Variational Network Inference: Strong and Stable with Concrete Support Supplementary Material

Amir Dezfouli [1 2] **Edwin V. Bonilla** [1] **Richard Nock** [3]

Abstract

This is the Supplementary Material to paper “Variational Network Inference: Strong and Stable with Concrete Support”.

I. Table of contents

[1] UNSW, Sydney. [2] Started work at Data61. [3] Data61, the Australian National University and the University of Sydney. Correspondence to: Amir Dezfouli <akdezfuli@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

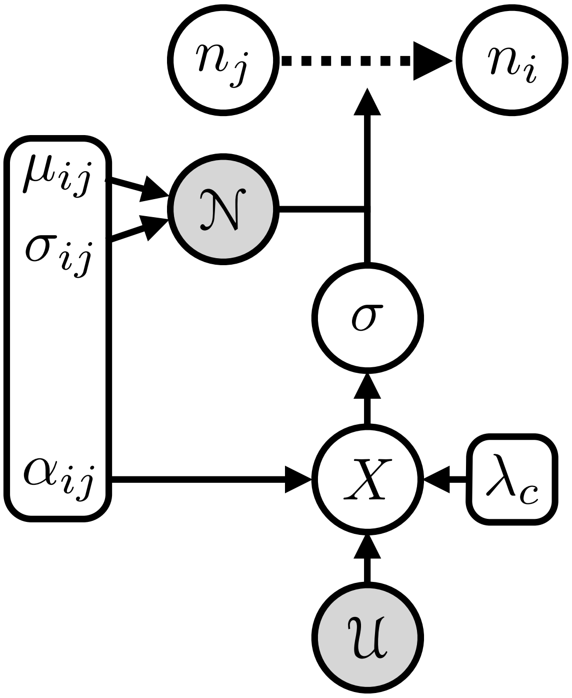

Figure 1. Sampling graph for arc going from node n_j to node n_i ($i \neq j$). Grey nodes denote random variables. Notations in part borrowed from (Maddison et al., 2016). λ_c is a constant that does not depend on the arc.

II. Supplementary material on proofs and algorithms

II.1. Proof of Theorem 1

Denote for short $\mathbf{G} \overset{\text{def}}{=} \mathbf{I} - \mathbf{B}$. The proof is split in three cases, (I) $\lambda_c > 0$ and $\alpha_{ij} > 0, \forall i \neq j$, (II) $\lambda_c = 0$ and $\alpha_{ij} > 0, \forall i \neq j$, and finally (III) $\lambda_c = 0$ and $\exists i \neq j, \alpha_{ij} = 0$.
(Case I: $\lambda_c > 0$, $\alpha_{ij} > 0, \forall i \neq j$) The coordinates g_{ij} take on constant values $g_{ii} = 1$ on the diagonal ($\forall i \in [N]$), and random values G_{ij} outside the diagonal ($i \neq j$). The density of G_{ij} equals $q(A_{ij}) \cdot q(W_{ij})$, where $q(W_{ij}) \overset{\text{def}}{=} \mathcal{N}(\mu_{ij}, \sigma^2_{ij})$ and $q(A_{ij}) \overset{\text{def}}{=} \sigma_{\alpha_{ij}, \lambda_c}(U)$ with

$$\sigma_{\alpha,\lambda_c}(U) \quad \overset{\text{def}}{=} \quad \frac{1}{1 + \exp\left(-\frac{\log \alpha + \log U - \log(1-U)}{\lambda_c}\right)} \ , \tag{1}$$

and $U \sim \mathcal{U}(0,1)$ is uniform on interval $(0,1)$ (Maddison et al., 2016). The proof that $\mathbf{G}$ is invertible adapts a standard argument (Tao, 2008, for example). For any[1] $N \geq 2$, denote $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N$ the columns of $\mathbf{G}$, that is, $\mathbf{G} = [\mathbf{g}_1 | \mathbf{g}_2 | ... | \mathbf{g}_N]$. Each of them can be thought of as a random vector where one coordinate takes value 1 with probability 1, an this coordinate is different for all vectors. $\mathbf{G}$ is non invertible iff $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N$ is linearly dependent. Remark that *none* of the $\mathbf{g}_j$s can be the null vector, so if $\mathbf{G}$ is not invertible, then

$$\exists j > 1 \quad : \quad \mathbf{g}_j \in \text{span}(\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1}) \ . \tag{2}$$

As a consequence,

$$\Pr(\det(\mathbf{G}) = 0) \quad \leq \quad \sum_j \Pr(\mathbf{g}_j \in \text{span}(\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1})) \ , \tag{3}$$

[1] Whenever $N = 1$, $\mathbf{G} \overset{\text{def}}{=} [1]$ is always invertible.

where the distribution is the product distribution over the columns of $\mathbf{G}$. Fix *any* $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1}$ belonging to the respective supports of the columns, and let

$$q_j \quad \stackrel{\mathrm{def}}{=} \quad \Pr(\mathbf{g}_j \in \mathrm{span}(\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1}) | \mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1}) \ . \tag{4}$$

Because the uniform and normal distributions are both absolutely continuous with respect to Lebesgue measure and $\sigma_{\alpha,\lambda_c}(x) \leq 1 \ll \infty$ (it is also Lipschitz) for any $\alpha > 0, \lambda_c \neq 0, U \in (0,1)$, so is the density of G_{ij} for any $i \neq j$, and thereby the density of $\mathbf{g}_j$ for any $j \geq 1$. Along with the fact that $\mathrm{span}(\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1})$ has strictly positive codimension for any $j \leq N$, it comes

$$q_j \quad = \quad 0, \forall j \geq 2, \forall \mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1} \text{ fixed} \ . \tag{5}$$

Integrating over the choices of $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1}$, we get $\Pr(\mathbf{g}_j \in \mathrm{span}(\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_{j-1})) = 0, \forall j \leq N$ and so $\Pr(\det(\mathbf{G}) = 0) = 0$ from ineq. (3). As a consequence, $\mathbf{I} - \mathbf{B}$ is non-singular with probability one, as claimed.
(Case II: $\lambda_c = 0$, $\alpha_{ij} > 0, \forall i \neq j$) this boils down to choosing a Bernoulli $B(p_{ij})$ distribution over A_{ij}, corresponding to the limit case $\lambda_c \rightarrow 0$ with (Maddison et al., 2016):

$$p_{ij} \quad = \quad \frac{\alpha_{ij}}{1 + \alpha_{ij}} \ . \tag{6}$$

In this case, the distribution of $\mathbf{g}_j$ is not absolutely continuous but a trick allows to truncate the distribution on a subset over which it is absolutely continuous, and therefore reduce to Case I to handle it.
The *only* atom eventually having non-zero probability is the canonical basis vector $\mathbf{1}_j$, which has probability $\prod_{i \neq j}(1 - p_{ij})$ to be sampled. We now perform a sequence of recursive row-column (row followed by column or the reverse) permutations, starting on $\mathbf{G}$, which by definition do not change its invertibility status but only the sign of its determinant. The first row-column permutation is carried out in such a way that the first column of the new matrix, $\Pi_1(\mathbf{G})$, is the first canonical basis vector, $\mathbf{1}_1$. We then repeat this operation to have the second canonical basis vector in the second column, and so on until until it cannot be done anymore to make appear on the left block a new canonical basis vector. Assuming we have done $N - k$ sequences, we obtain from $\mathbf{G}$ the final matrix $\Pi_1(\mathbf{G})$ with:

$$\Pi_1(\mathbf{G}) \quad = \quad \left[\begin{array}{c|c} \mathbf{I}_{N-k} & \mathbf{A}_{(N-k)\times k} \\ \mathbf{0}_{k\times(N-k)} & \hat{\mathbf{G}}_{1,k} \end{array} \right] \ . \tag{7}$$

Here, $\hat{\mathbf{G}}_{1,k} \in \mathbb{R}^{k \times k}$. Now, we are going to carry out Π_1 again, but on the lower-right block, $\hat{\mathbf{G}}_{1,k}$. Removing dimension-dependent indexes, we obtain matrix

$$\Pi_2(\mathbf{G}) \quad = \quad \left[\begin{array}{c|c} \mathbf{I} & \mathbf{A} \\ \mathbf{0} & \Pi_1(\hat{\mathbf{G}}_1) \end{array} \right] \tag{8}$$

$$= \quad \left[\begin{array}{c|c} \mathbf{I} & \mathbf{A}_1 \\ \mathbf{0} & \left[\begin{array}{c|c} \mathbf{I} & \mathbf{A}_2 \\ \mathbf{0} & \hat{\mathbf{G}}_2 \end{array} \right] \end{array} \right] \ . \tag{9}$$

We then keep on doing the same transformation on block $\hat{\mathbf{G}}_2$ until it is not possible anymore. When it is not possible anymore, we know that the current submatrix, say $\hat{\mathbf{G}}_n$, does not contain any canonical basis vector as column, as depicted in Figure 2.

Lemma A $|\det(\mathbf{G})| = |\det(\hat{\mathbf{G}}_n)|, \forall n \geq 1$.

Proof: We proceed by induction. The key observation is the following standard linear algebra identity. Denoting with a single index the order of a general square matrix, like $\mathbf{A}_p$, we have for any $\mathbf{A}_p$ non-singular,

$$\left[\begin{array}{c|c} \mathbf{A}_p & \mathbf{B} \\ \mathbf{C} & \mathbf{D}_q \end{array} \right] \underbrace{\left[\begin{array}{c|c} \mathbf{I}_p & -\mathbf{A}_p^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I}_q \end{array} \right]}_{\stackrel{\mathrm{def}}{=} \mathbf{E}} = \underbrace{\left[\begin{array}{c|c} \mathbf{I}_p & \mathbf{0} \\ \mathbf{C}\mathbf{A}_p^{-1} & \mathbf{I}_q \end{array} \right]}_{\stackrel{\mathrm{def}}{=} \mathbf{F}} \left[\begin{array}{c|c} \mathbf{A}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{D} - \mathbf{C}\mathbf{A}_p^{-1}\mathbf{B} \end{array} \right] , \tag{10}$$

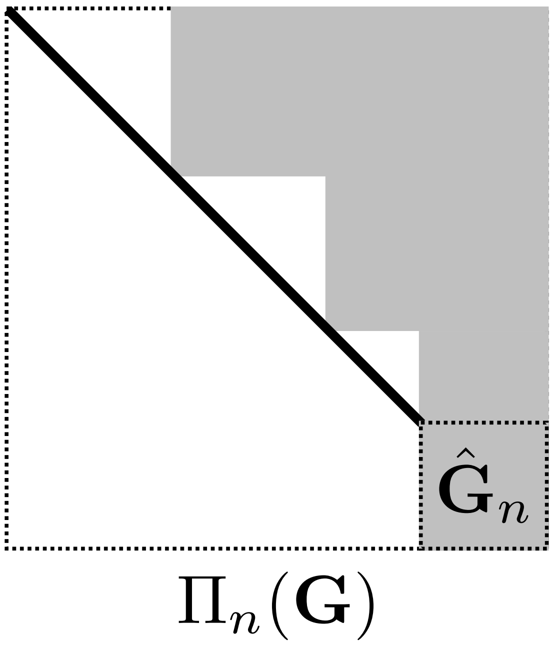

Figure 2. Final matrix $\Pi_n(\mathbf{G})$ obtained after recursively applying $\Pi_1(.)$ to the lower-right block. Here, white blocks mean all-zero, plain dark lines mean all-one, and grey is unspecified.

for any $p > 0, q > 0, p+q = n, \mathbf{B} \in \mathbb{R}^{p\times q}, \mathbf{C} \in \mathbb{R}^{q\times p}, \mathbf{D} \in \mathbb{R}^{q\times q}$. Taking determinants, we note that $\det(\mathbf{E}) = \det(\mathbf{F}) = 1$ because they are triangular with unit diagonal, and so

$$\begin{aligned} \det\left(\left[\begin{array}{c|c} \mathbf{A}_p & \mathbf{B} \\ \mathbf{C} & \mathbf{D}_q \end{array}\right]\right) &= \det\left(\left[\begin{array}{c|c} \mathbf{A}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{D} - \mathbf{C}\mathbf{A}_p^{-1}\mathbf{B} \end{array}\right]\right) &\quad (11) \\ &= \det(\mathbf{A}_p)\cdot\det(\mathbf{D} - \mathbf{C}\mathbf{A}_p^{-1}\mathbf{B}) \ , &\quad (12) \end{aligned}$$

because the right hand-side in eq. (11) is block diagonal. Matching the left hand-side of eq. (11) with eq. (7), so putting $\mathbf{A}_p = \mathbf{I}$ and $\mathbf{C} = \mathbf{0}$, we obtain $\det(\Pi_1(\mathbf{G})) = \det(\hat{\mathbf{G}}_1 - \mathbf{0}\mathbf{I}_{N-k}\mathbf{A}_p) = \det(\hat{\mathbf{G}}_1)$, and therefore $|\det(\mathbf{G})| = |\det(\hat{\mathbf{G}}_1)|$. We then just recursively use eq. (10) on the lower-right block ($\hat{\mathbf{G}}_j$, for $j = 1, 2, ..., n-1$) and get the statement of the Lemma. (End of the proof of Lemma A). □

So, $\mathbf{G}$ is invertible iff $\hat{\mathbf{G}}_n$ is invertible and:

$$\begin{aligned} \Pr(\det(\mathbf{G}) = 0) &\leq \Pr(\det(\Pi_1(\mathbf{G})) = 0) \\ &\quad = \Pr(\det(\hat{\mathbf{G}}_1) = 0) \\ &\leq \Pr(\exists k \in \{2, 3, ..., N\} : \det(\hat{\mathbf{G}}_k) = 0 | \hat{\mathbf{G}}_k \in \mathbb{R}^{k\times k} \wedge \mathcal{P}(\hat{\mathbf{G}}_k)) \\ &\leq \sum_{k=2}^{N} \Pr(\det(\hat{\mathbf{G}}_k) = 0 | \hat{\mathbf{G}}_k \in \mathbb{R}^{k\times k} \wedge \mathcal{P}(\hat{\mathbf{G}}_k)) \ , \quad (13) \end{aligned}$$

where $\mathcal{P}(\mathbf{G})$ is the property that no column of $\mathbf{G}$ is a canonical basis vector. Notice the change: no column in $\hat{\mathbf{G}}_k$ is allowed to be a canonical basis vector, and therefore the support for the density of the columns of $\hat{\mathbf{G}}_k$ is such that its distribution is now absolutely continuous. We are thus left with the same case as in Case I, which yields $\Pr(\det(\hat{\mathbf{G}}_k) = 0 | \hat{\mathbf{G}}_k \in \mathbb{R}^{k\times k} \wedge \mathcal{P}(\hat{\mathbf{G}}_k)) = 0, \forall k \in \{2, 3, ..., N\}$, and brings $\Pr(\det(\mathbf{G}) = 0) = 0$ as well.

(Case III: $\lambda_c \geq 0$, $\alpha_{ij} = 0$ for some $i \neq j$) Remark that $\lim_{\alpha\to 0} \sigma_{\alpha,\lambda_c}(x) = 0$ if $\lambda_c > 0$, and if $\lambda_c = 0$, this boils down from Case II (eq. (6)) to choosing a Bernoulli $B(0)$ distribution over A_{ij}, so both cases coincide with A_{ij} being chosen as $B(0)$, implying $G_{ij} = 0$. We are left with the same transformation as in Case II — the main difference being that some G_{ij} is surely zero, but it changes nothing to the reasoning done in case II. Therefore, $\Pr(\det(\mathbf{G}) = 0) = 0$ again.

II.2. Proof of Theorem 2

It comes from Theorem 1 that $\mathbf{G}^{-1}$ can always be computed with probability one with respect to the random sampling of $\mathbf{B}$, and there is no constraint on the parameterization of the concrete distribution for invertibility (Maddison et al., 2016). Interestingly perhaps, the story would be completely different for the invertibility of $\mathbf{B}$, as the argument for cases (II) and (III) break down because with positive probability that would be easy to lower-bound, $\mathbf{B}$ would in fact be not invertible.

The important consequence of Theorem 1 (main file) relies on the computation of the log likelihood, which we recall:

$$\log p(\mathbf{y}|\mathbf{W}, \mathbf{A}) \quad = \quad -\frac{1}{2}\log|\mathbf{\Sigma}_y| - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}_y^{-1}\mathbf{y} + C \ . \tag{14}$$

We now prove Theorem 2. We recall the main matrix component of eq. (14):

$$\mathbf{\Sigma}_y \quad = \quad ((\mathbf{I}-\mathbf{B})^\top(\mathbf{I}-\mathbf{B}))^{-1} \otimes \mathbf{K}_t + \sigma_f^2(\mathbf{I}-\mathbf{B})^{-1}\mathbf{B}((\mathbf{I}-\mathbf{B})^{-1}\mathbf{B})^\top \otimes \mathbf{\Sigma}_{\mathrm{I}} + \sigma_y^2\mathbf{I} \ . \tag{15}$$

We observe that the following two matrices are positive semi-definite[2]: $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{B}((\mathbf{I}-\mathbf{B})^{-1}\mathbf{B})^\top$, $\mathbf{K}_t$, while $\mathbf{\Sigma}_{\mathrm{I}}, \mathbf{I}, ((\mathbf{I}-\mathbf{B})^\top(\mathbf{I}-\mathbf{B}))^{-1}$ are positive definite (with probability 1 for that last one, see Section II.1). Hence, a sufficient condition for the combination in $\mathbf{\Sigma}_y$ to be positive definite is $\sigma_y^2 > 0$, as claimed. This brings the finiteness of $|\log p(\mathbf{y}|\mathbf{W}, \mathbf{A})|$ with probability one, the fact that $|\mathcal{L}_{\mathrm{ell}}| \ll \infty$, and the statement of Theorem 2.

II.3. Proof of Theorem 4

We split the proof in two main parts, the first of which focuses on a simplified version of the model in which the Bernoulli parameter ($\mathbf{A}$) is sampled according to a Dirac — *e.g.* in the context of inference, from the prior standpoint, it is maximally informed. The results might be useful outside our framework, if p is sampled from a distribution different from the ones we use.

