The Amazing Power of Randomness: NP=RP

András Faragó
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas
farago@utdallas.edu

Abstract

We (claim to) prove the extremely surprising fact that **NP = RP**. *It is achieved by creating a* **Fully Polynomial-Time Randomized Approximation Scheme (FPRAS)** *for approximately counting the number of independent sets in bounded degree graphs, with any fixed degree bound, which is known to imply* **NP = RP**. *While our method is rooted in the well known Markov Chain Monte Carlo (MCMC) approach for approximate sampling, we overcome the notorious problem of* **slow mixing** *by a new idea for generating a random sample from among the independent sets. A key tool that enables the result is a solution to a novel sampling task that we call* **Subset Sampling**. *In its basic form, a stationary sample is given from the (exponentially large) state space of a Markov chain, as input, and we want to transform it into another stationary sample that is conditioned on falling into a given subset, which is still exponentially large.*

In general, Subset Sampling can be both harder and easier than stationary sampling from a Markov chain. It can be harder, due to the conditioning on a subset, which may have more complex structure than the original state space. But it may also be easier, since a stationary sample is already given, which, in a sense, already encompasses "most of the hardness" of such sampling tasks, being already in the stationary distribution, which is hard to reach in a slowly mixing chain. We show that it is possible to efficiently balance the two sides: we can capitalize on already having a stationary sample from the original space, so that the complexity of confining it to a subset is mitigated.

We prove that an efficient approximation is possible for the considered sampling task, and then it will be applied recursively to create the FPRAS. Two key enabling principles of proving the properties of Subset Sampling are:

(1) "Concentration is almost as good as independence:" independent random variables can be well simulated for this purpose by dependent ones, if a certain function of the latter has good concentration properties,

(2) "Virtual proliferation:" we can gain surprising extra benefits if instead of handling a single random sequence, we consider an array of statistically independent versions of it, which is only virtual in the sense that the whole array is not actually available, we only have a single run. After exploring the structure, and carrying out a kind of structured sampling in the arising virtual matrix, we then return to the single-sequence case by picking one of the rows randomly, which will be statistically indistinguishable from the original, producing the same sample distribution. This approach of "move into a richer virtual domain, do something there, and then map it back to the real world" will serve as a key enabling principle for the proof. Of course, the algorithm itself stays fully in the "real world:" it only manipulates the actually available sequence.

1 Introduction

It has been well known for decades that the approximately uniform random sampling of combinatorial structures and the approximate counting of these structures are closely related, see e.g., Jerrum, Valiant, and Vazirani [19], Sinclair and Jerrum [37]. Moreover, it is also well known that efficient algorithms for such tasks, if exist, can imply a dramatic collapse of complexity classes.

Regarding random sampling, generating a uniform random sample from a family that is exponentially large in terms of its defining description is a hard problem for many such sets. A characteristic example is drawing a random independent set from a graph. It is hard in the following sense (Sinclair [36]): if we can generate a perfectly uniform random sample from the independent vertex sets of any graph in probabilistic polynomial time, then $\mathbf{NP} = \mathbf{RP}$.

The above hardness result is not difficult to prove (see Sinclair [36], Theorem 1.17), but it leads to other, more challenging questions. What if the graph is restricted to a special class, for example, graphs with a bounded maximum degree? Such a restriction is motivated by the hope that it may be easier to carry out the sampling in a simpler graph. What if we are satisfied with a random sample that is only approximately (not perfectly) uniform? How good approximation can be achieved efficiently, and in which classes of graphs? In many cases it is known that the task still remains hard, but the analysis is much more involved. We quote specific results about this in Section 5.

The standard and well studied approach to random sampling problems is the Markov Chain Monte Carlo (MCMC) method; for surveys see, e.g., Jerrum and Sinclair [16], Randall [35], Levin, Peres, and Wilmer [23], Miklós [31]. It works by designing an appropriate Markov chain on the set from which we want to sample, such that the stationary distribution of the chain is some desired distribution (typically uniform) and the chain is defined by simple local rules, which is usually easily achievable. Then, if the Markov chain is mixing fast enough, i.e., it approaches the stationary distribution rapidly (which may be challenging to prove), we can obtain an almost uniform random sample by running it for a polynomially bounded time.

Building on this approach has resulted in notable successes that provided polynomial-time randomized approximation schemes for a number of important problems, such as approximating the volume of convex bodies (Dyer, Frieze and Kannan [10], Kannan, Lovász and Simonovits [20, 26]), approximately counting perfect matchings or, equivalently, approximating the permanent of a 0-1 matrix (Jerrum and Sinclair [17]), and a good number of others. In hard cases (the ones that would yield $\mathbf{NP} = \mathbf{RP}$), however, this approach runs into the severe obstacle of *slow mixing,* that is, the Markov chain may require exponentially long time to get close enough to its stationary distribution. Specifically, this is the case for independent vertex sets in graphs. Dyer, Frieze and Jerrum [9] proved (unconditionally) that in graphs with maximum degree at least 6, *any* Markov chain on the independent sets has exponential mixing time in the worst case, given that the chain is defined with sufficient locality[1].

It is worth noting that rather than only approaching the stationary distribution of a Markov chain, it is possible to *perfectly* reach it in finite time. That is, there are algorithms that can generate a *precisely* stationary sample (without knowing or computing the distribution). Examples of such *perfect sampling* algorithms are the Coupling From the Past method of Propp and Wilson [34], Fill's algorithm [12] and the exact mixing algorithm of Lovász and Winkler [27]. Unfortunately, however, for such slowly mixing Markov chains, as the ones used for independent sets in graphs,

[1] What they require is that the Markov chain changes the status of at most 35% of vertices in a single step. This is satisfied by all known Markov chains that are used for this problem.

all the known perfect sampling methods need exponential time in the worst case.

The task of (near-)perfect sampling is a key ingredient in our solution, but we give a significant new twist to it. We "shift" the hardness by assuming that the Markov chain is *already in the stationary distribution*. Of course, without any other condition, this would make the sampling task trivial. But we also require that the output sample must be restricted to a given (not too small) *subset* of the state space that we call *target set.* Thus, in this task of *Subset Sampling*, we want to transform a stationary sample into another one that is conditioned on falling in the target set. This shifts the hardness in the sense that the hardness of reaching stationarity is already "contained" in the input, we only want to somehow "corner it" into the target set.

Let us mention that if we could *perfectly* accomplish subset sampling for slowly mixing chains, then it would allow an easy recursive algorithm for efficient random sampling from the independent vertex sets of a graph, which would yield **NP** = **RP** (see Sinclair [36]). One could start with an edgeless graph, where the task is trivial, and gradually restrict the family of allowed independent sets to smaller and smaller subfamilies, by adding the edges one by one, until the desired graph is reached. If we could always *perfectly* accomplish the "cornering" of the stationary sample to a target set, then a direct efficient solution would result. Unfortunately, we are not that lucky, as we can only approximate the perfect subset sample, which requires significant effort to keep the accumulating errors under control.

Note that the naive approach to Subset Sampling "run the chain, starting from the stationary input sample, until the target set is hit" may create a strongly biased output. A simple example is a random walk on a graph that is a large circle, and the target set consists of a path of 3 consecutive vertices along the circle. Let us start the walk from a vertex chosen uniformly at random. If the initial state does not happen to fall in the 3-node path (which is very likely if the circle is large), then we can only hit the target set first at one of its end-vertices. As a result, the first hitting place will be far from uniformly distributed on the target set, despite to starting the chain from its uniform stationary distribution.

