

On the exact maximum induced density of almost all graphs and their inducibility

Raphael Yuster *

Abstract

Let H be a graph on h vertices. The number of induced copies of H in a graph G is denoted by $i_H(G)$. Let $i_H(n)$ denote the maximum of $i_H(G)$ taken over all graphs G with n vertices.

Let $f(n, h) = \prod_{i=1}^h a_i$ where $\sum_{i=1}^h a_i = n$ and the a_i are as equal as possible. Let $g(n, h) = f(n, h) + \sum_{i=1}^h g(a_i, h)$. It is proved that for almost all graphs H on h vertices it holds that $i_H(n) = g(n, h)$ for all $n \leq 2^{\sqrt{h}}$. More precisely, we define an explicit graph property \mathcal{P}_h which, when satisfied by H , guarantees that $i_H(n) = g(n, h)$ for all $n \leq 2^{\sqrt{h}}$. It is proved, in particular, that a random graph on h vertices satisfies \mathcal{P}_h with probability $1 - o_h(1)$. Furthermore, all extremal n -vertex graphs yielding $i_H(n)$ in the aforementioned range are determined.

We also prove a stability result. For $H \in \mathcal{P}_h$ and a graph G with $n \leq 2^{\sqrt{h}}$ vertices satisfying $i_H(G) \geq f(n, h)$, it must be that G is obtained from a balanced blowup of H by adding some edges inside the blowup parts.

The *inducibility* of H is $i_H = \lim_{n \rightarrow \infty} i_H(n) / \binom{n}{h}$. It is known that $i_H \geq h! / (h^h - h)$ for all graphs H and that a random graph H satisfies almost surely that $i_H \leq h^{3 \log h} h! / (h^h - h)$. We improve upon this upper bound almost matching the lower bound. It is shown that a graph H which satisfies \mathcal{P}_h has $i_H = (1 + O(h^{-h^{1/3}})) h! / (h^h - h)$.

1 Introduction

We consider the basic problem of maximizing the number of induced copies of a given graph H in a graph with n vertices. This problem has been studied already for four decades by various researchers such as in [2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 14, 17] and its answer (even asymptotically) seems difficult already for some very small graphs. Our main result gives an exact answer to this problem for almost all graphs H , and for all n that are not too huge.

Let $i_H(G)$ denote the number of induced copies of H in a graph G . The related extremal graph-theoretic parameter of interest is therefore the maximum of $i_H(G)$ taken over all graphs G with n vertices, denoted by $i_H(n)$. It is a simple exercise to prove that the maximum induced density of H in graphs with n vertices, namely $i_H(n) / \binom{n}{h}$ tends to a limit, which is denoted by $i_H = \lim_{n \rightarrow \infty} i_H(n) / \binom{n}{h}$ and is called the *inducibility* of H .

*Department of Mathematics, University of Haifa, Haifa 31905, Israel. Email: raphy@math.haifa.ac.il. This research was supported in part by the Israel Science Foundation (grant No. 1082/16).

The first to define and study these natural parameters were Pippenger and Golumbic [17]. Let $f(n, h) = \prod_i^h a_i$ where $\sum_{i=1}^h a_i = n$ and the a_i are as equal as possible. Let $g(n, h)$ be the function recursively defined as $g(n, h) = 0$ if $n < h$ and otherwise by $g(n, h) = f(n, h) + \sum_{i=1}^h g(a_i, h)$. Having noticed that the so-called *nested blowups* of H on n vertices (see definition in the next section) contain at least $g(n, h)$ induced copies of H , namely that $i_H(n) \geq g(n, h)$, they used this fact to prove that $i_H \geq h!/(h^h - h)$ for every graph H . However, as it turns out, determining i_H and, moreover, determining $i_H(n)$ seems to be a very difficult task in general. In fact, there are only very few families of graphs and sporadic cases for which these are known.

It is easy to see that $i_H(n) = i_{H^C}(n)$ where H^C is the complement of H . Clearly, $i_{K_h}(n) = i_{I_h}(n) = \binom{n}{h}$ and that cliques and their complements are the only graphs with unit inducibility. Applying a result of Goodman [10] it is also a simple exercise to determine $i_{K_{1,2}}(n)$ and deduce that $i_{K_{1,2}} = 3/4$ thereby obtaining a complete solution for all graphs on three vertices. The inducibility of most graphs on four vertices is known, see [7, 8]. However, the inducibility of P_4 , the path on four vertices, is still not completely determined. See the website of the flag algebra software of Vaughan [19] for the best upper bound obtained using the flag algebra method of Razborov [18] and Even-Zohar and Linial [7] for the best lower bound. More generally, the inducibility and maximum induced density of complete bipartite graphs is well understood [4, 5, 17]. For example, $i_{K_{h,h}} = \binom{2h}{h}/4^h$. Bollobás, Egawa, Harris, and Jin [3] proved that if H is a sufficiently large balanced blowup of some complete graph K_r , then $i_H(n)$ is obtained by blowups of K_r . Recently, Hatami, Hirst, and Norine [11] proved that if H is a sufficiently large balanced blowup of some graph K , then any graph which attains $i_H(n)$ for n sufficiently large must itself essentially be a blow-up of K .

An intriguing conjecture raised in the paper of Pippenger and Golumbic [17], which is yet unsolved, is that the inducibility of the cycle C_h for $h \geq 5$ is equal to the aforementioned lower bound $h!/(h^h - h)$. The case $h = 5$ was only very recently solved by Balogh, Hu, Lidický, and Pfender [2] with a sophisticated application of flag algebra. In fact, they have proved that $i_{C_5}(5^k) = g(5^k, 5)$ and that the unique extremal graph is the corresponding nested blowup of C_5 . They also proved that for n sufficiently large, $i_{C_5}(n) = g(n, 5)$. So, in this case nested blowups (as opposed to balanced blowups) are extremal graphs. Nested blowups appear as extremal constructions also in many other extremal problems, as shown by Pikhurko [16].

Which other graphs have the property that their inducibility is close (or equal) to the generic lower bound $h!/(h^h - h)$? Clearly, if we can prove that the nested blowups are extremal for $c_H(n)$ for all n , then a by-product would be that $i_H = h!/(h^h - h)$. Likewise, if we can prove that the nested blowups are extremal for $c_H(n)$ for a very large n as a function of H , then we can get very close to $h!/(h^h - h)$. The main result of this paper does the latter, and does it for almost all graphs H .

Recall a random graph on h vertices is the distribution $G(h, 1/2)$ on h -vertex graphs where each pair of vertices is an edge with probability $1/2$. Recall also that a graph property \mathcal{P}_h is a family of h -vertex graphs closed under isomorphism. We say that a graph $H \sim G(h, 1/2)$ satisfies \mathcal{P}_h asymptotically almost surely (alternatively, that \mathcal{P}_h is satisfied by almost all graphs),

if $\Pr[H \in \mathcal{P}_h] = 1 - o_h(1)$. In Section 3, we describe an explicit graph property \mathcal{P}_h that we call *strongly asymmetric graphs*. The following two theorems together form the crux of our result.

Theorem 1 *Let $H \sim G(h, 1/2)$.*

$$\Pr[H \in \mathcal{P}_h] = 1 - o_h(1) .$$

Theorem 2

1. *For all $n \leq 2^{\sqrt{h}}$, if $H \in \mathcal{P}_h$, then $i_H(n) = g(n, h)$. Furthermore, the extremal graphs yielding $i_H(n)$ are precisely all the nested blowups of H on n vertices.*
2. *For all $n \leq 2^{\sqrt{h}}$, if $H \in \mathcal{P}_h$ and G is an n -vertex graph with $i_H(G) \geq f(n, h)$, then G must be a balanced blowup of H with some edges added inside the blowup parts.*

Notice that Theorem 2 is a completely deterministic statement. We emphasize this since there are graphs that satisfy \mathcal{P}_h and that are quite far from a typical element of $G(h, 1/2)$ (for example, there are graphs in \mathcal{P}_h that contain linear sized cliques). Another point to mention is that one can extend the range for which Theorem 2 holds to all $n \leq 2^{h^{1-\epsilon}}$ for a given $\epsilon > 0$, at the price of redefining $\mathcal{P}_h = \emptyset$ when h is sufficiently small, depending on ϵ .

It is known that random graphs H on h vertices satisfy $i_H \leq h^{O(\log h)} h! / (h^h - h)$ asymptotically almost surely, as has been observed by Even-Zohar and Linial. Using our result we can improve this upper bound by essentially removing the $h^{O(\log h)}$ factor, thereby coming extremely close to the generic lower bound.

Theorem 3 *If $H \in \mathcal{P}_h$, then*

$$i_H \leq \frac{h!}{h^h - h} \cdot \left(1 + \frac{4}{h^{1/3}} \right) .$$

The rest of this paper has the following structure. The next section consists of definitions and description of the basic objects required for the rest of the paper. Strongly asymmetric graphs are defined in Section 3, where we also prove that almost all graphs are strongly asymmetric, namely Theorem 1. The proof of the main result, Theorem 2, appears in Section 4. Its main ingredient uses a concept of *consistent large sets* defined there. The proofs of the properties of consistent large sets and the main lemma showing that they can be grouped to a single consistent large set are given in Sections 5 and 6 respectively. Section 7 shows how to apply our main result to inducibility, proving Theorem 3. The final section contains a conjecture and some concluding remarks.

2 Preliminaries

Throughout this paper we generally do not omit floors and ceilings as these are important for the proof to work for the case where n is close to h , as we do not want to treat this (seemingly easier) case separately. For positive integers h and n , a sequence of nonnegative integers a_1, \dots, a_h whose sum is n is an (h, n) -*partition*. We say that the sequence is *equitable* if any two elements in the

sequence differ by at most 1. Otherwise, we say that the sequence is non-equitable. Clearly, if $n \pmod{h} = t$ where $0 \leq t \leq h - 1$, then an equitable (h, n) -partition has t elements of order $\lceil n/h \rceil$ and $h - t$ elements of order $\lfloor n/h \rfloor$.

Let $f(n, h)$ be the product of the elements of an equitable (h, n) -partition. Then,

$$f(n, h) = \lceil n/h \rceil^t \lfloor n/h \rfloor^{h-t}.$$

We note that $f(n, h) = 0$ for $n < h$ and that $f(h, h) = 1$. More generally, $f(kh, h) = k^h$ for a positive integer k . The following lemma summarizes a few properties of $f(n, h)$ and non-equitable partitions.