We state the main notations involved in the Theorem. We define the total (squared) expected input (resp. output) to node i as $\mu_{i.}^+ \stackrel{\mathrm{def}}{=} \sum_j \mu_{ij}^2$ (resp. $\mu_{.i}^+ \stackrel{\mathrm{def}}{=} \sum_j \mu_{ji}^2$), and the total input (resp. output) variance as $\sigma_{i.}^+ \stackrel{\mathrm{def}}{=} \sum_j \sigma_{ij}^2$ (resp. $\sigma_{.i}^+ \stackrel{\mathrm{def}}{=} \sum_j \sigma_{ji}^2$). We also define averages, $\overline{\mu}_{i.}^+ \stackrel{\mathrm{def}}{=} \mu_{i.}^+/N$, $\overline{\sigma}_{i.}^+ \stackrel{\mathrm{def}}{=} \sigma_{i.}^+/N$ (same for outputs), and biased weighted proportions, $\tilde{p}_{i.}^\mu \stackrel{\mathrm{def}}{=} \sum_j p_{ij}\mu_{ij}^2/\mu_{i.}^+$, $\tilde{p}_{i.}^\sigma \stackrel{\mathrm{def}}{=} \sum_j p_{ij}\sigma_{ij}^2/\sigma_{i.}^+$ (again, same for outputs).

Now, we define two functions $U, E : \{1, 2, ..., 2N\} \rightarrow \mathbb{R}_+$ as:

$$U(i) \quad \stackrel{\mathrm{def}}{=} \quad \left\{ \begin{array}{ll} 2\tilde{p}_{i.}^\mu\overline{\mu}_{i.}^+ + 2\tilde{p}_{i.}^\sigma\overline{\sigma}_{i.}^+ & (i \leq N) \\ 2\tilde{p}_{.j}^\mu\overline{\mu}_{.j}^+ + 2\tilde{p}_{.j}^\sigma\overline{\sigma}_{.j}^+ : j \stackrel{\mathrm{def}}{=} i - N & (i > N) \end{array} \right. ,$$

$$E(i) \quad \stackrel{\mathrm{def}}{=} \quad \left\{ \begin{array}{ll} \phi(\tilde{p}_{i.}^\mu) \cdot \overline{\mu}_{i.}^+ + \overline{\sigma}_{i.}^+ & (i \leq N) \\ \phi(\tilde{p}_{.j}^\mu) \cdot \overline{\mu}_{.j}^+ + \overline{\sigma}_{.j}^+ : j \stackrel{\mathrm{def}}{=} i - N & (i > N) \end{array} \right. ,$$

where $\phi(z) \stackrel{\mathrm{def}}{=} 2\sqrt{z(1-z)}$ is Matsushita's entropy. For any diagonalizable matrix $\mathbf{M}$, we let $\lambda(\mathbf{M})$ denote its eigenspectrum, and $\lambda^\uparrow(\mathbf{M}) \stackrel{\mathrm{def}}{=} \max\lambda(\mathbf{M})$, $\lambda^\downarrow(\mathbf{M}) \stackrel{\mathrm{def}}{=} \min\lambda(\mathbf{M})$. Our simplified version of Theorem 4, which we first prove, is the following one.

Theorem B *Assume* $\mathsf{A}_{ij} \sim \mathcal{B}(\mathsf{\rho}_{ij})$ *with* $\mathsf{\rho}_{ij} \sim \mathrm{Dirac}(p_{ij})$, *and* $\mathsf{W}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$, $p_{ij}, \mu_{ij}, \sigma_{ij}$ *being fixed for any* i, j. *Fix any constants* $c > 0$ *and* $0 < \gamma < 1$ *and let*

$$\lambda_\circ \stackrel{\mathit{def}}{=} \frac{\lambda^\downarrow(\mathbf{K}_t)}{2} + \sigma_y^2 \ , \ \ \lambda_\bullet \stackrel{\mathit{def}}{=} 2\lambda^\uparrow(\mathbf{K}_t) + \sigma_f^2 + \sigma_y^2 \ . \tag{16}$$

Suppose that:

$$\max_i U(i) \quad \in \quad \left[\frac{\max_i E(i)}{N^\gamma}, \frac{1}{100N^2}\right] \ . \tag{17}$$

[2] As remarked above, depending on the choices of parameters λ_c and $\alpha_{..}$, the null space of $\mathbf{B}$ is indeed not always reduced to the null vector. Therefore, $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{B}((\mathbf{I}-\mathbf{B})^{-1}\mathbf{B})^\top$ may be not positive definite with strictly positive probability.

If N is larger than some constant depending on c and γ, then with probability $\geq 1-(1/N^c)$ over the sampling of $\mathbf{W}$ and $\mathbf{A}$, the following holds true:

$$\lambda(\mathbf{\Sigma}_y) \quad \subset \quad [\lambda_\circ, \lambda_\bullet] \ . \tag{18}$$

II.3.1. Helper tail bounds and properties for arcs, row and columns in matrix $\mathbf{A} \odot \mathbf{W}$

To obtain concentration bounds on $\log p(\mathbf{y}|\mathbf{W}, \mathbf{A})$, we need to map the arc signal onto the real line, including *e.g.* when $p = 0$ (in which case there cannot exist an arc between the two corresponding nodes, so there is no observable "weight" *per se*). We follow the convention for the Hawkes model of Linderman & Adams (2014), and associate to these "no signal" events the real zero, which makes sense since for example it matches the Dirac case when $\mu, \sigma \rightarrow 0$ — which corresponds to an arc with weight always zero —. Define for short

$$\mathbf{H} \quad \overset{\mathrm{def}}{=} \quad (\mathbf{I}-\mathbf{B})^\top(\mathbf{I}-\mathbf{B}) \ , \tag{19}$$

$$\mathbf{H}' \quad \overset{\mathrm{def}}{=} \quad (\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^\top \ , \tag{20}$$

$$\mathbf{J} \quad \overset{\mathrm{def}}{=} \quad \mathbf{B}\mathbf{B}^\top \ , \tag{21}$$

so that

$$\mathbf{\Sigma}_y \quad = \quad \mathbf{H}^{-1} \otimes \mathbf{K}_t + \sigma_f^2 (\mathbf{I}-\mathbf{B})^{-1}\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top} \otimes \mathbf{\Sigma}_{\mathrm{I}} + \sigma_y^2 \mathbf{I} \ . \tag{22}$$

We remark that the eigenspectrum of $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top}$ is the same as for $\mathbf{J}\mathbf{H}'^{-1}$: if $\mathbf{u}$ is an eigenvector of $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top}$, then $(\mathbf{I}-\mathbf{B})^{-1}\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top}\mathbf{u} = \lambda\mathbf{u}$ is equivalent to $\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top}\mathbf{u} = \lambda(\mathbf{I}-\mathbf{B})\mathbf{u}$, equivalent to $\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top}(\mathbf{I}-\mathbf{B})^{-1}\mathbf{v} = \lambda\mathbf{v}$ (letting $\mathbf{v} \overset{\mathrm{def}}{=} (\mathbf{I}-\mathbf{B})\mathbf{u}$), finally equivalent to $\mathbf{J}\mathbf{H}'^{-1}\mathbf{v} = \lambda\mathbf{v}$. Therefore, bounding the eigenspectra of $\mathbf{H}, \mathbf{H}', \mathbf{J}$, plus adequate assumptions on that of $\mathbf{K}_t$, shall lead to bounding the eigenspectra of $\mathbf{\Sigma}_y$, but to get al these bounds, we essentially need properties and concentration inequalities for the coordinates of $\mathbf{B}$ and their row- or column- sums. This is what we establish in this Section.

We first derive a tail bound for arc weight, removing indexes for clarity, and assuming $q(W) \overset{\mathrm{def}}{=} \mathcal{N}(\mu, \sigma^2)$ and $q(A) \overset{\mathrm{def}}{=} \mathcal{B}(p)$ (see Figure 1). Let W denote the random variable taking the arc weight. We recall that random variable X is $(k, \upbeta)$-sub-Gaussian ($k, \upbeta > 0$) iff (Vu, 2014):

$$\mathbb{E}_{\mathsf{X}}[\exp(\uplambda(\mathsf{X}-\mathbb{E}[\mathsf{X}])) \quad \leq \quad k \cdot \exp\left(\frac{\upbeta^2\uplambda^2}{2}\right) \ , \forall \uplambda \in \mathbb{R} \ . \tag{23}$$

Theorem C *Let $\mathsf{W} \sim q(W) \cdot q(A)$. The following holds true:*

$$\mathbb{E}_{\mathsf{W}}[\exp(\uplambda(\mathsf{W}-\mathbb{E}[\mathsf{W}])) \quad = \quad (1-p)\cdot\exp(-p\mu\uplambda) + p\cdot\exp\left(\mu(1-p)\uplambda + \frac{\sigma^2\uplambda^2}{2}\right) \ , \forall \uplambda \in \mathbb{R} \ . \tag{24}$$

Furthermore, W is $(1, \upbeta)$-sub-Gaussian with $\upbeta$ satisfying:

- $\upbeta^2 = p\sigma^2$ *if* $p \in \{0, 1\}$,
- $\upbeta^2 = 2\sqrt{p(1-p)}\mu^2 + \sigma^2$ *if* $p \in (0, 1)$.

Proof: Denote for short two random variables $\mathsf{N} \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathsf{B} \sim \mathcal{B}(p)$. We trivially have $\mathbb{E}[\mathsf{W}] = p\mu$ and:

$$\begin{aligned}
\mathbb{E}_{\mathsf{W}}[\exp(\uplambda(\mathsf{W}-\mathbb{E}[\mathsf{W}])) \quad &= \quad \mathbb{E}_{\mathsf{W}}[\exp((\mathsf{W}-p\mu)\uplambda)] \\
&= \quad (1-p)\cdot\mathbb{E}_{\mathsf{N}}[\exp(-p\mu\uplambda)] + p\cdot\mathbb{E}_{\mathsf{N}}[\exp((\mathsf{N}-p\mu)\uplambda)] \\
&= \quad (1-p)\cdot\exp(-p\mu\uplambda) + p\cdot\mathbb{E}_{\mathsf{N}}[\exp((\mathsf{N}-p\mu)\uplambda)] \\
&= \quad (1-p)\cdot\exp(-p\mu\uplambda) + p\cdot\exp(-p\mu\uplambda)\cdot\mathbb{E}_{\mathsf{N}}[\exp(\mathsf{N}\uplambda)] \\
&= \quad (1-p)\cdot\exp(-p\mu\uplambda) + p\cdot\exp(-p\mu\uplambda)\cdot\exp\left(\mu\uplambda + \frac{\sigma^2\uplambda^2}{2}\right) && (25) \\
&= \quad (1-p)\cdot\exp(-p\mu\uplambda) + p\cdot\exp\left(\mu(1-p)\uplambda + \frac{\sigma^2\uplambda^2}{2}\right) \ , && (26)
\end{aligned}$$

for any $\lambda \in \mathbb{R}$, as claimed for eq. (24). Eq. (25) comes from the moment generating function for Gaussian N. Now, it is clear that

- W is sub-Gaussian with parameter $\beta = \sigma$ in the following two cases: (i) $p = 1$, (ii) $\mu = 0$. For this latter case, we have indeed $\mathbb{E}_{\mathsf{W}}[\exp(\lambda(\mathsf{W} - \mathbb{E}[\mathsf{W}])) = (1-p) + p \cdot \exp(\sigma^2\lambda^2/2) \leq ((1-p)+p) \cdot \exp(\sigma^2\lambda^2/2) = \exp(\sigma^2\lambda^2/2)$ (using Jensen's inequality on $z \mapsto \exp(z)$). Furthermore, sub-Gaussian parameter σ cannot be improved in both cases.
- the trivial case $p = 0$ leads to sub-Gaussianity for any $\beta \geq 0$.

Otherwise (assuming thus $0 < p < 1$ and $\mu \neq 0$), we can immediately rule out the case $\beta \leq \sigma$ (for any $k > 0$), by noticing that, for $\beta = \sigma$, we have $p \cdot \exp(\mu(1-p)\lambda) = k$ for

$$\lambda = \frac{1}{(1-p)\mu} \log \frac{k}{p} \;\; (\ll \infty) \ , \tag{27}$$

and so, for this value of λ, $\mathbb{E}_{\mathsf{W}}[\exp(\lambda(\mathsf{W} - \mathbb{E}[\mathsf{W}])) > p \cdot \exp\left(\mu(1-p)\lambda + (\beta^2\lambda^2)/2\right) = k\exp(\beta^2\lambda^2/2)$. In the following, we therefore consider $0 < p < 1$, $\mu \neq 0$ and $\beta > \sigma$.

Lemma D $\forall p \in [0,1], \forall x > 0$, *we have*

$$p(x-1) + 1 \leq x^p \exp(\phi_u(p) \cdot \log^2 x) \ , \tag{28}$$

where $\phi_u(p) \overset{\text{def}}{=} \sqrt{p(1-p)}$ *is (unnormalized) Matsushita's entropy.*

Remark: ineq. (28) is probably close to be optimal analytically. Replacing $\phi_u(p)$ by a dominated entropy like Gini's $\phi_u(p) \propto p(1-p)$ (*i.e.* with finite derivatives on the right of 0 and left of 1) seems to break the result.
Proof: The proof makes use of several tricks to counter the fact that the right-hand side of ineq. (28) is essentially concave – but not always – in p, and essentially convex – but not always – in x, and matches the left-hand side as $p \rightarrow \{0, 1\}$. In a first step, we show that ineq. (28) holds for $\log x \in [-1, 1]$ (and any $p \in [0,1]$), then Step 2 shows that ineq. (28) holds for $\log x \geq -1$ (and $p \in [0,1]$). Step 3 uses a symmetry argument on the right-hand side of ineq. (28) to extend the result to any $x > 0$ (and any $p \in [0,1]$), thereby finishing the proof.

Step 1. We remark that $\phi_u(p) \overset{\text{def}}{=} \sqrt{p(1-p)}$ satisfies the following properties:

(i) $\lim_0 \phi'_u(p) = +\infty$, $\lim_1 \phi'_u(p) = -\infty$;

(ii) $\lim_{\{0,1\}} \phi''_u(p) + (1 + \phi'_u(p) \cdot k)^2 = -\infty$ for any k.

Denote for short $F(p, x) \overset{\text{def}}{=} x^p \exp(\phi_u(p) \cdot \log^2 x)$. We have:

$$\frac{\partial F}{\partial p} = \log x \cdot (1 + \phi'_u(p) \cdot \log x) \cdot F_x(p) \ , \tag{29}$$

$$\frac{\partial^2 F}{\partial p^2} = \log^2 x \cdot \left(\phi''_u(p) + (1 + \phi'_u(p)\log x)^2\right) \cdot F_x(p) \ . \tag{30}$$

It comes $\partial F/\partial p \sim_0 \phi'_u(p)\log^2 x \cdot F_x(p)$ and so $\lim_0 \partial F/\partial p = +\infty$ because of (i). Since $F(0, x) = 1$, we have $F(p, x) > p(x-1) + 1$ in a neighborhood of 0. Also, we can check as well that $\lim_0 \partial^2 F/\partial p^2 = -\infty$ because of (ii), so $F(p, x)$ is concave in a neighborhood of 0. For the same reasons, $F(p, x)$ is concave in a neighborhood of 1 and since $F(1, x) = x$, we also have $F(p, x) > p(x-1) + 1$ in a neighborhood of 1. Now, to zero the second derivative, we need equivalently:

$$\log x = \frac{1}{\phi'_u(p)} \cdot \left(\frac{\pm 1}{2\phi_u^{\frac{3}{2}}(p)} - 1\right) \ , \tag{31}$$

or, equivalently again:

$$G(p,x) \stackrel{\mathrm{def}}{=} 2\phi_u^{\frac{3}{2}}(p) + \log x(1-2p)\phi_u^{\frac{1}{2}}(p) \quad = \quad r \ , \tag{32}$$

with $r \in \{-1, 1\}$. We have (letting $z \stackrel{\mathrm{def}}{=} \log x$ for short and $h_1(z) \stackrel{\mathrm{def}}{=} \sqrt{8z^2+9}, h_2(z) \stackrel{\mathrm{def}}{=} (2z^2 + h_1(z) + 3)/(z^2+1)$),

$$\max_{p\in[0,1]} G(p,x) \quad = \quad \frac{(h_2(z))^{\frac{1}{4}}\left(\sqrt{h_2(z)(z^2+1)} + \sqrt{3+4z^2-h_1(z)z}\right)}{2^{\frac{5}{4}}3^{\frac{3}{4}}\sqrt{z^2+x1}} \ , \tag{33}$$

and we can check that $\max_{p\in[0,1]} G(p,x) < 1$ when $\log(x) \leq 1$. We can also check that $\min_{p\in[0,1]} G(p,x) > -1$ when $\log(x) \geq -1$, so eq. (31) has in fact no solution whenever $\log x \in [-1,1]$, regardless of the choice of r. Hence, in this case, $F(p,x)$ is concave in p and we get $F(p,x) \geq p(x-1)+1$, for any $\log x \in [-1,1]$.

