

## JOINT ESTIMATION OF MULTIPLE RELATED BIOLOGICAL NETWORKS

BY CHRIS. J. OATES<sup>1,\*</sup> JIM KORKOLA<sup>2</sup> JOE W. GRAY<sup>2</sup> AND SACH MUKHERJEE<sup>3,4,\*</sup>

<sup>1</sup>*University of Warwick*, <sup>2</sup>*OHSU Knight Cancer Institute*, <sup>3</sup>*MRC Biostatistics Unit* and <sup>4</sup>*University of Cambridge*

Graphical models are widely used to make inferences concerning interplay in multivariate systems. In many applications, data are collected from multiple related but non-identical units whose networks may differ but are likely to share many features. Here we present a hierarchical Bayesian formulation for joint estimation of multiple networks under an exchangeability assumption. The formulation is general and given a suitable class of graphical models can be used to provide a corresponding joint estimator. Motivated by emerging experimental designs in molecular biology, we focus on time-course data with interventions, using dynamic Bayesian networks as the graphical models. We introduce a computationally efficient, deterministic algorithm for exact joint inference in this setting. Theoretical results demonstrate that joint estimation offers gains relative to separate inference for individual networks. We present empirical results that support and extend the theory, including an extensive simulation study and an application to proteomic data from human cancer cell lines. Finally, we describe approximations that are still more computationally efficient than the exact algorithm and that also demonstrate good empirical performance.

**1. Introduction.** Graphical models are widely used to represent multivariate systems. Vertices in a graph (or network)  $G$  are identified with random variables and edges between the vertices describe conditional independence statements or, with suitable modeling and semantic extensions, causal influences between the variables. In many applications a key statistical challenge is to construct a network estimator  $\hat{G}(\mathbf{y})$ , based on data  $\mathbf{y}$ , that performs well in a sense appropriate to the application. Such “network inference” is increasingly a mainstream approach in many disciplines including neuroscience, sociology and biology.

Network inference methods usually assume that the data are identically distributed (specifically, that datasets satisfy an exchangeability assumption). However, in many applications, data are not identically distributed,

---

\* To whom correspondence should be addressed.

but are instead obtained from multiple related but non-identical units (or “individuals”, we use both terms interchangeably). This paper concerns network inference in this non-identically distributed setting.

Our work is motivated by biological networks in cancer. Multiple studies have demonstrated the remarkable genomic heterogeneity of cancer ([The 1000 Genomes Project Consortium, 2010](#); [The Cancer Genome Atlas Network, 2012](#)). At the same time, the question of how such heterogeneity is manifested at the level of biological networks has remained poorly understood. We focus in particular on protein signaling networks in human cancer cell lines. Signaling networks describe biochemical interplay between proteins and are central to cancer biology. However, sequence and transcript data alone are inadequate for the study of signaling, and indeed these data types can be discordant with the abundance of signaling proteins and post-translational modifications (including phosphorylation) that are key to the process ([Akbari \*et al.\*, 2014](#)). Recent developments in proteomics have started to allow data-driven study of signaling heterogeneity, notably reverse-phase protein arrays (RPPA; see e.g. [Hennessy \*et al.\*, 2010](#)) a technology that provides the data we analyse below.

To fix ideas, we begin by describing the specific application that motivates this work. We consider time course phospho-protein measurements obtained using RPPA technology (details appear below) for 6 cell lines. The goal of the study is to infer cell line-specific protein signaling networks  $G^j$ ,  $j = 1, \dots, 6$ , and additionally to highlight experimentally testable differences between them. Prior network information is available from the literature, but it is believed that cell line-specific genetic alterations may induce differences with respect to the “literature network” (and between cell lines). At the same time, the amount of data per cell line is limited (6 time points in each of 4 conditions, making a total of 24 data points per cell line  $j$ , constituting data  $\mathbf{y}^j$ ). Since the cell lines  $j$  are closely related, yet potentially different with respect to underlying networks, a key inferential question is how to “borrow strength” between the network estimation problems. That is, we seek a joint estimator of the cell line-specific networks  $\{G^1 \dots G^6\}$  based on the entire (non-identically distributed) dataset  $\{\mathbf{y}^1 \dots \mathbf{y}^6\}$  that shares information between the estimation problems whilst preserving the ability to identify cell line-specific network structure.

This application is an example of a more general class of biological applications, where individuals  $j$  could correspond to e.g. different patients or cell lines (or groups thereof; e.g. disease subtypes) and the networks themselves to gene regulatory or protein signaling networks that could depend on the genetic and epigenetic state of the individuals. Indeed, continuing reduction

in the unit cost of biochemical assays has led to an increase in experimental designs that include panels of potentially heterogeneous individuals (Barretina *et al.*, 2012; Cao *et al.*, 2011; Maher, 2012; The Cancer Genome Atlas Network, 2012). As in the signaling example above, in such settings, given individual-specific data  $\mathbf{y}^j$ , there is scientific interest in individual-specific networks  $G^j$  and their similarities and differences.

Following Werhli and Husmeier (2008); Penfold *et al.* (2012) and others, we focus on the case of directed networks  $G^j$  that are exchangeable in the sense that inference is invariant to permutation of individuals  $j \in \mathcal{J} = \{1, \dots, J\}$ . We model data on all individuals  $\{\mathbf{y}^j : j \in \mathcal{J}\}$  within a joint Bayesian framework. Regularization of individual networks is achieved by introducing a latent network  $G$  to couple inference across all individuals. We report posterior marginal inclusion probabilities for every possible edge in each individual network  $G^j$  in addition to the latent network  $G$ . The high-level formulation we propose is general and in principle essentially any graphical model of interest could be embedded within the framework proposed to enable joint estimation.

In general, the individual  $j$ 's could have complex, hierarchical relationships, for example with  $j$ 's belonging to groups and subgroups (e.g. corresponding to cancer types and subtypes; see Curtis *et al.*, 2012). The exchangeable case we consider corresponds in a sense to the simplest possible hierarchy in which each individual is dependent on a single latent graph (see Fig. 1). In settings where groups can be treated as approximately homogeneous, the approach presented in this paper can be trivially used to give group-level estimates, by using  $j$  to index groups rather than individuals, with all data for group  $j$  modeled as dependent on graph  $G^j$ . This corresponds to an assumption of identically distributed data within (but not between) groups. In the empirical study presented below we consider also robustness of our approach under violation of the exchangeability assumption.

For the application to time-course data from protein signaling that we focus on, we present a detailed development using directed graphical models called dynamic Bayesian networks (DBNs). These are directed acyclic graphs (DAGs) with explicit time-indices (Murphy, 2002). The main contributions of this paper are:

- **Bayesian computation.** For the time-course setting, we put forward an efficient and exact algorithm. This is done by exploiting factorization properties of the DBN likelihood, analytic marginalization over continuous parameters and belief propagation. In moderate-dimensional settings this allows exact joint estimation to be carried out

in seconds to minutes (we discuss computational complexity below).

- **Theory.** We provide a result that quantifies the statistical efficiency of joint relative to separate estimation and that gives a sufficient condition for improved performance.
- **Empirical investigation.** The availability of an efficient Bayesian algorithm enables, for the first time, a comprehensive empirical study of the statistical properties of joint estimators in the exchangeable network setting, including a wide range of simulation regimes and an application to protein signaling in a panel of breast cancer cell lines. We formulate joint estimators based on classical (non-joint) DBNs, including a recent causal variant suitable for interventional data (Spencer and Mukherjee, 2012). Joint estimation is often found to outperform the corresponding individual-level estimators. Some computationally favorable approximations to joint inference are described, that we find perform well under a range of conditions.

Related work includes Niculescu-Mizil and Caruana (2007); Werhli and Husmeier (2008); Dondelinger *et al.* (2012) who considered joint estimation of multiple related directed acyclic graphs (DAGs), but did not provide an exact inference algorithm. Joint estimation has recently been discussed in the Gaussian graphical model (GGM) literature, most recently by Danaher *et al.* (2014). In contrast to GGMs, motivated by our application in protein signaling, we focus on directed graphical models that admit a causal interpretation. Approaches to capturing context-specific conditional independence, based on the embellishment of Bayesian networks, include Boutilier *et al.* (1996); Geiger and Heckerman (1996). Our approach differs by regularizing based on network structure alone; we do not place exchangeability assumptions on the data-generating parameters. Our work is most closely related to Penfold *et al.* (2012) who also considered Bayesian joint estimation of directed graphs from time-course data, however our work differs in several respects. First, for the time-course setting, the exact algorithm we propose offers massive computational gains. As we discuss in detail below the methodology of Penfold *et al.* (2012) is prohibitively computationally expensive for the applications we consider here. Second, the computational efficiency of our approach allows us to present a much more extensive study of joint estimation than has hitherto been possible. Third, we allow for prior information regarding the network structure and ancillary information including individual-specific characteristics.

The remainder of the paper is organized as follows. In section 2 we describe a hierarchical Bayesian formulation and in section 3 we discuss computationally efficient joint inference in the case of DBNs. Empirical results

are presented in section 4, including an application to protein signaling in cancer. Finally we close with a discussion of our findings in section 5.

**2. Joint network inference: The general case.** We describe a general statistical formulation for joint network inference that can be coupled to essentially arbitrary classes of graphical models. For computational tractability it may be necessary to place restrictions on the class of graphical models; in section 3 we present a detailed development for DBNs, that are well-suited to our motivating application in breast cancer.

