



Improved baselines for causal structure learning on interventional data

Robin Richter¹ · Shankar Bhamidi² · Sach Mukherjee^{1,3}

Received: 14 June 2022 / Accepted: 15 May 2023
© The Author(s) 2023

Abstract

Causal structure learning (CSL) refers to the estimation of causal graphs from data. Causal versions of tools such as ROC curves play a prominent role in empirical assessment of CSL methods and performance is often compared with “random” baselines (such as the diagonal in an ROC analysis). However, such baselines do not take account of constraints arising from the graph context and hence may represent a “low bar”. In this paper, motivated by examples in systems biology, we focus on assessment of CSL methods for multivariate data where part of the graph structure is known via interventional experiments. For this setting, we put forward a new class of baselines called graph-based predictors (GBPs). In contrast to the “random” baseline, GBPs leverage the known graph structure, exploiting simple graph properties to provide improved baselines against which to compare CSL methods. We discuss GBPs in general and provide a detailed study in the context of transitively closed graphs, introducing two conceptually simple baselines for this setting, the observed in-degree predictor (OIP) and the transitivity assuming predictor (TAP). While the former is straightforward to compute, for the latter we propose several simulation strategies. Moreover, we study and compare the proposed predictors theoretically, including a result showing that the OIP outperforms in expectation the “random” baseline on a subclass of latent network models featuring positive correlation among edge probabilities. Using both simulated and real biological data, we show that the proposed GBPs outperform random baselines in practice, often substantially. Some GBPs even outperform standard CSL methods (whilst being computationally cheap in practice). Our results provide a new way to assess CSL methods for interventional data.

Keywords Causality · Causal structure learning · Interventional data · Transitively closed graphs · Gene regulatory networks · Null models

1 Introduction

Causal structure learning (CSL) refers to the task of estimating a graph encoding causal relationships from data (Pearl

2009; Spirtes 2010). CSL is an important and challenging topic in its own right and has attracted a great deal of recent research attention in a number of fields including statistics, machine learning and philosophy (reviewed in Heinze-Deml et al. 2018). Broadly speaking, given data X (which might be observational and/or interventional), CSL methods provide a graph estimate $\hat{G}(X)$ (or probabilistic analogue) with edges intended to encode causal relationships. The semantics of such graphs can be complex and depend on the precise model and application domain but for the present it is important only to emphasize that such estimators use data X to infer relationships between entities and can be viewed as encoding such information as a directed graph \hat{G} .

CSL methods necessarily require assumptions on the underlying causal system that may or may not hold in real applications and whose validity may be difficult to check in practice. As a result the behaviour of CSL methods under realistic conditions (noise levels, limited sample sizes etc.)

✉ Robin Richter
robin.richter@dzne.de
Shankar Bhamidi
bhamidi@email.unc.edu
Sach Mukherjee
sach.mukherjee@dzne.de

¹ Statistics and Machine Learning, Deutsches Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE), Venusberg-Campus 1/99, 53111 Bonn, Germany

² Department of Statistics and Operations Research, University of North Carolina, 304 Hanes Hall, Chapel Hill 27599-3260, NC, USA

³ MRC Biostatistics Unit, University of Cambridge, Robinson Way, Cambridge CB2 0SR, UK

may not be clear in advance. As such, in practical settings it is important to empirically assess the efficacy of CSL methods. To this end a number of studies have focused on such assessment (including, among others, Hill et al. 2016; Heinze-Deml et al. 2018; Eigenmann et al. 2020). In the empirical assessment of CSL methods, a common strategy is to compare the estimated graph \hat{G} with a “ground truth” graph G^* (depending on context either the true data-generating graph in a simulation, or an scientifically/experimentally-defined gold standard). Such quantitative comparisons are usually made alongside baselines, which provide a way to contextualize the performance of the estimator \hat{G} on the specific problem. Random baselines, such as the diagonal in an ROC analysis, are widely used, motivated by the idea that large deviations from the random case are an indicator that the estimator is successfully identifying causal structure.

In this paper we put forward a new class of baselines for the assessment of CSL methods in the setting that (some) interventional data is available. While random baselines are a good and useful tool, they ignore structure that might be inherent in the problem, in the sense of regularities in the ground truth graph G^* . In the interventional data setting, some information on G^* is available at the outset. We argue that such information can constrain possible solutions such that random baselines are in a way too general for this setting and provide only a “low bar” against which to assess CSL methods. Instead, we propose to exploit the knowledge of part of the ground truth graph in combination with straightforward graph properties, to define new baselines called *graph-based predictors* (GBPs), that share conceptual simplicity with classical baselines but that constitute a demonstrably stronger test.

A related line of work, developing and utilizing null models for networks seeks to contextualise interesting network features with reference to default, background models, see e.g. the surveys Fosdick et al. (2018) and Gauvin et al. (2018) as well as Chapter 11 in Fornito et al. (2016) and references therein. The key idea in these approaches is to understand whether a seemingly salient feature of a network (e.g. high levels of connectivity within specific subsets of the graph leading to the thriving area of community detection, see e.g. Newman and Girvan (2004) and the survey Fortunato (2010) and the references therein) is really unusual or noteworthy. In a similar fashion, we seek to contextualise the performance of CSL methods, using certain graph properties to define suitable baselines. However, a key difference is that in the null models literature the network itself is assumed known; in contrast, in our paper and CSL in general, the network itself is (partially) inferred.

Our work is motivated by, and illustrated in the context of, interventional experiments that have become feasible in recent years in molecular biology (see, among others, Sachs et al. 2005; Kemmeren et al. 2014; Shalem et al. 2015; Dixit

et al. 2016; Ursu et al. 2022). Such experiments are crucial for the inference of molecular networks, encoding causal relationships between entities such as genes or proteins, which in turn play a central role in disease and systems biology (see e.g. Phillips 2008; Parikhshak et al. 2015). The inference of molecular networks from data is a long-standing problem at the intersection of statistics, machine learning and systems biology (for introductions see e.g. Ideker et al. 2001; Babu et al. 2004; Sanguinetti and Huynh-Thu 2019; Nogueira et al. 2022).

In practice the interventional experiments in biology involve perturbation of molecular nodes (for example genes) and subsequent measurement of a high-dimensional readout (such as gene expression), specific examples of these include gene knock-out /-down, /-up, /-in experiments. Such data are relevant for causal learning because the measurement of a gene expression level for a gene B after perturbation of a gene A gives information on the (total) causal effect of A on B . Hence, if available, incorporating interventional data alongside observational data in CSL methods is desirable, and this has been studied from a number of perspectives (relevant literature includes Hauser and Bühlmann 2012; Rau et al. 2013; Spencer et al. 2015; Peters et al. 2016; Magliacane et al. 2016a,b; Meinshausen et al. 2016; Magliacane and van Ommen 2017; Wang et al. 2017; Hill et al. 2019; Rothenhäusler et al. 2019; Brouillard et al. 2020).

At the same time, interventional data are widely used to obtain gold standards to assess CSL methods (see e.g. Colombo and Maathuis 2014; Meinshausen et al. 2016; Wang et al. 2017). Notably, in practice, it is usually not feasible to perform all possible perturbation experiments due to time- and cost-constraints, rather only a subset can be performed. As we discuss in detail in Sect. 2, this can be viewed as providing information on a *partial observation* of the ground truth graph G^* and this practical scenario is the one we focus on.

A particularly interesting and relevant special case concerns transitively closed graphs. As noted above, in real-world gene perturbation experiments, one observes the total causal effect of perturbing one gene (the target A) on another gene B (usually many such genes are measured in contemporary “omics” designs, we refer to such data in the following as omics readouts or simply as omics data). An effect of A on B may be mediated by other genes intermediate in the underlying causal path. For this reason, such effects are transitive in the sense that if A has a causal edge to B (in the underlying causal graph) and B to C , then an intervention on A may change C (this corresponds to the total causal effect of A on C), resulting in an edge from A to C in a graph constructed directly from the perturbation experiments. The assumption of observing *transitively closed* or *ancestral* causal graphs has also been made in Magliacane et al. (2016a) who consider estimating transitively closed graphs and in Heinze-Deml

et al. (2018) and Eigenmann et al. (2020) where CSL methods were evaluated with respect to ancestral relations of this kind.

The contributions of this paper are as follows:

- *New class of baselines.* We propose a new class of baselines for CSL that take account of graph properties in the case that interventional data is available. The proposed baselines leverage structural properties rooted in the underlying causal graph.
- *Methods for transitively closed graphs.* Motivated by the nature of real-world gene perturbation experiments, we focus particular attention on transitivity and related properties and put forward specific baselines that exploit constraints derived from these properties.
- *Theoretical results of superiority and delineation.* We show for a particular baseline a superiority-statement in the context of latent-network models. Moreover, we delineate the proposed baselines from each other theoretically.
- *Empirical results using real gene/protein perturbation data.* Using real data from large-scale gene and protein perturbation experiments we study the behaviour of the proposed methods to understand whether they can actually provide improved baselines in practice.

Taken together, our results provide a framework for constructing improved baselines for CSL and thereby to more thoroughly assess the capabilities of CSL methods, with a focus on the use of interventional data, an area of key relevance for ongoing efforts at the interface between systems biology and large-scale perturbation designs.

The remainder of the paper is organized as follows. We begin in Sect. 2 with notation and background, defining the precise set-up for which the proposed baselines are intended. In Sects. 3.1 and 3.2 we introduce two general ways to construct graph-based predictors, based respectively on in-degree information and constraints rooted in transitivity. These two classes are illustrated with specific implementations – the *observed indegree predictor* (OIP) and several *transitivity assuming predictors* (TAPs) respectively – which are specifically derived for their use as baselines in system biology experiments. For the OIP a theoretical result of superiority over random baselines is given. Moreover, in Sect. 3.2 we propose simulation strategies for the TAPs as their direct computation is infeasible. In Sect. 3.3 combinations of the OIP and the TAPs are discussed. We detail in Sect. 3.4 the theoretical differences of all introduced candidate baselines and outline potential similarities. Section 4.1 provides detailed analysis of a simulation study of the proposed GBPs. In Sect. 4.2 we then study the behaviour of the proposed GBPs using real transcriptomics and proteomics data including observational and interventional experiments,

alongside application of standard CSL methods from the literature to the same data sets. We conclude with a brief discussion on open questions and possible future work in Sect. 5.

2 Notation and background

In this section we give some background on CSL and introduce notation and the general set-up. In particular, we detail the structure of the data X and its underlying causal graph G in the context of CSL on interventional data.

2.1 Contextualization within CSL

We focus on the setting in which *interventional* and *observational data* are included in X . For example in the case of omics data X includes rows of readouts after targeted gene perturbations (interventional) and after control experiments (observational). In practice a gold standard ground truth graph G^* might be obtained by comparing interventional and observational data, either in the current set of experiments or using previous experimental data. Given measurement of a variable B after perturbation of variable A , the causal relationship (A, B) (“from” A “to” B) is inferred by comparing the empirical distribution of B under the control experiments with the corresponding distribution under intervention on A . Since omics designs usually involve measuring many variables in parallel we consider here the common case that given a perturbation is performed on A we measure all other genes, i.e. each intervention experiment corresponds to a whole row of readouts in X . We consider only single interventions (i.e. only one node A is intervened upon in a given experiment). It is important in the below detailed set-up that we have access to interventional data in which some (but not all) genes are intervened upon, which is the common case in practice.

Some clarifications regarding our set-up are as follows: (1) We do not *a priori* rule out cycles in directed graphs. This is because in practice an intervention on a variable A may change B and vice versa (see also below). (2) For ease of discussion we assume that the type of intervention is fixed and that causal claims relate to the specific type of intervention. This is motivated by the fact that in practice, the precise nature of an intervention is defined by the experimental protocol, hence claims and predictions are limited to changes under the specific protocol. As a concrete example, if a knock-out of a gene A changes gene B , this does not imply that a knock-down of A would change B (since the latter experiment might induce a sub-threshold change to A) and so on. (3) For ease of computation we consider self-edges to be present at every node (compare $M[k, k] = 1$ for all k in (2.1) further below).

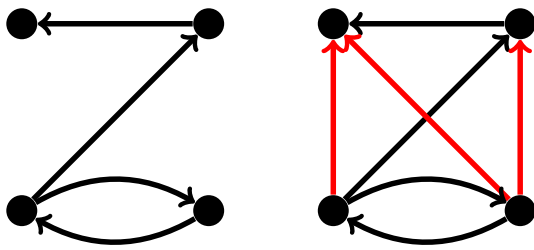


Fig. 1 A causal graph $G = (V, E)$ (left) and its induced ancestral causal graph $G^+ = (V, E^+)$ (right, “new” edges are depicted in red, i.e. $E^+[k, \ell] = 1 \neq 0 = E[k, \ell]$)

Point (1) stands in contrast to some of the classical CSL literature, in particular to methods based on directed acyclic graphs (DAGs), where the assumption of acyclicity plays a crucial role (Spirtes et al. 2000; Maathuis et al. 2009; Colombo and Maathuis 2014). Cyclic models have been discussed in the literature (see e.g. Richardson 1996; Hyttinen et al. 2014; Hill et al. 2019). In the applied context of perturbation omics experiments, cyclic models are natural, because an intervention on one gene A may lead to a change in another gene B , but an intervention on B may vice versa lead to a change in A . This is essentially due to the fact that real omics data are measurements at a given time in a dynamic system (with the causal effects always forward-in-time in the underlying system).

2.2 Notation and basic definitions

Denote a directed, unweighted graph by $G = (V, E)$ with vertex set $V = \{v_1, v_2, \dots, v_p\}$ and edge matrix $E \in \mathcal{E}$ with,

$$\mathcal{E} := \{M \in \{0, 1\}^{p \times p} : M[k, k] = 1, \text{ for all } k\}. \quad (2.1)$$

As the graphs of interest encode causal relationships between entities in V where useful we refer to them as *causal graphs*.

Definition 2.1 Let $G = (V, E)$ be a causal graph.

- (1) We say there exists a *causal path* from v_k to v_ℓ in G with $v_k \neq v_\ell \in V$, if, for some $T \in \mathbb{N}_0$ there exist vertices $v_k = w_0, w_1, \dots, w_T, w_{T+1} = v_\ell \in V$ such that

$$E[w_t, w_{t+1}] = 1 \text{ for all } 0 \leq t \leq T.$$

- (2) Call $G^+ = (V, E^+)$ the *ancestral causal graph* (or the *causal transitive closure*) of G if

$$(\exists \text{ causal path from } v_k \text{ to } v_\ell) \Leftrightarrow E^+[k, \ell] = 1$$

holds. Moreover, call G an *underlying* causal graph of G^+ . For an example see Fig. 1.

- (3) Call G *ancestral* or *transitively closed* if $G^+ = G$ holds.
- (4) For a node $v_k \in V$ define the *indegree* of v_k by

$$\text{deg}^-(v_k) := |\{v_\ell \in V \setminus \{v_k\} : E[\ell, k] = 1\}|$$

and the *outdegree* of v_k by

$$\text{deg}^+(v_k) := |\{v_\ell \in V \setminus \{v_k\} : E[k, \ell] = 1\}|.$$

We note that ancestral causality has been studied in the literature using a variety of models (see e.g. Zhang 2008; Magliacane et al. 2016b; Malinsky and Spirtes 2016; Mooij and Claassen 2020) and is a complex topic in its own right. The purpose of the above definition is simply to introduce the notion of a transitive closure and make the connection to indirect causation to facilitate introduction of specific, transitivity assuming baselines below.

We will use directed graphs that are random in an edge-wise Erdős-Rényi sense as defined next (such graphs are studied in Karp 1990).

Definition 2.2 Define a *random directed graph* (RDG) of size p , with edge probability q and denoted by $\text{RDG}_q(p) = (V, E)$, as a directed graph with $|V| = p$ nodes, where all off-diagonal entries of E are *iid* draws from a Bernoulli distribution with success probability q . Moreover, given a graph $\tilde{G} = (\tilde{V}, \tilde{E})$ and a subset of edges $K \subset \{[k, \ell] : 1 \leq k \neq \ell \leq p\}$ we construct as $\text{RDG}_{q,K}(\tilde{G}) = (V, E)$ the *partially* random directed graph with underlying \tilde{G} and edge probability q by drawing

$$E[k, \ell] \sim \begin{cases} \delta(\tilde{E}[k, \ell]) & \text{if } [k, \ell] \in K, \\ B(1, q) & \text{else.} \end{cases},$$

with δ denoting the Dirac delta distribution and with *iid* draws from the Bernoulli distribution $B(1, q)$.

Assumption 2.3 below specifies the set-up of the CSL problem on interventional data.

Assumption 2.3 Let $G = (V, E)$ be a causal graph with $|V| = p$. Given available interventional data $X_1 \in \mathbb{R}^{n_1 \times p}$ and observational data $X_2 \in \mathbb{R}^{m_1 \times p}$ as well as latent, unavailable interventional data $Y_1 \in \mathbb{R}^{n_2 \times p}$ and latent, unavailable observational data $Y_2 \in \mathbb{R}^{m_2 \times p}$, on the nodes V with $n_1, n_2, m_1, m_2 \in \mathbb{N}_{>0}$. We assume there exists a set of indices/vertices $\mathcal{I} \subset \{1, 2, \dots, p\}$, called the set of *available interventions*, such that all interventional measurements in X_1 correspond to an intervention on a node v_k with $k \in \mathcal{I}$ and all interventional measurements in Y_1 correspond to an intervention on a node v_ℓ with $\ell \notin \mathcal{I}$. Moreover, we assume

the existence of two *ground truth* functions

$$g_1 : (X_1, X_2) \mapsto (E[k, \ell])_{[k, \ell] \in S_1},$$

where $S_1 := ([k, \ell])_{k \in \mathcal{I}, \ell \in \{1, \dots, p\}, k \neq \ell}$,

$$g_2 : (Y_1, Y_2) \mapsto (E[k, \ell])_{[k, \ell] \in S_2},$$

where $S_2 := ([k, \ell])_{k \notin \mathcal{I}, \ell \in \{1, \dots, p\}, k \neq \ell}$.

Define by

$$X := \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \text{ and, } Y := \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

the *available data* and the *latent data*, respectively. We denote by

$$E_X := (E[k, \ell])_{[k, \ell] \in S_1} = g_1(X)$$

the *partial observation of G w.r.t. I*. Define analogously

$$E_Y := (E[k, \ell])_{[k, \ell] \in S_2} = g_2(Y)$$

the *unobserved causal relationships of G*. Note that we have after possibly reordering of the rows of E the relationship

$$E = \begin{pmatrix} E_X \\ E_Y \end{pmatrix},$$

by slight abuse of notation as we consider only off-diagonal entries.

Let a partial observation E_X of a causal graph G based on available observational and interventional data X be given. We call a predictor

$$\Theta : \{0, 1\}^{|S_1|} \rightarrow [0, 1]^{|S_2|},$$

$$E_X \mapsto \Theta(E_X) \in [0, 1]^{|S_2|}, \tag{2.2}$$

assigning to each unobserved causal relationship a *probability of its existence*, based solely on the partial observation E_X a *graph-based predictor* (GBP). Meanwhile, a predictor

$$\Phi : \mathbb{R}^{(n_1+m_1) \times p} \rightarrow [0, 1]^{|S_2|},$$

$$X \mapsto \Phi(X) \in [0, 1]^{|S_2|}, \tag{2.3}$$

assigning to each unobserved causal relationship a *probability of its existence*, based on the available data matrix X will be called a *data-based predictor* (DBP).

The foregoing assumptions essentially ensure that the graph estimand is operationally well-defined as it is assumed that there exists some oracle procedure by which the edge structure could be determined from idealized data. In the terms above, CSL methods would usually be classified as

DBPs, since they use empirical data to obtain a graph estimand.

For the sake of completeness, we introduce here notation and nomenclature for the ROC curve and the AUC in terms of our set-up, as it is a widely used performance measure for predictors such as Θ and Ψ given in (2.2) and (2.3), respectively. The ROC curve has to be defined with respect to a gold standard; accordingly for Definition 2.4 we assume that the entire graph is known for the purpose of computing the ROC curve and related quantities (of course only part of the graph is available to any estimator/CSL method; specifically, E_Y is unavailable).

Definition 2.4 Let E_X be a partial observation of a non-trivial causal graph $G = (V, E)$ and S_2 be the indices of the unobserved causal relationships. Let $R \in [0, 1]^{|S_2|}$ be the output of a predictor of E_Y . Let $1 = c_0 \geq c_1 \geq \dots \geq c_N \geq c_{N+1} = 0$ be the ordered, unique values of $\{R[k, \ell]\}_{(k, \ell) \in S_2} \cup \{0, 1\}$, with $N \leq |S_2|$. The *receiver operator characteristic (ROC) curve* $ROC(R)$ is given as the linear interpolation of the points,

$$\{(FPR_R(c_t), TPR_R(c_t))\}_{t=0}^{N+1},$$

where

$$FPR_R(c_t) = \frac{|\{[k, \ell] \in S_2 : R[k, \ell] > c_t \text{ and } E[k, \ell] = 0\}|}{|\{[k, \ell] \in S_2 : E[k, \ell] = 0\}|},$$

$$TPR_R(c_t) = \frac{|\{[k, \ell] \in S_2 : R[k, \ell] > c_t \text{ and } E[k, \ell] = 1\}|}{|\{[k, \ell] \in S_2 : E[k, \ell] = 1\}|},$$

for $c_t \neq 0$ and $FPR_R(0) = 1 = TPR_R(0)$, note that both denominators are not 0 by non-triviality of G . We define the *area under curve* (AUC) of the ROC curve as the finite area enclosed in $ROC(R)$, the x-axis and the line $\{x = 1\}$. Note, that by definition $(FPR_R(c_0), TPR_R(c_0)) = (0, 0)$, $(FPR_R(c_{N+1}), TPR_R(c_{N+1})) = (1, 1)$ and $FPR_R(c_t), TPR_R(c_t) \in [0, 1]$ and hence the AUC of $ROC(R)$ is well defined.