Step 2. Suppose now that $|\log x| > 1$. We have

$$\begin{aligned}\frac{\partial F}{\partial x} &= \frac{1}{x}\cdot(p + 2\phi_u(p)\log x)\cdot G_p(x) \ , &(34)\\ \frac{\partial^2 F}{\partial x^2} &= \frac{1}{x^2}\cdot\left(4\phi_u^2(p)\log^2 x + 2\phi_u(p)(2p-1)\log x + 2\phi_u(p) - p(1-p)\right)\cdot G_p(x) \ . &(35)\end{aligned}$$

We have $(\partial F/\partial x)(p,1) = p$ and convexity is ensured as long as

$$\log x \quad \not\in \quad \left[\frac{1-2p\pm\sqrt{1-8\phi_u(p)}}{4\phi_u(p)}\right] \stackrel{\mathrm{def}}{=} \mathcal{A} \ . \tag{36}$$

It happens that $\mathcal{A} \subset [-1,1]$, so whenever $|\log x| \geq 1$, $F(p,x)$ is convex in x. To finish Step 2, considering only the case $\log x \geq 1$, it is sufficient to show that $(\partial F/\partial x)(p,e) \geq p$, or equivalently,

$$H(p) \stackrel{\mathrm{def}}{=} (p + 2\phi_u(p))\exp(p+\phi_u(p)) \quad \geq \quad ep \ , \tag{37}$$

It can be shown that the first derivative,

$$H'(p) \quad = \quad \left(2 - p + \frac{2+p-6p^2}{2\phi_u(p)}\right)\cdot\exp(p+\phi_u(p)) \ , \tag{38}$$

is $\geq e$ for any $p < 0.7$ — so, since both limits in 0 for eq. (37) coincide, eq. (37) holds for any $p < 0.7$. The second derivative (fixing $Q(p) \stackrel{\mathrm{def}}{=} 2 - 13p + 34p^2 - 12p^3 - 8p^4 + \phi_u(p)((3p-2)(1-4p^2) + 4\phi_u^2(p)(1-p))$),

$$H''(p) \quad = \quad \phi_u''(p)\cdot Q(p)\cdot\exp(p+\phi_u(p)) \ , \tag{39}$$

is strictly negative for $p \geq 0.7$ — so, since both limits in 1 for eq. (37) coincide, eq. (37) is strictly concave for $p \geq 0.7$, it sits above its chord $[(0.7, H(0.7)), (1, e)]$ which itself sits above $p \mapsto ep$ for $p \leq 1$, so eq. (37) holds for any $p \geq 0.7$. This achieves the proof of Step 2.

Step 3. We now have that ineq. (28) holds for any $\log x \geq -1$ and any $p \in [0,1]$. To finish the argument, we just have to remark that $F(p,x)$ satisfies the following symmetry:

$$F(p,x) \quad = \quad x\cdot F\left(1-p, \frac{1}{x}\right) \ , \tag{40}$$

so assuming that $\log x < -1$, we have $\log(1/x) \geq 1$, so we reuse Steps 1 and 2 together with eq. (40) to obtain that for any $\log x < -1$,

$$\begin{aligned}F(p,x) &= x\cdot F\left(1-p,\frac{1}{x}\right)\\ &\geq x\cdot\left((1-p)\left(\frac{1}{x}-1\right)+1\right)\\ &\quad = (1-p)(1-x) + x = p(x-1)+1 \ , &(41)\end{aligned}$$

as claimed, where the inequality makes use of Steps 1, 2. This achieves the proof of Lemma D. □

To finish the proof of Theorem C, we make use of Lemma D as follows, starting from eq. (24):

$$\begin{aligned}
\mathbb{E}_{\mathsf{W}}[\exp(\lambda(\mathsf{W}-\mathbb{E}[\mathsf{W}])) &= (1-p)\cdot\exp(-p\mu\lambda)+p\cdot\exp\left(\mu(1-p)\lambda+\frac{\sigma^2\lambda^2}{2}\right) \\
&\leq \{(1-p)\cdot\exp(-p\mu\lambda)+p\cdot\exp\left(\mu(1-p)\lambda\right)\}\cdot\exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad (42)\\
&= \{(1-p)+p\cdot\exp\left(\mu\lambda\right)\}\cdot\exp(-p\mu\lambda)\cdot\exp\left(\frac{\sigma^2\lambda^2}{2}\right) \\
&\leq \exp(p\mu\lambda)\cdot\exp\left(\phi_u(p)\mu^2\lambda^2\right)\cdot\exp(-p\mu\lambda)\cdot\exp\left(\frac{\sigma^2\lambda^2}{2}\right) \quad (43)\\
&= \exp\left(\frac{(\sigma^2+2\phi_u(p)\mu^2)\lambda^2}{2}\right) \quad ,\forall\lambda\in\mathbb{R}\ . \quad (44)
\end{aligned}$$

Ineq. (42) uses the fact that $\sigma^2\lambda^2\geq 0$, and ineq. (43) uses Lemma D with $x=\exp(\mu\lambda)$. Hence, W is sub-Gaussian with parameters $k=1$ and $\beta^2=\sigma^2+2\phi_u(p)\mu^2=\sigma^2+2\sqrt{p(1-p)}\mu^2$, as claimed. This ends the proof of Theorem C. □

Theorem C leads to the following concentration inequality for the row- and column-sums of $\mathbf{B}$, which are key to bound eigenvalues.

Lemma E *Let $\mu^+_{i.}\overset{\mathit{def}}{=}\sum_j\mu^2_{ij}$, $\mu^+_{.j}\overset{\mathit{def}}{=}\sum_i\mu^2_{ij}$, $\sigma^+_{i.}\overset{\mathit{def}}{=}\sum_j\sigma^2_{ij}$, $\sigma^+_{.j}\overset{\mathit{def}}{=}\sum_i\sigma^2_{ij}$, and let $\overline{\mu}^+_{i.}\overset{\mathit{def}}{=}\mu^+_{i.}/N$ (and so on for the other averages $\overline{\sigma}^+_{i.},\overline{\sigma}^+_{.j}$). Finally, let $\tilde{\mathrm{p}}^\mu_{i.}\overset{\mathit{def}}{=}\sum_j p_{ij}\mu^2_{ij}/\mu^+_{i.}$, $\tilde{\mathrm{p}}^\mu_{.j}\overset{\mathit{def}}{=}\sum_i p_{ij}\mu^2_{ij}/\mu^+_{.j}$ and*

$$\begin{aligned}
\nu^r_i &\overset{\mathit{def}}{=} \overline{\mu}^+_{i.}\cdot\phi(\tilde{p}^\mu_{i.})+\overline{\sigma}^+_{i.}\ , \quad (45)\\
\nu^c_j &\overset{\mathit{def}}{=} \overline{\mu}^+_{.j}\cdot\phi(\tilde{p}^\mu_{.j})+\overline{\sigma}^+_{.j}\ , \quad (46)
\end{aligned}$$

where $\phi(p)\overset{\mathit{def}}{=}2\sqrt{p(1-p)}$ is (normalized) Matsushita's entropy. Then the following holds for any $t>0$:

$$\begin{aligned}
\mathbb{P}\left[\sum_i(\mathsf{W}_{ij}-p_{ij}\mu_{ij})\not\in(-Nt,Nt)\right] &\leq 2\exp\left(-\frac{Nt^2}{2\nu^c_j}\right)\ , \quad (47)\\
\mathbb{P}\left[\sum_j(\mathsf{W}_{ij}-p_{ij}\mu_{ij})\not\in(-Nt,Nt)\right] &\leq 2\exp\left(-\frac{Nt^2}{2\nu^r_i}\right)\ . \quad (48)
\end{aligned}$$

Proof: Since the sum of N independent random variables respectively (k,β_i)-sub-Gaussian ($i\in[N]$) brings a $(k,\sum_i\beta_i)$ sub-Gaussian random variable, Theorem C immediately yields:

$$\mathbb{P}\left[\frac{1}{N}\sum_j(\mathsf{W}_{ij}-\mathbb{E}[\mathsf{W}_{ij}])\geq t\right] \leq \exp\left(-\frac{Nt^2}{2\cdot\frac{1}{N}\sum_j(2\sqrt{p_{ij}(1-p_{ij})}\mu^2_{ij}+\sigma^2_{ij})}\right)\ . \quad (49)$$

Since $p\mapsto\sqrt{p(1-p)}$ is concave, we have:

$$\begin{aligned}
\sum_j\sqrt{p_{ij}(1-p_{ij})}\mu^2_{ij} &= \mu^+_{i.}\cdot\sum_j\frac{\mu^2_{ij}}{\mu^+_{i.}}\cdot\sqrt{p_{ij}(1-p_{ij})}\\
&\leq \mu^+_{i.}\cdot\sqrt{\tilde{p}^\mu_{i.}\,(1-\tilde{p}^\mu_{i.})}\ . \quad (50)
\end{aligned}$$

We finally obtain using ineq. (50),

$$\mathbb{P}\left[\frac{1}{N}\sum_j(\mathsf{W}_{ij}-\mathbb{E}[\mathsf{W}_{ij}])\geq t\right] \leq \exp\left(-\frac{Nt^2}{2\cdot(\overline{\mu}^+_{i.}\cdot 2\sqrt{\tilde{p}^\mu_{i.}(1-\tilde{p}^\mu_{i.})}+\overline{\sigma}^+_{i.})}\right)\ , \quad (51)$$

and we would obtain by symmetry:

$$\mathbb{P}\left[\frac{1}{N}\sum_j(\mathsf{W}_{ij}-\mathbb{E}[\mathsf{W}_{ij}])\leq -t\right] \quad\leq\quad \exp\left(-\frac{Nt^2}{2\cdot(\overline{\mu}^+_{i.}\cdot 2\sqrt{\tilde{p}^{\mu}_{i.}(1-\tilde{p}^{\mu}_{i.})}+\overline{\sigma}^+_{i.})}\right) \tag{52}$$

as well. This ends the proof of Lemma E. □
Let us define function $E:\{1,2,...,2N\}\rightarrow\mathbb{R}_+$ with:

$$E(i) \quad\overset{\mathrm{def}}{=}\quad \left\{\begin{array}{ll} 2\sqrt{\tilde{p}^{\mu}_{i.}(1-\tilde{p}^{\mu}_{i.})}\cdot\overline{\mu}^+_{i.}+\overline{\sigma}^+_{i.} & \text{if} \quad i\leq N\ , \\ 2\sqrt{\tilde{p}^{\mu}_{.(i-N)}(1-\tilde{p}^{\mu}_{i.})}\cdot\overline{\mu}^+_{.(i-N)}+\overline{\sigma}^+_{.(i-N)} & \text{otherwise} \end{array}\right. , \tag{53}$$

which collects the key parts in the concentration inequalities for row- / column-sums. We need in fact slightly more than Lemma E, as we do not just want to bound row- or column-sums, but we need to bound their L_1 norms (which, since $\|\mathbf{u}\|_1\geq|\mathbf{1}^\top\mathbf{u}|$ by the triangle inequality, yields a bound on row- or column-sums). It can be verified that $|\mathsf{W}_{ij}|$ is $(2,\beta)$-sug-Gaussian with the same β as for W_{ij}, but because $|\mathsf{W}_{ij}|$ now integrates a folded Gaussian random variable (Tsagris et al., 2014) instead of a Gaussian, its expectation is non trivial. We have not found any (simple) bound on the expectation of such a folded Gaussian, so we provide a complete one here for W_{ij}, which integrates as well Bernoulli parameter p_{ij}.

Lemma F *We have:*

$$\mathbb{E}[|\mathsf{W}_{ij}|] \quad\leq\quad p_{ij}\cdot\left(|\mu_{ij}|+\frac{1}{\gamma}\cdot\frac{\sigma^2_{ij}}{\sigma_{ij}+|\mu_{ij}|}\right)\ , \tag{54}$$

where $\gamma\overset{def}{=}\sqrt{\pi/2}$. Furthermore, (54) is optimal in the sense that both sides coincide when $\mu_{ij}=0$ (in this case, $\mathbb{E}[|\mathsf{W}_{ij}|]=\sigma_{ij}/\gamma$).

Proof: We now have (removing indices for readability, (Tsagris et al., 2014)):

$$\mathbb{E}[|\mathsf{W}|] \quad=\quad p\left(\sqrt{\frac{2}{\pi}}\cdot\sigma\exp\left(-\frac{\mu^2}{2\sigma^2}\right)+\mu\left[1-2\Phi\left(-\frac{\mu}{\sigma}\right)\right]\right)\ , \tag{55}$$

where Φ is the CDF of the standard Gaussian, so it is clear that the statement of the Lemma holds (and is in fact tight) when $\mu=0$, as in this case $\mathbb{E}[|\mathsf{W}|]=\sigma\sqrt{2/\pi}$. Otherwise, assume $\mu\neq 0$. For any $z>0$, let

$$f(z) \quad\overset{\mathrm{def}}{=}\quad \frac{1}{1+\sqrt{1+\frac{4}{z^2}}}\cdot\left(\sqrt{\frac{2}{\pi}}\cdot\frac{1}{z}\exp\left(-\frac{z^2}{2}\right)\right)\ , \tag{56}$$

where $u>0$ is a constant. It comes from Abramowitz & Stegun (1964, Inequality 7.1.3):

$$\Phi(z) \quad\leq\quad 1-f(z)\ , \tag{57}$$

and so, if $\mu<0$,

$$\begin{aligned}
\mathbb{E}[|\mathsf{W}|] \quad&=\quad p\left(\mu+\sqrt{\frac{2}{\pi}}\cdot\sigma\exp\left(-\frac{\mu^2}{2\sigma^2}\right)-2\mu\Phi\left(-\frac{\mu}{\sigma}\right)\right)\\
&\leq\quad p\left(-\mu+\sqrt{\frac{2}{\pi}}\cdot\sigma\exp\left(-\frac{\mu^2}{2\sigma^2}\right)+2\mu f\left(-\frac{\mu}{\sigma}\right)\right)\\
&\qquad=p\left(|\mu|+\sqrt{\frac{2}{\pi}}\cdot\sigma\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\left[1-\frac{2}{1+\sqrt{1+\frac{4\sigma^2}{\mu^2}}}\right]\right)\ ,
\end{aligned} \tag{58}$$

and we would obtain the same bound for $\mu > 0$. There just remains to remark that ($\forall z > 0$):

$$\begin{aligned} 1 - \frac{2}{1+\sqrt{1+\frac{1}{z}}} &\leq 1 - 2\sqrt{z} + 2z \ , \\ \left(1 - z + \frac{z^2}{2}\right) \cdot \exp\left(-\frac{z^2}{2}\right) &\leq \frac{1}{1+z} \ , \end{aligned}$$

and we obtain the statement of Lemma F. □

Using Lemma F , we can extend Lemma E and obtain the following Lemma.

Lemma G *Let* $E^\star \overset{def}{=} \max_i E(i)$ *and* $\mathbb{A}$ *denote the event:*

$$\mathbb{A} \equiv \left(\exists j \in [N] : \|\boldsymbol{c}_j\|_1 > \sum_i p_{ij}(|\mu_{ij}| + \delta_{ij}) + Nt\right) \vee \left(\exists i \in [N] : \|\boldsymbol{r}_i\|_1 > \sum_j p_{ij}(|\mu_{ij}| + \delta_{ij}) + Nt\right) \tag{59}$$

Then for any $t > 0$*,*

$$\mathbb{P}\left[\mathbb{A}\right] \leq 4N \exp\left(-\frac{Nt^2}{2E^\star}\right) \ , \tag{60}$$

where $\boldsymbol{r}_i \overset{def}{=} (\mathbf{B}\boldsymbol{1})_i$ *and* $\boldsymbol{c}_j \overset{def}{=} (\mathbf{B}^\top \boldsymbol{1})_j$ *are respectively row- and column-sums in* $\mathbf{B}$*,* $\delta_{ij} \overset{def}{=} \sigma_{ij}^2/(\sigma_{ij} + \gamma|\mu_{ij}|)$ *and* $\gamma \overset{def}{=} \sqrt{\pi/2}$*.*

The way we use Lemma G is the following: pick

$$t = \sqrt{\frac{2E^\star}{N} \cdot \log\frac{4N}{\delta}} \ . \tag{61}$$

We get that with probability $\geq 1 - \delta$, we shall have both

$$\|\mathbf{c}_j\|_1 \leq \sum_i \tilde{b}_{ij} + \sqrt{2E^\star N \cdot \log\frac{8N}{\delta}} \ , \forall j \in [N] \ , \tag{62}$$

$$\|\mathbf{r}_i\|_1 \leq \sum_j \tilde{b}_{ij} + \sqrt{2E^\star N \cdot \log\frac{8N}{\delta}} \ , \forall i \in [N] \ , \tag{63}$$

for all columns and rows in $\mathbf{B}$, with $\tilde{b}_{ij} \overset{\mathrm{def}}{=} p_{ij}(|\mu_{ij}| + \delta_{ij})$. There is a balance between the two summands in (62), (63) that we need to clarify to handle the upperbounds. This is achieved through the following Lemma.

Lemma H *For any* i, j*,*

$$\begin{aligned} \frac{1}{N}\sum_j \tilde{b}_{ij} &\leq \sqrt{2\tilde{p}^\mu_{i.}\overline{\mu}^+_{i.} + 2\tilde{p}^\sigma_{i.}\overline{\sigma}^+_{i.}} \ , \\ \frac{1}{N}\sum_i \tilde{b}_{ij} &\leq \sqrt{2\tilde{p}^\mu_{.j}\overline{\mu}^+_{.j} + 2\tilde{p}^\sigma_{.j}\overline{\sigma}^+_{.j}} \ , \end{aligned}$$

where $\tilde{p}^\sigma_{i.} \overset{def}{=} \sum_j p_{ij}\sigma^2_{ij}/\sigma^+_{i.}$, $\tilde{p}^\sigma_{.j} \overset{def}{=} \sum_i p_{ij}\sigma^2_{ij}/\sigma^+_{.j}$.

