, is Markov equivalent by definition). To circumvent this non-equivalence of factorized NML score, we propose to recover the expected XY-symmetry of $k^{\mathsf{NML}}_{X;Y|\{U_i\}}$ through the simple XY-symmetrization of Eq.8, leading to Eq.19.

References

Bouckaert, R. R. 1993. Probabilistic network construction using the minimum description length principle. *in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Clarke M, Kruse R, Moral S, eds)* 747:41–48.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15:3741–3782.

Hansen, M. H., and Yu, B. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96:746–774.

Kontkanen, P., and Myllymäki, P. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Inf. Process. Lett.* 103(6):227–233.

Kontkanen, P.; Buntine, W.; Myllymäki, P.; Rissanen, J.; and Tirri, H. 2003. Efficient computation of stochastic complexity. *in: C. Bishop, B. Frey (Eds.) Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics* 103:233–238.

Kontkanen, P. 2009. *Computationally Efficient Methods for MDL-Optimal Density Estimation and Data Clustering.* Ph.D. Dissertation.

Rissanen, J., and Tabus, I. 2005. Kolmogorovs structure function in mdl theory and lossy data compression. In *Adv. Min. Descrip. Length Theory Appl.* MIT Press. Chap. 10.

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

Roos, T.; Silander, T.; Kontkanen, P.; and Myllymäki, P. 2008. Bayesian network structure learning using factorized nml universal models. In *Proc. 2008 Information Theory and Applications Workshop (ITA-2008)*. IEEE Press. invited paper.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Shtarkov, Y. M. 1987. Universal sequential coding of single messages. *Problems of Information Transmission (Translated from)* 23(3):3–17.

Szpankowski, W. 2001. *Average case analysis of algorithms on sequences.* John Wiley & Sons.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* 65(1):31–78.

nodes	edges	$\langle k \rangle$	Model	$\langle k_{\max} \rangle$	$\langle k^{in}_{\max} \rangle$	$\langle k^{out}_{\max} \rangle$	N	Replicates
50	**20**	0.8	*1*	4	2	3	$[50-50,000]$	20
			2	4	2	2	$[50-50,000]$	20
			3	3	3	2	$[50-50,000]$	20
			4	3	3	2	$[50-50,000]$	20
			5	3	2	2	$[50-50,000]$	20
			Avg.	**3.4**	**2.4**	**2.2**		
50	**40**	1.6	*1*	5	3	5	$[50-50,000]$	20
			2	6	3	3	$[50-50,000]$	20
			3	5	3	3	$[50-50,000]$	20
			4	4	4	4	$[50-50,000]$	20
			5	5	3	3	$[50-50,000]$	20
			Avg.	**5**	**3.2**	**3.6**		
50	**60**	2.4	*1*	7	5	3	$[50-50,000]$	20
			2	6	6	3	$[50-50,000]$	20
			3	6	4	4	$[50-50,000]$	20
			4	6	5	3	$[50-50,000]$	20
			5	7	3	5	$[50-50,000]$	20
			Avg.	**6.4**	**4.6**	**3.6**		
50	**80**	3.2	*1*	7	5	7	$[50-50,000]$	20
			2	7	5	5	$[50-50,000]$	20
			3	6	5	5	$[50-50,000]$	20
			4	6	5	6	$[50-50,000]$	20
			5	6	4	5	$[50-50,000]$	20
			Avg.	**6.4**	**4.8**	**5.6**		
50	**120**	4.8	*1*	10	10	7	$[50-50,000]$	20
			2	13	10	7	$[50-50,000]$	20
			3	9	6	8	$[50-50,000]$	20
			4	13	9	7	$[50-50,000]$	20
			5	12	9	7	$[50-50,000]$	20
			Avg.	**11.4**	**8.8**	**7.2**		
50	**160**	6.4	*1*	12	10	9	$[50-50,000]$	20
			2	13	9	9	$[50-50,000]$	20
			3	14	7	9	$[50-50,000]$	20
			4	11	7	8	$[50-50,000]$	20
			5	11	10	8	$[50-50,000]$	20
			Avg.	**12.2**	**8.6**	**8.6**		

Table S1: **Description summary of the 30 benchmark networks used to evaluate the reconstruction methods.** The 30 benchmark networks of 50 nodes, and 20 to 160 edges, have been instantiated with the causal modeling tool Tetrad IV (http://www.phil.cmu.edu/tetrad/). For each model, 20 dataset replicates of size ranging between 50 and 50,000 were generated with Tetrad IV.