*2.1. Hierarchical model.* Consider a space  $\mathcal{G}$  of graphs on the vertex set  $\mathcal{P} = \{1, \dots, P\}$ . To keep the presentation general, we do not specify the type of graph nor restrictions on  $\mathcal{G}$  at this stage (the special case of DBNs for time-course data is described below). As shown in Fig. 1, each individual network  $G^j \in \mathcal{G}$  is modeled with dependence on a latent network  $G \in \mathcal{G}$  that in turn depends on a prior network  $G^0 \in \mathcal{G}$  (section 2.2). In this way, estimates of the individual networks  $G^j$  are regularized by shrinkage towards the common latent network  $G$  that, in turn, may be constrained by an informative network prior. As in any graphical model, observations  $\mathbf{Y}^j$  on individual  $j$  are dependent upon a graph  $G^j$  and parameters  $\theta^j$ . Here  $Z^j$  denotes any ancillary information available on individual  $j$ . The model is specified by

$$\begin{aligned} (1) \quad & p(G|G^0, \eta) \propto \exp(-\eta d(G, G^0)) \\ (2) \quad & p(G^j|G, \lambda^j, Z^j) \propto \exp(-\lambda^j d^j(G^j, G; Z^j)) \end{aligned}$$

and a suitably chosen graphical model likelihood  $p(\mathbf{Y}^j|G^j, \theta^j, Z^j)$ . Eqn. 1 follows the “network prior” approach of Mukherjee and Speed (2008) that was proposed for biological applications where subjective prior structural information is available. The functionals  $d^j, d : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}$  and hyperparameters  $\eta, \lambda^j$  must be specified (section 2.2). This paper restricts attention to exchangeable models; in particular we consider functionals  $d^j$  that are independent of the index  $j$ . We refer to the above formulation as *joint network inference* (JNI).

*2.2. Network prior.* The network prior (Eqn. 1) requires a penalty functional  $d : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}$  and a prior network  $G^0 \in \mathcal{G}$ , with the former capturing how close a candidate network  $G \in \mathcal{G}$  is to the latter. We discuss choice of  $G^0$  below. Given  $G^0$ , a simple choice of penalty function  $d$  is the structural Hamming distance (SHD) given by  $d(G, G^0) = \|G - G^0\|$  where  $\|M\| = \sum_{i,j} |m_{i,j}|$  is the  $\ell_1$ -norm of an adjacency matrix and the differential network  $G - G^0$

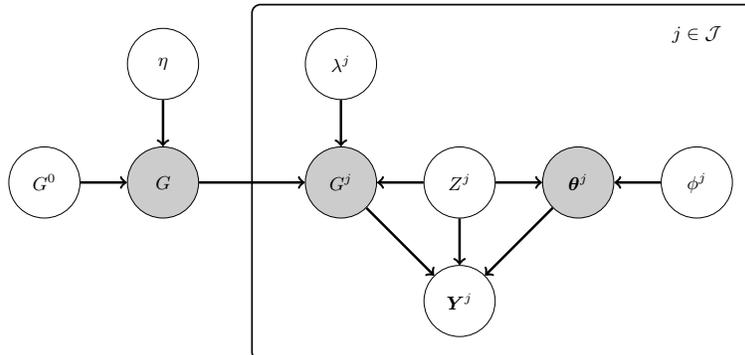


Fig 1: Joint network inference (JNI). A hierarchical model for analysis of multivariate data from multiple, nonidentical units or individuals, indexed by  $j$ . [Shaded nodes are unobserved.  $G^0$  = prior network,  $G$  = latent network,  $G^j$  = network specific to individual  $j$ ,  $\theta^j$  = parameters for individual  $j$ ,  $\mathbf{Y}^j$  = observables for individual  $j$ ,  $Z^j$  = ancillary information available on individual  $j$ ,  $\eta, \lambda^j$  = inverse temperature hyperparameters,  $\phi^j$  = hyperparameters defining a prior on  $\theta^j$ . Panel notation is used to indicate the presence of multiple individuals  $j \in \mathcal{J}$ .]

is defined to have edges that occur in exactly one of the networks  $G$ ,  $G^0$  (see also Ibrahim and Chen, 2000; Imoto *et al.*, 2003). The hyperparameter  $\eta$  controls the strength of the prior network  $G^0$  (Eqn. 1). Motivated by an application in cancer biology where prior structural information  $G^0$  is available, we follow Penfold *et al.* (2012) by restricting attention to SHD priors, however our statistical formulation is general (see below) and compatible with other penalty functionals. Alternatively, one could employ a beta-binomial prior as described in e.g. Dondelinger *et al.* (2012), that allows for the hyperparameters of the binomial to be integrated out. Note that in the latter case it is not possible to integrate specific prior structural information, making beta-binomial priors unsuitable for the application that this paper considers.

Given a latent network  $G$ , individual networks  $G^j$  are regularized in a similar way, as  $d^j(G^j, G) = \|G^j - G\|$ . In their work on combining multiple data sources, Werhli and Husmeier (2008) allow the  $\lambda^j$  to vary over individuals  $j \in \mathcal{J}$ . Likewise Penfold *et al.* (2012) learn the  $\lambda^j$  on a per-individual basis. However, in both studies, hyperparameter elicitation is nontrivial (see section 3.3). In the present paper, we consider only the special case where  $\lambda^1 = \lambda^2 = \dots = \lambda^J := \lambda$ .

A graph  $G \in \mathcal{G}$  can be characterized by (i) its adjacency matrix, or (ii) its

parent sets as  $G = (\pi_1, \dots, \pi_P)$  where  $\pi_p \subseteq \mathcal{P} = \{1 \dots P\}$  are the parents of vertex  $p$  in  $G$ . We write  $\mathcal{G}_p$  for the set of possible parent sets for  $p$ , such that formally  $\mathcal{G} = \mathcal{G}_1 \times \dots \times \mathcal{G}_P$ . Although we focus on SHD priors, the inference procedures presented in this paper apply to the more general class of “modular” priors, that may be factored over  $p \in \mathcal{P}$  and written in the form

$$(3) \quad d(G, G^0) = \sum_{p \in \mathcal{P}} d_p(\pi_p, \pi_p^0), \quad d^j(G^j, G; Z^j) = \sum_{p \in \mathcal{P}} d_p^j(\pi_p^j, \pi_p; Z^j)$$

for some functionals  $d_p, d_p^j : \mathcal{G}_p \times \mathcal{G}_p \rightarrow \mathbb{R}$ . Here  $\pi_p^0$  and  $\pi_p^j$  are parent sets for variable  $p$ , corresponding to  $G^0$  and  $G^j$  respectively.

In general inference for the JNI model (Eqns. 1,2) will be computationally intensive, as demonstrated in [Werhli and Husmeier \(2008\)](#); [Penfold et al. \(2012\)](#). In section 3 below we show that efficient, *exact* inference is nevertheless possible within the DBN class of graphical models.

**3. Joint network inference: DBNs.** The JNI model and network priors, as described above, are general. To apply the JNI framework in a particular context requires an appropriate likelihood at the individual level, that is, specification of the joint distribution  $p(\mathbf{Y}^j | G^j, \boldsymbol{\theta}^j, Z^j)$  of observables  $\mathbf{Y}^j$  given network  $G^j$ , ancillary information  $Z^j$  and parameters  $\boldsymbol{\theta}^j$ , together with a prior distribution  $p(\boldsymbol{\theta}^j | G^j, Z^j)$  over model parameters. We focus on time-course data, using DBNs and exploiting families of conjugate prior distributions. We show that factorization properties of the DBN likelihood permit computationally tractable joint inference and provide an explicit algorithm based on belief propagation.

**3.1. DBN formulation.** A DBN is a graphical model based on a DAG on the vertex set  $\mathcal{P} \times \mathcal{T}$  where  $\mathcal{T}$  is a set of time indices (Fig. 8(a); see [Murphy, 2002](#)). This DAG with  $PT$  vertices, is known as the “unrolled” DAG. Here, following [Hill et al. \(2012\)](#) and others, we use DBNs that permit only edges forwards in time and that are stationary in the sense that neither the network nor parameters change with time. For such DBNs, the network can be described by a directed graph  $G$  with exactly  $P$  vertices, with edges understood to go forward in time in the unrolled DAG (see Appendix B and Fig 8b). Note that  $G$  may have cycles. In what follows, all graphs (prior, latent and individual) describe the latter  $P$ -vertex representation.

Under a modular network prior, structural inference for DBNs can be carried out efficiently as described in [Hill et al. \(2012\)](#). In brief, the posterior  $G^j | \mathbf{y}$  factorizes into a product of local posteriors  $\pi_p^j | \mathbf{y}$ , one factor for each

target variable  $p$ . Background and assumptions for DBNs are summarized in Appendix B; for general background on DBNs we refer the interested reader to [Murphy \(2002\)](#) and for relevant details concerning the class of DBNs used here to [Hill et al. \(2012\)](#).

Write  $\mathbf{y}(t)$  for the matrix of observed data at time  $t$  for all individuals  $j$  and variables  $p$ . In order to simplify notation, we define a data-dependent functional

$$(4) \quad \mathfrak{P}(\mathbf{X}) = p(\mathbf{X}(1)) \prod_{t=2}^m p(\mathbf{X}(t) | \mathbf{y}(t-1))$$

that implicitly conditions upon observed history. Let  $y_p^j(t)$  denote the observed value of variable  $p$  in individual  $j$  at time  $t$ . The above notation allows us to conveniently summarize the product

$$(5) \quad p(y_p^j(1) | \pi_p^j) p(y_p^j(2) | \mathbf{y}(1), \pi_p^j) \dots p(y_p^j(m) | \mathbf{y}(m-1), \pi_p^j).$$

as  $\mathfrak{P}(\mathbf{y}_p^j | \pi_p^j)$ . Thus, we have that, for DBNs, the full likelihood also satisfies:

$$(6) \quad p(\mathbf{y} | G^1, \dots, G^J, Z^1, \dots, Z^J) = \prod_{j \in \mathcal{J}} \prod_{p \in \mathcal{P}} \mathfrak{P}(\mathbf{y}_p^j | \pi_p^j, Z^j)$$

where  $\mathbf{y}$  denotes the complete data (for all individuals, variables and times). In other words, the parent sets  $\pi_p^j$  for  $p \in \mathcal{P}$ ,  $j \in \mathcal{J}$  are mutually orthogonal in the Fisher sense, so that inference for each may be performed separately.