Proof: We have for any i, j,

$$\begin{aligned}
\left(\frac{1}{N}\sum_j \tilde{b}_{ij}\right)^2 &= \left(\frac{1}{N}\cdot\sum_j p_{ij}|\mu_{ij}|\left(1+\frac{\sigma_{ij}}{|\mu_{ij}|}\cdot\frac{1}{1+\gamma\frac{|\mu_{ij}|}{\sigma_{ij}}}\right)\right)^2\\
&\leq \left(\frac{1}{N}\cdot\sum_j p_{ij}|\mu_{ij}| + \frac{1}{N}\cdot\sum_j p_{ij}\sigma_{ij}\right)^2\\
&\leq 2\left(\frac{1}{N}\cdot\sum_j p_{ij}|\mu_{ij}|\right)^2 + 2\left(\frac{1}{N}\cdot\sum_j p_{ij}\sigma_{ij}\right)^2 \qquad (64)\\
&\leq 2\sum_j p_{ij}^2\mu_{ij}^2 + 2\sum_j p_{ij}^2\sigma_{ij}^2 \qquad (65)\\
&\leq 2\sum_j p_{ij}\mu_{ij}^2 + 2\sum_j p_{ij}\sigma_{ij}^2 \qquad (66)\\
&= 2\tilde{p}^{\mu}_{i.}\overline{\mu}^{+}_{i.} + 2\tilde{p}^{\sigma}_{i.}\overline{\sigma}^{+}_{i.}\ . \qquad (67)
\end{aligned}$$

Ineqs (64) and (65) follows from $(\sum_{u=1}^{v} a_u)^2 \leq v\sum_u a_u^2$. Ineq. (66) comes from $p_{ij}\in[0,1]$. We would have similarly

$$\left(\frac{1}{N}\sum_i \tilde{b}_{ij}\right)^2 \leq 2\tilde{p}^{\mu}_{.j}\overline{\mu}^{+}_{.j} + 2\tilde{p}^{\sigma}_{.j}\overline{\sigma}^{+}_{.j}\ . \qquad (68)$$

This ends the proof of Lemma H. □

II.3.2. Proof of Theorem B

Let us define function $U : \{1, 2, ..., 2N\} \rightarrow \mathbb{R}_+$ with:

$$U(i) = \left\{\begin{array}{ll} 2\tilde{p}^{\mu}_{i.}\overline{\mu}^{+}_{i.} + 2\tilde{p}^{\sigma}_{i.}\overline{\sigma}^{+}_{i.} & \text{if} \quad i\leq N\ ,\\ 2\tilde{p}^{\mu}_{.(i-N)}\overline{\mu}^{+}_{.(i-N)} + 2\tilde{p}^{\sigma}_{.(i-N)}\overline{\sigma}^{+}_{.(i-N)} & \text{otherwise}\end{array}\right. , \qquad (69)$$

which collects the bounds in ineqs (67) and (68), and let $U^\star \stackrel{\mathrm{def}}{=} \max_i U(i)$. Let

$$\ell \stackrel{\mathrm{def}}{=} N\sqrt{U^\star} + \sqrt{2E^\star N\cdot\log\frac{4N}{\delta}}\ . \qquad (70)$$

ℓ is be the quantity we need to handle all eigenspectra, but for this objective, let us define assumption (Z) as:

(Z) $(1+\epsilon)N\sqrt{U^\star}\leq 1/5$ (call it the domination assumption for short) *and*

$$\frac{E^\star}{U^\star} \leq \epsilon^2\cdot\frac{N}{2\log\frac{4N}{\delta}}\ . \qquad (71)$$

Assumption (Z) is a bit technical: we replace it by a simpler one, (A), which implies (Z). Suppose $\gamma\in(0,1)$ a constant, and assume $N\geq K^{1/(1-\gamma)}$ without loss of generality; fix for some *constant* $c>0$,

$$\begin{aligned}
\delta &= \frac{1}{N^c}\ , \qquad (72)\\
\epsilon^2 &= \frac{2}{N^{1-\gamma}}\cdot\log\frac{4N}{\delta}\\
&\geq \frac{2(c+4)}{N^{1-\gamma}}\cdot\log N\ . \qquad (73)
\end{aligned}$$

Condition (71) is now ensured provided

$$U^\star \quad \geq \quad \frac{1}{N^\gamma} \cdot E^\star \ , \tag{74}$$

while the domination condition is ensured, with $N = \Omega(c^{2+\kappa})$ ($\kappa > 0$ a constant) large enough so that $\epsilon \leq 1$, as long as

$$U^\star \quad \leq \quad \frac{1}{100N^2} \ . \tag{75}$$

So let us simplify assumption (Z) by the following assumption, which implies (Z):

(A) $c > 0$ and $0 < \gamma < 1$ being constants such that $N = \Omega(poly(c), 3^{1/(1-\gamma)})$, we have:

$$U^\star \quad \in \quad \left[\frac{E^\star}{N^\gamma}, \frac{1}{100N^2}\right] \ . \tag{76}$$

Again, (A) implies (Z).

Remark 1: the upperbound of (76) is quantitatively not so different from Linderman & Adams (2014)'s assumptions. They work with two assumptions, the first of which being

$$\sigma^2 \quad \leq \quad \frac{1}{N} \tag{77}$$

(we consider variances for the assumption to rely on same scales as ours), and also pick network parameters μ, σ in such a way that large deviations for edge weights are controlled with high probability, with a condition that roughly looks like:

$$\mu^2 + \frac{c}{N^2} \cdot \sigma^2 \quad = \quad O\left(\frac{1}{N^2}\right) \ , \tag{78}$$

for some constant $c > 3$. This constraint is relevant to the same stability issues as the ones we study here, and can be found in a slightly different form (but equivalent) in Hyvärinen & Smith (2013, Section 4), where it is mandatory for the estimation of ICA model parameters.
Finally, Linderman & Adams (2014) make the heuristic choice to enforce at least one of the two ineqs. (77, 78).

Remark 2: the sampling constraint akin to eq. (78) is in fact very restrictive for ICA estimation of models (Hyvärinen & Smith, 2013, Section 4), since typically **each** coordinate in $\mathbf{B}$ has to be bounded with high probability, whereas in our case, it is sufficient to control **sums** (L_1, row- or column-wise) with high probability. We can therefore benefit from concentration properties on large networks that such approaches may not have.

What is interesting from (76) is the hints that provide the *lowerbound* of (76) for Theorem B (main file) to hold. The main difference between $U^\star$ and $E^\star$ is indeed (omitting factor $2 \cdot \tilde{p}^\sigma \in [0,1]$ in variance terms) the switch between $z \mapsto 2z$ (for $U(.)$) and $z \mapsto \phi(z)$ (for $E(.)$). Figure 3 explains that the lowerbound may be violated essentially only on networks with very unlikely arcs almost everywhere, because ϕ has infinite derivative[3] as $z \rightarrow 0$. Also, it gives a justification for the name of the two functions E and U, where maximizing E tends to favor arcs with p close to $1/2$ (E stands for Equivocal), while maximizing U tends to favor arcs with p close to 1 (U stands for Unequivocal).

[3] And it seems that such entropy-like penalties with infinite derivatives in a neighborhood of zero are necessary to obtain Lemma D — as explained in the Lemma — if we want to keep the sub-Gaussian characterization of the W_{ij}s.

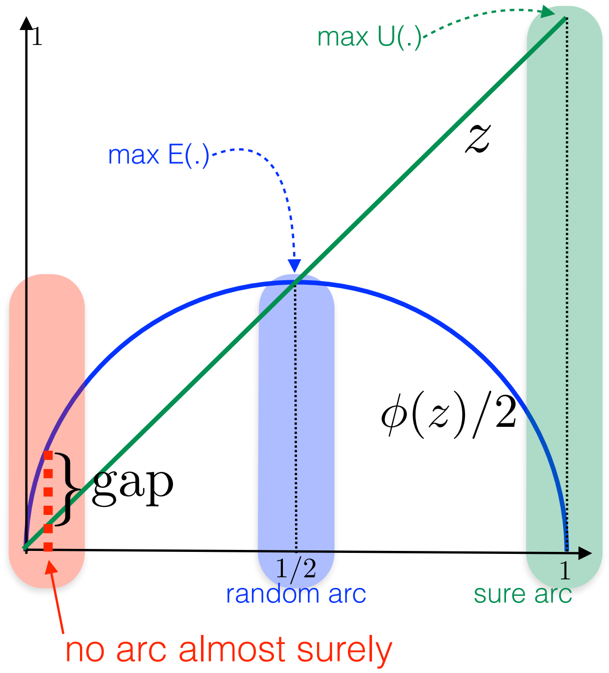

Figure 3. Region (red cartouche) for which the lowerbound in (76) may fail because $\phi(z)$ happens to be much larger than $2z$ — a worst case corresponding to networks where basically all ps are very small (*e.g.* $o(1/poly(N))$). The Figure also depicts the location of p for "ideal" maximizers of $E(.)$ (hence the name, Equivocal arcs) and $U(.)$ (hence the name, Unequivocal arcs).

($\star$) We now have all we need to bound the eigenspectra of $\mathbf{H}, \mathbf{H}'$. Let $\lambda^{\uparrow}(.)$ (resp. $\lambda^{\downarrow}(.)$) denote the maximal (resp. minimal) eigenvalue of the argument matrix. We obtain that with probability $\geq 1-\delta$,

$$\begin{aligned}
\lambda^{\uparrow}(\mathbf{H}) &\leq 1+\max_j \mathbf{c}_j^\top \sum_k \mathbf{c}_k - (\mathbf{r}_j+\mathbf{c}_j)^\top \mathbf{1} \\
&\leq 1+\max_j \|\mathbf{c}_j\|_1 \max_k |\mathbf{1}^\top \mathbf{r}_k| - (\mathbf{r}_j+\mathbf{c}_j)^\top \mathbf{1} \qquad (79)\\
&\leq 1+\max_j \|\mathbf{c}_j\|_1 \max_i \|\mathbf{r}_i\|_1 + \max_j \|\mathbf{c}_j\|_1 + \max_i \|\mathbf{r}_i\|_1 \\
&\leq (1+\ell)^2 \quad ,
\end{aligned}$$

(ineq. (79) comes from Hölder inequality) and similarly for the minimal eigenvalue,

$$\begin{aligned}
\lambda^{\downarrow}(\mathbf{H}) &\geq 1+\min_j \mathbf{c}_j^\top \sum_k \mathbf{c}_k - (\mathbf{r}_j+\mathbf{c}_j)^\top \mathbf{1} \\
&\geq 1-\ell^2-2\ell \quad ,
\end{aligned}$$

which implies that $\ell \leq \sqrt{2}-1$ for this latter bound not to be vacuous (ℓ is defined in eq. (70)). As long as $\delta = \Omega(1/poly(N))$, it is not hard to see that $N\sqrt{U^\star}$ dominates in ℓ for large networks so we can assume N large enough so that, for some small $\epsilon > 0$,

$$\frac{E^\star}{U^\star} \leq \epsilon^2 \cdot \frac{N}{2\log\frac{4N}{\delta}} \quad , \qquad (80)$$

which brings $\ell \leq (1+\epsilon)N\sqrt{U^\star}$. In this case, if $(1+\epsilon)N\sqrt{U^\star} \leq 1/5$, then $\lambda^{\downarrow}(\mathbf{H}) \geq 1/2$. Furthermore, it is not hard to check that we also get $\lambda^{\uparrow}(\mathbf{H}) \leq 3/2$. To summarize, as long as assumption (Z) (and so, as long as (A)) holds, the *complete* eigenspectra of $\mathbf{H}$, $\mathbf{H}^{-1}$ and by extension $\mathbf{H}'$, $\mathbf{H}'^{-1}$, all lie within $[1/2, 2]$ with high probability.

($\star$) We finish with the eigenspectrum of $\mathbf{J}$. We also easily obtain that

$$\begin{aligned}\lambda^{\uparrow}(\mathbf{J}) &\leq \max_j \mathbf{r}_j^\top \sum_k \mathbf{r}_k \\ &\leq \max_j \|\mathbf{r}_j\|_1 \max_k |\mathbf{1}^\top \mathbf{c}_k| \\ &\leq \ell^2 \ ,\end{aligned}$$

and obviously $\lambda^{\downarrow}(\mathbf{J}) \geq 0$, which is all we need.

($\star$) We now finish the proof of Theorem B, recalling that $\mathbf{\Sigma}_y$ can be summarized as:

$$\mathbf{\Sigma}_y = \mathbf{A} + \sigma_f^2 \mathbf{B} + \sigma_y^2 \mathbf{I} \ , \tag{81}$$

with $\mathbf{A} \stackrel{\mathrm{def}}{=} \mathbf{H}^{-1} \otimes \mathbf{K}_t$ has an eigensystem which is the (Minkowski) product of the eigensystems of its two matrices, and therefore is within $[\lambda^{\downarrow}(\mathbf{K}_t)/2, 2\lambda^{\uparrow}(\mathbf{K}_t)]$; on the other hand, $\mathbf{B} \stackrel{\mathrm{def}}{=} (\mathbf{I}-\mathbf{B})^{-1}\mathbf{J}(\mathbf{I}-\mathbf{B})^{-\top} \otimes \mathbf{\Sigma}_{\mathrm{I}}$ has eigensystem which is the one of $\mathbf{JH}'$ (eigenvalues have different algebraic multiplicity though), which therefore is within $[0, \ell^2 \cdot (1+\ell)^2] \subset [0, 2/25]$. Hence, simplifying a bit, we can bound the complete eigenspectrum of $\mathbf{\Sigma}_y$, $\lambda(\mathbf{\Sigma}_y)$, as:

$$\lambda(\mathbf{\Sigma}_y) \subset \left[\frac{\lambda^{\downarrow}(\mathbf{K}_t)}{2} + \sigma_y^2, 2\lambda^{\uparrow}(\mathbf{K}_t) + \sigma_f^2 + \sigma_y^2\right] \ , \tag{82}$$

under assumption (A), with probability $\geq 1 - \delta = 1 - 1/N^c$, as claimed. This ends the proof of Theorem B.

II.3.3. From Theorem B to Theorem 4

We now assume $\mathsf{A} \sim \mathcal{B}(\rho_{ij})$ with $\rho_{ij} \sim \mathcal{V}_{ij}(p_{ij})$, where $\mathcal{V}$ is a random variable satisfying $p_{ij} \stackrel{\mathrm{def}}{=} \mathbb{E}[\mathcal{V}_{ij}]$ and $\mathrm{supp}(\mathcal{V}) \subseteq [0,1]$ (the support of $\mathcal{V}$). The proof essentially follows that of Theorem B, with the following minor changes.

($\star$) The derivation of eq. (26) now satisfies, since $\phi_u(z)$ is maximal in $z = 1/2$,

$$\begin{aligned}\mathbb{E}_{\mathsf{W}_{ij}}[\exp(\lambda(\mathsf{W}_{ij} - \mathbb{E}[\mathsf{W}_{ij}]))] &\leq \int_{\mathrm{Supp}(\mathcal{V})} \exp\left(\frac{(\sigma_{ij}^2 + 2\phi_u(z)\mu_{ij}^2)\lambda^2}{2}\right) \mathrm{d}\mu(z) \\ &\leq \exp\left(\frac{(\sigma_{ij}^2 + 2\phi_u(1/2)\mu_{ij}^2)\lambda^2}{2}\right) = \exp\left(\frac{(\sigma_{ij}^2 + \mu_{ij}^2)\lambda^2}{2}\right) \ . \end{aligned} \tag{83}$$

($\star$) Assumption (A) now reads, for some constants $c > 0$ and $0 < \gamma < 1$ such that $N = \Omega(poly(c), 3^{1/(1-\gamma)})$, we have:

$$U^\star \in \left[\frac{S^\star}{N^\gamma}, \frac{1}{100N^2}\right] \ , \tag{84}$$

where U does not change but

$$S(i) \stackrel{\mathrm{def}}{=} \left\{\begin{array}{ll} \overline{\mu}_{i.}^{+} + \overline{\sigma}_{i.}^{+} & (i \leq N) \\ \overline{\mu}_{.j}^{+} + \overline{\sigma}_{.j}^{+} : j \stackrel{\mathrm{def}}{=} N - i & (i > N) \end{array}\right. .$$

II.4. Marginal Likelihood Given the Network Parameters

We first rewrite the prior in eq. 2 in the main paper as:

$$
\begin{aligned}
f_i(t) &= z_i(t) + \sum_{\substack{j=1\\ j\neq i}}^{N} A_{ij} W_{ij} \left[f_j(t) + \xi_{jt} \right], && (85)\\
f_i(t) &= z_i(t) + \sum_{\substack{j=1\\ j\neq i}}^{N} A_{ij} W_{ij} f_j(t) + \sum_{\substack{j=1\\ j\neq i}}^{N} A_{ij} W_{ij} \xi_{jt}, && (86)\\
f_i(t) - \sum_{\substack{j=1\\ j\neq i}}^{N} A_{ij} W_{ij} f_j(t) &= z_i(t) + \sum_{\substack{j=1\\ j\neq i}}^{N} A_{ij} W_{ij} \xi_{jt}, && (87)\\
(\mathbf{I} - \mathbf{A} \odot \mathbf{W}) \mathbf{f}(t) &= \mathbf{z}(t) + \mathbf{A} \odot \mathbf{W} \boldsymbol{\xi}_t, && (88)\\
\mathbf{f}(t) &= (\mathbf{I} - \mathbf{A} \odot \mathbf{W})^{-1} (\mathbf{z}(t) + \mathbf{A} \odot \mathbf{W} \boldsymbol{\xi}_t). && (89)
\end{aligned}
$$

In the main paper we refer to eq. 89 as the *inverse model*. With this, it is easy to see that we can write our complete model as:

$$
\begin{aligned}
z_i(t) &\sim \mathcal{GP}(\mathbf{0}, \kappa(t, t'; \boldsymbol{\theta})), && (90)\\
\boldsymbol{\xi}_t &\sim \mathcal{N}(0, \sigma_f^2 \mathbf{I}), && (91)\\
\epsilon_{it} &\sim \mathcal{N}(0, \sigma_y^2), && (92)\\
f_i(t) &= [\mathbf{G}]_{i,:} \left(\mathbf{z}(t) + \mathbf{B} \boldsymbol{\xi}_t \right), && (93)\\
y_i(t) &= f_i(t) + \epsilon_{it}. && (94)
\end{aligned}
$$

where $\mathbf{B} = \mathbf{A} \odot \mathbf{W}$; $\mathbf{G} = (\mathbf{I} - \mathbf{B})^{-1}$; $[\mathbf{M}]_{i,:}$ denotes the ith row of matrix $\mathbf{M}$. Here we analyse the conditional likelihood by integrating out everything but $\mathbf{A}, \mathbf{W}$. Clearly, for fixed $\mathbf{A}, \mathbf{W}$, since all the distributions are Gaussians, and we are only applying linear operators, the resulting distribution over f_i, and consequently over y_i, is also a Gaussian process. Hence, we only need to figure out the mean function and the covariance function of the resulting process. For the expectation we have that:

$$
\mu_i(t) = \mathbb{E}\left[f_i(t) \right] = 0, \qquad (95)
$$

since both $\mathbf{z}$ and ξ_{jt} are zero-mean processes. For the covariance function we have that:

$$
\begin{aligned}
\mathbb{C}\mathrm{ov}[f_i(t), f_j(t')] &= \mathbb{E}\left[(f_i(t) - \mu_i(t))(f_j(t) - \mu_j(t')) \right] && (96)\\
&= [\mathbf{G}\mathbf{G}^T]_{i,j} \kappa(t, t'; \boldsymbol{\theta}) + [\mathbf{G}\mathbf{B}\mathbf{B}^T\mathbf{G}^T]_{i,j} \sigma_f^2 && (97)\\
&= [\mathbf{K}_f]_{i,j} \kappa(t, t'; \boldsymbol{\theta}) + [\mathbf{E}]_{i,j} \sigma_f^2, && (98)
\end{aligned}
$$

where we have defined $[\mathbf{M}]_{i,j}$ the i, j entry of matrix $\mathbf{M}$ and the matrix of latent node covariances and noise covariances as:

$$
\begin{aligned}
\mathbf{K}_f &= \mathbf{G}\mathbf{G}^T && (99)\\
\mathbf{E} &= \mathbf{G}\mathbf{B}\mathbf{B}^T\mathbf{G}^T. && (100)
\end{aligned}
$$

