J. Appl. Probab. **62,** 136–152 (2025)
*doi:*10.1017/jpr.2024.72

MINIMAL H-FACTORS AND COVERS

JOEL LARSSON DANIELSSON,* *Chalmers University of Technology*
LORENZO FEDERICO,** *LUISS Guido Carli*

Abstract

Given a fixed small graph H and a larger graph G, an H-factor is a collection of vertex-disjoint subgraphs $H' \subset G$, each isomorphic to H, that cover the vertices of G. If G is the complete graph K_n equipped with independent $U(0,1)$ edge weights, what is the lowest total weight of an H-factor? This problem has previously been considered for $H = K_2$, for example. We show that if H contains a cycle, then the minimum weight is sharply concentrated around some $L_n = \Theta(n^{1-1/d^*})$ (where d^* is the maximum 1-density of any subgraph of H). Some of our results also hold for H-covers, where the copies of H are not required to be vertex-disjoint.

Keywords: Graph tiling; factor; cover; sharp concentration

2020 Mathematics Subject Classification: 60C05; 05B40; 05C70

1. Introduction

1.1. Threshold and minimum-weight problems

Let K_n denote the complete graph on n vertices, equipped with independent and identically distributed (i.i.d.) edge weights $\{X_e\}_{e\in\mathcal{E}(K_n)}$. We will use the terms 'weight' and 'cost' interchangeably. For now, let the weight distribution be uniform on [0, 1]; it will turn out that, for example, Exp(1) weights will give the same asymptotic behaviour. For details, see Section 2.4. For any family $\mathcal{F}$ of subgraphs of K_n, there are two closely related problems.

Threshold. What is the smallest p such that an $F \in \mathcal{F}$ is likely to appear in $G_{n,p}$? That is, if we define the random variable

$$T := \min_{F\in\mathcal{F}} \max_{e\in\mathcal{E}(F)} X_e,$$

what is its distribution? Is it sharply concentrated around its expected value?

Minimum weight. The minimal weight of an $F \in \mathcal{F}$ is a random variable

$$W := \min_{F\in\mathcal{F}} \sum_{e\in\mathcal{E}(F)} X_e.$$

What is its distribution? Is it sharply concentrated?

Received 24 April 2023; accepted 19 June 2024.
* Email address:joel.danielsson@chalmers.se
** Email address: lfederico@luiss.it

This pair of problems has been studied for many families $\mathcal{F}$, particularly for families where each $F \in \mathcal{F}$ is spanning, i.e. $\mathcal{V}(F) = \mathcal{V}(K_n)$. Threshold problems are generally more well studied than the corresponding minimum-weight problems. It has been observed that for many natural choices of $\mathcal{F}$, the property of $G_{n,p}$ containing some $F \in \mathcal{F}$ exhibits the sharp threshold phenomenon, that is, T is sharply concentrated around its mean. And for these families, this is often true of the minimum weight W as well.

For instance, if $\mathcal{F}$ is the family of spanning trees, then T is the threshold for connectivity in $G_{n,p}$, and W is the minimal cost of a spanning tree. It is well known that $p = \log n/n$ [3] is the threshold function for connectivity, and $W \overset{\mathbb{P}}{\to} \zeta(3)$ [6]. Closely related is the case when $\mathcal{F}$ is the family of perfect matchings. Here the threshold is again $p = \log n/n$ [4] (in both cases the minimal obstruction is local and it is the existence of an isolated vertex) and $W \overset{\mathbb{P}}{\to} \zeta(2)$ [1]. Similarly for Hamilton cycles, the threshold is $p = (\log n + \log\log n)/n$ [9, 10, 12] and $W \overset{\mathbb{P}}{\to} 2.04\ldots$ [16].

The goal of this paper is to consider the case when $\mathcal{F}$ is the family of either H-factors or H-covers. An H-factor is a collection of vertex-disjoint subgraphs of K_n, each isomorphic to H, which collectively cover all n vertices. H-covers are defined similarly, but the condition that the subgraphs are vertex-disjoint is dropped. While the threshold version of the H-factor problem has received much attention (e.g. [8], [13]), the minimum-weight version has (as far as we are aware) not yet been studied. We prove the following, as well as a similar result for partial factors, and weaker results for covers. These can all be found in Theorems 2 and 3.

Theorem 1. *Assume H is a fixed graph with at least one cycle, $d^* > 1$ is its maximum 1-density as defined in Section 2.2, and $\mathrm{O}_{\mathbb{P}}$ is as defined in Section 2.1.*

Let the random variable $F_H = F_H(n)$ be the minimum weight of an H-factor on K_n (equipped with i.i.d. uniform $[0, 1]$ or exponential $\mathrm{Exp}(1)$ edge weights). Then there exists $M = \Theta(n^{1-1/d^*})$ such that $|F_H - M| = \mathrm{O}_{\mathbb{P}}(M^{3/4})$, as $n \to \infty$.

1.2. Proof strategy

Our proof follows a significantly different strategy compared to the study of the minimal perfect matching. The condition that the graph H contains a cycle is equivalent to $d^* > 1$. For such d^*, note that the minimum weight of an H-factor scales like a positive power of n. This scaling enables the following divide-and-conquer approach, which is the main novel contribution of this paper. It is crucial for the two parts of our proof to work: the upper bound and sharp concentration of F_H.

A large partial H-factor Q, covering some $n - k$ vertices, can be completed by adding the lowest-weight H-factor Q' on the remaining k vertices. Any such Q has a weight of order at least n^{1-1/d^*}, while Q' has a weight of order at most k^{1-1/d^*}. So if $k \ll n$, we can complete a large partial factor at a relatively small extra cost. Note that for graphs H with $d^* = 1$ (such as $H = K_2$) this does not work, since the minimum weight there instead scales like $F_H = \Theta(1)$.

However, the Q above might have been picked based on the edge weights (e.g. as the lowest-weight such partial factor) so that the weights of Q and Q' are not independent. To avoid this dependence, we employ a variant of a trick originally due to Walkup [15] in Section 4.3: split every edge into a green and red edge, and put independent random weights on them, following $\mathrm{Exp}(1 - t)$ and $\mathrm{Exp}(t)$ distributions respectively, for some small $t > 0$. This ensures independence, and the minimum of the two weights on such a pair of edges follows the distribution

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

Exp(1). We can now find a large partial factor on the green edges (at a slightly inflated cost), and complete the factor using a small number of red edges (at a highly inflated cost).

For the upper bound, we use an upper bound on the cost of a partial factor (due to Ruciński [13]), and recursively apply the red–green split to find larger partial factors on the remaining vertices. To show concentration, we study a dual problem: For some $L = L(n)$, how large is the largest partial H-factor with weight at most L? We use Talagrand's concentration inequality to show that this size is sharply concentrated around a large value, and then the red–green split trick to complete this large partial factor at small additional cost.

1.3. Structure of the paper

We begin with some definitions in Section 2. In Section 3 we state our main results (Theorems 2 and 3) and one conjecture, and compare this with previous work. We then provide proofs in Section 4 under the assumption that the edge weights follow an exponential distribution. In Sections 4.1 and 4.2 we prove the lower bounds of Theorems 2 and 3 respectively. Section 4.3 is devoted to the red–green split trick mentioned in Section 1.2. This trick is then used in Sections 4.4 and 4.5, where we prove the upper bound and sharp concentration, respectively, of Theorem 2. In Section 5 we show that the (asymptotic) distribution of the minimum cost of an H-cover or H-factor is unchanged if the edge-weight distribution is changed from exponential to uniform or some other distribution of pseudo-dimension 1. Finally, in Section 6 we discuss some pathological examples that illustrate why the equivalent of Theorem 2 cannot hold for covers.