A natural next approach to try could be designing more sophisticated *stopping rules*, rather than just the first hit of the target set. For example, we could say "run the chain from the stationary initial state until the empirical distribution of hits in the target set is close enough to the conditioned stationary distribution." Unfortunately, however, no stopping rule is known that would carry out Subset Sampling efficiently in general for slowly mixing Markov chains. For results about stopping rules see, e.g., Lovász and Winkler [25, 28].

In view of all the negative results regarding hard sampling problems, and considering the unlikely consequences, it seems rather hopeless to efficiently accomplish Subset Sampling perfectly, and in full generality. Nevertheless, it turns out, surprisingly enough, that if we are satisfied with a near-perfect subset sample, and with a special target set, then it is still solvable efficiently, even in case of a slowly mixing chain, on a state space that is exponentially large in terms of its description. This will be enough to allow us to prove **NP** = **RP**.

A key observation that makes the proof possible is that independent random variables can be simulated by dependent ones for this purpose, if a certain function of them exhibits good concentration properties. We could phrase this phenomenon as "concentration is almost as good as independence." Another useful proof principle is what we call "virtual proliferation:" we can gain surprising extra benefits if instead of handling a single random sequence, we consider an array of statistically independent versions of it, which is only virtual in the sense that the whole array is not actually available. After exploring the structure, and carrying out a kind of structured sampling in the arising virtual matrix, we then return to the single-sequence case by picking one of the rows

randomly, which will be statistically indistinguishable from the original, producing the same sample distribution, as if the sampling algorithm is carried out on the original input sequence. Note that this "virtual proliferation" is a proof principle for analyzing the arising sample distribution. The algorithm itself only manipulates the actually available sequence.

Since we solve Subset Sampling only approximately, a great deal of effort has to go into estimating the errors, and their accumulation. To this end, a distance measure of probability distributions called *separation distance* will play an important role, as some of its properties, which we prove in Lemma 1, are surprisingly helpful.

As a final introductory note, we do not claim that our proof is the simplest possible, and the algorithm is the fastest possible.. Rather, our goal is to present all proofs in more details than what is customary in most papers, to make the checking of correctness as feasible as possible, to make sure that the main claim (**NP**=**RP**) indeed holds. Further, we also spell out the details of the numerical bounds required for the parameters, rather than relying on easier big-O notations. We do not claim that the parameters are optimal; our goal was merely to exhibit *specific* values, to demonstrate that they are not astronomically large.

2 Preliminaries

The ranges of random variables are always finite in this paper, so we do not repeat this with each statement and definition. The probability distribution of a random variable ξ with range S is represented by a finite dimensional vector. For such a vector α, its components are $\alpha(x) = \Pr(\xi = x)$, as x ranges over S in some order. The actual ordering will be irrelevant, so we just assume any fixed ordering. The inequality $\alpha \geq \beta$ means $\alpha(x) \geq \beta(x)$ for every $x \in S$. If $H \subseteq S$, then $\alpha(H) = \sum_{x \in H} \alpha(x)$.

For graphs the standard notation $G = (V, E)$ is used, where $V = V(G)$ is the set of vertices and $E = E(G)$ is the set of edges. If G, G' are graphs, then $G \subseteq G'$ means $V(G) = V(G')$ and $E(G) \subseteq E(G')$. That is, G, G' share the same vertices, but G' possibly has additional edges beyond the ones in G. The set of edges adjacent to a vertex v is denoted by $S(v)$ (the star of v), and the set of vertices adjacent to v is denoted by $\Gamma(v)$ (the neighborhood of v). If there are several graphs in the context, then the one we refer to is added as a subscript: $S_G(v)$ and $\Gamma_G(v)$.

For a bipartite graph, its *bi-adjacency matrix* is a matrix $[a_{ij}]$ in which each row represents a vertex on one side of the bipartition, say, left side, each column stands for a vertex on the right side, and $a_{ij} = 1$ if left-side vertex v_i is connected by an edge to right-side vertex v_j, otherwise $a_{ij} = 0$. A complete bipartite graph is one for which all entries in this matrix are 1. The bipartite graph is called *balanced* if its bi-adjacency matrix is a square matrix, i.e., there is an equal number of vertices on the two sides of the graph.

Another notation that we often apply: if N is a positive integer, then $[N]$ denotes the set $\{1, \ldots, N\}$.

3 Probabilistic Tools

3.1 Separation Distance of Probability Distributions

Since we are going to deal with a lot of approximations, it will be important to measure how different two probability distributions are. There are many ways to measure the the deviation

between probability distributions; for a comprehensive survey see, e.g., Deza and Deza [7]. For our purposes the concept known as *separation distance* will serve the best. It was used by Aldous and Diaconis [1] in the analysis of Markov chains; we define it here slightly differently.

Definition 1 (Separation distance) *For any two probability distributions α, β on the same set S, their* separation distance*, denoted by $s(\alpha, \beta)$, is defined as*

$$s(\alpha, \beta) = \min\{\varepsilon \geq 0 \mid \alpha \geq (1-\varepsilon)\beta\}. \tag{1}$$

For example, $s(\alpha, \beta) = 0.01$ means that for every $x \in S$ we have $\alpha(x) \geq 0.99\beta(x)$, but for any $c > 0.99$ there is an $x \in S$ with $\alpha(x) < c\beta(x)$.

If $s(\alpha, \beta) \leq \mu$ for some $\mu \geq 0$, then we say α is μ-close to β. In particular, if $\beta = U$, where U is the uniform distribution on S, then we say that α is μ-close to uniform, or μ-uniform, for short. A random variable with a μ-uniform distribution will be called a μ-uniform random variable.

Remark: Aldous and Diaconis [1] defined the same distance concept by

$$s(\alpha, \beta) = \max_x \left(1 - \frac{\alpha(x)}{\beta(x)}\right). \tag{2}$$

It is easy to see that the two definitions are equivalent[2], assuming that in (2) we take care of the case $\beta(x) = 0$. (For example, setting $\alpha(x)/0 = \infty$.) We prefer formula (1), because it remains well defined without paying attention to the possibility of $\beta(x) = 0$, or $\beta(x) = \alpha(x) = 0$.

Let us mention that the separation distance is not a metric in the usual sense, since it is not symmetric. It can also significantly differ from the well known total variation distance. For example, let $S = \{1, \ldots, n\}$, $\alpha(i) = 1/(n-1)$ for $i \leq n-1$, and $\alpha(n) = 0$. Let U be the uniform distribution on S. Then $s(\alpha, U) = 1$, while their total variation distance is $1/n$. In fact, $s(\alpha, \beta)$ is always an upper bound on the total variation distance, see [1].

In the next lemma we summarize some useful properties of the separation distance, based on Definition 1.

Lemma 1 *For arbitrary probability distributions on the same set S, their separation distance has the following properties:*

(i) ***Normalized Value:*** $0 \leq s(\alpha, \beta) \leq 1$.

(ii) ***Self-Identity:*** $s(\alpha, \beta) = 0$ *holds if and only if* $\alpha = \beta$.

(iii) ***Triangle Inequality:*** $s(\alpha, \beta) + s(\beta, \gamma) \geq s(\alpha, \gamma)$.