Lemma 2.1

$$f(n - 1, h) = \frac{f(n, h)}{\lceil n/h \rceil} (\lceil n/h \rceil - 1).$$

Consequently,

$$f(n - k, h) \leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil}\right)^k.$$

If a_1, \dots, a_h is a non-equitable (h, n) -partition, then

$$\prod_{i=1}^h a_i \leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil^2}\right).$$

Proof. The first part, as well as its obvious consequence are straightforward hence we prove only the third part. Consider placing n elements in h bins where the i th bin contains a_i elements. One can obtain such a placement by starting with an equitable (h, n) -partition of the elements into the bins and repeatedly moving elements from some bin whose current size is a to a bin whose current size is b where $a \leq b$. After each move the product of the bin sizes reduces by a factor of $(a - 1)(b + 1)/ab \leq 1 - 1/(ab)$. Before the first move we have $a \leq b \leq \lceil n/h \rceil$. So already after the first move the product reduces by a factor of at most $1 - \frac{1}{\lceil n/h \rceil^2}$. ■

We next define a recursive variant of $f(n, h)$. Let h and n be positive integers. Let $g(n, h) = 0$ for $n < h$. For $n \geq h$, define $g(n, h) = f(n, h) + \sum_{i=1}^h g(a_i, h)$ where a_1, \dots, a_h is an equitable (h, n) -partition. It is immediately observed that $g(n, h) = f(n, h)$ for all $n \leq h(h - 1)$ while $g(n, h) > f(n, h)$ for $n > h(h - 1)$. When $n = h^k$ for a positive integer k we obtain by induction that

$$g(h^k, h) = \sum_{i=1}^k h^{(k-1)h - (i-1)(h-1)} = \frac{h^{h(k-1)}(1 - h^{k(1-h)})}{1 - h^{1-h}}. \quad (1)$$

Let H be a graph on h vertices and assume that $V(H) = [h] = \{1, \dots, h\}$. Let a_1, \dots, a_h be an (h, n) -partition, where the a_i are positive integers. The graph $H(a_1, \dots, a_h)$ is defined as follows. Its vertex set is the disjoint union of independent sets A_1, \dots, A_h with $|A_i| = a_i$, hence it has n vertices. Its edges are defined as follows. For each pair i, j of distinct vertices of H , if $ij \in E(H)$ then the bipartite graph in $H(a_1, \dots, a_h)$ between A_i and A_j is complete, whereas if $ij \notin E(H)$ then the bipartite graph in $H(a_1, \dots, a_h)$ between A_i and A_j is empty. We say that $H(a_1, \dots, a_h)$ is

a *blowup* of H . We call A_i the *part* of the blowup corresponding to vertex $i \in V(H)$. We call a pair of distinct vertices of the blowup a *part pair* if both are in the same part and call them a *blowup pair* if they are in distinct parts. If a_1, \dots, a_h is an equitable (h, n) -partition, then $H(a_1, \dots, a_h)$ is called an n -vertex *balanced blowup* of H and is denoted by $H(n)$.

A few simple facts regarding blowups follow. An induced copy of H in $H(a_1, \dots, a_h)$ can be obtained by selecting one vertex from each part. This immediately shows that $i_H(H(a_1, \dots, a_h)) \geq \prod_{i=1}^h a_i$. In particular, an n -vertex balanced blowup of H gives that $i_H(H(n)) \geq f(n, h)$ which implies that $i_H(n) \geq f(n, h)$. Notice, however that the last inequality is not sharp for all $n > h(h-1)$. Indeed, consider $H(n)$. One of the parts has order at least h . Now, instead of letting this part be an independent set, add an induced copy of H to it. This adds at least one additional induced copy of H to the graph, while keeping all the induced copies having one vertex in each part. Hence, $i_H(n) > f(n, h)$ for $n > h(h-1)$.

The last example immediately triggers the following construction, so called the family of n -vertex *nested balanced blowups* of H , denoted by $H^*(n)$. For $n < h$ we define $H^*(n)$ to be the set of all graphs on n vertices. For $n \geq h$ we define $H^*(n)$ as follows. Take a balanced blowup $H(n)$. Now, replace each independent set A_i with an element of $H^*(a_i)$ (note: if $a_i = a_j$ we are allowed to replace A_i with an element of $H^*(a_i)$ that is different from the element of $H^*(a_i)$ that replaced a_j). An element of $H^*(n)$ is called a nested balanced blowup of H .

A few simple facts regarding nested balanced blowups follow from the definition. First notice that for all $h \leq n \leq h(h-1)$ we have that the elements of $H^*(n)$ are just the usual balanced blowups of H . Another obvious but interesting point to observe is that for a positive integer k , if $n = h^k$, then $H^*(n)$ has only one element. Also, by immediate induction, $i_H(X) \geq g(n, h)$ for any $X \in H^*(n)$ which implies that $i_H(n) \geq g(n, h)$. Notice, however, that this inequality is not always tight. In fact, in some cases, $H^*(n)$ can itself contain an element having more than $g(n, h)$ induced copies of H . Consider for example, $H = K_{1,2}$. It is immediate to check that the unique element of $K_{1,2}^*(9)$ contains more than $3^3 + 3 = g(9, 3)$ induced copies of $K_{1,2}$.

We end this section by recalling the definition of isomorphism between graphs. Let H_1 and H_2 be two graphs of the same order. A bijective function $f : V(H_1) \rightarrow V(H_2)$ is called an *isomorphism between H_1 and H_2* if $(u, v) \in E(H_1)$ if and only if $(f(u), f(v)) \in E(H_2)$. We say that H_1 and H_2 are *isomorphic* if there exists an isomorphism between them. An isomorphism between a graph and itself is called an *automorphism*. If f is a bijection between H_1 and H_2 , then a stationary point of f is a vertex v such that $f(v) = v$. Notice that the number of stationary points is at most $|V(H_1) \cap V(H_2)|$.

3 Strongly asymmetric graphs

The purpose of this section is to define the family of strongly asymmetric graphs and prove that almost all graphs reside in this family. However, before we do that, we need a couple of definitions. In what follows, *changing* a pair in a graph means either removing the edge connecting the pair, or adding an edge connecting the pair.

Definition 3.1 (blowup m -far; blowup m -close) Let H_1 and H_2 be two graphs of the same order, where H_1 is a blowup of some other graph H . We say that H_1 and H_2 are blowup m -close if one can change at most m blowup pairs of H_1 and arbitrarily change part pairs of H_1 such that after the change, the obtained graph is isomorphic to H_2 . We say that they are blowup m -far if they are not blowup m -close.

Notice that if H_1 is a trivial blowup (i.e. H_1 is just an arbitrary graph), then the latter definition just means that we can change at most m pairs of H_1 such that after the change, the obtained graph is isomorphic to H_2 . In this case we say that the graphs are m -close or m -far. More specifically, we need the following definition:

Definition 3.2 (m -far isomorphism; m -close isomorphism) Let H_1 and H_2 be two graphs of the same order, and let f be a bijection from $V(H_1)$ to $V(H_2)$. We say that f is m -close to an isomorphism between H_1 and H_2 if one can change at most m pairs of H_1 such that f is an isomorphism between the obtained graph and H_2 . Otherwise, we say that f is m -far from an isomorphism between H_1 and H_2 .

Definition 3.3 (agreement set; distinguishing set) Let H be a graph and let $S \subset V(H)$. Let $A(S) \subseteq V(H) \setminus S$ be the set of vertices with the following property. Each $w \in A(S)$ is either adjacent to all vertices of S or non-adjacent to all vertices of S . We call $A(S)$ the agreement set of S . We say that $Q \subset V(H)$ is a distinguishing set if Q is not contained in the agreement set of any two vertices outside of Q .

Let H be a graph on vertex set $[h]$. Recall from the previous section that a permutation $\pi : [h] \rightarrow [h]$ is an automorphism of H if $(\pi(i), \pi(j))$ is an edge of H if and only if (i, j) is an edge of H . Clearly, the set of all automorphisms of H is a group with respect to the composition operator, denoted by $Aut(H)$. Following Erdős and Rényi [6], we say that H is *asymmetric* if $Aut(H)$ consists only of the identity permutation and *symmetric* otherwise.

It is easy to verify that all graphs on at most 5 vertices (and more than one vertex) are symmetric. However, there are already asymmetric graphs on 6 vertices. The smallest one (with respect to the number of edges) is obtained from the path on vertices 1, 2, 3, 4, 5 (in this order) by adding vertex 6 and connecting it to vertices 3 and 4. Erdős and Rényi [6] proved that almost all graphs are asymmetric.

We will need a stronger notion of asymmetry in graphs. For a positive integer h we say that a graph H on h vertices is *strongly asymmetric*, or, synonymously, that it belongs to the family \mathcal{P}_h , if all the following conditions hold for H :

1. The degree of every vertex of H is larger than $0.4h$ and smaller than $0.6h$.
2. For every pair of vertices, the order of their agreement set is not larger than $0.55h$ and for any triple of vertices, the order of their agreement set is not larger than $0.3h$.
3. H has a distinguishing set of order at most $3 \log h$ ¹.

¹Throughout this paper all logarithms are in base 2.

4. **The blowup condition.** Let $S \subset [h]$ with $|S| \leq 0.7h$. Let B be any blowup of $H[S]$ with at least $0.8h$ vertices and at most h vertices, and let $K \subseteq [h]$ with $|K| = |B|$. Furthermore, assume that each blowup part of B contains no more than $h/100$ vertices. Then B is blowup $10^{-5}h^2$ -far from $H[K]$.
5. **The isomorphism condition.** Let $J \subset [h]$ and $K \subset [h]$ where $|J| = |K| = \lceil 0.7h \rceil$ and let π be a bijection from J to K having at least $0.1h$ non-stationary points. Then π is $10^{-5}h^2$ -far from an isomorphism between $H[J]$ and $H[K]$.

Qualitatively, the blowup condition states the large induced subgraphs of H are far from being blowups of significantly smaller induced subgraphs of H . The isomorphism condition states that two relatively large equicardinal subsets of vertices of H can be close to isomorphic only if they have small symmetric difference and isomorphism demonstrating their closeness must be almost constant. It should be noted that although not stated explicitly, strong asymmetry implies asymmetry. Indeed, suppose H is strongly asymmetric. Then, the isomorphism condition implies that any automorphism of H has at most $0.1h$ -non stationary points. But if an automorphism is non-trivial this means that there is a pair of distinct vertices whose agreement set is at least $0.9h - 2$, which is impossible by Condition 2.

We next show that almost all graphs on h vertices do in fact belong to \mathcal{P}_h . Recall that $\mathcal{G}(h, \frac{1}{2})$ is the probability space of all graphs on vertex set $[h]$ where each pair of vertices are connected with an edge with probability $\frac{1}{2}$, and the $\binom{h}{2}$ choices are independent.

Theorem 4 *For all h sufficiently large, if $H \sim G(h, 1/2)$, then*

$$\Pr[H \in \mathcal{P}_h] \geq 1 - \frac{2}{h}.$$

Hence Theorem 1 is just a more explicit restatement of Theorem 4. The proof of Theorem 4 follows from the five following lemmas, each considering one item in the definition of \mathcal{P}_h . While the first three are very elementary and their proofs are only presented for completeness, the latter two are more technical.

Lemma 3.4 *If h is sufficiently large and $H \sim G(h, 1/2)$, then with probability at least $1 - 1/(5h)$ all degrees of H are larger than $0.4h$ and smaller than $0.6h$.*

Proof. The degree of a vertex of H is a random variable with distribution $\text{Bin}(h - 1, 1/2)$ so its expectation is $(h - 1)/2$. The probability that it deviates from its expectation by more than a constant factor is exponentially small in h , and there are only h vertices to consider. ■

Lemma 3.5 *If h is sufficiently large and $H \sim G(h, 1/2)$, then with probability at least $1 - 1/(5h)$, H has the property that for every pair of vertices, the order of their agreement set is not larger than $0.55h$ and for any triple of vertices, the order of their agreement set is not larger than $0.3h$.*

Proof. Consider a pair of vertices. Their agreement set is a random variable with distribution $\text{Bin}(h - 2, 1/2)$ so its expectation is $(h - 2)/2$. The probability that it deviates from its expectation

by more than a constant factor is exponentially small in h , and there are less than h^2 pairs to consider. Consider a triple of vertices. Then their agreement set is a random variable with distribution $\text{Bin}(h-3, 1/4)$ so its expectation is $(h-3)/4$. The probability that it deviates from its expectation by more than a constant factor is exponentially small in h , and there are less than h^3 triples to consider. ■

Lemma 3.6 *If h is sufficiently large and $H \sim G(h, 1/2)$, then with probability at least $1 - 1/h$, H has a distinguishing set of at most $\lfloor 3 \log h \rfloor$ vertices.*