2. Definitions and notation

2.1. Notation

We will use $\overset{\mathbb{P}}{\to}$ to denote convergence in probability, and write $X \overset{d}{=} Y$ if the random variables X and Y follow the same distribution. We will also use both standard and probabilistic big-O notation. For sequences X_n, Y_n of random variables, the notations $X_n = \mathrm{O}_{\mathbb{P}}(Y_n)$ and $Y_n = \Omega_{\mathbb{P}}(X_n)$ are equivalent, and mean that for any $\varepsilon > 0$, there exists a $C = C(\varepsilon)$ such that $\mathbb{P}(|X_n| > C|Y_n|) < \varepsilon$ for all sufficiently large n. Let $X_n = \Theta_{\mathbb{P}}(Y_n)$ denote that $X_n = \mathrm{O}_{\mathbb{P}}(Y_n)$ and $X_n = \Omega_{\mathbb{P}}(Y_n)$. Similarly, the notations $X_n = \mathrm{o}_{\mathbb{P}}(Y_n)$, $Y_n = \omega_{\mathbb{P}}(X_n)$ and $X_n \ll Y_n$ are equivalent, and mean that $X_n/Y_n \overset{\mathbb{P}}{\to} 0$. When both X_n and Y_n are deterministic, these definitions agree with those for standard big-O notation.

For any graph G, we will use $\mathcal{V}(G)$ and $\mathcal{E}(G)$ to refer to its vertex set and edge set respectively, while $v_G := |\mathcal{V}(G)|$ and $e_G := |\mathcal{E}(G)|$. Since we will also frequently need to refer to Euler's number $\mathrm{e} \approx 2.718$, we will use a different font to avoid confusion: e instead of e. We will also use $\exp(x)$ for the exponential function, and $\mathrm{Exp}(\lambda)$ for the exponential distribution.

2.2. Density and balanced graphs

For any graph H with at least two vertices we define its density as $d_H := e_H/(v_H - 1)$. This quantity is sometimes called the 1-density (referring to the -1 in the denominator), but we will refer to it simply as the density. We call H *strictly balanced* if $d_G < d_H$ for every subgraph $G \subset H$. Furthermore, let $d^* := \max\{d_G \colon G \subseteq H\}$, and let $H^* \subseteq H$ be a subgraph which achieves this maximal density d^*.

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

2.3. Covers and factors

Definition 1. An (α, H)-cover Q is a collection of subgraphs of the complete graph K_n, each of which is isomorphic to H, and such that at most αn vertices of K_n are not covered by any copy H' of H, that is,

$$\left|\bigcup_{H' \in Q} \mathcal{V}(H')\right| \geq (1-\alpha)n.$$

An (α, H)-factor is an (α, H)-cover such that $\mathcal{V}(H')$ and $\mathcal{V}(H'')$ are disjoint for any two $H', H'' \in Q$ with $H' \neq H''$. For $\alpha = 0$ we will refer to $(0, H)$-covers and $(0, H)$-factors simply as H-covers and H-factors respectively. For $\alpha > 0$, we will also refer to (α, H)-covers and (α, H)-factors as partial covers and factors.

Note that by definition an H-factor over n vertices exists if and only if v_H divides n, and that for all the valid H-factors, $|Q| = n/v_H$. From now on, we tacitly assume all results about factors to hold only when v_H divides n.

2.4. Edge-weight distribution

It turns out that the precise distribution of the (positive) edge weights does not matter – only its asymptotic behaviour near 0. That is, our results will hold under the following condition: if F is the common CDF of the edge weights, and $F(x) = \lambda x + o(x)$ for some $\lambda > 0$ as $x \to 0$. This property is sometimes referred to as F having *pseudo-dimension* 1. For distributions without atoms, this corresponds to a density function tending to λ near 0. Some examples of such distributions are Uniform $U(0, 1)$, Exponential, and (for certain values of their parameters) Gamma, Beta, and Chi-squared.

We will prove this later in Section 5, but for the sake of convenience we will until then assume that the edge weights follow an exponential distribution Exp(1).

2.5. Minimum-weight covers and factors

We will also (with minor abuse of notation) let $\mathcal{E}(Q) := \bigcup_{H' \in Q} \mathcal{E}(H')$ denote the (multi-)set of edges that occur in some copy of H. For factors this is a set, while for covers this is a multiset where the multiplicity of an edge counts how many copies of H it occurs in. For every set Q of subgraphs of K_n, we define its weight as

$$W_Q := \sum_{e \in \mathcal{E}(Q)} X_e = \sum_{H' \in Q} \sum_{e \in \mathcal{E}(H')} X_e.$$

Note that if an edge appears in two or more subgraphs $H' \in Q$ (copies of H), its weight is counted again every time. We will let C_H and F_H denote the minimum weight of a partial cover and factor, respectively:

$$C_H(k, n) := \min\{W_Q \colon Q \text{ is an}(k/n, H)\text{-cover}\},$$

$$F_H(k, n) := \min\{W_Q \colon Q \text{ is an}(k/n, H)\text{-factor}\}.$$

In other words, $C_H(k, n)$ (or $F_H(k, n)$) is the minimal weight of a partial cover (or partial factor) on K_n that leaves at most k vertices uncovered. We will also (for technical purposes) sometimes need to keep track of upper bounds on the most expensive edge a (partial) cover or factor uses.

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

We therefore define

$$C_H^\varepsilon(k, n) := \min\Big\{W_Q \colon Q \text{ is an } (k/n, H)\text{-cover and } \max_{e\in\mathcal{E}(Q)} X_e \le \varepsilon\Big\},$$

$$F_H^\varepsilon(k, n) := \min\Big\{W_Q \colon Q \text{ is an } (k/n, H)\text{-factor and } \max_{e\in\mathcal{E}(Q)} X_e \le \varepsilon\Big\}.$$

Said cover (resp. factor) might not exist, in which case we set $C_H^\varepsilon(k, n) = \infty$ (resp. $F_H^\varepsilon(k, n) = \infty$). This means that $\mathbb{E}C_H^\varepsilon$, $\mathbb{E}F_H^\varepsilon$ are not well-defined, and we will simply avoid them. As we will show in the following subsection, the results from [8] allow us to determine a range of values of ε such that $F_H^\varepsilon(k, n) < \infty$ (and thus $C_H^\varepsilon(k, n) < \infty$) with very high probability. Note also that C_H^ε, F_H^ε are non-increasing (random) functions of ε: as ε increases, fewer edges become ‘forbidden’, which can only decrease the minimal cost.

As every H-factor is also a valid H-cover, by definition $C_H(k, n) \le F_H(k, n)$ and $C_H^\varepsilon(k, n) \le F_H^\varepsilon(k, n)$. We will also let $C_H(n) := C_H(0, n)$ and $F_H(n) := F_H(0, n)$ denote the minimal weight of an H-cover and H-factor, respectively, and similarly for $C_H^\varepsilon(n)$, $F_H^\varepsilon(n)$.

3. Results and conjectures

Our main results are the following two theorems, where we establish bounds on F_H and C_H, as well as prove that F_H is a sharply concentrated random variable.

Theorem 2. *For any graph H with $d^* > 1$ and $\alpha \in [0, 1)$, there are constants $0 < a < b$ such that*

$$an^{1-1/d^*} \le F_H(\alpha n, n) \le bn^{1-1/d^*}$$

with probability $1 - n^{-\omega(1)}$. Furthermore, $F_H(\alpha n, n)$ is sharply concentrated around its median value M:

$$|F_H(\alpha n, n) - M| = \mathrm{O}_{\mathbb{P}}(M^{3/4}).$$

Theorem 1 is a special case of this theorem, with $\alpha = 0$. The two parts of the theorem are more precise versions of the statements

$$F_H(\alpha n, n) = \Theta_{\mathbb{P}}(n^{1-1/d^*}) \quad \text{and} \quad F_H(\alpha n, n)/\mathbb{E}[F_H(\alpha n, n)] \xrightarrow{\mathbb{P}} 1,$$

respectively. Note, however, that together they do not guarantee that the limit (in probability) of $F_H(\alpha n, n)/n^{1-1/d^*}$ exists.

Conjecture 1. *For any graph H with at least two vertices, there is a continuous decreasing function $f_H \colon [0, 1] \to \mathbb{R}$ such that*

$$F_H(\alpha n, n)/n^{1-1/d^*} \xrightarrow{\mathbb{P}} f_H(\alpha).$$

See Remark 1 for a discussion of what this function f_H might be.