(iv) ***Mixture Representation:*** *For any $0 \leq \varepsilon \leq 1$, the inequality $s(\alpha, \beta) \leq \varepsilon$ holds if and only if there exists a probability distribution γ on S, satisfying*

$$\alpha = (1-\varepsilon)\beta + \varepsilon\gamma.$$

Furthermore, if $\varepsilon > 0$, then γ is uniquely determined by $\alpha, \beta, \varepsilon$. If $\varepsilon = 0$, then γ is arbitrary.

(v) ***Conditioning:*** *Let ξ be a random variable with distribution α on S. Let E be an event with $\Pr(E) \geq 1 - \varepsilon$ for some $0 \leq \varepsilon < 1$. Further, let α^E be the conditional distribution of ξ, given E, that is, $\alpha^E(x) = \Pr(\xi = x \mid E)$. Then $s(\alpha, \alpha^E) \leq \varepsilon$ and $s(\alpha^E, \alpha) \leq \varepsilon/q_0$, where q_0 is the smallest strictly positive value of $\alpha(x)$, i.e., $q_0 = \min\{x \mid x \in S, \alpha(x) > 0\}$.*

[2]We omit the formal proof of this equivalence, as we are not going to use it.

(vi) ***Restriction:*** *Let α_H be the distribution α conditioned on the event that the random variable falls in a set H, with $\alpha(H) > 0$, that is, $\alpha_H = \frac{\alpha}{\alpha(H)}$. Similarly, let $\beta_H = \frac{\beta}{\beta(H)}$, with $\beta(H) > 0$. Then*

$$s(\alpha_H, \beta_H) \leq s(\alpha, \beta) \left(1 + \frac{1}{\beta(H)}\right).$$

(vii) ***Convexity:*** *For $p_1, \ldots, p_n \geq 0$, $\sum_{i=1}^n p_i = 1$, and for arbitrary probability distributions $\alpha_1, \ldots, \alpha_n, \beta$ on S,*

$$s\left(\sum_{i=1}^n p_i \alpha_i,\ \beta\right) \leq \sum_{i=1}^n p_i s(\alpha_i, \beta).$$

Proof: See Appendix A.

3.2 Selecting Entries from a Random Sequence

Below we describe our sampling task in a general context, not restricted to Markov chains only.

3.2.1 Random Entries, H-Restricted Samples

We consider the following type of random sequences:

Definition 2 (Equidistributed random sequence) *A sequence of random variables is called* equidistributed*, if each entry of the sequence has the same distribution.*

A special example is a sequence of independent, identically distributed (i.i.d.) random variables[3]. Another special example is what is called exchangeable sequence, in which any sub-sequence of k entries have the same joint distribution, no matter which k entries are chosen, and this holds for any k (note that it does not imply independence). Yet another example is a Markov chain in stationary distribution. The latter will be our key application, but we develop some probabilistic tools for more general equidistributed sequences, without assuming any of the above special properties. These tools may also be of independent interest.

Let $X = (X_1, \ldots, X_T)$ be an equidistributed sequence of random variables, with common range S and common distribution π. Let us select a *random entry* of the sequence. This means, pick an index $\kappa \in \{1, \ldots, T\}$ uniformly at random, independently of X; then X_κ is called a random entry of the sequence. The distribution of X_κ is the mixture of the distributions of the individual entries with equal weights. Since, by assumption, each has distribution π, therefore, the random entry also has distribution π.

Let us now consider such a random entry, *conditioned* on the event that it falls in a subset $H \subseteq S$ (we assume $\pi(H) > 0$). We call this conditioned random entry an ***H-restricted sample.*** Note that, given X as input, an H-restricted sample may not be directly available to us, since if we select a random entry from the sequence, then it may not fall in H. Of course, we might try to somehow force it to fall in H, for example, we may make the selection dependent on the sample value, such as picking the first hit to H. Such forcing, however, may badly distort the distribution of the sample when the random variables are dependent.

[3]The reason we use the name *equidistributed,* rather than identically distributed, is that the latter could be easily confused with i.i.d.

Let the distribution of an H-restricted sample be denoted by π_H. It follows directly from the definition, that for any $a \in H$

$$\pi_H(a) = \Pr(X_\kappa = a \,|\, X_\kappa \in H) = \frac{\Pr(X_\kappa = a,\, X_\kappa \in H)}{\Pr(X_\kappa \in H)} = \frac{\Pr(X_\kappa = a)}{\Pr(X_\kappa \in H)} = \frac{\pi(a)}{\pi(H)}, \tag{3}$$

where we used that $\{X_\kappa = a\} \subseteq \{X_\kappa \in H\}$ for $a \in H$. Thus, the distribution of an H-restricted sample is just the common distribution π of the sequence, conditioned on H.

A problem that we need to solve to derive our main result is the task we refer to as ***Subset Sampling***: given a random realization of the sequence X, find an efficient algorithm to produce an H-restricted sample, i.e., a sample that is distributed according to π_H. In other words, can we somehow force a random entry to fall in H, *without* (significantly) distorting its distribution on H?

3.3 Random Hits

As illustrated in the Introduction by a simple example, the naive approach of taking the *first hit* to H can produce a very biased subset sample. What if, however, if we "smooth out" the hit by picking it randomly among the H-hits in a longer sequence? Although the first hit, or perhaps even the k^{th} hit for any fixed k, may show strong bias, they may average out in a random choice, so an *average hit* might give us a good approximation of a subset sample.

Consider the following sampling rule: let us *randomly* pick one of the entries among those that fall in H. We call the result a ***random H-hit.*** Let σ denote the distribution of the random H-hit, given that it exists, i.e., given that H is hit at least once. We want to compare σ and π_H, hoping to find useful cases when they are similar. It is worth noting that the key algorithmic difference between the H-restricted sample and the random H-hit is that the first is not directly available to us, while the random H-hit can be easily generated by simply picking one of the H-hits uniformly at random, from among all H-hits in the sequence.

A first hope could be that σ might serve as a good approximation of π_H, so this way we could accomplish Subset Sampling, at least approximately. Unfortunately, in general, this hope is false: without any further condition, σ is not a good approximation of π_H, they can be far from each other. Here is a simple example:

Set $S = \{1, 2, 3\}$ and $H = \{1, 2\}$. Let

$$A = (\underbrace{1, \ldots, 1}_{n}, \underbrace{2, \ldots, 2}_{n}), \quad B = (1, \underbrace{3, \ldots, 3}_{2n-1})$$

be two fixed sequences. Define a random sequence $Y = \xi A + (1 - \xi)B$, where ξ is a Bernoulli random variable with $\Pr(\xi = 0) = \Pr(\xi = 1) = 0.5$ (a coin flip), and then take an independent random permutation of Y to obtain the sequence X. One can directly see that X is equidistributed (due to the random permutation), and if π denotes the common distribution of its entries, then $\pi_H(1) = (n + 1)/(2n + 1) \to 1/2$ when $n \to \infty$, while $\sigma(1) = 3/4$. Thus, even asymptotically, the difference between π_H and σ does not vanish in this example.

Then one may wonder: are there special conditions that make the random H-hit a good approximation of the H-restricted sample? As a first important tool, we prove that to bound the separation distance between σ and π_H it is enough to assume that the number ν of H-hits is well concentrated. Interestingly, it turns out that only the concentration of ν matters, the dependencies among the random entries of the sequence can be completely arbitrary. To quantify the concentration of ν, we introduce the following definition.

Definition 3 (Concentration) *A random variable ν is called (ρ, ε)-concentrated, if*

$$\Pr\big((1-\rho)\mathrm{E}(\nu) \le \nu \le (1+\rho)\mathrm{E}(\nu) \big) \ge 1-\varepsilon$$

holds, where $\rho \ge 0$, $\varepsilon \in [0,1]$, and $\mathrm{E}(\nu)$ denotes the expected value of ν.