Remark 2.5 (Hanley and McNeil 1982; Cortes and Mohri 2004) Let $E_{Y,1} := \{[k, \ell] \in S_2 : E[k, \ell] = 1\}$ and $E_{Y,0} := \{[k, \ell] \in S_2 : E[k, \ell] = 0\}$, then the AUC of the ROC curve of predicted relationships $R \in [0, 1]^{|S_2|}$ is given by the *Wilcoxon-Mann-Whitney statistic*

$$AUC(R) = \frac{1}{|E_{Y,1}| |E_{Y,0}|} \times \sum_{[k, \ell] \in E_{Y,1}} \sum_{[k', \ell'] \in E_{Y,0}} \left(\delta_{R[k, \ell] > R[k', \ell']} + \frac{1}{2} \delta_{R[k, \ell] = R[k', \ell']} \right) \tag{2.4}$$

By the above definition the random predictor given by $R[k, \ell] = 0.5$ for all $[k, \ell] \in S_2$ induces a diagonal ROC curve, as it is the linear interpolation of the points $(0, 0)$ and $(1, 1)$, yielding an AUC of 0.5.

3 Construction and theory

In the following section we propose two general forms of graph-based predictors and derive special cases thereof. Moreover, we propose computation and simulation strategies and delineate the proposed GBPs from each other. R-code for the proposed GBPs is available at github.com/richterrob/GraphBasedPredictors.

3.1 Observed indegree predictor

We start in this subsection with the idea that a node-level statistic which is partially observed in E_X can carry non-trivial information about edge labels in E_Y . We go on to provide a specific instance of this general approach that uses the indegree as the node level statistic, leading to the *observed indegree predictor* (OIP).

GBPs based on a node-level statistic To utilize a node-level statistic to predict the unknown entries of E_Y , we need it to be both estimable from the partial observation E_X and to carry information about E_Y . Suppose $G = (V, E)$ is a causal graph and that we are given a statistic $\gamma_G : V \rightarrow W$ mapping the nodes of G to some feature space, e.g. $W = \mathbb{R}, \mathbb{Z}$. We desire of γ_G that it,

- (1.) depends only on $G = (V, E)$;
- (2.) is not constant on V ; and,
- (3.) given $\gamma_G(V) := (\gamma_G(v_1), \gamma_G(v_2), \dots, \gamma_G(v_p))^T$ there exists a predictor

$$\theta : W^p \rightarrow [0, 1]^{|S_2|}, \tag{3.1}$$

predicting the edge labels in E_Y “better than random” given (1.) and (2.) are satisfied, with “better than random” meaning that the AUC as defined in Definition 2.4 for $R = \theta(\gamma_G(V))$ is strictly larger than 0.5.

Examples of such a statistic γ_G might include

- mappings to the respective in- and outdegrees;
- mappings to the respective number of ancestors and/or descendants;

Let us give an example how (3.) might be satisfied for the above given node-level statistics. Consider a graph G with p nodes, featuring nodes $v_1, \dots, v_\ell \in V$ with no incoming and

no outgoing edge and nodes $v_{\ell+1}, \dots, v_p$ with at least one incoming and one outgoing edge (here $1 \leq \ell \leq p - 1$). In this case for v_1, \dots, v_ℓ the statistics mentioned above would either convey the information that there are no incoming edges or that there are no outgoing edges. Considering as an example the first case with γ_G assigning to each node the number of its ancestors, we can set

$$\theta(\gamma_G(V))[k, k'] = \begin{cases} 0 & \text{if } \gamma_G(V)[v_{k'}] = 0, \\ 0.5 & \text{else} \end{cases},$$

to obtain a predictor performing better than random with respect to the area under the curve of $R = \theta(\gamma_G(V))$. We formalize a graph-based predictor based on a node-level statistic in the following definition.

Definition 3.1 Let E_X be a partial observation of a causal graph G , $\gamma_G : V \rightarrow W$ a statistic on the nodes of G and θ as in (3.1). Define by $\gamma_X(v_k) := \gamma_{\tilde{G}}(v_k)$ the *partial observation of γ_G from available data X* , where \tilde{G} is the graph given by setting $E[k, \ell] = 0$ for all $[k, \ell] \in S_2$. Furthermore, assume there exists an estimator $\beta : W \rightarrow W$ of $\gamma_G(V)$ taking as an input $\gamma_X(V)$. A graph-based predictor based on a node-level statistic is defined by

$$\Theta_{\text{NLS}}(E_X) := \theta(\beta(\gamma_X(V))). \tag{3.2}$$

Assume that G , γ_G and θ of the above Definition 3.1 satisfy the desiderata (1.), (2.) and (3.) stated further above. Then, given that β predicts $\gamma_G(V)$ sufficiently well it is reasonable to claim Θ_{NLS} is performing better than random with respect to the AUC. A concrete example follows in the following subsection with the OIP including a discussion under which regime the given GBP performs better than random. For the moment let us make the following remark.

Remark 3.2 The construction of the GBP Θ_{NLS} as a general construct given in (3.2) encodes the idea “*The partial observation of a node-level statistic can carry information on unseen edges*”. Under which conditions the Θ_{NLS} performs “better” than the random baseline depends on its actual construction (i.e. choices of $\gamma_G, \theta, \beta, \mathcal{I}$) and is subject to an underlying distribution on the sets of graphs, i.e. $G \sim \mathcal{D}$.

Observed indegree predictor In the following we consider the indegree statistic by setting $\gamma_G(v_k) = \text{deg}^-(v_k)$. Consider the desiderata on γ_G of Sect. 3.1, then, given that (2.) is satisfied, we have by construction that γ_G satisfies (1.) and (3.). To see this for (3.) consider any predictor θ in (3.1) that is strictly increasing with respect to the indegree of the potential effect. It remains to assume (2.), given below as Assumption 3.3.

Assumption 3.3 Given the set-up of Assumption 2.3, deg^- is not a constant function on the vertex set V .

Note, that Assumption 3.3 is arguably a weak assumption, especially for large p . Thus, the indegree yields the following graph-based predictor, as a special case of (3.2).

Definition 3.4 Given a partial observation E_X of a causal graph G , define via

$$\Theta_{\text{OIP}}(E_X)[k, \ell] := \begin{cases} \frac{\text{deg}_X^-(v_\ell)}{|\mathcal{I}|} & \text{if } \ell \notin \mathcal{I}, \\ \frac{\text{deg}_X^-(v_\ell)}{|\mathcal{I}|-1} & \text{if } \ell \in \mathcal{I} \end{cases}, \tag{3.3}$$

the *observed indegree predictor* (OIP), where $\text{deg}_X^-(v_\ell) := |\{r \in \mathcal{I} \setminus \{\ell\} : E[r, \ell] = 1\}|$ is the *observed indegree*.

The OIP is a good candidate for a graph-based predictor under Assumption 3.3 due to the following heuristic. Assuming that the set of performed interventions \mathcal{I} was chosen independently of the edge matrix E , we have that $\text{deg}_X^-(v_k)$ is the sample mean of a hypergeometric distribution (population size $p-1$, number of success states $\text{deg}^-(v_k)$ and number of draws $|\mathcal{I}|$, with sample size 1), yielding in $(p/|\mathcal{I}|\text{deg}_X^-(v_k))$ an unbiased estimator of $\text{deg}^-(v_k)$ for all $1 \leq k \leq p$. In fact, for graphs with positive correlation structure we have the following result on the expected AUC of the OIP on a subset of S_2 .

Theorem 3.5 Let $G = (V, E)$ be such that E is drawn at random with marginal probabilities

$$E[k, \ell] \sim \begin{cases} \delta(1) & \text{if } k = \ell \\ B(1, q) & \text{else} \end{cases},$$

where $q \in (0, 1)$, with $E[k, \ell]$ and $E[k', \ell']$ drawn independently for all k, k' and all $\ell \neq \ell'$, and with a covariance structure given by

$$\text{Cov}\left(E[k, \ell], E[k', \ell'] \mid (E[\tilde{k}_j, \ell])_{j=1}^J\right) = \kappa_{N, J} > 0, \tag{3.4}$$

with $N := \sum_{j=1}^J E[\tilde{k}_j, \ell]$, for all ℓ and any pairwise distinct $k, k', \tilde{k}_1, \dots, \tilde{k}_J \in \{1, 2, \dots, p\}$, with $0 \leq J \leq p-2$. Let furthermore deg_Y^- be not constant on $V \setminus \mathcal{I}$.

Then, for any realization of the unknown relationships $M_Y \in \{0, 1\}^{|S_2|}$ we have

$$\mathbb{E}_{E_X | E_Y = M_Y} [\text{AUC}_{\mathcal{I}c}(\Theta_{\text{OIP}})] > 0.5, \tag{3.5}$$

where $\text{AUC}_{\mathcal{I}c}$ is the AUC on $\{[k, \ell] \in S_2 : \ell \notin \mathcal{I}\}$.

The proof of Theorem 3.5 can be found in ‘‘Appendix 1’’. Furthermore, we show that a subclass of latent network models (e.g. Hoff et al. 2002; Bollobás et al. 2007) fall in the setting of Theorem 3.5 (see Lemma 3 in ‘‘Appendix 1’’).

Remark 3.6 To extend Theorem 3.5 to the AUC on all of S_2 (the complete predicted E_Y by Θ_{OIP}) is at this point open. Considering the proof of Theorem 3.5 additional assumptions on the distributions of deg_Y^- and/or additional assumptions on q and $\kappa_{N, J}$ seem to be needed. For more details we refer the reader to ‘‘Appendix 1’’.

Notably, the outdegree on the other hand is not a suitable candidate for a graph-based predictor in the context of Assumption 2.3: Consider any unknown relationship $[k, \ell] \in S_2$, since E_X is formed by complete rows of E we have no observations on the outgoing edge-labels of v_k helping us to estimate $\text{deg}^+(v_k)$.

3.2 Transitivity assuming predictor

In this section we introduce a second way to construct a graph-based predictor by assuming that the graph satisfies some property relating to a non-trivial constraint(s) on its edge matrix such that E_X carries information on E_Y . Moreover, a special case of such a graph-based predictor based on transitive closedness will be derived.

GBPs based on a graph property Let the graph G in Assumption 2.3 satisfy some constraint(s) denoted by (C) , such that the partial observation E_X carries information on E_Y . We then construct a graph-based predictor via the matrix of expected values of the existence of an edge given a random draw from all graphs that satisfy (C) and are consistent with E_X . Examples of (C) might include

- the graph being transitively closed;
- the graph being a k -reachability graph;
- the nodes of the graph having an upper/lower bound on its in- and/or outdegrees.

Definition 3.7 Let \tilde{E}_X be a partial observation of a causal graph $\tilde{G} = (V, \tilde{E})$. Suppose \tilde{G} satisfies constraint(s) denoted by (C) . Then a graph-based predictor *based on a graph property (direct version)* is defined by

$$\Theta_{\text{d-GP}}(\tilde{E}_X)[k, \ell] := \frac{\left| \left\{ G : G \text{ satisfies } (C, \tilde{E}_X) \text{ and } E[k, \ell] = 1 \right\} \right|}{\left| \left\{ G : G \text{ satisfies } (C, \tilde{E}_X) \right\} \right|}, \tag{3.6}$$

where $G = (V, E)$ satisfying (\tilde{E}_X) is short for E is equal to \tilde{E} on S_1 .

We have at once the following remark.

Remark 3.8 Let $\tilde{G} = (V, \tilde{E})$ be drawn uniformly from the set of all graphs satisfying (C) , then

$$\Theta_{\text{d-GP}}(\tilde{E}_X) = \mathbb{E} \left[E_Y \mid E_X = \tilde{E}_X \right]. \tag{3.7}$$

In general $\Theta_{\text{d-GP}}$ can be very hard to compute or even to simulate. For a feasible example consider $(C) = (G \text{ is undirected (i.e. } E \text{ is symmetric) and features degree sequence } d \in \mathbb{R}^p)$ of prescribed edge degrees. In this case there exists a broad literature on how to draw (asymptotically) uniformly at random from the set $\{G : G \text{ satisfies } (C)\}$ (see e.g. Artzy-Randrup and Stone 2005; Newman 2003; Blitzstein and Diaconis 2011; Milo et al. 2003; Greenhill 2014), allowing in the worst case for Monte Carlo rejection sampling of (3.6), and, in the best case for direct sampling via a suitable adaptation of the Maslov–Sneppen MCMC algorithm. Unfortunately, similar strategies are not known, to the best of the authors’ knowledge, for drawing uniformly at random out of the set of all transitively closed graphs, not to mention the denominator set of (3.6) with $(C) = (G \text{ is transitively closed})$. However, as elaborated in the introduction, the case of transitively closed graphs is of particular interest in the context of omics readouts after gene perturbation experiments due to the fact that in conventional designs for such experiments, direct causal relationships are in general not easily distinguished from ancestral relationships. Thus, to the end of obtaining an easier to compute/simulate GBP we construct an indirect version of (3.6) described in Eq. (3.8), below.

Definition 3.9 Let \tilde{E}_X be a partial observation of a causal graph \tilde{G} . Let \tilde{G} satisfy constraint(s) denoted by (C) and let ϕ be a surjective mapping from the space of all graphs to the space of all graphs satisfying (C) . A graph-based predictor based on a graph property (indirect version) is defined by

$$\Theta_{\text{i-GP}}(\tilde{E}_X)[k, \ell] := \frac{\left| \left\{ G : \phi(G) \text{ satisfies } (\tilde{E}_X) \text{ and } \phi(E)[k, \ell] = 1 \right\} \right|}{\left| \left\{ G : \phi(G) \text{ satisfies } (\tilde{E}_X) \right\} \right|}, \tag{3.8}$$

where $\phi(E)$ is the edge matrix corresponding to $\phi(G)$.

The special case of (3.8) considered in the remainder of this Section is

$$(C) = (\tilde{G} \text{ is transitively closed})$$

for which we use $\phi(\tilde{G}) = \tilde{G}^+$. Moreover, also for the indirect version we can make a remark in the spirit of Remark 3.8.

Remark 3.10 Let $G = (V, \tilde{E})$ be drawn uniformly from the set of all graphs and let ϕ be as in Definition 3.9 mapping into the set of all graphs satisfying (C) , then

$$\Theta_{\text{i-GP}}(\tilde{E}_X) = \mathbb{E} \left[\phi(E)_Y \mid \phi(E)_X = \phi(\tilde{E})_X \right]. \tag{3.9}$$

Transitivity assuming predictors As an instance of a graph-based predictor arising from a graph property we consider in this section predicting E_Y of ancestral causal graphs. As mentioned earlier, this is motivated by the nature of omics readouts after intervention, since in such experiments what is seen is the total causal effect of perturbing gene A on gene B —potentially via mediators—rather than a necessarily direct causal effect.

Assumption 3.11 Given the set-up of Assumption 2.3, the causal graph G is ancestral.

Following Assumption 3.11, as a special case of (3.8), we define the following graph-based predictor.

Definition 3.12 Let E_X be a partial observation of an ancestral causal graph $G = (V, E)$ with S_2 being the indices of the unobserved causal relationship of G . Define by

$$\begin{aligned} \mathcal{X} &= \mathcal{X}(V, E_X) \\ &:= \{E_0 \in \mathcal{E} : E_0^+[k, \ell] = E[k, \ell] \text{ for all } [k, \ell] \in S_1\}, \end{aligned}$$

the set of all edge matrices E_0 , whose transitive closure E_0^+ coincides with E on the index set S_1 , i.e. the set of all edge matrices that are consistent with the partial observation E_X . We define

$$\Theta_{\text{TAP}}(E_X)[k, \ell] = \frac{\left| \{E_0 \in \mathcal{X} : E_0^+[k, \ell] = 1\} \right|}{|\mathcal{X}|}, \tag{3.10}$$

calling Θ_{TAP} the *transitivity assuming predictor* (TAP).

In contrast to the OIP, for which computation is straightforward, computing/simulating the TAP is non-trivial. Given a non-trivial scenario, i.e. $S_1, S_2 \neq \emptyset$, the set \mathcal{X} is determined by constraints on the $(p-1)$ -th power of E_0 . Concretely, two types of constraints surface, in detail we have $E_0 \in \mathcal{X}$ if and only if Eqs. (3.11) and (3.12) below are both satisfied.

$$E_0^{p-1}[k, \ell] = 0, \forall [k, \ell] \in S_1 \text{ s.t. } E[k, \ell] = 0, \tag{3.11}$$

$$E_0^{p-1}[k, \ell] \neq 0, \forall [k, \ell] \in S_1 \text{ s.t. } E[k, \ell] = 1. \tag{3.12}$$

A closed form for (3.10) can, to the best of the authors’ knowledge, only be given for those entries $[k, \ell] \in S_2$ which features $\Theta_{\text{TAP}}(E_X)[k, \ell] = 0$, as they are induced by the constraint (3.11) as Lemma 3.13 below implies, the proof of which is given in “Appendix 1”.

Algorithm 1 TAP - Monte-Carlo Rejection Sampling

Input: E_X partial observation of G , $T \in \mathbb{N}$ number of required successful draws, $q \in [0, 1]$ edge probability for drawing the partial RDG.

1. Set $\tau = 0$ and $\hat{\Theta}_{\text{TAP}}(E_X)^{(0,q)}[k, \ell] = 0$ for all $[k, \ell] \in S_2$.
2. Compute the set of impossible edges $K \subset S_1 \cup S_2$ using the characterization in Lemma 3.13.

while $\tau < T$ **do**

3.A Let G be an edgeless graph with p vertices. Draw E_0 as $\text{RDG}_{q,K}(G)$, using Definition 2.2.

if $E_0 \in \mathcal{X}$ **then**

3.B Set for all $[k, \ell] \in S_2$:

$$\hat{\Theta}_{\text{TAP}}(E_X)^{(\tau+1,q)}[k, \ell] := \frac{\tau \hat{\Theta}_{\text{TAP}}(E_X)^{(\tau,q)}[k, \ell] + E_0^+[k, \ell]}{\tau + 1},$$

3.C Set $\tau = \tau + 1$

end if

end while

Output: $\hat{\Theta}_{\text{TAP}}(E_X)^{(T,q)}$

Lemma 3.13 Given E_X a partial observation of an ancestral causal graph $G = (V, E)$ and let Θ_{TAP} be the TAP defined in (3.10). Then we have

$$\Theta_{\text{TAP}}(E_X)[k, \ell] = 0 \iff \mathcal{A}_{v_k} \not\subseteq \mathcal{A}_{v_\ell}, \tag{3.13}$$

where \mathcal{A}_v denotes the set of known parents of $v \in V$ given in E_X . We call edges satisfying the right hand side of (3.13) impossible edges.

To compute $\Theta_{\text{TAP}}(E_X)[k, \ell]$ beyond impossible edges, we are left with brute-force calculation with unfavourable computational complexity such that already for $p \gg 10$ calculations may be intractable. In the remainder of the chapter we propose simulation strategies of the TAP and variants thereof, which are computationally less expensive.

Rejection sampling and choice of q Algorithm 1, given below, simulates for $q = 0.5$ the TAP defined in (3.10) by straightforward Monte Carlo rejection sampling, with edge probability 0 for impossible edges, cf. Lemma 3.13. In general, it sets impossible edges to zero, draws the rest of the edge matrix entries as a partial RDG with edge probability $q \in (0, 1)$, see Definition 2.2, and, rejects the so drawn edge matrix E_0 if $E_0 \notin \mathcal{X}$. This procedure is repeated until a fixed number of $T \in \mathbb{N}$ non-discarded graphs have been drawn. By construction the so obtained $\hat{\Theta}_{\text{TAP}}^{(T,0.5)}$ is a consistent estimator of Θ_{TAP} .

The rationale for introducing parameter q in Algorithm 1 is as follows. Since the probability that an RDG features the complete graph as its transitive closure goes to 1 as $p \rightarrow \infty$ (see Karp 1990; Krivelevich and Sudakov 2013), we have to scale the parameter T with p for sufficient convergence, increasing the computational costs. Meanwhile, letting $q \rightarrow$

Algorithm 2 B-TAP - Biased Sampling from \mathcal{X}

Input: E_X partial observation of G , $T \in \mathbb{N}$ number of draws, $q \in [0, 1]$ edge probability for drawing the partial RDG.

1. Set $\hat{\Theta}_{\text{B-TAP}}(E_X)^{(0,q)}[k, \ell] = 0$ for all $[k, \ell] \in S_2$.
2. Compute the set of impossible edges $K \subset S_1 \cup S_2$ using the characterization in Lemma 3.13.
3. Let $G = (V, E)$ be given by $|V| = p$ and

$$E[k, \ell] = \begin{cases} 0 & \text{if } [k, \ell] \in K, \\ 1 & \text{else.} \end{cases}$$

for $\tau = 1, 2, \dots, T$ **do**

4.A Draw E_{-1} as $\text{RDG}_{q,K}(G)$, using Definition 2.2.

4.B For each node v_k with $k \in \mathcal{I}$ let G_k be the subgraph on the nodes $\{v_k\} \cup \mathcal{D}_{v_k}$, where \mathcal{D}_{v_k} denotes the set of known descendants of v_k . Draw modified RSTs $\tau_{\text{mod}}^{(k)}$ of G_k rooted in v_k , cf. Sect. 3.2.