As mentioned in the introduction, the corresponding limits do exist for several similar problems, including the travelling salesman and minimum-weight perfect matching (i.e. K_2-factor) problems. Theorems corresponding to Conjecture 1 for these two problems have been proved using a local graph limit method in, for example, [11], [16] and [17]. In broad terms, what these papers show is that the ‘local’ structure of the optimal solution is ‘locally’ determined. That is, it is possible to determine with high certainty whether a given edge participates in the

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

optimal solution by only inspecting a large but bounded neighbourhood of it. Unfortunately, this approach cannot be directly translated into our setting where H has density $d^* > 1$, but it would be interesting to see if Conjecture 1 holds for similar reasons.

Moving on to the related problem of covers, since $C_H \le F_H$ we automatically get an upper bound by Theorem 2 above. We also have the following lower bound.

Theorem 3. *Assume H has at least two vertices and let $\Delta := \max_{H' \subset H} (e_{H'}/v_{H'})$. Then, for any $\alpha \in [0, 1)$, we have $C_H(\alpha n, n) = \Omega_{\mathbb{P}}(n^{1-1/\max\{d_H, \Delta\}})$.*

Note the different exponents in the upper and lower bounds on C_H. They match if (for instance) H is *balanced*, so that $d_H = d^*$. In Section 6 we discuss examples where H is not balanced, only one of these bounds is sharp, and where C_H is not sharply concentrated. We might still conjecture that the H-cover equivalent of Theorem 2 or Conjecture 1 holds for balanced H.

Although we work with graphs throughout this paper, in principle our proof method should work for hypergraphs as well, under suitable conditions. However, some theorems we cite have only been proved in the graph setting and would need to be adapted to work for hypergraphs; see the discussion at the end of Section 3.1. Furthermore, we have not yet defined H-factors in a hypergraph setting. For graphs, an H-factor consists of a collection of copies of H that are vertex-disjoint and spans the vertex set of K_n. The disjointness condition can be generalized to hypergraphs by requiring that no two copies H', H'' of the r-uniform hypergraph H overlap in more than k vertices for some $k < r$ (with 'vertex-disjoint' corresponding to $k = 1$). Similarly, for the spanning condition we can require that the copies of H cover all the k'-sets of vertices in $K_n^{(r)}$ for some $k' < r$. Any pair (k, k') leads to a different generalization of the H-factor problem.

3.1. Related work

Before we move on to the proofs, we will briefly discuss some related work. First, we discuss a 2008 paper by Johansson, Kahn, and Vu [8] on the threshold version of the H-factor problem. We will use one of their theorems in our proof of the upper bound in Theorem 2. Second, we discuss a recent paper by Frankston, Kahn, Narayan, and Park [5], which provides a general and flexible framework for proving upper bounds on both threshold and minimum-weight problems.

In [8], the threshold function for the appearance of an H-factor for strictly balanced H was determined (up to a constant factor), as well as slightly less precise bounds on the threshold for general H.

Theorem 4. (Theorems 2.1 and 2.2 in [8].) *Assume H has at least two vertices.*

(i) *If H is strictly balanced, the threshold for the appearance of a H-factor in $G_{n,p}$ is $th_H := n^{-1/d_H}(\log n)^{1/e_H}$. That is,*

$$\mathbb{P}(G_{n,p} \text{ contains an } H\text{-factor}) = \begin{cases} n^{-\omega(1)}, & \text{if } p \ll th_H, \\ 1 - n^{-\omega(1)}, & \text{if } p \gg th_H. \end{cases}$$

(ii) *For general H the threshold is $n^{-1/d^*+o(1)}$. More precisely, for any $\varepsilon > 0$,*

$$\mathbb{P}(G_{n,p} \text{ contains an } H\text{-factor}) = \begin{cases} n^{-\omega(1)}, & \text{if } p \ll n^{-1/d^*}, \\ 1 - n^{-\omega(1)}, & \text{if } p \gg n^{-1/d^*+\varepsilon}. \end{cases}$$

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

This immediately implies the following upper bound on F_H, only a factor n^{ε} worse than the upper bound in Theorem 2.

Corollary 1. *For any* $\varepsilon > 0$ *and* $A \gg n^{-1/d^*+\varepsilon}$, $F_H^A(n) \leq n^{1-1/d^*+\varepsilon}$ *with probability* $1 - n^{-\omega(1)}$.

More recently, significant progress in the study of the general threshold type and minimum-weight type problems discussed in Section 1.1 was made in the 2019 breakthrough paper by Frankston, Kahn, Narayanan, and Park [5]. For a large class of families $\mathcal{F}$ which includes H-factors, they proved upper bounds on both the threshold for the appearance of an $F \in \mathcal{F}$ in $G_{n,p}$ as well as the minimum weight of an $F \in \mathcal{F}$ in a randomly weighted K_n. Among other applications, this leads to a much simpler proof of the upper bound in Theorem 4(i) than that in [8] – albeit for a slightly weaker upper bound, with a larger exponent for the logarithmic factor.

Using the results in [5], an alternative (and slightly shorter) proof of Proposition 3 can be obtained. Interestingly, the proof in [5] also uses a sprinkling method. They essentially sprinkle edges in multiple stages, and keep track of how many $F \in \mathcal{F}$ we keep making 'good progress' towards building.

If the reader wants to prove Theorem 2 for some generalization of H-factors to hypergraphs, using the results from [5] is the route we would recommend. The theorems in [5] are general enough to be applicable to both graphs and hypergraphs directly. In comparison, our proof of Theorem 2 depends on one theorem from the Johansson, Kahn, and Vu paper [8] and one by Ruciński [13]. Both of these are only proved explicitly for graphs, although the former paper mentions that its proofs remain essentially unchanged for hypergraphs.

4. Proofs

In this section we state and prove several propositions from which our main theorems follow: Theorem 2 follows from Propositions 1, 3, and 4, and Theorem 3 follows from Propositions 2 and 3.

4.1. Lower bound: H-factors

In this section we establish a lower bound on the minimum cost of H-factors, and then in Section 4.2 we do the same for H-covers. Although any lower bound on C_H-covers is also a lower bound on F_H, our lower bound for H-factors holds with probability $1 - 2^{-\Omega(n)}$, while the lower bound for H-covers is only shown to hold with probability $1 - \varepsilon$. For this reason we consider it worthwhile to include both.

Proposition 1. *Assume* $\alpha \in [0, 1)$ *is fixed (not depending on n). There exists a* $c > 0$ *such that the minimal cost of an* (α, H)*-factor is* $F_H(\alpha n, n) \geq cn^{1-1/d^*}$, *with probability* $1 - 2^{-\Omega(n)}$.

To prove this, we need the following simple bound (which will also be useful several times more throughout the paper).

Lemma 1. *If* $x > 0$, $X_1, X_2, \ldots X_k$ *are i.i.d.* Exp(1)*-distributed random variables and* $X := \sum_i X_i$, *then*

$$1 - x \leq \frac{\mathbb{P}(X \leq x)}{x^k/k!} \leq 1.$$

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

Proof. X follows a Gamma distribution with shape parameter k and scale parameter 1, with density function $t^{k-1}\mathrm{e}^{-t}/(k-1)!$. Since $\mathrm{e}^{-t} \geq 1 - x$ on the interval $t \in [0, x]$,

$$\mathbb{P}(X \leq x) \geq (1-x)\int_0^x \frac{t^{k-1}}{(k-1)!}\mathrm{d}t = (1-x)x^k/k!.$$

Similarly, using $\mathrm{e}^{-t} \leq 1$ gives $\mathbb{P}(X \leq x) \leq x^k/k!$. □

We can now prove Proposition 1.

Proof of Proposition 1. Assume without loss of generality that αn is an integer multiple of v_H. Let t be the smallest number of copies of H an (α, n)-factor can have. Since $(1-\alpha)n$ vertices of K_n are covered, each by a unique copy of H, $v_H t = (1-\alpha)n$.

We will first prove that $F_H(\alpha n, n) \geq cn^{1-1/d_H}$ with high probability by applying a first moment method to the following random variable. For any $L = L(n)$, let Y_L be the number of (α, H)-factors Q that have precisely t copies of H and that have a weight $W_Q \leq L$. Note that if $Y_L = 0$ then $F_H(\alpha n, n) > L$, because any (α, H)-factor that has more than t copies of H contains one with precisely t copies.