While this will lead to a useful and non-trivial tool (see the details in Section B1 of Appendix B), it is not sufficient for our main task. The reason is that the strong enough concentration of the number of H-hits, unfortunately, does not hold in the application where we need it (slowly mixing Markov chains on the independent sets of a graph). Yet, it helps in the development of a more powerful tool, which is based on the new concept of *p-robust subset.* This concept serves two purposes. First, it makes sure that the existence of an H-hit in a random sequence is sufficiently likely. More importantly, it can act as a "quasi-substitute" for independence, without *a priori* assuming any kind of concentration. Here is the formal definition:

Definition 4 (p-robust subset) *Let $X = (X_1, \ldots, X_T)$ be a sequence of random variables, with a common range S, and let $p \in (0,1)$ be a real number. A nonempty subset $H \subseteq S$ is called a* p-robust subset (for X), *if each entry of X falls in H with probability at least p, and this remains so even if we condition on arbitrary values of any preceding entries. Formally, $\forall t \in [T]:\ \Pr(X_t \in H) \ge p$, and for any $k \in [T-1]$, $t, t_1, \ldots, t_k \in [T]$ with $t > t_1 \ldots > t_k \ge 1$*

$$\Pr(X_t \in H \,|\, X_{t_1} = a_{t_1}, \ldots, X_{t_k} = a_{t_k}) \ge p$$

holds for any $a_{t_1}, \ldots, a_{t_k} \in S$, whenever $\Pr(X_{t_1} = a_{t_1}, \ldots, X_{t_k} = a_{t_k}) > 0$.

In words, a set $H \subseteq S$ is p-robust if every X_t falls in H with probability at least p, regardless of any previous history. Note that it does not mean X_t is independent from the preceding entries, as the actual probability that X_t falls into H may depend on them, just it never drops below p. In this special sense, it is robust against the effect of history. It turns out that this property, which holds in our application, allows an approximation of independence, regarding the events $\{X_t \in H\}$, $t = 1, \ldots, T$, and this will do a very valuable service.

Now we are ready to present a theorem, which is the "heart" of our approach, and may be of independent interest, too. Let us emphasize again that it assumes neither any kind of concentration, nor independence, not even Markovian dependence, about the considered sequence.

Theorem 1 (Subset Sampling Theorem) *Let $X = (X_1, X_2, \ldots, X_T)$ be an equidistributed sequence of random variables, with common range S, and common distribution π. Further, let $H \subseteq S$ be a p-robust subset for X. Pick a sample Z from the sequence X, as follows: if there is an H-hit in X, then select Z as a random H-hit from X. If there is no H-hit in X, then set $Z = z_0$, where $z_0 \in H$ is any fixed element of H. Let α denote the distribution of Z, and let $\pi_H = \pi/\pi(H)$ denote the distribution π conditioned on H, i.e., the distribution of the H-restricted sample. Then*

$$s(\alpha, \pi_H) \ \le\ \frac{2}{n^{k/3}}$$

holds, where n, k are integer parameters that satisfy the following conditions:

$$T = n^k, \quad |S| \le 2^n, \quad \min_{x \in S} \pi(x) \ge 2^{-n}, \quad p \ge 1/n, \quad k \ge 7, \quad n \ge 3k^{1/5}.$$

Proof: See Appendix B. The proof builds on the concentration based bound, further complementing it with other tools that we develop, eventually achieving that the need for concentration is eliminated from the final result (but, interestingly, it is used in the proof). Note that the overall proof is not easy, with the accompanying explanations it fills approximately 19 pages.

4 FPRAS for Counting Independent Sets in Bounded Degree Graphs

Before presenting the actual FPRAS, we need to develop some more tools.

4.1 A Markov Chain on the Independent Sets of a Graph

Let us introduce a Markov chain defined on the independent sets of a graph $G = (V, E)$. We call it *Basic Insert/Delete Chain (BIDC)*. It is similar to what is called Insert/Delete Chain by Dyer and Greenhill in [11][4]. We distinguish our version with the adjective "basic", as we do not need the additional parameter λ (sometimes referred to as *fugacity*, which weighs a set of size k with k^λ), and also we will not need to keep the chain in the current state with probability $\geq 1/2$.

The transition probabilities of our Markov chain are given as follows. Let $\mathcal{I} = \mathcal{I}(G)$ denote the family of independent sets of G (including the empty set). Let $X_t \in \mathcal{I}$ be the state of the chain at time t. (The time is discrete, so t is assumed an integer.) The next state $X_{t+1} \in \mathcal{I}$ is obtained by the following rule.

Basic Insert/Delete Chain (BIDC) transition rule

1. Draw a vertex $u \in V$ uniformly at random.
2. Set

$$X_{t+1} = \begin{cases} X_t - \{u\} & \text{if} \quad u \in X_t \\ X_t \cup \{u\} & \text{if} \quad u \notin X_t \text{ and } X_t \cup \{u\} \in \mathcal{I} \\ X_t & \text{otherwise.} \end{cases}$$

BIDC can be interpreted as a random walk on an undirected graph $\widetilde{G}$, with $V(\widetilde{G}) = \mathcal{I}(G)$. As each vertex of $\widetilde{G}$ represents an independent set of G, therefore, $\widetilde{G}$ is typically exponentially large in terms $|V| = n$. Two different vertices X, Y are connected by an edge in $\widetilde{G}$ if and only if $|X \Delta Y| = 1$, where Δ denotes symmetric difference. There are also self-loops, making the graph regular of degree n, since u can be chosen n different ways (a self-loop contributes 1 to the degree). This graph is always connected, since all independent sets can be reached from $\emptyset$. It follows from standard results on Markov chains and random walks on graphs (see, e.g., the survey by Lovász [24]) that this Markov chain has a unique stationary distribution, and it is the uniform distribution on $\mathcal{I}(G)$. On the other hand, the mixing time is exponentially large in n in the worst case, if the graph has maximum degree at least 6, as implied by the results of Dyer, Frieze and Jerrum [9].

4.2 Extending a Random Independent Set to a Somewhat Larger Graph

A key tool in our solution is a method to transform an almost uniform random independent set over a graph into another almost uniform random independent set over a graph that has the same vertices, but it is slightly larger in terms of edges: it has a bounded number of extra edges, all adjacent to the same vertex. We call the algorithm *Random Independent Set Lifting (RISL)*.

[4]Related models are also frequently investigated in the statistical physics literature, under various names, such as hard-core lattice gas model, Glauber dynamics, Potts model, Ising model, etc. Our intent, however, is to abstract away from the statistical physics background, and focus rather on the graph theoretic connection, with as precise proofs as possible.

We use the following notation: if $G = (V, E)$ is a graph and $v_0 \in V$ is a vertex in it, then $S(v_0)$ stands for the set of edges adjacent to v_0 in G, and $G - S(v_0)$ denotes the graph obtained by removing the edges in $S(v_0)$ from G. The vertex v_0 then remains an isolated vertex in $G - S(v_0)$.

Recall that a random variable is called μ-uniform, for some $\mu \in [0, 1]$, if its separation distance is at most μ from the uniform distribution. Thus, a random independent set with distribution σ is μ-uniform over $\mathcal{I}(G)$, if $s(\sigma, U) \leq \mu$, where U is the uniform distribution on $\mathcal{I}(G)$. Now we address the following question: given a graph G, and a vertex $v_0 \in V(G)$, how to transform a μ-uniform independent set over $\mathcal{I}(G - S(v_0))$ into a μ'-uniform independent set over $\mathcal{I}(G)$, so that μ' is close to μ, there is only a small additional deviation from uniformity.