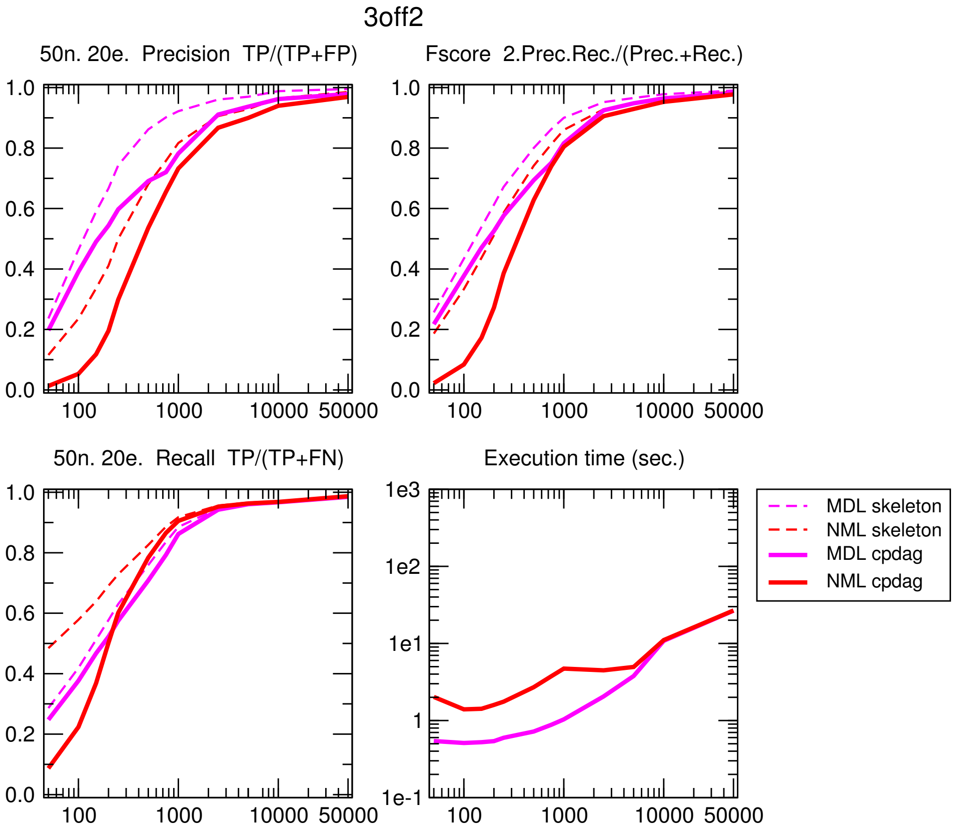

Figure S1: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

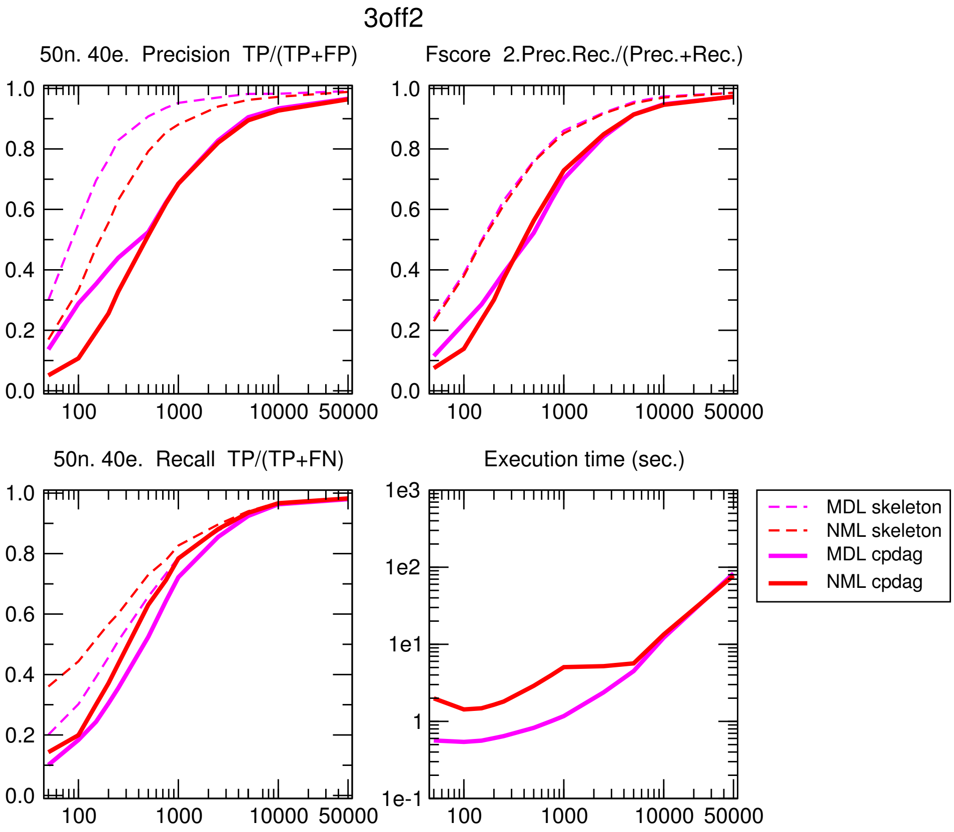

Figure S2: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

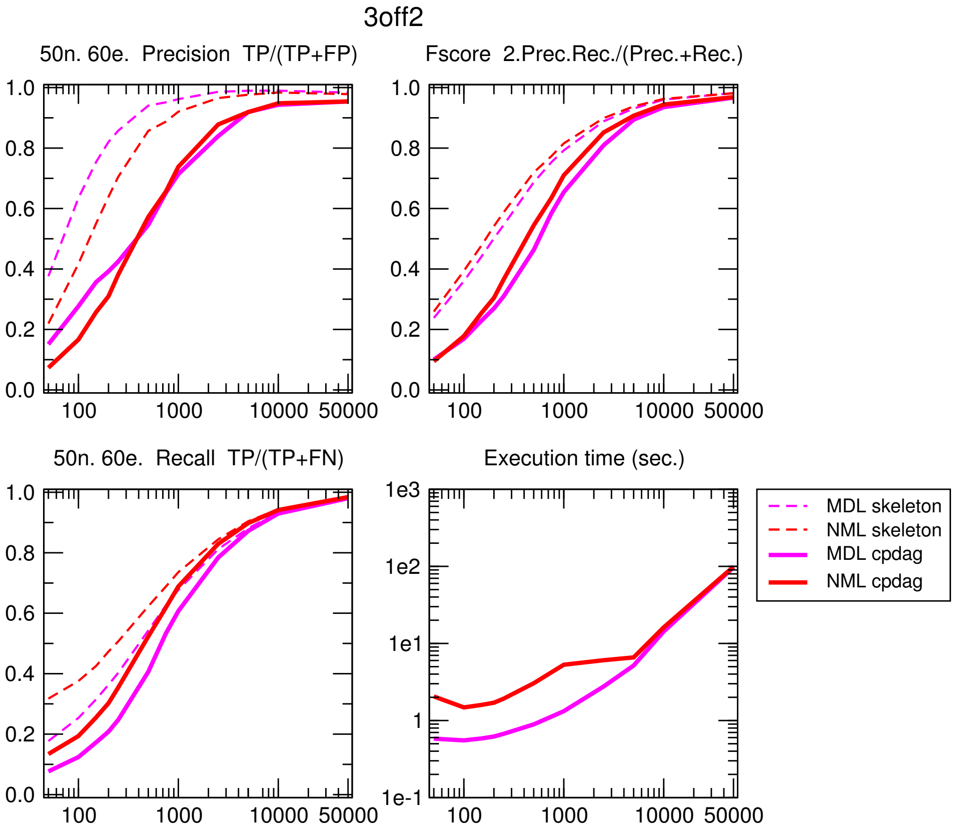

Figure S3: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

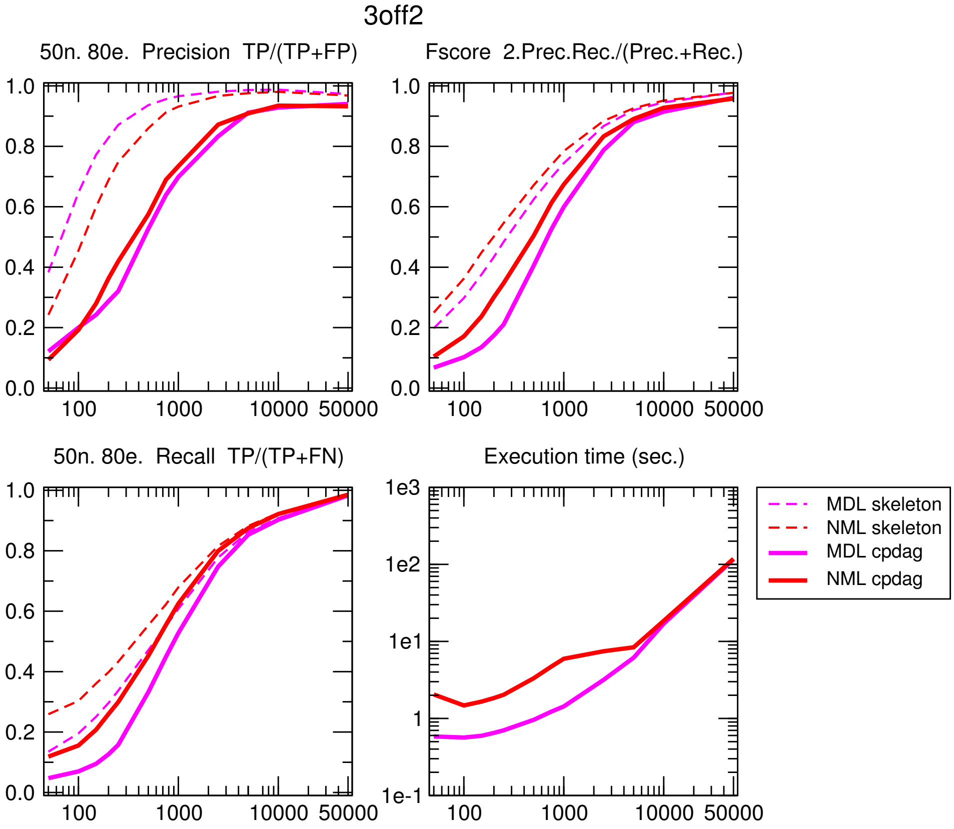

Figure S4: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

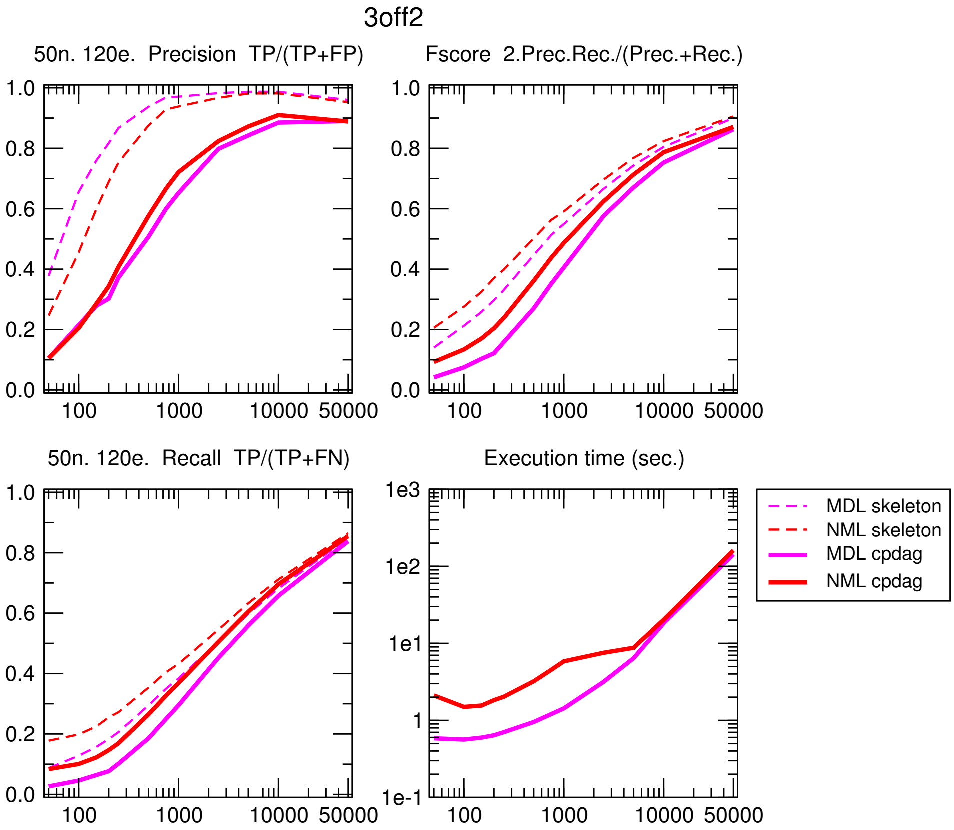

Figure S5: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

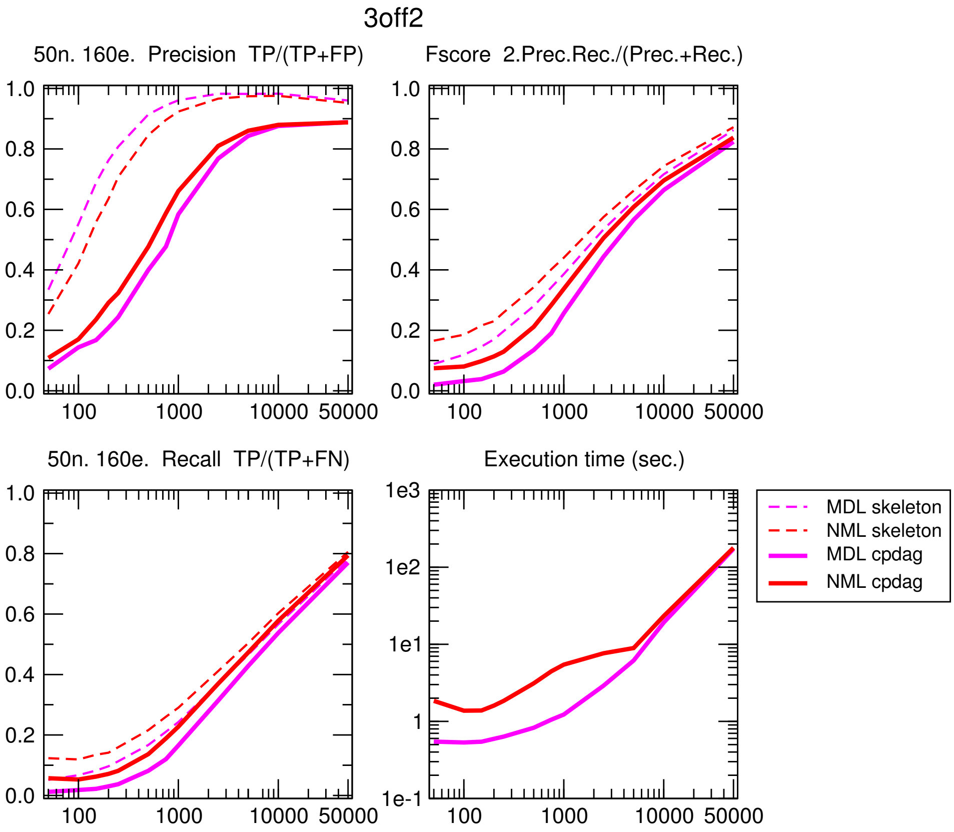

Figure S6: 3off2 **reconstruction, effect of complexity MDL and NML.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. The change of slop in execution time at sample size N=1000 for NML corresponds to the use of the Szpankowski approximation (see Supplementary Methods).