*3.2. Efficient, exact joint estimation.* We carry out exact inference in this setting using belief propagation ([Pearl, 1982](#)). Belief propagation is an iterative procedure in which messages are passed between variables in such a way as to compute exact marginal distributions; in this respect belief propagation belongs to a more general class of iterative algorithms known as “sum-product” algorithms ([Kschischang et al., 2001](#)). Our algorithm is summarized as follows (for simplicity we suppress dependence upon ancillary information  $Z^j$ ):

1. We begin by marginalizing over parameters  $\theta^j$  and caching the local scores  $\mathfrak{P}(\mathbf{y}_p^j | \pi_p^j)$  for all parent sets  $\pi_p^j \in \mathcal{G}_p$ , all variables  $p \in \mathcal{P}$  and all individuals  $j \in \mathcal{J}$ ; these could be obtained using any DBN likelihood. In this paper we exploited conjugate priors to obtain exact expressions for marginal likelihoods (Eqn. 33, see Appendix C for details).
2. Following marginalization the JNI graphical model collapses to the discrete Bayesian network shown in Fig. 2, whose nodes are themselves graphs.

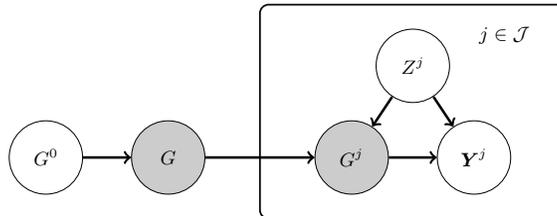


Fig 2: Marginalization of JNI over continuous (unknown) parameters  $\theta^j$ . [Shaded nodes are unobserved.  $G^0$  = prior network,  $G$  = latent network,  $G^j$  = network specific to individual  $j$ ,  $\mathbf{Y}^j$  = observables for individual  $j$ ,  $Z^j$  = ancillary information available on individual  $j$ . Hyperparameters  $\eta$ ,  $\lambda^j$ ,  $\phi^j$  are suppressed for clarity. Panel notation is used to indicate the presence of multiple individuals  $j \in \mathcal{J}$ .]

3. Posterior marginal distributions  $p(\pi_p | \mathbf{y}_p, \pi_p^0)$  and  $p(\pi_p^j | \mathbf{y}_p, \pi_p^0)$  are then computed using belief propagation on this discrete Bayesian network. Pseudocode for this step is provided in Algorithm 1 in Appendix D.

Let  $\mathbf{A}_{\text{JNI}}$  denote the  $P \times P$  matrix of marginal posterior inclusion probabilities for edges in the latent network  $G$ , i.e.  $(\mathbf{A}_{\text{JNI}})_{ip} := p(i \in \pi_p | \mathbf{y}, G^0)$ . These quantities are analogous to posterior inclusion probabilities in Bayesian variable selection and are computed, using Bayesian model averaging, as

$$(7) \quad (\mathbf{A}_{\text{JNI}})_{ip} = p(i \in \pi_p | \mathbf{y}, G^0) = \sum_{\pi_p \in \mathcal{G}_p} \mathbf{1}_{i \in \pi_p} p(\pi_p | \mathbf{y}, \pi_p^0)$$

where  $\mathbf{1}_A$  is the indicator of the event  $A$  and similarly for individual networks  $(\mathbf{A}_{\text{JNI}}^j)_{ip} := p(i \in \pi_p^j | \mathbf{y}, G^0)$ .

Following Hill *et al.* (2012) we reduced the space of parent sets  $\mathcal{G}_p$  using an in-degree sparsity restriction of the form  $|\pi_p^j| \leq c$  for all  $\pi_p^j \in \mathcal{G}_p$ ,  $p \in \mathcal{P}$ ,  $j \in \mathcal{J}$ . Thus the cardinality of the space of parent sets  $|\mathcal{G}_p| = \mathcal{O}(P^c)$  is polynomial in  $P$ , where it was previously super-exponential.

Caching of selected probabilities is used to avoid redundant recalculation. Pseudocode is provided in Algorithm 1 in Appendix D, consisting of three phases of computation. Storage costs are dominated by Phases I and II, each requiring the caching of  $\mathcal{O}(JP^{1+c})$  terms. Phase II dominates computational effort, with total (serial) algorithmic complexity  $\mathcal{O}(J^2P^{1+2c})$ . However, within-phase computation is “embarrassingly parallel” in the sense that all calculations are independent (indicated by square parentheses notation in the pseudocode). In practice we have found that problems of size  $P \leq 20$ ,  $J \leq 20$ ,  $c \leq 3$  can be solved within minutes using serial computation on a

standard laptop computer. We provide serial and parallel MATLAB R2014a implementations in Supplement B.

**3.3. Network prior elicitation.** Elicitation of hyperparameters for network priors is an important and nontrivial issue. Here we specify the hyperparameters  $\lambda, \eta$  in a subjective manner. We do so due to reported difficulties in estimation of hyperparameters for related models (Werhli and Husmeier, 2008; Dondelinger et al., 2012; Penfold et al., 2012). We present three criteria below that, for the special case of SHD, are simple to implement and can be used for expert elicitation. These heuristics seek to relate the hyperparameters to more directly interpretable measures of the similarity and difference that they induce between prior, latent and individual networks: (i) Firstly, we note the following formula for the probability of maintaining the status (present/absent) of a candidate parent  $i \in \mathcal{P}$  between the latent network  $G$  and an individual network  $G^j$ :

$$(8) \quad h_\lambda := p(i \notin \pi_p^j \Delta \pi_p) = \frac{1}{1 + e^{-\lambda}}.$$

This probability provides an interpretable way to consider the influence of  $\lambda$ . For example a prior confidence of  $h_\lambda \approx 0.73$  that a given edge status in  $G$  is preserved in a particular individual  $G^j$  translates into an odds ratio  $h_\lambda/(1 - h_\lambda) \approx 2.7$  and a hyperparameter  $\lambda \approx 1$  (see SFig. 1). An analogous equation relates  $\eta$  and  $h_\eta := p(i \notin \pi_p \Delta \pi_p^0)$ , allowing prior strength to be set in terms of the probability that an edge status in the prior network  $G^0$  is maintained in the latent network  $G$ . (ii) A second, related approach is to consider the expected total SHD between an individual network  $G^j$  and the latent network  $G$ :

$$(9) \quad \mathbb{E}(\|G^j - G\|) = P^2(1 - h_\lambda)$$

This can be interpreted as the average number of edge changes needed to obtain  $G^j$  from  $G$ . An analogous equation holds for  $\eta$  and  $h_\eta$ . (iii) Thirdly, in certain applications, the latent network  $G$  may not have a direct scientific interpretation, in which case the criteria presented above may be unintuitive. Then, hyperparameters can be elicited by consideration of (a) similarity between individual networks  $G^j, G^k$ , and (b) concordance of individual networks  $G^j$  with the prior network  $G^0$  (see Supplement A for further discussion).

**3.4. An information sharing bound.** Below we consider the extent to which information can be shared between individuals within JN1, providing an upper bound that is attained as the number of individuals  $J$  grows

large. To formalize the contribution to inference from information sharing we consider the case in which no data is available on a specific individual (without loss of generality, individual  $j = 1$ ) and analytically quantify the extent to which JNI can estimate the true network  $\overline{G}^1$  by “borrowing strength” from the data  $\mathbf{Y}^2, \dots, \mathbf{Y}^J$  that represent observations on the remaining individuals. (Over-lines will be used to signify the “true” data-generating networks.) As a baseline, write  $\mathbf{A}_0^j = p(i \in \pi_p^j | \mathbf{Y}^j)$  for the (naive) estimator that prohibits the sharing of information between individuals. For simplicity we restrict attention to the case where no network prior is used ( $\eta = 0$ ), the data-generating hyperparameter  $\lambda$  is known and in-degree restrictions are not in place ( $c = P$ ). Then, with neither data nor prior information available on individual 1, it trivially follows that

$$(10) \quad \mathbb{E}_{\mathbf{Y}, \overline{G}, \overline{G}^1, \dots, \overline{G}^J | \eta, \lambda} \left[ \frac{\|\mathbf{A}_0^1 - \overline{G}^1\|}{P^2} \right] = \frac{1}{2}$$

where the expectation is taken over all possible data-generating networks and corresponding data.

From standard, independent network inference we know that consistent estimation requires unbiasedness of the likelihood function  $p(\mathbf{Y}^j | G^j)$ , in the sense that  $\mathbb{E}_{\mathbf{Y}^j | \overline{G}^j} p(\mathbf{Y}^j | G^j)$  is maximized by  $G^j = \overline{G}^j$ . We therefore begin by constructing the analogous regularity condition for joint estimation: Write  $\mathbf{R}$  for the matrix that encodes the prior metric on  $\mathcal{G}$  as  $(\mathbf{R})_{G, G'} = \exp(-\lambda \|G - G'\|) / C(\lambda)$  where  $C(\lambda) = \sum_{G \in \mathcal{G}} \exp(-\lambda \|G\|)$ . Write  $\mathbf{S}$  for the matrix of expected Bayesian scores  $(\mathbf{S})_{G^j, \overline{G}^j} = \mathbb{E}_{\mathbf{Y}^j | \overline{G}^j} p(\mathbf{Y}^j | G^j)$ .

**ASSUMPTION (Joint regularity).** *For each column of the matrix  $\mathbf{M} = (\mathbf{R}\mathbf{S}\mathbf{R})_{G, \overline{G}}$ , the non-diagonal entries are strictly smaller than the diagonal entry. i.e.  $M_{G, \overline{G}} < M_{\overline{G}, \overline{G}}$  for all  $G \neq \overline{G}$ .*

To gain intuition for the joint regularity assumption, consider the special case where  $\lambda \rightarrow \infty$ ; here  $\mathbf{R} = \mathbf{I}$  and we only require that the expected local Bayesian score  $(\mathbf{S})_{G^j, \overline{G}^j}$  is maximized by  $G^j = \overline{G}^j$ ; i.e. we recover the unbiasedness condition from standard network inference.

**THEOREM.** *Under the joint regularity assumption, there exists  $0 < \epsilon < 1$  such that*

$$(11) \quad \mathbb{E}_{\mathbf{Y}, \overline{G}, \overline{G}^1, \dots, \overline{G}^J | \eta, \lambda} \left[ \frac{\|\mathbf{A}_{\text{JNI}}^1 - \overline{G}^1\|}{P^2} \right] = f(J) + \frac{1}{1 + e^\lambda}$$

where  $f(J) \leq 2P^2 \epsilon^{J-1} \rightarrow 0$  as  $J \rightarrow \infty$ .