4.C Set

$$E_0[k, \ell] := \begin{cases} 1 & \text{if } \exists r \in \mathcal{I} \text{ s.t. } [k, \ell] \in \mathcal{T}_{\text{mod}}^{(r)} \\ E_{-1}[k, \ell] & \text{else} \end{cases}. \tag{3.13}$$

4.D Set for all $[k, \ell] \in S_2$:

$$\hat{\Theta}_{\text{B-TAP}}(E_X)^{(\tau+1,q)}[k, \ell] := \frac{\tau \hat{\Theta}_{\text{B-TAP}}(E_X)^{(\tau,q)}[k, \ell] + E_0^+[k, \ell]}{\tau + 1}.$$

end for

Output: $\hat{\Theta}_{\text{B-TAP}}(E_X)^{(T,q)}$

0 as $p \rightarrow \infty$ reduces the convergence time of Algorithm 1, as we will see in Fig. 12 in ‘‘Appendix 1’’ (in particular with regards to Algorithm 2 further below) where q is chosen with respect to the sparsity of the observed graph. The caveat of choosing $q \neq 0.5$ is that $\hat{\Theta}_{\text{TAP}}^{(T,q)} \xrightarrow{T \rightarrow \infty} \Theta_{\text{TAP}}^{(q)}$ which in general is not equal to Θ_{TAP} , i.e. $\hat{\Theta}_{\text{TAP}}^{(T,q)}$ is for $q \neq 0.5$ not a consistent estimator of Θ_{TAP} , as shown in Lemma 3.22 in Sect. 3.4.

Biased sampling from \mathcal{X} Even for q selected smaller and smaller as the size of the graph p grows, since the rejection sampler of Algorithm 1 draws an ever growing number of discarded edge matrices, the computational costs of Algorithm 1 sill grow with $p \rightarrow \infty$, see Fig. 12 in ‘‘Appendix 1’’. To the end of reducing computational costs of Algorithm 1 further, consider Algorithm 2 avoiding rejections all together. Additional to the exclusion of impossible edges, Algorithm 2 includes a step drawing spanning trees to ensure the inequality constraints of (3.12) are met by pasting them in the partial RDGs drawn in Step 3.A. To this end introduce the Broder Algorithm below.

Definition 3.14 (Broder 1989) Given an un-directed graph $G = (V, E)$, i.e. a graph as introduced in Sect. 2.2 with symmetric E . Assume G to be connected. Draw a random spanning tree (RST) rooted in v_1 by simulating a random walk $x_1, x_2, x_3, \dots, x_T$ on G with $x_1 = v_1$ and stopping time $T \in \mathbb{N}$ such that every vertex is visited at least once.

Denote for each vertex $v_k \neq v_1$ the index t_k featuring $x_{\min(\{1 \leq \ell \leq T: x_\ell = v_k\})-1} = v_{t_k}$, i.e. v_{t_k} is the predecessor of the first visit of the random walk to v_k . Then, the RST is given by the set of edges

$$\mathcal{T} := \{[t_k, k] : 2 \leq k \leq p\}.$$

In Broder (1989) it is shown that an RST of Definition 3.14 is drawn uniformly at random out of the set of all spanning trees of G rooted in v_1 . However, for a directed graph G which is not strongly connected, a random walk as in Definition 3.14 could get “stuck” (compare also Anari et al. 2020). Consider in the following a directed graph $G = (V, E)$ featuring a path from v_1 to any other vertex. To the end of drawing a computationally feasible spanning tree in G rooted in v_1 we use a **modified version of the RST**:

1. Set $\mathcal{Y} := \{[k, \ell] \in \{1, 2, \dots, p\}^2 : E[k, \ell] = 0\}$ to be the set of all non-edges in G .
2. Set $W = \{2, 3, \dots, p\}$ to be the set of all non-visited vertices.
3. Set $\kappa = 0$ and set $\mathcal{T}_{\text{mod}} := \emptyset$.
4. **while** $\kappa = 0$:
 - (a) Consider the complete graph \tilde{G} on V .
 - (b) Draw a RST of \tilde{G} rooted in v_1 denoted by $\mathcal{T}_0 := \{[k_1, \ell_1], [k_2, \ell_2], \dots, [k_{p-1}, \ell_{p-1}]\}$ sorted by their appearance in the random walk of Definition 3.14.
 - (c) Let

$$m := \min(\{1 \leq r \leq p - 1 : \ell_r \in W \text{ and } [k_r, \ell_r] \in \mathcal{Y} \cup \{p\}\})$$

and set

$$\mathcal{T}_{\text{mod}} = \mathcal{T}_{\text{mod}} \cup \{[k_s, \ell_s] \in \mathcal{T}_0 : \ell_s \in W \text{ and } s < m\}.$$

- (d) Set

$$W = \{2 \leq \ell \leq p : \nexists k \text{ s.t. } [k, \ell] \in \mathcal{T}_{\text{mod}}\}$$

and if $W = \emptyset$ set $\kappa = 1$.

We call the so obtained \mathcal{T}_{mod} a modified RST (m-RST). Note that the above construction does not vouch for \mathcal{T}_{mod} being drawn uniformly at random out of the set of all spanning trees rooted in v_1 . In the case that $\mathcal{Y} = \emptyset$ however the draw of the modified RST coincides with the draw of an RST. Since, as we will show in Sect. 3.4, Algorithm 2 does not draw uniform at random from \mathcal{X} even if the spanning tree was drawn uniformly at random from all spanning trees, we except this caveat for the sake of computational simplicity. In particular, we have that $\hat{\Theta}_{\text{B-TAP}}^{(T,q)} \xrightarrow{T \rightarrow \infty} \text{B-TAP}^{(q)}$ which is in general

not equal to $\Theta_{\text{TAP}}^{(q)}$, for any $q \in [0, 1]$. We call the predictor $\text{B-TAP}^{(q)}$ *biased transitivity assuming predictor* (B-TAP) with edge-probability $q \in (0, 1)$. As for Algorithm 1 with growing p we propose to choose q according to the sparsity of the observed graph for feasible run times.

3.3 Extensions

The graph-based predictors defined in (3.2), (3.6) and (3.8) are related. Furthermore, additional graph-based predictors could be constructed. In the following section we exemplify this.

First, given Assumption 3.11 the graph G has to stem from a quite restrictive subset of all graphs in order not to satisfy Assumption 3.3, as Lemma 3.15 below shows.

Lemma 3.15 *Let $G = (E, V)$ be a transitively closed graph such that $\text{deg}^-(v_k) = n$ for all $v_k \in V$. Then there exists $m, K \in \mathbb{N}$ with $Km = n$ such that G has K strongly connected components of cardinality m that each form complete subgraphs.*

The proof of Lemma 3.15 can be found in “Appendix 1”. Due to Lemma 3.15 we can motivate the OIP not only by the type of observations E_X – complete rows – but also by the heuristic of observing ancestral graphs. This leads to a combination of the TAP and the OIP given below.

Definition 3.16 Given a partial observation E_X of a causal graph G , let K be the set of impossible edges as given by Lemma 3.13. Define the *transitivity-assuming observed indegree predictor* (T-OIP) by

$$\Theta_{\text{T-OIP}}(E_X)[k, \ell] := \begin{cases} 0 & \text{if } [k, \ell] \in K, \\ \frac{\text{deg}_{\mathcal{I}}^-(v_\ell) + 1}{|\mathcal{I}| + 1} & \text{if } [k, \ell] \notin K \text{ and } \ell \notin \mathcal{I}, \\ \frac{\text{deg}_{\mathcal{I}}^-(v_\ell) + 1}{|\mathcal{I}|} & \text{else} \end{cases} \quad (3.14)$$

Note, that to define the T-OIP, the assumption of transitivity is not needed.

Second, we can extend the definition of the TAPs, from ancestral causal graphs to all possible causal graphs. This is particularly important for omics data: First, because the TAPs should be computable even if Assumption 3.11 does not hold, for example when assuming that the causal effect dies out over long causal chains. Second, because we need to be able to compute TAPs also in the case of faulty assignments in E_X , e.g. due to measurement errors. To this end we introduce the following relaxed versions.

Definition 3.17 Given a partial observation E_X of a causal graph G , let $\tilde{G} = (V, \tilde{E})$ be given by

$$\tilde{E}[k, \ell] := \begin{cases} 0 & \text{if } [k, \ell] \in S_2, \\ E[k, \ell] & \text{else} \end{cases}, \tag{3.15}$$

and let $\tilde{E}_X^+ = (\tilde{E}^+[k, \ell])_{[k, \ell] \in S_1}$. Define the TAP of E_X by

$$\Theta_{\text{TAP}}(E_X) = \Theta_{\text{TAP}}(\tilde{E}_X^+), \tag{3.16}$$

and define analogously $\Theta_{\text{TAP}}^{(q)}$ and $\Theta_{\text{TAP}}^{(q)}$. Moreover, using (3.15) we can define *pre-processed* versions of the Θ_{OIP} and the $\Theta_{\text{T-OIP}}$ by

$$\Theta_{\text{P-OIP}}(E_X) := \Theta_{\text{OIP}}(\tilde{E}_X^+),$$

and,

$$\Theta_{\text{P-T-OIP}}(E_X) := \Theta_{\text{T-OIP}}(\tilde{E}_X^+).$$

3.4 Non-equivalence of the proposed predictors

Having introduced multiple predictors, using closely related heuristics, cf. Lemma 3.15, the question arises whether the respective ROC curves of the predictors are related or even coincide. To this end, we provide a set of counterexamples demonstrating the differences in predicted values and, when applicable, differences in induced ROC curves between the predictors. The first example shows that $\Theta_{\text{TAP}} \neq \Theta_{\text{B-TAP}}^{(0.5)}$ and in particular that the random draw from \mathcal{X} described in Algorithm 2 is not uniform even if the RST are drawn uniformly at random. To compare the marginal distribution on the edges of the drawn graphs from \mathcal{X} introduce

$$\begin{aligned} \theta_{\text{TAP}}(E_X)[k, \ell] &= \mathbb{P}[E_0[k, \ell] = 1 | E_0 \in \mathcal{X}] \\ &= \frac{|\{E_0 \in \mathcal{X} : E_0[k, \ell] = 1\}|}{|\mathcal{X}|}, \end{aligned} \tag{3.17}$$

as the marginal conditional probability of the existence of an edge when drawing E_0 according to Algorithm 1 with $q = 0.5$ conditioned on $E_0 \in \mathcal{X}$. We have at once the following Corollary to Lemma 3.13, for a proof see ‘‘Appendix 1’’.

Corollary 3.18 Given E_X a partial observation of an ancestral causal graph G and let $\theta_{\text{TAP}}(E_X)$ be given as in (3.17). Then we have for $[k, \ell] \in S_2$ that

$$\theta_{\text{TAP}}(E_X)[k, \ell] = 0 \Leftrightarrow \mathcal{A}_{v_k} \not\subseteq \mathcal{A}_{v_\ell}$$

holds, where \mathcal{A}_v is the set of known parents of $v \in V$ in G .

Moreover, the following Lemma shows that edges that are ‘‘not-impossible’’ edges between nodes without a known ancestor in common feature $\theta_{\text{TAP}} = 1/2$.

Lemma 3.19 Given E_X a partial observation of an ancestral causal graph G and let θ_{TAP} be as in (3.17). Then we have for $[k, \ell] \in S_2$ with $\mathcal{A}_{v_k} \subseteq \mathcal{A}_{v_\ell}$ that

$$\theta_{\text{TAP}}(E_X)[k, \ell] = \frac{1}{2} \Leftrightarrow (\mathcal{A}_{v_k} \setminus \{v_\ell\}) \cap \mathcal{A}_{v_\ell} = \emptyset,$$

where \mathcal{A}_v is the set of known parents of $v \in V$ in G given in E_X .

The proof is given in ‘‘Appendix 1’’. Given the above we state below the counterexample for $\Theta_{\text{TAP}} \neq \Theta_{\text{B-TAP}}^{(0.5)}$. Note that in the following, for the sake of readability, we will augment the image space of the predictors to $[0, 1]^{p \times p}$ instead of $[0, 1]^{|S_2|}$.

Example Given a set of nodes $V = \{v_1, v_2, v_3, v_4\}$ and a partial observation E_X of $G = (V, E)$ as depicted in Fig. 2a. Brute force calculation of all graphs in \mathcal{X} yields

$$\begin{aligned} (\theta_{\text{TAP}}(E_X)[k, \ell])_{k, \ell=1}^4 &= \frac{1}{24} \begin{pmatrix} 24 & 18 & 18 & 0 \\ 12 & 24 & 15 & 0 \\ 12 & 15 & 24 & 0 \\ 12 & 12 & 12 & 24 \end{pmatrix} \\ &\neq \frac{1}{24} \begin{pmatrix} 24 & 20 & 20 & 0 \\ 12 & 24 & 16 & 0 \\ 12 & 16 & 24 & 0 \\ 12 & 12 & 12 & 24 \end{pmatrix} = (\theta_{\text{B-TAP}}^{(0.5)}(E_X)[k, \ell])_{k, \ell=1}^4, \end{aligned} \tag{3.18}$$

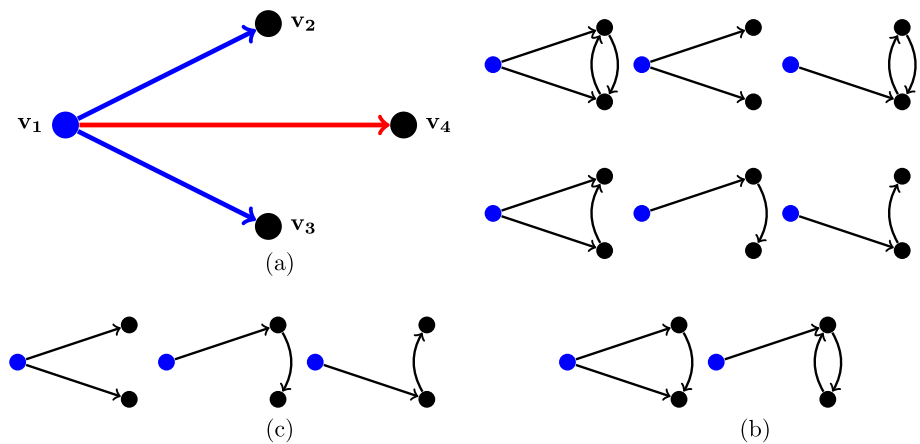
where $\theta_{\text{B-TAP}}^{(0.5)}(E_X)[k, \ell]$ denotes the marginal probability of $E_0[k, \ell] = 1$ when drawing E_0 according to Algorithm 2. Note that there are no impossible edges present in the sub-graph on $\{v_1, v_2, v_3\}$ and thus when drawing a tree from Fig. 2c we draw uniformly at random from the set of all spanning trees rooted in v_1 , cf. Sect. 3.2. Computing furthermore the predictors Θ_{TAP} and $\Theta_{\text{B-TAP}}^{(0.5)}$ yields

$$\begin{aligned} \Theta_{\text{TAP}}(E_X) &= \frac{1}{768} \begin{pmatrix} 768 & 768 & 768 & 0 \\ 504 & 768 & 624 & 0 \\ 504 & 624 & 768 & 0 \\ 582 & 654 & 654 & 768 \end{pmatrix} \\ &\neq \frac{1}{768} \begin{pmatrix} 768 & 768 & 768 & 0 \\ 512 & 768 & 640 & 0 \\ 512 & 640 & 768 & 0 \\ 584 & 656 & 656 & 768 \end{pmatrix} = \Theta_{\text{B-TAP}}^{(0.5)}(E_X). \end{aligned} \tag{3.19}$$

A detailed computation of the above matrices is given in ‘‘Appendix 1’’, consider to this end also (b) and (c) of Fig. 2. In this example we observe:

1. The marginal distributions $\theta_{\text{B-TAP}}^{(0.5)}(E_X)$ of Algorithm 2 and the resulting prediction $\Theta_{\text{B-TAP}}^{(0.5)}(E_X)$ are not equal to $\theta_{\text{TAP}}(E_X)$ and $\Theta_{\text{TAP}}(E_X)$, respectively. Hence, $\Theta_{\text{B-TAP}}^{(T, 0.5)}(E_X)$ does not converge to Θ_{TAP} for $T \rightarrow \infty$.

Fig. 2 The partial observation E_X of the ancestral causal graph G given in the Example (\mathcal{I} and known edges in blue, known non-edges in red) (a). The eight possible edge label configurations on $\{[1, 2], [1, 3], [2, 3], [3, 2]\}$ for an edge matrix $E_0 \in \mathcal{X}$, i.e. consistent with the ancestral causal relationships given in E_X (b). Possible spanning trees on $\mathcal{D}_{v_1} \cup \{v_1\}$ ensuring consistency with E_X (c)



2. In this example, the order of matrix entries of $\Theta_{TAP}(E_X)$ is preserved by $\Theta_{B-TAP}^{(0.5)}(E_X)$, hence, the induced ROC curves and thus AUC scores are the same by Definition 2.4.

To the best of the authors’ knowledge a counterexample of different ROC curves for the TAP and the B-TAP is not known. As a consequence we make the following conjecture.

Conjecture 3.20 *Given a partial observation E_X of an ancestral causal graph G . Under (possibly quite restrictive) conditions on the descendant sets \mathcal{D}_{v_k} for $k \in \mathcal{I}$ we have that the ROC curves induced by Θ_{TAP} and $\Theta_{B-TAP}^{(0.5)}$ coincide.*

Staying with the above Example we show $\Theta_{T-OIP}(E_X) \neq \Theta_{TAP}(E_X)$ and, even more, that the induced ROC curves might differ.

Example (cont’d) Let G and its partial observation E_X be as before. Compute

$$\begin{aligned}
 (\Theta_{T-OIP}(E_X)[k, \ell])_{k, \ell=1}^4 &= \frac{1}{6} \begin{pmatrix} 6 & 6 & 6 & 0 \\ 3 & 6 & 6 & 0 \\ 3 & 6 & 6 & 0 \\ 3 & 6 & 6 & 6 \end{pmatrix} \\
 &= \frac{1}{768} \begin{pmatrix} 768 & 768 & 768 & 0 \\ 384 & 768 & 768 & 0 \\ 384 & 768 & 768 & 0 \\ 384 & 768 & 768 & 768 \end{pmatrix} \neq \Theta_{TAP}(E_X).
 \end{aligned}$$

This yields,

$$\Theta_{TAP}(E_X)[4, 1] > \Theta_{TAP}(E_X)[2, 1],$$

while,

$$\Theta_{T-OIP}(E_X)[4, 1] = \Theta_{T-OIP}(E_X)[2, 1].$$

Hence, given a that $E[4, 1] = 1 \neq 0 = E[2, 1]$ the T-OIP and the TAP induce different ROC curves.

Analogously to Conjecture 3.20 we conjecture that T-OIP is a “coarser” predictor than the TAP in the following.

Conjecture 3.21 *Given a partial observation E_X of an ancestral causal graph G underlying (possibly quite restrictive) conditions. We have for edges $[k, \ell], [r, s] \in S_2$ (possibly underlying some condition) that*

$$\begin{aligned}
 &(\Theta_{T-OIP}(E_X)[k, \ell] > \Theta_{T-OIP}(E_X)[r, s]) \\
 &\Rightarrow (\Theta_{TAP}(E_X)[k, \ell] > \Theta_{TAP}(E_X)[r, s]).
 \end{aligned}$$

Last, Lemma 3.22, below, shows that changing q in Algorithm 1 may lead to potentially different ROC curves.

Lemma 3.22 *There exists an ancestral causal graph G and a partial observation E_X of G , as well as, $q_0 \in (0, 1)$ and edges $[k, \ell], [k', \ell'] \in S_2$, such that*

$$\Theta_{TAP}^{(0.5)}[k, \ell] > \Theta_{TAP}^{(0.5)}[k', \ell'],$$

and,

$$\Theta_{TAP}^{(q_0)}[k, \ell] < \Theta_{TAP}^{(q_0)}[k', \ell'].$$

The proof can be found in “Appendix 1”. Similarly we can deduce that changing q in Algorithm 2 with RST drawn uniformly at random may lead to potentially different ROC curves, leading us to conjecture that one can also find a counterexample for Algorithm 2 with RSTs drawn as m-RSTs, we refer again to “Appendix 1” for details.

4 Simulation study

In this section we study the use of the graph-based predictors as baselines in the case where the underlying ground truth graph satisfies Assumption 3.11 and beyond. To this end, we use simulated and real graphs.

4.1 On simulated graphs

We simulate graphs of cardinality p as transitive closures of RDGs with edge probability α/p governed by a sparsity parameter $\alpha \in (0, 1)$. Note that the dependence of the sparsity on p is needed in order not to draw only graphs featuring the complete graph as their transitive closure (Krivelevich and Sudakov 2013).

