PAPER • OPEN ACCESS

The planted *k*-factor problem

To cite this article: Gabriele Sicuro and Lenka Zdeborová 2021 *J. Phys. A: Math. Theor.* **54** 175002

View the article online for updates and enhancements.

The planted *k*-factor problem

Gabriele Sicuro[1,3,*] and Lenka Zdeborová[2]

[1] IdePHICS Laboratory, EPFL, 1015 Lausanne, Switzerland
[2] SPOC Laboratory, EPFL, 1015 Lausanne, Switzerland
[3] Department of Mathematics, King's College London, London, WC2R 2LS, United Kingdom

E-mail: gabriele.sicuro@kcl.ac.uk

CrossMark

Abstract

We consider the problem of recovering an unknown *k*-factor, hidden in a weighted random graph. For $k = 1$ this is the planted matching problem, while the $k = 2$ case is closely related to the planted traveling salesman problem. The inference problem is solved by exploiting the information arising from the use of two different distributions for the weights on the edges inside and outside the planted sub-graph. We argue that, in the large size limit, a phase transition can appear between a full and a partial recovery phase as function of the signal-to-noise ratio. We give a criterion for the location of the transition.

Keywords: inference, graph theory, cavity method, disordered systems, belief propagation, planting

(Some figures may appear in colour only in the online journal)

1. Introduction

The problem of recovering a hidden structure (the signal) in a graph on the basis of the observation of its edges and the weights on edges and vertices appears in many, diverse contexts. Community detection [1], group testing [2], certain types of error correcting codes [3], particle tracking [4] are some example of statistical inference problems formulated on graphs in which some underlying pattern has to be identified. The feasibility of the hidden structure recovery depends, of course, on the amount of 'noise' in the problem. It turns out that, in the limit of large system sizes, sharp recovery thresholds with respect to the signal-to-noise ratio can appear. These sharp thresholds separate no recovery phases (in which no information on the

signal is accessible), partial recovery phases (in which an output correlated with the signal can be obtained) and full recovery phases (in which the signal is recovered with a vanishing error).

These algorithmic transitions, analogous to phase transitions in physics, can be of different orders (depending on the number of derivatives of the order parameter that exist). The order of the transition seems to have important algorithmic implications. First order transitions, in particular, have been related to the presence of a computationally hard phases [1]. In the stochastic block model on sparse graphs, for example, a phase transition between a no-recovery phase and a partial recovery phase is found [1]. In [5, 6] it has been shown that a partial to full recovery transition can also appear in the same model if denser topologies are considered. A partial to full recovery first order transition appears in low-density-parity-check error correcting codes [3], where the target is to correctly recover a code-word. Somewhat uncommonly, a partial to full recovery infinite order phase transition has been recently found in the planted matching problem, in which a hidden matching has to be detected in a weighted random graph [7–9]. This is a new type of a phase transition found in inference problems that has been put in relation with the presence of instabilities of the belief propagation (BP) fixed points. It is thus of interest to investigate whether it appears in other problems than the planted matching.

In this paper we study a generalization of the planted matching problem—the so-called planted k-factor problem. A k-factor of a graph $\mathcal{G}$ is a spanning subgraph with fixed degree k, i.e., a spanning k-regular graph. In this problem, the (weighted) k-regular graph is hidden (planted) by adding new weighted edges to it. The weights of the planted and non-planted edges are random quantities with different distributions. The goal is to recover the planted k-factor knowing k and the two weight distributions.

In analogy with percolation being continuous while the appearance of a k-core is a discontinuous (first order) phase transition [10], we may have anticipated that the transition in the k-factor problem will be of a different type than in the planted matching. We will show instead that the planted k-factor problem manifests a partial-full recovery continuous phase transition akin the one in the planted matching, and we will give a criterion for the threshold between the two phases.

The planted k-factor problem is related to many interesting inference problems. For example, it shares some similarity with the problem of structure-detection in small-world graphs [11]. In the latter case, a k-regular ring lattice is hidden in a (unweighted) random graph by a rewiring procedure. Full recovery is possible depending on the parameters of the rewiring process.

The planted one-factor problem corresponds to the aforementioned planted matching problem, introduced in [4] as a model for particle tracking. In this model, a weighted perfect matching, hidden in a graph, has to be recovered. In reference [4] the problem was studied numerically for a particular case of the distribution of weights. The results suggested the existence of a phase transition between a full recovery phase and a partial recovery phase. More recently, Moharrami and coworkers [7] rigorously analyzed another special setting of the problem assuming that $\mathcal{G}$ is the complete graph and the weights are exponentially distributed, and proved the existence of a partial/full recovery phase transition. The results of reference [7] have been generalized to sparse graphs and general weight distributions in reference [8], via heuristic arguments based on a correspondence between the problem and branching random walk processes. A proof of the results in [8] has been recently given in [9].

The planted two-factor problem is a relaxation of the planted traveling salesman problem [12]. In this problem, a unique, hidden Hamiltonian cycle visiting (exactly once) all vertices of a graph has to be recovered. In the planted two-factor problem, instead, the planted subgraph is more generally given by a set of cycles. Solving the two-factor problem can be, however, informative on a hidden Hamiltonian cycle. In reference [12] the Hamiltonian cycle recovery

problem has been studied on the complete graph. Therein, a sufficient condition for the full recovery of the planted Hamiltonian cycle has been derived. Moreover, it is proved therein that, within the full recovery phase, the solution obtained searching for a two-factor coincides (with high probability and in the large size limit) with the hidden Hamiltonian cycle.

In this paper, we generalize the available results for the planted one-factor problem and planted two-factor problem to the general planted k-factor problem with arbitrary distributions for the edge weights and including sparse graphs in the analysis. Unlike reference [11], we will assume no prior knowledge on the structure of the planted k-factor (except for the degree of its vertices). Our approach, based on the standard cavity method and the corresponding BP equations [13], allows us to obtain, at the level of rigor of theoretical physics, a criterion for the full recovery of the planted subgraph in the large size limit of the problem. The threshold criterion is derived studying the recursive distributional equations (RDE) corresponding to the cavity equations. It turns out that a 'drift' in their solutions can appear under iteration. If this is the case, the full-recovery solution is the only stable one, and full recovery is possible. The study of such drift has been tackled in analogy with the analysis in reference [8] for the $k = 1$ case, and with the phenomenon of front propagation for reaction-diffusion equations [14–18]. We give an explicit criterion for the threshold between a partial recovery phase and a full recovery phase of the planted k-factor. Our results recover, as special cases, the ones obtained in references [7–9] for the planted one-factor. In the limit of dense graphs, they provide a sharper characterization of the phase transition for $k = 2$ with respect to the analysis in [12], as discussed in more detail in section 6.1.

The rest of the paper is organized as follows. In section 2 we define the problem under study and introduce two adopted statistical estimators, namely the block maximum *a posteriori* (MAP) and the symbol MAP. In section 3 we present the BP equations for the solution of the problem and their corresponding probabilistic description. In section 4 we numerically study a specific case, observing that a transition between a full recovery and a partial recovery phase can appear as function of the parameters of the problem. In section 5 we give a heuristic derivation of the criterion for the location of the transition for arbitrary weight distributions for the block MAP estimator, equation (52), which is the main result of the paper. In section 6 we compare our theoretical predictions with the numerical results obtained for different variants of the problem, including the case considered in section 4 and the Hamiltonian cycle recovery problem considered in reference [12]. Finally, conclusions and perspectives for future work are given in section 7.

2. Definitions

2.1. Planting a k-factor

Let us assume that an integer $k \in \mathbb{N}$ and two probability densities p and $\hat{p}$ on the real line are given. We will focus on an ensemble of weighted simple graphs denoted $\mathscr{G}_0 = (\mathscr{V}_0, \mathscr{E}_0, \underline{w})$, containing by construction a planted k-factor to be recovered. Here $\mathscr{V}_0 = \{1, \ldots, N\}$ is the set of $N \in \mathbb{N}$ vertices such that kN is even, and $\mathscr{E}_0$ is the set of edges (unordered pairs of distinct vertices of $\mathscr{V}_0$). A weight $w_e \in \mathbb{R}$ is associated to each edge e of the graph, so that $\underline{w} = \{w_e : e \in \mathscr{E}_0\}$ is the set of such weights. We introduce a probability measure over the set of weighted graphs by means of the following generation steps of $\mathscr{G}_0$ (see figure 1(a)).

(a) One first constructs a k-regular graph having vertex set $\mathscr{V}_0$. The graph is chosen uniformly among all possible k-regular graph with N vertices [19]. This can be achieved using fast algorithms available in the literature [20]. The obtained graph has $\frac{1}{2}kN$ edges and
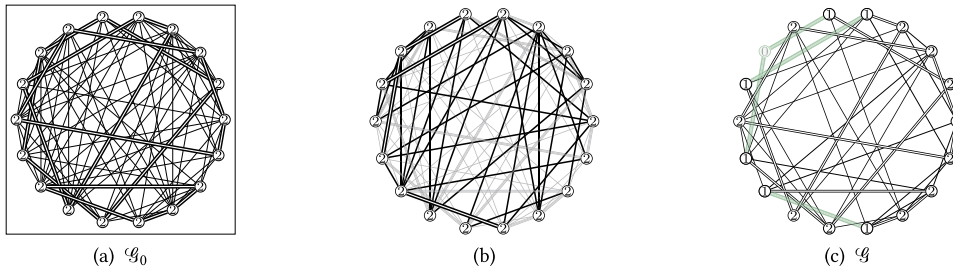
Figure 1. Pictorial representation of the preliminary pruning in a planted two-factor problem. (a) A two-regular graph on a vertex set of N vertices is generated (thick double lines; in the picture, $N = 18$), with edge weights distributed as $\hat{p}$. Random edges are added with some probability c/N (thin lines), with edge weights distributed as p. The obtained graph is $\mathscr{G}_0$.s A capacity k is assigned to each vertex. (b) Edges in $\mathrm{supp}(\hat{p})\backslash\Gamma$ (double lines) can be identified as planted and removed, decreasing the capacity of the endpoints by 1. Edges in $\mathrm{supp}(p)\backslash\Gamma$ (thick single lines) can be identified and removed. Vertices with zero capacity can be removed. (c) We call $\mathscr{G}$ the obtained pruned graph, and we call $\mathscr{F}_k^{(0)}$ the set of identified edges by means of this first pruning process (green).

edge-set $\mathscr{F}_k^*$. A random weight w_e, generated independently of all the others with density $\hat{p}$, is associated to each edge $e \in \mathscr{F}_k^*$.