PROOF. See Appendix A. □

Comparing Eqn. 11 to Eqn. 10 we see that information sharing offers gains in estimation, agreeing with intuition, with larger gains when the true networks are almost homogeneous ( $\lambda$  large). Moreover the statistical power of JNI to estimate  $\overline{G^1}$  converges to its maximum exponentially quickly as  $J \rightarrow \infty$ .

**4. Results.** The proposed methodology was compared against several existing network inference algorithms. We restricted attention to methods that are compatible with time course data and, following the majority of the literature, carry out estimation for each individual separately. The computational demands of Niculescu-Mizil and Caruana (2007); Werhli and Husmeier (2008); Penfold *et al.* (2012) precluded application in this setting. Specifically, in the simulated data examples we report below, over 3000 rounds of inference were performed in total, on problems larger than DREAM4 ( $P = 10$ ,  $J = 5$ ). Using the approach of Penfold *et al.* (2012), these experiments would have required more than 10 years serial computational time; in contrast our approach required less than 24 hours serial computation on a standard laptop. Thus, we consider the following methods:

- (i) *DBN*. A dynamic Bayesian network, as described in Hill *et al.* (2012), including nonlinear interaction terms. For this choice of model it is possible to construct a fully conjugate set of priors, delivering a closed form expression for the local Bayesian score  $\mathfrak{P}(\mathbf{y}_p^j | \pi_p^j, Z^j)$ . The model is summarized in Appendix B.
- (ii) *IDBN*. Spencer and Mukherjee (2012) recently proposed an extension of Hill *et al.* (2012) that allows analysis of datasets that contain interventions; this is outlined in Appendix B. Interventional DBNs (IDBNs) inherit the computational advantages of DBNs, in the sense that there is a closed form expression for the local Bayesian score, but extend DBNs in a causal direction. We considered two alternative implementations of IDBNs: (i) *IDBN*. The approach of Spencer and Mukherjee (2012) was applied to each individual separately. (ii) *Mono IDBN*. Data on all individuals were pooled together and fed into a single IDBN analysis, an approach that Werhli and Husmeier (2008) described as “monolithic”.
- (iii) *Rel Nets*. A popular approach within the bioinformatics community is to score edges based on Pearson correlation of participating nodes (“relevance networks”; see e.g. Butte *et al.*, 2000). Here, we used a time-course analogue in which the correlation is calculated between

successive time points.

- (iv) *LASSO*. An  $\ell_1$ -penalized likelihood was used to obtain estimates for coefficients in a linear autoregressive model. Coefficients were estimated for each variable independently, taking each variable in turn as the response. The penalty parameters  $\lambda_p$  were each selected using leave-one-out cross validation. Non-zero coefficients indicated presence of edges. Further details appear in Supplement A.

Note that DBN and IDBN are able to integrate a prior network  $G^0$ , whereas Rel Nets and LASSO are not. JNI facilitates joint estimation given a suitable graphical model likelihood. We applied JNI to the DBN and IDBN models described above. This resulted in several proposed estimators:

- (v) *J-DBN*. JNI applied to DBN.
- (vi) *J-IDBN*. JNI applied to IDBN.
- (vii) *Fixed IDBN*. Here we formed the likelihood assuming a single graph for all individuals and the latent network (i.e.  $G^1 = \dots = G^J = G$ ) but with parameters allowed to differ. This can be considered a joint analogue of Mono IDBN that allows individual-specific parameter values.
- (viii) *AJ-IDBN*. A computationally efficient approximation to J-IDBN, in which the latent network topology is first estimated using Fixed IDBN. This is in turn used as an informative network prior within  $J$  independent rounds of IDBN. In this way information sharing is allowed to occur, but at the expense of a coherent joint posterior.

In the empirical study below we compare JNI variants (v-viii) against existing methods (i)-(iv).

4.1. *Performance metrics*. The proposed methodology addresses three questions, some or all of which may be of scientific interest depending on the application; (i) estimation of the latent network  $G$ , (ii) estimation of individual networks  $G^1, \dots, G^J$ , and (iii) estimation of differences between individual networks (“differential networks”; [Ideker and Krogan, 2012](#)). We quantify performance for each task using the area under the receiver operating characteristic (ROC) curve (AUR). This metric, equivalent to the probability that a randomly chosen true edge is preferred by the inference scheme to a randomly chosen false edge, summarizes, across a range of thresholds, the ability to select edges in the data-generating network. AUR may be computed relative to the true latent network  $G$ , or relative to the true individual networks  $G^j$ , quantifying performance on tasks (i) and (ii) respectively. Both sets of results are presented below, in the latter case averaging

AUR over all individual networks. For (iii), in order to assess ability to estimate differential networks, we computed AUR scores based on the statistics  $F_{ip}^j = |p(i \in \pi_p^j | \mathbf{y}, G^0, Z^j) - p(i \in \pi_p | \mathbf{y}, G^0, Z^1, \dots, Z^J)|$  that should be close to one if  $i \in \pi_p^j \Delta \pi_p$ , otherwise  $F_{ip}^j$  should be close to zero.

It is easy to show that inference for the latent network, under only the prior (i.e.  $\hat{G} = G^0$ ), attains mean AUR equal to  $h_\eta$ . Similarly, prior inference for the individual networks (i.e.  $\hat{G}^j = G^0$ ) attains mean AUR equal to  $1 - h_\eta - h_\lambda + 2h_\eta h_\lambda$ . This provides a baseline for the proposed methodology at tasks (i) and (ii) and allows performance to be decomposed into AUR due to prior knowledge and AUR contributed through inference.

Using a systematic variation of data-generating parameters, we defined 15 distinct data generating regimes described below. For all 15 regimes we considered 50 independent datasets; standard errors accompany average AUR scores. Results presented below use a computationally favorable in-degree restriction  $c = 3$ . In order to check robustness to  $c$ , a subset of experiments were repeated using  $c = 4$ , with close agreement observed (SFig. 4).

#### 4.2. Simulation study.

4.2.1. *Data generation.* Data were generated according to DBN models (Appendix B) as described in detail in Supplement A. This data-generating scheme was extended to mimic interventional experiments that are a feature of our application to breast cancer. In this case, for each time course, a randomly chosen variable is marked as the target of an interventional treatment. Data were then generated according to the augmented likelihood described in Appendix B (fixed effects were taken to be zero).

4.2.2. *Model misspecification and nonlinear data generating models.* We assume exchangeable networks; it is therefore interesting to explore the performance of the proposed estimators when the assumption of exchangeability is violated. Specifically, we consider a “worst case” scenario where individual networks  $G^1, \dots, G^J$  are sampled from a mixture model with two distinct components. Moreover we consider the extreme case where networks in distinct mixture components share only a few edges in common; it is expected that exchangeable estimators will exhibit poor performance in this scenario. Further, in order to investigate the impact of model mis-specification at the level of the time-series model itself, we considered time course data generated from a computational model of protein signaling, based on nonlinear ODEs (Xu *et al.*, 2010). In order to extend this model, which is for a single cell type, to simulate a heterogeneous population, we selected three pro-

tein species per individual (at random) and deleted their outgoing edges to obtain the data-generating networks  $G^j$  (see Supplement A).

4.2.3. *Estimator performance.* We consider the three estimation tasks:

**Latent network:** We investigated ability to recover the latent network  $G$ . The existing approaches (i)-(iv) estimate only individual-specific networks. For estimation of the latent, shared network using these methods we simply took an unweighted average of the  $J$  estimated adjacency matrices. The proposed joint estimators (v)-(viii) were assigned hyperparameter values  $\eta = 1, \lambda = 2$  ( $\lambda = 3$  for Xu *et al.*, 2010) based on the heuristic of Eqn. 8; sensitivity to misspecified hyperparameter values is investigated later in section 4.2.4. Results based on simulated data with interventions are displayed in STable 3. We found little difference in the ability of J-IDBN, Fixed IDBN and AJ-IDBN to recover the latent network structure across a wide range of regimes, though J-IDBN achieved best performance in 9 out of 15 regimes. Interestingly we found that the IDBN estimator, which performs an unweighted average of  $J$  independent inferences, performed significantly worse than each of J-IDBN, Fixed-IDBN and AJ-IDBN in respectively 15, 13 and 11 out of 15 regimes. Similarly, all above approaches clearly outperformed Mono IDBN and Rel Nets, which were in turn outperformed by inference based on the prior alone, demonstrating the importance of accounting for individual-specific parameter values. The joint formulation of DBNs (J-DBN) significantly outperformed standard DBNs, with higher AUR in all 15 regimes. LASSO performed best in the regime with long time series ( $n = 10$ ) but failed in other regimes to outperform inference based on the prior alone. We obtained qualitatively similar results for both alternative data-generating schemes (STables 4-5).

**Individual networks:** At this task, J-IDBN outperformed all other approaches in 9 out of 15 regimes. AJ-IDBN offered a similar level of performance and together these estimators demonstrated better performance compared to alternatives in 13 out of 15 regimes. Since AJ-IDBN avoids intensive computation, this may provide a practical estimator of individual networks in higher dimensional settings. Again, the joint approaches J-IDBN and J-DBN both outperformed the standard approaches IDBN and DBN respectively, demonstrating an increase in statistical power resulting from the proposed methodology. Rel Nets and LASSO performed poorly at this task. Similar results were observed using the alternative data-generating schemes (STables 4-5).