In Fig. 3 box plots of the AUC performance of the ROC curves for 20 runs are presented for varying graph size p . In Fig. 3 the parameter α was set to 0.7 and the amount of known rows was given by $|\mathcal{I}| = p/5$. Compared are the predictors TAP (Algorithm 1, $(T, q) = (100, 0.5)$), TAP-q (Algorithm 1, $(100, \alpha/p)$), B-TAP (Algorithm 2, $(100, 0.5)$), B-TAP-q (Algorithm 2, $(100, \alpha/p)$), T-OIP and the OIP. The TAP, TAP-q, B-TAP and B-TAP-q are only computed until $p = 25, p = 100, p = 1000$ and $p = 1000$, respectively, due to their exploding computational costs (see “Appendix 1”). For all p and all predictors the respective performance is on average better than random. While the variability in AUC performance decreases with growing p , the mean performance increases for all but the B-TAP and TAP. For large p the B-TAP and the TAP suffer from their slow convergence, which is especially visible when compared to the B-TAP-q and TAP-q, respectively. It stands out that the OIP, T-OIP and the B-TAP-q have a similar performance and substantially outperform the classic random baseline (at 0.5 AUC). Moreover, the OIP and the T-OIP were by a margin the fastest to compute, see for a comparison of computation times Fig. 12 in “Appendix 1”.

For the influence of α and $|\mathcal{I}|$ on the B-TAB, B-TAB-q, OIP and T-OIP performance we refer the reader to Figs. 9 and 10 in “Appendix 1”. In summary, the order in performance of the methods remains mainly unchanged. Furthermore, for some example mean ROC curves of Fig. 3 we refer to Fig. 8 in “Appendix 1”.

In Fig. 4 we present the AUC performance of the B-TAP, B-TAP-q, OIP and T-OIP in the case the ground truth graph is a k -reachability graph and thus violates Assumption 3.11 to various extends. A graph $G = (V, E)$ is the k -reachability graph of a graph $\tilde{G} = (V, \tilde{E})$ if we have

$$(E[k, \ell] = 1) \Leftrightarrow (\exists \text{ a path of length } \leq k \text{ from } v_k \text{ to } v_\ell).$$

In particular, $k = 1$ yields $G = \tilde{G}$ and $k \geq p - 1$ yields $G = \tilde{G}^+$. For Fig. 4 we drew a RDG with edge probability $0.7/p$ and graph size $p = 1000$ and computed the respective k -reachability graph. For each, the number of known rows was set to $|\mathcal{I}| = 200$. We observe that already for $k = 25$ the AUC performance was comparable to the AUC performance on the transitively closed graph ($k = 1000$). Meanwhile,

performance did not decrease drastically for $k = 2, 5$ and prediction performance for all predictors remains better than random. One reason might be that Assumption 3.3 continues to hold even if Assumption 3.11 is violated. Additionally, drawing k -reachability graphs in this way the probabilities of existence of incoming edges at a particular node are positively correlated relating to our findings in Theorem 3.5. Note that the T-OIP loses its advantage over the OIP from incorporating the impossible edges the more Assumption 3.11 is violated. As in Fig. 3 we see that the B-TAP performs significantly worse compared to the B-TAP-q due to its slower convergence with respect to T . Last, for $k = 1$ we see a performance of all predictors around random, which could be expected, as for randomly drawn graphs the expected indegree of each node is equal, possibly violating Assumption 3.3.

4.2 On graphs derived from “omics”-data

In the following we test the new predictors on real yeast gene expression data¹ from Kemmeren et al. (2014) (used for CSL by Meinshausen et al. 2016) and on proteomics data² from Sachs et al. (2005) (used for CSL by Wang et al. 2017). Compared are the baselines proposed in this paper with the performance of the PC and IDA algorithms (see Spirtes et al. 2000; Maathuis et al. 2009, respectively), the MCMC-Mallow approach by Rau et al. (2013), the GIES algorithm (Hauser and Bühlmann 2012) (using the R-package pcalg Kalisch et al. 2012) and the IGSP algorithm of Wang et al. (2017).

As the backgrounds of all approaches vary let us make some remarks on their usage in this study:

- For PC, GIES and IGSP the output is an estimated graph (rather than a matrix of scores) and as such only points (one for each run) on the ROC plane are depicted in Fig. 7 and comparison via the AUC is not possible.
- The PC and IDA algorithms are considering any measurement as observational, as they are not designed to deal with interventional measurements. To the end of a fair comparison, we report their performance when only the available observational measurements (i.e. X_2) are passed to the algorithms (denoted by (*obs*)) and their performance when all available measurements (i.e. X) are passed to the respective algorithms (denoted by (*int-obs*)). Note, that even when interventional measurements are passed, they are treated by PC and IDA as observational.

¹ The data can be found at <https://deleteome.holstegelab.nl/> under the tab Downloads>Causal inference.

² The data can be found at <https://github.com/yuhaow/sp-intervention>.

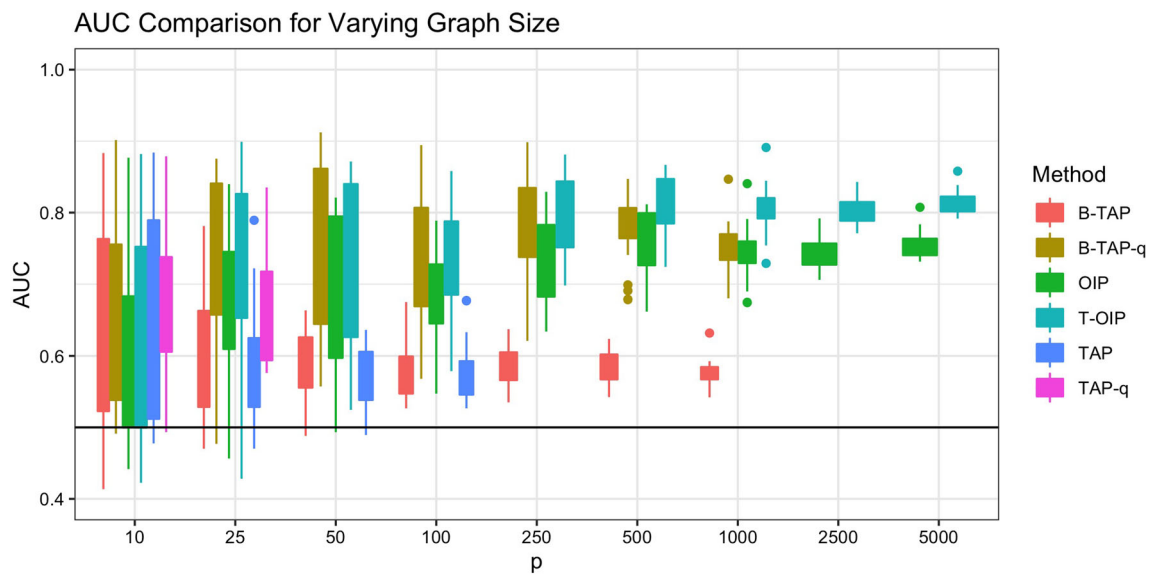


Fig. 3 Simulation study, AUC performance for varying graph size p

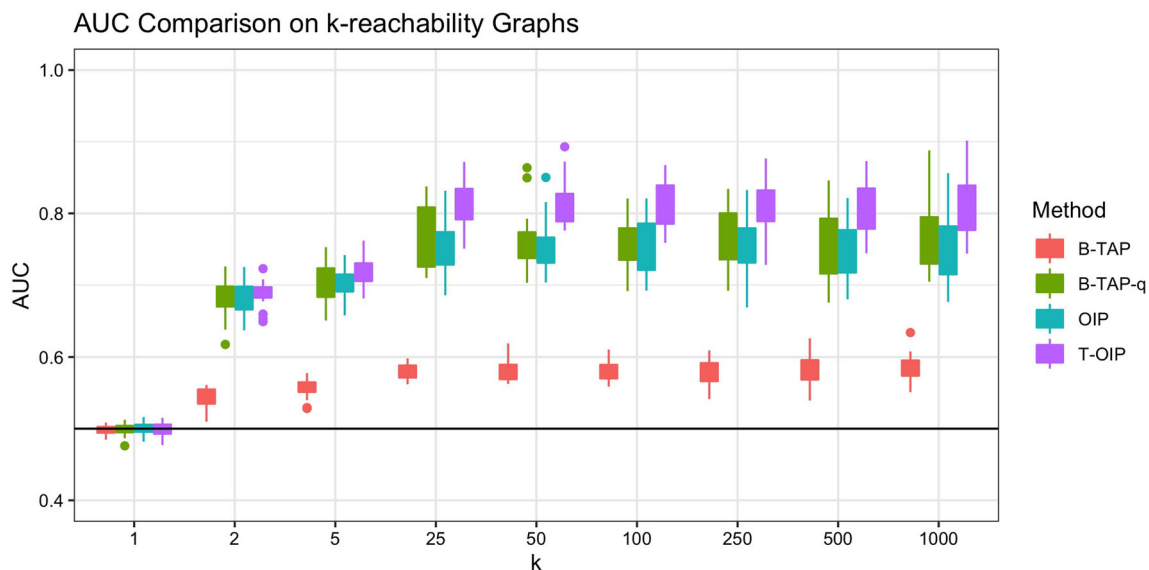


Fig. 4 Simulation study, AUC performance on k -reachability graphs

- The IGSP requires more than one interventional measurement per intervention, as this is not available for the Kemmeren et al. (2014) dataset the IGSP is only evaluated on the Sachs et al. (2005) dataset.
- Default parameter choices have been used. In detail for PC (and thus IDA) we chose α_{PC} to be 0.01 as proposed in Kalisch et al. (2012), for IGSP α_{IGSP} has been set to 0.2 as it was among the best performing α 's in the corresponding experiment in Wang et al. (2017) and the MCMC-Mallow algorithm has been used with constants set as in the accompanying R-code of Rau et al. (2013).

Transcriptomics data (Kemmeren et al.) The data consists of gene expression readouts of 262 observational experiments (i.e. with no intervention) and 1479 interventional experiments (each interventions is on a single gene, specifically knock-outs; each intervention targeting a different gene), measured are 6170 genes in total (including the 1479 intervened upon genes). We consider in this evaluation the “square” graph using only the readouts of the 1479 genes that have been intervened upon. Denote by $X_1 \in \mathbb{R}^{\tilde{p} \times p}$ the available interventional measurements and by $X_2 \in \mathbb{R}^{N_1 \times p}$ the available observational measurements, denote furthermore

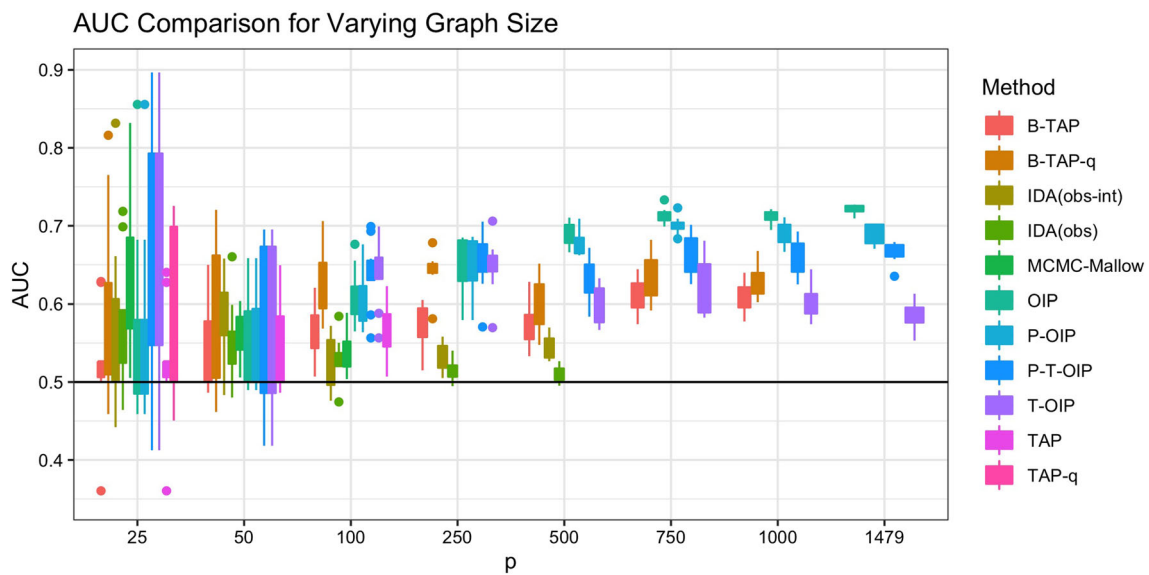


Fig. 5 AUC on the yeast transcriptomics data of Kemmeren et al. (2014) performance with varying graph size, the gold-standard-threshold is fixed at $Z = 5$

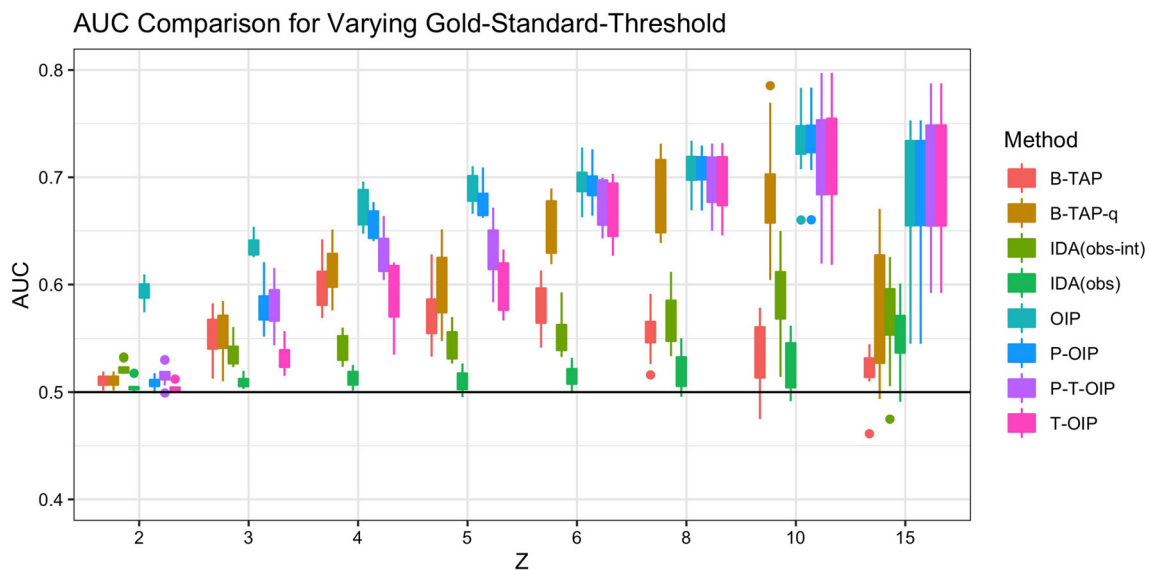


Fig. 6 AUC performance on the yeast transcriptomics data of Kemmeren et al. (2014) with varying gold-standard-threshold, the size of the graph is fixed at $p = 500$

by $Y_1 \in \mathbb{R}^{(p-\tilde{p}) \times p}$ and $Y_2^{N_2 \times p}$ their unavailable counterparts, cf. Assumption 2.3, and assume (if necessary via reordering) that row k and column k correspond to gene v_k . Then the partial observation E_X is constructed by the following gold-standard-rule:

$$(E[k, \ell] = 1) \Leftrightarrow (|X_1[k, \ell] - \text{Med}(X_2[\cdot, \ell])| > Z \cdot \text{IQR}(X_2[\cdot, \ell])),$$

where $X_1[k, \ell]$ is the readout of gene v_ℓ after the intervention on v_k , $\text{Med}(\cdot)$ assigns its median to a vector, $Z > 0$ is the *gold-standard-threshold* and $\text{IQR}(\cdot)$ assigns its interquartile-distance to a vector, i.e. there exists an edge from A

to B if and only if the readout of B under intervention on A has an absolute z -score higher than Z with respect to the empirical distribution of readouts of B under no intervention. The unobserved causal relationships E_Y are constructed analogously via Y_1 and Y_2 with the same gold-standard-threshold Z . Given a graph size p , the following protocol was used to obtain available and unavailable data:

1. Pick p of the 1479 genes at random and discard the rest.
2. Pick $\tilde{p} = \lceil p/5 \rceil$ rows of the interventional readouts at random, those constitute X_1 . The remaining rows constitute Y_1 .

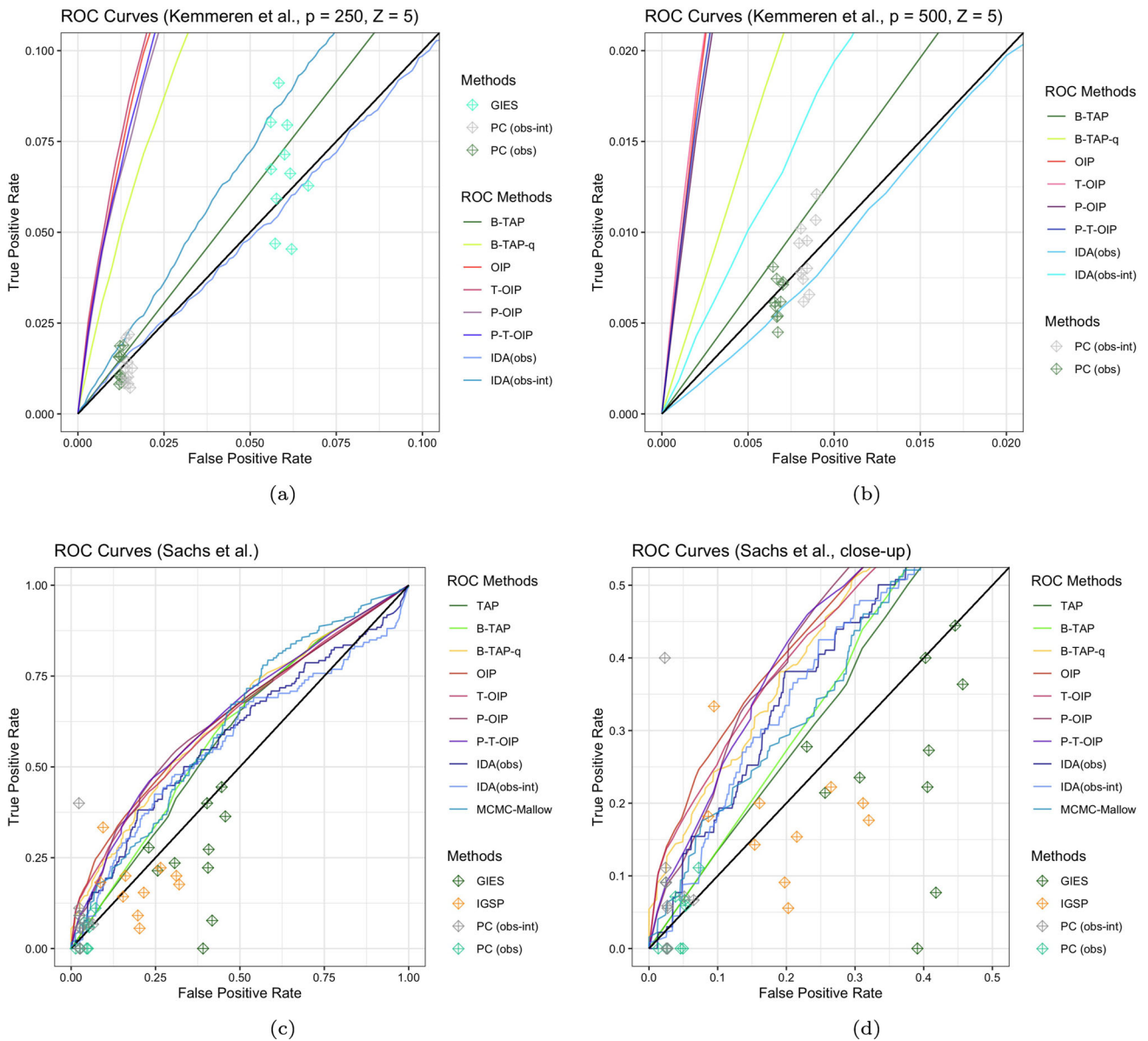


Fig. 7 Mean ROC curves on the Kemmeren et al. (2014) data for $p = 250$ (a zoomed in at $[0, 0.02]^2$) and $p = 500$ (b zoomed in at $[0, 0.1]^2$) and on the Sachs et al. (2005) data, complete (c) and zoomed in at $[0, 0.5]^2$

3. Pick $N_1 = 131$ (half) of the rows of the observational readouts at random, those constitute X_2 . The remaining rows constitute Y_2 .

For Figs. 5 and 6 the above protocol was repeated 10 times, T was set to 100 for TAP, TAP-q, B-TAP and B-TAP-q, and q of TAP-q and B-TAP-q was set to the sparsity of the partial observation provided. While in Fig. 5 the graph size p varies and Z is set to 5, in Fig. 6 p is set to 500 and Z varies. Due to computational demands it was not feasible to apply all methods for all p . Figures 5 and 6 show that the proposed graph-based predictors clearly outperform the classical random baseline on the given data set. Moreover, they outper-

form IDA and MCMC-Mallow (where the latter ones were computed). The order of performance holds generally also for varying Z , in particular the OIP consistently outperforms IDA. Interestingly, for large Z corresponding to considering only “large” effects the differences in performance between the OIP and TAPs seem to slightly diminish, while as Z decreases only the OIP achieves a performance clearly better than random. This suggests that Assumption 3.11 may hold in practice in particular when considering larger effects.

In (a) and (b) of Fig. 7 close-ups of the mean ROC curves for $p = 250$ and $p = 500$ are displayed. For methods producing an estimated graph results are shown as points on the ROC plane. For both, PC and GIES, we observe a perfor-

mance slightly above random which is outperformed by the OIPs and the TAPs. Moreover, on closer inspection the ascent of the OIPs and TAPs is particularly steep at the start of the ROC curves in the bottom left corner, a region often considered important when CSL methods are used for hypothesis generation (see e.g. Colombo et al. 2012; Meinshausen et al. 2016).