Informally, our strategy to efficiently solve the outlined transformation problem consists of two steps: (1) Given the μ-uniform independent set over $\mathcal{I}(G - S(v_0))$, we run the BIDC Markov chain from it as initial state, to obtain a trajectory $Y = (X_1, \ldots, X_T)$. (2) Then we apply Theorem 1 to pick a subset sample from the target set $H = \mathcal{I}(G) \subseteq \mathcal{I}(G - S(v_0))$. This sample will serve as the desired μ'-uniform independent set over $\mathcal{I}(G)$. With a careful analysis of the errors, we obtain that μ' will be only only slightly larger than μ. Now we can formally present the algorithm. The parameter k in the description is an integer constant, not part of the input.

Algorithm *Random Independent Set Lifting (RISL)*

Input: a graph $G = (V, E)$, a specified vertex $v_0 \in V$, and an independent set $X \in \mathcal{I}(G - S(v_0))$. The algorithm also uses a parameter k, which is a fixed constant.

Output: an independent set $Z \in \mathcal{I}(G)$.

The Algorithm

Step 1 *(Trajectory construction)* Run BIDC on $\mathcal{I}(G - S(v_0))$ from initial state X. Let $Y = (X_1, \ldots, X_T)$ be the obtained trajectory, where $X_1 = X$, $T = n^k$, $n = |V|$.

Step 2 *(Subset sampling)* Set $H = \mathcal{I}(G)$. If Y contains an H-hit, then select an H-hit Z from Y uniformly at random. If there is no H-hit in Y, then set $Z = \emptyset$. Output Z and HALT.

Theorem 2 *Let k be a fixed positive integer constant with $T = n^k$ in the algorithm RISL, and $n = |V(G)|$. Then the algorithm has the following properties:*

(i) *The running time is $O(n^{k+1})$, using $O(n^k \log n)$ random bits.*

(ii) *Assume the input X is a μ-uniform random independent set over $\mathcal{I}(G - S(v_0))$, for some $\mu \in [0, 1)$. Then the output Z is a μ'-uniform random independent set over $\mathcal{I}(G)$, with*

$$\mu' \leq \mu + \frac{2}{n^{k/3}},$$

whenever the parameters satisfy the following conditions: $n \geq 3k^{1/5}$, $k \geq 7$, and $n \geq \Delta + 2$, where Δ is any upper bound on the maximum degree of G.

(iii) *The algorithm receives G, v_0, X as input, and it also uses the constant k. But the algorithm has no access to the parameter Δ, and the running time is also independent of Δ, assuming that the conditions listed in* (ii) *hold.*

Proof: See Appendix C.

4.3 Independence Ratio

The independence ratio of a vertex expresses that what fraction of independent sets contain the vertex, from a given family of independent sets. Since later we are going to allow that the sets in the family are chosen randomly, therefore, it can happen that the same set occurs more than once. Therefore, in this context, we use the following notational convention:

Notational convention. By a *family* of independent sets we mean a *multiset*, in which the same element can occur multiple times. Note that multisets are routinely used in many contexts. According to Knuth [21], the name multiset was coined by de Bruijn in the 1970s, but the concept has been around for centuries. The notation $\mathcal{F} \sqsubseteq \mathcal{I}(G)$ will mean that each element of $\mathcal{F}$ is an independent set in the graph G, but $\mathcal{F}$ may possibly contain the same set with multiplicity higher than one. Furthermore, the cardinality $|\mathcal{F}|$ takes the multiplicities into account. At the same time, $\mathcal{I}(G)$ will still denote the *set* of all independent sets, in the usual sense (without multiplicities).

Definition 5 (Independence ratio of a vertex) *Let $G = (V, E)$ be a graph, and $\mathcal{F} \sqsubseteq \mathcal{I}(G)$ be multiset of independent sets in G. The* $\mathcal{F}$-independence ratio of v, *denoted by $p(v \,|\, \mathcal{F})$, is defined as the fraction of sets in $\mathcal{F}$ that contain v, that is,*

$$p(v \,|\, \mathcal{F}) = \frac{|\{F \in \mathcal{F} \,|\, v \in F\}|}{|\mathcal{F}|},$$

where both the numerator and the denominator refer to cardinalities of multisets. If $\mathcal{F}$ contains each independent set of G precisely once, i.e., $\mathcal{F} = \mathcal{I}(\mathcal{G})$, then we simply call the above quantity independence ratio *and denote it by $p(v)$.*

Remark: Observe that $p(v)$ is precisely the probability that vertex v falls into an independent set that is chosen uniformly at random from $\mathcal{I}(G)$.

The next concept captures the possibility that the independence ratio $p(v)$ may be approximated by $\mathcal{F}$-independence ratio $p(v \,|\, \mathcal{F})$, via restricting ourselves to some multiset $\mathcal{F} \sqsubseteq \mathcal{I}(G)$. The important case for us will be when $|\mathcal{F}| \ll |\mathcal{I}(G)|$.

Definition 6 (δ-approximator) *Let $G = (V, E)$ be a graph, $\mathcal{F} \sqsubseteq \mathcal{I}(G)$ be a multiset of independent sets in G, and $\delta \in (0, 1)$. We say that $\mathcal{F}$ is a* δ-approximator *of $\mathcal{I}(G)$, if*

$$(1 - \delta)p(v) \leq p(v \,|\, \mathcal{F}) \leq (1 + \delta)p(v) \tag{4}$$

holds for every $v \in V$. We refer to δ as the tolerance parameter.

Remark: Due to $\emptyset \in \mathcal{I}(G)$, as well as $\{v\} \in \mathcal{I}(G)$ for every $v \in V$, we always have $0 < p(v) < 1$ for all $v \in V$. In the next lemma we show that more is true: $p(v) \leq 1/2$ always holds, and if the graph has bounded maximum degree, then $p(v)$ is bounded from below by a positive *constant*, which is independent of the size of the graph.

Lemma 2 *Let $G = (V, E)$ be a graph with maximum degree $\leq \Delta$. Then for every $v \in V$ the following holds:*

$$\frac{1}{2^{\Delta} + 1} \leq p(v) \leq \frac{1}{2}.$$

Proof. Let $\mathcal{I}_v \subseteq \mathcal{I}(G)$ be the set of independent sets that contain v, and $\overline{\mathcal{I}}_v = \mathcal{I}(G) - \mathcal{I}_v$. Let $\Gamma(v)$ be the neighborhood of v in G, and let $A_v = V - \Gamma(v) - \{v\}$. For any $H \in \overline{\mathcal{I}}_v$ we can assign an independent set $F(H) \in \mathcal{I}_v$ by $F(H) = (H \cap A_v) \cup \{v\}$. Moreover, every set $B \in \mathcal{I}_v$ arises this way, since $B = F(B - \{v\})$. Therefore, the mapping $F : \overline{\mathcal{I}}_v \mapsto \mathcal{I}_v$ is onto, implying that at most half of all independent sets can belong to $\mathcal{I}_v$, i.e., $p(v) \leq 1/2$. On the other hand, the mapping F may not be 1-1, since different sets $H_1, \ldots, H_k \in \overline{\mathcal{I}}_v$ may have $F(H_1) = \ldots = F(H_k)$. By the definition of F, however, this can only happen, if $H_1 \cap A_v = \ldots = H_k \cap A_v$. That means, $H_1, \ldots, H_k$ can only differ within $\Gamma(v)$. Since $|\Gamma(v)| \leq \Delta$, therefore, there can be at most 2^Δ different sets in it, so $k \leq 2^\Delta$ must hold. Thus, F is a "$\leq 2^\Delta$-to-1" mapping from $\overline{\mathcal{I}}_v$ onto $\mathcal{I}_v$. Consequently, at least the $1/(2^\Delta + 1)$ fraction of all independent sets must belong to $\mathcal{I}_v$, implying $p(v) \geq 1/(2^\Delta + 1)$.