**Differential networks:** Since JNI regularizes between individuals we sought to test whether it could eliminate spurious differences and thereby improve

estimation of differential networks. Differential networks may also be estimated using existing methods (i)-(iv); to do so in each case we compared individual network estimates with the estimate of the latent network obtained as described in section 4.2.3 above. We found that, whilst estimation of differential networks appears to be more challenging than the other tasks, J-IDBN outperformed the other approaches in 7 out of 15 regimes. Moreover the J-IDBN and J-DBN methods outperformed IDBN and DBN respectively in all 15 regimes. These results suggest that coherence of joint analysis aids in suppressing spurious features for estimation of differential network topology. Rel Nets performed poorly at this task and LASSO performed slightly better. Intriguingly, AJ-IDBN performed well in estimating differential networks, performing best in 7 out of 15 regimes. This suggests that the approximate joint estimator may be suited to estimation of differential networks. Results on the non-interventional datasets supported this conclusion (STable 4). On the Xu *et al.* (2010) datasets, however, IDBN and Rel Nets were among the best performing estimators (STable 5), despite being misspecified for the nonlinear data-generating model.

4.2.4. *Robustness.* We assess three aspects of robustness:

**Hyperparameter misspecification:** For the above investigation we used Eqn. 8 to elicit hyperparameters  $\eta, \lambda$ . This was possible because the data-generating parameters were known by design; however in general this will not be the case. We therefore sought to empirically investigate the effect of hyperparameter misspecification. SFig. 3 displays how performance of the J-IDBN estimator for latent networks depends on the choice of hyperparameters  $\lambda, \eta$ . Performance does not appear to be highly sensitive to the precise hyperparameter values used and there is a large region in which AUR remains high.

**Outliers and batch effects:** The biological datasets that motivate this study often contain outliers. At the same time, experimental design may lead to batch effects. In order to probe estimator robustness, we generated data as described above, with the addition of outliers and certain batch effects. Specifically, Gaussian noise from the contamination model  $0.95\mathcal{N}(0, 0.1^2) + 0.05\mathcal{N}(0, 10^2)$  was added to all data prior to inference. At the same time, one individual’s data were replaced entirely by Gaussian white noise to simulate a (strong) batch effect that could arise e.g. if preparation of a specific biological sample was incorrect. The relative decrease in performance at feature detection is reported in SFig. 5. We found that J-IDBN remained the best-performing estimator for all three estimation problems. However, for the differential network estimation task in particular the

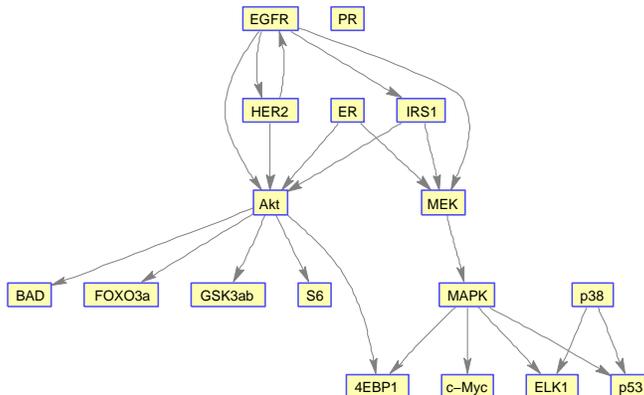


Fig 3: Signaling downstream of the epidermal growth factor receptor (EGFR). The graph shown summarizes known causal links characterized by extensive biochemistry. [Note that edges in the graph represent high-level summaries of often complex molecular interactions that may involve latent chemical species.]

decrease in performance was pronounced for joint methods.

**Nonexchangeability:** SFig. 6 displays the result of inference on data where the exchangeability assumption is violated. It can be seen that the performance of all (exchangeable) estimators decreases in these circumstances, but the magnitude of the decrease is small (e.g. for estimation of individual networks, J-IDBN experiences a 0.01 decrease in AUR). We note that the proposed estimators can be extended to nonexchangeable settings where elements of the structure that relates individuals are known; see [Oates and Mukherjee \(2014\)](#) for further details.

4.3. *Protein signaling networks in breast cancer.* We consider experimental data derived from human breast cancer cell lines, focusing on protein signaling networks for which a substantial proportion of wild type network topology has been characterized by extensive biochemistry (Fig. 3). The investigation presented below serves three purposes: Firstly, it allows us to investigate the applicability of the proposed joint approaches to experimental data. Secondly, it allows us to investigate the use of ancillary information, in the form of mutational status and histological information. Finally, the results and approach are relevant to the topical question of exploring signaling heterogeneity across cancer cell lines.

Data were obtained using reverse-phase protein arrays ([Hennessy et al., 2010](#)) from  $J = 6$  breast cancer cell lines (AU 565, HCC 1569, MCF 7,

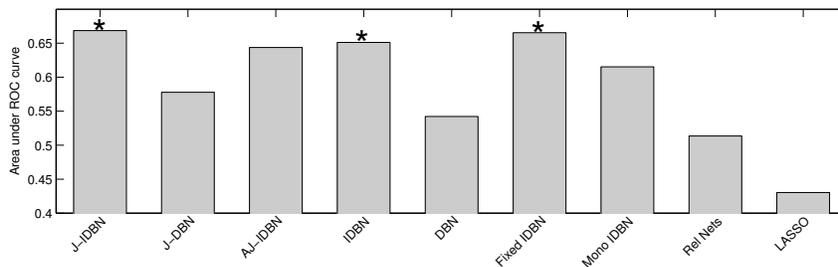


Fig 4: Results from breast cancer cell line data, comparison with network based on literature. The methods shown were used to estimate a latent network; AUR is with respect to the literature-based network shown in Fig. 3; the latter was not used to provide prior information in these experiments. [Asterisks denote AUR scores which were significant at the 1% level under a permutation test with AUR as the statistic and 10,000 samples used to obtain an empirical null distribution.]

MDA MB 231, SKBR3 and SUM 190PT; experimental protocol is described in brief in Supplement A). Data comprised observations for the  $P = 17$  proteins shown in Fig. 3 (see also STable 1; these data form part of a larger, ongoing study including further cell lines and proteins). Specifically,  $\mathbf{y}$  contains the logarithms of the measured concentrations. Data were acquired under treatment with an EGFR/HER2 inhibitor Lapatinib (“EGFRi”), an Akt inhibitor (“Akti”), EGFRi and Akti in combination, and without inhibition (“DMSO”) at 0.5,1,2,4,8 and 24 hours following Serum stimulation, giving a total of  $n_j = 24$  observations of each variable in each individual cell line.

4.3.1. *Ancillary information.* For the cancer cell lines analyzed here, ancillary information is available in the form of genetic aberrations (mutation statuses) and histological profiling. These were obtained from published sources (Neve *et al.*, 2006) and online databases (Forbes *et al.*, 2011) and reproduced in STable 2. We integrated this information into a prior; in brief we considered two main factors: (i) Loss-of-function mutations in kinase domains; these induced zero prior probability on edges emanating from the mutant protein. Where the mutation also affects the ability of a protein to be phosphorylated, then incoming edges were also assigned zero prior probability. (ii) Cell lines with ectopic expression of the receptor HER2 are known to depend heavily upon EGFR signaling. In this case the network prior did not penalize edges emanating from the EGFR receptor nodes. A

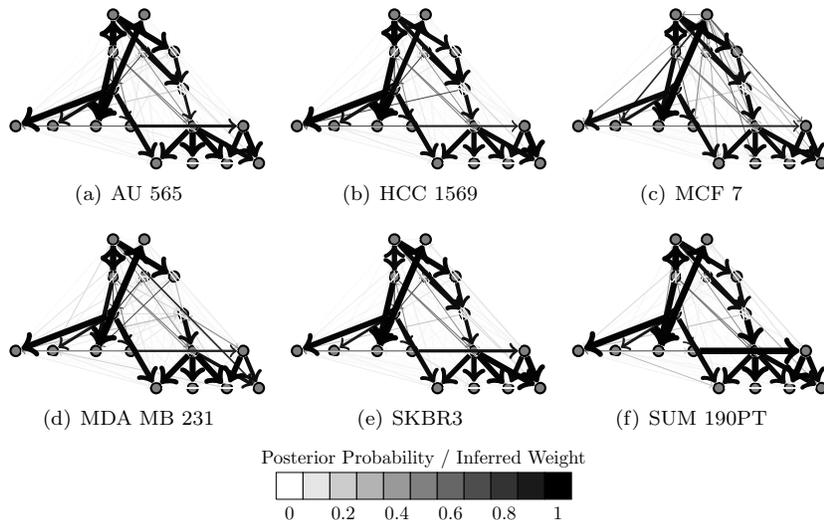


Fig 5: Breast cancer data; cell line-specific networks inferred by J-IDBN. [Edge width and color are proportional to posterior marginal inclusion probabilities. The layout of vertices is congruent to Fig. 3, which can be used as a key.]

full discussion of ancillary data appears in Supplement A.

Extensive biochemistry on normal cell lines has provided a wealth of knowledge concerning causal links between the signaling proteins considered here. We used this information to specify the prior graph  $G^0$  (shown in Fig. 3). However, networks specific to individual breast cancer cell lines remain comparatively unexplored such that most of our prior knowledge on cell lines derives from assumed similarity with  $G^0$ . Whilst cancer signaling may differ with respect to wild type signaling, we expect the differences to be small in number. In light of these observations, we used subjective elicitation (section 3.3) to select hyperparameters  $\lambda = 4, \eta = 5$ , corresponding to  $\mathbb{E}(\|G^j - G\|) \approx 5$ ,  $\mathbb{E}(\|G - G^0\|) \approx 2$ .