(b) Given a pair of vertices that are not neighbors in $\mathscr{F}_k^*$, a link is added between them with probability cN^{-1}. Let $\mathcal{E}_0$ be the final set of edges of the obtained graph. A weight w_e, independently generated from all the others with distribution p, is assigned to each $e \in \mathcal{E}_0\backslash\mathscr{F}_k^*$.

We shall call planted (resp. non-planted) edges those in $\mathscr{F}_k^*$ (resp. in $\mathcal{E}_0\backslash\mathscr{F}_k^*$). The parameters of this ensemble of weighted random graphs are thus the integers N and k, the parameter c controlling the density of non-planted edges, and the two distributions $\hat{p}$ and p for the generation of the weights of the planted and non-planted edges respectively. Given $\mathscr{F}_k^*$, the probability to generate a graph $\mathscr{G}_0$ is therefore

$$\mathbb{P}(\mathscr{G}_0|\mathscr{F}_k^*) = \mathbb{I}(\mathscr{F}_k^* \subseteq \mathcal{E}_0) \prod_{e\in\mathscr{F}_k^*} \hat{p}(w_e) \prod_{e\in\mathcal{E}_0\backslash\mathscr{F}_k^*} p(w_e) \left(\frac{c}{N}\right)^{|\mathcal{E}_0|-k\frac{N}{2}} \left(1 - \frac{c}{N}\right)^{\binom{N}{2}-|\mathcal{E}_0|}, \quad (1)$$

where here and in the following $\mathbb{I}(A)$ denotes the indicator function of the event A. The edges in $\mathcal{E}_0\backslash\mathscr{F}_k^*$ form essentially an Erdós–Rényi random graph of average degree c. The edge-set $\mathscr{F}_k^*$, on the other hand, is a k-factor of $\mathscr{G}_0$ by construction, i.e., a spanning subgraph of $\mathscr{G}_0$ in which all the vertices have the same degree k. The resulting graph has average coordination $c + k$. Note that, for $k = 1$, $\mathscr{F}_1^*$ is a matching on $\mathscr{G}_0$, and the introduced ensemble of graphs coincides with the one studied in reference [8] for the planted matching problem.

2.2. The inference problem

Given a graph $\mathscr{G}_0$ in the ensemble described above, we wonder if it si possible to infer the k-factor $\mathscr{F}_k^*$ hidden in it. We assume that the generation rules are known, alongside with k, c, p and $\hat{p}$. All the exploitable information is contained in the posterior probability $\mathbb{P}(\mathscr{F}_k|\mathscr{G}_0)$ that a certain k-factor $\mathscr{F}_k$ in $\mathscr{G}_0$ is the planted k-factor $\mathscr{F}_k^*$. From Bayes theorem we obtain the following expression for the posterior:

$$\mathbb{P}(\mathscr{F}_k|\mathscr{G}_0) \propto \mathbb{I}(\mathscr{F}_k \text{ is a } k - \text{factor}) \prod_{e\in\mathscr{F}_k} \hat{p}(w_e) \prod_{e\in\mathcal{E}_0\backslash\mathscr{F}_k} p(w_e)\mathbb{I}(\mathscr{F}_k \subseteq \mathcal{E}_0), \quad (2)$$

where the symbol $\propto$ hides a normalization constant independent of $\mathscr{F}_k$. To parametrize the probability measure above, it is convenient to introduce, for each edge e, the binary variable $m_e \in \{0, 1\}$, so that $\underline{m} = \{m_e = \mathbb{I}(e \in \mathscr{F}_k) : e \in \mathcal{E}_0\} \in \{0, 1\}^{|\mathcal{E}_0|}$, and rewrite the posterior as

$$\mathbb{P}(\underline{m}|\mathscr{G}_0) \propto \prod_{e \in \mathcal{E}_0} \left(\frac{\hat{p}(w_e)}{p(w_e)}\right)^{m_e} \prod_{i=1}^N \mathbb{I}\left(\sum_{e \in \partial i} m_e = k\right), \tag{3}$$

where ∂i denotes the set of edges incident to the vertex i. We want to compute an estimator $\hat{\mathscr{F}}_k$ that is 'close' to the hidden k-factor $\mathscr{F}_k$. The estimator is associated to the set of binary variables $\underline{\hat{m}}$ that encodes the set of edges in $\hat{\mathscr{F}}_k$. With a slight abuse of notations, in the following we identify a set of edges $\mathscr{F}_k$ with its corresponding $\underline{m}$. We will denote $\underline{m}^*$ the set of variables associated to $\mathscr{F}_k^*$ and $\underline{\hat{m}}$ the set of variables associated to an estimator $\hat{\mathscr{F}}_k$. In this paper, we will quantify the distance between an estimator $\hat{\mathscr{F}}_k$ and the true planted k-factor $\mathscr{F}_k^*$ in terms of the cardinality of the symmetric difference $\mathscr{F}_k^* \triangle \hat{\mathscr{F}}_k$ between $\mathscr{F}_k^*$ and $\hat{\mathscr{F}}_k$,

$$\varrho(\mathscr{F}_k^*, \hat{\mathscr{F}}_k) := \frac{|\mathscr{F}_k^* \triangle \hat{\mathscr{F}}_k|}{2|\mathscr{F}_k^*|} = \frac{1}{kN}\sum_{e \in \mathcal{E}_0}\mathbb{I}(\hat{m}_e \neq m_e^*), \tag{4}$$

or equivalently the Hamming distance between the binary string $\underline{m}^*$ encoding $\mathscr{F}_k^*$ and the binary string $\underline{\hat{m}}$ encoding $\hat{\mathscr{F}}_k$. We will consider two 'maximal *a posteriori*' (MAP) estimators, each one minimizing a 'measure of distance' with the planted k-factor.

- A first possibility is to choose as estimator the k-factor that minimizes the probability $\mathbb{P}(\mathscr{F}_k^* \neq \hat{\mathscr{F}}_k)$ over all realizations of the problem,

$$\hat{\mathscr{F}}_k^{(b)} = \arg\max_{\underline{m}} \mathbb{P}(\underline{m}|\mathscr{G}_0). \tag{5}$$

  This estimator is usually called 'block MAP' [3].
- A different estimator, called 'symbol MAP' and denoted in the following $\hat{\mathscr{F}}_k^{(s)}$, is obtained requiring that the distance to be minimized is precisely the error in equation (4). In this case, for each edge $e \in \mathcal{E}_0$, we choose

$$\hat{m}_e = \arg\max_{m_e} \mathbb{P}_e(m_e|\mathscr{G}_0), \tag{6}$$

  with $\mathbb{P}_e$ the marginal of the posterior probability (3) for the edge e. Observe, however, that this estimator is not necessarily a k-factor.

In the following, we will discuss both the estimators defined above, in the thermodynamic limit $N \to \infty$, as a function of the parameters of the model. As in the planted matching problem [4, 7, 8], the possibility of identifying the planted edges will depend on the similarity between the distribution p and the distribution $\hat{p}$, and on the parameter c, that corresponds itself to a noise level expressing the number of confusing non-planted edges introduced in the graph.

3. Cavity equations

3.1. Pruning the graph

To efficiently solve the problem, and possibly reduce the size of the input, it is convenient to proceed with a preliminary 'pruning' of the graph. Before proceeding further, let us assign a 'capacity variable' $\kappa_i = k$ to each vertex i. The capacity of each node i will take into account

the number of unidentified planted edges incident at i. Let

$$\mathrm{supp}(\hat{p}) := \{w \in \mathbb{R} : \hat{p}(w) > 0\}, \tag{7a}$$

$$\mathrm{supp}(p) := \{w \in \mathbb{R} : p(w) > 0\}, \tag{7b}$$

be the support of $\hat{p}$ and p respectively, and let

$$\Gamma := \mathrm{supp}(p) \cap \mathrm{supp}(\hat{p}) \tag{7c}$$

be the intersection of these supports. We suppose that $\Gamma \neq \emptyset$ (the inference problem is otherwise trivial).

If an edge e bears a weight $w_e \in \mathrm{supp}(p) \backslash \Gamma$, then it can be immediately identified as 'non-planted', and removed from $\mathscr{G}_0$. This event will happen with probability $1 - \mu$, where

$$\mu := \int_\Gamma p(w)\mathrm{d}w. \tag{7d}$$

On the other hand, the set of edges

$$\mathscr{F}_k^{(0)} := \{e \in \mathcal{E}_0 : w_e \in \mathrm{supp}(\hat{p}) \backslash \Gamma\} \tag{7e}$$

surely belongs to the planted configuration, $\mathscr{F}_k^{(0)} \subseteq \mathscr{F}_k^*$. These edges can be correctly classified as 'planted' with no algorithmic effort, except for the inspection of their weights. A planted edge e can be therefore identified, solely on the basis of the value of its weight, with probability $1 - \hat{\mu}$, where

$$\hat{\mu} := \int_\Gamma \hat{p}(w)\mathrm{d}w. \tag{7f}$$

We can remove from the graph the identified planted edges, see figure 1(b). We must take care, however, of reducing at the same time by 1 the capacity of the endpoints of a planted edge that is removed. At the end of this process, the capacity of a generic vertex i is $0 \leqslant \kappa_i \leqslant k$, see figure 1(c). For large N, the capacity of the vertices after this pruning has binomial distribution $\mathrm{Bin}(k, \hat{\mu})$. In particular, $(1 - \hat{\mu})^k N$ vertices have zero capacity, meaning that all their incident planted edges have been identified. These vertices can also be removed, alongside with all their remaining (non-planted) incident edges.