♠

Let us now introduce the concept of a *random approximator*, which is a stochastic version of the δ-approximator. It is a *random* family of independent sets, statistically independent of each other as random variables, each having approximately uniform distribution, and they act as a δ-approximator with high probability. As an auxiliary concept, let us first define an approximately uniform i.i.d. random family of independent sets.

Definition 7 (μ-uniform i.i.d. family) *Let $\mathcal{F} = \{X_1, \ldots, X_m\} \sqsubseteq \mathcal{I}(G)$ be a nonempty family (multiset) of random independent sets in G, and $\mu \in [0, 1]$. We say that $\mathcal{F}$ is a* μ-uniform i.i.d. family *over $\mathcal{I}(G)$ if each $X_i \in \mathcal{F}$ is a μ-uniform random independent set from $\mathcal{I}(G)$, and the $X_1, \ldots, X_m$ are i.i.d., as random variables.*

Definition 8 (Random approximator) *Let $G = (V, E)$ be a graph, and μ, δ, h non-negative real numbers. We say that $\mathcal{F} \sqsubseteq \mathcal{I}(G)$ is a* (μ, δ, h)-approximator *of $\mathcal{I}(G)$, if $\mathcal{F}$ is a μ-uniform i.i.d. family over $\mathcal{I}(G)$, and for every $v \in V$ it holds that*

$$\Pr\big((1-\delta)p(v) \leq p(v \,|\, \mathcal{F}) \leq (1+\delta)p(v) \big) \geq 1 - h \tag{5}$$

4.4 Random Approximator Extension to a Slightly Larger Graph

The next result serves as a central tool: it provides a probabilistic polynomial time algorithm that extends a random approximator of $\mathcal{I}(G - S(v_0))$ to a random approximator of $\mathcal{I}(G)$, with only minor change in the parameters. The algorithm itself is a simple application of the RISL algorithm, presented in Section 4.2.

Algorithm *Random Approximator Extension (RAE)*

Input: a graph $G = (V, E)$, a specified vertex $v_0 \in V$, and a random (μ, δ, h)-approximator $\mathcal{F}_0$ of $\mathcal{I}(G - S(v_0))$, with $|\mathcal{F}_0| = m$.
The algorithm also uses a parameter k, which is a fixed constant, inherited from algorithm RISL (see Section 4.2), which is called here as a subroutine.

Output: a random (μ', δ', h')-approximator $\mathcal{F}$ of $\mathcal{I}(G)$, with $|\mathcal{F}| = m$.

The Algorithm

Step 1 Let $X_1, \ldots, X_m$ be the independent sets in $\mathcal{F}_0$. For every X_i, $i = 1, \ldots, m$, run algorithm RISL with input G, v_0, X_i. Let $Z_i \in \mathcal{I}(G)$ be the corresponding output of RISL.

Step 2 Set $\mathcal{F} = \{Z_1, \ldots, Z_m\}$ (multiset). Output $\mathcal{F}$ and HALT.

Theorem 3 *Let k be a fixed positive integer constant, used in algorithm RISL (see* Section 4.2*), whenever it is called as a subroutine. Let $n = |V(G)|$, $m = |\mathcal{F}_0|$. Then algorithm RAE has the following properties:*

(i) *The algorithm runs in $O(mn^{k+1})$ time, using $O(mn^k \log n)$ random bits.*

(ii) *Assume that $\mathcal{F}_0$ is a (μ, δ, h)-approximator of $\mathcal{I}(G - S(v_0))$. Then the output $\mathcal{F}$ of the algorithm is a (μ', δ', h')-approximator of $\mathcal{I}(G)$, with $|\mathcal{F}| = m$, and*

$$\mu' \le \mu + \frac{2}{n^{k/3}}, \qquad \delta' = \delta, \qquad h' \le \mu + \frac{3}{n^{k/3}},$$

whenever the following conditions are satisfied:

$$\delta \ge m^{-1/3}, \quad m \ge (2^\Delta + 1)^3(6 + k \ln n)^3, \quad k \ge 7, \quad n \ge 3k^{1/5}, \quad n \ge \Delta + 2,$$

where Δ is any upper bound on the maximum degree of G.

(iii) *The algorithm receives $G, \mathcal{F}_0, v_0$ as input, and it also uses the constant k. But the algorithm has no access to the parameters Δ, μ, δ, h, and the running time is also independent of these parameters, assuming that the conditions listed in* (ii) *hold.*

Remarks: (1) Observe that μ', h' can become slightly larger than μ, h, respectively, as a result of extending the random approximator, but the tolerance parameter δ remains unchanged, as long as $\delta \ge m^{-1/3}$. (2) In order to have polynomial running time in n only (rather than m and n), it is necessary that $\Delta = O(\log n)$, as otherwise the condition $m \ge (2^\Delta + 1)^3(6 + k \ln n)^3$ in (ii) would make m superpolynomial in n, leading to superpolynomial running time in n.

Proof of Theorem 3: See Appendix D.

4.5 The FPRAS

Let us first define what exactly is that we are looking for. We specify the definition of FPRAS for the case of independent sets in graphs, because that is the problem considered here, but the definition can be directly modified to other counting problems, as well.

Definition 9 (FPRAS for independent sets) *Let G be a graph with $n = |V(G)|$ and let $I(G) = |\mathcal{I}(G)|$ denote the number of independent sets in G. Further, let $\delta > 0$ be a tolerance parameter. We say that a (randomized) algorithm is a* Fully Polynomial Approximation Scheme (FPRAS) *for $I(G)$, if it satisfies the following requirements:*

(i) *Upon input G, δ, the algorithm outputs a number $\widetilde{I}$, such that*

$$\Pr\big((1-\delta)I(G) \le \widetilde{I} \le (1+\delta)I(G)v) \big) \ge \frac{3}{4} \tag{6}$$

where the probability is meant solely with respect to the internal randomization of the algorithm (the input is not random).