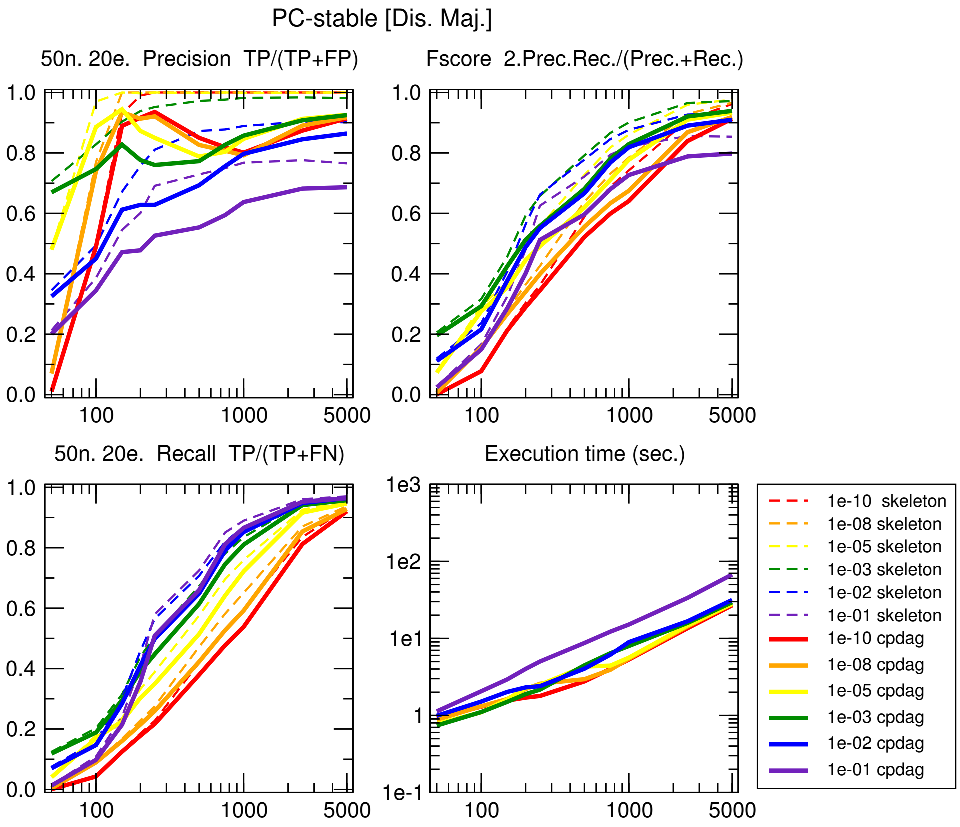

Figure S7: **PC, effect of independence test parameter** α**.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

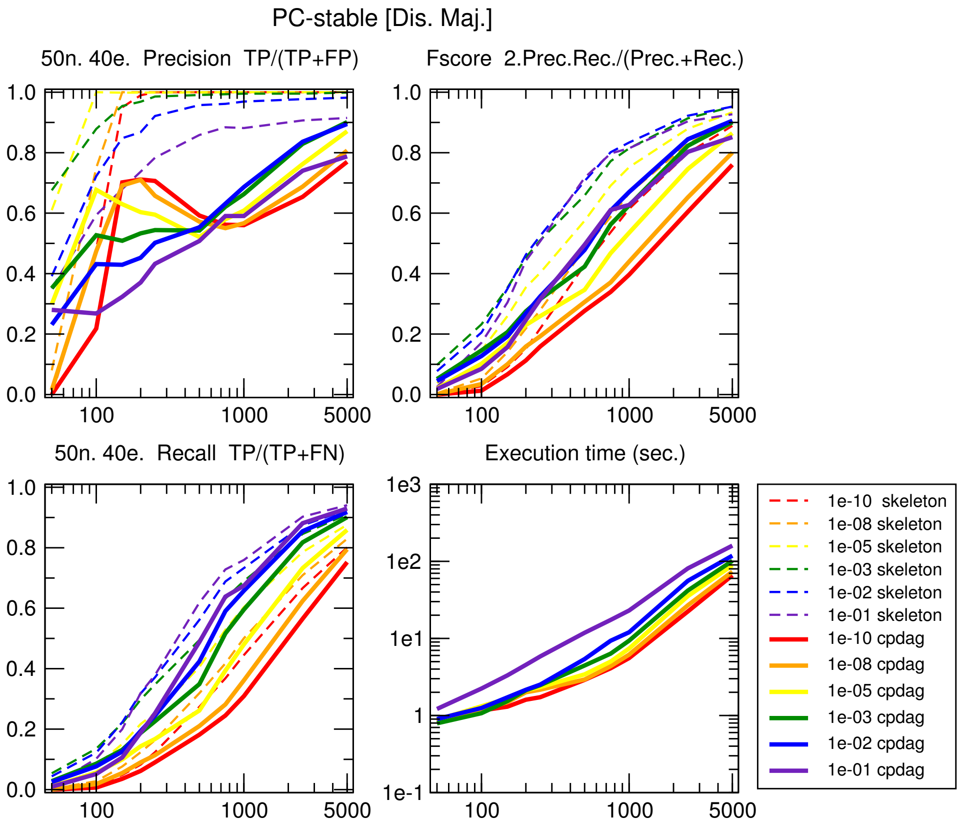

Figure S8: **PC, effect of independence test parameter** α**.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

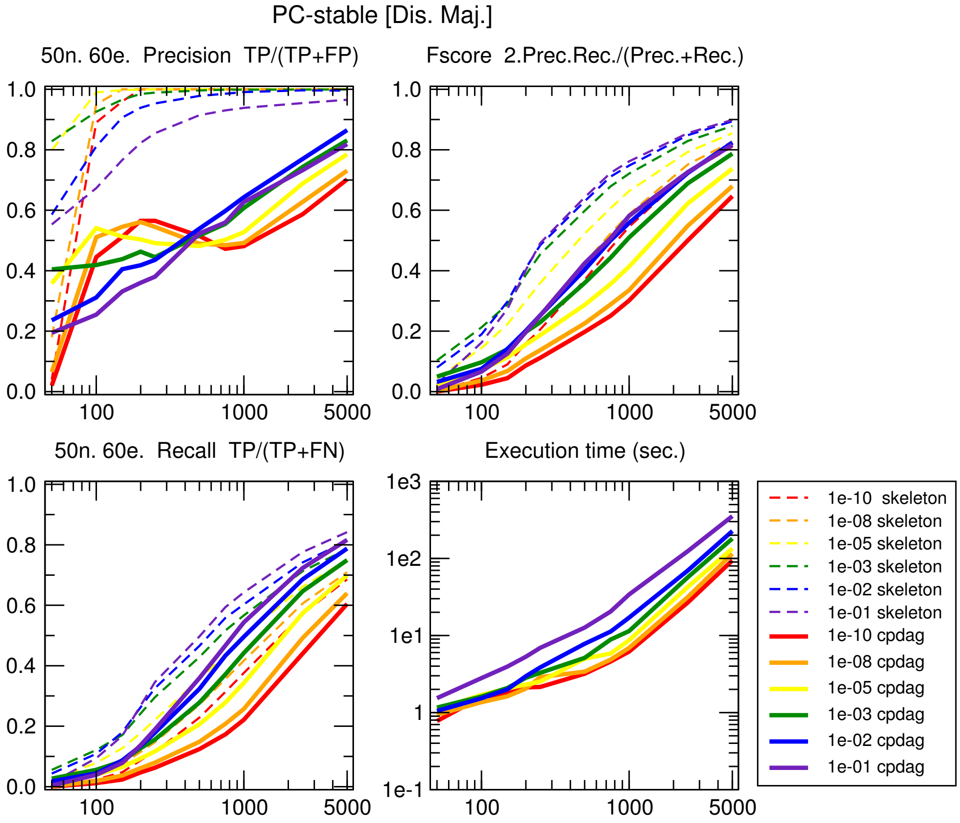

Figure S9: **PC, effect of independence test parameter** α**.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k_{\max}^{in} \rangle = 4.6$ and $\langle k_{\max}^{out} \rangle = 3.6$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

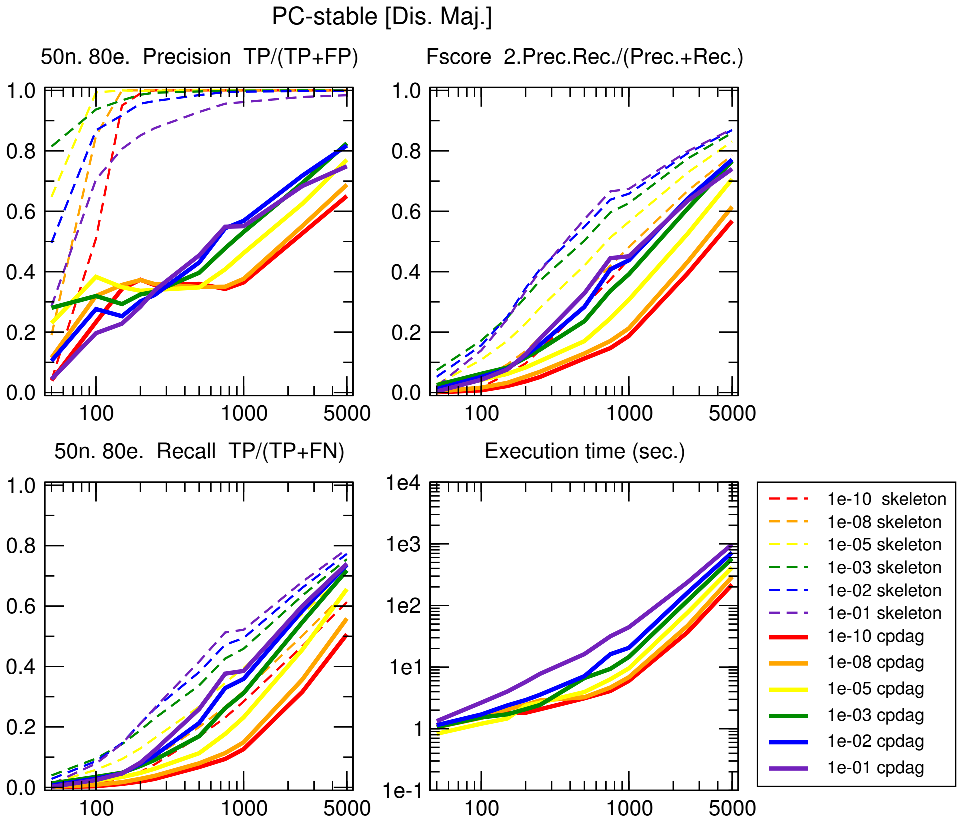

Figure S10: **PC, effect of independence test parameter** α**.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k_{\max}^{in} \rangle = 4.8$ and $\langle k_{\max}^{out} \rangle = 5.6$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