In the resulting pruned graph, a vertex has K incident planted edges and Z non-planted edges, where K and Z are two random variables having distribution

$$\mathbb{P}[\mathsf{K} = \kappa] = \binom{k}{\kappa} \frac{\hat{\mu}^\kappa (1 - \hat{\mu})^{k-\kappa}}{\hat{\mu}_k}, \quad \kappa = 1, \ldots, k, \tag{8a}$$

$$\mathbb{P}[\mathsf{Z} = z] = \gamma^z \frac{\mathrm{e}^{-\gamma}}{z!}, \quad z \in \mathbb{N}_0, \quad \gamma := c\mu\hat{\mu}_k, \tag{8b}$$

where $\hat{\mu}_k := 1 - (1 - \hat{\mu})^k$, so that the pruned graph has average degree $\mathbb{E}[\mathsf{K} + \mathsf{Z}] = \gamma + k\hat{\mu}\hat{\mu}_k^{-1}$. The distributions of the weights of the surviving edges are obtained from the original ones conditioning the weights to be in Γ, i.e.,

$$P(w) := \frac{p(w)}{\mu} \mathbb{I}(w \in \Gamma), \tag{9a}$$

$$\hat{P}(w) := \frac{\hat{p}(w)}{\hat{\mu}} \mathbb{I}(w \in \Gamma), \tag{9b}$$

for the non-planted and planted edges, respectively. We will denote $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \underline{w})$ the pruned graph, with $\mathcal{V} \subseteq \mathcal{V}_0$ and $\mathcal{E} \subseteq \mathcal{E}_0$ the new vertex and edge-sets. For large N, $\mathcal{V}$ has cardinality $\hat{N} := \hat{\mu}_k N$, each vertex $i \in \mathcal{V}$ having capacity $1 \leqslant \kappa_i \leqslant k$ distributed as K. The graph will have a total of $\frac{1}{2}k\hat{\mu}N$ surviving planted edges and $\frac{1}{2}\gamma\hat{\mu}_k N$ surviving non-planted edges.

### 3.2. The belief propagation equations

To write down a BP algorithm for the planted k-factor problem, we start from the probability distribution over the configurations $\underline{m} = \{m_e : e \in \mathcal{E}\} \in \{0,1\}^{|\mathcal{E}|}$ on the edges of the pruned graph $\mathscr{G}$,

$$\nu(\underline{m}) \propto \exp\left(-\beta \sum_{e \in \mathcal{E}} m_e \omega_e\right) \prod_{i \in \mathcal{V}} \mathbb{I}\left(\sum_{e \in \partial i} m_e = \kappa_i\right), \tag{10}$$

where $\beta > 0$ and $\omega_e = \omega(w_e)$, with

$$\omega(w) := -\ln \frac{\hat{P}(w)}{P(w)}. \tag{11}$$

The quantities ω_e play the role of effective weights on the edges of the graph. The introduction of β is convenient because the measure in equation (10) coincides, for $\beta = 1$, with the posterior defined in equation (3). On the other hand, for $\beta \to \infty$ the measure concentrates on the configurations maximizing the posterior, i.e., on the block MAP. Equation (10) can also be associated to a graphical model. In particular, we can associate a variable vertex m_e ($\bigcirc$) to each edge e of $\mathscr{G}$. We also introduce two types of interaction vertices. A first type of interaction vertex ($\blacksquare$) is associated to each $i \in \mathcal{V}$, and linked to all variable vertices m_e such that $e \in \partial i$. Such vertex imposes that κ_i variables m_e, $e \in \partial i$, are equal to 1. A second interaction vertex expresses the contribution $e^{-\beta m_e \omega_e}$ ($\square$) for each e, and it is linked to the variable vertex m_e. Pictorially,
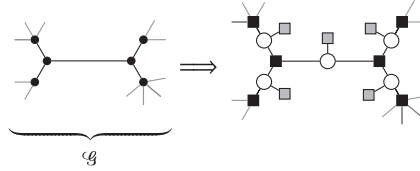

The BP algorithm [13] provides a recipe for the computation of the marginals of ν on such factor graph. The idea is to approximate the marginal $\nu_e(m)$ for the edge $e = (i,j)$ as

$$\nu_e(m) := \sum_{\{m_{\tilde{e}}\}_{\tilde{e} \neq e}} \nu(\underline{m}) \simeq \nu_{i \to e}(m)\nu_{j \to e}(m)e^{-\beta m \omega_e},$$

where $\nu_{i \to e}(m)$ (respectively $\nu_{i \to e}(m)$) mimics a marginal probability in graphical models in which j (respectively i) is absent. Such factorization is exact on infinite trees. The algorithm goal is the computation of such messages, and it is conjectured to be exact in the large size limit for sparse random graphs. The messages obey the following equations (one for each directed edge of the graph),

$$\nu_{i \to e}(m) \propto \sum_{\{m_{\tilde{e}}\}_{\tilde{e} \in \partial i \backslash e}} \mathbb{I}\left(m + \sum_{\tilde{e} \in \partial i \backslash e} m_{\tilde{e}} = \kappa_i\right) \prod_{\substack{\tilde{e}=(r,i) \\ \tilde{e} \in \partial i \backslash e}} \nu_{r \to \tilde{e}}(m_{\tilde{e}})e^{-\beta m_{\tilde{e}} \omega_{\tilde{e}}}. \tag{12}$$

We adopt the convention $\sum_{a \in A} f(a) = 0$ and $\prod_{a \in A} f(a) = 1$ if $A = \emptyset$ for any function f. Pictorially, equation (12) can be rendered as
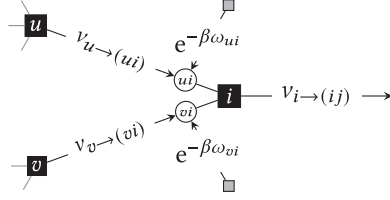
where the arrows indicate the directions of 'propagation' of the messages. Taking advantage of the binary nature of the variables m_e, it is convenient to parametrize the marginals in terms of 'cavity fields' $h_{i \to e}$,

$$\nu_{i \to e}(m) = \frac{e^{\beta m h_{i \to e}}}{1 + e^{\beta h_{i \to e}}}, \tag{13}$$

so that an equation for the cavity fields $h_{i \to e}$ can be written as

$$
\begin{aligned}
h_{i \to e} = -\frac{1}{\beta} \ln \frac{\nu_{i \to e}(0)}{\nu_{i \to e}(1)} = \frac{1}{\beta} \ln \sum_{\{m_{\hat{e}}\}_{\partial i \backslash e}} \mathbb{I} \left(\sum_{\hat{e} \in \partial i \backslash e} m_{\hat{e}} = \kappa_i - 1 \right) \\
\times \prod_{\substack{\hat{e} = (r, i) \\ \hat{e} \in \partial i \backslash e}} e^{\beta m_{\hat{e}}(h_{r \to \hat{e}} - \omega_{\hat{e}})} \\
- \frac{1}{\beta} \ln \sum_{\{m_{\hat{e}}\}_{\partial i \backslash e}} \mathbb{I} \left(\sum_{\hat{e} \in \partial i \backslash e} m_{\hat{e}} = \kappa_i \right) \\
\times \prod_{\substack{\hat{e} = (r, i) \\ \hat{e} \in \partial i \backslash e}} e^{\beta m_{\hat{e}}(h_{r \to \hat{e}} - \omega_{\hat{e}})}.
\end{aligned}
\tag{14}
$$

Once the equations have been solved for all the fields on the graph edges, the marginal probability of the variable m on the edge $e = (i, j)$ is given by

$$\nu_e(m) = \frac{e^{\beta m(h_{i \to e} + h_{j \to e} - \omega_e)}}{1 + e^{\beta(h_{i \to e} + h_{j \to e} - \omega_e)}}, \tag{15}$$

i.e. $\nu_e(1)$ evaluated with $\beta = 1$ is the probability that $e \in \mathscr{F}_k^*$. The BP approximation to the symbol MAP estimator in (6) is obtained computing the messages, and then the marginal, with $\beta = 1$, and then selecting the set

$$
\begin{aligned}
\hat{\mathscr{F}}_k^{(s)}(\mathscr{G}) &:= \left\{ e \in \mathcal{E} : \nu_e(1) > \frac{1}{2} \right\} \\
&= \left\{ e = (i, j) \in \mathcal{E} : h_{i \to e} + h_{j \to e} > \omega_e \right\}.
\end{aligned}
\tag{16}
$$

Observe that proceeding in this way the selected edge-set $\hat{\mathscr{F}}_k^{(s)} \cup \mathscr{F}_k^{(0)}$ is not an actual k-factor in general.

The block MAP estimator is obtained taking $\beta \to +\infty$ in the equations for the marginals: in this limit the measure in equation (10) concentrates on the configuration $\{m_e\}_e$ that maximizes the likelihood. For $\beta \to +\infty$, equation (14) simplify, and we obtain

$$h_{i\to e} = \min^{(\kappa_i)}\left[\{\omega_{\hat{e}} - h_{r\to\hat{e}}\}_{\substack{\hat{e}=(r,i)\\ \hat{e}\in\partial i\setminus e}}\right],\tag{17}$$

where $\min^{(r)}[A]$ outputs the rth smallest element of the set A. The block MAP estimator $\hat{\mathscr{F}}_k^{(b)}$ is found using the same criterion given in equation (16) upon convergence of the algorithm, so that the final estimator for $\mathscr{F}_k^*$ is $\hat{\mathscr{F}}_k^{(b)} \cup \mathscr{F}_k^{(0)}$.

3.3. Recursive distributional equations

In this section we will study the average error on the considered ensemble by analysing the statistical properties of the solutions of the BP equations by the cavity method [21]. We introduce the following random variables.

- $\hat{\mathsf{H}}$ is a random variable that has the law of the cavity field $h_{i\to e}$ given that e is a randomly chosen planted edge;
- H is a random variable that has the law of the cavity field $h_{i\to e}$ given that e is a randomly chosen non-planted edge;
- $\hat{\Omega}$ is a random variable that has the law of the effective weight ω_e given that e is a randomly chosen planted edge;
- Ω is a random variable that has the law of the effective weight ω_e given that e is a randomly chosen non-planted edge.