(ii) *The running time of the algorithm is bounded by a polynomial of n and $1/\delta$.*

Remark. The value 3/4 in (6) is somewhat arbitrary, it could be replaced by any constant in $(1/2, 1)$, so that the definition still remains equivalent with the original. In fact, it still equivalent with the following, apparently more demanding definition: given G, δ, and, in addition, also a confidence parameter ϵ, output a number $\widetilde{I}$, such that the algorithm runs in probabilistic polynomial time in $n, 1/\delta, \ln(1/\epsilon)$, and it satisfies

$$\Pr\Big((1-\delta)I(G) \leq \widetilde{I} \leq (1+\delta)I(G)v)\Big) \geq 1-\epsilon. \tag{7}$$

This, apparently more demanding definition explains why the scheme is called *fully polynomial:* the running time does not become exponential even for exponentially small ε, as the running time is polynomial in $\ln(1/\epsilon)$, rather than in $1/\varepsilon$. It turns out, however, that it is sufficient to satisfy (7) with a *constant* $\epsilon = 1/4$, which gives rise to the simpler formulation (6). The reason is that once we have an algorithm for the simpler version (6), we can boost the confidence parameter to an arbitrary given $\epsilon \in (0, 1/2)$. It can be achieved this way: given an arbitrary $\epsilon \in (0, 1/2)$, run the algorithm $k = \lceil 16 \ln(2/\epsilon) \rceil$ times with independent random bits, each time with confidence parameter $\epsilon = 1/4$, and then take the median of the obtained estimates. The median will satisfy (6) with the desired value of ϵ, while the running time increases only by a factor of $O(\ln(1/\epsilon))$, so it is polynomial in $n, 1/\delta, \ln(1/\varepsilon)$. This claim appears to be part of folklore. For a proof see, e.g., Vigoda [39]; it is a straightforward application of the Chernoff bound. Because of this equivalence, usually the simpler definition (6) is used in the literature, we also follow this tradition.

Now we are ready to present our main algorithm:

Algorithm *FPRAS for Independent Sets*

Input: a graph $G = (V, E)$, and a tolerance parameter $\delta > 0$.

Output: A number $\widetilde{I}$, serving as an approximation of $I(G)$.

The Algorithm

Step 1 Let $v_1, \ldots, v_n$ be the vertices of G. Define the graphs H_ℓ, $\ell = 1, \ldots, n$, such that H_ℓ consists of the subgraph of G induced by $v_1, \ldots, v_\ell$, plus adding the remaining $n - \ell$ vertices $v_{\ell+1}, \ldots, v_n$, as isolated vertices.

Step 2 Set $\ell = 1$, $m = \left\lceil 64n^6 \left(1 + \frac{1}{\delta}\right)^3 \right\rceil$.

Step 3 Create a random approximator $\mathcal{F}_\ell$ of $\mathcal{I}(H_\ell)$, as follows:

(a) Set $j = 1$. Pick m random independent sets in H_1 (which consists of n isolated vertices), using an independent random bit to decide for each vertex whether it is put in the random set or not. Call the resulting system $\mathcal{F}_1$.

(b) If $j < \ell$, then create $\mathcal{F}_{j+1}$ from $\mathcal{F}_j$ by running the RAE algorithm (see Section 4.4), with input graph H_j, distinguished vertex $v_0 = v_j$, and random approximator $\mathcal{F}_j$. The output is the random approximator $\mathcal{F}_{j+1}$.

(c) Set $j = j + 1$. If $j < \ell$, then go to (b). Else compute $\widetilde{q}_\ell$ by counting what fraction of independent sets in $\mathcal{F}_\ell$ contain v_ℓ in the graph H_ℓ. Go to Step 4.

Step 4 If $\ell < n$ then set $\ell = \ell + 1$, and repeat from Step 3. Else output the value

$$\widetilde{I} = 2 \prod_{\ell=1}^{n-1} \frac{1}{1 - \widetilde{q}_{\ell+1}}$$

and HALT.

Theorem 4 *The algorithm* FPRAS for Independent Sets *satisfies Definition 9 for input graphs with maximum degree bounded by an arbitrary constant. Specifically,*

(i) *Upon input graph G with $n = |V(G)|$, and tolerance parameter $\delta > 0$, the algorithm produces the output number $\widetilde{I}$ in probabilistic polynomial time in n and $1/\delta$. In particular, the algorithm has the the following complexity:*

$$\text{if } \delta \geq 1 \text{ then it runs in time } O\left(n^{k+9}\right), \text{using } O\left(n^{k+8}\log n\right) \text{ random bits;}$$

$$\text{if } \delta < 1 \text{ then it runs in time } O\left(n^{k+9}(1/\delta)^3\right), \text{using } O\left(n^{k+8}(1/\delta)^3\log n\right) \text{ random bits,}$$

where k is the fixed parameter used in algorithm RAE (see Section 4.4*), which is run here as a subroutine in* Step 3(b).

(ii) *Suppose that*

$$k \in \{7,8,9\}, \quad m \geq (2^{\Delta}+1)^3(6+k\ln n)^3, \quad \text{and} \quad n \geq \Delta+2,$$

where Δ is any constant upper bound on the maximum degree of G. Then the output $\widetilde{I}$ of the algorithm, as required by Definition 9, satisfies

$$\Pr\left(\,(1-\delta)I(G) \leq \widetilde{I} \leq (1+\delta)I(G)v)\,\right) \;\geq\; \frac{3}{4}, \tag{8}$$

assuming that he input graph has at least $n \geq n_0(k)$ vertices, where

$$n_0(k) = \begin{cases} 24 & \text{if} \quad k=9 \\ 118 & \text{if} \quad k=8 \\ 13,824 & \text{if} \quad k=7. \end{cases}$$

Furthermore, the probability in (8) *is meant solely with respect to the internal randomization of the algorithm.*

Proof: See Appendix E.

Remarks.

- The parameter k serves the role to control the trade-off between the running time and the minimum required instance size, in terms of the number of vertices. For example, if $k = 9$, then the minimum instance size is 24, and the exponent of n in the running time is 18. If we choose $k = 7$, then the exponent of n in the running time decreases by 2, but the price is that the minimum instance size grows to 13,824.

- It turns out from the proof that considering $k > 9$ does not bring further significant advantage, that is why we restricted the value of k to $\{7,8,9\}$ (less than 7 is not allowed by Theorem 1, and this propagates to all subsequent poofs).

- While we state Theorem 4 with requiring a constant bound Δ on the maximum degree, in fact the proof also works for any $\Delta = O(\log n)$. It is not needed, however, for our main goal, since the FPRAS for bounded degree graphs is already enough to prove **NP=RP**, see Section 5. Therefore, we stay with constant degree bound. Note that a maximum degree with $\Delta/\log n \to \infty$ would already lead to super-polynomial running time, due to the requirement $m \geq (2^{\Delta}+1)^3(6+k\ln n)^3$ in (ii).

- As mentioned in the Introduction, we do not claim that the numerical parameters are optimized, so further improvement in the values are likely possible. Our goal was merely to exhibit *specific* values, rather than using big-O notations for the parameters, to demonstrate that they are not astronomically large.

5 NP vs. RP

Finally, we are ready to state our main result.

Theorem 5 $\mathbf{NP} = \mathbf{RP}$.

Proof. By Theorem 4, an FPRAS exists for approximately counting the independent sets in bounded degree graphs, with any constant degree bound Δ. This is known to imply $\mathbf{NP} = \mathbf{RP}$, with at least three different proofs for it in the literature. Specifically, Luby and Vigoda [29] prove in 1997, building on an approach presented by Sinclair [36][5], that there is a constant Δ, such that the existence of the FPRAS for graphs with maximum degree $\leq \Delta$ implies $\mathbf{NP} = \mathbf{RP}$. The unspecified constant Δ is brought down to $\Delta = 25$ by Dyer, Frieze, and Jerrum [9] in 2002. Finally, Sly [38] further reduces the constant to $\Delta = 6$ in 2010. All the three papers prove, with different methods, that the existence of an FPRAS for approximately counting independent sets in graphs with the respective constant degree bound Δ implies $\mathbf{NP} = \mathbf{RP}$.