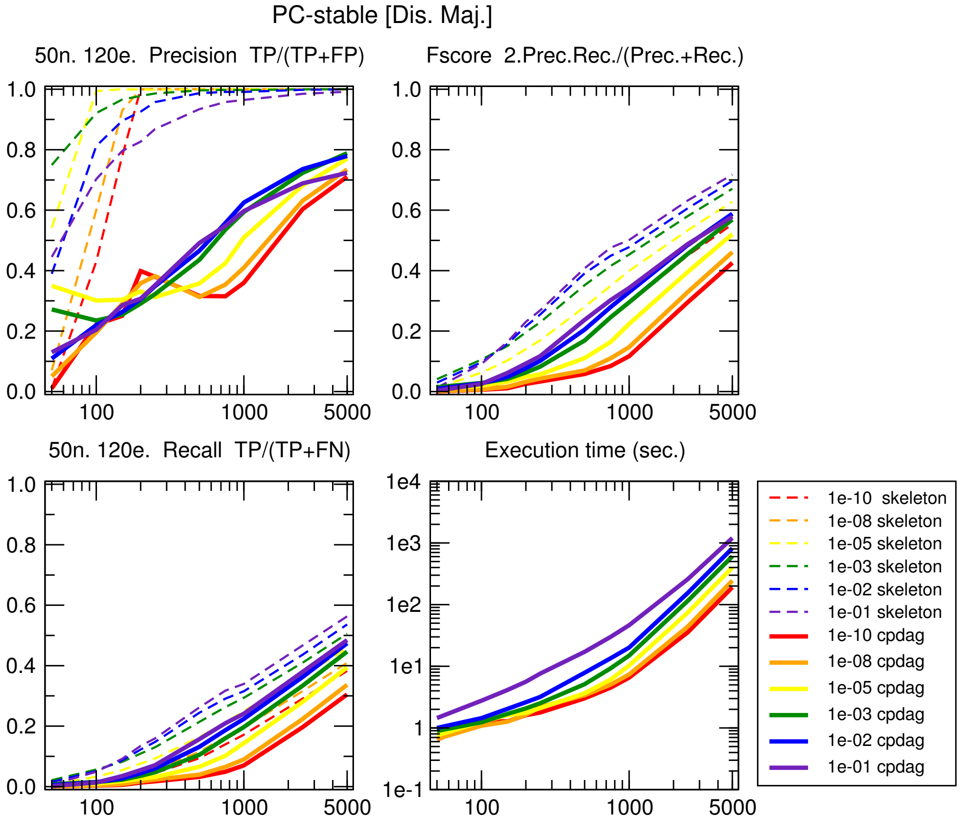

Figure S11: **PC, effect of independence test parameter** α**.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

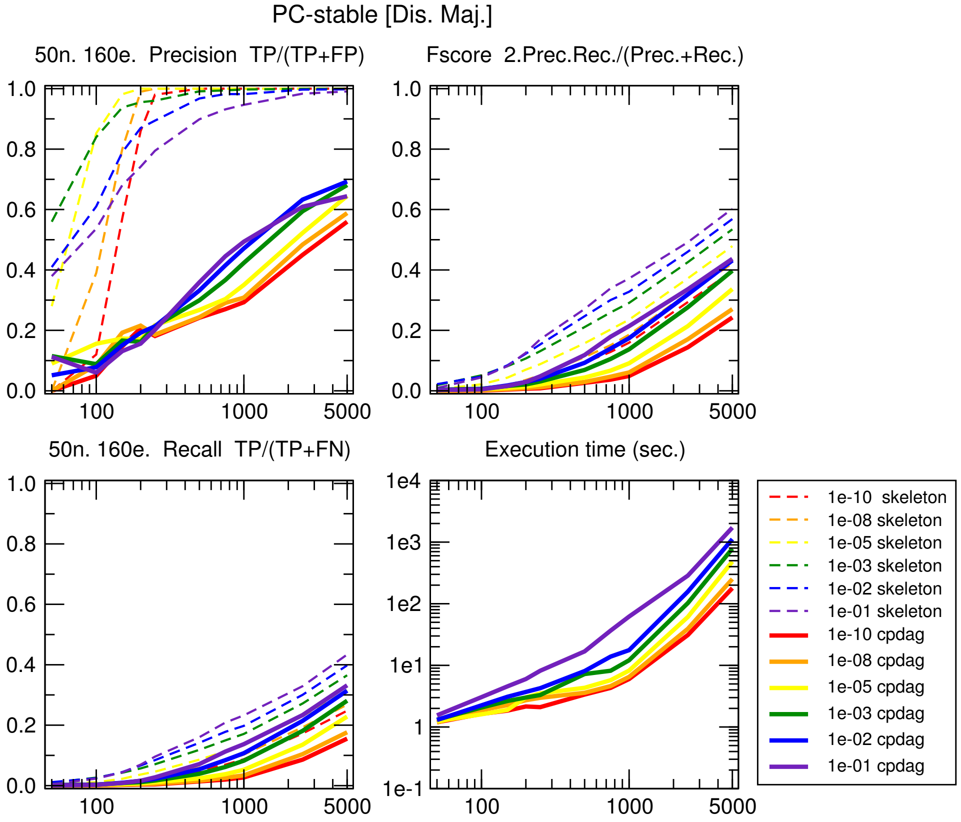

Figure S12: **PC, effect of independence test parameter** α**.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. G^2 independence test; PC-stable, majority rule (Colombo and Maathuis, 2014).

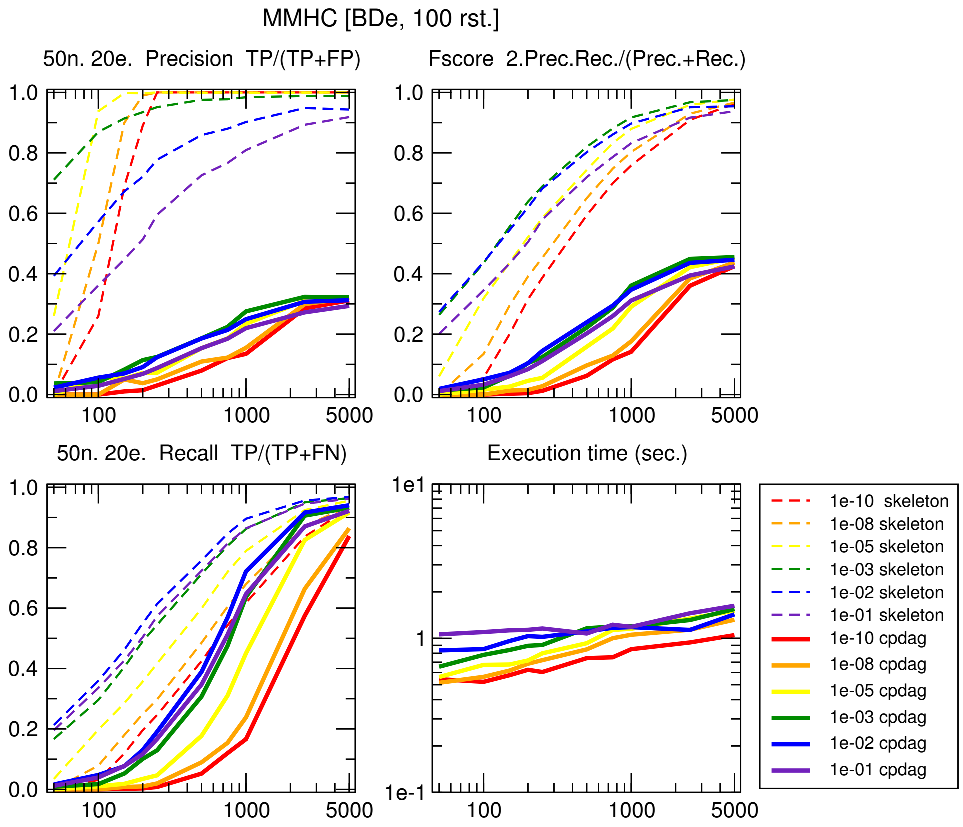

Figure S13: **MMHC, effect of independence test parameter** α. 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k_{\max}^{in} \rangle = 2.4$ and $\langle k_{\max}^{out} \rangle = 2.2$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

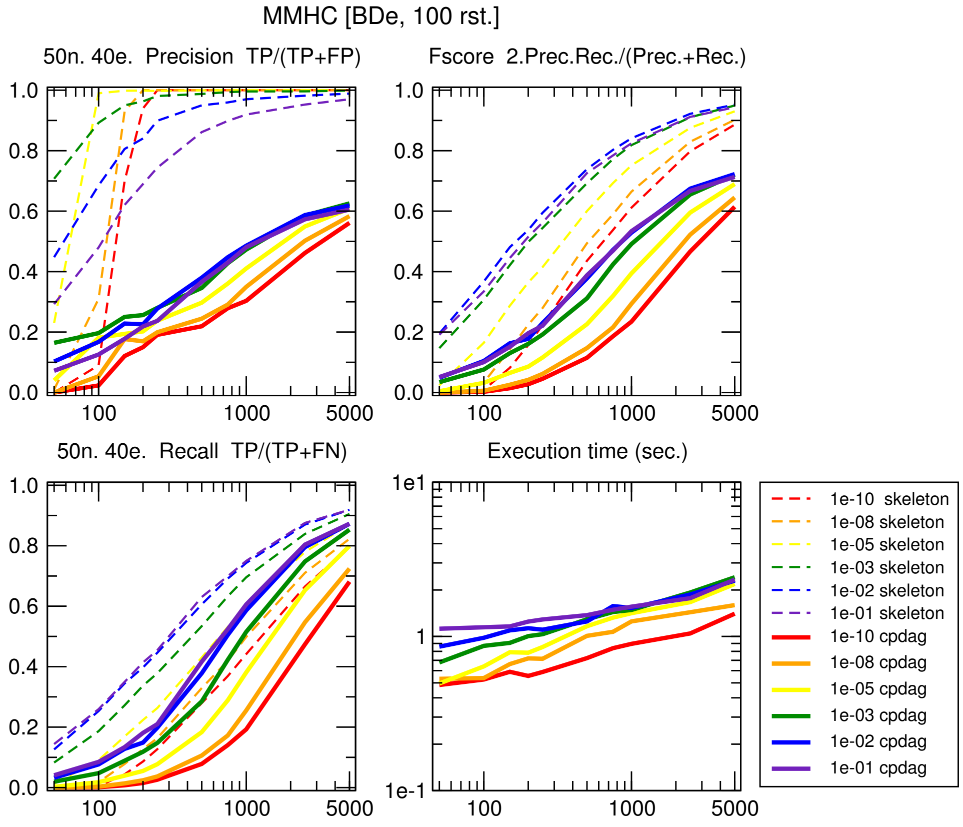

Figure S14: **MMHC, effect of independence test parameter** α. 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k_{\max}^{in} \rangle = 3.2$ and $\langle k_{\max}^{out} \rangle = 3.6$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