In the hypothesis that the replica symmetric hypothesis holds (i.e., typical realizations of ν have no long-range correlations), then (14) translates into RDEs involving the introduced random variables H and $\hat{\mathsf{H}}$. To write down this set of RDEs, first observe that an endpoint i of a planted edge e is incident to Z non-planted edges, plus a set $\mathsf{K}' - 1$ of other planted edges. The random variable K', however, is not simply distributed as K, but instead as [13]

$$\mathbb{P}[\mathsf{K}' = \kappa] = \frac{\kappa\mathbb{P}[\mathsf{K} = \kappa]}{\mathbb{E}[\mathsf{K}]} = \binom{k}{\kappa}\frac{\kappa\hat{\mu}^\kappa(1-\hat{\mu})^{k-\kappa}}{k\hat{\mu}}.\tag{18}$$

This is because the planted subgraph, having $\hat{\mu}_k N \gg 1$ vertices, contains $\frac{1}{2}\hat{\mu}_k N\kappa\mathbb{P}[\mathsf{K} = \kappa]$ edges adjacent to a vertex of capacity κ, so that the probability of picking a planted edge that is adjacent to a vertex with capacity κ is proportional to $\kappa\mathbb{P}[\mathsf{K} = \kappa]$. For the sake of brevity, here and in the following, given a random variable X, we will denote by X' a random variable distributed as

$$\mathbb{P}[\mathsf{X}' = x] = \frac{x\mathbb{P}[\mathsf{X} = x]}{\mathbb{E}[\mathsf{X}]}.\tag{19}$$

Similarly, if e is a non-planted edge and i is one of its endpoints, there will be K planted edge incident to i, and other $\mathsf{Z}' - 1$ non-planted edges. Within the replica-symmetric assumption of independence of the incoming cavity fields, one thus obtains from equation (14):

$$\hat{\mathsf{H}} \overset{\mathrm{d}}{=} -\frac{1}{\beta}\ln\frac{\sum_{\{m\},\{\tilde{m}\}}\mathbb{I}\left(\sum_{i=1}^{\mathsf{K}'-1}m_i + \sum_{j=1}^{\mathsf{Z}}\tilde{m}_j = \mathsf{K}'\right)\prod_{i=1}^{\mathsf{K}'-1}\mathrm{e}^{\beta m_i(\hat{\mathsf{H}}_i-\hat{\Omega}_i)}\prod_{j=1}^{\mathsf{Z}}\mathrm{e}^{\beta\tilde{m}_j(\mathsf{H}_j-\Omega_j)}}{\sum_{\{m\},\{\tilde{m}\}}\mathbb{I}\left(\sum_{i=1}^{\mathsf{K}'-1}m_i + \sum_{j=1}^{\mathsf{Z}}\tilde{m}_j = \mathsf{K}'-1\right)\prod_{i=1}^{\mathsf{K}'-1}\mathrm{e}^{\beta m_i(\hat{\mathsf{H}}_i-\hat{\Omega}_i)}\prod_{j=1}^{\mathsf{Z}}\mathrm{e}^{\beta\tilde{m}_j(\mathsf{H}_j-\Omega_j)}},$$

$$\tag{20a}$$

$$\mathsf{H} \overset{\mathrm{d}}{=} -\frac{1}{\beta} \ln \frac{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{\mathsf{K}} m_i + \sum_{j=1}^{\mathsf{Z}'-1} \tilde{m}_j = \mathsf{K}\right) \prod_{i=1}^{\mathsf{K}} \mathrm{e}^{\beta m_i(\hat{\mathsf{H}}_i - \hat{\Omega}_i)} \prod_{j=1}^{\mathsf{Z}'-1} \mathrm{e}^{\beta \tilde{m}_j(\mathsf{H}_j - \Omega_j)}}{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{\mathsf{K}} m_i + \sum_{j=1}^{\mathsf{Z}'-1} \tilde{m}_j = \mathsf{K}-1\right) \prod_{i=1}^{\mathsf{K}} \mathrm{e}^{\beta m_i(\hat{\mathsf{H}}_i - \hat{\Omega}_i)} \prod_{j=1}^{\mathsf{Z}'-1} \mathrm{e}^{\beta \tilde{m}_j(\mathsf{H}_j - \Omega_j)}}.$$

(20b)

The equalities have to be considered in distribution. In the equations above all random variables are independent, Z is Poisson distributed with mean γ, the Ω_i's have the same law as Ω, H_i are independent copies of H, and, similarly, $\hat{\mathsf{H}}_i$ of $\hat{\mathsf{H}}$. Finally, the variable K is distributed as in equation (8a). Observe that being Z Poisson distributed, $\mathsf{Z}' - 1 \overset{\mathrm{d}}{=} \mathsf{Z}$. Note also that, in the $\beta \to +\infty$ limit, equation (20) simplify, giving

$$\hat{\mathsf{H}} \overset{\mathrm{d}}{=} \min_{ij}^{(\mathsf{K}')} \left\{ \{\hat{\Omega}_i - \hat{\mathsf{H}}_i\}_{i=1}^{\mathsf{K}'-1} \cup \{\Omega_i - \mathsf{H}_j\}_{j=1}^{\mathsf{Z}} \right\},$$

(21a)

$$\mathsf{H} \overset{\mathrm{d}}{=} \min_{ij}^{(\mathsf{K})} \left\{ \{\hat{\Omega}_i - \hat{\mathsf{H}}_i\}_{i=1}^{\mathsf{K}} \cup \{\Omega_i - \mathsf{H}_j\}_{j=1}^{\mathsf{Z}'-1} \right\}.$$

(21b)

Recalling the inclusion rule (16), the average reconstruction error is given as

$$\mathbb{E}[\varrho] = \frac{\hat{\mu}}{2} \mathbb{P}[\hat{\mathsf{H}}_1 + \hat{\mathsf{H}}_2 \leqslant \hat{\Omega}] + \frac{\gamma \hat{\mu}_k}{2k} \mathbb{P}[\mathsf{H}_1 + \mathsf{H}_2 > \Omega].$$

(22)

3.4. Hard fields

At this point, we aim at evaluating $\mathbb{E}[\varrho]$ by solving, possibly numerically, the RDEs given in equation (20). It is, however, convenient to first isolate the contribution of 'hard-fields'. It is indeed not difficult to see that the events $\hat{\mathsf{H}} = +\infty$ and $\mathsf{H} = -\infty$ have finite probability. This follows from the fact that $\mathbb{P}[\mathsf{Z} = 0] > 0$ in (20a), which leads to $\hat{\mathsf{H}} = +\infty$, and this event can lead to $\mathsf{H} = -\infty$ in (20b). Let therefore be $\hat{q} := \mathbb{P}[\hat{\mathsf{H}} < +\infty]$ and $q := \mathbb{P}[\mathsf{H} > -\infty]$, probability that the fields are finite. We introduce two new random variables $\hat{H}$ and H that have the law of $\hat{\mathsf{H}}$ and H conditional on being finite:

$$\mathsf{H} \overset{\mathrm{d}}{=} \begin{cases} -\infty & \text{with prob. } 1 - q, \\ H & \text{with prob. } q, \end{cases}$$

(23a)

$$\hat{\mathsf{H}} \overset{\mathrm{d}}{=} \begin{cases} +\infty & \text{with prob. } 1 - \hat{q}, \\ \hat{H} & \text{with prob. } \hat{q}. \end{cases}$$

(23b)

To obtain the equations obeyed by q, $\hat{q}$, H and $\hat{H}$, it is convenient, in this sense, to observe that

$$\hat{\mathsf{H}} \overset{\mathrm{d}}{=} \overset{(\mathsf{K}')}{\min} \left\{ \{\hat{\Omega}_i - \hat{\mathsf{H}}_i\}_{i=1}^{\mathsf{K}'-1} \cup \{\Omega_i - \mathsf{H}_j\}_{j=1}^{\mathsf{Z}} \right\} + O(1/\beta),$$

(24a)

$$\mathsf{H} \overset{\mathrm{d}}{=} \overset{(\mathsf{K})}{\min} \left\{ \{\hat{\Omega}_i - \hat{\mathsf{H}}_i\}_{i=1}^{\mathsf{K}} \cup \{\Omega_i - \mathsf{H}_j\}_{j=1}^{\mathsf{Z}} \right\} + O(1/\beta),$$

(24b)

the correction terms being finite. From these equations we easily get that

$$1 - \hat{q} = \mathrm{e}^{-\gamma q},$$

(25a)

$$1 - q = \sum_{\kappa=1}^{k} \mathbb{P}[\mathsf{K} = \kappa](1 - \hat{q})^{\kappa} = 1 - \frac{1 - (1 - \hat{\mu}\hat{q})^k}{1 - (1 - \hat{\mu})^k},$$

(25b)

that is a set of equations for q and $\hat{q}$. Introducing the variables Z distributed as

$$\mathbb{P}[Z = z] = \frac{1 - \hat{q}}{\hat{q}} \frac{(\gamma q)^z}{z!}, \quad z \in \mathbb{N} \tag{26a}$$

and the variable K distributed as

$$\mathbb{P}[K = \kappa] = \binom{k}{\kappa} \frac{(1 - \hat{\mu}\hat{q})^{k-\kappa}(\hat{\mu}\hat{q})^\kappa}{1 - (1 - \hat{\mu}\hat{q})^k}, \quad \kappa = 1, \ldots, k, \tag{26b}$$

we can reduce ourselves to equations involving 'soft fields' H and $\hat{H}$ only. Using the notation introduced in equation (19) we can write

$$\hat{H} \overset{\mathrm{d}}{=} -\frac{1}{\beta} \ln \frac{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{K'-1} m_i + \sum_{j=1}^{Z} \tilde{m}_j = K'\right) \prod_{i=1}^{K'-1} e^{\beta m_i(\hat{H}_i - \hat{\Omega}_i)} \prod_{j=1}^{Z} e^{\beta \tilde{m}_j(H_j - \Omega_j)}}{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{K'-1} m_i + \sum_{j=1}^{Z} \tilde{m}_j = K'-1\right) \prod_{i=1}^{K'-1} e^{\beta m_i(\hat{H}_i - \hat{\Omega}_i)} \prod_{j=1}^{Z} e^{\beta \tilde{m}_j(H_j - \Omega_j)}},$$
$$\tag{27a}$$

$$H \overset{\mathrm{d}}{=} -\frac{1}{\beta} \ln \frac{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{K} m_i + \sum_{j=1}^{Z'-1} \tilde{m}_j = K\right) \prod_{i=1}^{K} e^{\beta m_i(\hat{H}_i - \hat{\Omega}_i)} \prod_{j=1}^{Z'-1} e^{\beta \tilde{m}_j(H_j - \Omega_j)}}{\sum_{\{m\},\{\tilde{m}\}} \mathbb{I}\left(\sum_{i=1}^{K} m_i + \sum_{j=1}^{Z'-1} \tilde{m}_j = K-1\right) \prod_{i=1}^{K} e^{\beta m_i(\hat{H}_i - \hat{\Omega}_i)} \prod_{j=1}^{Z'-1} e^{\beta \tilde{m}_j(H_j - \Omega_j)}}.$$
$$\tag{27b}$$