The covariance function of the observations is then given by:

$$
\mathbb{C}\mathrm{ov}[y_i(t), y_j(t')] = [\mathbf{K}_f]_{i,j} \kappa(t, t'; \boldsymbol{\theta}) + [\mathbf{E}]_{i,j} \sigma_f^2 + \delta_{ij} \sigma_y^2. \qquad (101)
$$

For further understanding of this model, let us assume that the observations lie on a grid in time, $t = 1, \ldots, T$ and $\mathbf{Y}$ is a $N \times T$ matrix of observations with $\mathbf{y} = \mathrm{vec}\mathbf{Y}$ hence the likelihood of all observations is:

$$
\begin{aligned}
p(\mathbf{y} | \mathbf{W}, \mathbf{A}) &= \mathcal{N}(\mathbf{y}; \mathbf{0}, \boldsymbol{\Sigma}_y), \text{ with} && (102)\\
\boldsymbol{\Sigma}_y &= \mathbf{K}_f \otimes \mathbf{K}_t + \mathbf{E} \otimes \sigma_f^2 \mathbf{I} + \mathbf{I} \otimes \sigma_y^2 \mathbf{I}, && (103)
\end{aligned}
$$

where $\otimes$ denotes the Kronecker product; If we use this setting then we obtain:

$$\mathbf{\Sigma}_y = \mathbf{K}_f \otimes \mathbf{K}_t + (\sigma_f^2 \mathbf{E} + \sigma_y^2 \mathbf{I}) \otimes \mathbf{I}. \tag{104}$$

Interestingly, the model above has been studied in statistics and in machine learning (see e.g. Bonilla et al., 2008; Rakitsch et al., 2013). Furthermore, inference and hyperparameter estimation can be done efficiently by exploiting properties of the Kronecker product, e.g. an evaluation of the marginal likelihood can be done in $\mathcal{O}(N^3 + T^3)$. Nevertheless, unless there is a substantial overlapping between the locations of the observations across the nodes (i.e. times), the Kronecker formulation becomes intractable.

Naïve Computational Cost Assuming the general case (i.e. non-grid observations), let us refer to $\mathbf{\Sigma}_y = \mathbf{K} + \sigma_y^2 \mathbf{I}$ as the covariance of the marginal process over $\mathbf{y}$, as induced by the covariance function in Equation (101), where $\mathbf{K}$ is the covariance matrix induced by the covariance function in Equation (98). Therefore, the prior, conditional likelihood, and marginal likelihood of the model are:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K}), \tag{105}$$

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma_y^2 \mathbf{I}), \tag{106}$$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{\Sigma}_y), \tag{107}$$

where we have omitted the dependencies of the above equation on the network parameters $\mathbf{A}, \mathbf{W}$. Because of the marginalization property of GPs it is easy to see that all the above distributions are n-dimensional, where $n = \sum_{i=1}^{N} T$, where T is the number of observations per node. Hence the cost of evaluating the exact marginal likelihood is $\mathcal{O}(n^3)$.

II.5. Efficient Computation of Marginal Likelihood Given Network Parameters

For simplicity, we consider here the synchronized case where all the N nodes in the network have T observations at the same times. i.e. the total number of observations is $n = N \times T$. Here we show an efficient expression for the log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{W}, \mathbf{A}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_y| - \frac{1}{2}\mathbf{y}^T \mathbf{\Sigma}_y^{-1} \mathbf{y}, \text{ where} \tag{108}$$

$$\mathbf{\Sigma}_y = \mathbf{K}_f \otimes \mathbf{K}_t + \mathbf{\Omega} \otimes \mathbf{\Sigma}_{\mathrm{I}}, \text{ with} \tag{109}$$

$$\mathbf{\Omega} = (\sigma_f^2 \mathbf{E} + \sigma_y^2 \mathbf{I}) \text{ and} \tag{110}$$

$$\mathbf{\Sigma}_{\mathrm{I}} = \mathbf{I} \tag{111}$$

The main difficulty of computing this expression is the calculation of the log determinant of an n dimensional matrix, as well as solving an n-dimensional system of linear equations. Our goal is to show that we never need to solve these operations on an n-dimensional matrix, which are $\mathcal{O}(n^3)$ but instead use $\mathcal{O}(N^3 + T^3)$ operations.

Given the eigen-decomposition of the above matrices

$$\mathbf{\Omega} = \mathbf{Q}_{\Omega} \mathbf{\Lambda}_{\Omega} \mathbf{Q}_{\Omega}^T \tag{112}$$

$$\mathbf{\Sigma}_{\mathrm{I}} = \mathbf{Q}_{\mathrm{I}} \mathbf{\Lambda}_{\mathrm{I}} \mathbf{Q}_{\mathrm{I}}^T = \mathbf{I}, \tag{113}$$

It is possible to show that the marginal covariance is given by

$$\mathbf{\Sigma}_y = (\mathbf{Q}_{\Omega} \mathbf{\Lambda}_{\Omega}^{1/2} \otimes \mathbf{Q}_{\mathrm{I}} \mathbf{\Lambda}_{\mathrm{I}}^{1/2}) \left(\tilde{\mathbf{K}}_f \otimes \tilde{\mathbf{K}}_t + \mathbf{I} \otimes \mathbf{I} \right) (\mathbf{Q}_{\Omega} \mathbf{\Lambda}_{\Omega}^{1/2} \otimes \mathbf{Q}_{\mathrm{I}} \mathbf{\Lambda}_{\mathrm{I}}^{1/2})^T, \text{ where} \tag{114}$$

$$\tilde{\mathbf{K}}_f = \mathbf{\Lambda}_{\Omega}^{-1/2} \mathbf{Q}_{\Omega}^T \mathbf{K}_f \mathbf{Q}_{\Omega} \mathbf{\Lambda}_{\Omega}^{-1/2} \tag{115}$$

$$\tilde{\mathbf{K}}_t = \mathbf{\Lambda}_{\mathrm{I}}^{-1/2} \mathbf{Q}_{\mathrm{I}}^T \mathbf{K}_t \mathbf{Q}_{\mathrm{I}} \mathbf{\Lambda}_{\mathrm{I}}^{-1/2} = \mathbf{K}_t \tag{116}$$

For these matrices we also define their eigen-decomposition analogously to above:

$$\tilde{\mathbf{K}}_f = \tilde{\mathbf{Q}}_f \tilde{\mathbf{\Lambda}}_f \tilde{\mathbf{Q}}_f^T \tag{117}$$

$$\tilde{\mathbf{K}}_t = \mathbf{K}_t = \tilde{\mathbf{Q}}_t \tilde{\mathbf{\Lambda}}_t \tilde{\mathbf{Q}}_t^T \tag{118}$$

II.5.1. Log-determinant Term

$$\begin{aligned}\log|\boldsymbol{\Sigma}_y| &= \log|\boldsymbol{\Omega}\otimes\boldsymbol{\Sigma}_{\mathrm{I}}| + \log|\tilde{\mathbf{K}}_f\otimes\tilde{\mathbf{K}}_t + \mathbf{I}\otimes\mathbf{I}| && (119)\\
&= T\sum_{i=1}^{N}\log\lambda_{\Omega}^{(i)} + N\sum_{j=1}^{T}\log\lambda_{\mathrm{I}}^{(j)} + \sum_{i=1}^{N}\sum_{j=1}^{T}\log(\tilde{\lambda}_f^{(i)}\tilde{\lambda}_t^{(j)}+1) && (120)\\
&= T\sum_{i=1}^{N}\log\lambda_{\Omega}^{(i)} + \sum_{i=1}^{N}\sum_{j=1}^{T}\log(\tilde{\lambda}_f^{(i)}\tilde{\lambda}_t^{(j)}+1) && (121)\end{aligned}$$

II.5.2. Quadratic Term

$$\mathbf{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y} = \mathbf{y}^T(\boldsymbol{\Lambda}_{\Omega}^{1/2}\mathbf{Q}_{\Omega}^T\otimes\boldsymbol{\Lambda}_{\mathrm{I}}^{1/2}\mathbf{Q}_{\mathrm{I}}^T)^{-1}\left(\tilde{\mathbf{K}}_f\otimes\tilde{\mathbf{K}}_t+\mathbf{I}\otimes\mathbf{I}\right)^{-1}(\mathbf{Q}_{\Omega}\boldsymbol{\Lambda}_{\Omega}^{1/2}\otimes\mathbf{Q}_{\mathrm{I}}\boldsymbol{\Lambda}_{\mathrm{I}}^{1/2})^{-1}\mathbf{y} \quad (122)$$

$$\mathbf{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y} = \mathbf{y}^T(\mathbf{Q}_{\Omega}\boldsymbol{\Lambda}_{\Omega}^{-1/2}\otimes\mathbf{Q}_{\mathrm{I}}\boldsymbol{\Lambda}_{\mathrm{I}}^{-1/2})\left(\tilde{\mathbf{K}}_f\otimes\tilde{\mathbf{K}}_t+\mathbf{I}\otimes\mathbf{I}\right)^{-1}(\boldsymbol{\Lambda}_{\Omega}^{-1/2}\mathbf{Q}_{\Omega}^T\otimes\boldsymbol{\Lambda}_{\mathrm{I}}^{-1/2}\mathbf{Q}_{\mathrm{I}}^T)\mathbf{y} \quad (123)$$

Let us define

$$\begin{aligned}\tilde{\mathbf{y}} &= (\boldsymbol{\Lambda}_{\Omega}^{-1/2}\mathbf{Q}_{\Omega}^T\otimes\boldsymbol{\Lambda}_{\mathrm{I}}^{-1/2}\mathbf{Q}_{\mathrm{I}}^T)\mathbf{y} && (124)\\
&= \mathrm{vec}(\boldsymbol{\Lambda}_{\mathrm{I}}^{-1/2}\mathbf{Q}_{\mathrm{I}}^T\mathbf{Y}\mathbf{Q}_{\Omega}\boldsymbol{\Lambda}_{\Omega}^{-1/2}) && (125)\\
&= \mathrm{vec}(\mathbf{Y}\mathbf{Q}_{\Omega}\boldsymbol{\Lambda}_{\Omega}^{-1/2}) && (126)\end{aligned}$$

Hence the quadratic form above becomes:

$$\begin{aligned}\mathbf{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{y} &= \tilde{\mathbf{y}}^T\left(\tilde{\mathbf{K}}_f\otimes\tilde{\mathbf{K}}_t+\mathbf{I}\otimes\mathbf{I}\right)^{-1}\tilde{\mathbf{y}} && (127)\\
&= \mathrm{tr}(\tilde{\mathbf{Y}}^T\tilde{\mathbf{Q}}_t\tilde{\mathbf{Y}}_{tf}\tilde{\mathbf{Q}}_f^T), \text{ where} && (128)\end{aligned}$$

$$[\tilde{\mathbf{Y}}_{tf}]_{i,j} = \left(\frac{1}{[\tilde{\boldsymbol{\lambda}}_t\tilde{\boldsymbol{\lambda}}_f^T+1]_{i,j}}\right)[\tilde{\mathbf{Q}}_t^T\tilde{\mathbf{Y}}\tilde{\mathbf{Q}}_f]_{i,j} \quad (129)$$

and $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\tilde{\mathbf{y}} = \mathrm{vec}(\tilde{\mathbf{Y}})$ are the vectors obtained by stacking the columns of the $T\times N$ matrices $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ respectively.

III. Supplementary material on experiments

As mentioned in the main paper, the choice of baseline comparisons was based on Peters et al. (2014). Other than the methods discussed in the main paper, there are four other methods considered by Peters et al. (2014): (1) Brute-force search; (2) Greedy DAG Search (GDS, see e.g. Chickering, 2002); (3) Greedy equivalence search (GES, Chickering, 2002; Meek, 1997); (4) Regression with subsequent independence test (RESIT, Peters et al., 2014).

In the experiments reported in section 5.1, since the ground truth is known, the evaluation criteria is AUC (area under the ROC curve). Calculating AUC values requires a discriminative threshold to generate ROCs. In the case of GDS and GES there was no clear parameter that could be considered as the discriminative threshold, and therefore results for these algorithms are not reported. In the case of RESIT, there is a threshold, but the threshold values for which the method produces different results were not provided, making it infeasible to calculate AUC, and therefore the output of this algorithm is not reported. In the experiments reported in section 5.2, the implementations of GES, GDS and RESIT that we used returned an error (possibly because the number of nodes was greater than the observations from each node). Therefore their results are not reported. Finally, for the experiment in section 5.3 we compared the results with CPC, which provided comparatively good performance in other experiments. Also, we did not include the brute-force method, which is not feasible to perform in networks with more than four nodes, and therefore makes it inapplicable in the experiments studied here.

The PC and CPC algorithms are constrain-based structure learning methods for directed acyclic graphs (DAG). The algorithms require a conditional independence test, for which we used the test for zero partial correlation between variables. The IAMB method is a two-phase algorithm for Markov blanket discovery. Linear correlation is used for the test of conditional independence required by this algorithm. The LiNGAM method is a Linear non-Gaussian Additive Model (LiNGAM) for estimating structural equation models. PW-LINGAM provides the direction of connection between the two connected nodes. We used partial correlation for determining whether two nodes are connected, and the magnitude of the correlation was used as the discriminative threshold. For connected nodes at the threshold PW-LINGAM was used to determine the direction of the connection.

For [PC, CPC GES], IAMB and LiNGAM implementations provided by R packages Kalisch et al. (2012), Marco (2010), Kalisch et al. (2012) were used respectively. For PW-LINGAM the code provided by the authors was re-implemented in R and was used. For GDS and RESIT implementation provided by authors of Peters et al. (2014) in R was used.

In the following, we define p_{ij} as the probability of the connection from node j to node i, which is calculated as follows:

$$p_{ij} = \frac{\alpha_{ij}}{\alpha_{ij} + 1}. \tag{130}$$

III.1. Prior setting and optimization specifics

We used the squared exponential covariance function and optimized variational parameters, hyperparameters, and likelihood parameters in an iterative fashion using Adam (Kingma & Ba, 2014). Similarly to Maddison et al. (2016), different λ_c values are used for the prior and posterior distributions. For experiments with $N > 15$, following Maddison et al. (2016) we used $\lambda_c = 0.5$ for priors and $\lambda_c = 2/3$ for posterior distributions. For the experiments in section 5.1, in which $N \leq 15$, we used the first subject ($T = 200$) as the validation data and selected $\lambda_c = 1.0$ for priors and $\lambda_c = 0.15$ for posterior distributions. The number of Monte Carlo samples was selected based on computational constraints, and were 200, 20 and 2 samples for small-scale (§5.1, 5.2), medium-scale (§5.3), and large-scale (§5.3) experiments respectively.

Prior over W_{ij} is assumed to be zero-mean Gaussian distribution with variance $\sigma_w^2 = 2/N$ similar to Linderman & Adams (2014). Prior over A_{ij} is assumed to be Concrete$(1, \lambda_c)$, which implies that the probability that a link exists between two nodes is 0.5:

$$p(A_{ij}) = \text{Concrete}(1, \lambda_c), \quad p(W_{ij}) = \mathcal{N}(0, 2/N). \tag{131}$$

III.2. Brain functional connectivity data

AUC computation. This is obtained by varying the discrimination threshold and drawing the false-positive rate (fpr) vs true-positive rate (tpr). In the case of LATNET, this threshold is the absolute expected value of the overall connection strength between the nodes ($|\mu_{ij} p_{ij}|$). In the case of PC, CPC and IAMB algorithms, the discrimination threshold is the p-value (target type I error rate) of the conditional independence test, and in the case of LiNGAM and PW-LINGAM, absolute values of the estimated linear coefficients and partial correlation coefficients are used as the discrimination thresholds respectively.

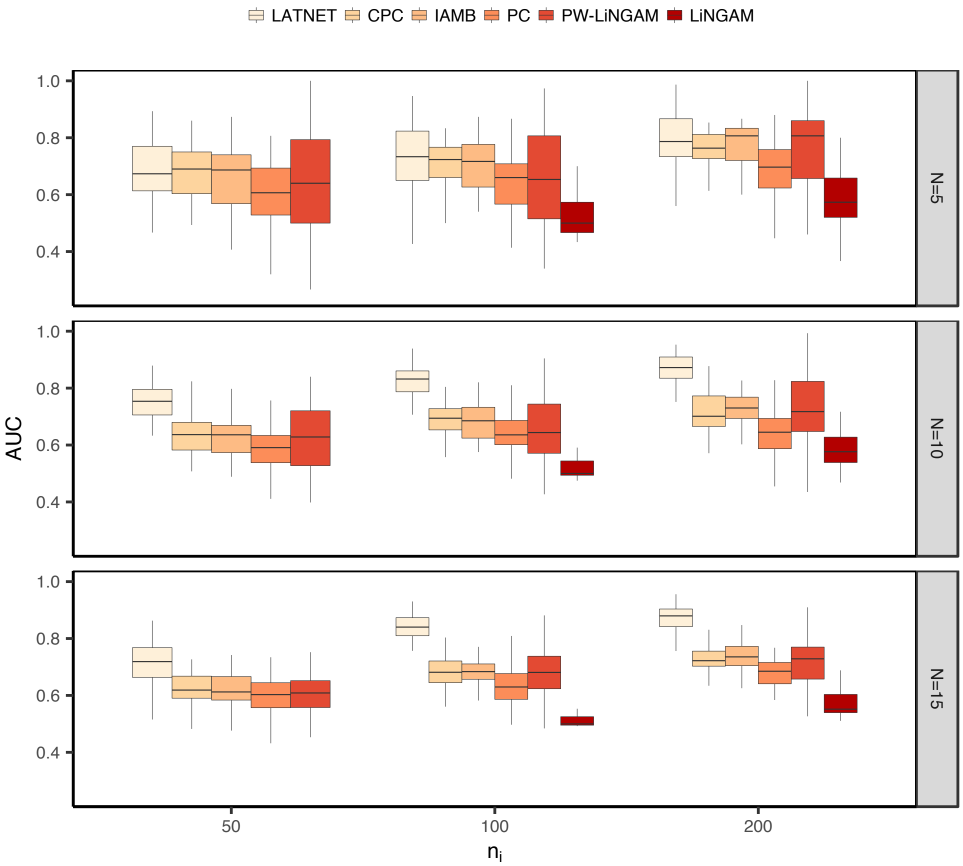

Figure 4. Performance of the methods in link prediction on the brain functional connectivity data in terms of AUC. N is the number of nodes in the network, and T is the number of observations in each node.