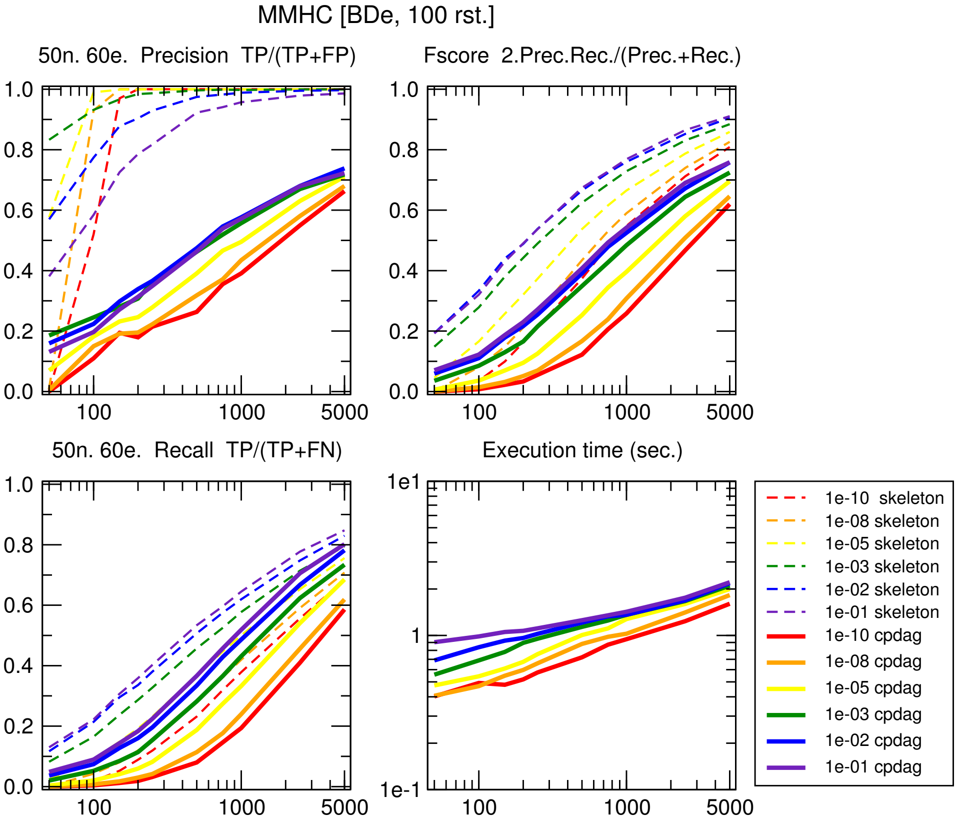

Figure S15: **MMHC, effect of independence test parameter** α**.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k\rangle = 2.4$, $\langle k^{in}_{\max}\rangle = 4.6$ and $\langle k^{out}_{\max}\rangle = 3.6$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

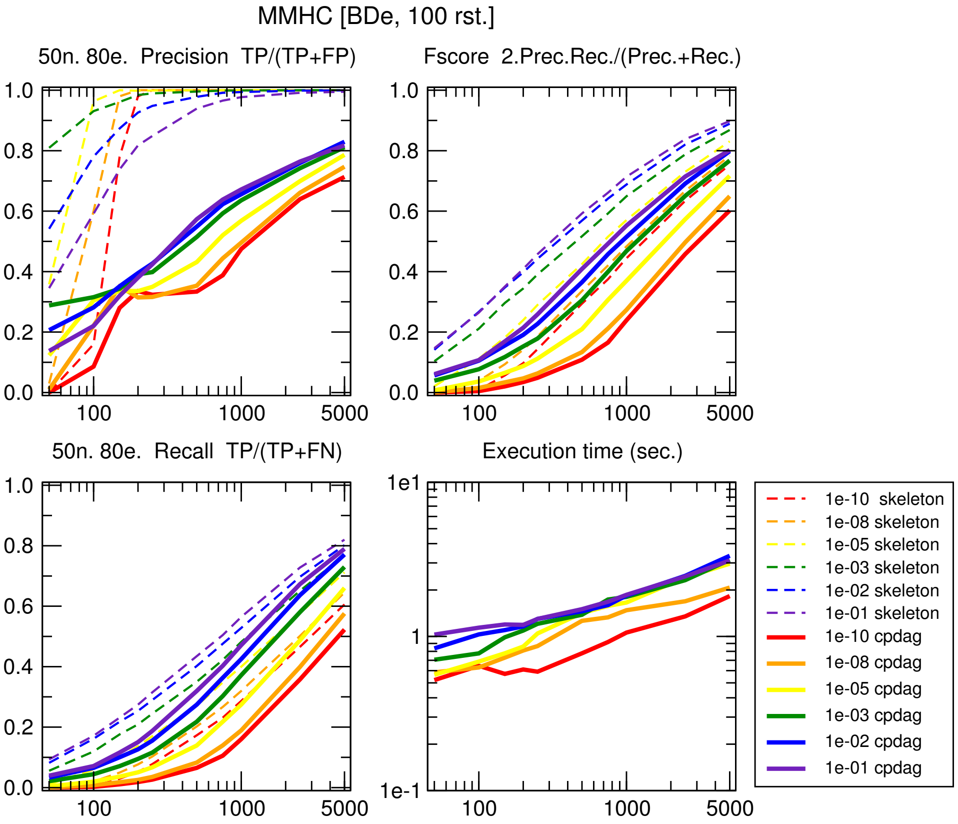

Figure S16: **MMHC, effect of independence test parameter** α**.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k\rangle = 3.2$, $\langle k^{in}_{\max}\rangle = 4.8$ and $\langle k^{out}_{\max}\rangle = 5.6$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

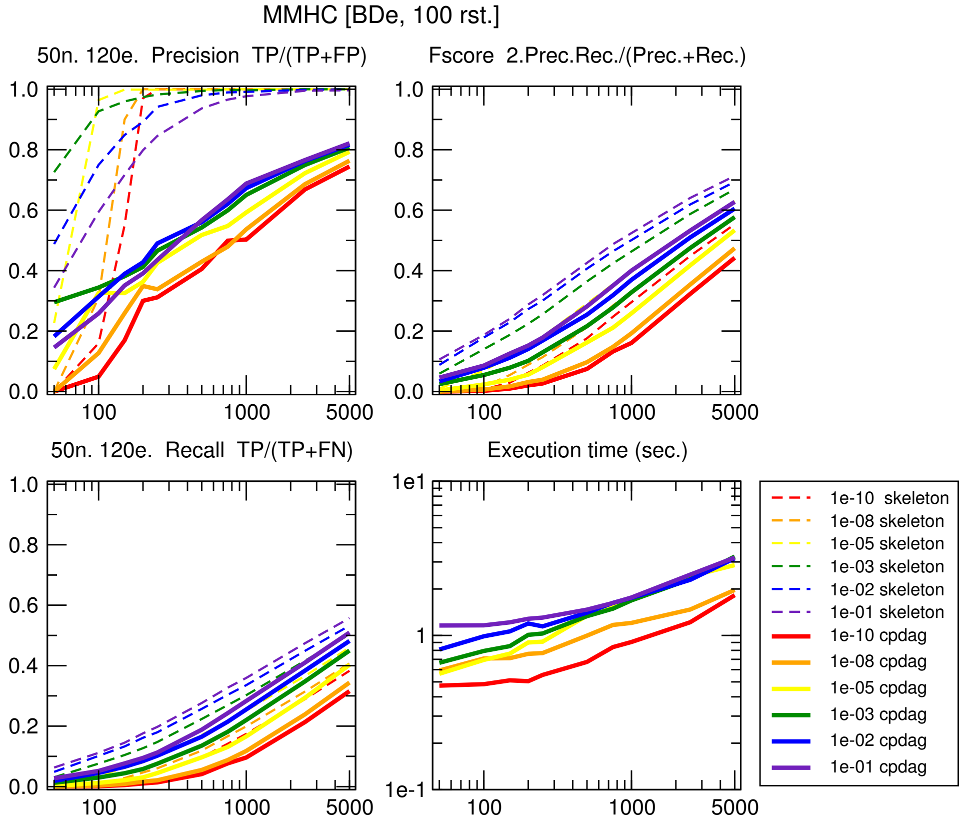

Figure S17: **MMHC, effect of independence test parameter** α**.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k^{in}_{\max} \rangle = 8.8$ and $\langle k^{out}_{\max} \rangle = 7.2$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

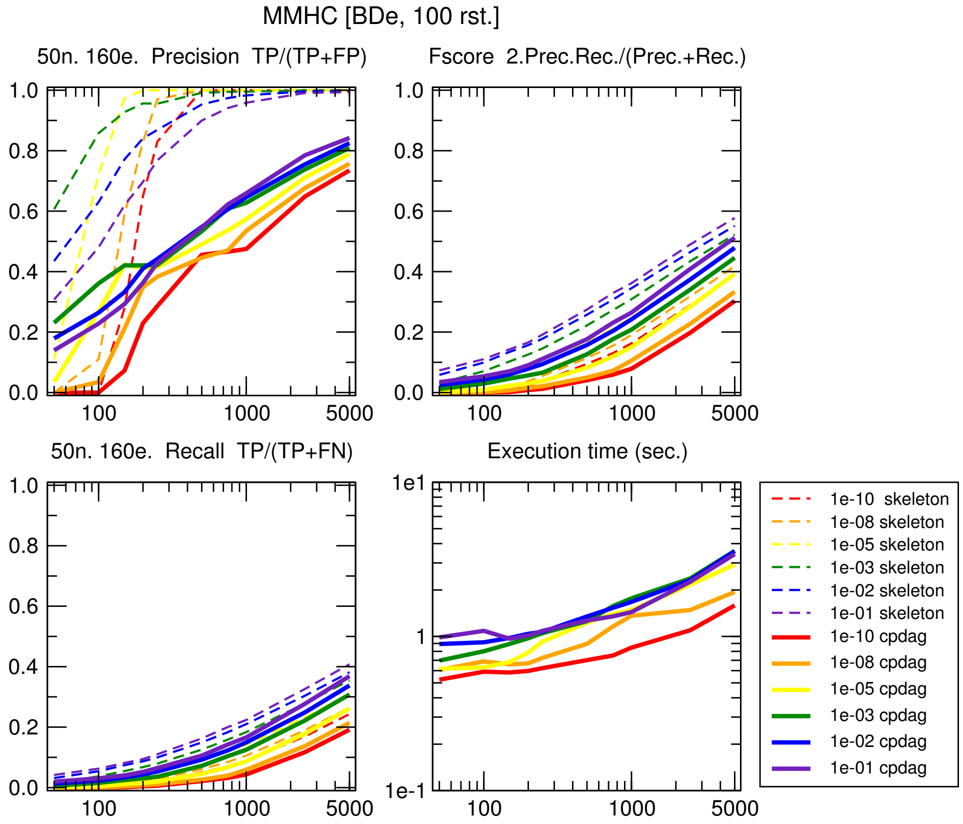

Figure S18: **MMHC, effect of independence test parameter** α**.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k^{in}_{\max} \rangle = 8.6$ and $\langle k^{out}_{\max} \rangle = 8.6$. G^2 independence test; MMHC, BDe score (Tsamardinos, Brown, and Aliferis, 2006).