Proof. Fix some $Q \subset [h]$ with $|Q| = \lfloor 3 \log h \rfloor$. We prove that with probability at least $1 - 1/h$, Q is a distinguishing set. For two given vertices outside of Q , the probability that Q is contained in their agreement set is precisely $2^{|-Q|}$. As there are less than $h^2/2$ pairs of vertices outside of Q , the probability that $|Q|$ is not distinguishing is (by the union bound) at most $(h^2/2)2^{|-Q|}$ which satisfies

$$\frac{h^2}{2} 2^{|-Q|} \leq \frac{h^2}{2} 2^{1-3 \log h} = \frac{1}{h}.$$

■

Lemma 3.7 *If h is sufficiently large and $H \sim G(h, 1/2)$, then the following holds. Let $S \subset [h]$ with $|S| = s \leq 0.7h$. Let (a_1, \dots, a_s) be an (s, k) -partition with $h \geq k \geq 0.8h$ and with $a_i \leq h/100$ for $i = 1, \dots, s$. Let $K \subseteq [h]$ with $|K| = k$. Then, the probability that $B = H[S](a_1, \dots, a_s)$ is blowup $10^{-5}h^2$ -close to $H[K]$ is at most $e^{-0.001h^2}$.*

Proof. Denote the parts of B by A_1, \dots, A_s where $|A_i| = a_i \leq h/100$. Let us fix a set M of $m \leq \lfloor 10^{-5}h^2 \rfloor$ blowup pairs of B and fix a bijection $\pi : V[B] \rightarrow K$. Let B_M be the graph obtained from B after changing M (making pairs in M that are edges into non-edges and making pairs in M that are non-edges into edges). Call a vertex of B *problematic* if it appears in at least $h/30$ elements of M . Notice that being problematic has nothing to do with the structure of H , since it only depends on S and M (S is just a subset of $[h]$ and M is just a subset of blowup pairs, while blowup pairs only depend on the partition). As each pair in M contributes two to the count towards being problematic, the number of problematic vertices is at most

$$\frac{2m}{h/30} \leq 0.0006h. \tag{2}$$

Consider now the following maximum set of pairs $\{x_1, y_1\}, \dots, \{x_\ell, y_\ell\}$ that are pairwise disjoint (i.e altogether they contain 2ℓ vertices) and have the following property. For all $j = 1, \dots, \ell$, $\{x_j, y_j\}$ both belong to the same part (no matter which) and both are non problematic. We lower bound ℓ as follows. Since ℓ is maximum, from each part A_i for $i = 1, \dots, s$, there is at most one uncovered non-problematic vertex in our set of pairs. Thus,

$$\ell \geq \frac{k - 0.0006h - s}{2} \geq 0.04h \tag{3}$$

where we have used that $k \geq 0.8h$ and $s \leq 0.7h$. Consider some $\{x_j, y_j\}$ pair and suppose they belong to the part A_i . Although both x_j and y_j are non-problematic, they can still occur each in at most $h/30$ elements of M , so after ignoring the at most $2 \cdot h/30 = h/15$ vertices in pairs in M to which at least one of x_j or y_j belongs and after ignoring all the problematic vertices we still have a set Z of vertices of B of size at least

$$k - 0.0006h - \frac{h}{15} - a_i \geq 0.8h - 0.006h - \frac{h}{15} - \frac{h}{100} \geq 0.7h$$

that has the following property: For each $z \in Z$, (x_j, z) and (y_j, z) are blowup pairs that are not in M . Again, we stress that Z has nothing to do with the structure of H since it only depends on S , on M , and on problematic vertices. In other words, regardless of the structure of H , in the graph B_M , the set Z will be in the agreement set of x_j and y_j . Now consider $H[K]$. In order for π to be an isomorphism between B_M (regardless of any additional changes to part pairs of B) and $H[K]$ we need for the agreement set of each pair $\{\pi(x_1), \pi(y_1)\}, \dots, \{\pi(x_\ell), \pi(y_\ell)\}$ to be of size at least $0.7h$. What is the probability of this occurring?

Consider some pair $\{\pi(x_j), \pi(y_j)\}$. Since $H[K]$ is a random graph on k vertices, their expected agreement set in $H[K]$ is a random variable X with distribution $\text{Bin}(k-2, 1/2)$. The probability that it is larger than $0.7h$ is bounded by the standard Chernoff bound (see appendix in [1]):

$$\Pr[X > 0.7h] \leq \Pr[X > 0.7(k-2)] < e^{-2 \cdot 0.2^2(k-2)} < e^{-0.06h}$$

where we have used our assumption that $h \geq k \geq 0.8h$. Since agreement sets of distinct pairs among $\{\pi(x_1), \pi(y_1)\}, \dots, \{\pi(x_\ell), \pi(y_\ell)\}$ are pairwise independent, we have from the last inequality and from (3) that the probability that π is an isomorphism between B_M and $H[K]$ is at most

$$e^{-0.06h\ell} < e^{-0.002h^2}.$$

Now, to complete the proof of the lemma, we need to show that *no matter* which set of at most $\lfloor 10^{-5}h^2 \rfloor$ pairs we take to play the role of M , and no matter which bijection π we take, we still have that π is not an isomorphism between B_M (regardless of any additional changes to part pairs of B) and $H[K]$, with high probability. How many possible sets M can we have? As we are selecting at most $\lfloor 10^{-5}h^2 \rfloor$ pairs from a pool of at most $\binom{h}{2} < h^2/2$ possible pairs, the number of possible M is smaller than

$$\sum_{i=0}^{\lfloor 10^{-5}h^2 \rfloor} \binom{h^2/2}{i} < h^2 \binom{h^2/2}{\lfloor 10^{-5}h^2 \rfloor} < e^{0.00007h^2} \quad (4)$$

where we have used the standard Stirling approximation. It follows from the union bound and the fact that there are less than $h!$ possible π , that the probability that B is blowup $10^{-5}h^2$ -close to $H[K]$ is at most

$$h!e^{0.00007h^2} \cdot e^{-0.002h^2} < e^{-0.001h^2}.$$

■

Lemma 3.8 *If h is sufficiently large and $H \sim G(h, 1/2)$, then the following holds. Let $K \subset [h]$ and $J \subset [h]$ with $|K| = |J| = \lceil 0.7h \rceil$ and let π be a bijection from J to K having at least $0.1h$ non-stationary points. The probability that π is $10^{-5}h^2$ -close to an isomorphism between $H[J]$ and $H[K]$ is at most $e^{-0.12h^2}$.*

Proof. Let $k = \lceil 0.7h \rceil$ be the order of J and K . Let $A \subset J$ be a set of $s = \lceil 0.1h \rceil$ non-stationary points and let $B = \pi(A) \subset K$ be the set of their images. Then, π restricted to A is a bijection from A to B with no stationary points. By Lemma 3.9, we can pick $t = \lceil s/3 \rceil$ pairs $(a_1, \pi(a_1)), \dots, (a_t, \pi(a_t))$ such that all $2t$ elements in these pairs are distinct.

As in the proof of the previous lemma, let us fix a set M of $m \leq \lfloor 10^{-5}h^2 \rfloor$ pairs of vertices of J . Let $H[J]_M$ be the graph obtained from $H[J]$ after changing M (making pairs in M that are edges into non-edges and making pairs in M that are non-edges into edges). Call a vertex of J *problematic* if it appears in at least $h/30$ elements of M . Observe that being problematic has nothing to do with the structure of $H[J]$, as it only depends on the choice of M and the choice of the subset J . As in (2), the number of problematic vertices is at most $0.0006h$.

Now, remove from the set of pairs $(a_1, \pi(a_1)), \dots, (a_t, \pi(a_t))$ all the pairs for which a_i is problematic. The number of pairs that remain is at least

$$t - 0.0006h = \lceil s/3 \rceil - 0.0006h \geq 0.1h/3 - 0.0006h \geq 0.03h .$$

Without loss of generality, let these pairs be $(a_1, \pi(a_1)), \dots, (a_\ell, \pi(a_\ell))$ where $\ell = \lceil 0.03h \rceil$. So, all of these 2ℓ vertices are distinct, and all the a_1, \dots, a_ℓ are non-problematic. Consider the set $S = \{a_1, \pi(a_1), \dots, a_\ell, \pi(a_\ell)\}$. Notice that $|J \setminus S| \geq k - 2\ell$. Now, let $w \in J \setminus S$ be such that $(a_i, w) \notin M$. Since a_i is non-problematic the number of choices for w is at least

$$|J \setminus S| - h/30 \geq k - 2\ell - h/30 \geq 0.7h - 2\lceil 0.03h \rceil - h/30 \geq 0.6h .$$

consider the two pairs $\{a_i, w\}$ and $\{\pi(a_i), \pi(w)\}$. In order for π to be an isomorphism between $H[J]_M$ and $H[K]$ we must have that $\{a_i, w\}$ and $\{\pi(a_i), \pi(w)\}$ agree (both are edges or both are non-edges). Since agreement occurs with probability $\frac{1}{2}$ and since all the choices of i and w result in independent events, since they correspond to distinct pairs, we obtain that

$$\Pr[\pi \text{ is an isomorphism between } H[J]_M \text{ and } H[K]] \leq 2^{-\ell \cdot 0.6h} \leq 2^{-0.18h^2} . \quad (5)$$

We note that a similar idea to the one in the last inequality has been used in the proof of Theorem 3.1 in [15]. To complete the proof of the lemma, we must again, as in the previous lemma, consider all possible choices for M . As in (4), the number of possible M is smaller than $e^{0.00007h^2}$. It follows from the union bound that the probability that π is $10^{-5}h^2$ -close to an isomorphism between $H[J]$ and $H[K]$ is at most

$$e^{0.00007h^2} \cdot 2^{-0.18h^2} < e^{-0.12h^2} .$$

■

Recall that in the proof of Lemma 3.8 we have used the following lemma.

Lemma 3.9 *Let f be a bijection between sets A and B of size s and with no stationary points. Then there are at least $t = \lceil s/3 \rceil$ pairs $(a_1, f(a_1)), \dots, (a_t, f(a_t))$ such that all $2t$ elements in these pairs are distinct.*

Proof. A *chain* of f is a maximal length sequence (x_1, \dots, x_ℓ) such that $f(x_i) = x_{i+1}$ for $i = 1, \dots, \ell - 1$ and for which $x_\ell \notin A$ or $f(x_\ell) = x_1$ (in the latter case, the chain forms a *cycle*). Clearly, $A \cup B$ can be partitioned into chains. Furthermore, no chain is a singleton since there are no stationary points. So, from each chain (x_1, \dots, x_ℓ) we can pick at least $\lfloor \ell/2 \rfloor$ pairs $(x_1, x_2 = f(x_1)), (x_3, x_4 = f(x_3))$ and so on, such that all the elements in the picked pairs are distinct. Thus, if the partition has c odd length chains, we have picked $(|A \cup B| - c)/2$ pairs. Now, since $|A \cup B| \geq s$ and since $c \leq |A \cup B|/3$ the lemma follows. ■

Proof of Theorem 4. Assume that h is sufficiently large to satisfy all five lemmata 3.4 until 3.8. Consider $H \sim G(h, 1/2)$. The probability that it does not satisfy the first condition for \mathcal{P}_h is at most $1/(5h)$ by Lemma 3.4. The probability that it does not satisfy the second condition for \mathcal{P}_h is at most $1/(5h)$ by Lemma 3.5. The probability that it does not satisfy the third condition for \mathcal{P}_h is at most $1/h$ by Lemma 3.6. There are less than 2^h options for $S \subset [h]$, at most 2^h options for $K \subseteq [h]$ and less than h^h options for B to be the blowup of $H[S]$. Hence, the probability that H does not satisfy the fourth condition (the blowup condition) for \mathcal{P}_h is, by Lemma 3.7 at most

$$2^h \cdot 2^h \cdot h^h \cdot e^{-0.001h^2} < \frac{1}{5h}$$

(where here we assume that h is also sufficiently large to satisfy the last inequality). There are less than 2^h options for each of J and K in the fifth condition (the isomorphism condition) of \mathcal{P}_h , and less than $h!$ possible bijections π . Hence, the probability that H does not satisfy the fifth condition for \mathcal{P}_h is, by Lemma 3.8 at most

$$2^h \cdot 2^h \cdot h! \cdot e^{-0.12h^2} < \frac{1}{5h}.$$

Hence the probability that $H \notin \mathcal{P}_h$ is at most

$$4 \cdot \frac{1}{5h} + \frac{1}{h} < \frac{2}{h}.$$

■

4 Proof of the main result

We will prove a stronger theorem which immediately implies Theorem 2 by just using $\epsilon = \frac{1}{2}$ in the following statement and by redefining $\mathcal{P}_h = \emptyset$ for $h \leq h_0(\frac{1}{2})$.