Proteomics data (Sachs et al.) The data consists of protein measurements of 992 observational experiments (i.e. with no interventions) and in total 13435 interventional experiments, each targeting a single protein, spread over 8 target-proteins (the number of interventional measurements per target-protein varies between 301 and 3602). In total 24 proteins are measured (among them the 8 targeted in the interventions).

As sample size for the interventional experiments is far larger compared to the data from Kemmeren et al. the two-sided Wilcoxon-ranksum test is used to construct the ground truth as done in Wang et al. (2017). In detail, given available observational measurements X_2 and available interventional measurements $X_{1,k}$ with k corresponding to the targeted intervention, i.e. $X_1 = (X_{1,1}^T \cdots X_{1,\tilde{m}}^T)^T$ (for some $1 \leq \tilde{m} \leq 7$), we say that there is an edge from protein k to protein ℓ , i.e. $E_X[k, \ell] = 1$, if the two-sided Wilcoxon-ranksum test rejects (at significance level 0.05) the null hypothesis that the samples $(X_2[\cdot, \ell])$ and $(X_{1,k}[\cdot, \ell])$ stem from the same distribution. Via the same gold standard rule E_Y is constructed from Y_1 and Y_2 . We followed the protocol below:

1. Pick $\tilde{m} = 4 = 8/2$ interventional targets at random, all of their interventional measurements combined constitute X_1 . The remaining measurements, namely those targeting one of the other four interventional targets, constitute Y_1 .
2. Pick $496 = 992/2$ rows of the observational measurements at random, those constitute X_2 . The remaining rows constitute Y_2 .

In (c) and (d) of Fig. 7 the mean ROC curves over 10 runs of the protocol are compared. Again, for methods producing an estimated graph results are shown as points on the ROC plane. Even on this graph with a few number of nodes and with only $|\mathcal{I}| = 4$ we observe a better performance than random of the GBPs, in particular the variants of the OIP and the TAP even outperform the IDA and perform comparably or slightly better than the MCMC-Mallow approach, compare also the AUC comparison in Fig. 11 in “Appendix 1”. Moreover, we see that CSL methods outputting an estimated graph in fact

lie only in a minority of runs over the mean OIP ROC curve (PC (obs-int) (2-3/10), IGSP (1-2/10)), or in fact, never as is the case for PC (obs) and GIES.

Furthermore, in Fig. 12 of “Appendix 1” the computational costs of Figs. 5 and 7 are reported. In particular the OIPs have very low computation times, while the MCMC-Mallow and IGSP take considerable longer to compute.

5 Discussion

In this paper we have argued for new baselines to evaluate causal structure learning methods on interventional data, as a complement to random baselines that in some settings may represent a “low bar”. The inclusion of interventional measurements carries information not only on the edges of the causal graph corresponding to the available interventional measurements, but also, to some extent, on remaining edges in the graph. This is why in settings where such data are available, simple heuristics to account for the available information can provide improved baselines. For these settings we introduced three general graph-based predictors, cf. (3.2), (3.6) and (3.8). Motivated by large-scale systems biology experiments we went on to consider special cases of (3.2) and (3.8) in the observed indegree predictor (OIP) and the transitivity assuming predictor (TAP) and extensions thereof. We showed that the OIP will perform under quite general conditions better than the random baseline and we showed theoretical differences of the introduced predictors. The potential of the OIPs and TAPs as more challenging baselines were demonstrated in a simulation study as well as on real data. In fact on real data the newly defined baselines can outperform standard CSL methods (with default tuning parameter values), although it should be emphasized that in the particular application studied, the assumptions underpinning some of the methods may not hold and furthermore in some examples we had to apply the methods in ways that deviate from their intended use.

In the future new graph-based predictors could be defined for specific use-cases. Moreover, an evaluation of the baseline’s performance on further metrics, beyond the ROC, might be desirable. In its general nature, this paper focussed on ROC curves and their accompanying AUCs. As GBPs estimate only the graph structure and not underlying distributions, recently proposed evaluations of CSL methods taking in account estimated distributions of the measurements X can not be considered (O’Donnell et al. 2021). However, for particular use-cases evaluation on a more specific metric and/or

forcing the GBPs to predict binary graphs - as PC, IGSP and GIES do, for example via cross-validation - might be insightful. This is in particular true for the OIP as it performed best on the real data in Sect. 4.2.

Regarding the computation of the TAP, it remains to be seen whether for large p one can devise a feasible, consistent simulation procedure, or, if resorting to the B-TAP or a changed q remains necessary. Moreover, it would be of interest to study whether the resulting ROC curves of the TAP, B-TAP and OIP can in general be related as conjectured in Sect. 3.4.

Acknowledgements We would like to thank Bernd Taschler of the University of Oxford for feedback and help in data processing and in testing the R code. Moreover, we thank the anonymous reviewers and the associate editor for their help in improving this contribution. This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) project “MechML”. S. Bhamidi was supported in part by the National Science Foundation (NSF) grants DMS-2113662 and DMS-2134107.

Author contributions All authors contributed to the conceptualization and methodology of the paper. R.R. conducted the formal analysis including the derivation of the proofs, wrote the software and prepared the figures. Proofs and statements of Theorem 3.4, Lemma A.1 and Lemma A.3 were derived equally by R.R. and S.B. R.R. and S.M. wrote the main manuscript text with input from S.B.. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: Proofs

Appendix A.1: Proof of Theorem 3.5

To prove Theorem 3.5 we need the following preliminary result.

Lemma 1 *Let $G = (V, E)$ be such that E is drawn at random with marginal probabilities*

$$E[k, \ell] \sim \begin{cases} \delta(1) & \text{if } k = \ell \\ B(1, q) & \text{else} \end{cases},$$

where $q \in (0, 1)$, with $E[k, \ell]$ and $E[k', \ell']$ drawn independently for all k, k' and all $\ell \neq \ell'$, and with a covariance structure given by

$$\text{Cov} \left(E[k, \ell], E[k', \ell'] \middle| (E[\tilde{k}_j, \ell])_{j=1}^J \right) = \kappa_{N, J} > 0, \quad (\text{A1})$$

with $N := \sum_{j=1}^J E[\tilde{k}_j, \ell]$, for all ℓ and any pairwise distinct $k, k', \tilde{k}_1, \dots, \tilde{k}_J \in \{1, 2, \dots, p\}$.

Fix $\ell \neq \ell'$ and disjoint sets $Q_1, Q_2 \subset \{1, 2, \dots, p\}$ and disjoint sets $\tilde{Q}_1, \tilde{Q}_2 \subset \{1, 2, \dots, p\}$ such that $\ell \notin Q_1 \cup Q_2, \ell' \notin \tilde{Q}_1 \cup \tilde{Q}_2, |Q_1| = |\tilde{Q}_1|$ and $|Q_2| = |\tilde{Q}_2|$. Furthermore, fix $\mathbf{m}, \tilde{\mathbf{m}} \in \{0, 1\}^{|Q_2|}$.

Then, we have

$$\begin{aligned} & \mathbb{P} \left[\sum_{k \in Q_1} E[k, \ell] > \sum_{k \in \tilde{Q}_1} E[k, \ell'] \middle| \right. \\ & \quad \left. (E[k, \ell])_{k \in Q_2} = \mathbf{m}, (E[k, \ell'])_{k \in \tilde{Q}_2} = \tilde{\mathbf{m}} \right] \\ & > \mathbb{P} \left[\sum_{k \in Q_1} E[k, \ell] < \sum_{k \in \tilde{Q}_1} E[k, \ell'] \middle| \right. \\ & \quad \left. (E[k, \ell])_{k \in Q_2} = \mathbf{m}, (E[k, \ell'])_{k \in \tilde{Q}_2} = \tilde{\mathbf{m}} \right], \end{aligned} \quad (\text{A2})$$

if and only if

$$\|\mathbf{m}\|_1 > \|\tilde{\mathbf{m}}\|_1.$$

Proof First note that by construction we have for $\mathbf{x}, \tilde{\mathbf{x}} \in \{0, 1\}^{|Q_1|}$ and $\mathbf{y}, \tilde{\mathbf{y}} \in \{0, 1\}^{|Q_2|}$ that

$$\begin{aligned} & \mathbb{P} \left[(E[k, \ell])_{k \in Q_1} = \mathbf{x} \middle| (E[r, \ell])_{r \in Q_2} = \mathbf{y} \right] \\ & = \mathbb{P} \left[(E[k, \ell'])_{k \in \tilde{Q}_1} = \tilde{\mathbf{x}} \middle| (E[r, \ell'])_{r \in \tilde{Q}_2} = \tilde{\mathbf{y}} \right]. \end{aligned} \quad (\text{A3})$$

Second, let ℓ and $Q \subsetneq \{1, \dots, p\}$ be arbitrary such that $\ell \notin Q$, and $\mathbf{m} = (m_j) \in \{0, 1\}^{|Q|}$ a vector such that $\sum_j m_j \geq 1$. Suppose furthermore without loss of generality that $1 \in Q$ and $m_1 = 1$. We have for $\ell \neq k \notin Q$ that

$$\begin{aligned} & \mathbb{P} [E[k, \ell] = 1 | (E[r, \ell])_{r \in Q} = \mathbf{m}] \\ & = \frac{\mathbb{P} [E[k, \ell] = 1, E[1, \ell] = 1 | (E[r, \ell])_{r \in Q \setminus \{1\}} = \tilde{\mathbf{m}}]}{\mathbb{P} [E[1, \ell] = 1 | (E[r, \ell])_{r \in Q \setminus \{1\}} = \tilde{\mathbf{m}}]}, \end{aligned} \quad (\text{A4})$$

where $\tilde{\mathbf{m}}$ is the vector \mathbf{m} without its first entry. For ease of notation denote

$$Z_0 = \{(E[r, \ell])_{r \in \mathcal{Q} \setminus \{1\}} = \tilde{\mathbf{m}}\}.$$

Then, using (A1) we have

$$\begin{aligned} & \frac{\mathbb{P}[E[k, \ell] = 1, E[1, \ell] = 1 | Z_0]}{\mathbb{P}[E[1, \ell] = 1 | Z_0]} \\ & > \frac{\mathbb{P}[E[k, \ell] = 1 | Z_0] \mathbb{P}[E[1, \ell] = 1 | Z_0]}{\mathbb{P}[E[1, \ell] = 1 | Z_0]} \\ & = \frac{\mathbb{P}[E[k, \ell] = 1 | Z_0] \mathbb{P}[E[1, \ell] = 0 | Z_0]}{\mathbb{P}[E[1, \ell] = 0 | Z_0]} \\ & > \frac{\mathbb{P}[E[k, \ell] = 1, E[1, \ell] = 0 | Z_0]}{\mathbb{P}[E[1, \ell] = 0 | Z_0]} \\ & = \mathbb{P}[E[k, \ell] = 1 | Z_0 \text{ and } E[1, \ell] = 0]. \end{aligned} \tag{A5}$$

By symmetry of the covariance structure yielding (A3) we follow that for any k, k', ℓ, ℓ' s.t. $k \neq \ell$ and $k' \neq \ell'$ and subsets $\mathcal{Q}, \tilde{\mathcal{Q}}$ with $|\mathcal{Q}| = |\tilde{\mathcal{Q}}|$ that

$$\begin{aligned} & \mathbb{P}[E[k, \ell] = 1 | (E[r, \ell])_{r \in \mathcal{Q}} = \mathbf{m}] \\ & > \mathbb{P}[E[k', \ell'] = 1 | (E[r, \ell'])_{r \in \tilde{\mathcal{Q}}} = \tilde{\mathbf{m}}], \end{aligned} \tag{A6}$$

if and only if

$$\|\mathbf{m}\|_1 > \|\tilde{\mathbf{m}}\|_1. \tag{A7}$$

In particular, by (A1) we have that if equality holds in (A6), equality holds in (A7), yielding the “if and only if” part of the statement.

Moreover, we obtain analogously the opposite statement that for any k, k', ℓ, ℓ' s.t. $k \neq \ell$ and $k' \neq \ell'$ and subsets $\mathcal{Q}, \tilde{\mathcal{Q}}$ with $|\mathcal{Q}| = |\tilde{\mathcal{Q}}|$ we have

$$\begin{aligned} & \mathbb{P}[E[k, \ell] = 0 | (E[r, \ell])_{r \in \mathcal{Q}} = \mathbf{m}] \\ & > \mathbb{P}[E[k', \ell'] = 0 | (E[r, \ell'])_{r \in \tilde{\mathcal{Q}}} = \tilde{\mathbf{m}}], \end{aligned} \tag{A8}$$

if and only if

$$\|\mathbf{m}\|_1 < \|\tilde{\mathbf{m}}\|_1.$$

Again, with equality holding either in both, or in none of the equations by virtue of (A3).

Let us in the following assume without loss of generality by symmetry of the covariance structure and independence between columns that $\mathcal{Q}_x = \tilde{\mathcal{Q}}_x$ for $x = 1, 2$.

The claim of Lemma 1 can now be proven via induction on the size of the set \mathcal{Q}_1 , while keeping \mathcal{Q}_2 fixed. To this end, $\ell, \ell', \mathbf{m}, \tilde{\mathbf{m}}, \mathcal{Q}_1$ and \mathcal{Q}_2 be as in the assumption (including $\ell' \notin \mathcal{Q}_1, \mathcal{Q}_2$). We initialize the induction hypothesis with $|\mathcal{Q}_1| = |\{k\}| = 1$. Define

$$\begin{aligned} A & := \mathbb{P}\left[E[k, \ell] > E[k, \ell'] \mid \right. \\ & \quad \left. (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}, (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\right] \\ & = \mathbb{P}\left[E[k, \ell] = 1 \mid (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}\right] \\ & \quad \times \mathbb{P}\left[E[k, \ell'] = 0 \mid (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\right]. \end{aligned}$$

In the case $\|\mathbf{m}\|_1 > \|\tilde{\mathbf{m}}\|_1$ we have by (A6) and (A8) that

$$\begin{aligned} A & > \mathbb{P}\left[E[k, \ell'] = 1 \mid (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\right] \\ & \quad \times \mathbb{P}\left[E[k, \ell] = 0 \mid (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}\right] \\ & = \mathbb{P}\left[E[k, \ell] < E[k, \ell'] \mid \right. \\ & \quad \left. (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}, (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\right]. \end{aligned} \tag{A9}$$

In the same way we have that if $\|\mathbf{m}\|_1 = \|\tilde{\mathbf{m}}\|_1$ we obtain equality in (A9) by using that equality in (A7) yields equality in (A6), yielding the base case of the induction.

It remains to show the induction step. Let the claim be shown for $|\mathcal{Q}_1| = N \in \mathbb{N}$ and consider now $|\mathcal{Q}_1| = |\{k_1, \dots, k_{N+1}\}| = N + 1$ and $\|\mathbf{m}\|_1 > \|\tilde{\mathbf{m}}\|_1$. For ease of notation let us define for $x, y = 0, 1$ the events

$$\begin{aligned} Z & := \{(E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}, (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\}, \\ \tilde{Z}_{(x,y)} & := \{E[k_{N+1}, \ell] = x, E[k_{N+1}, \ell'] = y\}, \\ Z_{(x,y)} & := Z \cap \tilde{Z}_{(x,y)}. \end{aligned}$$

We have

$$\begin{aligned} B & := \mathbb{P}\left[\sum_{n=1}^{N+1} E[k_n, \ell] > \sum_{n=1}^{N+1} E[k_n, \ell'] \mid \right. \\ & \quad \left. (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}, (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}}\right] \\ & = \mathbb{P}\left[\underbrace{\sum_{n=1}^N E[k_n, \ell] > \sum_{n=1}^N E[k_n, \ell']}_{:=P_{(0,0)}} \mid Z_{(0,0)}\right] \mathbb{P}\left[\tilde{Z}_{(0,0)} \mid Z\right] \\ & \quad + \mathbb{P}\left[\underbrace{\sum_{n=1}^N E[k_n, \ell] > \sum_{n=1}^N E[k_n, \ell']}_{:=P_{(1,1)}} \mid Z_{(1,1)}\right] \mathbb{P}\left[\tilde{Z}_{(1,1)} \mid Z\right] \\ & \quad + \mathbb{P}\left[\underbrace{\sum_{n=1}^N E[k_n, \ell] \geq \sum_{n=1}^N E[k_n, \ell']}_{:=P_{(1,0)}} \mid Z_{(1,0)}\right] \mathbb{P}\left[\tilde{Z}_{(1,0)} \mid Z\right] \\ & \quad + \mathbb{P}\left[\underbrace{\sum_{n=1}^N E[k_n, \ell] > 1 + \sum_{n=1}^N E[k_n, \ell']}_{:=P_{(0,1)}} \mid Z_{(0,1)}\right] \mathbb{P}\left[\tilde{Z}_{(0,1)} \mid Z\right] \end{aligned} \tag{A10}$$

Using the induction assumption we have immediately

$$\begin{aligned}
 P_{(0,0)} &> \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,0)} \right], \\
 P_{(1,1)} &> \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,1)} \right].
 \end{aligned}
 \tag{A11}$$

Moreover we obtain by induction assumption as $\|\mathbf{m}\|_1 + 1 > \|\tilde{\mathbf{m}}\|_1$ that

$$\begin{aligned}
 P_{(1,0)} &> \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] \leq \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right] \\
 &= \mathbb{P} \left[1 + \sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right] \\
 &\quad + \underbrace{\mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right]}_{:=\gamma_1} \\
 &\quad + \underbrace{\mathbb{P} \left[1 + \sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right]}_{:=\gamma_2},
 \end{aligned}
 \tag{A12}$$

and by $\|\mathbf{m}\|_1 \geq \|\tilde{\mathbf{m}}\|_1 + 1$ we have,

$$\begin{aligned}
 P_{(0,1)} &= \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] > \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right] \\
 &\quad - \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = 1 + \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right] \\
 &\geq \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] \leq \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right] \\
 &\quad - \underbrace{\mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right]}_{:=\tilde{\gamma}_1} \\
 &\quad - \underbrace{\mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = 1 + \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right]}_{:=\tilde{\gamma}_2},
 \end{aligned}
 \tag{A13}$$

Now we need to show that

$$\mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| Z \right] (\gamma_1 + \gamma_2) - \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| Z \right] (\tilde{\gamma}_1 + \tilde{\gamma}_2) \geq 0,$$

in order that we can drop it. We start by using the induction assumption to show:

$$\begin{aligned}
 &\mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| Z \right] \gamma_1 \\
 &= \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'], \tilde{Z}_{(1,0)} \middle| Z \right] \\
 &= \mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| \sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'], Z \right] \\
 &\quad \times \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'] \middle| Z \right] \\
 &> \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| \sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'], Z \right] \\
 &\quad \times \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'] \middle| Z \right] \\
 &= \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| Z \right] \tilde{\gamma}_1.
 \end{aligned}
 \tag{A14}$$

For the second term we use the symmetry of the covariance structure, yielding (A3), and the independence between the columns of the edge matrix:

$$\begin{aligned}
 &\mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| Z \right] \gamma_2 \\
 &= \mathbb{P} \left[1 + \sum_{n=1}^N E[k_n, \ell] = \sum_{n=1}^N E[k_n, \ell'], \tilde{Z}_{(1,0)} \middle| Z \right] \\
 &= \sum_{t=0}^{N-1} \left(\mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = t, E[k_{N+1}, \ell] = 1 \middle| Z \right] \right. \\
 &\quad \left. \times \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell'] = t + 1, E[k_{N+1}, \ell'] = 0 \middle| Z \right] \right) \\
 &= \sum_{t=0}^{N-1} \left(\mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell'] = t, E[k_{N+1}, \ell'] = 1 \middle| Z \right] \right. \\
 &\quad \left. \times \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] = t + 1, E[k_{N+1}, \ell] = 0 \middle| Z \right] \right) \\
 &= \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| Z \right] \tilde{\gamma}_2
 \end{aligned}
 \tag{A15}$$

Using (A14) and (A15) and plugging (A11), (A12) and (A13) in (A10) we get

$$\begin{aligned}
 B &> \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,0)} \right] \mathbb{P} \left[\tilde{Z}_{(0,0)} \middle| Z \right] \\
 &+ \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,1)} \right] \mathbb{P} \left[\tilde{Z}_{(1,1)} \middle| Z \right] \\
 &+ \mathbb{P} \left[1 + \sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right] \mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| Z \right] \\
 &+ \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] \leq \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right] \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| Z \right] \\
 &= \mathbb{P} \left[\sum_{n=1}^{N+1} E[k_n, \ell] < \sum_{n=1}^{N+1} E[k_n, \ell'] \middle| \right. \\
 &\quad \left. (E[r, \ell])_{r \in \mathcal{Q}_2} = \mathbf{m}, (E[r, \ell'])_{r \in \mathcal{Q}_2} = \tilde{\mathbf{m}} \right], \tag{A16}
 \end{aligned}$$

showing the induction step for $\|\mathbf{m}\|_1 > \|\tilde{\mathbf{m}}\|_1$. Which leaves the case $\|\mathbf{m}\|_1 = \|\tilde{\mathbf{m}}\|_1$. First, by induction assumption we have equality in both displays of (A11). Second, by (A3) and symmetry of construction we have that

$$\begin{aligned}
 \mathbb{P} \left[\tilde{Z}_{(1,0)} \middle| Z \right] &= \mathbb{P} \left[\tilde{Z}_{(0,1)} \middle| Z \right], \\
 P_{(1,0)} &= \mathbb{P} \left[\sum_{n=1}^N E[k_n, \ell] \leq \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(0,1)} \right], \\
 P_{(0,1)} &= \mathbb{P} \left[1 + \sum_{n=1}^N E[k_n, \ell] < \sum_{n=1}^N E[k_n, \ell'] \middle| Z_{(1,0)} \right].
 \end{aligned}$$

Hence, putting both observations together we have equality in (A16) for $\|\mathbf{m}\|_1 = \|\tilde{\mathbf{m}}\|_1$, finishing the proof.