In the $\beta \to +\infty$ limit,

$$\hat{H} \overset{\mathrm{d}}{=} \min_{ij}^{(K')} \left\{ \{\hat{\Omega}_i - \hat{H}_i\}_{i=1}^{K'-1} \cup \{\Omega_i - H_j\}_{j=1}^{Z} \right\}, \tag{28a}$$

$$H \overset{\mathrm{d}}{=} \min_{ij}^{(K)} \left\{ \{\hat{\Omega}_i - \hat{H}_i\}_{i=1}^{K} \cup \{\Omega_j - H_j\}_{j=1}^{Z'-1} \right\}. \tag{28b}$$

It is important to note that the solution $\hat{H} = +\infty$ and $H = -\infty$ is an admissible solution of the RDEs for any value of β. This solution corresponds to a full recovery phase, in which the fraction of planted edges that are not correctly recovered vanishes as the system size grows.

The average reconstruction error (22) can be rewritten as

$$\mathbb{E}[\varrho] = \frac{\hat{\mu}\hat{q}^2}{2}\mathbb{P}[\hat{H}_1 + \hat{H}_2 \leqslant \hat{\Omega}] + \frac{\hat{\mu}_k q^2 \gamma}{2k}\mathbb{P}[H_1 + H_2 \geqslant \Omega]. \tag{29}$$

This procedure of 'hard-fields' elimination on the RDEs admits also an interpretation on a single graph instance. Infinite fields on the planted edges may appear in the BP equation (14): if a vertex i has coordination equal to its capacity κ_i, then $h_{i \to e} = +\infty$ for all its incident edges e. This is not surprising: in this case, indeed, all edges incident to i surely belong to the planted k-factor. The vertex i and all the κ_i edges incident to it can be removed and the capacity of its neighbors updated. This removal procedure can be iterated until either $\mathscr{F}_k^*$ has been entirely recovered, or a non-trivial core survives in which every vertex i has $|\partial i| > \kappa_i$. The BP algorithm in equation (14) can be runned on the obtained core.

The described pruning appears as a generalization of the known pruning procedure adopted for the study of optimal matching on Erdós–Rényi graphs [22] and planted matching problems [8]. It corresponds to a process of identification of the planted edges merely based on the topology of the graph $\mathscr{G}$ (and therefore not related to the weights on the edges). A 'percolation

transition' occurs in the parameter of the problem between a phase in which the graph is completely pruned (and the k-factor completely recovered), and a phase in which an (extensive) core survives. The threshold is obtained studying the equation

$$q + \frac{\left(1 - \hat{\mu} + \hat{\mu}\,\mathrm{e}^{-\gamma q}\right)^k - 1}{\hat{\mu}_k} = 0 \tag{30}$$

that has $q = 0$ as attractor for

$$ck\mu\hat{\mu} \leqslant 1. \tag{31}$$

In the considered problem, this condition implies full recovery of the planted configuration by simple topological considerations, i.e., iterative pruning.

4. A numerical experiment: the exponential distribution case

We shall now numerically investigate the planted k-factor problem on the ensemble of graphs introduced above, using the tools described in the previous section. We will take an uniform distribution for the non-planted weights,

$$p(w) = \frac{1}{c}\mathbb{I}(0 \leqslant w \leqslant c), \tag{32}$$

and an exponential distribution

$$\hat{p}(w) = \lambda\,\mathrm{e}^{-\lambda w}\mathbb{I}(w \geqslant 0), \tag{33}$$

for the planted weights. It follows that in this case $\Gamma = [0, c]$, $\mu = 1$, $\hat{\mu} = 1 - \mathrm{e}^{-c\lambda}$ and $\gamma = c(1 - \mathrm{e}^{-kc\lambda})$. In the $c \to +\infty$ limit, $\hat{\mu} = \mu = 1$ and $q = \hat{q} = 1\,\forall\,k \in \mathbb{N}$.

The large N limit results for $\mathbb{E}[\varrho]$ have been obtained by a numerical resolution of the RDEs for the soft fields for $\beta = 1$ and $\beta \to +\infty$ (see equation (28)) via a population dynamics (PD) algorithm [21]. The approach consists in introducing an iterative version of the RDEs discussed above that defines a new set of random variables $(\hat{H}^{(n)}, H^{(n)})_n$, $n = 0, 1, \ldots$ These new random variables satisfy, for $\beta \to +\infty$,

$$\hat{H}^{(n+1)} \overset{\mathrm{d}}{=} \min_{ij}{}^{(K')} \left\{ \{\hat{\Omega}_i - \hat{H}_i^{(n)}\}_{i=1}^{K'-1} \cup \{\Omega_i - H_j^{(n)}\}_{j=1}^{Z} \right\}, \tag{34a}$$

$$H^{(n)} \overset{\mathrm{d}}{=} \min_{ij}{}^{(K)} \left\{ \{\hat{\Omega}_i - \hat{H}_i^{(n)}\}_{i=1}^{K} \cup \{\Omega_j - H_j^{(n-1)}\}_{j=1}^{Z'-1} \right\}, \tag{34b}$$

with initial condition $\hat{H}^{(0)} = H^{(-1)} = 0$. A similar set of iterative RDEs can be written for $\beta = 1$ starting from equation (27). The underlying assumption is that, if a non-trivial solution of the original RDEs exists, such solution will be reached in probability by equation (34) for $n \to +\infty$.

From the algorithmic point of view, the law of the random variable $H^{(n)}$ is represented by an empirical distribution of a sample $\{h_1, \ldots, h_{\mathcal{N}}\}$ of its representants, with $\mathcal{N} \gg 1$, so that

$$\mathbb{P}[H^{(n)} \leqslant h] \approx \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbb{I}(h_i^{(n)} \leqslant h). \tag{35}$$

Similarly, an empirical distribution is adopted to approximate the law of $\hat{H}^{(n)}$. The RDEs (34) are used to update the population representing $(\hat{H}^{(n)}, H^{(n)})$ to a new population representing the

variables $(\hat{H}^{(n+1)}, H^{(n+1)})$, each update corresponding to an 'iteration' of the algorithm. The algorithm stops when some convergence criterion (usually the convergence of the moments of the populations) is satisfied (see, e.g., reference [13] for additional details).

The PD predictions have been compared with actual BP results, obtained solving the inference problem on a large number of instances (see also section 6 for further details about the BP implementation).

In the block MAP case ($\beta \to +\infty$) it is observed that, for a given pair k and c, there exists an interval $\Lambda^{(b)} := (\lambda_-^\star, \lambda_+^\star)$ of values of λ, such that for $\lambda \in \Lambda^{(b)}$ the PD algorithm converges to a finite solution. The corresponding value of $\mathbb{E}[\varrho]$ is found to be non-zero and, within numerical precision, $\mathbb{E}[\varrho] \to 0$ smoothly on the boundary of the interval. In particular, we numerically observe that

$$\lim_{\lambda \to \lambda_+^\star} \sqrt{\lambda_+^\star - \lambda} \, \ln \mathbb{E}[\varrho] = -\alpha, \quad \alpha > 0. \tag{36}$$

This implies that $\mathbb{E}[\varrho]$ approaches zero exponentially fast as $\lambda \to \lambda_+^\star$, i.e., the transition is of infinite order. Remarkably, the very same behavior has been observed in the $k = 1$ and $c \to +\infty$ case [8,9], where it is shown that $\lambda_+^\star = 4$ and

$$\ln \mathbb{E}[\varrho] = -\frac{2\pi}{\sqrt{4 - \lambda}} - \frac{3}{2} \ln(4 - \lambda) + O(1). \tag{37}$$

Finally, for $c \to +\infty$ it is observed that $\lambda_-^\star \to 0$, whereas $\lambda_+^\star$ approaches a finite limit. The obtained prediction for $\mathbb{E}[\varrho]$ is fully compatible with the BP results, and $\mathbb{E}[\varrho] \to 0$ for $\lambda \to \lambda_\pm^\star$ with the size of the considered graphs.

On the other hand, for $\lambda \notin \Lambda^{(b)}$ the PD algorithm does not converge. To be more precise, the population is subject to a 'drift' towards the full recovery solution, $H \to -\infty$ and $\hat{H} \to +\infty$. This can be seen, for example, in figure 2(c), where some numerical results for $c = k = 2$ are given in the $\beta \to +\infty$ case for $\hat{H}$. It is seen that the numerically estimated mean $\mathbb{E}[\hat{H}]$ is population-size-dependent, and in particular diverges with the population size, whereas the variance is not. Moreover, larger populations correspond to larger values of $\mathbb{E}[\hat{H}]$.