Theorem 5 *For every $\epsilon > 0$ there exists $h_0 = h_0(\epsilon)$ such that the following hold for all $h > h_0$ and for all $n \leq 2^{h^{1-\epsilon}}$.*

1. If $H \in \mathcal{P}_h$, then $i_H(n) = g(n, h)$. Furthermore, the family of extremal graphs is precisely $H^*(n)$.
2. if G has n vertices and $i_H(G) \geq f(n, h)$ then G must be obtained from a balanced blowup of H by adding some edges inside the blowup parts.

Proof. The fact that $i_H(n) \geq g(n, h)$ was already shown in Section 2. Thus, we must prove that $i_H(n) \leq g(n, h)$ and that the only graphs which attain this bound are the nested balanced blowups of H . Throughout the proof we shall assume that h is sufficiently large as a function of ϵ to satisfy various inequalities, thereby establishing the existence of h_0 . The proof will be by induction on n where the base case $n \leq h$ is trivial since $g(h, h) = 1$ and since $g(n, h) = 0$ for all $n < h$.

Let $n \leq 2^{h^{1-\epsilon}}$ and assume that we have already established the result for all values smaller than n . Let G be a graph with n vertices. Let \mathcal{H} denote the set of all induced copies of H in G . In other words, considering G and H as labeled graphs, each $F \in \mathcal{H}$ is identified with an injective function from $V(H)$ to $V(G)$ which preserves adjacencies and non-adjacencies (thus, we view F as such a function). We must therefore prove that $|\mathcal{H}| \leq g(n, h)$.

Assume that $V(H) = [h]$. Since $H \in \mathcal{P}_H$, it has a distinguishing set $Q \subset [h]$ with $|Q| = q \leq 3 \log h$. Recall that Q has the following property. For any two distinct vertices $u, v \in [h] \setminus Q$, there exists $w \in Q$ such that exactly one of u or v is adjacent to w . For the rest of this proof we fix Q .

Definition 4.1 [Role consistent] Let $F_1, F_2 \in \mathcal{H}$ be two induced copies of H in G . We say that F_1 and F_2 are role-consistent if for all $x \in \text{Im}(F_1) \cap \text{Im}(F_2)$ we have $F_1^{-1}(x) = F_2^{-1}(x)$. A subset $\mathcal{F} \subseteq \mathcal{H}$ is role-consistent if any pair of elements of \mathcal{F} are role-consistent.

A role-consistent subset \mathcal{F} induces a partition $P(\mathcal{F}) = \{R, P_1, \dots, P_h\}$ of $V(G)$ as follows. For each $v \in V(G)$, if v is in no element of \mathcal{F} (namely, in no image of any $F \in \mathcal{F}$), then assign v to R . Otherwise, let $v \in \text{Im}(F)$ for some $F \in \mathcal{F}$. Then, place v in P_j where $F^{-1}(v) = j$. Since \mathcal{F} is role-consistent, this assignment is consistent regardless of the choice of $F \in \mathcal{F}$ for which $v \in \text{Im}(F)$. Stated otherwise, every copy of H in \mathcal{F} has the property that it contains precisely one vertex from each of P_1, \dots, P_h where the vertex of the copy that plays the role of i belongs to P_i .

Definition 4.2 [Role partition; redundant part] We call $P(\mathcal{F})$ the role partition of \mathcal{F} and call R the redundant part of the role partition.

It immediately follows that if \mathcal{F} is role-consistent and $P(\mathcal{F}) = \{R, P_1, \dots, P_h\}$ is its role partition, then

$$|\mathcal{F}| \leq \prod_{i=1}^h |P_i| \leq f(n - |R|, h) \leq f(n, h). \quad (6)$$

Definition 4.3 [Q-partition] Let \mathcal{I} be the set of all injective functions from Q to $V(G)$. For $I \in \mathcal{I}$ let \mathcal{H}_I be the set of all elements $F \in \mathcal{H}$ which are consistent with I . Namely, $F(x) = I(x)$ for all $x \in Q$. Clearly $\{\mathcal{H}_I \mid I \in \mathcal{I}\}$ is a partition of \mathcal{H} which we call the Q -partition.

Observe also that since $|Q| = q$ we have $|\mathcal{I}| < n^q$.

Our high level approach is to prove that a delicate modification of the Q -partition has the property that most of its elements are small, and that the remaining (not so small) elements of this partition can be absorbed into one large part that is still role-consistent. We first need the following lemma.

Lemma 4.4 *Each element of the Q -partition is role-consistent. Namely, for each $I \in \mathcal{I}$, the set \mathcal{H}_I is role-consistent.*

Proof. Let $F_1, F_2 \in \mathcal{H}_I$. Let $x \in \text{Im}(F_1) \cap \text{Im}(F_2)$. Suppose $F_1^{-1}(x) = i$ and $F_2^{-1}(x) = j$. Now, if $i \in Q$ or $j \in Q$, then by the definition of \mathcal{F}_I we must have $i = j$. Otherwise, if $i \notin Q$ and $j \notin Q$, then for every vertex $w \in Q$, either both i and j are adjacent to w or both are non-adjacent. Thus, by the property of Q , i and j must be the same vertex of H , namely $i = j$. \blacksquare

By Lemma 4.4, by (6) and by the fact that $|\mathcal{I}| < n^q$ we have the following crude upper bound.

$$|\mathcal{H}| \leq f(n, h)n^q \leq f(n, h)n^{3 \log h}. \quad (7)$$

Notice that this bound applies to all $n \geq h$. Recall however, that the upper bound obtained in (7) is still far from what we require as we would like to essentially eliminate the $n^{3 \log h}$ factor.

Consider the following process performed on a role-consistent set \mathcal{F} which we call *core production*. Start by defining $X_0 = \mathcal{F}$. As long as X_i has the property that there is a vertex v of G that appears in at most $f(n, h)n^{-3-3 \log h}$ elements of X_i (X_i is a subset of \mathcal{F} hence elements of X_i are induced copies of H in G), then remove from X_i all elements containing v and denote the remaining set by X_{i+1} . When the process ends at some stage t we either have $X_t = \emptyset$ or else X_t is a subset of \mathcal{F} having the property that each vertex appearing in any element of X_t appears in at least $f(n, h)n^{-3-3 \log h}$ elements of X_t . Set $X_t = \mathcal{F}^*$.

Definition 4.5 [core; leftover; large] *Let \mathcal{F} be role-consistent. We call \mathcal{F}^* a core of \mathcal{F} and call $\mathcal{F}^\# = \mathcal{F} \setminus \mathcal{F}^*$ a leftover of \mathcal{F} . We say that a nonempty \mathcal{F} is large if $\mathcal{F}^* = \mathcal{F}$.*

Notice that since \mathcal{F}^* is a subset of \mathcal{F} , it is also role-consistent. Also observe that $(\mathcal{F}^*)^* = \mathcal{F}^*$ so every nonempty core is large. Finally, observe that if \mathcal{F} is large, then every vertex of G that appears in an element of \mathcal{F} appears in at least $f(n, h)n^{-3-3 \log h}$ other elements, so in particular $|\mathcal{F}| \geq f(n, h)n^{-3-3 \log h}$.

Lemma 4.6 *Suppose we performed core production on each element of the Q -partition, thereby obtaining cores $\{\mathcal{H}_I^* \mid I \in \mathcal{I}\}$ and leftovers $\{\mathcal{H}_I^\# \mid I \in \mathcal{I}\}$. Let $\mathcal{H}^\# = \cup_{I \in \mathcal{I}} \mathcal{H}_I^\#$. Then,*

$$|\mathcal{H}^\#| \leq \frac{f(n, h)}{n^2}.$$

Proof. In the core production of a role-consistent \mathcal{F} we add at each step at most $f(n, h)n^{-3-3 \log h}$ elements to the leftover set, where all the added elements of the step contain some vertex $v \in V(G)$. Since there are n vertices, the final leftover of \mathcal{F} satisfies

$$|\mathcal{F}^\#| \leq n \cdot f(n, h)n^{-3-3 \log h} = f(n, h)n^{-2-3 \log h}.$$

Since the number of elements $I \in \mathcal{I}$ is less than $n^q \leq n^{3 \log h}$, we have

$$|\mathcal{H}^\#| \leq \sum_{I \in \mathcal{I}} |\mathcal{H}_I^\#| \leq n^{3 \log h} f(n, h) n^{-2-3 \log h} = \frac{f(n, h)}{n^2}.$$

■

Our main lemma (whose proof is deferred to Section 6) is the following.

Lemma 4.7 *Suppose that \mathcal{F} and \mathcal{E} are two large sets. Then their union is role-consistent.*

Once we establish Lemma 4.7 we can complete the proof of Theorem 5 as follows. Consider the set of cores $\{\mathcal{H}_I^* \mid I \in \mathcal{I}\}$ and leftovers $\{\mathcal{H}_I^\# \mid I \in \mathcal{I}\}$. Recall that if $\mathcal{H}_I^* \neq \emptyset$, then \mathcal{H}_I^* is large. Hence, by Lemma 4.7 used repeatedly,

$$\mathcal{H}^* = \cup_{I \in \mathcal{I}} \mathcal{H}_I^*$$

is role-consistent (observe that if \mathcal{F} and \mathcal{E} are large and their union is role-consistent then this union is large as well). There are now two cases to consider:

Case 1: $|\mathcal{H}^*| < f(n, h)$. Consider the role-partition $P(\mathcal{H}^*) = \{R, P_1, \dots, P_h\}$. Now, $|\mathcal{H}^*| < f(n, h)$ could happen for three reasons. (i) Either $R \neq \emptyset$. (ii) Else, the partition is not equitable. (iii) Else, the partition is equitable, but there is some pair of vertices u, v in G such that $u \in P_i, v \in P_j, i \neq j$, and the following occurs. either $uv \in E(G)$ but $ij \notin E(H)$ or $uv \notin E(G)$ but $ij \in E(H)$. Equivalently, no element of \mathcal{H}^* contains both u and v .