How many (α, H)-factors in K_n with precisely t copies of H are there (regardless of weight)? There are $\binom{n}{\alpha n} = 2^{\mathrm{O}(n)}$ ways to pick which αn vertices will not be covered, and then at most $(v_H t)!/t! = 2^{\mathrm{O}(n)}n^{(v_H-1)t}$ ways to construct an H-factor on the remaining $v_H t = (1-\alpha)n$ vertices. Rewriting the exponent of n as $v_H - 1 = e_H/d_H$, we can upper-bound the number of such factors by $(c_1 n^{1/d_H})^{e_H t}$ for some constant c_1. Now consider an (α, H)-factor Q with t copies of H. It consists of $e_H t$ edges, so by Lemma 1

$$\mathbb{P}(W_Q \leq L) \leq L^{e_H t}/(e_H t)! \leq (c_2 L/n)^{e_H t}, \tag{1}$$

for some $c_2 > 0$. We therefore obtain $\mathbb{E}Y_L \leq (c_1 c_2 L n^{-1+1/d_H})^{e_H t}$. Since c_1, c_2 are constants, we can ensure that the expression within brackets is at most $1/2$ by letting $L := cn^{1-1/d_H}$ for a sufficiently small $c = c(\alpha, H)$. Then $\mathbb{E}Y_L \leq 2^{-e_H t} = 2^{-\Omega(n)}$, whence $F_H(\alpha n, n) \geq cn^{1-1/d_H}$ with probability $2^{-\Omega(n)}$.

Now, if $d^* > d_H$ we can improve this lower bound. Let $H^* \subseteq H$ be a subgraph of the maximal density d^*. Consider Q as above: an (α, H)-factor which consists of t copies of H, with $v_H t = (1-\alpha)n$. This partial H-factor will contain a partial H^*-factor Q^* consisting of t copies of H^* and hence covering tv_{H^*} vertices: just remove the superfluous vertices and edges from each copy of H in Q. This Q^* is an (α^*, H^*)-factor, with α^* such that the number of vertices covered by Q^* is $(1-\alpha^*)n = v_{H^*}t = \Omega(n)$. By the previous argument (and since $\alpha^* \in [0, 1)$), $F_H(\alpha, n) \geq F_{H^*}(\alpha^*, n) \geq c(\alpha^*, H^*)n^{1-1/d^*}$ with probability $2^{-\Omega(n)}$. □

In the following remark we discuss some possible optimizations of this result.

Remark 1. With some more care taken, we can find minimal c_1, c_2 in the proof above. The number of H-factors is $n!/(\alpha n)!t!\mathrm{Aut}(H)^t$ (where $\mathrm{Aut}(H)$ is the number of automorphisms of H). Applying Stirling's approximation to this and to $(e_H t)!$ in (1) leads to $c_1 c_2 = (r/e_H)\mathrm{e}^{1-1/d_H} \cdot (r\alpha^{-\alpha r}\mathrm{Aut}(H))^{-1/e_H}$, where $r := n/t = v_H/(1-\alpha)$. It is a tempting conjecture that the resulting bound with $c^{-1} := c_1 c_2$ is tight, at least for strictly balanced H. In other words, $F_H(n)/n^{1-1/d_H}$ should converge in probability to this c.

4.2. Lower bound: *H*-covers

We now prove the less sharp lower bound on the minimal cost of an H-cover.

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

Proposition 2. *For any fixed $\alpha > 0$ there exists a $K > 0$ such that, for any $t > 0$ fixed or tending to 0 as $n \to \infty$, we have the following.*

(i) $C_H(\alpha, n) \geq tn^{1-1/d_H}$ *with probability at least* $1 - Kt^{e_H}$.

(ii) *Let* $\Delta := \max_{G\subseteq H}(e_G/v_G)$. *Then* $C_H(\alpha, n) \geq tn^{1-1/\Delta}$ *with probability at least* $1 - Kt^{e_G}$, *where G is the graph that attains the maximum* Δ.

Proof. For any $b > 0$, call a copy $H' \subset K_n$ of H *b-cheap* if $W_{\{H'\}} < b$, i.e. if the total weight of the edges in H' is at most b. Let N_b be the total number of b-cheap H'. We want to estimate $\mathbb{E}[N_b]$. For a given H', by Lemma 1 the probability that it is b-cheap is at most $b^{e_H}/e_H!$. Furthermore, there are less than n^{v_H} copies of H in K_n. Then, by Markov's inequality, for any $\lambda > 0$,

$$\mathbb{P}(N_b \geq \lambda) \leq \frac{\mathbb{E}[N_b]}{\lambda} \leq \frac{n^{v_H} b^{e_H}}{\lambda e_H!}. \tag{2}$$

Now suppose that there exists an (α, H)-cover Q with $W_Q \leq tn^{1-1/d_H}$. This Q consists of at least $(\alpha/v_H)n$ copies of H, since each copy of H covers at most v_H vertices not covered by another copy.

The number of $H' \in Q$ that are not b-cheap can be at most W_Q/b. In particular for $b := 2v_H n^{-1/d_H}/\alpha$, there can be at most $\alpha n/2v_H$ that are not b-cheap, or in other words at most half of the $H' \in Q$. Hence Q must contain at least $(\alpha/2v_H)n$ b-cheap copies H', which implies that $N_b \geq (\alpha/2v_H)n$. By (2),

$$\mathbb{P}\left(N_b \geq \frac{\alpha}{2v_H} n\right) \leq \frac{2vn^{v_H} b^{e_H}}{\alpha n e_H!} = \frac{(2vt/\alpha)^{e_H}}{\alpha e_H!}.$$

This immediately implies part (i). For part (ii), consider the subgraph G that attains the maximum $\Delta := \max_{G\subset H}(e_G/v_G)$. As noted earlier, any (α, H)-cover Q contains at least $(\alpha/v_H)n$ copies of H. Let $H_1, H_2, \ldots$ be an enumeration of them, and let $G_i \subset H_i$ be copies of G in each. Note that we might have $G_i = G_j$ for some $i \neq j$, as two distinct copies of H might overlap in a copy of G. We have

$$W_Q = \sum_i W_{H_i} \geq \sum_i W_{G_i} \geq \frac{\alpha}{v_H} n \min W_{G'}, \tag{3}$$

where the last minimum is taken over all copies $G' \subset K_n$ of G. Applying (3) with $\lambda = 1$, G instead of H and $b = t\,n^{-1/\Delta}$ for a small $t > 0$, we see that $\mathbb{P}(N_b \geq 1)$ is at most $t^{v_G}/e_G!$. In other words, with probability at least 1-$t^{v_G}/e_G!$ there is no b-cheap copy of H, from which part (ii) follows. □

Remark 2. For strictly balanced H, Proposition 2(i) can be sharpened by a second moment argument to hold with probability $1 - o(1)$ rather than $1 - Kt^{e_H}$.

4.3. Red–green split lemma

In this section we introduce the red–green split trick mentioned in Section 1.2. This lemma will be useful both to prove the upper bound on F_H, as well as to prove that it is sharply concentrated. It is also used in Section 5.

We state and prove Lemma 2 (as well as Proposition 3) not only for F_H but for F_H^A: the minimum weight of an H-factor using no edge of weight more than A, then considering F_H as the particular case where $A = \infty$. Keeping track of upper bounds on the most expensive edge in an H-factor makes statements and proofs slightly more involved. While such bounds will be of use in Theorem 7, they are not necessary for our main results, Theorems 2 and 3. We therefore suggest that readers who are only interested in the latter theorems simply ignore the superscript in F_H^A, and any inequalities involving A, A_k, B, and C.

Lemma 2. *Let $n > m > k \geq 0$ be integer multiples of v_H.*

(i) *For any $t \in (0, 1)$, the random variables $F_H^A(m, n)$, $F_H^B(k, m)$, and $F_H^C(k, n)$ (where $C \geq \max(A/t, B/(1-t))$) can be coupled such that surely*

$$F_H^C(k, n) \leq \frac{F_H^A(m, n)}{t} + \frac{F_H^B(k, m)}{1-t}.$$

(ii) *Let $a, b, A, B > 0$ and let $C \geq (a+b)\max(A/a, B/b)$. Then*

$$\mathbb{P}\big(F_H^C(k, n) > (a+b)^2\big) \leq \mathbb{P}\big(F_H^A(m, n) > a^2\big) + \mathbb{P}\big(F_H^B(k, m) > b^2\big).$$

Both of these inequalities also hold when $A = B = C = \infty$, i.e. with F_H instead of F_H^A, F_H^B, and F_H^C.