The numerical results suggest therefore that a nontrivial, attractive fixed point exists for $\lambda \in \Lambda^{(b)}$ only, otherwise the only attractor being the trivial fixed point $\hat{H} = -H = +\infty$ corresponding to the full recovery phase. In reference [8] it is argued that, for $k = 1$, an infinite order phase transition takes place between a full recovery phase and a partial recovery phase, and in particular full recovery is obtained for $\lambda \in \mathbb{R}^+ \backslash \Lambda^{(b)}$. The conjectures in reference [8] about the location of the transition and its nature have been recently rigorously proved in reference [9]. Our results strongly suggest that the same phenomenology extends to the $k > 1$ case. As in the $k = 1$ case, the accurate numerical determination of the endpoints of $\Lambda^{(b)}$ is heavily affected by finite-population-size effects using PD (and finite-size effects using BP). Indeed, the transition manifests itself as a front propagation in the cumulative distribution function that drifts towards large values of the fields. Such front propagations are generically driven by the behavior in the exponentially small tail far away from the front [16]. The finite population size induces a cutoff on the smallest representable value of the cumulative distribution function, that translates, assuming an exponential decay of the cumulative, into logarithmic finite population size effects on the velocity of the front and the location of the transition.

The very same phenomenology is observed for $\beta = 1$, i.e., for the symbol MAP, where a partial recovery phase $\lambda \in \Lambda^{(s)} = (\lambda_-^\star, \lambda_+^\star)$ is surrounded by a full recovery phase. For a given pair k and c of parameters, the symbol MAP transition points are found to be very close to the block MAP transition points obtained for the same values of k and c. Also in this case,
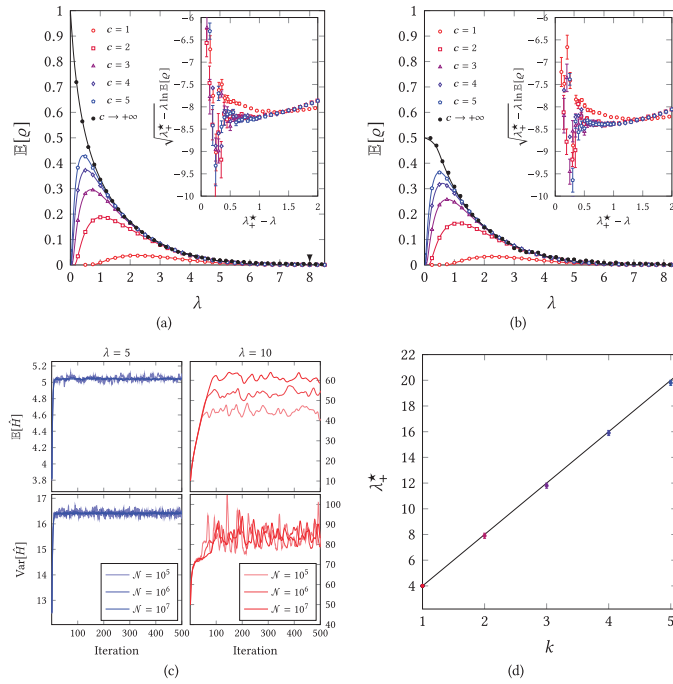
Figure 2. (a) and (b) Reconstruction error using the block MAP (a) and the symbol MAP (b) estimators for the planted two-factor problem with exponential planted weights and different values of average degree c of the non-planted edges. The lines have been obtained numerically solving the RDEs (28), corresponding to $\beta = +\infty$, and (27) with $\beta = 1$ respectively, using a PD algorithm with $\mathcal{N} = 10^6$ fields. The dots are obtained solving 10^3 instances of the problem using the BP algorithm in equations (17) and (14) with $\beta = 1$ on graphs of $N = 10^3$ vertices. The $c \to +\infty$ curves are estimated using PD with $c = 20$ and are compared with BP results obtained solving the problem on complete graphs with $N = 10^2$ vertices. In the block MAP case, we marked with an arrow the partial to full recovery threshold in the $c \to +\infty$ limit predicted to be at $\lambda = 4k$. *In the insets*, plot of $\sqrt{\lambda_+^\star - \lambda} \ln \mathbb{E}[\varrho]$ given by the PD algorithm, with $\lambda_+^\star$ estimated using the condition in equation (54). Both in the block MAP case and in the symbol MAP case, the limiting value for $\lambda \to \lambda_+^\star$ is finite, suggesting that the transition is of infinite order. (c) Mean and variance of the variable $\hat{H}$ estimated using a PD algorithm for $k = c = 2$ for $\beta \to +\infty$. In this case, we numerically find $\lambda_+^\star = 7.9(1)$, whereas the prediction given by equation (54) is $\lambda_+^\star = 7.9946 \dots$. The results are obtained assuming all the $\mathcal{N}$ fields equal to zero as initial condition. An 'iteration' of the algorithm corresponds to an update of all fields of the population by means of the RDEs. For $\lambda = 5$, inside the partial recovery phase, the algorithm rapidly converges to asymptotic values that do not depend on the size $\mathcal{N}$ of the population. For $\lambda = 10$, i.e., in the full recovery phase, the algorithm converges to values of $\mathbb{E}[\hat{H}]$ that are $\mathcal{N}$-dependent and diverge with $\mathcal{N}$, whereas the variance of the distribution goes to an $\mathcal{N}$-independent value (although noisy). The described phenomenology appeared for all the investigated values of c and k, and for both the block MAP and the symbol MAP. (d) Transition point $\lambda_+^\star$ for the block MAP in the exponential model for $c \to +\infty$. The continuous line correspond to the prediction $\lambda_+^\star = 4k$ given in section 6.2, whereas the dots are the numerical estimation of the transition point obtained using PD with $c = 50$ and a population of $\mathcal{N} = 10^8$ fields.

it is found that $(\lambda_+^\star - \lambda)^{1/2} \ln \mathbb{E}[\varrho] \to -\hat{\alpha}$ for $\lambda \to \lambda_+^\star$ for some $\hat{\alpha} > 0$, suggesting that the transition between the partial recovery phase and the full recovery phase is of infinite order also in this case. In figure 2(b) we give the PD results for $\mathbb{E}[\varrho]$ in this case, alongside with the results of the BP simulations.

5. A criterion for the block MAP transition

In this section we give a heuristic criterion for the transition between the partial recovery phase $\mathbb{E}[\varrho] > 0$, and the full recovery phase $\mathbb{E}[\varrho] = 0$ in the case of the block MAP. Our reasoning will follow and generalize the one given in reference [8] for the $k = 1$ case. Our approach is inspired by the physics literature on front propagation in reaction-diffusion systems and equations of the FKPP type [16–18]. Before applying it, however, an additional simplification of equation (28) must be performed. We will proceed in generality, assuming that $\hat{p}$ depends on a parameter, let us call it λ. Moreover, we will assume that a special value $\lambda^\star$ exists such that for $\lambda < \lambda^\star$ we are in a partial-recovery phase, whereas for $\lambda > \lambda^\star$ we are in a full-recovery phase. We will also assume that the transition is *continuous*, i.e., $\mathbb{E}[\varrho] \to 0$ smoothly as $\lambda \to \lambda^{\star-}$. Observe that $\mathbb{E}[\varrho] \to 0$ means that $\Pr[H_1 + H_2 < \Omega] \to 1$ and $\Pr[\hat{H}_1 + \hat{H}_2 > \hat{\Omega}] \to 1$, see equation (29). The first property implies that, approaching the transition, $H_1 < \Omega - H_2$ almost surely, i.e., in equation (28b) the minimum picks almost surely one of the 'planted contributions'. Similarly, the second property implies that in the same limit the minimum in equation (28a) is almost surely picked in the set of 'non-planted contributions'. These observations lead us to introduce a new set of random variables $(\hat{U}^{(n)}, U^{(n)})_n$, satisfying the iterative RDEs,

$$\hat{U}^{(n+1)} \stackrel{\mathrm{d}}{=} \min_{1 \leqslant i \leqslant Z} \{\Omega_i - U_i^{(n)}\}, \tag{38}$$

$$U^{(n)} \stackrel{\mathrm{d}}{=} \max_{1 \leqslant i \leqslant K} \{\hat{\Omega}_i - \hat{U}_i^{(n)}\}, \tag{39}$$

with initial condition $\hat{U}^{(0)} = U^{(0)} = 0$, corresponding to the expected 'effective' behavior of equation (34) near the transition. The new set of auxiliary variables is informative on the behavior of the random variables $(\hat{H}^{(n)}, H^{(n)})_n$. Indeed, we can prove that

$$\hat{U}^{(n)} \preceq \hat{H}^{(n)}, \quad H^{(n-1)} \preceq U^{(n)}, \quad \forall\, n. \tag{40}$$

Given two random variables X and Y, we say that $X \preceq Y$ if $\mathbb{P}[X > z] \leqslant \mathbb{P}[Y > z]$ for all z [23]. The proof proceeds by induction. Equation (40) are satisfied for $n = 0$. Assuming that they are satisfied for given n, it is easily proved that they are satisfied for $n + 1$, being

$$H^{(n)} \preceq \max_{1 \leqslant i \leqslant K} \{\hat{\Omega}_i - \hat{H}_i^{(n-1)}\} \preceq \max_{1 \leqslant i \leqslant K} \{\hat{\Omega}_i - \hat{U}_i^{(n)}\} = U^{(n+1)}, \tag{41a}$$

$$\hat{U}^{(n+1)} = \min_{1 \leqslant j \leqslant Z} \{\Omega_j - U_j^{(n)}\} \preceq \min_{1 \leqslant j \leqslant Z} \{\Omega_j - H_j^{(n)}\} \preceq \hat{H}^{(n+1)}. \tag{41b}$$

The result stated above implies that, if $\Pr[\hat{U}^{(n)} > z] \to 1$ for $n \to +\infty$, then $\Pr[\hat{H}^{(n)} > z] \to 1$ as well in the same limit. We will obtain now a sufficient condition to have $\Pr[\hat{U}^{(n)} > z] \to 1$ that we will give us, therefore, a (sufficient) criterion to be in the full recovery phase. We define

$$F(x; n) := \mathbb{P}[\hat{U}^{(n)} < x], \tag{42}$$

$$\Phi(x; n) := \mathbb{P}[U^{(n)} < x]. \tag{43}$$

Denoting by $\mathbb{E}_X[\bullet]$ the expectation with respect to the random variable X, we have that

$$F(x; n+1) = 1 - \mathbb{E}_Z\left[\left(\mathbb{E}_\Omega\left[\Phi(\Omega - x; n)\right]\right)^Z\right], \tag{44}$$

$$\Phi(x; n) = \mathbb{E}_K\left[\left(1 - \mathbb{E}_{\hat\Omega}\left[F(\hat\Omega - x; n)\right]\right)^K\right], \tag{45}$$

and therefore

$$F(x; n+1) = 1 - \mathbb{E}_Z\left[\left(\mathbb{E}_\Omega\left[\left(1 - \mathbb{E}_{\hat\Omega}\left[F(\hat\Omega - \Omega + x; n)\right]\right)^K\right]\right)^Z\right]. \tag{46}$$