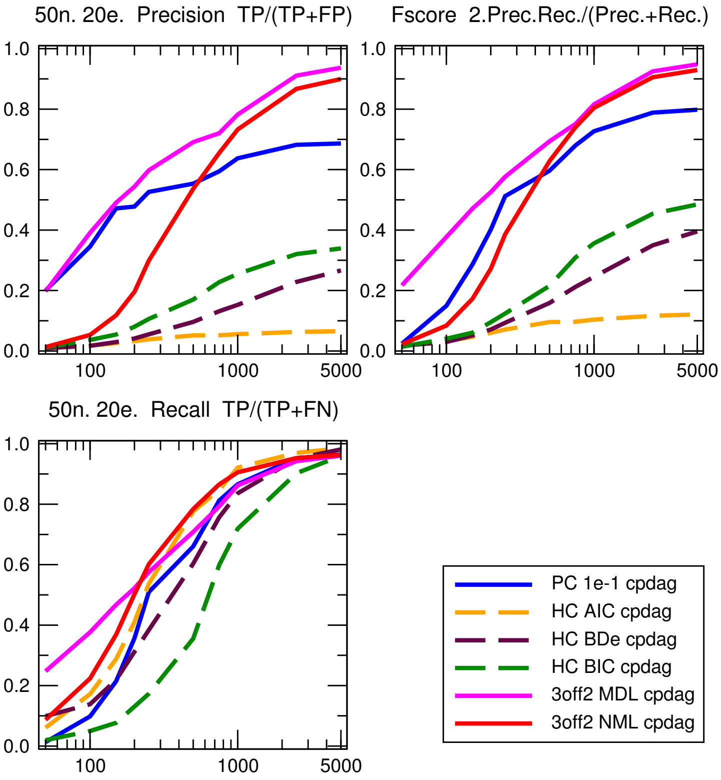

Figure S19: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k\rangle = 0.8$, $\langle k^{in}_{\max}\rangle = 2.4$ and $\langle k^{out}_{\max}\rangle = 2.2$. Bayesian scores: AIC, BDe and BIC.

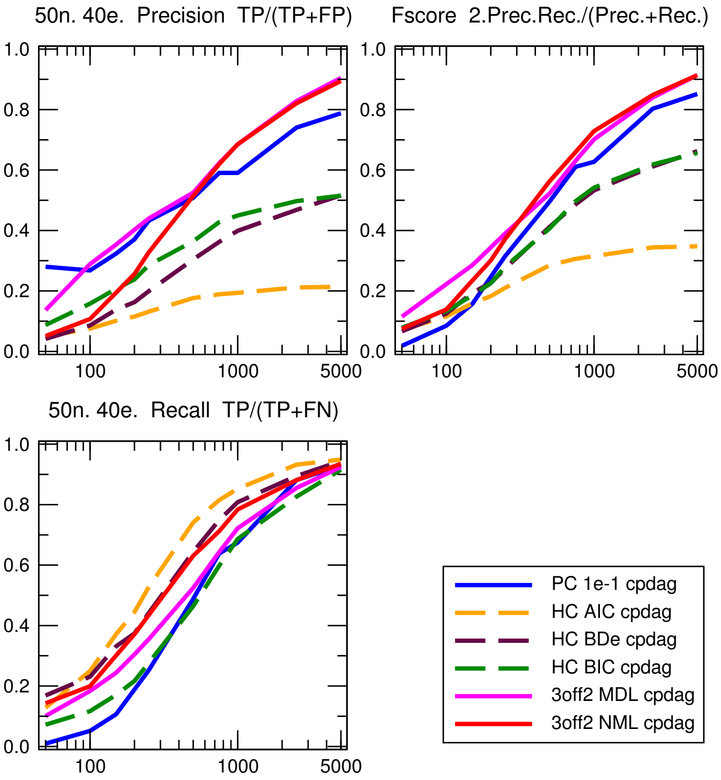

Figure S20: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k\rangle = 1.6$, $\langle k^{in}_{\max}\rangle = 3.2$ and $\langle k^{out}_{\max}\rangle = 3.6$. Bayesian scores: AIC, BDe and BIC.

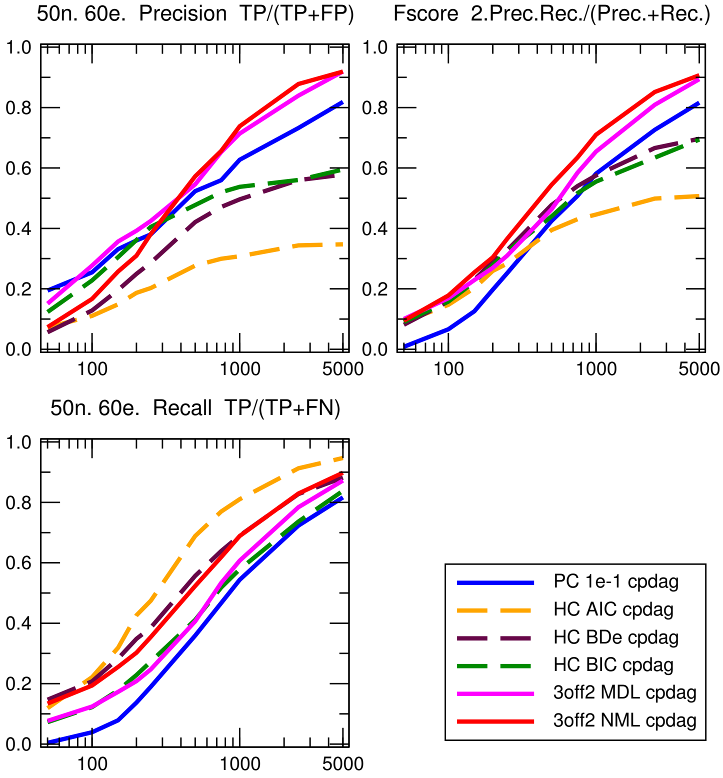

Figure S21: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$. Bayesian scores: AIC, BDe and BIC.

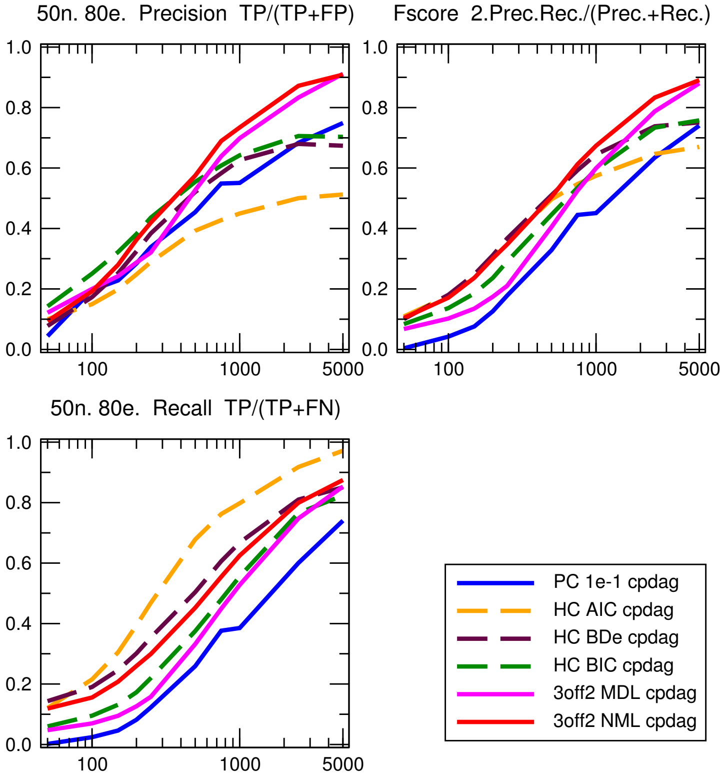

Figure S22: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$. Bayesian scores: AIC, BDe and BIC.

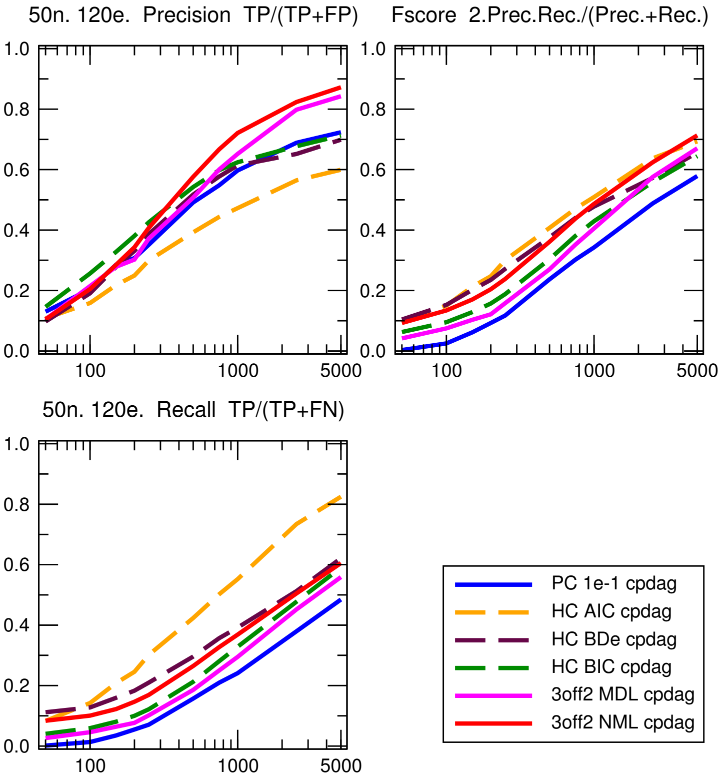

Figure S23: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k\rangle = 4.8$, $\langle k^{in}_{\max}\rangle = 8.8$ and $\langle k^{out}_{\max}\rangle = 7.2$. Bayesian scores: AIC, BDe and BIC.