Proof of Theorem 3.5 Let us start by stating the expected value for the AUC_{TC} derived from Remark 2.5.

$$\begin{aligned}
 \mathbb{E}_{E_X|E_Y=M_Y} [AUC_{TC}(\Theta_{OIP}(E_X))] &= \frac{1}{|\tilde{E}_{Y,1}||\tilde{E}_{Y,0}|} \\
 &\times \sum_{[k,\ell] \in \tilde{E}_{Y,1}} \sum_{[k',\ell'] \in \tilde{E}_{Y,0}} \mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k, \ell], [k', \ell'])], \tag{A17}
 \end{aligned}$$

where

$$\begin{aligned}
 \sigma_X([k, \ell], [k', \ell']) &= \delta_{\Theta_{OIP}[k,\ell] > \Theta_{OIP}[k',\ell']} + \frac{1}{2} \delta_{\Theta_{OIP}[k,\ell] = \Theta_{OIP}[k',\ell']} \tag{A18}
 \end{aligned}$$

and $\tilde{E}_{Y,x} = \{[k, \ell] \in E_{Y,x} : \ell \notin \mathcal{I}\}$ for $x = 0, 1$.

Case 1: $\ell = \ell'$:

We have by construction $\Theta_{OIP}(E_X)[k, \ell] = \Theta_{OIP}(E_X)[k', \ell']$ for all $k \neq k'$, hence

$$\mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k, \ell], [k', \ell'])] = \frac{1}{2}, \tag{A19}$$

for all $k \neq k'$ such that $E[k, \ell] \neq E[k', \ell]$.

Case 2: $\ell \neq \ell'$:

By assumption we have $\ell, \ell' \notin \mathcal{I}$. Let us first consider the case

$$\sum_{\substack{r \notin \mathcal{I} \\ r \neq \ell}} M_Y[r, \ell] > \sum_{\substack{r \notin \mathcal{I} \\ r \neq \ell'}} M_Y[r, \ell'], \tag{A20}$$

then we have for any $k_0, k'_0 \notin \mathcal{I}$ such that $E[k_0, \ell] = 1 \neq 0 = E[k'_0, \ell']$ that

$$\begin{aligned}
 \mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k_0, \ell], [k'_0, \ell'])] &= \mathbb{E}_{E_X|E_Y=M_Y} [\delta_{\deg_X^-(\ell) > \deg_X^-(\ell')}] \\
 &+ \frac{1}{2} \mathbb{E}_{E_X|E_Y=M_Y} [\delta_{\deg_X^-(\ell) = \deg_X^-(\ell')}] \\
 &= \mathbb{P} [\deg_X^-(\ell) > \deg_X^-(\ell') | E_Y = M_Y] \\
 &+ \frac{1}{2} \mathbb{P} [\deg_X^-(\ell) = \deg_X^-(\ell') | E_Y = M_Y] \\
 &= \mathbb{P} \left[\sum_{r \in \mathcal{I}} E[r, \ell] > \sum_{r \in \mathcal{I}} E[r, \ell'] \middle| E_Y = M_Y \right] \\
 &+ \frac{1}{2} \mathbb{P} \left[\sum_{r \in \mathcal{I}} E[r, \ell] = \sum_{r \in \mathcal{I}} E[r, \ell'] \middle| E_Y = M_Y \right] \\
 &> \mathbb{P} \left[\sum_{r \in \mathcal{I}} E[r, \ell] < \sum_{r \in \mathcal{I}} E[r, \ell'] \middle| E_Y = M_Y \right] \\
 &+ \frac{1}{2} \mathbb{P} \left[\sum_{r \in \mathcal{I}} E[r, \ell] = \sum_{r \in \mathcal{I}} E[r, \ell'] \middle| E_Y = M_Y \right], \tag{A21}
 \end{aligned}$$

where the last inequality holds true by Lemma 1. Since the sum of the last two lines in (A21) is 1 by construction, we have

$$\mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k_0, \ell], [k'_0, \ell'])] > \frac{1}{2}. \tag{A22}$$

Moreover we have by the last line of (A21) that for $k_1, k'_1 \notin \mathcal{I}$ with $E[k_1, \ell] = 0 \neq 1 = E[k'_1, \ell']$ the following holds:

$$\begin{aligned}
 \mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k_1, \ell], [k'_1, \ell'])] &= 1 - \mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k_0, \ell], [k'_0, \ell'])], \tag{A23}
 \end{aligned}$$

recall to this end also Eq. (A3) of Lemma 1.

By (A20) there exist more pairs (k_0, k'_0) yielding (A22) than pairs (k_1, k'_1) yielding (A23), hence,

$$\begin{aligned} & \sum_{\substack{[k,s] \in E_{Y,1} \\ s \in \{\ell, \ell'\}}} \sum_{\substack{[k,s] \in E_{Y,0} \\ s \in \{\ell, \ell'\}}} \mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k, \ell], [k', \ell])] \\ & > \sum_{\substack{[k,s] \in E_{Y,1} \\ s \in \{\ell, \ell'\}}} \sum_{\substack{[k,s] \in E_{Y,0} \\ s \in \{\ell, \ell'\}}} \frac{1}{2}. \end{aligned} \tag{A24}$$

Note, that in the case the inequality in (A20) is inverted Eq. (A24) still holds.

In the case that

$$\sum_{\substack{r \notin \mathcal{I} \\ r \neq \ell}} M_Y[r, \ell] = \sum_{\substack{r \notin \mathcal{I} \\ r \neq \ell'}} M_Y[r, \ell'], \tag{A25}$$

We have by Lemma 1 equality in (A21) and thus

$$\mathbb{E}_{E_X|E_Y=M_Y} [\sigma_X([k, \ell], [k', \ell])] = \frac{1}{2}. \tag{A26}$$

Plugging Eqs. (A19), (A24) and (A26) into (A17) and since by assumption there exists at least one pair $\ell, \ell' \notin \mathcal{I}$ such that (A20) holds, we obtain

$$\mathbb{E}_{E_X|E_Y=M_Y} [AUC_{\mathcal{I}^C}(\Theta_{\text{OIP}}(E_X))] > \frac{1}{2}.$$

Last, let us give an example of a graph generation process that falls under Theorem 3.5.

Definition 2 (compare e.g. Hoff et al. 2002; Bollobás et al. 2007) We define a directed latent network model with fixed outgoing and node depending incoming propensities (LNM-fix-O) by drawing

$$(z_\ell)_{\ell=1}^p \stackrel{iid}{\sim} \mathcal{D}((0, 1)), \tag{A27}$$

from some non-degenerate distribution \mathcal{D} on $(0, 1)$ (i.e. \mathcal{D} is not a Dirac delta distribution). Subsequently, draw $G = (V, E)$ by iid draws

$$E[k, \ell] \sim \begin{cases} \delta(1) & \text{if } k = \ell \\ B(1, z_\ell) & \text{else} \end{cases}.$$

Lemma 3 The probability distribution of the LNM-fix-O given in Definition 2 satisfies the assumptions of Theorem 3.5 and Lemma 1.

Proof First, by construction the marginal distributions are Bernoulli random draws, moreover, we have

$$E[k, \ell] \sim B(1, \mathbb{E}_{\mathcal{D}}[z_\ell]),$$

for all $1 \leq \ell \leq p$. Hence, by (A27) we have, setting $q := \mathbb{E}_{\mathcal{D}}[z_\ell]$, that all marginal distributions feature the same success probability.

Second, by construction $E[k, \ell]$ and $E[k', \ell']$ are independent for all $\ell \neq \ell'$.

Third, we are left to show (3.4). Let $1 \leq k, k', k_1, \dots, k_J \leq p$ be pairwise different and $\mathbf{m} \in \{0, 1\}^J$ with $\|\mathbf{m}\|_1 = N$, we have

$$\begin{aligned} & cov\left(E[k, \ell], E[k', \ell] \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right) \\ & = \mathbb{E}\left[E[k, \ell]E[k', \ell] \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \\ & \quad - \left(\mathbb{E}\left[E[k, \ell] \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \right. \\ & \quad \left. \times \mathbb{E}\left[E[k', \ell] \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right]\right) \\ & = \mathbb{P}\left[E[k, \ell] = 1, E[k', \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \\ & \quad - \left(\mathbb{P}\left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \right. \\ & \quad \left. \times \mathbb{P}\left[E[k', \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right]\right) \\ & = \left(\mathbb{P}\left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}, E[k', \ell] = 1\right] \right. \\ & \quad \left. - \mathbb{P}\left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right]\right) \\ & \quad \times \mathbb{P}\left[E[k', \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \end{aligned} \tag{A28}$$

Let us assume for now that \mathcal{D} is a discrete random variable. In this case consider

$$\begin{aligned} & \mathbb{P}\left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] \\ & = \int_0^1 \left(\mathbb{P}\left[E[k, \ell] = 1 \mid z_\ell = z\right] \right. \\ & \quad \left. \times \mathbb{P}\left[z_\ell = z \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right]\right) d\mu_{\mathcal{D}}(z) \tag{A29} \\ & = \int_0^1 z \mathbb{P}\left[z_\ell = z \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right] d\mu_{\mathcal{D}}(z) \\ & = \mathbb{E}\left[z_\ell \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m}\right]. \end{aligned}$$

Second by Bayes theorem we have

$$\begin{aligned} & \mathbb{P} \left[z_\ell = z \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right] \\ &= \frac{\mathbb{P} \left[(E[k_j, \ell])_{j=1}^J = \mathbf{m} \mid z_\ell = z \right] \mathbb{P} [z_\ell = z]}{\mathbb{P} \left[(E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]} \tag{A30} \\ &= \frac{z^N (1-z)^{J-N} \mathbb{P} [z_\ell = z]}{\int_0^1 \tilde{z}^N (1-\tilde{z})^{J-N} d\mu_{\mathcal{D}}(\tilde{z})}, \end{aligned}$$

we have that (A29) only depends on J and N and not on the exact configuration of \mathbf{m} . Hence the covariance of (A28) depends only on J and N and by construction not on k, k', ℓ and the exact configuration of \mathbf{m} . It remains to show that (A28) is strictly greater than 0. To this end, consider $\tilde{\mathbf{m}} = (\mathbf{m}, 1) \in \{0, 1\}^{J+1}$. We have by (A29), (A30) and renaming $k' = k_{J+1}$,

$$\begin{aligned} & \mathbb{P} \left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \right] \\ &= \mathbb{E} \left[z_\ell \mid (E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \right] \\ &= \frac{\int_0^1 z \mathbb{P} \left[(E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \mid z_\ell = z \right] \mathbb{P} [z_\ell = z] d\mu_{\mathcal{D}}(z)}{\mathbb{P} \left[(E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \right]} \\ &= \frac{\int_0^1 z \mathbb{P} \left[(E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \mid z_\ell = z \right] \mathbb{P} [z_\ell = z] d\mu_{\mathcal{D}}(z)}{\mathbb{P} \left[E[k', \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right] \mathbb{P} \left[(E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]} \\ &= \frac{\int_0^1 z \mathbb{P} \left[(E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \mid z_\ell = z \right] \mathbb{P} [z_\ell = z] d\mu_{\mathcal{D}}(z)}{\mathbb{E} \left[z_\ell \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right] \mathbb{P} \left[(E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]} \\ &= \frac{\mathbb{E} \left[z_\ell^2 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]}{\mathbb{E} \left[z_\ell \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]}, \tag{A31} \end{aligned}$$

where the last equality is by virtue of

$$\begin{aligned} & \mathbb{P} \left[(E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \mid z_\ell = z \right] = z^{N+1} (1-z)^{N-J} \\ &= z \mathbb{P} \left[(E[k_j, \ell])_{j=1}^J = \mathbf{m} \mid z_\ell = z \right]. \end{aligned}$$

By the non-degenerate nature of \mathcal{D} we can use the strict form of Jensen’s inequality to conclude from (A31) that

$$\begin{aligned} & \mathbb{P} \left[E[k, \ell] = 1 \mid (E[k_j, \ell])_{j=1}^{J+1} = \tilde{\mathbf{m}} \right] \\ &> \mathbb{E} \left[z_\ell \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right] \tag{A32} \\ &= \mathbb{P} \left[E[k', \ell] = 1 \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right]. \end{aligned}$$

By plugging (A32) into (A28) we obtain

$$\text{cov} \left(E[k, \ell], E[k', \ell] \mid (E[k_j, \ell])_{j=1}^J = \mathbf{m} \right) > 0, \tag{A33}$$

finishing the proof for discrete distributions \mathcal{D} . For general distributions the proof follows analogously using standard

conditional expectation operations. For space considerations we omit the details.

Appendix A.2: Proof of Lemma 3.13

Proof Let $[k, \ell] \in S_2$ and $\mathcal{A}_{v_k} \not\subseteq \mathcal{A}_{v_\ell}$, then there exists a parent $v_r \in \mathcal{A}_{v_k}$ such that $E[r, \ell] = 0$ and $E[r, k] = 1$. Let \tilde{E} be an edge matrix whose transitive closure is consistent with E_X , then there exists a causal path from v_r to v_k . Assume moreover $\tilde{E}^+[k, \ell] = 1$ then there would exist a causal path from v_r to v_ℓ violating the constraint $\tilde{E}^+[r, \ell] = E[r, \ell] = 0$. Hence, the nominator of (3.10) is equal to 0.

Vice versa, let $[k, \ell] \in S_2$ be such that $\mathcal{A}_{v_k} \subseteq \mathcal{A}_{v_\ell}$. Then, define

$$\tilde{E}[r, s] = \begin{cases} E[r, s] & \text{if } [r, s] \in S_1 \\ 1 & \text{if } [r, s] = [k, \ell]. \\ 0 & \text{else} \end{cases}$$

It is straight-forward to check that $\tilde{E} \in \mathcal{X}$ and features $\tilde{E}^+[k, \ell] = 1$. Hence $\Theta_{\text{TAP}}(E_X)[k, \ell] > 0$.

Appendix A.3: Proof of Lemma 3.15

Proof Assume that E is not symmetric: Let $v_k, v_\ell \in V$ such that $E[k, \ell] = 1$ and $E[\ell, k] = 0$. Since G is transitively closed we have $\mathcal{P}_k \subset \mathcal{P}_\ell \setminus \{v_k\}$, where \mathcal{P}_k is the set of parents of the node v_k . Yielding $\text{deg}^-(v_k) \leq \text{deg}^-(v_\ell) - 1$, a contradiction to the assumption $\text{deg}^-(v_k) = \text{deg}^-(v_\ell)$. Hence if Assumption 3.3 holds, E is symmetric. By G being transitively closed we have $\mathcal{A}_k = \mathcal{P}_k = \mathcal{D}_k$, where \mathcal{A}_k and \mathcal{D}_k denote the respective sets of ancestors and descendants, for every $v_k \in V$. Hence, for any $v_k \in V$ we have that $\{v_k\} \cup \mathcal{P}_k$ form a connected component and a complete subgraph, yielding the claim by the fact that deg^- is constant on V .

Appendix A.4: Proof of Corollary 3.18

Proof By definition of $\theta_{\text{TAP}}(E_X)$ we have that from $\Theta_{\text{TAP}}(E_X)[k, \ell] = 0$ follows $\theta_{\text{TAP}}(E_X)[k, \ell] = 0$, yielding one direction. For the other direction consider \tilde{E} from the proof of Lemma 3.13 to yield $\theta_{\text{TAP}}(E_X)[k, \ell] > 0$ if $\mathcal{A}_{v_k} \subseteq \mathcal{A}_{v_\ell}$.

Appendix A.5: Proof of Lemma 3.19

Proof Let $[k, \ell] \in S_2$ be an edge with $\mathcal{A}_{v_k} \subseteq \mathcal{A}_{v_\ell}$. Define

$$\begin{aligned} \Phi_1 &:= \{E_0 \in \mathcal{X}(V, E_X) : E_0[k, \ell] = 1\}, \\ \Phi_2 &:= \{E_0 \in \mathcal{X}(V, E_X) : E_0[k, \ell] = 0\}, \end{aligned}$$

Let $M \in \Phi_2$, then we define by $M_{+[k,\ell]}$ the edge matrix with

$$M_{+[k,\ell]}[r, s] := \begin{cases} 1 & \text{if } [r, s] = [k, \ell] \\ M[r, s] & \text{else} \end{cases}, \quad (\text{A34})$$

i.e. adding the edge $[k, \ell]$ to M . Note that by definition of Φ_2 we have $M[k, \ell] = 0$. Define furthermore a mapping

$$\phi : \Phi_2 \rightarrow \Phi_1, \quad \phi : M \mapsto M_{+[k,\ell]}.$$

Since $M \in \mathcal{X}(V, E_X)$ and $\mathcal{A}_{v_k} \subseteq \mathcal{A}_{v_\ell}$ adding the edge $[k, \ell]$ will not interfere with the zero-constraints given by (3.11) and as adding edges can never interfere with the $\neq 0$ constraints in (3.12), we have $M_{+[k,\ell]} \in \mathcal{X}(V, E_X)$ and thus ϕ is well-defined. Moreover, ϕ is by definition injective.

Given $(\mathcal{A}_{v_k} \setminus \{v_\ell\}) \cap \mathcal{A}_{v_\ell} = \emptyset$ we define additionally

$$\psi : \Phi_1 \rightarrow \Phi_2, \quad \psi : M \mapsto M_{-[k,\ell]},$$

with $M_{-[k,\ell]}$ being defined as in A34 with $M[k, \ell]$ set to 0, i.e. deleting the edge $[k, \ell]$ from M . If deleting the edge $[k, \ell]$ would break a path from v_r with $r \in \mathcal{I}$ to $v_s \in V$ we would have $v_k, v_\ell \in \mathcal{D}_{v_r}$ since $k \notin \mathcal{I}$, a contradiction to $(\mathcal{A}_{v_k} \setminus \{v_\ell\}) \cap \mathcal{A}_{v_\ell} = \emptyset$. Hence, $\psi(E_X)$ is well defined, since deleting an edge can not interfere with the zero-constraints given by E_X . By definition ψ is the inverse function of ϕ , making ϕ a bijection. Hence, we have

$$\theta_{\text{TAP}(E_X)}[k, \ell] = \frac{1}{2}.$$

Vice versa, if there were to exist $v_r \in (\mathcal{A}_{v_k} \setminus \{v_\ell\}) \cap \mathcal{A}_{v_\ell}$ define M_0 by

$$M_0[x, y] = \begin{cases} 1 & \text{if } x \in \mathcal{I}, v_y \in \mathcal{A}_{v_x} \text{ and } [x, y] \neq [r, \ell] \\ 1 & \text{if } [x, y] = [k, \ell] \\ 0 & \text{else} \end{cases}.$$

Then by definition $M_0 \in \Phi_1$, but $(M_0)_{-(v,w)} \notin \mathcal{X}(V, E_X)$, hence M_0 is not in the image of ϕ , yielding by injectivity of ϕ ,

$$\#\Phi_1 > \#\Phi_2,$$

i.e. $\theta_{\text{TAP}}[k, \ell](E_X) > \frac{1}{2}$.

Appendix A.6: Computation of the example

Consider G and E_X of the Example in Sect. 3.2 detailed in Fig. 2a. We have by Lemma 3.19 that

$$\begin{aligned} \theta_{\text{TAP}(E_X)}[2, 1] &= \theta_{\text{TAP}(E_X)}[3, 1] \\ &= \theta_{\text{TAP}(E_X)}[4, 1] \\ &= \theta_{\text{TAP}(E_X)}[4, 2] \\ &= \theta_{\text{TAP}(E_X)}[4, 3] = \frac{1}{2}. \end{aligned}$$

Moreover, we can compute by Fig. 2b that $\theta_{\text{TAP}(E_X)}[2, 3] = \theta_{\text{TAP}(E_X)}[3, 2] = 5/8$ and $\theta_{\text{TAP}(E_X)}[1, 2] = \theta_{\text{TAP}(E_X)}[1, 3] = 6/8$. Last, by Corollary 3.18 we have that the edge matrix entries $[1, 4], [2, 4]$ and $[3, 4]$ are zero, yielding

$$(\theta_{\text{TAP}(E_X)}[k, \ell])_{k,\ell=1}^4 = \frac{1}{8} \begin{pmatrix} 8 & 6 & 6 & 0 \\ 4 & 8 & 5 & 0 \\ 4 & 5 & 8 & 0 \\ 4 & 4 & 4 & 8 \end{pmatrix}.$$

As given in Fig. 2b the edge matrix entries $\{[1, 2], [1, 3], [2, 3], [3, 2]\}$ of an edge matrix $E_0 \in \mathcal{X}(V, E_X)$ have 8 different possible configurations. Additionally, the five edge matrix entries $[2, 1], [3, 1], [4, 1], [4, 2], [4, 3]$ can be in $2^5 = 32$ different configurations. Hence, we have $\#\mathcal{X}(V, E_X) = 32 \cdot 8 = 256$. In the following we collect for each edge all graphs $E \in \mathcal{X}$ featuring the given edge in their respective transitive closure.