Suppose now that the cumulative $F(x; n)$ is subject to a 'drift', i.e., there exists a velocity v such that $F(x + vn; n) \to F(x)$ as $n \to \infty$. This is in line with the numerical result, which suggest that $\mathbb{E}[\hat H] \to +\infty$ and $\mathbb{E}[H] \to -\infty$ approaching the transition, whereas the higher order cumulants remain finite. Figure 2(c), in particular, shows that the means of the distributions are subject to a constant drift velocity that (at the leading order in n) does not depend on n. Moreover, this assumption is compatible with what has been observed in [8], and rigorously proved in [14, 15], in the study of equation (38) for $K \equiv 1$. Then, for $n \to +\infty$,

$$F(x - v) = 1 - \mathbb{E}_Z\left[\left(\mathbb{E}_\Omega\left[\left(1 - \mathbb{E}_{\hat\Omega}\left[F(\hat\Omega - \Omega + x)\right]\right)^K\right]\right)^Z\right]. \tag{47}$$

For $x \to -\infty$, $F(x) \to 0$ by definition, and in this limit at first order in F

$$F(x - v) \simeq \mathbb{E}[Z]\mathbb{E}[K]\mathbb{E}\left[F(\hat\Omega - \Omega + x)\right] \tag{48}$$

(we have dropped the subscripts implying an average over all variables in the argument). This linear (integral) equation has a solution in the form $F_v(z) = e^{\theta z}$, with $\theta > 0$ to respect the increasing character of distribution functions. Indeed, plugging this solution into the linearized equation we have

$$v(\theta) \simeq -\frac{\ln\left(\mathbb{E}[Z]\mathbb{E}[K]\mathbb{E}\left[\exp(\theta\hat\Omega - \theta\Omega)\right]\right)}{\theta}. \tag{49}$$

The choice of the appropriate θ to estimate the drift velocity is, at this point, not obvious. It can be shown [14–18] that the relevant value θ^* is the one that corresponds to the *maximum* velocity, i.e., $\theta^* = \arg\sup_{\theta>0} v(\theta)$, and therefore

$$v = -\inf_{\theta>0} \frac{\ln\left(\mathbb{E}[Z]\mathbb{E}[K]\mathbb{E}\left[\exp(\theta\hat\Omega - \theta\Omega)\right]\right)}{\theta} \tag{50a}$$

$$= -\inf_{\theta>0} \frac{\ln\left[\mathcal{I}(\theta)\mathcal{I}(1-\theta)\right]}{\theta} \tag{50b}$$

where

$$\mathcal{I}(\theta) := \sqrt{ck}\int_\Gamma \hat p^\theta(w)p^{1-\theta}(w)\mathrm{d}w. \tag{51}$$

If $v = v(\theta^*) > 0$, then the distribution drifts towards $+\infty$ and we are in a full recovery phase ($\hat H \to +\infty$). We postulate that the marginal condition $v(\theta^*) = 0 \Rightarrow \ln\left[\mathcal{I}(\theta^*)\mathcal{I}(1-\theta^*)\right] =$

0 corresponds to the transition point. Being $\ln(\mathcal{I}(\theta)\mathcal{I}(1-\theta))$ a convex function symmetric around $\theta = 1/2$, one has $v = 0$ when $\mathcal{I}(1/2) = 1$. This condition can be written as

$$\frac{D_{1/2}(\hat{p}\|p)}{\ln(ck)} = 1, \tag{52}$$

where

$$D_\alpha(p\|q) := \frac{1}{\alpha - 1} \ln \int p^\alpha(x) q^{1-\alpha}(x) \mathrm{d}x, \tag{53}$$

is the Rényi divergence of order α. In an equivalent form, equation (52) can be written as

$$\int_\Gamma \sqrt{\hat{p}(w)p(w)} \mathrm{d}w = \frac{1}{\sqrt{ck}}. \tag{54}$$

The condition above generalizes the one obtained for the planted matching problem [8], that is recovered for $k = 1$.

As a final comment, observe that, being by construction $\mathbb{E}[\varrho(\hat{\mathscr{F}}_k^{(s)}, \mathscr{F}_k^*)] \leqslant \mathbb{E}[\varrho(\hat{\mathscr{F}}_k^{(b)}, \mathscr{F}_k^*)]$, full recovery by means of the block MAP implies full recovery by means of the symbol MAP, and the partial recovery interval obtained using the symbol MAP is contained in the partial recovery interval of the block MAP.

6. Examples of results

In this section, we consider some special formulations of the planted k-factor problem, and we compare our numerical results with the theoretical predictions obtained from the general criterion given in section 5.

The numerical results are obtained studying the RDEs (20) for the problem by means of a PD algorithm for $\beta = 1$ (symbol MAP) and $\beta \to +\infty$ (block MAP).

We also implemented a BP algorithm for the solution of the problem on actual graphs by means of the algorithm in equation (14). In particular, a random weighted graph with N vertices is generated according to the ensemble introduced in section 2. This graph is first subject to the pruning procedures described in sections 3.1 and 3.4. Given an edge $e = (i, j)$ of the resulting graph, we associate two fields $h_{i\to e}$ and $h_{j\to e}$ to it, initialized to random values. The fields are updated using equation (14) if $\beta > 0$ or using equation (17) if $\beta = +\infty$ for a large number of iterations. A candidate solution $\hat{\mathscr{F}}_k$ is then selected using the criterion in equation (16). In all cases, we stopped the algorithm after $5N$ updates, or before if the set $\hat{\mathscr{F}}_k$ does not change for at least 50 iterations. The error ϱ is then obtained using equation (4), and the average error $\mathbb{E}[\varrho]$ is estimated considering a large number of independent instances of the problem.

The BP algorithm for the estimation of the block MAP given in equation (17) coincides with the BP algorithm for the minimum weight k-factor introduced by Bayati and coworkers [24]. They proved therein that the algorithm converges in polynomial time to the correct minimizer as long as there are no fractional solutions, i.e., solutions with non-integer values of m_e. A worst-case analysis of the convergence properties of the BP algorithm with $\beta = 1$, on the other hand, is still missing. Observe that the BP algorithm has longer running time for $c \gg k$ when β is finite with respect to the $\beta \to +\infty$ case. In this case, indeed, given a node with valence κ, each step in equation (14) requires the sum of $O(c^\kappa)$ contributions. Equation (17), instead, asks only for to the κth incoming field, an operation that requires $O(c)$ steps. On complete graphs, having $c = N - k - 1$, this means that the algorithm running time is increased by a factor N^{k-1} in the finite β case with respect to the $\beta = +\infty$ case.

6.1. The fully-connected case

Dense models are recovered in our setting considering the $c \to +\infty$ limit, to be taken after the $N \to +\infty$ limit. Assuming that $\mu\hat{\mu}_k \neq 0$ (the problem is otherwise trivial) this implies $\gamma \to +\infty$ and therefore $\hat{q} = q = 1$ because of equation (25), meaning that in the thermodynamic limit there are (almost surely) no hard fields.

Equation (52) also implies that, if $\lim_{c\to+\infty} D_{1/2}(\hat{p}\|p) < +\infty$ no transition can take place in the fully connected limit. To get nontrivial results in this limit, it is therefore necessary to scale the weights with c, so that at the transition

$$D_{1/2}(\hat{p}\|p) = \ln c + o(\ln c) \quad \text{for } c \gg 1. \tag{55}$$

Suppose, for example, that p is c-independent, and $\hat{p}(w) \equiv c^{-a} f(wc^{-a}; b)$ with $a > 0$ and b parameters, and some function f such that $f_0(b) := \lim_{x\to0} f(x; b) \in (0, +\infty)$. Then the condition (55) implies that the threshold for $c \to +\infty$ is at

$$a = 1. \tag{56}$$

If instead $\hat{p}(w) \equiv c^{-1} f(wc^{-1}; b)$, then the asymptotic formula (55) is not sufficient anymore and equation (54) must be considered. The threshold condition becomes

$$\sqrt{f_0(b)} \int \sqrt{p(w)} \mathrm{d}w = \frac{1}{\sqrt{k}}. \tag{57}$$

The observations above are compatible with the rigorous results obtained by Bagaria and coworkers for the planted two-factor problem on complete graphs of N vertices for $N \to +\infty$ [12]. In their paper, they prove that, for $k = 2$, on the threshold the following limit holds

$$\liminf_{N\to+\infty} \frac{D_{1/2}(\hat{p}\|p)}{\ln N} = 1 \tag{58}$$

under some assumptions on the distributions p and $\hat{p}$ (fulfilled, e.g., by Gaussian or exponential distributions, see reference [12] for details). This condition corresponds to equation (55), observing that on a complete graph $c = N - k - 1$. As we will show below, however, equation (58) can be not sufficient to recover the transition point.