Suppose (i) occurs. In this case, $|\mathcal{H}^*| \leq f(n-1, h)$. But then by Lemma 2.1 and Lemma 4.6 we obtain:

$$|\mathcal{H}| \leq |\mathcal{H}^*| + |\mathcal{H}^\#| \leq f(n-1, h) + \frac{f(n, h)}{n^2} = \left(1 - \frac{1}{\lceil n/h \rceil}\right) f(n, h) + \frac{f(n, h)}{n^2} < f(n, h).$$

Suppose (ii) occurs. In this case, $|\mathcal{H}^*|$ is at most $\prod_{i=1}^h |P_i|$ and the partition is not equitable, hence by Lemma 2.1 and Lemma 4.6 we obtain:

$$|\mathcal{H}| \leq |\mathcal{H}^*| + |\mathcal{H}^\#| \leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil^2}\right) + \frac{f(n, h)}{n^2} < f(n, h).$$

Suppose (iii) occurs. In this case, no element of \mathcal{H}^* contains both u and v and the partition is equitable. Thus,

$$|\mathcal{H}^*| \leq f(n, h) \left(1 - \frac{1}{|P_i||P_j|}\right) \leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil^2}\right).$$

So, again by Lemma 4.6 we obtain:

$$|\mathcal{H}| \leq |\mathcal{H}^*| + |\mathcal{H}^\#| \leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil^2}\right) + \frac{f(n, h)}{n^2} < f(n, h).$$

Case 2: $|\mathcal{H}^*| = f(n, h)$. First observe that the structure of \mathcal{H}^* is unique. Clearly we must have $P(\mathcal{H}^*) = \{\emptyset, P_1, \dots, P_h\}$ and the partition is equitable. Furthermore, for each choice of h vertices,

one from each part (i.e. a transversal), we obtain an element of \mathcal{H}^* , as this is the only way to get $f(n, h)$ elements in \mathcal{H}^* . Observe also that this means that the bipartite graph between P_i and P_j is either empty (if $ij \notin E(H)$) or complete (if $ij \in E(H)$). Notice that as a bonus we have now proved the second part of Theorem 5, since in the previous case where $|\mathcal{H}^*| < f(n, h)$ we have already shown that $|\mathcal{H}| < f(n, h)$.

We claim that every element of $\mathcal{H}^\#$ is entirely contained in some P_i . Assume otherwise, and let $X \in \mathcal{H}^\#$ have its vertices in more than one part. Observe that it cannot have each vertex in a distinct part, as in this case we are repeating a copy of H which is already in \mathcal{H}^* , while $\mathcal{H}^* \cap \mathcal{H}^\# = \emptyset$. So, X has vertices in more than one part but not in all parts. Consider the smallest non-singleton part, say P_i . Thus, $|P_i \cap V(X)| \geq 2$. If $|P_i \cap V(X)| < 0.4h$, then any two vertices from $P_i \cap V(X)$ have all the vertices in $V(X) \setminus P_i$, namely more than $0.6h$ vertices, in their agreement set. This contradicts the assumption that $H \in \mathcal{P}_H$. If $|P_i \cap V(x)| > 0.6h$ then any vertex in $V(X) \setminus P_i$ has either all the vertices of $P_i \cap V(x)$ as its neighbors, or as its non-neighbors. But this contradicts the assumption that for $H \in \mathcal{P}_H$, the degree of any vertex is between $0.4h$ and $0.6h$. Finally, if $0.4h \leq |P_i \cap V(X)| \leq 0.6h$ then any three vertices from $P_i \cap V(X)$ have all the vertices in $V(X) \setminus P_i$, namely more than $0.4h$ vertices, in their agreement set, contradicting the assumption that $H \in \mathcal{P}_H$.

Having proved that every element of $\mathcal{H}^\#$ is entirely contained in some P_i , we can now use the induction hypotheses to obtain that

$$|\mathcal{H}^\#| \leq \sum_{i=1}^h g(|P_i|, h).$$

Therefore,

$$|\mathcal{H}| \leq |\mathcal{H}^*| + |\mathcal{H}^\#| \leq f(n, h) + \sum_{i=1}^h g(|P_i|, h) = g(n, h).$$

Finally, notice that the only way to obtain an equality is to have each P_i induce a graph with the maximum number of induced copies of H , namely, by induction, with $g(|P_i|, h)$ induced copies of H . Thus, each P_i induces an element of the nested balanced blowup on $|P_i|$ vertices. Hence, by the definition of $H^*(n)$, G is a nested balanced blowup of H with n vertices, namely $G \in H^*(n)$. ■

5 Properties of large sets

In order to prove Lemma 4.7, we first need to establish some properties that hold for every large set. Let \mathcal{F} be a large set and let $P(\mathcal{F}) = \{R, P_1, \dots, P_h\}$ be the role partition of \mathcal{F} . Our first lemma quantifies the fact that in a large set, a small fraction of parts cannot contain too many elements.

Lemma 5.1 *Let $J \subset [h]$ with $|J| = \lfloor 0.00001h \rfloor$. Then, $\sum_{j \in J} |P_j| \leq 0.0004n$.*

Proof. Consider the following process. Let $n \geq h$. Start with an equitable partition of n elements into h parts. The product of the sizes of the parts is thus $f(n, h)$. Let $0 < k < h$. Designate the

largest k parts of the equitable partition (so we prefer take parts of size $\lceil n/h \rceil$ to the designated parts as long as there are enough, and if there are less than k such parts, then the remaining designated parts are of size $\lfloor n/h \rfloor$). Now, suppose $t \geq k\lceil n/h \rceil$. Note that the designated parts contain altogether at most t elements. Move elements from the non-designated parts to the designated parts so that after each move, the designated parts form an equitable partition of k parts and the non-designated parts form an equitable partition into $h - k$ parts. Stop moving after the designated parts contain precisely t elements. The number of moves that we have made is therefore at least

$$t - \lceil n/h \rceil \cdot k .$$

What happened to the product of the sizes of the parts after each move? Suppose we have moved an element from a non-designated part whose current size is a to a designated part whose current size is b , hence $b \geq a$. So the ratio between the product of the part sizes after the move and the product before the move is precisely

$$\frac{(a-1)(b+1)}{ab} = \left(1 - \frac{b-a+1}{ab}\right) < 1 - \frac{1}{a} + \frac{1}{b} .$$

Notice that at any point, we have $a \leq \lceil n/h \rceil$. Also, after half of the moves are made, each element in a designated set contains at least $\lfloor t/2k \rfloor$ elements. Hence for the last half of the moves, we have $b \geq \lfloor t/2k \rfloor$. In these moves the aforementioned ratio is at most

$$1 - \frac{1}{\lceil n/h \rceil} + \frac{1}{\lfloor t/2k \rfloor} .$$

So after all moves are completed, the product of the sizes of the parts is now at most

$$f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil} + \frac{1}{\lfloor t/2k \rfloor}\right)^{t/2 - \frac{\lceil n/h \rceil \cdot k}{2}} . \quad (8)$$

Let us now return to the statement of the lemma. We will consider the product of the sizes of the parts $\prod_{i=1}^h |P_i|$. Let $k = \lfloor 0.00001h \rfloor$ and let $t > 0.0004n$. Suppose the lemma does not hold. Then we have k parts that together contain t elements. By convexity, the maximum of the product of the sizes of the parts subject to this condition is obtained when $R = \emptyset$ (no redundant part), and when the k parts form an equitable partition of t into k parts, and the remaining $h - k$ parts form an equitable partition of $n - t$ into $h - k$ parts. By (8) we have that

$$\begin{aligned} \prod_{i=1}^h |P_i| &\leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil} + \frac{1}{\lfloor t/2k \rfloor}\right)^{t/2 - \frac{\lceil n/h \rceil \cdot k}{2}} \\ &\leq f(n, h) \left(1 - \frac{1}{\lceil n/h \rceil} + \frac{1}{\lfloor 5n/h \rfloor}\right)^{0.0002n - \frac{\lceil n/h \rceil \cdot \lfloor 0.00001h \rfloor}{2}} \\ &\leq f(n, h) \left(1 - \frac{h}{4n}\right)^{0.0001n} \\ &\leq f(n, h) e^{-0.000025h} \\ &\leq f(n, h) 2^{-h^{1-\epsilon} 4 \log h} \\ &< f(n, h) n^{-4 \log h} . \end{aligned}$$

Here we have used $t/(2k) \geq 5n/h$ and $\frac{1}{\lceil n/h \rceil} - \frac{1}{\lceil 5n/h \rceil} \geq h/(4n)$ while recalling also our assumptions that h is sufficiently large as a function of ϵ and that $n \leq 2^{h^{1-\epsilon}}$. Now, since $|\mathcal{F}| \leq \prod_{i=1}^h |P_i|$ we have reached a contradiction to the assumption that \mathcal{F} is large, as large sets have at least $f(n, h)n^{-3-3\log h}$ elements. \blacksquare

Call a part P_i *small* if it contains at most $\lfloor n/(10h) \rfloor$ vertices.

Lemma 5.2 *The number of small parts is less than $\lfloor 0.0002h \rfloor$.*

Proof. The proof is similar to the previous one, only now we wish to focus on moving elements away from a set of designated parts so the bounds we obtain on the product of the set sizes fits better to this scenario. Consider the following process. Let $n \geq h$. Start with an equitable partition of n elements into h parts. The product of the sizes of the parts is thus $f(n, h)$. Let $0 < k < h$. Designate the smallest k parts of the equitable partition. Now, suppose $2t \leq k \lfloor n/h \rfloor$. Note that the designated parts contain altogether at least $2t$ elements. Move elements from designated parts to the non-designated parts so that after each move, the designated parts form an equitable partition of k parts and the non-designated parts form an equitable partition into $h - k$ parts. Stop moving after the designated parts contain precisely t elements.

What happened to the product of the sizes of the parts after each move? Suppose we have moved an element from a designated part whose current size is a to a non-designated part whose current size is b , hence $b \geq a$. So the ratio between the product of the part sizes after the move and the product before the move is smaller than $1 - \frac{1}{a} + \frac{1}{b}$. Notice that at any point, we have $b \geq \lfloor n/h \rfloor$. Consider the last t moves. In each such move, the total number of elements in the designated sets prior to the move is at most $2t$. Hence, for each such move we have $a \leq \lceil 2t/k \rceil$. So in these moves the aforementioned ratio is at most

$$1 - \frac{1}{\lceil 2t/k \rceil} + \frac{1}{\lfloor n/h \rfloor}.$$

So after all moves are completed, the product of the sizes of the parts is now at most

$$f(n, h) \left(1 - \frac{1}{\lceil 2t/k \rceil} + \frac{1}{\lfloor n/h \rfloor} \right)^t. \quad (9)$$

Let us now return to the statement of the lemma. We will consider the product of the sizes of the parts $\prod_{i=1}^h |P_i|$. First, observe that we may assume that $n \geq 10h$ as otherwise having a small part means having an empty part, meaning that $\mathcal{F} = \emptyset$ which is impossible as we assume that \mathcal{F} is large.

Let $k = \lfloor 0.0002h \rfloor$. Suppose the lemma does not hold. Then we have k small parts that together contain t elements. By convexity, the maximum of the product of the sizes of the parts subject to this condition is obtained when $R = \emptyset$ (no redundant part), when the k parts form an equitable partition of t into k parts, the remaining $h - k$ parts form an equitable partition of $n - t$ into $h - k$

parts, and t is maximized namely $t = k \lfloor n/10h \rfloor$. By (9) we have that

$$\begin{aligned}
\Pi_{i=1}^h |P_i| &\leq f(n, h) \left(1 - \frac{1}{\lfloor 2 \lfloor n/10h \rfloor \rfloor} + \frac{1}{\lfloor n/h \rfloor} \right)^{\lfloor 0.0002h \rfloor \lfloor n/10h \rfloor} \\
&\leq f(n, h) \left(1 - \frac{h}{n} \right)^{10^{-5}n} \\
&\leq f(n, h) e^{-10^{-5}h} \\
&\leq f(n, h) 2^{-h^{1-\epsilon} 4 \log h} \\
&< f(n, h) n^{-4 \log h} .
\end{aligned}$$

Here we have used $\frac{1}{\lfloor 2 \lfloor n/10h \rfloor \rfloor} - \frac{1}{\lfloor n/h \rfloor} \geq h/n$ while recalling also our assumptions that h is sufficiently large as a function of ϵ and that $n \leq 2^{h^{1-\epsilon}}$. Now, since $|\mathcal{F}| \leq \Pi_{i=1}^h |P_i|$ we have reached a contradiction to the assumption that \mathcal{F} is large, as large sets have at least $f(n, h) n^{-3-3 \log h}$ elements. \blacksquare

Let $u \in V(G) \setminus R$ and let $\hat{\mathcal{F}}_u$ be the set of elements of \mathcal{F} that contain u . Let $B(u, \mathcal{F}) \subset V(G)$ be the set of vertices of G that do not appear in any element of $\hat{\mathcal{F}}_u$. Clearly, $R \subseteq B(u, \mathcal{F})$ as vertices in the redundant part do not appear in any element of \mathcal{F} . Since \mathcal{F} is large, we have that any vertex of G which is not in R , appears in at least $f(n, h) n^{-3-3 \log h}$ elements of \mathcal{F} . Thus, $|\hat{\mathcal{F}}_u| \geq f(n, h) n^{-3-3 \log h}$. Let \mathcal{F}_u be the set of elements of \mathcal{F} which do not contain vertices from $B(u, \mathcal{F})$. Notice that the role partition of \mathcal{F}_u is precisely $P(\mathcal{F}_u) = \{R', P'_1, \dots, P'_h\}$ where $R' = B(u, \mathcal{F})$ and $P'_i = P_i \setminus B(u, \mathcal{F})$. Also, clearly, $\hat{\mathcal{F}}_u \subseteq \mathcal{F}_u$ but the inclusion may be proper as there may be elements in \mathcal{F} that do not contain u nor any element from $B(u, \mathcal{F})$. Given this inclusion, we have

$$|\mathcal{F}_u| \geq |\hat{\mathcal{F}}_u| \geq f(n, h) n^{-3-3 \log h} .$$

Our next lemma bounds the order of $B(u, \mathcal{F})$.