III.3. Spellman's sentinels of the yeast cell cycle

We have analyzed the signals of 799 (one gene was missing in our data, out of the 800 tagged in the original paper) sentinels of the yeast cell cycle (YCC) from Spellman et al. (1998), for a total of ≈13,600 data points. Figure 5 presents the counting histograms for μ and p found among all inferred arcs. Let us denote as *strong* arcs arcs that jointly belong to the red areas of both curves (meaning that both p is in top 99.9% quantile *and* $|\mu|$ in top 99% quantile). We remark that the scale for σ is roughly in the tenth of that for μ, so that for strong arcs, distributions with $|\mu|$ in its top 99% quantile can be considered encoding non-void arc connection (even when small in an absolute scale). We also notice that the distribution in p admits relatively large values (≈ 0.7), so that its top 99.9% percentile can be encoding arc probability strictly larger than $1/2$. In the *Left* picture of Figure 2, we plot *all* strong arcs with $p_{ij} > 0.62$; among these, the *very strong* arcs (displayed with thick arcs) are those for which $p > 0.65$ (vs $p \in (0.62, 0.65]$ for the other arcs displayed).

We have analyzed arcs belonging to at least one of these categories (p is in top 99.9% quantile *or* $|\mu|$ in top 99% quantile), the intersection of both representing strong arcs. Intuitively, this top list should contain most of the (much shorter) "A-lists" of cell-cycle genes as recorded in the litterature. One of these lists (Cho et al., 1998) has been curated and can be retrieved from Rowicka et al. (2007, Table 4 SI). It contains 106 genes. Table A1 gives the genes we retrieve, meaning that at least one *significant* arc appear for each of them (p is in top 99.9% quantile *or* $|\mu|$ in top 99% quantile). The values given in the Table allow to concluce that almost 68% of the 106 genes are retrieved as having at least one significant arc. Since the total number of genes with strong arcs we retrieve is 177, out of the 799, the probability that the result observed in Table A1 is due to chance is zero up to more than *thirty* digits. Hence, assuming the list of genes in Table A1 is indeed a most important one, we can conclude in the reliability of our technique for network discovery for this domain.

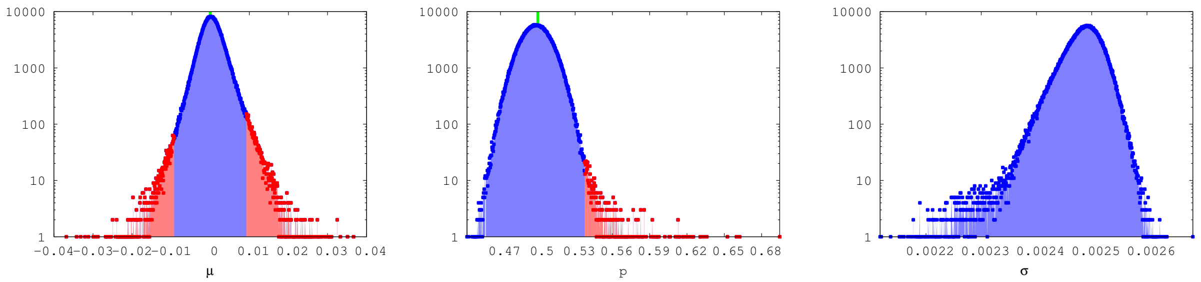

Figure 5. Counting histograms (y, blue + red) for the values of μ (left), p (center) and σ (right, y-scales are log-scales). The vertical green segment indicates $\mu = 0$ (left) and $p = .5$ (center). The red part displays the upper 99% percentile for $|\mu|$ (left) and upper 99.9% percentile for p (center).

G1(P)	FKS1, CLN3, CDC47, RAD54, PCL2, MNN1, RAD53, CLB5	8/16
G1/S	DPB2, CDC2, PRI2, POL12, CDC9, CDC45, CDC21, RNR1, CLB6, POL1, MSH2, RAD27, ASF1, POL30, RFA2, PMS1, MST1, RFA1, MSH6, SPC42, CLN2, PCL1, RFA3	23/28
S	MCD1, HTA2, SWE1, HTB1, KAR3, HSL1, HHF2, HHT1, HTB2, CIK1, CLB4	11/17
G2	CLB1, CLB2, BUD8, CDC5	4/4
G2/M	SWI5, CWP1, CHS2, FAR1, DBF2, MOB1, ACE2, CDC6	8/9
M(P)	CDC20	1/2
M(M)	TEC1, RAD51, NUM1	3/4
M(A)	TIP1, SWI4, KIN3, ASF2, ASH1, SIC1, PCL9, EGT2, SED1	9/15
M(T)	$\emptyset$	0/1
M/G1	PSA1, RME1, CTS1	3/3
G1	HO	1/4
late G1	$\emptyset$	0/3

Table A1. Genes found in at least one arc with p in top 99.9% quantile or $|\mu|$ in top 99% quantile, in the list of 106 documented genes of the cell cycle in Cho et al. (1998); Rowicka et al. (2007), as a function of the phase as defined in Cho et al. (1998) (left). The right column mentions the number of genes retrieved / total number of genes in the original list (for example, *all* G2 genes appear).

As a next step, Table A2 presents the breakdown for the relative distribution of *strong* arcs in the YCC as a function of the YCC phase, using as reference the original one from Spellman et al. (1998), collapsing the vertices in their respective phase of the YCC to obtain a concise graph of within and between phase dependences (Figure 6 gives a schematic view of the most significant part of the distribution — arcs between different genes of the *same* YCC phase create the loops observed). We can draw two conclusions: (i) the graph of dependences between phases is not symmetric. Furthermore, (ii) M and G1 appear as the phases which concentrate more than half of the strong arcs, which should be expected given the known regulatory importance in these two phases (Spellman et al., 1998). To make more precise in observation (i) that the network is indeed imbalanced, we have computed the ratio out-degree / in-degree for all genes admitting *strong* edges of both kinds (*i.e.* with the gene as in- / out- node). Table A3 presents all genes collected. A total of 100 genes is found, the majority of which (68) is imbalanced. We also remark that roughly 80% of them is associated to M and/or G1 (only 19 are associated to phases S or G2), which is consistent with the findings of Table A2.

To finish with the quantitative analyses, Tables A5 and A4 present the main *strong* genes in term of in or out degree (genes with in or out-degree < 3 are not shown). Notice the preeminence of two well known cell-cycle regulated genes, HO and WSC4.

To catch a glimpse at the overall network found from a more qualitative standpoint, we have learned a coordinate system for genes based on a popular manifold learning technique (Meila & Shi, 2001). Since this technique requires the graph to be symmetric, we have symmetrized the network by taking the max of the p-values to weight each edge. Figure 7 presents the results obtained, for the two leading coordinates — excluding the coordinate associated to eigenvalue 1, which encodes the stationary distribution of the Markov chain and is therefore trivial —. It is clear that the first coordinates splits key YCC genes from the rest of the crowd (*Cf* Tables A5 and A4), also highlighting the importance of WSC4 and strong edges to create the manifold. HO is used to switch mating type and WSC4 is required for maintenance of cell wall integrity (Simon

	M	M/G1	G1	S	G2
M	18	4	7	3	6
M/G1	5	7	3	0	1
G1	5	4	23	1	0
S	2	0	0	2	1
G2	4	1	0	0	0

Table A2. Distribution of strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile) with respect to phases in the YCC. Each entry has been rounded to the nearest integer for readability.

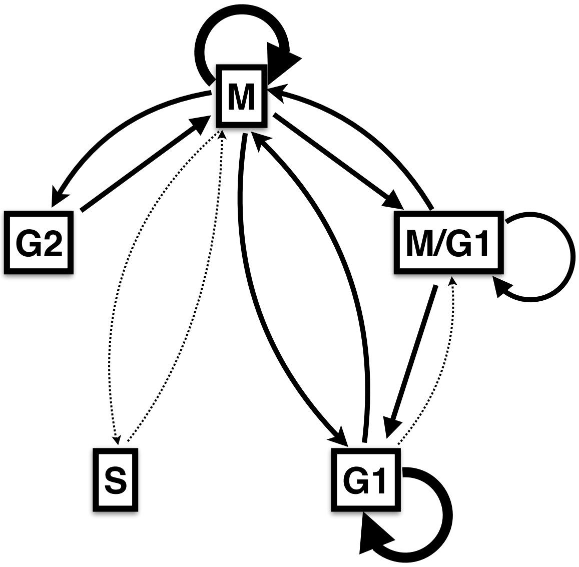

Figure 6. Distribution of strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile) with respect to phases in the YCC (clockwise), displayed as follows: thick plain $\geq 8\%$, plain $\in [4\%, 8\%)$, dashed $\in (2\%, 4\%)$. Reference values in Table A2.

et al., 2001).

Interestingly, the most prominent genes belong to a small set of chromosomes (essentially 13, 15, 16). What is quite striking is the fact that SPS4 and SFG1 are in fact neighbors on chromosome XV[4]. It is far beyond the scope of our paper to eventually relate the netork structure and associated causal influence in expression — which we aim to capture — to the proximity in the (physical) loci of genes, but this is eventually worthwhile noticing and exploiting with respect to the already known coexpression of neighboring genes in yeast (Santoni et al., 2013).

Comparison with CPC Last, we have compared our results to those of CPC. Results are shown in Figure 9 for the manifold (compare to Figure 7 for our technique), and in Figure 9 for the distribution of strong arcs found (strong in the case of CPC means $p \geq 0.05$). The graph found is much closer to a complete graph, which is a quite irrealistic observation since cell cycles are extremely inbalanced in terms of importance with respect to regulation. Furthermore, as remarked below in a more quantitative way, it is known that *Saccharomyces cerevisiae* tends to have a predominant gap phase G1 compared to G2, which is clearly less visible from the CPC results compared to LATNET's results.

III.4. Analysis of the complete yeast genome

We have analyzed the complete set of 6178 genes (representing now $100,000+$ data points) in the yeast genome data (Spellman et al., 1998). Not that this time, this represents a maximum of more than 38 millions arcs in total in the network.

Table A6 presents the breakdown in percentages between YCC genes and non-YCC genes. Clearly, the graph is heavily non symmetric: while roughly 11% of strong arcs come from outside of the YCC to inside the YCC, more than 25% of

[4] http://www.yeastgenome.org/locus/YOR315W/overview

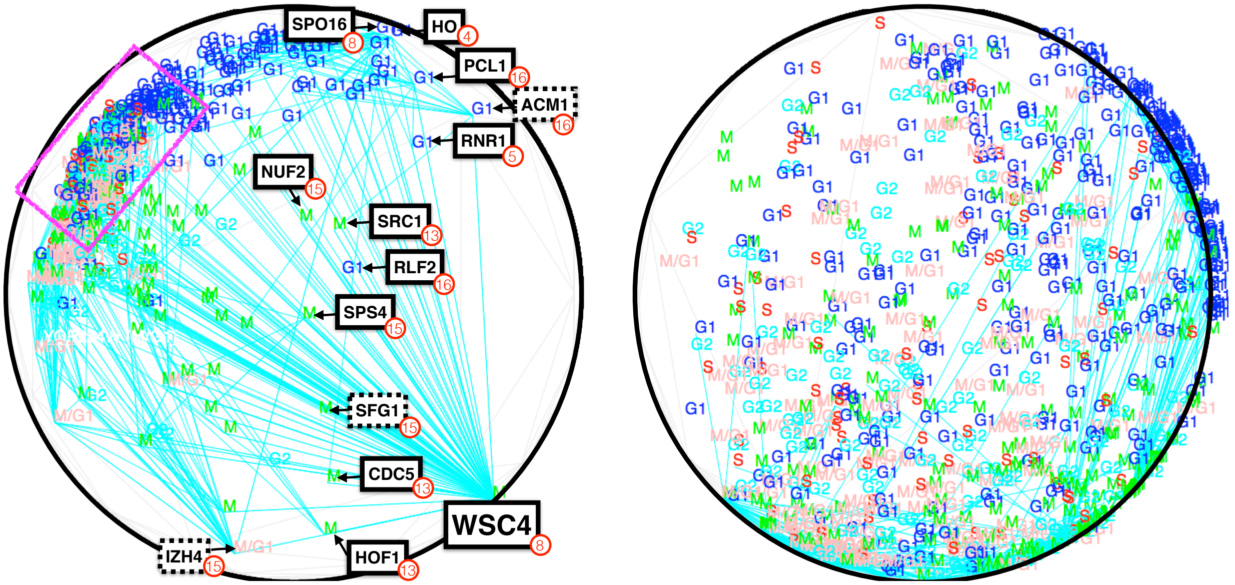

Figure 7. Manifold coordinates learned from the symmetrized p-values graph using Meila & Shi (2001), on Klein disk (we chose it to ease reading: the representation is conformal and geodesics are straight lines — WSC4, which is in fact far away from all other genes, does not prevent a visually meaningful display of the other main genes). Segments are *strong* arcs (arrowheads not represented). *Left*: Major genes influencing the computation and known to be Cell Cycle Transcriptionally Regulated (CCTR Rowicka et al., 2007) are displayed in plain boxes. Chromosomes are shown in red. *Right*: zoom over the pink area in the left plot, showing few strong edges belong to this area, and therefore strong edges guide the construction of the manifold's main coordinates.

these strong arcs come from inside the YCC to outside the YCC. The largest percentage of arcs between YCC - nonYCC is obtained from G1 onto the non YCC genes ($> 10\%$), which seems to be plausible, since G1 is a gap phase involving a lot of interactions with the environment, testing for nutrient supply and growth availability. Interestingly, the strong arcs are recalibrated to take into account the complete set of genes (strong arcs are defined with respect to quantiles in data), yet the relative proportions in the YCC still denote the predominance of M and G1, and the very small percentage of strong arcs for phases S and G2. We notice that the predominance of phase G1 compared to G2 is in perfect accordance with the fact that Spellman et al. (1998) picked the yeast *Saccharomyces cerevisiae* which is indeed known to possess long G1 phases (compared to *e.g. Saccharomyces pombe*).

Finally, Figure 11 displays the manifold obtained for the complete genome. We represent only a corner of the manifold of 6K+ genes, which displays this time the importance of other YCC genes, including in particular YPR204W. This comes at no surprise: this gene codes for a DNA helicase, a motor protein tht separates DNA strands. DNA helicases are involved in a number of processes and not just the YCC. We do not show strong arcs in the picture, but it is worthwhile remarking that the relative predominance of the most prevalent YCC genes is still here, in the whole genome analysis: WSC4, SPO16 and SLD2 are in the top-5 of out-degree measures with strong arcs.

III.5. Sydney property prices data

Performance measures. Firstly, with regards to spatial coherence, it is reasonable to assume that the underlying network is spatially localized, as property prices in nearby suburbs are likely be related. Therefore, our first performance measure is the air distance between the suburbs that are discovered to be connected (shorter distances are better). Secondly, concerning network stability, if we apply a method to different time windows we expect to see some overlaps between the discovered networks. Therefore, our second measure of performance is the proportion (r) of networks in which a connection was present (for each discovered connection). For more stable connections we expect r to be higher. To compute the above measures, we kept the analysis window to five years ($T = 20$ since data is quarterly) and starting from 1995–1999 the window is shifted by one year each time until 2010-2014. Connections between the nodes were obtained for each window, and the discriminate thresholds were chosen so that each method finds 16-18 connections on average.

We set the discrimination threshold for each method so that on average each method finds 17-19 edges in the network

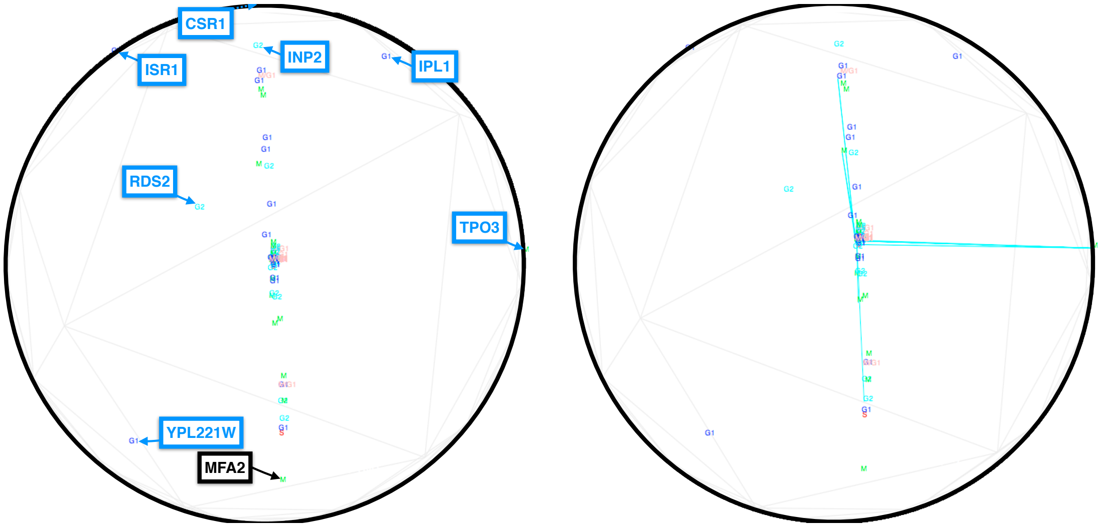

Figure 8. Left: manifold learned from CPC, using the same convention as for LATNET. Remark that none of the *edges* learned by CPC appears, because they are all concentrated inside several blobs that belong to the visible vertical line in the center. Genes displayed in blue are those extremely localed genes that do *not* belong to the genes in Rowicka et al. (2007, Table 4 SI). Right: we have substituted the edges learned by CPC by ours, showing that the most important genes in fact belong to the central blob of the picture, therefore not discriminative of the YCC genes.