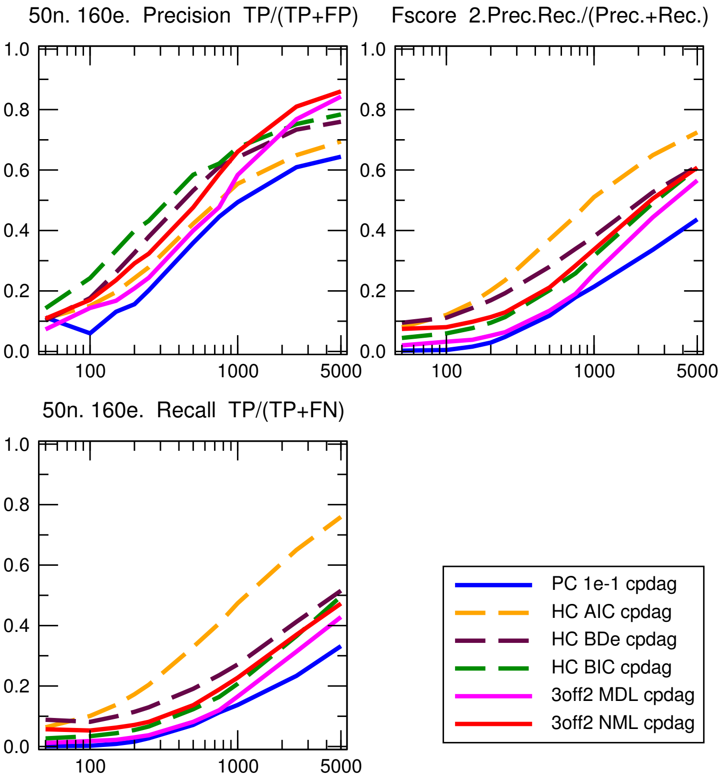

Figure S24: **CPDAG comparison between 3off2, PC and Bayesian hill climbing.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k\rangle = 6.4$, $\langle k^{in}_{\max}\rangle = 8.6$ and $\langle k^{out}_{\max}\rangle = 8.6$. Bayesian scores: AIC, BDe and BIC.

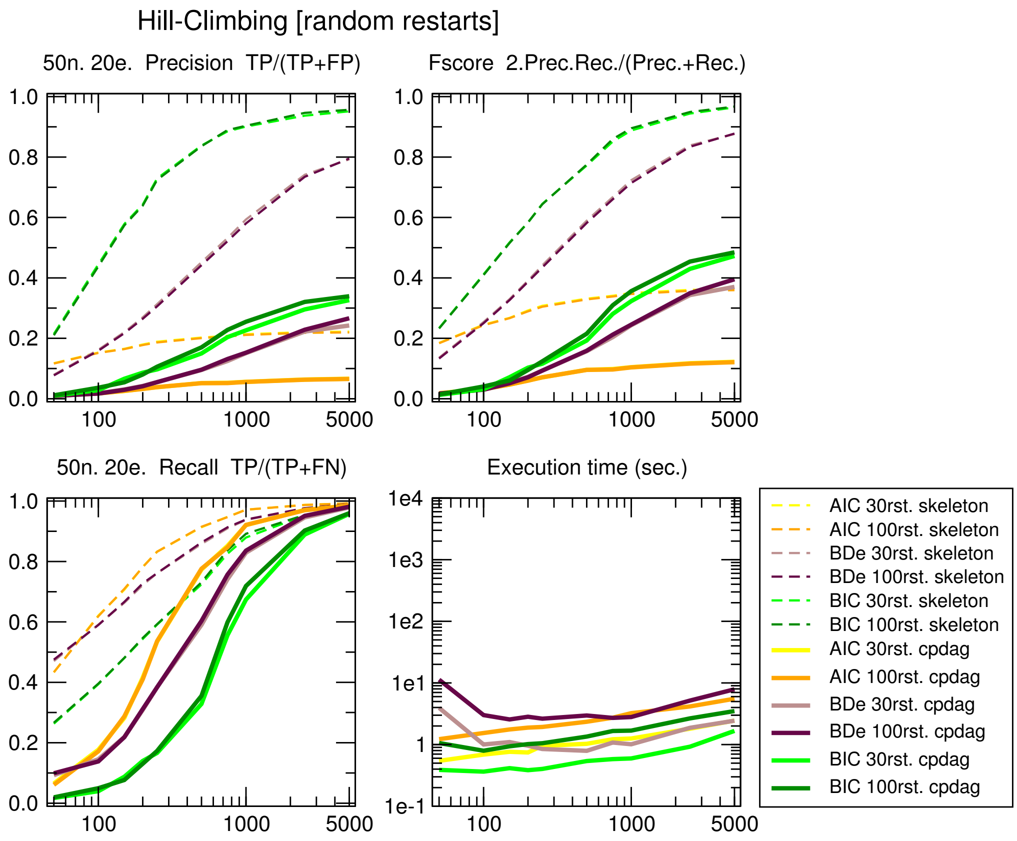

Figure S25: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 20 edge benchmark networks generated using Tetrad. $\langle k \rangle = 0.8$, $\langle k^{in}_{\max} \rangle = 2.4$ and $\langle k^{out}_{\max} \rangle = 2.2$.

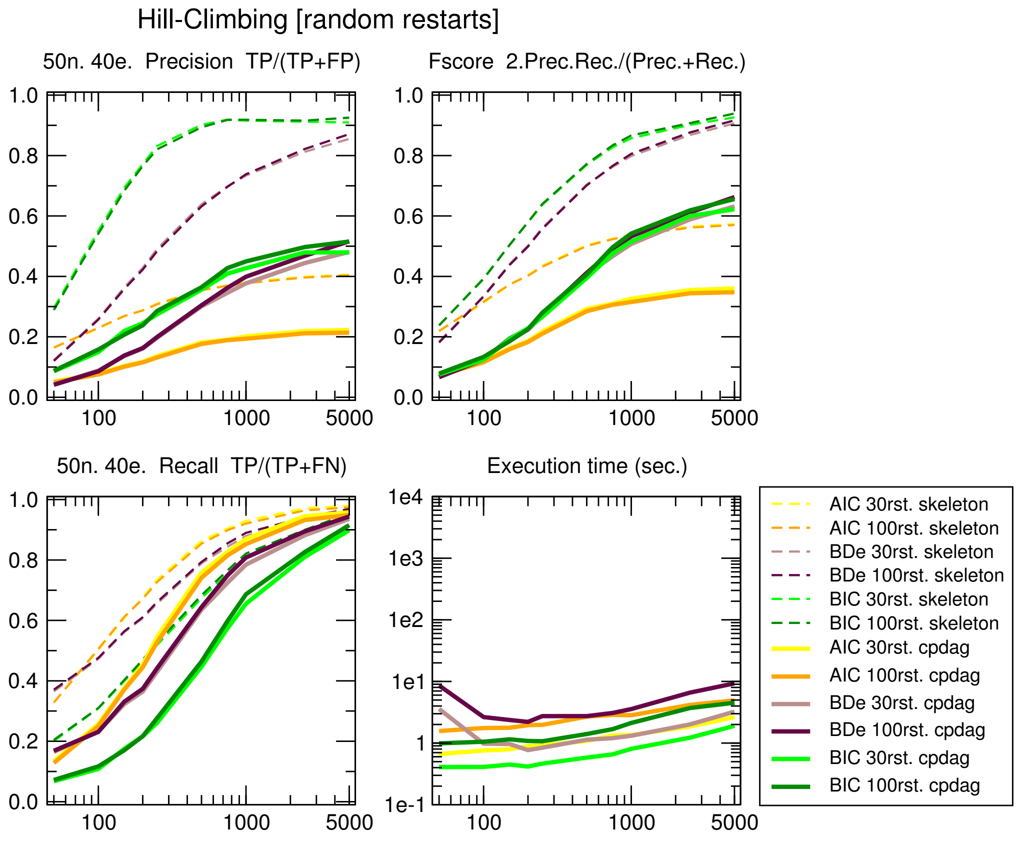

Figure S26: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 40 edge benchmark networks generated using Tetrad. $\langle k \rangle = 1.6$, $\langle k^{in}_{\max} \rangle = 3.2$ and $\langle k^{out}_{\max} \rangle = 3.6$.

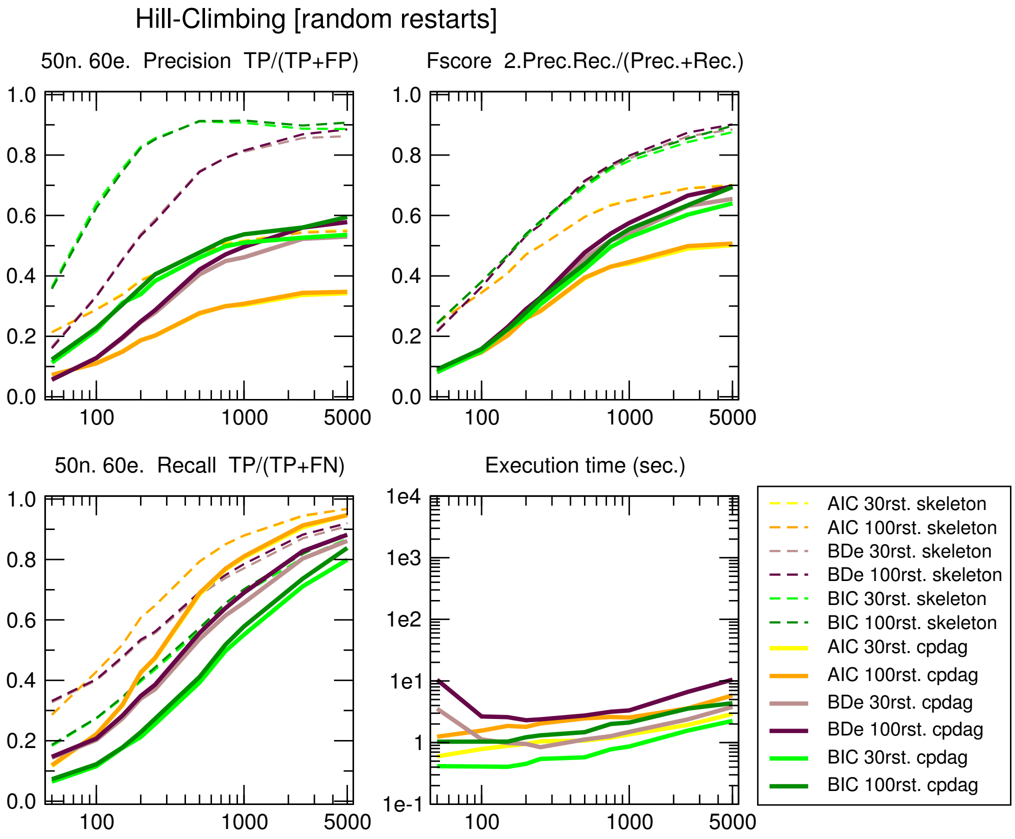

Figure S27: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 60 edge benchmark networks generated using Tetrad. $\langle k \rangle = 2.4$, $\langle k^{in}_{\max} \rangle = 4.6$ and $\langle k^{out}_{\max} \rangle = 3.6$.