4.3.2. *Validation of estimators.* In order to test estimator performance, we first considered the latent network  $G$ , comparing estimates to the literature network shown in Fig. 3. For a fair assessment we used an empty prior network  $G^0$ . Inferred networks are displayed in SFig. 7. Results demonstrated good recovery of the literature causal network, with J-IDBN attaining the highest AUR (0.67,  $p < 0.01$ , Fig. 4). As in the simulation study, J-IDBN outperformed IDBN, with AJ-IDBN and Fixed IDBN representing

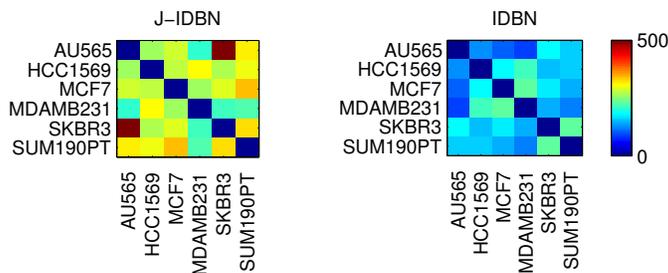


Fig 6: Breast cancer data; pairwise similarity between cell-line-specific networks inferred by J-IDBN (left) and IDBN (right). J-IDBN identifies AU 565 and SKBR 3 as having the most similar networks; these cell lines were originally derived from the same patient. In contrast, IDBN does not do so. [Colors denote Bonferroni  $-\log(p)$  values based on the Pearson correlation coefficient of posterior inclusion probabilities for pairs of cell lines, so that red indicates a high degree of similarity. For presentation the diagonal is set to zero.]

good alternative estimators and the remaining estimators performing poorly. This suggests the conclusions drawn in section 4.2 apply also to the analysis of biological time series data. In particular, modeling of interventions appears to be crucial in this setting, in line with the conclusions of [Spencer and Mukherjee \(2012\)](#).

*4.3.3. Inference for cell line networks.* We investigated inference for cell line specific networks  $G^j$  (Fig. 5), taking the prior network  $G^0$  from literature (Fig. 3). In order to assess results, we exploited the fact that cell lines AU565 and SKBR3 derive from the same patient. We would therefore expect these two cell lines to be most similar at the network level. J-IDBN networks for AU565 and SKBR3 were indeed the most similar, maximizing the Pearson correlation coefficient between corresponding posterior marginal inclusion probabilities over all  $\binom{6}{2} = 15$  pairs of cell lines. In contrast standard IDBNs did not do so (Fig. 6). Fig. 7 compares posterior inclusion probabilities (or analogous edge weights for the non-Bayesian methods) for AU565 against SKBR3. We find posterior edge probabilities from these two lines are closer under JNI estimators compared with standard, independent estimators. However, a thorough assessment of the accuracy of the individual networks requires additional experimental work and is beyond the scope of this paper.

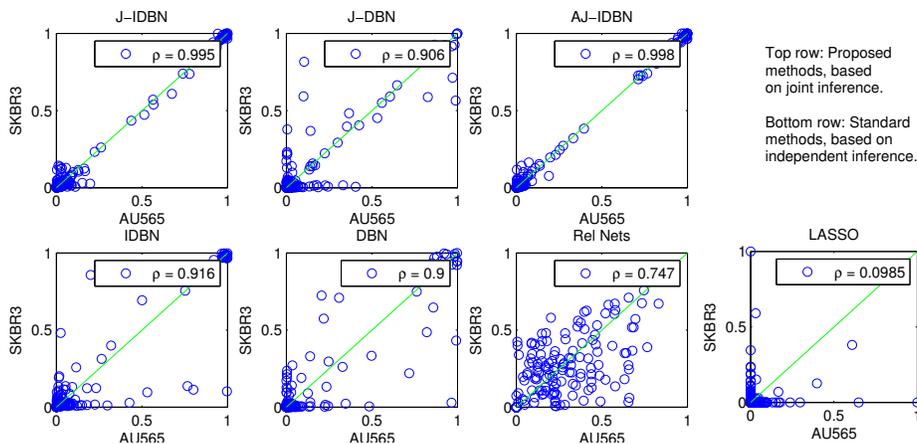


Fig 7: Comparison of posterior edge probabilities obtained from analysis of data from two breast cancer cell lines (AU 565 and SKBR 3) that were originally derived from the same patient. The joint estimators J-IDBN and J-DBN improve the Spearman correlation coefficient (“rho”) between posterior edge probabilities compared to independent inference using IDBN and DBN.

**5. Discussion.** We focused on three related structure learning problems arising in the context of a set of nonidentical but exchangeable units or individuals:

1. Estimation of a shared network from the heterogeneous data.
2. Estimation of networks for specific individuals.
3. Learning features specific to individuals (“differential networks”).

Each problem may be of independent scientific interest; the joint approaches investigated here address all three problems simultaneously within a coherent statistical framework. We considered simulated data, with and without model misspecification, as well as proteomic data obtained from cancer cell lines. For all three problems we demonstrated that a joint analysis performs at least as well as independent or simpler aggregate analyses.

We considered modular priors (that factorise over nodes) in order to facilitate efficient computation. However, it may be useful to consider richer priors for joint estimation. One possibility that is pertinent to applications in cancer biology would be hierarchical regularization that allows entire pathways to be either active or inactive. However, we note that this would require revisiting hyperparameter elicitation since the heuristics we described are specific to SHD priors. We restricted the joint model to have equal inverse

temperatures  $\lambda^1 = \dots = \lambda^J := \lambda$ . Relaxing this assumption may improve robustness to batch effects that target single individuals, since then weak informativeness ( $\lambda^j \approx 0$ ) may be learned from data. It would also be interesting to distinguish between  $G \setminus G^j$  (“loss of function”) and  $G^j \setminus G$  (“gain of function”) features. In this work we did not explore information sharing through parameter values  $\theta^j$ , yet this may yield more powerful estimators of network structure in settings where individuals’ parameters  $\theta^j, \theta^k$  are not independent.

The case of exchangeable networks that we considered here represents the simplest of a more general class of models for related networks. For DBNs, the computationally efficient joint estimation we propose could be extended to more complex settings. Oates and Mukherjee (2014) discuss the case where multiple individuals are related according to a known tree structure. In this more general setting, efficient algorithms based on belief propagation continue to apply, since the tree constraint ensures that the corresponding factor graph is acyclic and so the sum-product lemma continues to hold (Kschischang *et al.*, 2001). However prior elicitation is complicated by the requirement that each edge of the tree induces an additional hyperparameter that must be specified. Prior specification in these circumstances can be extremely challenging and requires correspondingly different methodology that are distinct from the present paper. Oates and Mukherjee (2014) proposes strategies that avoid this elicitation problem, including a subset constraint that edges in individual networks are inherited along the edges of the tree. In the future, joint learning of networks and the hierarchical structure that relate them to one another may also be possible.

**Acknowledgements.** We are grateful to the Editor and anonymous referees for feedback that has improved the content and presentation of this paper. We would also like to thank J.D. Aston, F. Dondelinger, C.A. Penfold, S.E.F. Spencer and S.M. Hill for helpful discussion and comments. Financial support was provided by NCI U54 CA112970, UK EPSRC EP/E501311/1 and EP/D002060/1, and the Cancer Systems Biology Center grant from the Netherlands Organisation for Scientific Research.

**Appendix A: Proof of Theorem.** The following Lemma shows that, under the joint regularity assumption, JN1 is a consistent estimator of the true latent network  $\bar{G}$  in the limit  $J \rightarrow \infty$ :

LEMMA. *Let  $\eta = 0$ . Then under the joint regularity assumption there exists  $0 < \epsilon < 1$  such that  $\mathbb{E}_{\mathbf{Y}, \bar{G}, \bar{G}^1, \dots, \bar{G}^J | \eta, \lambda} p(\bar{G} | \mathbf{Y}) > 1 - |\mathcal{G}| \epsilon^J$ .*

PROOF. Since we are using a flat prior ( $\eta = 0$ ) on  $G$  we have, suppressing dependence upon  $\lambda$ ,

$$(12) \quad p(\bar{G}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\bar{G})}{\sum_{G \in \mathcal{G}} p(\mathbf{Y}|G)}$$

so from Jensen's inequality

$$(13) \quad \mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\bar{G}|\mathbf{Y}) \geq \frac{\mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|\bar{G})}{\sum_{G \in \mathcal{G}} \mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|G)}$$

$$(14) \quad = \left[ 1 + \sum_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} \frac{\mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|G)}{\mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|\bar{G})} \right]^{-1}$$

$$(15) \quad > 1 - \sum_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} \frac{\mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|G)}{\mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\mathbf{Y}|\bar{G})}$$

$$(16) \quad = 1 - \sum_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} \prod_{j \in \mathcal{J}} \frac{\mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|G)}{\mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|\bar{G})}.$$

The joint regularity assumption is equivalent to the requirement that  $\mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|G)$  has a unique maximum at  $G = \bar{G}$ , since

$$(17) \quad \mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|G) = \mathbb{E}_{\bar{G}^j | \bar{G}, \lambda} \mathbb{E}_{\mathbf{Y}^j | \bar{G}^j} \sum_{G^j \in \mathcal{G}} p(\mathbf{Y}^j|G^j) p(G^j|G)$$

$$(18) \quad = \sum_{\bar{G}^j \in \mathcal{G}} p(G^j|G) \sum_{G^j \in \mathcal{G}} [\mathbb{E}_{\mathbf{Y}^j | \bar{G}^j} p(\mathbf{Y}^j|G^j)] p(\bar{G}^j | \bar{G})$$

$$(19) \quad = \sum_{\bar{G}^j \in \mathcal{G}} \sum_{G^j \in \mathcal{G}} (\mathbf{R}^T)_{G, G^j} (\mathbf{S})_{G^j, \bar{G}^j} (\mathbf{R})_{\bar{G}^j, \bar{G}}$$

$$(20) \quad = (\mathbf{R}^T \mathbf{S} \mathbf{R})_{G, \bar{G}} = (\mathbf{R} \mathbf{S} \mathbf{R})_{G, \bar{G}}$$

where we have used that  $\mathbf{R}$  is symmetric. It follows that

$$(21) \quad \epsilon := \max_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} \frac{\mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|G)}{\mathbb{E}_{\mathbf{Y}^j, \bar{G}^j | \bar{G}, \lambda} p(\mathbf{Y}^j|\bar{G})} < 1.$$

We therefore conclude that

$$(22) \quad \mathbb{E}_{\mathbf{Y}, \bar{G}^1, \dots, \bar{G}^J | \bar{G}, \lambda} p(\bar{G}|\mathbf{Y}) > 1 - |\mathcal{G}| \epsilon^J.$$