($p_{ij} = 0.597$ for LATNET; p-value=0.015 for PC; p-value=0.012 for CPC; p-value=10^{-7} for IAMB; partial correlation= 0.5 for PW-LINGAM).

Figure 12 shows the top-6 of these arcs inferred by LATNET. They clearly indicate that one area of Sydney, Woollahra, acts like an authority in the network, since it receives lots of arcs from other major areas (Hunters Hill, Manly, Mosman, Pittwater). These areas all share common features: they are in central-north Sydney, all have coastal areas, and they happen to be well-known prestigious areas with the highest median property price in Sydney (Campion, 2011), so the observed percolation is no surprise.

Figures 13,14,15,17,16 show the results of LATNET, CPC, PC PW-LINGAM and IAMB algorithms on Sydney property price data. Suburbs were ranked geographically according to their latitude and longitude coordinates, and their locations in the graphs are assigned according to their ranks. We used the ranks instead of actual coordinates of the suburbs in order to be able to better visualize connections in the inner ring. Each panel in the graphs shows the results for a certain period of time, indicated by the label above the panel. Suburbs in Sydney are divided into four groups according to their locations: inner ring (red points), middle ring (green points), outer ring (blue points), and Greater Metropolitan Region (GMR; yellow points).

Data is downloaded from:
http://www.housing.nsw.gov.au/about-us/reports-plans-and-papers/rent-and-sales-reports/back-issues/issue-111

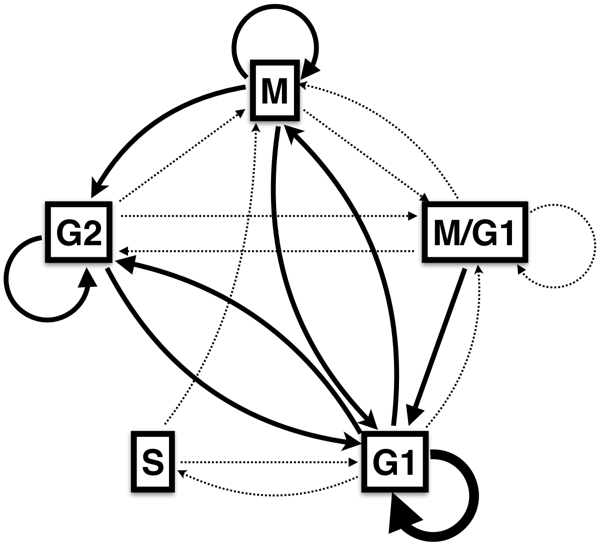

Figure 9. Distribution of strong arcs ($p \geq 0.05$) found by CPC, following Figure 6. Remark that the figure fails to carry the importance of phases M and G1, as Figure 6 for LATNET — in particular, phase M roughly carries the same weights distribution as phase G2, which does not conform to observations (G2 is not even mandatory for the YCC while M obviously is).

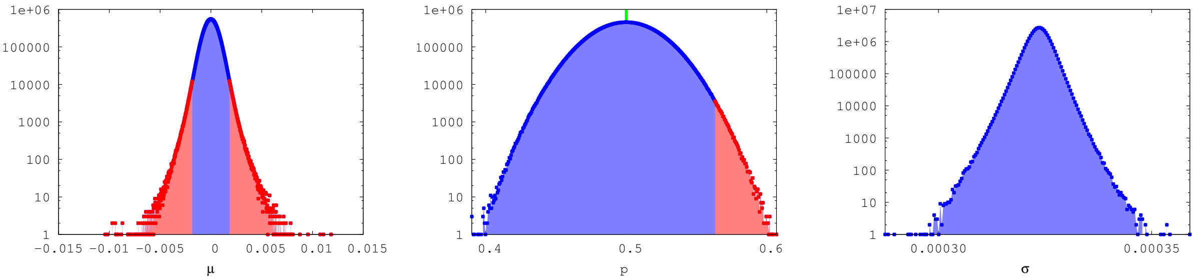

Figure 10. Counting histograms (y, blue + red) for the values of μ (left), p (center) and σ (right, y-scales are log-scales). Conventions follow Figure 5. A tiny fraction (< 1‰) of arcs found have p or σ close to zero; they are not shown to save readability.

	Gene	Phase	out/in-degree ratio
in-degree $<$ out-degree	ASH1	M/G1	5.0
	YIL158W	M	3.0
	MSH6	G1	3.0
	SWI5	M	3.0
	RAD53	G1	3.0
	YNR009W	S	2.5
	MET3	G2	2.3333333333333335
	YOX1	G1	2.0
	SVS1	G1	2.0
	YOL007C	G1	2.0
	CDC20	M	2.0
	YKR041W	G2	2.0
	CDC5	M	2.0
	MET28	S	2.0
	YML034W	M	2.0
	SMC3	G1	2.0
	HHO1	S	2.0
	YDL039C	M	2.0
	RAD27	G1	2.0
	FAR1	M	2.0
	DIP5	M	1.75
	YPL267W	G1	1.6
	CDC45	G1	1.5
	RNR1	G1	1.5
	PCL9	M/G1	1.5
	LEE1	S	1.5
	YOR314W	M	1.5
	YIL025C	G1	1.4444444444444444
	AGP1	G2	1.3333333333333333
	CWP1	G2	1.3333333333333333
	ALD6	M	1.2
	YOL132W	M	1.1666666666666667
	YNR067C	M/G1	1.0666666666666667
in-degree $=$ out-degree	YNL078W	M/G1	1.0
	YCL013W	G2	1.0
	RME1	G1	1.0
	CLB1	M	1.0
	RPI1	M	1.0
	YIL141W	G1	1.0
	BUD4	M	1.0
	YLR235C	G1	1.0
	YOR315W	M	1.0
	YER124C	G1	1.0
	YPR156C	M	1.0
	YGL028C	G1	1.0
	BUD3	G2	1.0
	STE3	M/G1	1.0
	HST3	M	1.0
	ALK1	M	1.0
	CHS2	M	1.0
	YLL061W	S	1.0
	YFR027W	G1	1.0
	LAP4	G1	1.0
	YNL173C	M/G1	1.0
	YML033W	M	1.0
	SEO1	S	1.0
	YOR264W	M/G1	1.0
	NUF2	M	1.0
	YOR263C	M/G1	1.0
	YBR070C	G1	1.0
	YNL300W	G1	1.0
	YPR045C	M	1.0
	YOR248W	G1	1.0
	MYO1	M	1.0
	RLF2	G1	1.0
in-degree $>$ out-degree	YOL101C	M/G1	0.9444444444444444
	YDR355C	S	0.8
	HTB2	S	0.75
	YRO2	M	0.7142857142857143
	YDR380W	M	0.6666666666666666
	FET3	M	0.6666666666666666
	YDL163W	G1	0.6666666666666666
	CLB6	G1	0.6666666666666666
	ECM23	G2	0.6666666666666666
	YBR089W	G1	0.6666666666666666
	YGR221C	G1	0.6666666666666666
	YDL037C	M	0.6
	MF(ALPHA)2	G1	0.6
	YLR183C	G1	0.5714285714285714
	PDR12	M	0.5555555555555556
	YER150W	M/G1	0.5555555555555556
	POL1	G1	0.5
	YHR143W	G1	0.5
	SPS4	M	0.5
	PCL1	G1	0.5
	YGL184C	S	0.5
	EGT2	M/G1	0.5
	CTS1	G1	0.5
	YDR149C	G2	0.5
	GAP1	G2	0.5
	HO	G1	0.5
	WSC4	M	0.4673913043478261
	SPO16	G1	0.46153846153846156
	YMR032W	M	0.4
	YGP1	M/G1	0.4
	SPH1	G1	0.3333333333333333
	YCL022C	G1	0.3333333333333333
	YCLX09W	G2	0.3333333333333333
	PIR1	M/G1	0.25
	ARO9	M	0.16666666666666666

Table A3. Imbalancedness of the network: genes in decreasing ratio out-degree/in-degree, computed using strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile). Only those with >0 out-degree and in-degree are shown. A star (*) indicates reported targets for cell-cycle activators (Simon et al., 2001).

Gene	Phase	out-degree
WSC4	M	43
YOL101C	M/G1	17
YNR067C	M/G1	16
YIL025C	G1	13
YDL037C	M	12
YOR264W	M/G1	10
YER124C	G1	10
HO	G1	9
YPL267W	G1	8
YLR183C	G1	8
MET3	G2	7
DIP5	M	7
SEO1	S	7
YOL132W	M	7
ALD6	M	6
YGL028C	G1	6
PCL9	M/G1	6
SPO16	G1	6
YDL039C	M	6
YOL007C	G1	6
YER150W	M/G1	5
YRO2	M	5
PDR12	M	5
PCL1	G1	5
YNR009W	S	5
RME1	G1	5
ASH1	M/G1	5
AGP1	G2	4
GAP1	G2	4
YOR263C	M/G1	4
YNL173C	M/G1	4
CWP1	G2	4
FAR1	M	4
MCD1	G1	4
YOX1	G1	4
YDR355C	S	4
YOR314W	M	3
MSH6	G1	3
SPT21	G1	3
LEE1	S	3
YIL158W	M	3
YLR049C	G1	3
RNR1	G1	3
HTB2	S	3
GLK1	M/G1	3
SWI5	M	3
MF(ALPHA)2	G1	3
CDC45	G1	3
RAD53	G1	3

Table A4. Genes in decreasing out-degree for strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile). Only those with out-degree ≥ 3 are shown.

Gene	Phase	in-degree
WSC4	M	92
YDL037C	M	20
HO	G1	18
YOL101C	M/G1	18
YNR067C	M/G1	15
YLR183C	G1	14
SPO16	G1	13
YOR264W	M/G1	10
YER124C	G1	10
PCL1	G1	10
YIL025C	G1	9
PDR12	M	9
YER150W	M/G1	9
GAP1	G2	8
SEO1	S	7
YRO2	M	7
ARO9	M	6
SPH1	G1	6
YOL132W	M	6
YGL028C	G1	6
MF(ALPHA)2	G1	5
YMR032W	M	5
ALD6	M	5
YPL267W	G1	5
YGP1	M/G1	5
YDR355C	S	5
RME1	G1	5
YOR263C	M/G1	4
YGL184C	S	4
PCL9	M/G1	4
PIR1	M/G1	4
HTB2	S	4
DIP5	M	4
YNL173C	M/G1	4
SPS4	M	4
YCL022C	G1	3
YOL007C	G1	3
MET3	G2	3
YGR221C	G1	3
YDL039C	M	3
CWP1	G2	3
YCLX09W	G2	3
YDR380W	M	3
AGP1	G2	3
CLB6	G1	3
YBR089W	G1	3
FET3	M	3
YDL163W	G1	3
ECM23	G2	3

Table A5. Genes in decreasing in-degree for strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile). Only those with in-degree ≥ 3 are shown.

	M	M/G1	G1	S	G2	N
M	0.6	ϵ	1.2	ϵ	ϵ	2.1
M/G1	ϵ	ϵ	0.2	ϵ	ϵ	1.4
G1	1.0	ϵ	1.0	0.2	ϵ	4.5
S	0.2	ϵ	0.2	ϵ	0.2	2.0
G2	0.2	ϵ	0.2	ϵ	ϵ	1.0
N	6.6	3.5	10.3	2.3	2.9	58.5

Table A6. Distribution of strong arcs (p in top 99.9% quantile, $|\mu|$ in top 99% quantile) for the *complete genome* of the yeast, including the breakdown for the YCC phases (see *e.g.* Table A2). "ϵ" means $< 0.1\%$ and "N" stands for "None" (Gene not in the sentinels of the YCC").

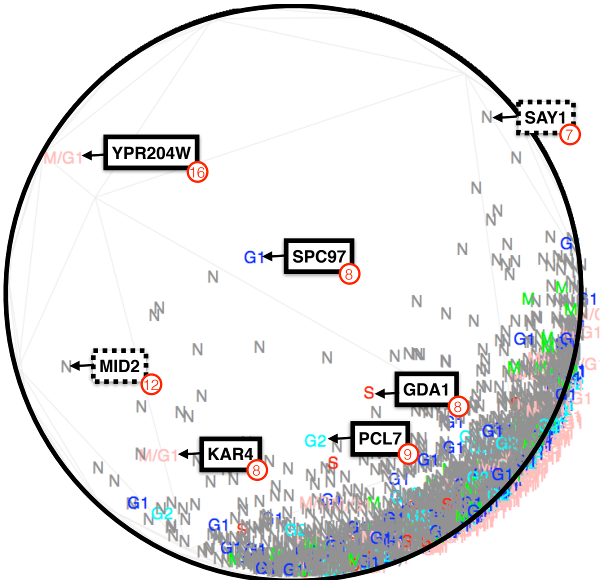

Figure 11. manifold obtained for LATNET in the whole yeast genome, conventions follow Figure 7 (strong arcs not displayed for readability).

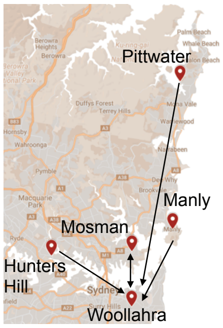

Figure 12. Arcs with highest r values discovered by LATNET.

.

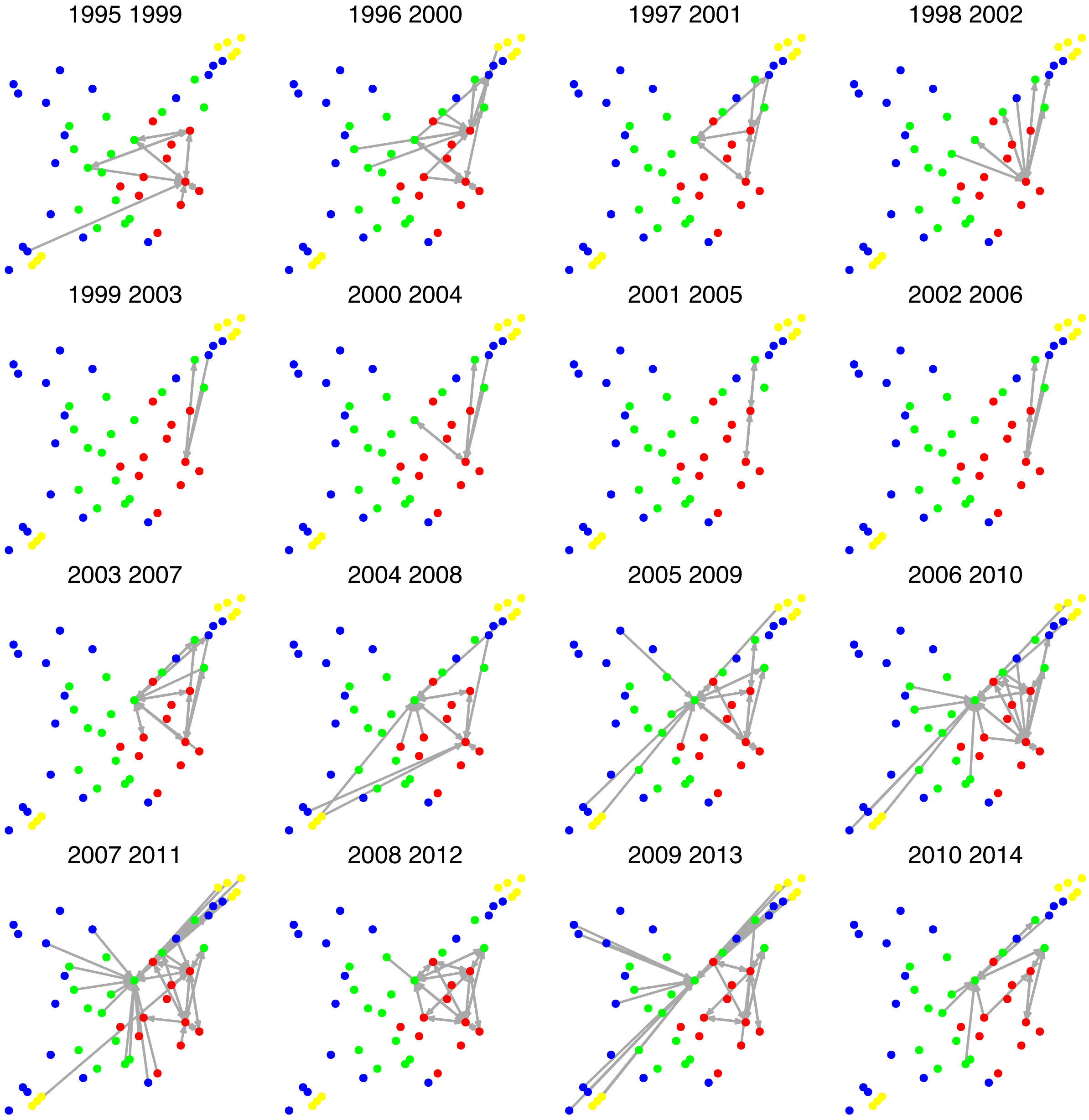

Figure 13. Associations between median house prices in different suburbs discovered by LATNET
.