- $\Theta_{\text{TAP}(E_X)}[2, 1]$:
Count all the edge matrices in $\mathcal{X}(V, E_X)$ featuring $E^+[2, 1] = 1$:
 - All 128 edge matrices with $E[2, 1] = 1$.
 - Among the 128 edge matrices with $E[2, 1] = 0$, there is none with $E[2, 4] = 1$, and there are $5/8 \cdot 128 = 80$ with $E[2, 3] = 1$. Among those 80 we have 40 with $E[3, 1] = 1$.
 - Among the remaining 40 with $E[2, 3] = 1$ and $E[2, 1] = E[3, 1] = 0$ we have none with $E[3, 4] = 1$.

Hence, we obtain

$$\Theta_{\text{TAP}(E_X)}[2, 1] = \frac{128 + 40}{256} = \frac{168}{256} = \frac{21}{32}.$$

- $\Theta_{\text{TAP}(E_X)}[4, 1]$:
Count all the edge matrices in $\mathcal{X}(V, E_X)$ featuring $E^+[4, 1] = 1$:
 - All 128 edge matrices with $E[4, 1] = 1$.
 - Among the 128 edge matrices with $E[4, 1] = 0$ we have 64 graphs with $E[4, 2] = 1$ from which...

- ... 32 feature $E[2, 1] = 1$, hence $E^+[4, 1] = 1$.
- ... 32 feature $E[2, 1] = 0$. Of those there are
 - $5/8 \cdot 32 = 20$ which feature $E[2, 3] = 1$. And half of those 20 feature $E[3, 1] = 1$. Hence, there are 10 additional edge matrices with $E^+[4, 1] = 1$.
 - Of the remaining 12 which feature $E[2, 3] = 0$ we have 3 featuring $E[4, 3] = E[3, 1] = 1$.
- Among the 64 with $E[4, 1] = E[4, 2] = 0$ there are 32 with $E[4, 3] = 1$ from which...
 - ... 16 feature $E[3, 1] = 1$, hence $E^+[4, 1] = 1$.
 - ... 16 feature $E[3, 1] = 0$. Of those there are $5/8 \times 16 = 10$ which feature $E[3, 2] = 1$. And half of those 10 feature $E[2, 1] = 1$. Yielding 5 additional edge matrices with $E^+[4, 1] = 1$.

Hence, we obtain

$$\begin{aligned} \Theta_{\text{TAP}}(E_X)[4, 1] &= \frac{128 + 32 + 10 + 3 + 16 + 5}{256} \\ &= \frac{194}{256}. \end{aligned}$$

• $\Theta_{\text{TAP}}(\mathbf{E}_X)[4, 2]$:

Count all the edge matrices in $\mathcal{X}(V, E_X)$ featuring $E^+[4, 2] = 1$:

- All 128 edge matrices with $E[4, 2] = 1$.
- Among the 128 graphs with $E[4, 2] = 0$ we have 64 graphs with $E[4, 1] = 1$, yielding $E^+[4, 2] = 1$ from the constraint $E^+[1, 2] = 1$.
- Among the 64 with $E[4, 1] = E[4, 2] = 0$ there are 32 with $E[4, 3] = 1$ from which...
 - ... 16 feature $E[3, 1] = 1$, hence $E^+[4, 2] = 1$ by the constraint $E^+[1, 2] = 1$.
 - ... 16 feature $E[3, 1] = 0$. Of those there are $5/8 \times 16 = 10$ which feature $E[3, 2] = 1$, hence $E^+[4, 2] = 1$.

Hence, we obtain

$$\Theta_{\text{TAP}}(E_X)[4, 2] = \frac{128 + 64 + 16 + 10}{256} = \frac{218}{256}.$$

• $\Theta_{\text{TAP}}(\mathbf{E}_X)[2, 3]$:

Count all the edge matrices in $\mathcal{X}(V, E_X)$ featuring $E^+[2, 3] = 1$:

- All $5/8 \cdot 256 = 160$ edge matrices with $E[2, 3] = 1$.
- Among the 96 graphs with $E[2, 3] = 0$ we have no graph with $E[2, 4] = 1$ and 48 edge matrices with $E[2, 1] = 1$, which yield by constraints $E^+[2, 3] = 1$.

Hence, we obtain

$$\Theta_{\text{TAP}}(E_X)[2, 3] = \frac{160 + 48}{256} = \frac{208}{256} = \frac{13}{16}.$$

By symmetry we obtain

$$(\Theta_{\text{TAP}}(E_X)[k, \ell])_{k, \ell=1}^4 = \frac{1}{256} \begin{pmatrix} 256 & 256 & 256 & 0 \\ 168 & 256 & 208 & 0 \\ 168 & 208 & 256 & 0 \\ 194 & 218 & 218 & 256 \end{pmatrix}.$$

To compute $\Theta_{\text{B-TAP}}^{(0.5)}(E_X)$ consider the three spanning trees ensuring edges in $[1, 2]$ and $[1, 3]$ given in Fig. 2c. Each exists with probability $1/3$, all other edges are then added on with a probability of $1/2$. Hence, we have that $\theta_{\text{B-TAP}}(E_X)[k, \ell]$ denoting the marginal distribution of $E_0[k, \ell] = 1$ is given by

$$(\theta_{\text{B-TAP}}(E_X)[k, \ell])_{k, \ell=1}^4 = \frac{1}{6} \begin{pmatrix} 6 & 5 & 5 & 0 \\ 3 & 6 & 4 & 0 \\ 3 & 4 & 6 & 0 \\ 3 & 3 & 3 & 6 \end{pmatrix}$$

Let us now count how often we draw a graph featuring $E^+[k, \ell] = 1$ for Algorithm 2.

• $\Theta_{\text{B-TAP}}(\mathbf{E}_X)[2, 1]$:

When drawing E there is a $1/2$ probability that $E[2, 1] = 1$. Given $E[2, 1] = 0$ we have...

- ... with probability $1/3$ that the spanning tree \mathcal{T} features $[2, 3] \in \mathcal{T}$. Then, with probability $1/2$ we have $E[3, 1] = 1$ (yielding an additional $1/12$). If, however $E[3, 1] = 0$, by $E[2, 4] = E[3, 4] = 0$ we have $E^+[2, 1] = 0$.
- ... with probability $2/3$ that the spanning tree \mathcal{T} features $[2, 3] \notin \mathcal{T}$. Then, with probability $1/4$ we have $E[2, 3] = E[3, 1] = 1$ (yielding an additional $1/12$). If, however either $E[3, 1] = 0$ or $E[2, 3] = 0$, by $E[2, 4] = E[3, 4] = 0$ we have $E^+[2, 1] = 0$.

Yielding in total a probability of

$$\Theta_{\text{B-TAP}}(E_X)[2, 1] = \frac{24 + 4 + 4}{48} = \frac{32}{48}.$$

• $\Theta_{\text{B-TAP}}(\mathbf{E}_X)[4, 1]$:

When drawing E there is a $1/2$ probability that $E[4, 1] = 1$. Given $E[4, 1] = 0$ we have...

- ... with probability $1/4$ that $E[4, 2] = E[2, 1] = 1$ (yielding an additional $1/8$). However, if $E[4, 2] = 1 \neq 0 = E[2, 1]$ we have with probability $1/6$ that $[2, 3] \in \mathcal{T}$ and $E[3, 1] = 1$ (yielding an additional

1/48), or, with probability 1/6 that $[2, 3] \notin \mathcal{T}$, but $E[2, 3] = E[3, 1] = 1$ (yielding an additional 1/48), or last, we have with probability 1/12 that $[2, 3] \notin \mathcal{T}$, $E[2, 3] = 0$ and $E[4, 3] = E[3, 1] = 1$ (yielding an additional 1/96).

- ... with probability 1/8 that $E[4, 2] = 0$, but $E[4, 3] = E[3, 1] = 1$ (yielding an additional 1/16). However, if $E[3, 1] = 0$ we have as above with probability 1/3 that $E[2, 3] = E[2, 1] = 1$ (yielding an additional 1/48).

Yielding in total a probability of

$$\Theta_{\text{B-TAP}}(E_X)[4, 1] = \frac{48 + 12 + 2 + 2 + 1 + 6 + 2}{96} = \frac{73}{96}.$$

• $\Theta_{\text{B-TAP}}(\mathbf{E}_X)[4, 2]$:

When drawing E there is a 1/2 probability that $E[4, 2] = 1$. Given $E[4, 2] = 0$ we have...

- ... with probability 2/3 that $[1, 2] \in \mathcal{T}$ (which automatically means $[3, 2] \notin \mathcal{T}$). Then, with probability 1/2 we have $E[4, 1] = 1$ (yielding an additional 1/6), with probability 1/8 that $E[4, 1] = 0$ and $E[4, 3] = E[3, 2] = 1$ (yielding an additional 1/24) and with probability 1/16 we have $E[4, 1] = E[3, 2] = 0$ and $E[4, 3] = E[3, 1] = 1$ (yielding an additional 1/48).
- ... with probability 1/3 that $\mathcal{T} = \{[1, 3][3, 2]\}$. Then, with probability 1/2 we have $E[4, 3] = 1$ (yielding an additional 1/12), with probability 1/4 that $E[4, 3] = 0$ and $E[4, 1] = 1$ (yielding an additional 1/24).

Yielding in total a probability of

$$\Theta_{\text{B-TAP}}(E_X)[4, 2] = \frac{24 + 8 + 2 + 1 + 4 + 2}{48} = \frac{41}{48}.$$

• $\Theta_{\text{B-TAP}}(\mathbf{E}_X)[2, 3]$:

When drawing E there is...

- ... a 1/3 probability that $[2, 3] \in \mathcal{T}$.
- ... a 1/3 probability that $[2, 3] \notin \mathcal{T}$ and $E[2, 3] = 1$ (yielding an additional 1/3).
- ... a 1/3 probability that $[2, 3] \notin \mathcal{T}$ and $E[2, 3] = 0$. Then, with probability 1/2 we have $E[2, 1] = 1$ (yielding an additional 1/6).

Yielding in total a probability of

$$\Theta_{\text{B-TAP}}(E_X)[4, 2] = \frac{16 + 16 + 8}{48} = \frac{40}{48}.$$

Putting everything together we obtain

$$(\Theta_{\text{B-TAP}}(E_X)[k, \ell])_{k, \ell=1}^4 = \frac{1}{96} \begin{pmatrix} 96 & 96 & 96 & 0 \\ 64 & 96 & 80 & 0 \\ 64 & 80 & 96 & 0 \\ 73 & 82 & 82 & 96 \end{pmatrix}.$$

Appendix A.6: Proof of Lemma 3.22

Proof Let G_q be an ancestral causal graph with node set $V_s := \{v_0, v_1, v_2, v_3, w_0, w_1, \dots, w_s\}$ for $s \geq 2$. The set of available interventions is given by $\mathcal{I} = \{v_0, v_3, w_0\}$ and the partial observation $E_X^{(s)}$ is given by

$$\begin{aligned} E[v_0, v_k] &= 1 \quad \text{for } k = 1, 2, 3, \\ E[v_\ell, w_k] &= 0 \quad \text{for } \ell = 0, 3 \text{ and } k = 0, 1, \dots, s, \\ E[v_3, v_k] &= 0 \quad \text{for } k = 0, 1, \\ E[v_3, v_2] &= 1, \\ E[w_0, v_k] &= 0 \quad \text{for } k = 0, 1, 2, 3, \\ E[w_0, w_k] &= 1 \quad \text{for } k = 1, \dots, s, \end{aligned}$$

by abuse of notation. We observe the following impossible edges by Lemma 3.13

$$\begin{aligned} K &= \{[v_k, w_\ell] : \forall k, \ell\} \cup \{[w_\ell, v_k] : \forall k, \ell\} \\ &\cup \{[v_k, v_\ell] : k = 2, 3, \ell = 0, 1\}. \end{aligned}$$

Let $E_{(q,s)}$ be a partial RDG with edge probability q as drawn in step (3.A) in Algorithm 1. Define

$$\begin{aligned} \gamma_0(q, s) &:= \mathbb{P}[E_{(q,s)}^+[v_1, v_2] = 1 | E_{(q,s)} \in \mathcal{X}], \\ \gamma_1(q, s) &:= \mathbb{P}[E_{(q,s)}^+[w_1, w_2] = 1 | E_{(q,s)} \in \mathcal{X}]. \end{aligned}$$

To show the claim, it suffices to show that there exist q_0 and s_0 such that $\gamma_0(0.5, s_0) < \gamma_1(0.5, s_0)$ and $\gamma_0(q_0, s_0) > \gamma_1(q_0, s_0)$.

1. **Claim:** There exists a $s_0 \geq 2$ such that $\gamma_0(0.5, s_0) < \gamma_1(0.5, s_0)$.

First, note that by definition $\gamma_0(q, s)$ is independent of the choice of s . Hence there exists $\gamma_0(q) \in [0, 1]$ such that

$$\gamma_0(q) = \gamma_0(q, s) \quad \text{for all } s \geq 2.$$

In particular by construction $\gamma_0(q) \leq c$ for some $c < 1$. Moreover, by Karp (1990) we have that for a fixed q and $s \rightarrow \infty$ the number of nodes in the largest component of the random subgraph on $\tilde{V} = \{w_0, w_1, \dots, w_s\}$ converges exponentially to s , in particular faster than $s^{-1/s}$.

Thus, we have that

$$\mathbb{P}[E[w_k, w_\ell] = 1 \text{ for all } k, \ell = 0, 1, \dots, s] \xrightarrow{s \rightarrow \infty} 1.$$

Hence,

$$\gamma_1(q, s) \xrightarrow{s \rightarrow \infty} 1,$$

yielding the existence of an $s_0 \geq 2$ such that $\gamma_0(0.5, s_0) = \gamma_0(0.5) < \gamma_1(0.5, s_0)$.

- Claim:** For fixed $s \geq 2$ we have for $q \rightarrow 0$ that $\gamma_0(q, s) \rightarrow 1/2$.

For any $s \geq 2, q < 1/2$ and $x, y \in V$ we compute

$$\begin{aligned} \mathbb{P}[E^+[x, y] = 1 | E_{(q,s)} \in \mathcal{X}] &= \frac{\mathbb{P}[E^+[x, y] = 1 \text{ and } E_{(q,s)} \in \mathcal{X}]}{\mathbb{P}[E_{(q,s)} \in \mathcal{X}]} \\ &= \frac{\sum_{E_0 \in \mathcal{X} \text{ s.t. } E_0^+[x,y]=1} \mathbb{P}[E_{(q,s)} = E_0]}{\sum_{E_0 \in \mathcal{X}} \mathbb{P}[E_{(q,s)} = E_0]} \tag{A35} \\ &= \frac{\sum_{E_0 \in \mathcal{X} \text{ s.t. } E_0^+[x,y]=1} q^{A(E_0)}(1-q)^{B(E_0)}}{\sum_{E_0 \in \mathcal{X}} q^{A(E_0)}(1-q)^{B(E_0)}} \\ &\xrightarrow{q \rightarrow 0} \frac{|\{E_0 \in \mathcal{X}_0 : E_0^+[x, y] = 1\}|}{|\mathcal{X}_0|}, \end{aligned}$$

where $\mathcal{X}_0 \subset \mathcal{X}$ is the subset of graphs with minimal edges and

$$\begin{aligned} A(E_0) &= |\{[k, \ell] : E_0[k, \ell] = 1\}|, \\ B(E_0) &= |\{[k, \ell] : E_0[k, \ell] = 0\}| \end{aligned}$$

To satisfy the constraints of the partial observation the graphs with the minimal amount of edges in \mathcal{X} are given by those that feature an edge $E[v_3, v_2] = 1$, that feature either $E[v_0, v_1] = E[v_0, v_3] = 1$ or $E[v_0, v_1] = E[v_1, v_3] = 1$ and that feature a spanning tree rooted in w_0 on the subset \tilde{V} . Thus, we obtain

$$\gamma_0(q_0, s) \xrightarrow{q \rightarrow 0} \frac{1}{2}.$$

- Claim:** For fixed $s \geq 2$ there exists a $c < 1/2$ such that for $q \rightarrow 0$ we have $\gamma_1(q, s) \rightarrow c$.

Consider that for any spanning tree on \tilde{V} rooted in w_0 that features a path from w_1 to w_2 we can switch the labels of w_1 and w_2 to obtain a spanning tree without a path from w_1 to w_2 . By construction, the above assignment is injective yielding that there are at most as many spanning trees featuring a path from w_1 to w_2 as there are spanning trees who do not feature such a path. Furthermore, consider the spanning tree given by $E[w_0, w_k] = 1$ for

all $k = 1, \dots, q$ to obtain that there exist spanning trees that feature neither a path from w_1 to w_2 , nor from w_2 to w_1 . Hence by (A35) and acyclicity of spanning trees there exists a $c < 1/2$ such that

$$\gamma_1(q_0, s) \xrightarrow{q \rightarrow 0} c.$$

By Claim 1 we can choose $s_0 \geq 2$ such that $\gamma_0(0.5, s_0) < \gamma_1(0.5, s_0)$. Given this s_0 , by Claim 2. and 3. we can choose $q_0 \in (0, 1)$ small enough such that $\gamma_0(q_0, s_0) > \gamma_1(q_0, s_0)$, finishing the proof.

To adapt the above proof for Algorithm 2 when the RSTs in Step 4.B are drawn uniformly at random consider the following step:

- The first claim follows analogously.
- Instead of the second claim it can be shown that $\gamma_0(q, s) \rightarrow 5/8$ by considering that $E^+[v_1, v_2] = 1$ for 5 of 8 possible spanning trees rooted in v_0 (note that $E[v_3, v_2] = 1$ in any case). Note moreover, that in the case the RSTs are drawn via the modified Broder algorithm as given in Sect. 3.2 $\gamma_0(q, s) \rightarrow c < 1/2$.
- The third claim can be shown to hold also for Algorithm 2 by the analogous arguments. This is true even for the modified version of the Broder algorithm, as it coincides on \tilde{V} with the classical one.

Appendix B: Additional plots

Appendix B.1: Additional simulations

In Fig. 8 mean ROC curves are displayed for $p = 25, 50, 250, 500$ relating to Fig. 3. In Fig. 9 the effect of varying the sparsity parameter α on the performance of the proposed graph-based predictors is shown for a graph featuring $p = 1000$ and $\#\mathcal{I} = 200$. In Fig. 10 the effect of varying the number of known rows on the performance of the proposed graph-based predictors is shown for a graph featuring $p = 1000$ and $\alpha = 0.7$. Last, in Fig. 11 the AUC performs for all predictors outputting a matrix of scores on the Sachs et al. (2005) proteomics data is reported via box-plots.