Since Eqn. 22 is independent of  $\bar{G}$ , the result follows.  $\square$

PROOF OF THEOREM. Since no observables are available on the first individual ( $\mathbf{Y}^1 = \emptyset$ ) we have

$$(23) \quad \mathbf{A}_{\text{JNI}}^1 = \sum_{G \in \mathcal{G}} p(G|\mathbf{Y}) \sum_{G^1 \in \mathcal{G}} p(G^1|G)G^1.$$

We also require the ‘‘oracle’’ estimator (O-JNI); this is simply JNI but with  $\bar{G}$  fixed and known. i.e.

$$(24) \quad \mathbf{A}_{\text{O-JNI}}^1 = \sum_{G^1 \in \mathcal{G}} p(G^1|\bar{G})G^1.$$

Note that  $\mathbb{E}_{\bar{G}|\eta,\lambda} \|\mathbf{A}_{\text{O-JNI}}^1 - \bar{G}^1\| = \mathbb{E}_{\bar{G}^1, \bar{G}|\lambda} \|\bar{G} - \bar{G}^1\| = P^2(1 - h_\lambda)$ . We begin by showing that JNI approximates O-JNI:

$$(25) \quad \begin{aligned} \mathbf{A}_{\text{O-JNI}}^1 - \mathbf{A}_{\text{JNI}}^1 &= (1 - p(\bar{G}|\mathbf{Y})) \sum_{G^1 \in \mathcal{G}} p(G^1|\bar{G})G^1 \\ &\quad - \sum_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} p(G|\mathbf{Y}) \sum_{G^1 \in \mathcal{G}} p(G^1|G)G^1 \end{aligned}$$

and by the triangle inequality

$$(26) \quad \begin{aligned} \|\mathbf{A}_{\text{O-JNI}}^1 - \mathbf{A}_{\text{JNI}}^1\| &\leq \left\| (1 - p(\bar{G}|\mathbf{Y})) \sum_{G^1 \in \mathcal{G}} p(G^1|\bar{G})G^1 \right\| \\ &\quad + \left\| \sum_{\substack{G \in \mathcal{G} \\ G \neq \bar{G}}} p(G|\mathbf{Y}) \sum_{G^1 \in \mathcal{G}} p(G^1|G)G^1 \right\| \end{aligned}$$

$$(27) \quad \leq (1 - p(\bar{G}|\mathbf{Y})) \sup_{G^1 \in \mathcal{G}} \|G^1\| + (1 - p(\bar{G}|\mathbf{Y})) \sup_{G^1 \in \mathcal{G}} \|G^1\|$$

$$(28) \quad \leq 2(1 - p(\bar{G}|\mathbf{Y}))P^2.$$

Again, by the triangle inequality

$$(29) \quad \|\mathbf{A}_{\text{JNI}}^1 - \bar{G}^1\| \leq \|\mathbf{A}_{\text{JNI}}^1 - \mathbf{A}_{\text{O-JNI}}^1\| + \|\mathbf{A}_{\text{O-JNI}}^1 - \bar{G}^1\|.$$

Taking expectations and applying the Lemma produces

$$(30) \quad \mathbb{E}_{\mathbf{Y}, \bar{G}|\eta,\lambda} \|\mathbf{A}_{\text{JNI}}^1 - \bar{G}^1\| \leq 2P^2|\mathcal{G}|\epsilon^{J-1} + P^2(1 - h_\lambda)$$

as required.  $\square$

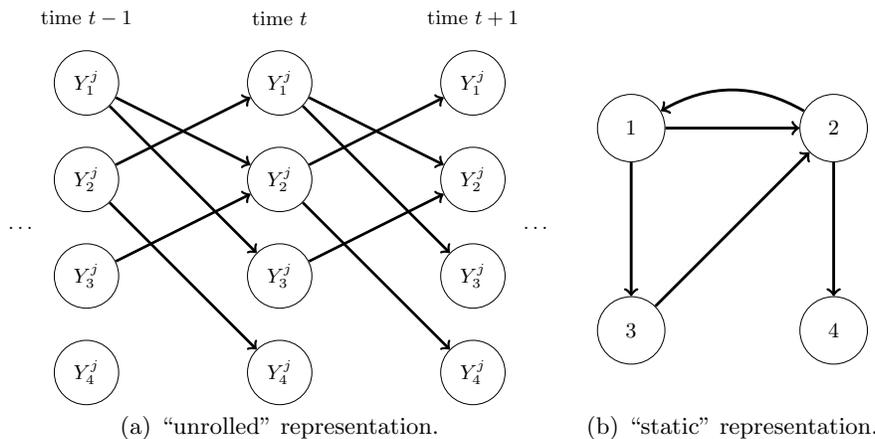


Fig 8: Dynamic Bayesian networks (DBNs). (a) An “unrolled” dynamic Bayesian network (DBN) showing each variable at successive time points. (b) The corresponding “static” representation of DBN (a) with exactly one vertex for each variable.

**Appendix B: Dynamic Bayesian networks.** For the DBNs used here, an edge  $(p, q)$  from  $p \in \mathcal{P}$  to  $q \in \mathcal{P}$  in  $G^j \in \mathcal{G}$  implies that  $Y_q^j(t)$ , the observed value of variable  $q$  in individual  $j$  at time  $t$ , depends directly upon  $Y_p^j(t-1)$ , the observed value of  $p$  in individual  $j$  at time  $t-1$  (Fig. 8(a); note that  $t$  indexes sample index, rather than actual sampling time). Let  $\mathbf{Y}^j$  denote a vector containing all observations for individual  $j$ . Then  $\mathbf{Y}^j(t)$  is conditionally independent of  $\{\mathbf{Y}^j(t-\tau) : \tau \geq 2\}$  given  $\mathbf{Y}^j(t-1)$ ,  $\theta^j$ ,  $G^j$  and  $Z^j$  (first-order Markov assumption). These conditional independence relations are conveniently summarized as a (static) network  $G^j$  with exactly  $P$  vertices (Fig. 8(b)); note that this latter network need not be acyclic.

Hill *et al.* (2012) describe a DBN rooted in the Bayesian linear model. Specifically the response  $Y_p^j(t)$  is predicted by covariates  $\mathbf{Y}^j(t-1)$ . i.e.

$$(31) \quad \mathbf{Y}_p^j = \mathbf{X}_0 \boldsymbol{\alpha} + \mathbf{X}_{\pi_p^j}^j \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where  $\boldsymbol{\epsilon} \sim N(\mathbf{0}_{n \times 1}, \sigma^2 \mathbf{I}_{n \times n})$ . In many cases multiple time series will be available. In this case the vector  $\mathbf{Y}_p^j$  contains the concatenated time series. The matrix  $\mathbf{X}_0 = [\mathbf{1}_{\{t=1\}} \mathbf{1}_{\{t>1\}}]_{n \times 2}$  contains a term for the initial time point in each experiment. The elements of  $\mathbf{X}_{\pi_p^j}^j$  corresponding to initial observations  $Y_p^j(1)$  are simply set to zero. Parameters  $\theta_p^j = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma\}$  are specific to model  $\pi_p^j$ , variable  $p$  and individual  $j$ . In the simplest case, given data

$\mathbf{Y} = \mathbf{y}$ , the model-specific component  $\mathbf{X}_{\pi_p}^j$  of the design matrix consists of the raw predictors  $\mathbf{y}_{\pi_p}^j(t-1)$  where  $\mathbf{y}_Z^j$  denotes the elements of the vector  $\mathbf{y}^j(t-1)$  belonging to the set  $A$ , though more complex basis functions may be used, including interaction terms. For experiments performed in this paper, interaction terms were taken to be all possible products of parent variables, as recommended by Hill *et al.* (2012).

Spencer and Mukherjee (2012) modeled interventional data by modification to the DAG using ideas from causal inference (Pearl, 2000). We mention briefly some of the key ideas and refer the interested reader to the references for full details. A “perfect intervention” corresponds to 100% removal of the target’s activity with 100% specificity. In the context of protein phosphorylation, kinases may be intervened upon using chemical agents. Spencer and Mukherjee (2012) make the simplifying assumptions that these interventions are perfect (the “perfect out fixed effects” (POFE) approach). We refer the reader to Spencer and Mukherjee (2012) for an extended discussion of POFE. This changes the DAG structure to model the intervention and also estimates an additional fixed effect parameter to model the change under intervention in the log-transformed data. When generating data for the simulation study in section 4.2 we take fixed effects to equal zero.

### Appendix C: Exact marginal likelihood for DBN and IDBN.

Hill *et al.* (2012) employed an exact Bayesian approach to capture the suitability of the candidate parent set  $\pi_p^j$ . In brief, a Jeffreys prior  $p(\boldsymbol{\alpha}, \sigma | \pi_p^j, \phi^j, Z^j) \propto 1/\sigma$  for  $\sigma > 0$  was placed over the common parameters. Prior to inference, the non-interventional components of the design matrix are orthogonalized using the transformation  $(\mathbf{X}_{\pi_p^j}^j)_{ik} \mapsto \sum_l (\mathbf{I}_n - \mathbf{P}_0)_{il} (\mathbf{X}_{\pi_p^j}^j)_{lk}$ , where  $\mathbf{P}_0 = \mathbf{X}_0(\mathbf{X}_0^T \mathbf{X}_0)^{-1} \mathbf{X}_0^T$  (Bayarri *et al.*, 2012). A  $g$ -prior was placed on regression coefficients (Zellner, 1986), given by

$$(32) \quad \boldsymbol{\beta} | \boldsymbol{\alpha}, \sigma, \pi_p^j, \phi^j, Z^j \sim N(\mathbf{0}_{b \times 1}, \phi^j \sigma^2 (\mathbf{X}_{\pi_p^j}^T \mathbf{X}_{\pi_p^j})^{-1})$$

where  $b = \dim(\boldsymbol{\beta})$ . Using these priors alongside either DBNs or IDBNs as outlined above, the marginal likelihood can be obtained in closed-form:

$$(33) \quad \mathfrak{P}(\mathbf{y}_p^j | \pi_p^j, \phi^j, Z^j) \propto \frac{1}{(\phi^j + 1)^{b/2}} \left( \mathbf{y}_p^{jT} \left( \mathbf{I}_{n \times n} - \mathbf{P}_0 - \frac{\phi^j}{\phi^j + 1} \mathbf{P}_{\pi_p^j} \right) \mathbf{y}_p^j \right)^{-\frac{n-a}{2}}$$

where  $\mathbf{P}_{\pi_p^j} = \mathbf{X}_{\pi_p^j}(\mathbf{X}_{\pi_p^j}^T \mathbf{X}_{\pi_p^j})^{-1} \mathbf{X}_{\pi_p^j}^T$ ,  $a = \dim(\boldsymbol{\alpha})$  and  $b = \dim(\boldsymbol{\beta})$ . Empirical investigations have previously demonstrated good results for network infer-

ence based on the above marginal likelihood (Hill *et al.*, 2012; Spencer and Mukherjee, 2012).