6.2. The exponential model

Let us now briefly revisit the exponential case discussed in section 4. If we apply equation (54) to derive the block MAP threshold, we obtain

$$2 - 2\,\mathrm{e}^{-\frac{c\lambda}{2}} = \sqrt{\frac{\lambda}{k}}. \tag{59}$$

Introducing $s = ck$ and $t = \lambda k^{-1}$, one parameter can be absorbed obtaining

$$2 - 2\,\mathrm{e}^{-\frac{st}{2}} = \sqrt{t}. \tag{60}$$

This equation is always solved by $t = \lambda = 0$. For $s = ck \gtrsim 1.2277\dots$, two additional solutions for t, and therefore λ, appear, let us call them $\lambda_-^\star$ and $\lambda_+^\star$, delimiting the partial recovery phase, see figure 3. In the $c \to +\infty$ limit, $\lambda_-^\star \simeq 1/c \to 0$ and only one transition point is found
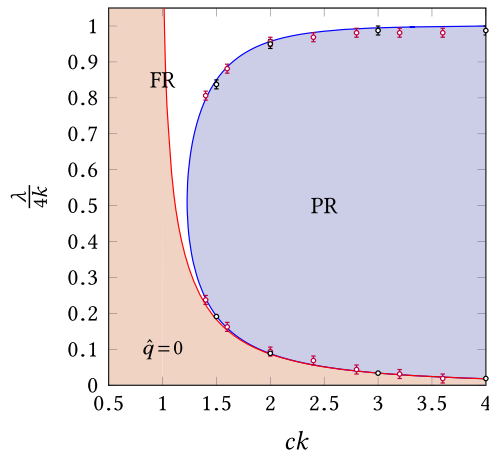
$$\lambda_+^\star = 4k. \tag{61}$$

Figure 3. Phase diagram of the planted k-factor with exponential planted weights on sparse graphs. Our argument in section 5 predicts a partial recovery (PR) phase (in blue) and a full recovery phase, corresponding to the remaining portion of the plane, depending on the values of λ and c. Within the full recovery phase, the red area corresponds to the set of parameters for which full recovery is possible by means of pruning, i.e., $\hat{q} = q = 0$. The red dots have been obtained from the numerical resolution of the RDEs (17) for the two-factor by a population dynamics algorithm with $\mathcal{N} = 10^7$. The black dots correspond instead to the $k = 1$ case and are taken from reference [8].

This result is confirmed by the numerics, see figure 2(d). For $k = 1$ we recover the known result $\lambda_+^\star = 4$, rigorously proved in [7]. The criterion predicts that $4k - \lambda_+^\star$ approaches zero as e^{-ck} for $ck \to +\infty$. Observe that in this case, considering $c = N - k - 1$, $D_{1/2}(\hat{p}\|p) = \ln N + O(N)$ for all values of k and $\lambda > 0$. In other words, equation (58) of [12], although verified on the transition, is not enough to recover the threshold.

6.3. Hidden Hamiltonian cycle recovery

In this section we are interested in solving a special type of planted two-factor problem, namely the hidden Hamiltonian cycle recovery (HC) problem. This is a planted two-factor problem in which the hidden two-factor is connected, i.e., it is a Hamiltonian cycle of the graph. The very same BP algorithms discussed for the planted two-factor can be applied to recover the hidden Hamiltonian cycle.

This problem was studied in reference [12], with the planted weights being assumed to be normal variables, $\hat{p} = \mathcal{N}(\lambda, 1)$, whereas the non-planted weights have distribution $p = \mathcal{N}(0, 1)$. In this case, therefore, $D_{1/2}(\hat{p}\|p) = \frac{1}{4}\lambda^2$. Applying equation (58), a nontrivial transition in the block MAP estimator is expected at $\lambda^2 = 4 \ln N + o(\ln N)$. Parametrizing $\hat{p}$ with $\lambda^2 = \hat{\lambda}^2 \ln N$, the transition is then at $\hat{\lambda} = 2$. This is rigorously proved and numerically verified in reference [12], see figure 4. In figure 4 we also plot, for the block MAP case, the probability that the estimator $\hat{\mathscr{F}}_2$ provided by the BP algorithm is actually connected, i.e., it is a single Hamiltonian cycle. Our numerics suggest that, for $\hat{\lambda} > 2$, this probability goes to 1 as $N \to +\infty$.

As discussed in section 2, however, the block MAP is not the estimator that minimizes the error in equation (4). If we aim at minimizing the error ϱ, then the optimal estimator is the symbol MAP. We recall that this estimator does not provide a k-factor in general. Running our BP algorithm on graphs with hidden Hamiltonian cycles in the ensemble considered in [12], we
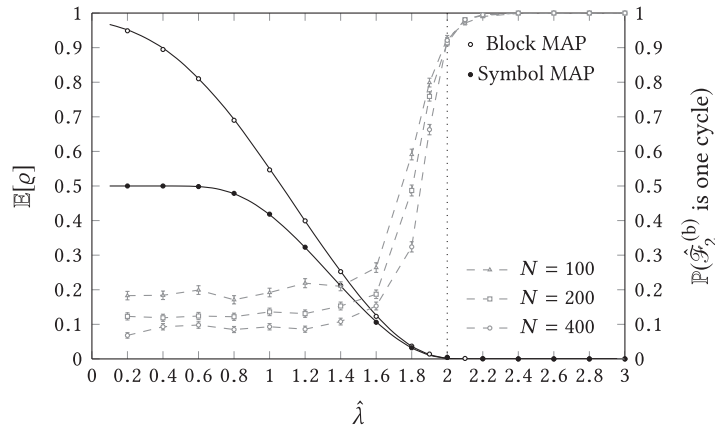
Figure 4. Recovery using the block MAP and the symbol MAP in the HC problem with $\hat{p} = \mathcal{N}(\hat{\lambda}\sqrt{\ln N}, 1)$ and $p = \mathcal{N}(0, 1)$. The dots are obtained running BP on 10^3 instances of complete graphs with $N = 100$ with a hidden Hamiltonian cycle to be recovered. The black lines correspond to the PD prediction for the planted two-factor with $c = 100$, with same planted and non-planted weight distributions. The vertical line indicates the transition point between partial and full recovery predicted by the theory for the block MAP. In gray, we plot the probability that the block MAP estimator $\hat{\mathscr{F}}^{(b)}_{k=2}$ is a Hamiltonian cycle (instead of a union or two or more cycles) for different sizes of the problem.

obtained the results in figure 4. The average error obtained using the symbol MAP is found to be smaller than the one obtained using the block MAP, as expected. On the other hand, finding the symbol MAP is computationally more expensive, as discussed above. For comparison, in figure 4 we plot also the PD results for the two-factor, finding a good agreement between the infinite-size prediction of the two-factor problem and the BP results of the HC problem also in the partial recovery phase.

7. Perspectives

The transition appearing in the planted k-factor problem is of the same type as found in the planted matching problem [7–9] and separates a partial recovery phase from a full recovery phase. Using heuristic arguments based on the literature on front propagation for reaction–diffusion equations, we have been able to obtain a simple and explicit criterion for the transition. We numerically tested the transition criterion, and we checked its consistency with the known results on the recovery thresholds of the planted two-factor problem. A rigorous proof of this transition criterion remains, however, as an open problem.

The heuristic argument is based on the fact (numerically observed) that the phase transition is continuous. It is not excluded *a priori* that first order transitions are possible for some nontrivial choice of the weight distributions or degree distributions of the graph, allowing the presence of multiple BP fixed points [25].

Finally, the threshold criterion obtained in the paper concerns the block MAP and only provides a bound for full recovery by means of the symbol MAP. In the numerically investigated cases, the recovery thresholds of the symbol MAP are observed to be very close to the ones of

the block MAP. A formula for the exact location of the symbol MAP transition (and possibly its relation with the block MAP transition) is however still missing.

Acknowledgments

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Gabriele Sicuro https://orcid.org/0000-0002-9258-2436

References

[1] Decelle A, Krzakala F, Moore C and Zdeborová L 2011 *Phys. Rev.* E **84** 066106
[2] Mézard M, Tarzia M and Toninelli C 2008 *J. Stat. Phys.* **131** 783–801
[3] Richardson T and Urbanke R 2008 *Modern Coding Theory* (Cambridge: Cambridge University Press)
[4] Chertkov M, Kroc L, Krzakala F, Vergassola M and Zdeborová L 2010 *Proc. Natl Acad. Sci. USA* **107** 7663–8
[5] Abbe E and Sandon C 2015 Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery *Annual IEEE Symp. on Foundations of Computer Science, FOCS* vol 2015 (December) (IEEE Computer Society) pp 670–88
[6] Abbe E, Bandeira A S and Hall G 2016 *IEEE Trans. Inform. Theory* **62** 471–87
[7] Moharrami M, Moore C and Xu J 2019 The planted matching problem: phase transitions and exact results (arXiv:1912.08880)
[8] Semerjian G, Sicuro G and Zdeborová L 2020 *Phys. Rev.* E **102** 022304
[9] Ding Jing, Wu Yihong, Xu Jiaming and Yang Dana 2021 The planted matching problem: Sharp threshold and infinite-order phase transition (arXiv:2103.09383)
[10] Stauffer D and Aharony A 2018 *Introduction to Percolation Theory* (Boca Raton, FL: CRC Press)
[11] Cai T, Liang T and Rakhlin A 2017 *IEEE Trans. Netw. Sci. Eng.* **4** 165–76
[12] Bagaria V, Ding J, Tse D, Wu Y and Xu J 2020 *Oper. Res.* **68** 53–70
[13] Mézard M and Montanari A 2009 *Information, Physics, and Computation* (*Oxford Graduate Texts*) (Oxford: Oxford University Press)
[14] Kingman J F C 1975 *Ann. Probab.* **3** 790–801
[15] Biggins J D 1977 *J. Appl. Probab.* **14** 630–6
[16] Brunet E and Derrida B 1997 *Phys. Rev.* E **56** 2597–604
[17] Majumdar S N and Krapivsky P L 2000 *Phys. Rev.* E **62** 7735–42
[18] Ebert U and van Saarloos W 2000 *Physica* D **146** 1–99
[19] Bollobás B 2001 *Random Graphs* (*Cambridge Studies in Advanced Mathematics*) 2nd edn (Cambridge: Cambridge University Press)
[20] Gao P and Wormald N 2017 *SIAM J. Comput.* **46** 1395–427
[21] Mézard M and Parisi G 2001 *Eur. Phys. J.* B **20** 217–33

[22] Karp R and Sipser M 1981 Maximum matchings in sparse random graphs *Proc. of the 22nd Annual IEEE Symp. on Foundations of Computing* pp 364–75

[23] Lindvall T 2012 *Lectures on the Coupling Method* (*Dover Books on Mathematics*) (New York: Dover)

[24] Bayati M, Borgs C, Chayes J and Zecchina R 2011 *SIAM J. Discrete Math.* **25** 989–1011

[25] Bordenave C, Lelarge M and Salez J 2013 *Probab. Theor. Relat. Fields* **157** 183–208