Lemma 5.3 $|B(u, \mathcal{F})| \leq 10^{-5}n$.

Proof. Suppose $|B(u, \mathcal{F})| = k$. Since each element of $\hat{\mathcal{F}}_u$ contains no vertex of $B(u, \mathcal{F})$, we have that $|\hat{\mathcal{F}}_u| \leq f(n - k, h)$. Using Lemma 2.1 we have

$$|\hat{\mathcal{F}}_u| \leq f(n - k, h) \leq f(n, h) \left(1 - \frac{1}{\lfloor n/h \rfloor} \right)^k .$$

We must show that $k \leq 10^{-5}n$. Assume otherwise, then

$$\frac{|\hat{\mathcal{F}}_u|}{f(n, h)} \leq \left(1 - \frac{1}{\lfloor n/h \rfloor} \right)^k \leq \left(1 - \frac{h}{2n} \right)^{10^{-5}n} < e^{-10^{-6}h} < 2^{-h^{1-\epsilon} 4 \log h} < n^{-4 \log h}$$

contradicting the fact that $|\hat{\mathcal{F}}_u| \geq f(n, h) n^{-3-3 \log h}$. \blacksquare

Call a part P'_i *very small* if it contains at most $n/(15h)$ vertices.

Lemma 5.4 *The number of very small parts is at most $0.0005h$.*

Proof. By Lemma 5.2, the number of small parts P_i is less than $\lfloor 0.0002h \rfloor$. To make P_i that were not small to be very small parts in $P'_i = P_i \setminus B(u, F)$, P_i must lose at least $n/(10h) - n/(15h) = n/(30h)$ of its vertices because they joined $B(u, F)$. But since by Lemma 5.3 $|B(u, \mathcal{F})| \leq 10^{-5}n$, we have that the number of such parts is at most $(n/100000)/(n/(30h)) = 0.0003h$. Hence, overall there are at most $0.0003h + \lfloor 0.0002h \rfloor \leq 0.0005h$ very small parts. \blacksquare

Let $u \in P'_i$ and $v \in P'_j$ with $i \neq j$. We say that the pair is *inconsistent* if either $ij \in E(H)$ and $uv \notin E(G)$ or $ij \notin E(H)$ and $uv \in E(G)$. Otherwise, the pair is *consistent*. Let $z > 0$ be a real parameter. We define the z -consistency graph $Z_{\mathcal{F}_u}(z)$ as follows. The set of vertices of $Z_{\mathcal{F}_u}(z)$ is $[h]$ (namely, it has the same set of vertices as H). Two vertices i and j are adjacent in $Z_{\mathcal{F}_u}(z)$ if and only if the number of consistent pairs with one endpoint in P'_i and one endpoint in P'_j is more than $(1-z)|P'_i||P'_j|$.

Our next lemma shows that for relatively small z , $Z_{\mathcal{F}_u}(z)$ has large cliques.

Lemma 5.5 *$Z_{\mathcal{F}_u}(h^{-\epsilon/2})$ has a clique of size at least $0.99999h + 1$.*

We will actually prove that the complement of $Z_{\mathcal{F}_u}(h^{-\epsilon/2})$ does not have a matching of size at least $0.000005h - 1$, as this implies that $Z_{\mathcal{F}_u}(h^{-\epsilon/2})$ has a clique of size at least $h - 2 \cdot 0.000005h + 2 = 0.99999h + 2$. Denote this complement by C .

Suppose C has a matching of size $k = \lceil 0.000005h \rceil - 1$. Denote it by $(x_1, y_1), \dots, (x_k, y_k)$. Any copy of H in \mathcal{F}_u contains precisely one vertex from each of the parts P'_1, \dots, P'_h . By the definition of $Z_{\mathcal{F}_u}(z)$, each such copy must avoid at least $z|P'_{x_i}||P'_{y_i}|$ of the pairs of vertices from $P'_{x_i} \times P'_{y_i}$, as it must avoid inconsistent pairs and this holds for all $i = 1, \dots, k$. Hence, for $z = h^{-\epsilon/2}$ we have

$$\begin{aligned} |\mathcal{F}_u| &\leq (1-z)^k \prod_{i=1}^h |P'_i| \\ &\leq \left(1 - \frac{1}{h^{\epsilon/2}}\right)^{0.000005h-1} \prod_{i=1}^h |P'_i| \\ &\leq e^{-0.000004h^{1-\epsilon/2}} f(n, h) \\ &\leq f(n, h) 2^{-h^{1-\epsilon} 4 \log h} \\ &\leq f(n, h) n^{-4 \log h} \end{aligned}$$

contradicting the fact that $|\mathcal{F}_u| \geq f(n, h) n^{-3-3 \log h}$. \blacksquare

As we will be interested in the case $z = h^{-\epsilon/2}$, we will denote for simplicity $Z_{\mathcal{F}_u} = Z_{\mathcal{F}_u}(h^{-\epsilon/2})$. As Lemma 5.5 suggests, we have a clique of order $\lceil 0.99999h \rceil + 1$ in $Z_{\mathcal{F}_u}$ so let $K^* \subset [h]$ denote such a clique. By Lemma 5.4, there are at most $0.0005h$ very small parts. Hence, there is $K(u, \mathcal{F}) \subseteq K^*$ with $|K(u, \mathcal{F})| \geq (\lceil 0.99999h \rceil + 1) - 0.0005h - 1 \geq 0.999h$ such that $K(u, \mathcal{F})$ is a clique in $Z_{\mathcal{F}_u}$ and for any $j \in K(u, \mathcal{F})$ we have that $|P'_j| \geq n/(15h)$. Furthermore, if $u \in P_\ell$ then $\ell \notin K(u, \mathcal{F})$.

How many vertices of G appear in parts P'_i for which $i \in K(u, \mathcal{F})$ (namely, in the clique parts that are not very small, and not containing the part of u). We claim that these are most of the vertices. Denote this set by $W(u, \mathcal{F}) = \cup_{i \in K(u, \mathcal{F})} P'_i$.

Lemma 5.6 $|W(u, \mathcal{F})| \geq 0.9995n$.

Proof. A vertex $v \notin W(u, \mathcal{F})$ has to satisfy one of the following conditions. Either $v \in B(u, \mathcal{F})$, else $v \in P'_i$ but $i \notin K^*$, else $v \in P'_i$, $i \in K^*$ but P'_i is very small, else v is in the same part as u . Summing the sizes of these forbidden sets we obtain by Lemmas 5.1, 5.3, 5.4 and the fact that $|\lfloor h \rfloor \setminus K^*| + 1 \leq \lfloor 0.00001h \rfloor$ that it is at most

$$\frac{n}{100000} + 0.0004n + 0.0005h \cdot (n/(15h)) \leq 0.0005n.$$

■

Let $C^*(u, \mathcal{F})$ denote the set of inconsistent pairs x, y with $x \in P'_i$, $y \in P'_j$ and both $i, j \in K(u, \mathcal{F})$.

Lemma 5.7 $|C^*(u, \mathcal{F})| \leq n^2 h^{-\epsilon/2}$.

Proof. Clearly, by the definition of $K(u, \mathcal{F})$, the number of inconsistent pairs with one endpoint in P'_i and the other in P'_j is at most $|P'_i||P'_j|h^{-\epsilon/2}$ when $i, j \in K(u, \mathcal{F})$ (in fact, this also holds in the larger set K^*). Hence,

$$\begin{aligned} |C^*(u, \mathcal{F})| &\leq h^{-\epsilon/2} \sum_{i, j \in K(u, \mathcal{F}), i \neq j} |P'_i||P'_j| \\ &\leq h^{-\epsilon/2} \sum_{i=1}^h \sum_{j=i+1}^h |P'_i||P'_j| \\ &\leq h^{-\epsilon/2} h^2 \frac{n^2}{h^2} \leq n^2 h^{-\epsilon/2}. \end{aligned}$$

■

6 Proof of Lemma 4.7

We now return to the proof of Lemma 4.7. Recall that we want to prove that if \mathcal{F} and \mathcal{E} are two large sets, then their union is role-consistent. Let $P(\mathcal{F}) = \{R, P_1, \dots, P_h\}$ and $P(\mathcal{E}) = \{T, S_1, \dots, S_h\}$ be the role partitions of \mathcal{F} and \mathcal{E} . We must therefore prove that if $u \in P_i$ and $u \in S_j$, then $i = j$.

Assume otherwise, then there exists some $u \in V(G)$ such that $u \in P_i$ and $u \in S_j$ and $i \neq j$. Using the aforementioned notations we focus on the objects $B(u, \mathcal{F})$, \mathcal{F}_u , $P(\mathcal{F}_u) = \{R', P'_1, \dots, P'_h\}$, $Z_{\mathcal{F}_u}$, $K(u, \mathcal{F})$, $W(u, \mathcal{F})$, $C^*(u, \mathcal{F})$ and correspondingly $B(u, \mathcal{E})$, \mathcal{E}_u , $P(\mathcal{E}_u) = \{T', S'_1, \dots, S'_h\}$, $Z_{\mathcal{E}_u}$, $K(u, \mathcal{E})$, $W(u, \mathcal{E})$, $C^*(u, \mathcal{E})$. Also recall that $R' = B(u, \mathcal{F})$ and $T' = B(u, \mathcal{E})$.

Call an index $i \in K(u, \mathcal{E})$ *problematic for \mathcal{F}* if more than $|S'_i|/10$ of the vertices of S'_i do not belong to $W(u, \mathcal{F})$. Notice that unlike the definitions in the previous section which only depend on a given large set and a vertex of its partition, the definition of being problematic involves two large sets. Informally, being non-problematic means that 90 percent of the vertices of the part appear in “good sets” of the other partition. How many problematic parts are there?

Lemma 6.1 *The number of vertices of $K(u, \mathcal{E})$ that are problematic for \mathcal{F} is at most $0.075h$.*

Proof. Suppose $i \in K(u, \mathcal{E})$. Then, in particular S'_i is not very small so it contains at least $n/(15h)$ vertices. If it were problematic for \mathcal{F} , then it would have at least $0.1n/(15h) = n/(150h)$ vertices not in $W(u, \mathcal{F})$. But the number of vertices not in $W(u, \mathcal{F})$ is at most $0.0005n$ by Lemma 5.6. So, the overall number of vertices of $K(u, \mathcal{E})$ that are problematic for \mathcal{F} is at most $0.0005n/(n/(150h)) \leq 0.075h$. ■

Let J denote the set of vertices of $K(u, \mathcal{E})$ that are non-problematic for \mathcal{F} . Recall that $|K(u, \mathcal{E})| \geq 0.999h$. Hence, by Lemma 6.1, $|J| \geq (0.999 - 0.075)h \geq 0.9h$.