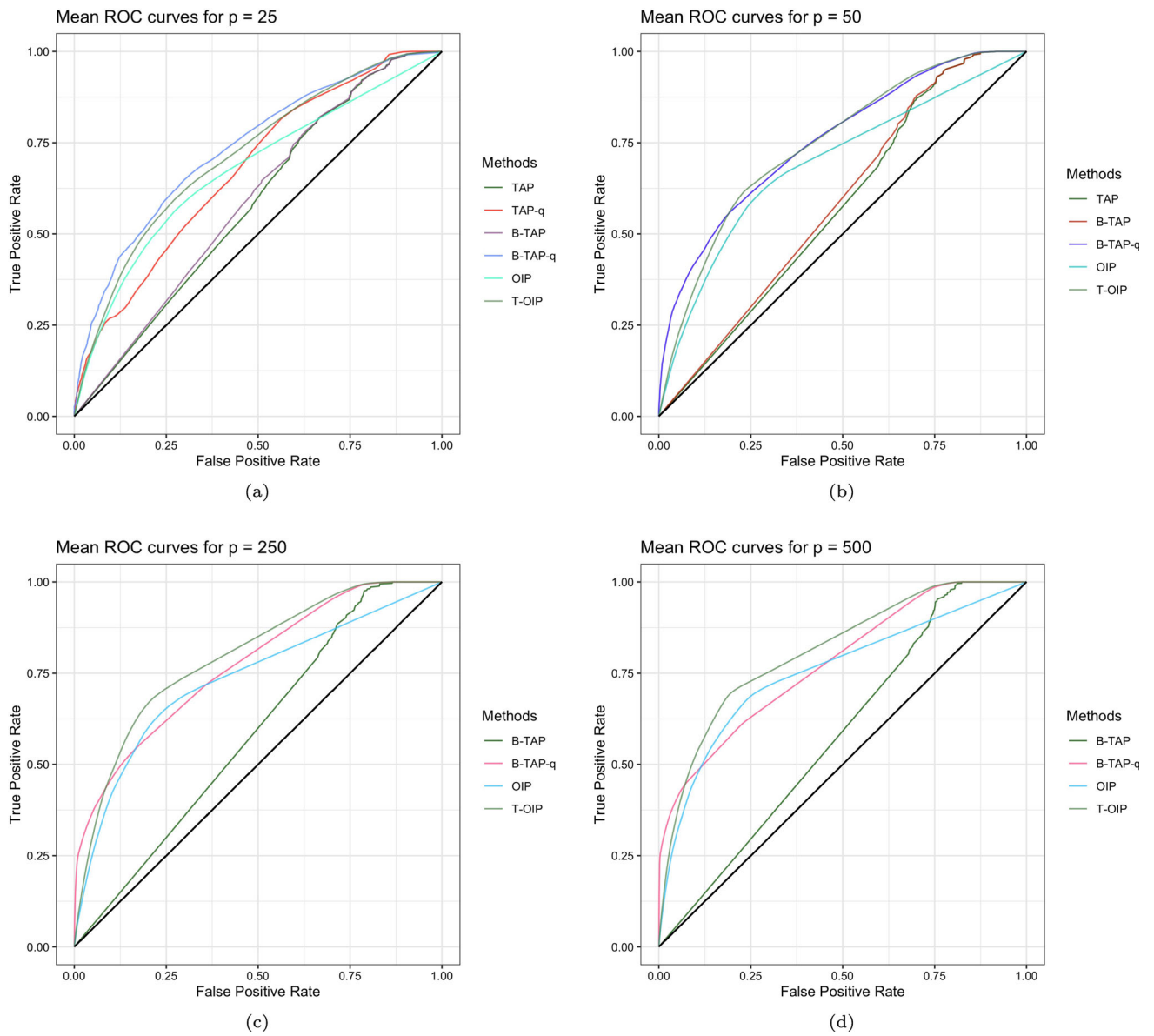


Fig. 8 Mean ROC curves for Fig. 3

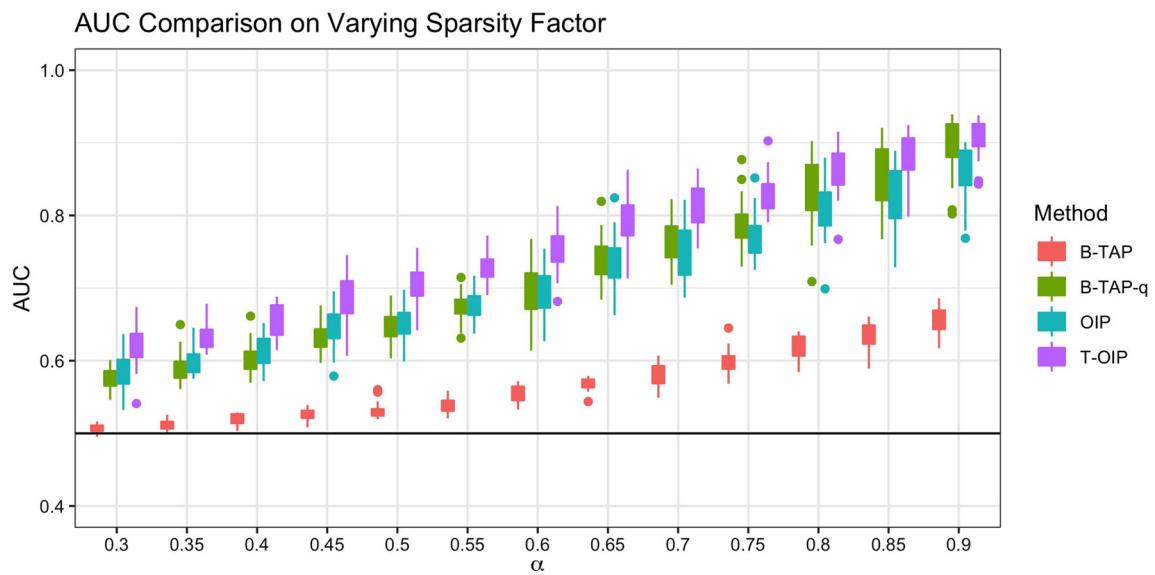


Fig. 9 AUC performance for varying sparsity parameter α . Each experiment was repeated 20 times and all graphs featured $p = 1000$ and $|\mathcal{I}| = 200$

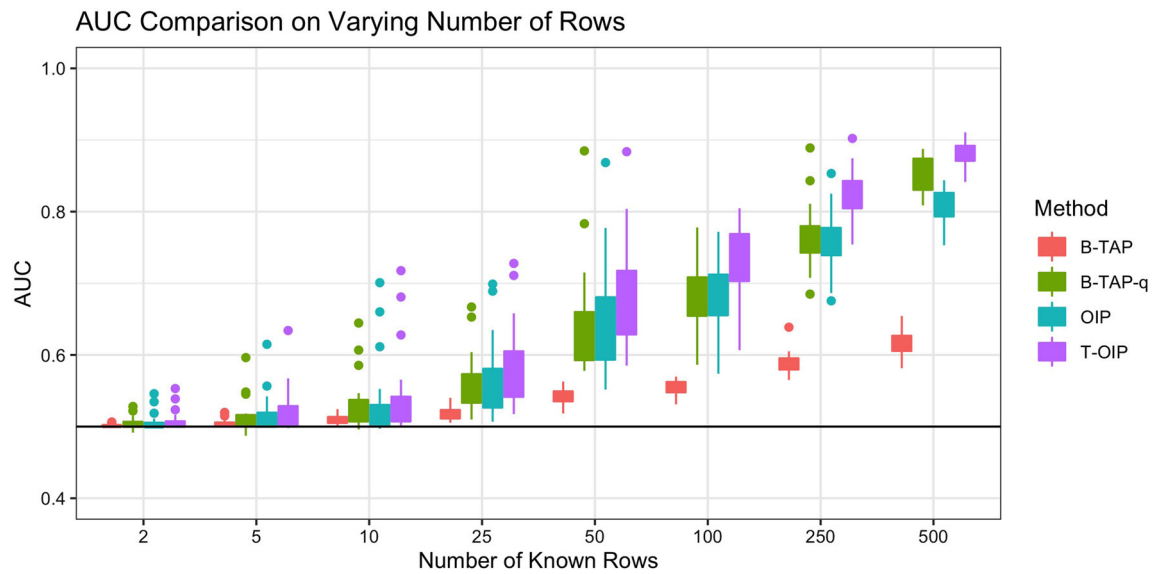


Fig. 10 AUC performance for varying number of known rows $|\mathcal{I}|$. Each experiment was repeated 20 times and all graphs featured $p = 1000$ and $\alpha = 0.7$

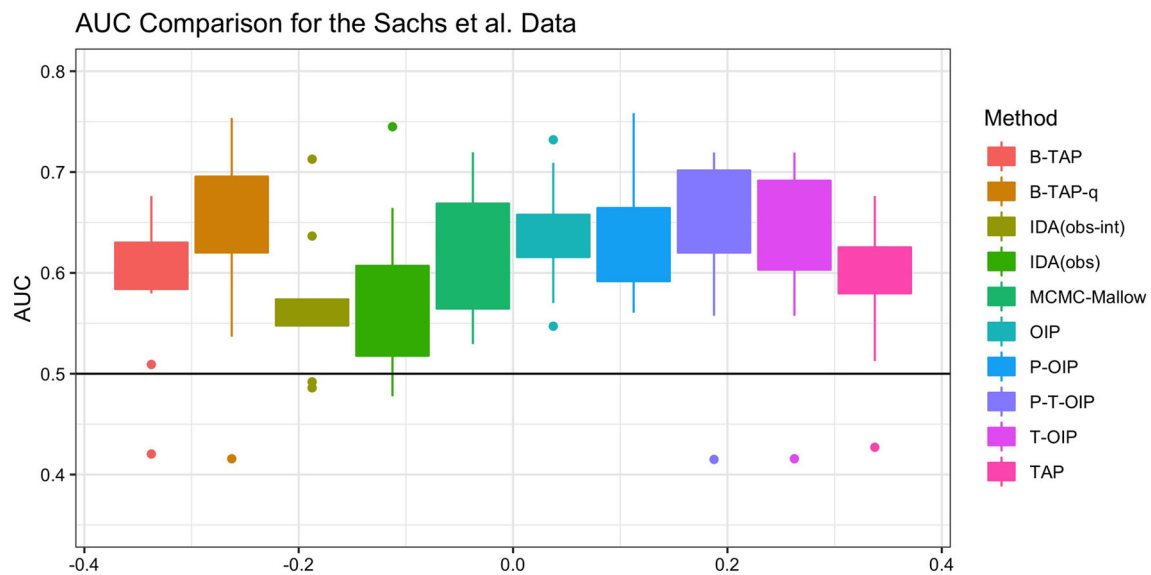


Fig. 11 AUC performance for the Sachs et al. (2005) data

Appendix B.2: Computational costs

In Fig. 12a–c the computation times³ for all predictors regarding Figs. 3, 5 and 11 are given, respectively. Note that the computational costs for the IDA (obs) and IDA (int-obs) are basically the same as for PC (obs) and PC (int-obs),

respectively, and thus for the ease of presentation the PC results are omitted in (a) and (b). Moreover, in Fig. 12a the computation time of a transitive closure is included as a reference.

³ All computations were run on a HP Z840 workstation.

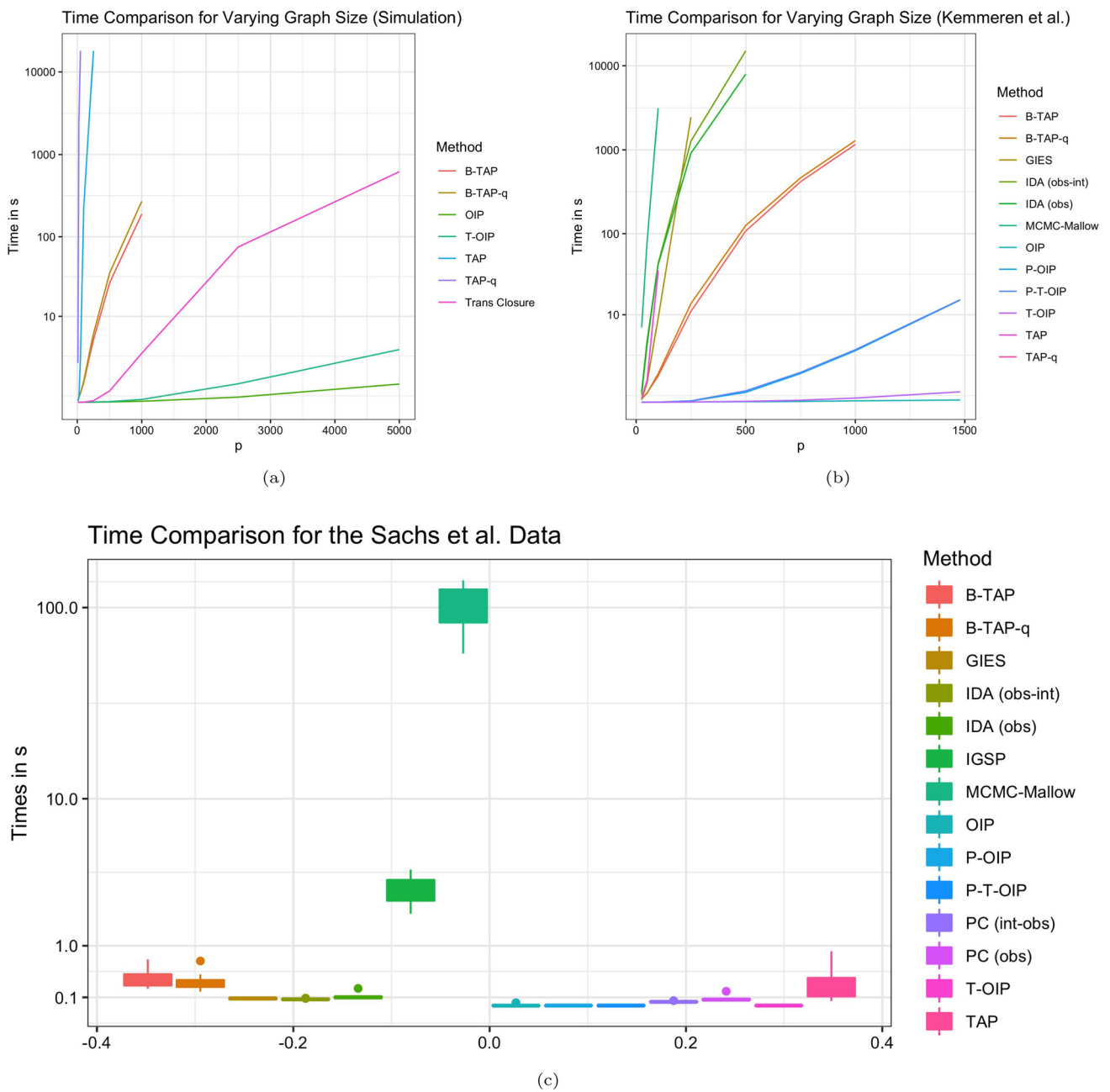


Fig. 12 Mean computation times for Fig. 3 in panel (a) TAP and TAP-q reach the *time out* at 18,000s, Fig. 5 in panel (b) and Fig. 11 in panel (c)

References

- Anari, N., Hu, N., Saberi, A., Schild, A.: Sampling arborescences in parallel (2020). [arXiv:2012.09502](https://arxiv.org/abs/2012.09502)
- Artzy-Randrup, Y., Stone, L.: Generating uniformly distributed random networks. *Phys. Rev. E* **72**(5), 056708 (2005)
- Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., Teichmann, S.A.: Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**(3), 283–291 (2004). <https://doi.org/10.1016/j.sbi.2004.05.004>
- Blitzstein, J., Diaconis, P.: A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Math.* **6**(4), 489–522 (2011). <https://doi.org/10.1080/15427951.2010.557277>
- Bollobás, B., Janson, S., Riordan, O.: The phase transition in inhomogeneous random graphs. *Random Struct. Algorithms* **31**(1), 3–122 (2007)
- Broder, A.Z.: Generating random spanning trees. *FOCS* **89**, 442–447 (1989)
- Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., Drouin, A.: Differentiable causal discovery from interventional data. *Adv. Neural Inf. Process. Syst.* **33**, 21865–21877 (2020)
- Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3741–3782 (2014)
- Colombo, D., Maathuis, M.H., Kalisch, M., Richardson, T.S.: Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* **40**(1), 294–321 (2012). <https://doi.org/10.1214/11-AOS940>
- Cortes, C., Mohri, M.: Confidence intervals for the area under the roc curve. *Adv. Neural Inf. Process. Syst.* **17** (2004)
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Aron, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N., Regev, A.: Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**(7), 1853–1866.e17 (2016). <https://doi.org/10.1016/j.cell.2016.11.038>
- Eigenmann, M., Mukherjee, A., Maathuis, M.: Evaluation of causal structure learning algorithms via risk estimation. In: UAI, pp. 151–160. PMLR (2020). <http://proceedings.mlr.press/v124/eigenmann20a.html>
- Fornito, A., Zalesky, A., Bullmore, E.: *Fundamentals of Brain Network Analysis*. Academic Press, Cambridge (2016). <https://doi.org/10.1016/C2012-0-06036-X>
- Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010)
- Fosdick, B.K., Larremore, D.B., Nishimura, J., Ugander, J.: Configuring random graph models with fixed degree sequences. *SIAM Rev.* **60**(2), 315–355 (2018). <https://doi.org/10.1137/16M1087175>
- Gauvin, L., Génois, M., Karsai, M., Kivela, M., Takaguchi, T., Valdano, E., Vestergaard, C.L.: Randomized reference models for temporal networks (2018). [arXiv:1806.04032](https://arxiv.org/abs/1806.04032)
- Greenhill, C.: The switch Markov chain for sampling irregular graphs. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1564–1572. SIAM (2014)
- Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**(1), 29–36 (1982)
- Hauser, A., Bühlmann, P.: Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13**(1), 2409–2464 (2012)
- Heinze-Deml, C., Maathuis, M.H., Meinshausen, N.: Causal structure learning. *Annu. Rev. Stat. Appl.* **5**(1), 371–391 (2018). <https://doi.org/10.1146/annurev-statistics-031017-100630>
- Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L.V., Favorov, A.V., Lee, W.S., Taylor, D., Hu, C.W., Long, B.L., Noren, D.P., Bisberg, A.J., Consortium, H.-D., Mills, G.B., Gray, J.W., Kellen, M., Norman, T., Friend, S., Qutub, A.A., Fertig, E.J., Guan, Y., Song, M., Stuart, J.M., Spellman, P.T., Koepl, H., Stolovitzky, G., Saez-Rodriguez, J., Mukherjee, S.: Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**(4), 310–318 (2016). <https://doi.org/10.1038/nmeth.3773>
- Hill, S.M., Oates, C.J., Blythe, D.A., Mukherjee, S.: Causal learning via manifold regularization. *J. Mach. Learn. Res.: JMLR* **20**, 127 (2019). <https://doi.org/10.17863/cam.44718>
- Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**(460), 1090–1098 (2002)
- Hyttinen, A., Eberhardt, F., Järvisalo, M.: Constraint-based causal discovery: Conflict resolution with answer set programming. UAI (2014). http://www.its.caltech.edu/~fehardt/papers/HEJ_UAI2014.pdf
- Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001). <https://doi.org/10.1146/annurev.genom.2.1.343>
- Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47**(11), 1–26 (2012)
- Karp, R.M.: The transitive closure of a random digraph. *Random Struct. Algorithms* **1**(1), 73–93 (1990). <https://doi.org/10.1002/rsa.3240010106>
- Kemmeren, P., Sameith, K., van de Pasch, L.A.L., Benschop, J.J., Lenstra, T.L., Margaritis, T., O’Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C.W., van Heesch, S., Kashani, M.M., Ampatzidis-Michailidis, G., Brok, M.O., Brabers, N.A.C.H., Miles, A.J., Bouwmeester, D., van Hooff, S.R., van Bakel, H., Sluiter, E., Bakker, L.V., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M.J.A., Holstege, F.C.P.: Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**(3), 740–752 (2014). <https://doi.org/10.1016/j.cell.2014.02.054>
- Krivelevich, M., Sudakov, B.: The phase transition in random graphs: a simple proof. *Random Struct. Algorithms* **43**(2), 131–138 (2013). <https://doi.org/10.1002/rsa.20470>
- Maathuis, M.H., Kalisch, M., Bühlmann, P.: Estimating high-dimensional intervention effects from observational data. *Ann. Stat.* **37**(6A), 3133–3164 (2009). <https://doi.org/10.1214/09-AOS685>
- Magliacane, S., van Ommen, T.: Causal transfer learning (2017). <https://staff.science.uva.nl/j.m.mooij/articles/1707.06422.pdf>
- Magliacane, S., Claassen, T., Mooij, J.: Joint causal inference on observational and experimental datasets (2016a). <https://staff.fnwi.uva.nl/j.m.mooij/articles/1611.10351v2.pdf>
- Magliacane, S., Claassen, T., Mooij, J.M.: Ancestral causal inference. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc. (2016b). <https://proceedings.neurips.cc/paper/2016/file/f3d9de86462c28781cbe5c47ef22c3e5-Paper.pdf>
- Malinsky, D., Spirtes, P.: Estimating causal effects with ancestral graph Markov models. In: *Conference on Probabilistic Graphical Models*, pp. 299–309. PMLR (2016)
- Meinshausen, N., Hauser, A., Mooij, J., Peters, J., Versteeg, P., Bühlmann, P.: Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl. Acad. Sci. U.S.A.* **113**(27), 7361–7368 (2016). <https://doi.org/10.1073/pnas.1510493113>

- Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.: On the uniform generation of random graphs with prescribed degree sequences (2003). <https://arxiv.org/abs/cond-mat/0312028>
- Mooij, J.M., Claassen, T.: Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In: Conference on Uncertainty in Artificial Intelligence, pp. 1159–1168. PMLR (2020)
- Newman, M.E.: Mixing patterns in networks. *Phys. Rev. E* **67**(2), 026126 (2003)
- Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113 (2004)
- Nogueira, A.R., Pugnana, A., Ruggieri, S., Pedreschi, D., Gama, J.: Methods and tools for causal discovery and causal inference. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **12**(2), e1449 (2022)
- O'Donnell, R.T., Korb, K.B., Allison, L.: Causal KL: evaluating causal discovery (2021). [arXiv:2111.06029](https://arxiv.org/abs/2111.06029)
- Parikshak, N.N., Gandal, M.J., Geschwind, D.H.: Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Rev. Genet.* **16**(8), 441–458 (2015). <https://doi.org/10.1038/nrg3934>
- Pearl, J.: *Causality*. Cambridge University Press, Cambridge (2009). <https://doi.org/10.1017/CBO9780511803161>
- Peters, J., Bühlmann, P., Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **78**(5), 947–1012 (2016). <https://doi.org/10.1111/rssb.12167>
- Phillips, P.C.: Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**(11), 855–867 (2008). <https://doi.org/10.1038/nrg2452>
- Rau, A., Jaffrézic, F., Nuel, G.: Joint estimation of causal effects from observational and intervention gene expression data. *BMC Syst. Biol.* **7**, 111 (2013). <https://doi.org/10.1186/1752-0509-7-111>
- Richardson, T.: A discovery algorithm for directed cyclic graphs. In: Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence, pp. 454–461 (1996)
- Rothenhäusler, D., Bühlmann, P., Meinshausen, N.: Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *Ann. Stat.* **47**(3), 1688–1722 (2019). <https://doi.org/10.1214/18-AOS1732>
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**(5721), 523–529 (2005). <https://doi.org/10.1126/science.1105809>
- Sanguinetti, G., Huynh-Thu, V.A. (eds.): *Gene Regulatory Networks: Methods and Protocols*. Methods in Molecular Biology, vol. 1883. Springer, New York (2019). <https://doi.org/10.1007/978-1-4939-8882-2>
- Shalem, O., Sanjana, N.E., Zhang, F.: High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**(5), 299–311 (2015)
- Spencer, S.E., Hill, S.M., Mukherjee, S.: Inferring network structure from interventional time-course experiments. *Ann. Appl. Stat.* **9**, 507–524 (2015)
- Spirtes, P.: Introduction to causal inference. *J. Mach. Learn. Res.* **11**(5) (2010)
- Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: *Causation, Prediction, and Search*. MIT Press, Cambridge (2000)
- Ursu, O., Neal, J.T., Shea, E., Thakore, P.I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M.M., et al.: Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nat. Biotechnol.* **40**, 1–10 (2022)
- Wang, Y., Solus, L., Yang, K., Uhler, C.: Permutation-based causal inference algorithms with interventions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates Inc., New York (2017)
- Zhang, J.: Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.* **9**, 1437–1474 (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.