Remark 3. The lemma also holds for H-covers, and in that case the requirement that n, m and k are integer multiples of v_H is not necessary. The proof for H-covers is identical, *mutatis mutandis*. However, we will only prove and use the lemma for factors.

Proof. We will begin by proving part (i) of the lemma. Let G be the multigraph on $[n]$ given by connecting every pair of vertices by two parallel edges, one green and one red. Independently for all edges, assign to each green edge an Exp(t)-distributed random weight and to each red edge an Exp($1-t$)-distributed random weight. We will use the following properties of the exponential distribution:

(a) if $X \sim \text{Exp}(t)$ and $Y \sim \text{Exp}(1-t)$ are independent, then $\min(X, Y) \sim \text{Exp}(1)$,

(b) if $X \sim \text{Exp}(t)$, then $tX \sim \text{Exp}(1)$.

Let Z be the cost of the cheapest $(k/n, H)$-factor in G that uses no edge more expensive than C. (If no such factor exists, $Z = \infty$.) It will always use the cheaper of two parallel edges, so by property (a) we see that $Z \overset{d}{=} F_H^C(k, n)$. Our aim is now to construct a fairly cheap (but not necessarily optimal) such factor in G. First, we pick the cheapest green $(m/n, H)$-factor that uses no edge more expensive than A/t, and let Z_{green} be its cost. Note that by the rescaling property (b), $tZ_{\text{green}} \overset{d}{=} F_H^A(m, n)$.

We are left with a random set of m uncovered vertices. Crucially, this random set is independent of the weights on the red edges. Pick the cheapest red $(k/m, H)$-factor (i.e. a partial factor leaving at most k out of m vertices uncovered) on this set that uses no edge more expensive than $B/(1-t)$, and let its cost be Z_{red}. Again by (b), $(1-t)Z_{\text{red}} \overset{d}{=} F_H^B(k, m)$.

Combining the green copies of H from the first step with the red copies of H in the second step gives us a partial H-factor Q on G covering all but at most k vertices, i.e. a $(k/n, H)$-factor. No edge in Q costs more than $\max(A/t, B/(1-t)) \leq C$, whence $Z \leq W_Q = Z_{\text{green}} + Z_{\text{red}}$. Thus (by an appropriate coupling) the following inequality holds:

$$F_H^C(k, n) \leq \frac{F_H^A(m, n)}{t} + \frac{F_H^B(k, m)}{1-t}.$$

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

For part (ii), it follows from part (i) that if $F_H^A(m, n) \le a^2$ and $F_H^B(k, m) \le b^2$, then $F_H^C(k, n) \le a^2/t + b^2/(1-t)$. Minimizing over t gives that the right-hand side is $(a+b)^2$ for $t = a/(a+b)$, and for this t we obtain $C = \max(A/t, B/(1-t)) = (a+b)\max(A/a, B/b)$. Hence $F_H^C(k, n) \le (a+b)^2$, unless $F_H^A(m, n) > a^2$ or $F_H^B(k, m) > b^2$. Using the union bound on these two events gives the inequality in part (ii).

For the case $A = B = C = \infty$ the proof is nearly identical, except that we do not need to keep track of the cost of the most expensive edges. □

4.4. Upper bound

In this section we prove the following upper bound on the total cost of an H-factor, both unconstrained and limited to using only edges of weight at most A.

Proposition 3. *For any fixed graph H with $d^* > 1$ and any $\varepsilon > 0$, there exists a $c > 0$ such that if $A \ge n^{-1/d^*+\varepsilon}$, then $F_H^A(n) \le cn^{1-1/d^*}$ with probability at least $1 - n^{-\omega(1)}$. In particular, this holds for $A = \infty$.*

To prove this proposition, we will need the following theorem from [7, Theorem 4.9], originally due to Ruciński [13].

Theorem 5. *For any $\alpha \in (0, 1)$ there exist constants $c, t > 0$ such that $G_{n,p}$ with $p = cn^{-1/d^*}$ contains an (α, H)-factor with probability at least $1 - 2^{-tn}$.*

In [7], the existence of such a partial factor is only stated to hold with probability $1 - o(1)$, but in the proof the probability is shown to be $1 - 2^{-\Omega(n)}$.

Proof of Proposition 3. The proof strategy is essentially as follows. For some small fixed number $\alpha > 0$, we will find a cheap H-factor on n vertices by iteratively using the red–green split trick from Lemma 2. This will give a cheap (α, H)-factor on n_i vertices (starting with $n_0 := n$), then a cheap (α, H)-factor on the remaining n_{i+1} vertices, and so on, for a total of k steps. On the remaining n_k vertices, it suffices to find a not too expensive H-factor.

More precisely, pick α so that $\alpha^{1-1/d^*} = \frac{1}{4}$ (and hence $\alpha < \frac{1}{4}$). Let $n_0 := n$ and let n_i be the largest multiple of v_H such that $n_i \le \alpha n_{i-1}$. Also, for some small fixed $\delta > 0$ to be determined later, let k be an integer such that $\alpha^k \le n^{-\delta} \le \alpha^{k-1}$. For this choice of n_i and k, we have $\alpha^{i+1}n \le n_i \le \alpha^i n$ and $\alpha n^{1-\delta} \le n_k \le n^{1-\delta}$. Also, $4^k < n^\delta$.

Applying part (i) of Lemma 2, with $t = 1/2$ and $A_i := 2^i A$, repeatedly to $F_H^{A_i}(n_i)$ for $i = 0, 1, \ldots, k-1$, we find that there exists a coupling such that

$$\begin{aligned} F_H^{A_0}(n_0) &\le 2F_H^{A_1}(n_1, n_0) + 2F_H^{A_1}(n_1) \\ &\le 2F_H^{A_1}(n_1, n_0) + 4F_H^{A_2}(n_2, n_1) + 4F_H^{A_2}(n_2) \\ &\le 2F_H^{A_1}(n_1, n_0) + 4F_H^{A_2}(n_2, n_1) + 8F_H^{A_3}(n_3, n_2) + 8F_H^{A_3}(n_3) \\ &\ \ldots \\ &\le \underbrace{\sum_{i=0}^{k-1} 2^{i+1} F_H^{A_{i+1}}(n_{i+1}, n_i)}_{\text{(4a)}} + \underbrace{2^k F_H^{A_k}(n_k)}_{\text{(4b)}}. \end{aligned} \tag{4}$$

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

First we will bound the sum (4a). By Theorem 5, there exist constants c, t (depending only on α, H) such that if $A_{i+1} \geq cn_i^{-1/d^*}$ then $F_H^{A_{i+1}}(n_{i+1}, n_i) \leq cn_i^{1-1/d^*}$ with probability at least $1-2^{-tn_i} \geq 1-2^{-tn_k}$. To check whether this lower bound on A_{i+1} holds, note that since $A_i \geq A$ and $n_i \geq n_k$, it suffices to show that $An_k^{1/d^*} \geq c$. Using $n_k \geq \alpha^{k+1}n$ and $\alpha^k \geq \alpha n^{-\delta}$, we get

$$An_k^{1/d^*} = n^{-1/d^*+\varepsilon} n_k^{1/d^*} \geq n^{\varepsilon}(\alpha^{k+1})^{1/d^*} \geq n^{\varepsilon}(\alpha^2 n^{-\delta})^{1/d^*} \gg n^{\varepsilon/2}, \tag{5}$$

where the last inequality holds by picking δ sufficiently small. Hence the conditions of Theorem 5 are met, and it then follows (by a union bound) that with probability at least $1-k2^{-tn_k} = 1-n^{-\omega(1)}$, we have (4a) $\leq 2c\sum_{i=0}^{k-1} 2^i n_i^{1-1/d^*}$. Since $n_i \leq \alpha^i n$ and $\alpha^{1-1/d^*} = \frac{1}{4}$ (by the choice of α), we can bound the terms in this sum by

$$2^i n_i^{1-1/d^*} \leq (2\alpha^{1-1/d^*})^i \cdot n^{1-1/d^*} \leq 2^{-i} n^{1-1/d^*}. \tag{6}$$