The hyperparameter  $\phi^j$ , that is related to the weight of the parameter prior  $p(\beta|\alpha, \sigma)$  relative to the data  $\mathbf{y}_p^j$ , was selected, in this paper, using the conditional empirical Bayes procedure outlined in George and Foster (2000), corresponding to

$$(34) \quad \hat{\phi}^j(\pi_p^j) = \arg \max_g \mathfrak{P}(\mathbf{y}_p^j | \pi_p^j, g, Z^j).$$

In order to retain computational efficiency, we evaluated the argument over a finite set of eight candidate values corresponding to prior weight of 0,10,20,30,40,50% and  $(100/n)\%$  (the unit information prior). Alternative strategies for eliciting  $g$ -priors are discussed in Bayarri *et al.* (2012); Liang *et al.* (2008).

**Appendix D: Belief propagation for JNI.** Exact inference for JNI is based on belief propagation (Pearl, 2000). Algorithm 1 displays pseudocode for exact joint model averaging. We also display computational complexity in terms of the number  $M = |\mathcal{G}_p|$  of possible parent sets and the number  $J$  of individuals. Computational complexity of calculating marginal likelihoods  $\mathfrak{P}(\mathbf{y}_p^j | \pi_p^j)$  will partly depend upon sample size  $n$ ; scaling exponents shown here assume  $\mathcal{O}(n) = \mathcal{O}(1)$ . Algorithm 1 contains pseudocode for computation of posterior marginal inclusion probabilities for edges in both the latent network  $G$  and individual-specific networks  $G^j$ . For simplicity we suppress dependence upon ancillary data  $Z^j$  throughout.

---

**Algorithm 1** Belief propagation for JNI.

---

```

1: for  $p \in \mathcal{P}$  do
Phase 0:
2:   Compute and cache  $\mathfrak{P}(\mathbf{y}_p^j | \pi_p^j)$  [ $\forall j \in \mathcal{J}$ ] [ $\forall \pi_p \in \mathcal{G}_p$ ]
Phase I:
3:   Compute and cache [ $\forall j \in \mathcal{J}$ ] [ $\forall \pi_p \in \mathcal{G}_p$ ]
4:    $\mathfrak{P}(\mathbf{y}_p^j | \pi_p) = \sum_{\pi_p^j \in \mathcal{G}_p} \mathfrak{P}(\mathbf{y}_p^j | \pi_p^j) p(\pi_p^j | \pi_p)$  [ $\mathcal{O}(M)$ ]
Phase II:
5:   Compute and cache [ $\forall j \in \mathcal{J}$ ] [ $\forall \pi_p, \pi_p^j \in \mathcal{G}_p$ ]
6:    $p(\pi_p | \mathbf{y}_p, \pi_p^0) \propto p(\pi_p | \pi_p^0) \prod_{j \in \mathcal{J}} \mathfrak{P}(\mathbf{y}_p^j | \pi_p)$  [ $\mathcal{O}(J)$ ]
7:    $p(\pi_p^j | \mathbf{y}_p, \pi_p^0) \propto \sum_{\pi_p \in \mathcal{G}_p} p(\pi_p | \pi_p^0) \mathfrak{P}(\mathbf{y}_p^j | \pi_p^j) p(\pi_p^j | \pi_p) \prod_{k \in \mathcal{J} \setminus \{j\}} \mathfrak{P}(\mathbf{y}_p^k | \pi_p)$  [ $\mathcal{O}(MJ)$ ]
Phase III:
8:   Compute and cache [ $\forall j \in \mathcal{J}$ ] [ $\forall i \in \mathcal{P}$ ]
9:    $p(i \in \pi_p | \mathbf{y}, G^0) = \sum_{\pi_p \in \mathcal{G}_p} \mathbf{1}_{i \in \pi_p} p(\pi_p | \mathbf{y}_p, \pi_p^0)$  [ $\mathcal{O}(M)$ ]
10:   $p(i \in \pi_p^j | \mathbf{y}, G^0) = \sum_{\pi_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in \pi_p^j} p(\pi_p^j | \mathbf{y}, \pi_p^0)$  [ $\mathcal{O}(M)$ ]
11: end for

```

---

## SUPPLEMENTARY MATERIAL

**Supplement A: Additional results and protocols**

(<http://???>). Includes: alternative data generating models; robustness to in-degree restriction, outliers, batch effects and nonexchangeability; ancillary information for breast cancer; inferred wild type networks for breast cancer.

**Supplement B: Computational implementation**

(<http://???>). MATLAB R2014a code (serial and parallel) implementing joint network inference.

**References.**

- Akbani *et al.* (2014) A pan-cancer proteomic perspective on the Cancer Genome Atlas. *Nat. Commun.*, to appear.
- Barretina *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**:603-607.
- Bayarri *et al.* (2012) Criteria for Bayesian model choice with application to variable selection. *Ann. Stat.* **40**(3):1550-1577.
- Boutilier *et al.* (1996) Context-specific independence in Bayesian networks. *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, 115-123.
- Butte *et al.* (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97**(22):12182-12186.
- Cao *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**:956-963.
- Curtis *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403):346-352.
- Danaher, P., Wang, P., Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc. B* **76**(2):373-397.
- Dondelinger, F., Lebre, S., Husmeier, D. (2012) Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Mach. Learn.* **90**(2):191-230.
- Forbes *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* **39**(Suppl 1):D945-D950.
- Geiger, D., Heckerman, D. (1996) Knowledge representation and inference in similarity networks and Bayesian multinets. *Artif. Intell.* **82**(1-2):45-74.
- George, E.I., Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika* **87**(4):731-747.
- Hennessy *et al.* (2010) A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Nonmicrodissected Human Breast Cancer. *Clin. Proteom.* **6**:129-151.
- Hill *et al.* (2012) Bayesian Inference of Signaling Network Topology in a Cancer Cell Line. *Bioinformatics* **28**(21):2804-2810.
- Ibrahim, J.G., Chen, M.-H. (2000) Power prior distributions for regression models. *Stat. Sci.* **15**(1):46-60.
- Ideker, T., Krogan, N.J. (2012) Differential network biology. *Mol. Syst. Biol.* **8**:565.
- Imoto *et al.* (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings of the IEEE Computer Society Bioinformatics Conference*, 104-113.

- Kschischang, F.R., Frey, B.J., Loeliger, H.-A. (2001) Factor Graphs and the Sum-Product Algorithm. *IEEE T. Inform. Theory* **47**(2):498-519.
- Liang *et al.* (2008) Mixtures of  $g$  Priors for Bayesian Variable Selection. *J. Am. Stat. Assoc.* **103**(481):410-423.
- Lu *et al.* (2011) Kinome siRNA-phosphoproteomic screen identifies networks regulating Akt signaling. *Oncogene* **30**:4567-4577.
- Maher, B. (2012) ENCODE: The human encyclopaedia. *Nature* **489**(7414):46-48.
- Mukherjee, S., Speed, T.P. (2008) Network inference using informative priors. *Proc. Nat. Acad. Sci. USA* **105**(38):14313-14318.
- Murphy, K. (2002) Dynamic Bayesian Networks: Representation, Inference and Learning, PhD Thesis, University of California, Berkeley.
- Neve *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**(6):515-527.
- Niculescu-Mizil, A., Caruana, R. (2007) Inductive Transfer for Bayesian Network Structure Learning. *J. Mach. Learn. Res. Workshop and Conference Proceedings* **27**:339-346.
- Oates, C., Mukherjee, S. (2014) Joint Structure Learning of Multiple Non-Exchangeable Networks. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, in press.
- Pearl, J. (1982) Reverend Bayes on inference engines: A distributed tree approach. *Proceedings of the Second National Conference on Artificial Intelligence*, 133-136.
- Pearl, J. (2000) Causality: models, reasoning and inference. Cambridge: MIT press.
- Penfold *et al.* (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics* **28**(12):i233-i241.
- Spencer, S., Hill, S.M., Mukherjee, S. (2012) Dynamic Bayesian networks for interventional data. *CRiSM Working Paper Series, The University of Warwick, UK* **12**:24.
- The 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* **467**(7319):1061-1073.
- The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418):61-70.
- Werhli, A.V., Husmeier, D. (2008) Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *J. Bioinf. Comput. Biol.* **6**(3):543-572.
- Xu *et al.* (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Sig.* **3**(113):ra20.
- Zellner, A. (1986) On Assessing Prior Distributions and Bayesian Regression Analysis With  $g$ -Prior Distributions, *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, 233-243.

DEPARTMENT OF STATISTICS  
UNIVERSITY OF WARWICK  
COVENTRY  
CV4 7AL  
UK  
E-MAIL: [c.oates@warwick.ac.uk](mailto:c.oates@warwick.ac.uk)

KNIGHT CANCER INSTITUTE  
OREGON HEALTH AND SCIENCE UNIVERSITY  
PORTLAND  
OR 97239  
USA  
E-MAIL: [korkola@ohsu.edu](mailto:korkola@ohsu.edu); [grayjo@ohsu.edu](mailto:grayjo@ohsu.edu)

MRC BIostatistics UNIT AND  
CRUK CAMBRIDGE INSTITUTE  
UNIVERSITY OF CAMBRIDGE  
CAMBRIDGE  
CB2 0SR  
UK  
E-MAIL: [sach@mrc-bsu.cam.ac.uk](mailto:sach@mrc-bsu.cam.ac.uk)