Let $X \subset V(G)$ be a set of $|J|$ vertices obtained by *randomly* selecting precisely one vertex from each part S'_j for $j \in J$. All $|J|$ choices are performed independently. Thus, $G[X]$ is a subgraph of G on $|J|$ vertices. We would like to first see how close $G[X]$ is to a subgraph of H on $|J|$ vertices.

Lemma 6.2 *With probability larger than $3/4$, $G[X]$ is $4h^{2-\epsilon/2}$ -close to $H[J]$.*

Proof. Denote $X = \{x_j \mid j \in J\}$. So x_j is the vertex selected at random from S'_j . Consider a pair of distinct indices i, j such that $i, j \in J$. What is the probability that x_i and x_j are consistent? First notice that since $i, j \in J$, then $|S'_i|, |S'_j| \geq n/(15h)$ as they are not very small. Next, notice that ij is a clique edge in $Z_{\mathcal{E}_u}$, and hence the number of pairs between S'_i and S'_j that are consistent is at least $(1 - h^{-\epsilon/2})|S'_i||S'_j|$. So the probability that x_i and x_j are inconsistent is at most $h^{-\epsilon/2}$. Thus, the expected number of inconsistent pairs in X is at most

$$\binom{|J|}{2} h^{-\epsilon/2} < h^{2-\epsilon/2}.$$

By Markov's inequality the probability of the number of inconsistent pairs being more than $4h^{2-\epsilon/2}$ is less than $1/4$. The proof is complete by noticing that changing the inconsistent pairs results in a graph that is isomorphic to $H[J]$ where x_j plays the role of j in $H[J]$. ■

Lemma 6.3 *Let $X^* = X \cap W(u, \mathcal{F})$. Then with probability larger than $3/4$, $|X^*| > 0.8h$.*

Proof. Recall that $X = \{x_j \mid j \in J\}$ where x_j is the vertex selected at random from S'_j . Since S'_j is non-problematic for \mathcal{F} , at least a fraction of 0.9 of its vertices are in $W(u, \mathcal{F})$. Hence the probability that $x_j \in W(u, \mathcal{F})$ is at least 0.9. The expected value of $|X^*|$ is therefore at least $0.9|J| \geq 0.9 \cdot 0.9h = 0.81h$. As $|X^*|$ is the sum of independent indicator random variables, the probability that it is smaller by a constant factor than its expectation is exponentially small in h . Hence, $\Pr[|X^*| \leq 0.8h]$ is exponentially small in h . ■

We next need to bound the number of elements of $C^*(u, \mathcal{F})$ with both endpoints in X^* .

Lemma 6.4 *With probability larger than $3/4$, $|C^*(u, \mathcal{F}) \cap (X^* \times X^*)| \leq 1000h^{2-\epsilon/2}$.*

Proof. We will actually upper bound the size of the potentially larger set $C^*(u, \mathcal{F}) \cap (X \times X)$. Let x, y be a pair of vertices in $C^*(u, \mathcal{F})$. If $x \in B(u, \mathcal{E})$, then trivially $x \notin X$. Similarly, if $y \in B(u, \mathcal{E})$, then $y \notin X$. Hence, we can assume that $x \in S'_i$ and $y \in S'_j$. Now if $i = j$ then at least one of them is not selected to X , since from each part we selected at most one vertex to X . So we can

assume $i \neq j$. If $i \notin J$ then $x \notin X$ since we selected no vertex of S'_i to X . Likewise, if $j \notin J$ then $y \notin X$ since we selected no vertex of S'_j to X . So we can assume that $i \neq j, i, j \in J$. But recall that $J \subseteq K(u, \mathcal{E})$, so $|S'_i| \geq n/(15h)$ and $|S'_j| \geq n/(15h)$ (i.e. they are not very small). Since the vertex from S'_i was chosen at random and the vertex of S'_j was chosen at random, the probability of the pair x, y to be in $X \times X$ is at most $(15h/n)^2$. By Lemma 5.7, $|C^*(u, \mathcal{F})| \leq n^2 h^{-\epsilon/2}$. Hence the expected value of $|C^*(u, \mathcal{F}) \cap (X \times X)|$ is at most

$$n^2 h^{-\epsilon/2} \cdot \left(\frac{15h}{n}\right)^2 \leq 225 h^{2-\epsilon/2}.$$

By Markov's inequality the probability of being larger than four times the expectation is less than $1/4$. \blacksquare

We now want to show that the vertices of X^* are not too concentrated among some parts of P'_1, \dots, P'_h , namely that no part P'_i contains too many vertices of X^* .

Lemma 6.5 *With probability larger than $3/4$, for all $i = 1, \dots, h$ we have $|P'_i \cap X^*| \leq h/100$.*

Proof. We will actually prove the slightly stronger statement bounding $|P'_i \cap X|$. Consider some part P'_i and the disjoint parts $P'_{i,j} = P'_i \cap S'_j$ for $j \in J$ (notice that these disjoint parts might not cover all the vertices of P'_i , but we don't care about these uncovered vertices since they cannot be selected to X). Now, since we selected one vertex from each S'_j , we have that $|P'_{i,j} \cap X| \in \{0, 1\}$. It is therefore an indicator random variable with probability of success

$$\Pr[|P'_{i,j} \cap X| = 1] = \frac{|P'_{i,j}|}{|S'_j|} \leq \frac{|P'_{i,j}|}{n/(15h)}$$

where we have used the fact that S'_j is not very small. It now follows that the expectation of $|P'_i \cap X|$ is at most

$$\sum_{j \in J} \frac{|P'_{i,j}|}{n/(15h)} \leq \frac{15h}{n} |P'_i| \leq \frac{15h}{n} \cdot 0.0004n \leq 0.006h$$

where we have used the fact that no single set P_i (moreover $P'_i \subseteq P_i$) can contain more than $0.0004n$ vertices, by Lemma 5.1. But now observe that $|P'_i \cap X|$ is the sum of $|J|$ independent indicator random variables (recall, from each S'_j for $j \in J$ we select a single vertex to X independently), so by a Chernoff bound (see appendix in [1]), the probability that $|P'_i \cap X|$ is larger than its expectation by a constant factor is exponentially small in $|J| \geq 0.9h$, so in particular, exponentially small in h . So, for h sufficiently large, with probability smaller than $1/(4h)$, $|P'_i \cap X| > h/100$. Now, by the union bound, we have that with probability larger than $3/4$, for all $i = 1, \dots, h$ it holds that $|P'_i \cap X| \leq h/100$. \blacksquare

Since each of the four lemmas 6.2, 6.3, 6.4 6.5 states that the corresponding event happens with probability larger than $3/4$, we have that with positive probability, all four events hold. Thus, we can pick a set $X = \{x_j \mid j \in J\}$ such that X and its subset $X^* = X \cap W(u, \mathcal{F})$ satisfy the following four properties:

P1 $G[X]$ is $4h^{2-\epsilon/2}$ -close to $H[J]$. In particular, $G[X^*]$ is $4h^{2-\epsilon/2}$ -close to $H[J^*]$ where $J^* \subset J$ satisfies $J^* = \{j \mid x_j \in X^*\}$.

P2 $|X^*| = |J^*| > 0.8h$.

P3 $|C^*(u, \mathcal{F}) \cap (X^* \times X^*)| \leq 1000h^{2-\epsilon/2}$.

P4 For all $i = 1, \dots, h$ we have $|P'_i \cap X^*| \leq h/100$.

We now consider the way the vertices of X^* appear in P'_1, \dots, P'_h . Let $S \subset [h] = \{i \mid X^* \cap P'_i \neq \emptyset\}$. So $i \in S$ means that it contains a least one vertex of X^* and $i \notin S$ means that it does not contain vertices of X^* . There are two cases to consider:

Case 1: $|S| \leq 0.7h$. In this case we obtain by P1 on the one hand, that $G[X^*]$ is $4h^{2-\epsilon/2}$ -close to $H[J^*]$. On the other hand, we obtain by P3 that there is a blowup of $H[S]$ that is blowup $1000h^{2-\epsilon/2}$ -close to $G[X^*]$. Furthermore, $|S| \leq 0.7h$ while $|X^*| \geq 0.8h$ by P2 and no blowup part contains more than $h/100$ vertices by P4. So, this blowup is blowup $(4h^{2-\epsilon/2} + 1000h^{2-\epsilon/2})$ -close to $H[J^*]$. Since $4h^{2-\epsilon/2} + 1000h^{2-\epsilon/2} < 10^{-5}h^2$ this contradicts the blowup condition in the definition of \mathcal{P}_h , thus our assumption that $H \in \mathcal{P}_h$.

Case 2: $|S| \geq 0.7h$. Recall that we assume that $u \in P_i$ and $u \in S_j$ where $i \neq j$. So we fix the two indices i and j . Let $S^* \subseteq S \cup \{i\}$ have precisely $\lceil 0.7h \rceil$ elements where we force $i \in S^*$. Now consider the following set of vertices X^{**} of G . For each $t \in S^*$, if $t \neq i$ then add to X^{**} some vertex of X^* which belongs to P'_t (this is always possible since $P'_t \cap X^* \neq \emptyset$ by the definition of S). If $t = i$, then add u to X^{**} . Observe that $|X^{**}| = |S^*| = \lceil 0.7h \rceil$ and that u is always in X^{**} . We will prove that $G[X^{**}]$ is very close to $H[S^*]$ and that it is also very close to $H[Y^*]$ where $Y^* = \{p \mid x_p \in X^{**}\} \cup \{j\}$. Observe also that $|Y^*| = \lceil 0.7h \rceil$.

Consider the natural bijection π from S^* to X^{**} defined as follows. Let $t \in S^*$. If $t \neq i$ let $\pi(t)$ be the vertex of X^{**} which belongs to P'_t . For $t = i$ define $\pi(i) = u$. Why is π close to an isomorphism between $H[S^*]$ and $G[X^{**}]$? Every inconsistent pair has to be in $C^*(u, \mathcal{F})$ and also has to be in $(X^* \times X^*)$, unless that pair involves u . But by the definition of \mathcal{F}_u , all the vertices in the parts P'_1, \dots, P'_h are consistent with u . Hence, by P3, π is $1000h^{2-\epsilon/2}$ -close to an isomorphism between $H[S^*]$ and $G[X^{**}]$.

Consider the natural bijection σ from X^{**} to Y^* defined as follows. Suppose $x_p \in X^{**}$ then $\sigma(x_p) = p$. Also define $\sigma(u) = j$. Why is π close to an isomorphism between $G[X^{**}]$ and $H[Y^*]$? Every inconsistent pair that does not involve u , has to be in $G[X]$. Also, $Y^* \setminus \{j\} \subset J^*$. But by the definition of \mathcal{E}_u , all the vertices in the parts S'_1, \dots, S'_h are consistent with u . Hence, by P1, σ is $4h^{2-\epsilon/2}$ -close to an isomorphism between $G[X^{**}]$ and $H[Y^*]$.

It follows that the composition $\pi\sigma$ is $4h^{2-\epsilon/2} + 1000h^{2-\epsilon/2}$ close to an isomorphism between $H[S^*]$ and $H[Y^*]$. Since $4h^{2-\epsilon/2} + 1000h^{2-\epsilon/2} < 10^{-5}h^2$, we must have by the isomorphism condition in the definition of \mathcal{P}_h , that $\pi\sigma$ has less than $0.1h$ non-stationary points.