Hence (4a) is at most $4cn^{1-1/d^*}$ with high probability. For the term (4b) of equation (4), the slightly rougher bound in Corollary 1 suffices: for any $\delta' > 0$, if $A_k \gg n_k^{1-1/d^*+\delta'}$ then $F_H^{A_k}(n_k) \leq n_k^{1-1/d^*+\delta'}$ with probability $n_k^{-\omega(1)}$. But by (5), $A_k \geq A \gg n_k^{-1/d^*+\varepsilon/2}$, so the condition on A_k is met if we pick $\delta' < \varepsilon/2$. Then

$$\text{(4b)} = 2^k F_H^{A_k}(n_k) \leq 2^k n_k^{1-1/d^*+\delta'} \leq 2^{-k} n^{1-1/d^*+\delta'},$$

where the last inequality uses inequality (6) and $n_k^{\delta'} \leq n^{\delta'}$. From the choice of k, $2^{-k} \leq n^{-\delta/|\log_2 \alpha|}$, and we can therefore ensure that the right-hand side above is $\mathrm{o}(n^{1-1/d^*})$ by picking δ' sufficiently small ($\delta' < \delta/|\log_2 \alpha|$). It follows that (4a) + (4b) $\leq (4c + \mathrm{o}(1))n^{1-1/d^*}$ with probability $1-n^{-\omega(1)}$. □

Remark 4. Strictly speaking, use of the theorem from [8] is not necessary here. However, it allows the recursion in (4) to end after fewer steps, which helps keep the error probability in Proposition 3 low, as well as the upper bound A on the most expensive edge used.

4.5. Concentration

We will now move on to show that F_H is sharply concentrated.

Proposition 4. *For any graph H with $d_H > 1$, $\varepsilon > 0$ and $\alpha \in [0, 1)$, there exists a $c > 0$ such that if we let $M = M(\alpha, n, H)$ denote the median of $F_H(\alpha n, n)$, then for all sufficiently large n and with probability at least $1-\varepsilon$,*

$$|F_H(\alpha n, n) - M| < cM^{3/4}.$$

We will consider a dual problem: How large is the largest partial factor that costs at most L, for some $L = L(n)$? More precisely, let the random variable $Z_H = Z_H(n, L)$ be defined by

$$Z_H := \max\{\alpha n \colon \text{there exists a } (1-\alpha, H)\text{-factor } Q \text{ with } W_Q \leq L\}.$$

In other words, Z_H is the largest number of vertices that a partial factor costing at most L can cover. Note that $Z_H(n, L) \geq n-m$ if and only if $F_H(m, n) \leq L$. Our first step is to apply Talagrand's concentration inequality to Z_H. To do so we need the definitions of f-certifiable and Lipschitz random variables.

Definition 2. (*f-certifiable random variable.*) Let $X\colon \Omega^n \to \mathbb{R}$ be a random variable. For a function f on $\mathbb{R}$ we say that X is f-certifiable if, for any $\omega \in \Omega^n$ with $X(\omega) \geq s$, there is a set $I \subseteq [n]$ of at most $f(s)$ coordinates such that $X(\omega') \geq s$ for all ω' which agree with ω on I. (That is, ${\omega'}_i = \omega_i$ for all $i \in I$.)

Definition 3. (*Lipschitz random variable.*) Let X be as above. We say that X is K-Lipschitz if, for every ω, ω' with $\omega_i = {\omega_i}'$ for all but one i, $|X(\omega) - X(\omega')| \leq K$.

We can now state Talagrand's inequality. While it was first established in [14], we will use the following more 'user-friendly' version from [2].

Theorem 6. (Talagrand's concentration inequality.) *Assume* Ω *is a probability space. If* X *is a* K*-Lipschitz,* f*-certifiable random variable* $X\colon \Omega^n \to \mathbb{R}$*, where* Ω^n *is equipped with the product measure, then for any* $b, t \geq 0$*,*

$$\mathbb{P}(X \leq b)\mathbb{P}(X \geq b + tK\sqrt{f(b)}) \leq \exp(-t^2/4).$$

The following lemma finds the appropriate values of f and K so that we can apply this inequality to the random variable Z_H.

Lemma 3. Z_H *is* v_H*-Lipschitz and* f*-certifiable with* $f(s) = e_H\lceil s/v_H\rceil \leq (e_H/v_H)n$*.*

Proof. To show that Z_H is f-certifiable, pick an integer $s \in [n]$ and a tuple of edge weights $\omega \in \Omega^{\binom{n}{2}}$ such that $Z_H(\omega) \geq s$. Then there exists a partial H-factor Q with $W_Q(\omega) \leq L$ and which covers at least s vertices. Assume without loss of generality that Q is one of the smallest such partial H-factors. It then contains $\lceil s/v_H\rceil$ copies of H and $f(s) := e_H\lceil s/v_H\rceil$ edges. For any ω' which agrees with ω on the $f(s)$ edges of Q, $W_Q(\omega') = W_Q(\omega) \leq L$. Hence $Z_H(\omega') \geq s$. (It might be that $Z_H(\omega) \neq Z_H(\omega')$; here we only care whether they are $\geq s$.)

To show the Lipschitz condition, pick an edge e and condition on all other edge weights. Consider Z_H as a function of just $x = X_e$. Note first that $Z_H(x)$ is a non-increasing function, i.e. $Z_H(x) \leq Z_H(0)$ for any $x \geq 0$. Let Q be a partial H-factor achieving the maximum size $Z_H(0)$. That is, Q covers $Z_H(0)$ vertices and has weight $W_Q = W_Q(x)$ such that $W_Q(0) \leq L$. Is $e \in \mathcal{E}(Q)$?

(a) If $e \in \mathcal{E}(Q)$, let H_e be the copy of H in Q which contains e. Then $Q - H_e$ is a partial H-factor with weight at most $W_{Q-H_e}(x) < W_Q(0) \leq L$ (for any x), and it covers $Z_H(0) - v_H$ vertices. Hence $Z_H(x) \geq Z_H(0) - v_H$.

(b) If $e \notin \mathcal{E}(Q)$, then $W_Q(x)$ is a constant function and $W_Q(x) = W_Q(0) \leq L$. Hence $Z_H(x) \geq Z_H(0)$.

In either case, $Z_H(0) - v_H \leq Z_H(x) \leq Z_H(0)$. Thus Z_H is v_H-Lipschitz. □

Remark 5. This is where our proof would fail for the corresponding cover problem. For covers, some edges might belong to a large number of copies of H, leading to a large Lipschitz constant. This is the case in our example in Section 6.

Before proceeding with the proof of Proposition 4, we will need two small lemmas.

Lemma 4. *If* $k < m < n$*,*

$$F_H(m, n) \leq \frac{n-m}{n-k} F_H(k, n).$$

Proof. $F_H(m, n)$ is the lowest cost of a partial H-factor covering at least $n - m$ vertices of K_n. We can construct a cheap such partial factor in two steps. First, let Q be the optimal $(k/n, H)$-factor (which consists of $(n-k)/v_H$ copies of H), i.e. the factor such that $W_Q = F_H(k, n)$.

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

Next, let Q' be the partial factor obtained by removing all but the $(n-m)/v_H$ cheapest copies of H in Q, leaving a $(m/n, H)$-factor. Then Q' contains a fraction $(n-m)/(n-k)$ of the copies of H in Q. Hence it costs

$$W_{Q'} \leq \frac{n-m}{n-k} W_Q.$$

□

Lemma 5. *For any m and n, the random variable $F_H(m, n)$ follows a continuous distribution (i.e. it has no atoms).*

Proof. For any partial factor Q and $t \geq 0$, $\mathbb{P}(W_Q = t) = 0$. And since there are only finitely many such Q, $\mathbb{P}(F_H(m, n) = t) \leq \mathbb{P}(\exists Q\colon W_Q = t) = 0$. □

We can now finally prove that the cost of a (partial) H-factor concentrates around its median.