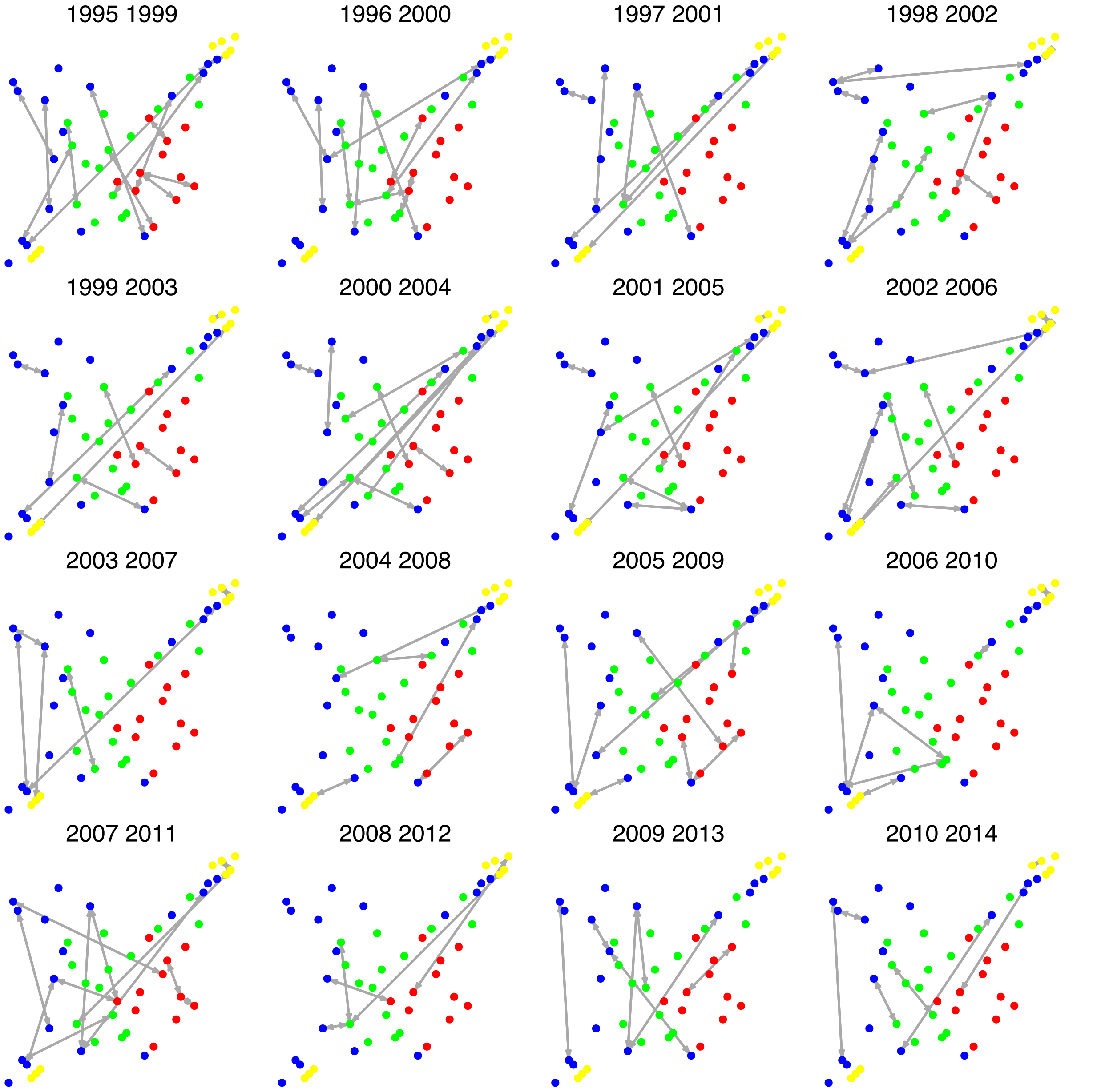

Figure 14. Associations between median house prices in different suburbs discovered by CPC

.

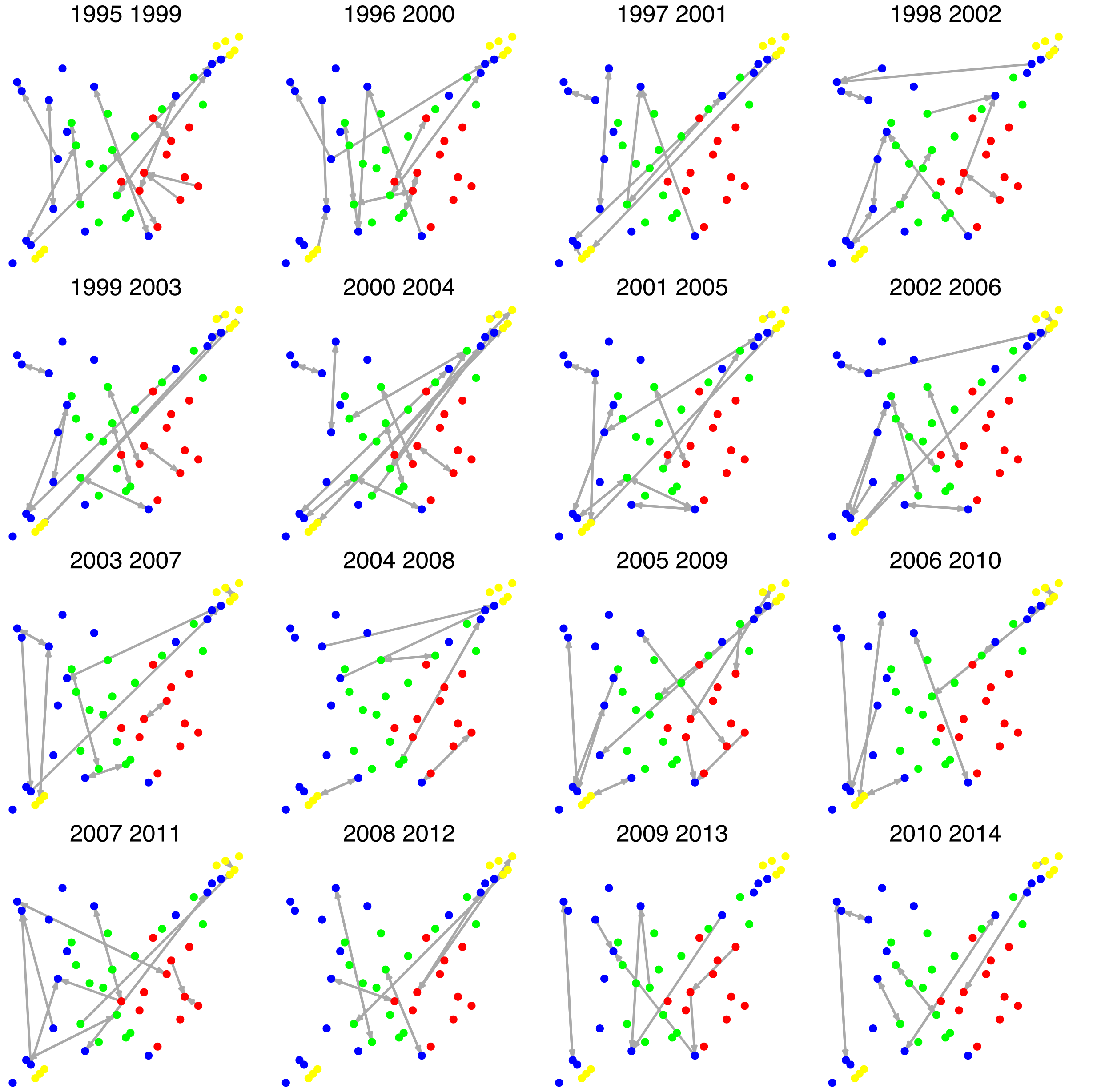

Figure 15. Associations between median house prices in different suburbs discovered by PC

.

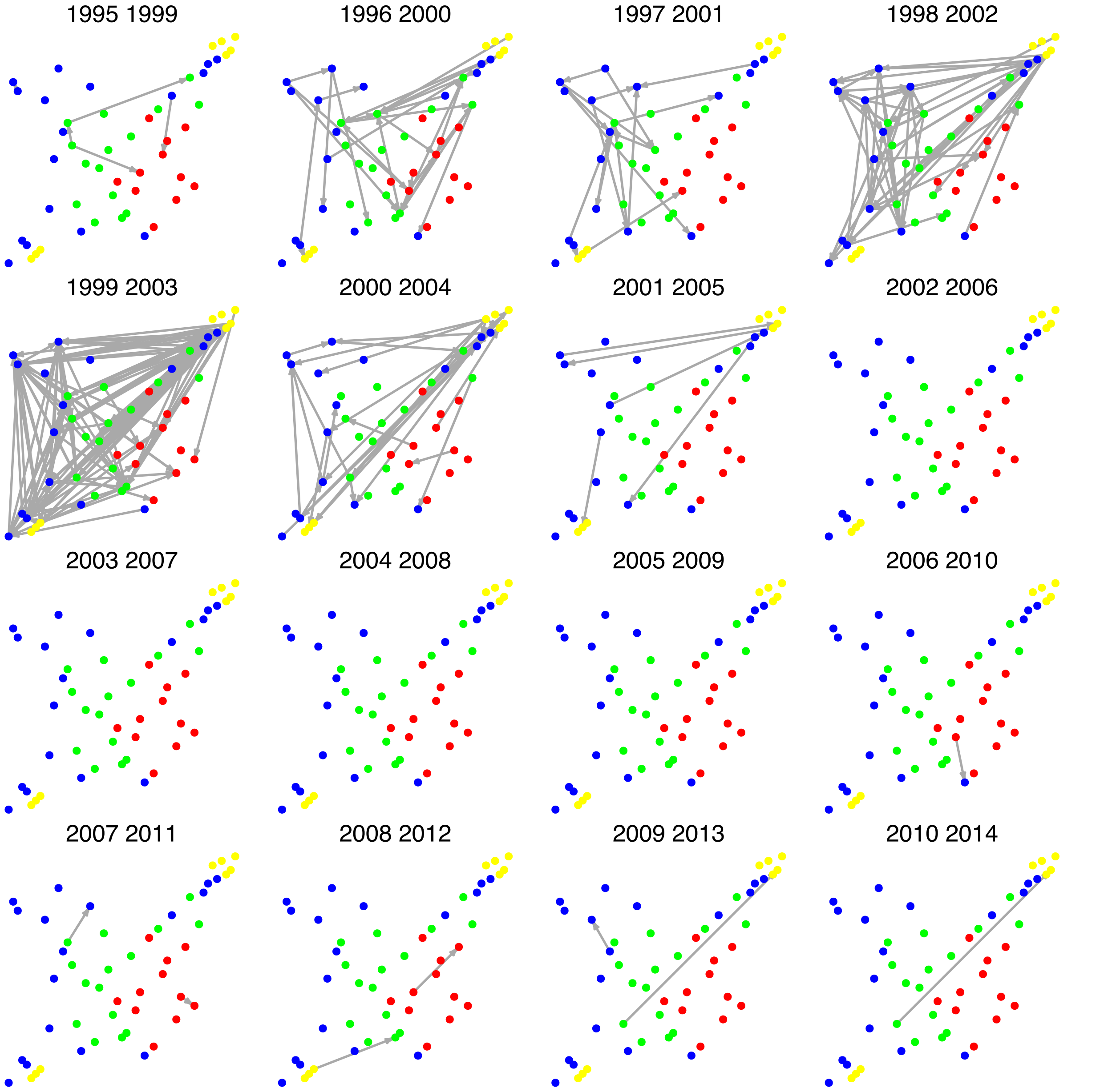

Figure 16. Associations between median house prices in different suburbs discovered by PW-LINGAM

.

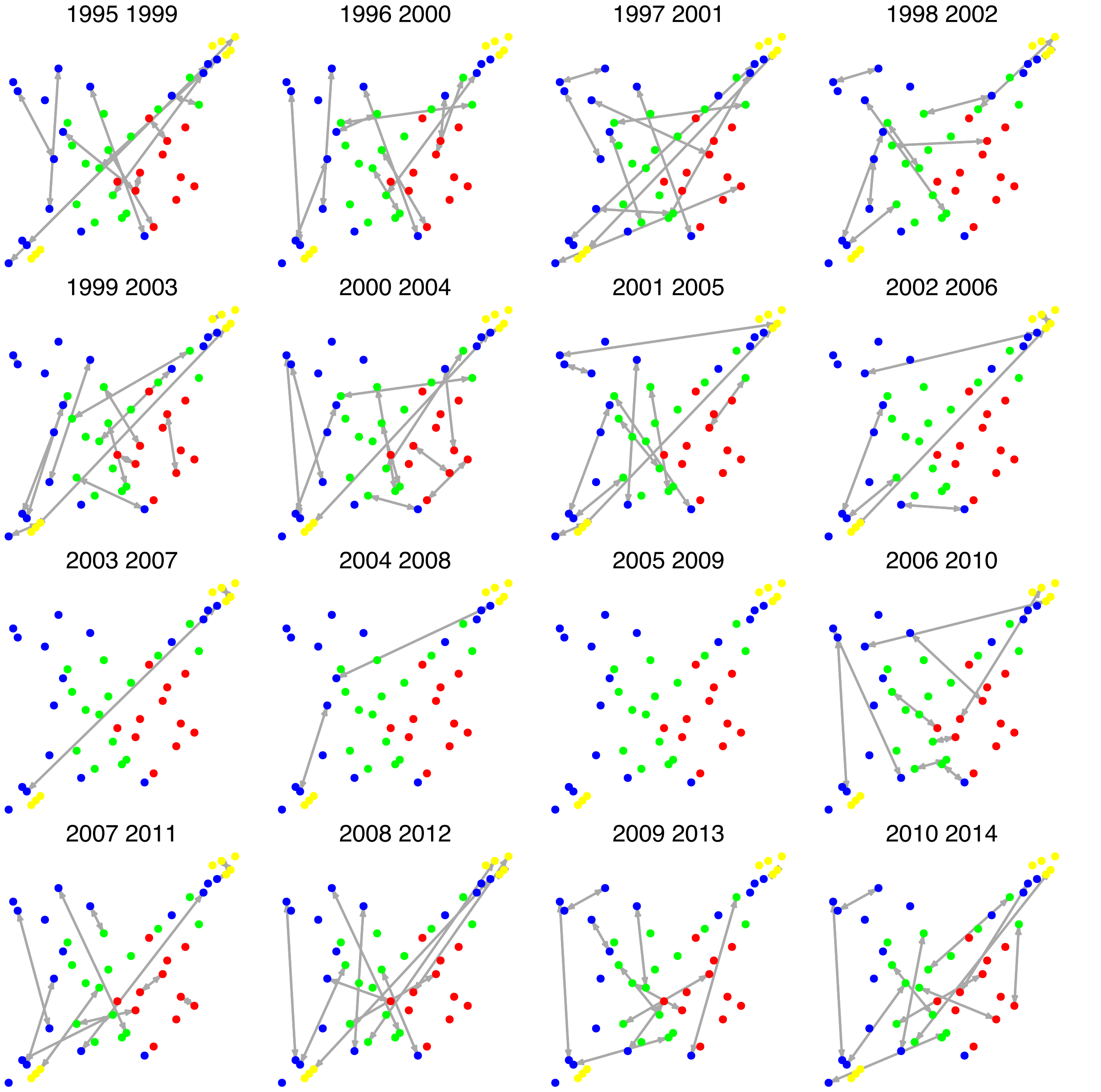

Figure 17. Associations between median house prices in different suburbs discovered by IAMB

.

References

Abramowitz, M. and Stegun, I.-A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U. S. Government Printing Office, 1964.

Bonilla, E.-V., Chai, K.-M. A, and Williams, C.-K.-I. Multi-task Gaussian process prediction. In *NIPS*, 2008.

Campion, Vikki. http://www.dailytelegraph.com.au/archive/money/property-prices-in-sydneys-traditional-blue-collar-suburbs-are-booming-with-cabramatta-best-performing-residex-reports/news-story/6971cf79862fdf4f2b3fea3cf2917b5f, 2011.

Chickering, D.-M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554, 2002.

Cho, R.-J., Campbell, M.-J., Winzeler, E.-A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.-G., Gabrielian, A.-E., Landsman, D., Lockhart, D.-J., and Davis, W. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

Hyvärinen, A. and Smith, S.-M. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *JMLR*, 14:111–152, 2013.

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

Kingma, D.-P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014.

Linderman, S.-W. and Adams, R.-P. Discovering Latent Network Structure in Point Process Data. In *ICML*, 2014.

Maddison, C.-J., Mnih, A., and Teh, Y.-W. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv:1611.00712*, 2016.

Marco, S. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3), 2010.

Meek, C. *Graphical Models: Selecting causal and statistical models*. PhD thesis, PhD thesis, Carnegie Mellon University, 1997.

Meila, M. and Shi, J. Learning segmentation by random walks. In *NIPS*, volume 14, 2001.

Peters, J., Mooij, J.-M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *JMLR*, 15 (1):2009–2053, 2014.

Rakitsch, B., Lippert, C., Borgwardt, K., and Stegle, O. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *NIPS*, 2013.

Rowicka, M., Kudlicki, A., Tu, B.-P., and Otwinowski, Z. High-resolution timing of cell cycle-regulated gene expression. *PNAS*, 104(43):16892–16897, 2007.

Santoni, D., Castiglione, F., and Paci, P. Identifying correlations between chromosomal proximity of genes and distance of their products in protein-protein interaction networks of yeast. *PLoS ONE*, 8, 2013.

Simon, I., Barnett, J., Hannett, N., Harbison, C.-T., Rinaldi, N.-J., Volkert, T.-L., Wyrick, J.-J., Zeitlinger, J., Gifford, D.-K., Jaakkola, T.-S., and Young, R.-A. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

Spellman, P.-T., Sherlock, G., Zhang, M.-Q., Iyer, V.-R., Anders, K., Eisen, M.-B., Brown, P.-O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

Tao, T. Singularity and determinant of random matrices, 2008. Lewis Memorial Lecture.

Tsagris, M., Beneki, C., and Hassani, H. On the folded normal distribution. *Mathematics*, 2:12–28, 2014.

Vu, V.-H. *Modern Aspects of Random Matrix Theory*. Proceedings of Symposia in Applied Mathematics. American Mathematical Society, 2014.