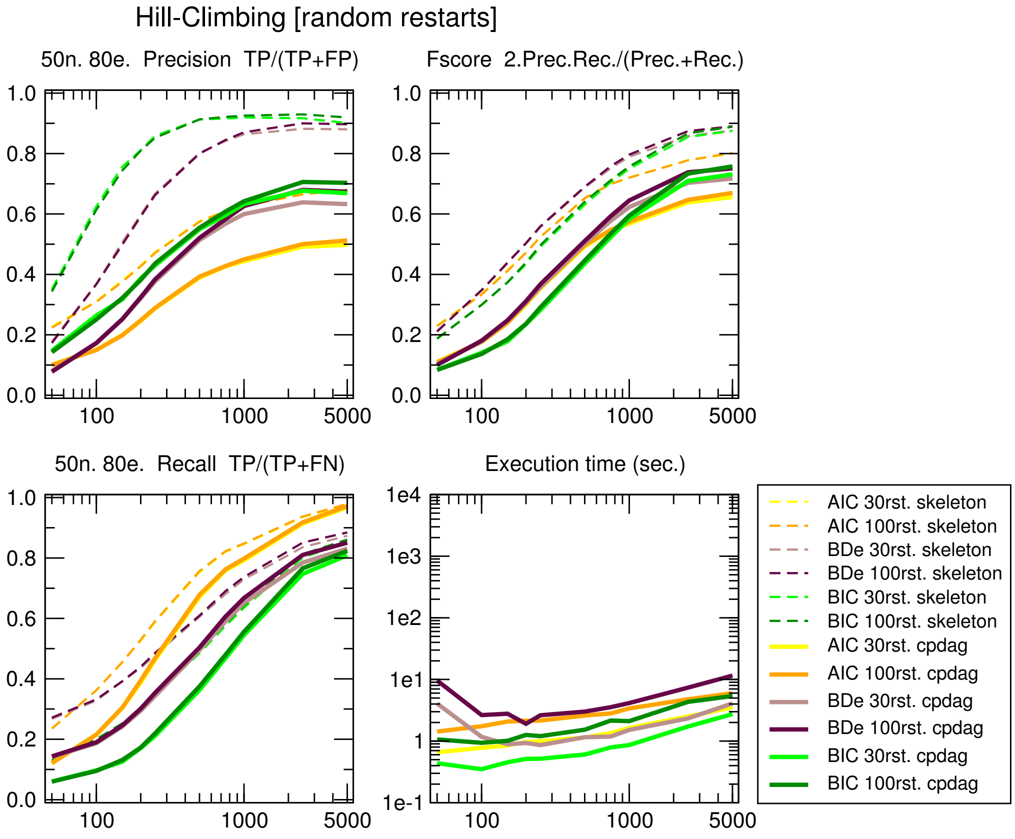

Figure S28: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 80 edge benchmark networks generated using Tetrad. $\langle k \rangle = 3.2$, $\langle k^{in}_{\max} \rangle = 4.8$ and $\langle k^{out}_{\max} \rangle = 5.6$.

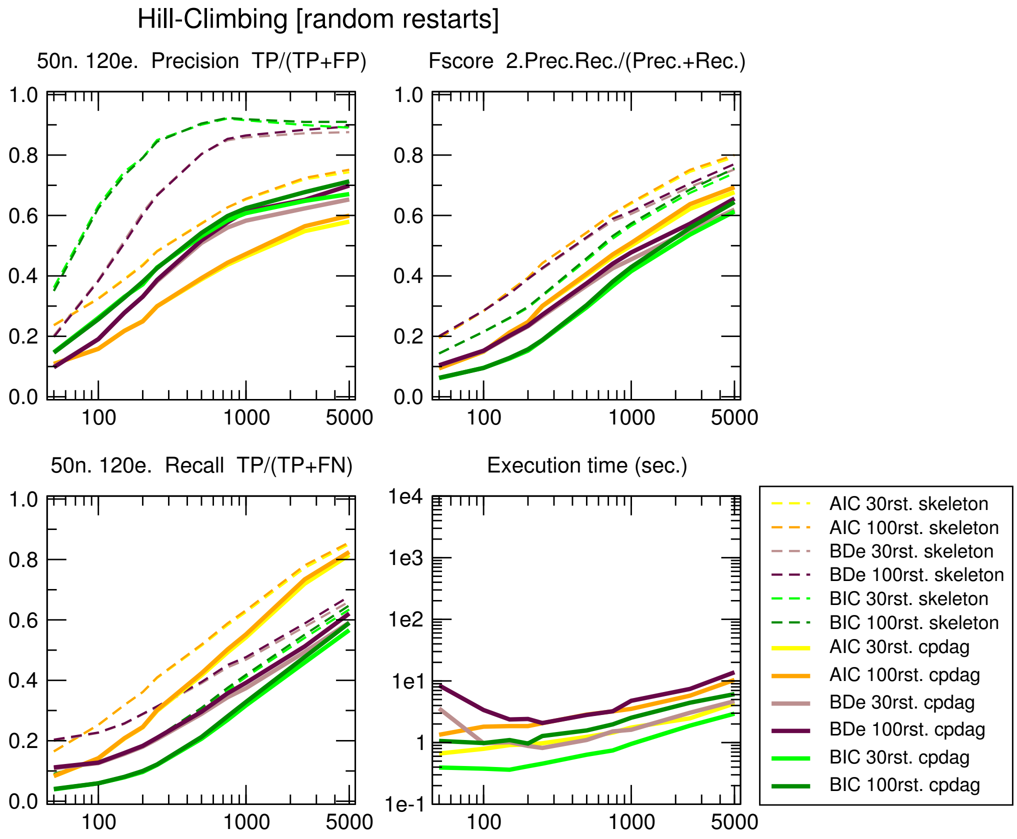

Figure S29: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 120 edge benchmark networks generated using Tetrad. $\langle k \rangle = 4.8$, $\langle k_{\max}^{in} \rangle = 8.8$ and $\langle k_{\max}^{out} \rangle = 7.2$.

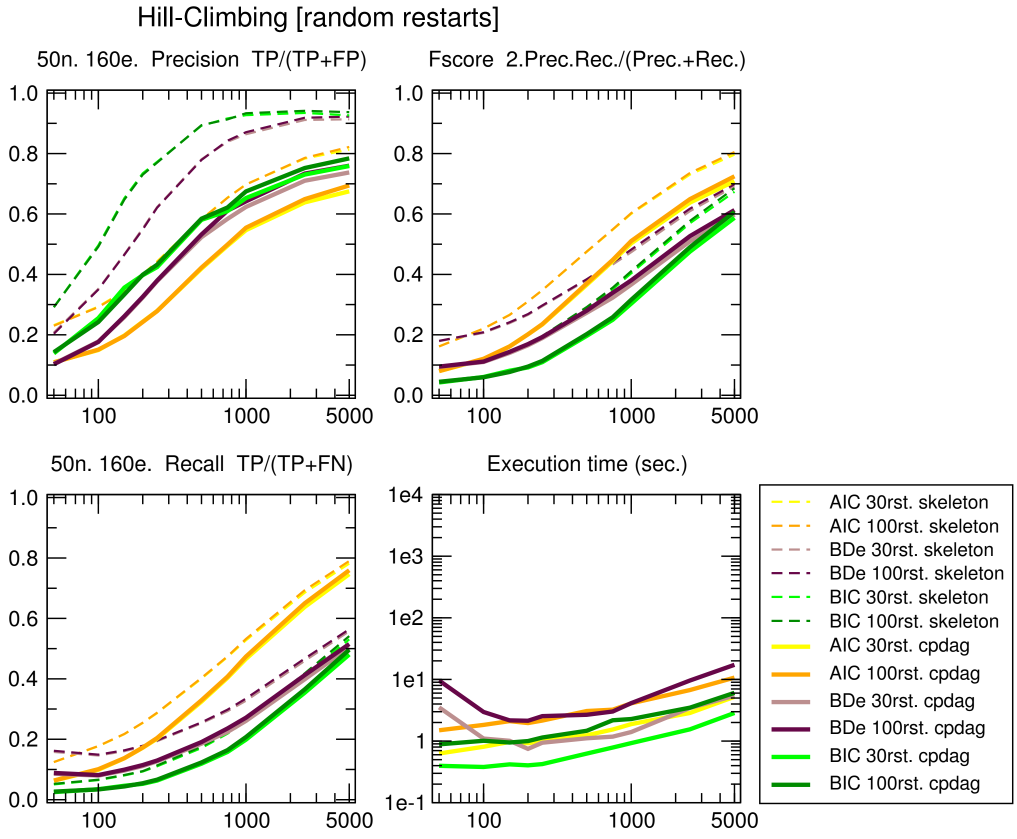

Figure S30: **Bayesian Hill-Climbing, effect of Bayesian score AIC, BDe and BIC.** 50 node, 160 edge benchmark networks generated using Tetrad. $\langle k \rangle = 6.4$, $\langle k_{\max}^{in} \rangle = 8.6$ and $\langle k_{\max}^{out} \rangle = 8.6$.