Proof of Proposition 4. Let m be the largest multiple of v_H such that $m \leq \alpha n$. By Lemma 5, $F_H(m, n)$ is a continuous random variable, whence we can find L such that $\mathbb{P}(F_H(m, n) \leq L) = \varepsilon$. Using the upper bound (Proposition 3) and lower bound (Proposition 1) on F_H, we see that in order for $\mathbb{P}(F_H(m, n) \leq L) = \varepsilon$ to hold, we must have $L = \Theta(n^{1-1/d^*})$. (For the lower bound, the condition $\alpha < 1$ is used.) We will now apply the Talagrand inequality to the v_H-Lipschitz, $e_H n/v_H$-certifiable random variable $Z_H(L, n)$. Choose $t > 0$ such that $\exp(-t^2/4) = \varepsilon^2$ and let $b := n - m - k$, where $k := \lceil t\sqrt{e_H v_H n}\rceil$. Then

$$\mathbb{P}(Z_H \leq n - m - k) \cdot \mathbb{P}(Z_H \geq n - m) \leq \varepsilon^2. \tag{7}$$

By the choice of L and recalling that $Z_H(L, n)$ is the largest $n - m$ such that $F_H(m, n) \leq L$, the second probability in the left-hand side of (7) is ε. Hence the first probability is

$$\mathbb{P}(F_H(m + k, n) \geq L) = \mathbb{P}(Z_H \leq n - m - k) \leq \varepsilon. \tag{8}$$

So with probability at least $1 - \varepsilon$, there is a partial H-factor of cost at most L and that leaves at most $m + k$ vertices uncovered. What is the cost of a partial factor covering k out of the remaining $m + k$ vertices? By Lemma 4 and Proposition 3,

$$F_H(m, m + k) \leq \frac{k}{m + k} F_H(m + k) \leq ck(m + k)^{-1/d^*} \leq ck^{1-1/d^*} =: \ell, \tag{9}$$

with probability $1 - k^{-\omega(1)} \geq 1 - \varepsilon$ for some constant c (since $k \gg 1$). Using part (ii) of Lemma 2,

$$\mathbb{P}\big(F_H(m, n) > \big(\sqrt{L} + \sqrt{\ell}\big)^2\big) \leq \mathbb{P}(F_H(m + k, n) > L) \tag{10}$$

$$+\mathbb{P}(F_H(m, m + k) > \ell). \tag{11}$$

Note that $\ell = \Theta(\sqrt{L})$, since $L = \Theta(n^{1-1/d^*})$, $\ell = \Theta(k^{1-1/d^*})$, and $k = \Theta(\sqrt{n})$. Thus $\big(\sqrt{L} + \sqrt{\ell}\big)^2 \leq L + bL^{3/4}$ for some constant b. The right-hand side of (10) is at most ε by (8), and (11) is at most ε by (9). Thus (10) and (11) give

$$\mathbb{P}(F_H(m, n) > L + bL^{3/4}) \leq 2\varepsilon,$$

and by the choice of L, $\mathbb{P}(F_H(m, n) < L) = \varepsilon$. Assuming without loss of generality that $\varepsilon < 1/4$, this also implies that the median M of $F_H(m, n)$ lies in the interval $[L, L + bL^{3/4}]$, and in particular $M = \Theta(L)$. Hence $|F_H(m, n) - M| = \mathrm{O}_{\mathbb{P}}(M^{3/4})$. □

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

5. Other edge-weight distributions

As mentioned in Section 2.4, the exact edge-weight distribution does not matter – only its asymptotic behaviour near 0. Here we prove this fact.

Theorem 7. *Assume K_n is equipped with positive i.i.d. edge weights Z_e with some common CDF $\tilde{G}$ satisfying $\lim_{x\to 0}\tilde{G}(x)/x=1$ (i.e. $\tilde{G}(x)=x+\mathrm{o}(x)$). Let $\tilde{F}_H(m,n)$ be the minimum-weight $(m/n,H)$-factor with respect to these weights (and similarly for $\tilde{F}_H(n)$, $\tilde{C}_H(n)$, $\tilde{C}_H(m,n)$). Then these edge weights can be coupled to i.i.d.* Exp(1) *edge weights in such a way that for any $m=m(n)$ with $\lim_{n\to\infty} m/n<1$, $\tilde{F}_H(m,n)/F_H(m,n)\xrightarrow{\mathbb{P}}1$ and $\tilde{C}_H(m,n)/C_H(m,n)\xrightarrow{\mathbb{P}}1$.*

Remark 6. If instead $\tilde{G}(x)=\lambda x+\mathrm{o}(x)$ for some $\lambda>0$, we can replace the edge weights Z_e with weights λZ_e. This changes the optimal cost by a factor λ, and since $\mathbb{P}(\lambda Z_e\le x)=\tilde{G}(x/\lambda)=x+\mathrm{o}(x)$, we have $\tilde{F}_H(m,n)/F_H(m,n)\xrightarrow{\mathbb{P}}\lambda$.

Proof of Theorem 7. We will prove this for $m=0$ and F_H; the proof is essentially identical for $m>0$ or C_H, but the notation becomes messier.

Let $X_e\sim\mathrm{Exp}(1)$, and let $G(x)=1-\mathrm{e}^{-x}$ be the CDF of this distribution. Then $G(X_e)$ is uniformly distributed in the interval [0,1], and we can therefore couple it to Z_e by letting $Z_e:=\tilde{G}^{-1}(G(X_e))$.

Pick a small fixed $\varepsilon>0$. Since both $\tilde{G}(x)$ and $G(x)$ are asymptotically $x+\mathrm{o}(x)$ as $x\to 0$, we can find a $C=C(\varepsilon)>\varepsilon$ such that for any $x\in[0,3C]$, both $G(x)\le\tilde{G}((1+\varepsilon)x)$ and $\tilde{G}(x)\le G((1+\varepsilon)x)$ holds. So whenever either X_e or Z_e is at most $2C$, the other is at most $2C(1+\varepsilon)<3C$, and hence

$$(1-\varepsilon)X_e\le Z_e\le(1+\varepsilon)X_e.$$

We will prove that the following chain of inequalities holds with high probability:

$$1-4\varepsilon\le\frac{\tilde{F}_H(n)}{F_H^{2C}(n)}\le 1+\varepsilon. \tag{12}$$

For the second inequality of (12), consider $F_H^{2C}(n)$. This is finite if and only if there exists an H-factor Q that uses no edge of weight more than $2C$. We know from Corollary 1 that such a Q exists with probability $1-n^{-\omega(1)}$, so it is enough to prove that (12) holds with high probability under the assumption that there is such a Q, or equivalently that $F_H^{2C}(n)<\infty$. Pick Q as the cheapest such H-factor, so that $W_Q=F_H^{2C}(n)$. An edge e in Q has edge weight $X_e\le 2C$ by construction, whence $Z_e\le(1+\varepsilon)X_e$, and

$$\tilde{F}_H(n)\le\sum_{e\in\mathcal{E}(Q)}Z_e\le\sum_{e\in\mathcal{E}(Q)}(1+\varepsilon)X_e=(1+\varepsilon)F_H^{2C}(n).$$

For the first inequality of (12), let Q instead be the optimal H-factor with respect to the edge weights Z_e, i.e. $\sum_{e\in\mathcal{E}(Q)}Z_e=\tilde{F}_H(n)$. We will use it to construct a cheap H-factor (with respect to X_e). Call a copy $H'\in Q$ 'bad' if it contains at least one edge e with cost $Z_e\ge C$. The total number of such edges in Q is at most $\tilde{F}_H(n)/C$, so there are at most this many bad copies, and at most $v_H\tilde{F}_H(n)/C$ vertices are covered by a bad copy.