There now two options to consider. If $|S^* \triangle Y^*| \geq 0.2h$, then trivially $\pi\sigma$ has at least $|S^* \triangle Y^*|/2 = 0.1h$ non-stationary points, a contradiction. If $|S^* \triangle Y^*| < 0.2h$, then $|S^* \cap Y^*| \geq 0.6h$. But since $\pi(i) = u$ and $\sigma(u) = j$ this means that in H , the distinct vertices i and j have more

than $0.6h - 2 > 0.55h$ vertices in their agreement set, contradicting the assumption that $H \in \mathcal{P}_h$. ■

7 Inducibility

In this section we prove Theorem 3. Let $H \in \mathcal{P}_h$ from Theorem 2. By Theorem 2 it holds for $n \leq 2^{\sqrt{h}}$ that $i_H(n) = g(n, h)$. It will be profitable to use the largest k such that $n = h^k$ and still $n \leq 2^{\sqrt{h}}$. We can therefore assume that $k \geq h^{1/3} + 2$ as

$$h^{\lceil h^{1/3} + 2 \rceil} \leq 2^{\sqrt{h}}.$$

Now, by (1) we have that

$$i_H(h^k) = g(h^k, h) = \frac{h^{h(k-1)}(1 - h^{k(1-h)})}{1 - h^{1-h}}.$$

Since $g(h^k, h) / \binom{h^k}{h}$ serves as a trivial upper bound for i_H we obtain that

$$i_H \leq \frac{h^{h(k-1)}(1 - h^{k(1-h)})}{(1 - h^{1-h})\binom{h^k}{h}}.$$

Clearly if we would have let $k \rightarrow \infty$ we would have obtained $i_H = h! / (h^h - h)$. However we can only assume that k is bounded from below by $h^{1/3} + 2$. Nevertheless, we have that

$$\frac{h^{h(k-1)}(1 - h^{k(1-h)})}{(1 - h^{1-h})\binom{h^k}{h}} \leq \frac{h! h^{h(k-1)}(1 - h^{k(1-h)})}{(1 - h^{1-h})(h^k - h)^h} \leq \frac{h!}{h^h - h} \cdot \frac{1}{(1 - h^{1-k})^h}.$$

But notice that

$$\frac{1}{(1 - h^{1-k})^h} \leq e^{\frac{2}{h^{k-2}}} \leq 1 + \frac{4}{h^{k-2}} \leq 1 + \frac{4}{h^{h^{1/3}}}.$$

It follows that

$$i_H \leq \frac{h!}{h^h - h} \cdot \left(1 + \frac{4}{h^{h^{1/3}}}\right).$$
■

8 Concluding remarks and open problems

We conjecture that the statement of Theorem 5 and hence of Theorem 2 can be extended to all n . Moreover:

Conjecture 1 *For all strongly asymmetric graphs H it holds that $i_H(n) = g(n, h)$ for all $n \in \mathbb{N}$. In particular,*

$$i_H = \frac{h!}{h^h - h}.$$

It would be extremely interesting to determine all graphs for which $i_H = \frac{h!}{h^h - h}$. As mentioned in the introduction, all cycles of length at least 5 are conjectured to be in this family. Likewise, determining the set of graphs for which $i_H(n) = g(n, h)$ for, say, all $n \leq h^3$, would also be gratifying as such graphs have $i_H = (1 + o_h(1)) \frac{h!}{h^h - h}$, as can be seen from the proof of Theorem 3. Given that this is true (for a much larger range of n) for all strongly asymmetric graphs, one might stipulate that Theorem 2 (or maybe even Conjecture 1) holds for all asymmetric graphs. In fact, as the next proposition shows, this is very far from true.

Proposition 8.1 *For all h sufficiently large, there are asymmetric graphs H on h vertices for which $i_H(n) > g(n, h)$ for $n < h \log h$.*

Proof. Recall that a random graph from $G(q, 1/2)$ is almost surely asymmetric [6]. Furthermore, almost surely a random graph from $G(q, 1/2)$ has all its degrees at most $0.55q$ and the number of common neighbors of every two vertices is at least $0.2q$ (see the proofs of Lemma 3.4 and Lemma 3.5). Hence, for q sufficiently large, we can find (much more than) 2^q pairwise non-isomorphic graphs H_1, \dots, H_{2^q} such that each H_i is an asymmetric graph on q vertices, its maximum degree is at most $0.55q$, and the number of common neighbors in H_i of every two vertices is at least $0.2q$.

Now let H be the graph on $h = q2^q$ vertices obtained by taking a balanced blowup of the clique K_{2^q} where each blowup part is of size q . Suppose the parts are A_1, \dots, A_{2^q} . Now make each A_i induce a copy of H_i .

We claim that H is asymmetric. Indeed, consider some automorphism π . We first notice that $\pi(A_i) = A_j$ for some j (possibly $i = j$). Suppose not, then there are $u, v \in A_i$ such that $\pi(u) \in A_j$ and $\pi(v) \in A_k$ for $j \neq k$. Since π is an automorphism, the number of common neighbors of u and v must equal the number of common neighbors of $\pi(u)$ and $\pi(v)$. But the number of common neighbors of u and v in H is at least $h - q + z \geq h - q + 0.2q$ where z is the number of common neighbors of u and v in H_i . On the other hand, the number of common neighbors of $\pi(u)$ and $\pi(v)$ in H is $h - 2q + x + y \leq h - 2q + 0.55q + 0.55q = h - q + 0.1q$ where x is the degree of $\pi(v)$ in H_k and y is the degree of $\pi(u)$ in H_j . This contradicts that π is an automorphism. Hence, we must have $\pi(A_i) = A_i$ for some j .

But now we claim that we must have $\pi(A_i) = A_i$. Indeed, if $\pi(A_i) = A_j$, then since π is an automorphism, this means that H_i and H_j are isomorphic. Since H_i is not isomorphic to H_j whenever $i \neq j$, this implies that $\pi(A_i) = A_i$ for all $i = 1, \dots, 2^q$. But if $\pi(A_i) = A_i$ this means that π restricted to A_i is an automorphism of H_i . Since H_i is asymmetric, π must be trivial on A_i . As this holds for all $i = 1, \dots, 2^q$, we must have that π is trivial, proving that H is asymmetric.

Designate a set $X \subset V(H)$ order 2^q obtained by taking one vertex from each A_i for $i = 1, \dots, 2^q$. Notice in fact that $H[X]$ is a clique of order 2^q . Now consider the graph G on $n = hq$ vertices which is the balanced blowup of H . So the parts of this blowup are $\{P_x \mid x \in V(H)\}$ and $|P_x| = q$. Now, replace each independent set P_x with a copy of H_i where $x \in A_i$ (recall that the A_i are a partition of H and each induce a copy of H_i in H). Now consider the subgraph of G induced by the parts $\{P_x \mid x \in X\}$. Notice that this subgraph is isomorphic to H . On the other hand, it is not an induced copy of H obtained by taking one vertex from each part of G , so it is an additional

induced copy to the $f(n, h) = g(n, h) = q^h$ induced copies that can be obtained as such. Hence, $i_H(G) > g(n, h)$ while $n = hq < h \log h$. ■

Theorem 5, and therefore Theorem 2 have variants which apply to other graph densities and some other combinatorial structures. Indeed, suppose $0 < p < 1$ is given. One can adjust the definition of \mathcal{P}_h to obtain a graph property $\mathcal{P}_h(p)$ which contains almost all graphs of $G(h, p)$. Indeed, Condition 1 in the definition could be changed to requiring that the degree be between $p(1 - \delta)n$ and $p(1 + \delta)n$ for an absolute constant δ , that the agreement sets in Condition 2 would be at most $(1 + \delta)(p^2 + (1 - p)^2)$ and $(1 + \delta)(p^3 + (1 - p)^3)$ respectively, that the distinguishing set in Condition 3 be of size at most $-3 \log(p^2 + (1 - p)^2) \log h$, and that the constant $10^{-5}h^2$ in Conditions 4,5 be replaced with $c \min\{p, 1 - p\}h^2$ for some absolute constant c . The proof of Theorem 5 and its lemmas stay essentially the same, after adjusting constants everywhere. Similarly, a variant of the theorems applies to the family of tournaments. We can naturally define a tournament to be strongly asymmetric by modifying the definition of \mathcal{P}_h to suit tournaments (namely, degree requirements are replaced with in-degree and out-degree requirements, changing an edge means flipping its direction, etc.). Likewise, instead of considering $G(h, 1/2)$ we consider the probability space of all tournaments on h vertices.

9 Acknowledgment

Independently of our work, Jacob Fox, Hao Huang, and Choongbum Lee (private communication) have sent the author a manuscript proving that almost all graphs H have $i_H(n) = g(n, h)$ for all $n \in \mathbb{N}$. Once their paper is publicly available, a reference will be given in the archived version of this manuscript.

References

- [1] N. Alon and J. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.
- [2] J. Balogh, P. Hu, B. Lidický, and F. Pfender. Maximum density of induced 5-cycle is achieved by an iterated blow-up of 5-cycle. *European Journal of Combinatorics*, 52:47–58, 2016.
- [3] B. Bollobás, Y. Egawa, A. Harris, and G. Jin. The maximal number of induced r -partite subgraphs. *Graphs and Combinatorics*, 11(1):1–19, 1995.
- [4] B. Bollobás, C. Nara, and S. Tachibana. The maximal number of induced complete bipartite graphs. *Discrete mathematics*, 62(3):271–275, 1986.
- [5] J. Brown and A. Sidorenko. The inducibility of complete bipartite graphs. *Journal of Graph Theory*, 18(6):629–645, 1994.
- [6] P. Erdős and A. Rényi. Asymmetric graphs. *Acta Mathematica Hungarica*, 14(3-4):295–315, 1963.

- [7] C. Even-Zohar and N. Linial. A note on the inducibility of 4-vertex graphs. *Graphs and Combinatorics*, 31(5):1367–1380, 2015.
- [8] G. Exoo. Dense packings of induced subgraphs. *Ars Combinatoria*, 22:5–10, 1986.
- [9] V. Falgas-Ravry and E. Vaughan. Turán H -densities for 3-graphs. *Electronic Journal of Combinatorics*, 19(3):P40, 2012.
- [10] W. Goodman. On sets of acquaintances and strangers at any party. *The American Mathematical Monthly*, 66(9):778–783, 1959.
- [11] H. Hatami, J. Hirst, and S. Norine. The inducibility of blow-up graphs. *Journal of Combinatorial Theory, Series B*, 109:196–212, 2014.
- [12] D. Hefetz and M. Tyomkyn. On the inducibility of cycles. *arXiv preprint arXiv:1702.07342*, 2017.
- [13] J. Hirst. The inducibility of graphs on four vertices. *Journal of Graph Theory*, 75(3):231–243, 2014.
- [14] H. Huang. On the maximum induced density of directed stars and related problems. *SIAM Journal on Discrete Mathematics*, 28(1):92–98, 2014.
- [15] J. H. Kim, B. Sudakov, and V. H. Vu. On the asymmetry of random regular graphs and random graphs. *Random Structures & Algorithms*, 21(3-4):216–224, 2002.
- [16] O. Pikhurko. On possible Turán densities. *Israel Journal of Mathematics*, 201(1):415–454, 2014.
- [17] N. Pippenger and M. C. Golumbic. The inducibility of graphs. *Journal of Combinatorial Theory, Series B*, 19(3):189–203, 1975.
- [18] A. Razborov. Flag algebras. *Journal of Symbolic Logic*, 72(4):1239–1282, 2007.
- [19] E. Vaughan. Flagmatic. <http://www.maths.qmul.ac.uk/~ev/flagmatic>, 2012.