Using the second inequality of (12) together with Proposition 3, we see that $\tilde{F}_H(n)\le(1+\varepsilon)F_H^{2C}(n)\le Kn^{1-1/d^*}$ with high probability for some constant K, and then at most

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

$k := v_H \cdot \lfloor Kn^{1-1/d^*}/C \rfloor \ll n$ vertices are covered by a bad copy. Removing every bad copy then gives a $(k/n, H)$-factor using no edge more expensive than C (with high probability). Hence $\tilde{F}_H(n) \geq (1-\varepsilon)F_H^C(k, n)$. By Lemma 2,

$$F_H^{2C}(n) \leq \frac{F_H^C(k, n)}{1-\varepsilon} + \frac{F_H^{C\varepsilon}(k)}{\varepsilon}. \tag{13}$$

Pick some a_n, b_n with $k^{1-1/d^*} \ll a_n \ll b_n \ll n^{1-1/d^*}$. The second term on the right-hand side of (13) is by Theorem 5 at most a_n with high probability. On the other hand, by Proposition 2 the first term is at least b_n with high probability. Hence $F_H^{C\varepsilon}(k) \leq a_n \ll b_n \leq F_H^C(k, n)$ with high probability, and

$$F_H^{2C}(n) \leq \frac{1+\varepsilon}{1-\varepsilon} F_H^C(k, n) \leq \frac{1+\varepsilon}{(1-\varepsilon)^2} \tilde{F}_H(n),$$

with high probability, which gives the first inequality of (12).

But since (12) is valid for any CDF $\tilde{G}$ with $\tilde{G}(x) = x + \mathrm{o}(x)$ as $x \to 0$, in particular it is valid for G, and thus $1 - 4\varepsilon \leq \tilde{F}_H(n)/F_H^{2C}(n) \leq 1 + \varepsilon$ as well. It follows that $1 - 6\varepsilon \leq \tilde{F}_H(n)/F_H(n) \leq 1 + 6\varepsilon$ with high probability. Since ε was arbitrary, $\tilde{F}_H(n)/F_H(n) \to 1$ in probability. □

6. Examples of unbalanced cover

We will conclude with examples of cover problems where the upper and lower bounds on C_H do not match, and where C_H is not sharply concentrated. Recall that the lower bound on C_H was of order $n^{1-1/\max(d_H, \Delta)}$, while for factors it was n^{1-1/d^*} (with $d_H, \Delta \leq d^*$).

Why are the lower bounds for factors and covers different? If $d_H < d^*$, then H has a denser subgraph H^*, and the minimal H-cover might have many copies of H overlapping in the same copy of H^*. In an H-factor there are at least $\Omega(n)$ vertices lying in some copy of H^* (because t disjoint copies of H contain at least t disjoint copies of H^*), while in an H-cover only one copy of H^* is guaranteed.

For the sake of simplicity, consider instead the threshold for the appearance of an H-cover in $G_{n,p}$. The threshold for the existence of a collection of copies of H^* that cover at least $\Omega(n)$ vertices is $p = n^{-\beta}$ with $\beta = 1/d^* = \min_{H' \subseteq H} (v_H - 1)/e_H$. But the threshold for the appearance of at least one copy of H^* is lower, with $\beta = \min_{H' \subseteq H} v_H/e_H$.

For example, consider $H = K_4 + K_2$ (disjoint union of the complete graph on four vertices and an edge). Here the 1-density of H is $d_H = 1.4$, while the maximum 1-density of a subgraph is $d^* = 2$ (the K_4). The maximum 0-density is 1.5 (again, the K_4). So $\max(d_H, \Delta) = 1.5$, and Proposition 2 gives the lower bound $C_H(n) = \Omega_{\mathbb{P}}(n^{1/3})$, while Proposition 3 gives the upper bound $C_H(n) = \mathrm{O}_{\mathbb{P}}(n^{1/2})$.

For this H, the lower bound is tight: the cheapest H-factor will typically be the cheapest K_4 together with the cheapest cover of the remaining $n - 4$ vertices by edges. Define the random variable Z by the lowest weight of a copy of K_4 in K_n. With a first and second moment method counting the number of K_4 cheaper than $cn^{-2/3}$, one can show that $Z = \Theta_{\mathbb{P}}(n^{-2/3})$, but $\mathbb{P}(Z \leq cn^{-2/3})$ is bounded away from both 0 and 1 for any c. In other words, Z is not sharply concentrated.

A red–green split argument as in Lemma 2 with $t = 1/2$ leads to a coupling such that $C_H \leq (n-4)Z + 2C_{K_2}$ (where $C_H = C_H(n)$ and $C_{K_2} = C_{K_2}(n-4)$), because the smallest number of copies of H that can overlap in the same copy of K_4 while also covering all n vertices is $(n-4)/2$. For C_{K_2}, note that $C_{K_2} \leq F_{K_2} = \mathrm{O}_{\mathbb{P}}(1)$ (by [1]). On the other hand, $C_H(n) \geq nZ/6$,

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press

because any cover contains at least $n/6$ copies of H that each must contain a K_4, and each such copy has weight at least Z. Together this gives us that $\frac{1}{6} \le C_H(n)/nZ \le 1 + \mathrm{o}_{\mathbb{P}}(1)$ with high probability. Hence $C_H(n) = \Theta_{\mathbb{P}}(n^{1/3})$, but since Z is not sharply concentrated, neither is C_H.

One might guess that this pathological behaviour is due to H being disconnected, but it occurs even for some connected graphs. For instance, if H is a (5,2)-lollipop graph: a complete graph K_5, with a path P_2 away from one of the vertices of the clique. There are 7 vertices and 12 edges, so $d_H = 2$. Since the densest subgraph is the K_5, $\Delta = 2$ and $d^* = 5/2$. From Theorem 3, the asymptotics of C_H is then between $n^{0.5}$ and $n^{0.6}$. Here a near-optimal H-cover can be found that is a single K_5 together with a large collection of paths from this clique. Consider $G_{n,p}$ with $p = n^{-1/2+\varepsilon}$ for some small $\varepsilon > 0$. With high probability, this graph contains a K_5 and has diameter 2. Hence $C_H = \mathrm{O}_{\mathbb{P}}(np) = \mathrm{O}_{\mathbb{P}}(n^{1/2+\varepsilon})$, which is arbitrarily close to the lower bound from Theorem 3.

Funding information

There are no funding bodies to thank relating to this creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Aldous, D. J. (2001). The $\zeta(2)$ limit in the random assignment problem. *Random Structures Algorithms* **18**, 381–418.

[2] Alon, N. and Spencer, J. H. (2016). *The Probabilistic Method*. John Wiley.

[3] Erdős, P. and Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen* **6**, 290–297.

[4] Erdős, P. and Rényi, A. (1966). On the existence of a factor of degree one of a connected random graph. *Acta Math. Acad. Sci. Hungar.* **17**, 359–368.

[5] Frankston, K., Kahn, J., Narayanan, B. and Park, J. (2021). Thresholds versus fractional expectation-thresholds. *Ann. of Math.* **194**, 475–495.

[6] Frieze, A. M. (1985). On the value of a random minimum spanning tree problem. *Discrete Appl. Math.* **10**, 47–56.

[7] Janson, S., Łuczak, T. and Ruciński, A. (2000). *Random Graphs* (Wiley-Interscience Series in Discrete Mathematics and Optimization). Wiley-Interscience, New York.

[8] Johansson, A., Kahn, J. and Vu, V. (2008). Factors in random graphs. *Random Structures Algorithms* **33**, 1–28.

[9] Komlós, J. and Szemerédi, E. (1983). Limit distribution for the existence of Hamiltonian cycles in a random graph. *Discrete Math.* **43**, 55–63.

[10] Koršunov, A. (1976). Solution of a problem of P. Erdős and A. Rényi on Hamiltonian cycles in undirected graphs. *Dokl. Akad. Nauk SSSR* **228**, 529–532.

[11] Larsson, J. (2021). The minimum perfect matching in pseudo-dimension $0 < q < 1$. *Combinatorics Prob. Comput.* **30**, 374–397.

[12] Pósa, L. (1976). Hamiltonian circuits in random graphs. *Discrete Math.* **14**, 359–364.

[13] Ruciński, A. (1992). Matching and covering the vertices of a random graph by copies of a given graph. *Discrete Math.* **105**, 185–197.

[14] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques* **81**, 73–205.

[15] Walkup, D. W. (1979). On the expected value of a random assignment problem. *SIAM J. Comput.* **8**, 440–442.

[16] Wästlund, J. (2010). The mean field traveling salesman and related problems. *Acta Math.* **204**, 91–150.

[17] Wästlund, J. (2012). Replica symmetry of the minimum matching. *Ann. of Math.* **175**, 1061–1091.

https://doi.org/10.1017/jpr.2024.72 Published online by Cambridge University Press