♠

6 Some Notes on Other Complexity Classes

The intriguing question still remains whether **RP** can be derandomized or not. In any case, if our result is correct, then now the two conjectures "$\mathbf{P} \neq \mathbf{NP}$" and "$\mathbf{P} = \mathbf{RP} = \mathbf{BPP}$" contradict to each other. The author's personal opinion, contrary to the majority view, is that $\mathbf{P} \neq \mathbf{RP}$, i.e., complete derandomization is not possible (which, of course, implies $\mathbf{P} \neq \mathbf{NP}$). If this is indeed true, then we have $\mathbf{P} \neq \mathbf{RP} = \mathbf{NP}$. This would also give some hint why the **P** vs. **NP** problem is so difficult:

- If our result is correct, then **NP** is *almost* equal to **P**, the *only* barrier is the need for random bits. Therefore, separating **P** and **NP** is hard, as they are *very close* to each other. Yet, if derandomization is not possible, then they are still different, so collapsing them is impossible.
- If $\mathbf{P} \neq \mathbf{RP} = \mathbf{NP}$ is actually the case, then, in a sense, it is just the opposite of what has been widely believed, which is $\mathbf{P} = \mathbf{RP} \neq \mathbf{NP}$. While the statements $\mathbf{P} \neq \mathbf{RP} = \mathbf{NP}$ and $\mathbf{P} = \mathbf{RP} \neq \mathbf{NP}$ are not formal negations of each other, they intuitively point in opposite directions. Thus, one could possibly argue that previous research, no matter how involved, could not reach the goal, because it progressed in the wrong direction.

Some consequences of $\mathbf{NP} = \mathbf{RP}$ immediately follow from existing results. For example, it is well known (proved by Zachos [40]), that $\mathbf{NP} \subseteq \mathbf{BPP}$ implies the collapse of the entire Polynomial Hierarchy to **BPP**, so our result (if correct) implies $\mathbf{PH} = \mathbf{BPP}$.

[5] Sinclair [36] attributes the idea of boosting the number of solutions to Jerrum, Valiant and Vazirani [19].

Another interesting fact is the issue of *average case complexity* of **NP**, in the sense of Levin [22], assuming the input is chosen randomly from any reasonable (polynomial-time computable) distribution. According to current knowledge, $\mathbf{P} \neq \mathbf{NP}$ alone is not known to rule out that all problems in **NP** could still be solved in deterministic polynomial time *on the average*, over random inputs. In other words, in case of $\mathbf{P} \neq \mathbf{NP}$, we could still have that **NP** is "easy on average" for deterministic algorithms. However, Buhrman, Fortnow and Pavan [6] prove that if **NP** is easy on average then $\mathbf{P} = \mathbf{BPP}$. Since, by our result, the latter would now imply $\mathbf{P} = \mathbf{NP}$, therefore, whether **NP** is hard in the worst case becomes *equivalent* to whether it is hard in the average case.

It is also interesting to look into that if $\mathbf{P} \neq \mathbf{NP}$, then, combined with the obtained collapse $\mathbf{NP} = \mathbf{RP}$, could we conclude something new about complexity class *separations*? Here is a consequence that we could call "disjunctive separation," (conditional on $\mathbf{P} \neq \mathbf{NP}$) because it proves that *at least one* of two major separations would follow. (The same has not been known to follow from $\mathbf{P} \neq \mathbf{NP}$ alone.) We use the common notations $\mathbf{E} = DTIME(2^{O(n)})$ and $\mathbf{NE} = NTIME(2^{O(n)})$.

Theorem 6 *If* $\mathbf{P} \neq \mathbf{NP}$*, then at least one of the following holds:*

(1) $\mathbf{NP} \neq co\text{–}\mathbf{NP}$

(2) $\mathbf{E} \neq \mathbf{NE}$.

Proof. By our result $\mathbf{NP} = \mathbf{RP}$ holds, which implies $\mathbf{PH} = \mathbf{BPP}$ (see Zachos [40]). Let **TALLY** be the set of all tally (unary) languages. Let us consider $\mathbf{TALLY} \cap \mathbf{PH} = \mathbf{TALLY} \cap \mathbf{BPP}$, which is nonempty, as there are tally languages in **PH**. By a result of Babai, Fortnow, Nisan and Wigderson [2] if all tally languages in **PH** are decidable in polynomial time, then $\mathbf{P} = \mathbf{BPP}$. Since $\mathbf{P} \neq \mathbf{NP}$ is assumed, therefore, by our result, $\mathbf{P} = \mathbf{BPP}$ is impossible. Consequently, it must be that $\mathbf{TALLY} \cap (\mathbf{BPP} - \mathbf{P}) \neq \emptyset$. Then $\mathbf{NP} = \mathbf{RP} \subseteq \mathbf{BPP}$ implies that at least one of $\mathbf{TALLY} \cap (\mathbf{NP} - \mathbf{P})$ and $\mathbf{TALLY} \cap (\mathbf{BPP} - \mathbf{NP})$ is nonempty. If there is a tally language in $\mathbf{NP} - \mathbf{P}$ then it is well known to imply $\mathbf{E} \neq \mathbf{NE}$ (see Book [5]). If $\mathbf{BPP} - \mathbf{NP} \neq \emptyset$, then $\mathbf{NP} \neq co\text{–}\mathbf{NP}$ must hold, since otherwise **BPP** would collapse into **ZPP**, due to $\mathbf{NP} = \mathbf{RP}$. If, however, **BPP** collapses into **ZPP**, then we get the contradiction $\emptyset \neq \mathbf{BPP} - \mathbf{NP} = \mathbf{ZPP} - \mathbf{RP} = \emptyset$.

♠

7 Conclusion

We (claim to) have proved $\mathbf{NP} = \mathbf{RP}$. The most important task is now the validation of the correctness by the expert community. Therefore, we present the proofs in the appendices in as much details as reasonably possible.

References

[1] D. Aldous and P. Diaconis, "Strong Uniform Times and Finite Random Walks", *Advances in Applied Math.*, 8(1987), pp, 69–97.

[2] L. Babai, L. Fortnow, N. Nisan and A. Wigderson, "BPP has Subexponential Time Simulations Unless EXPTIME has Publishable Proofs", *Computational Complexity*, 3(1993), pp. 307–318.

[3] J. Bather, "A Conversation with Herman Chernoff," *Statistical Science,* Vol. 11, No. 4, 1996, pp. 335–350.

[4] B. Bollobás, *Random Graphs,* Cambridge University Press, 2001.

[5] R. Book, "Tally Languages and Complexity Classes", *Information and Control*, 26(1974), pp. 186–193.

[6] H. Buhrman, L. Fortnow, and A. Pavan, "Some Results on Derandomization" *Theory of Computing Systems,* 38(2005), pp. 211–227.

[7] M.M. Deza and E.Deza, *Encyclopedia of Distances*, Springer, 2009.

[8] D.P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms,* Cambridge Univresity Press, 2009.

[9] M. Dyer, A. Frieze and M. Jerrum, "On Counting Independent Sets in Sparse Graphs", *SIAM Journal on Computing*, 31(2002), pp. 1527–1541.

[10] M. Dyer, A. Frieze, and R. Kannan, "A Random Polynomial Time Algorithm for Approximating the Volume of Convex Bodies", *Journal of the ACM*, 38(1991) pp. 1–17.

[11] M. Dyer and C. Greenhill, "On Markov Chains for Independent Sets", *Journal of Algorithms*, 35(2000), pp. 17–49.

[12] J.A. Fill, "An Interruptible Algorithm for Perfect Sampling via Markov Chains", *Annals of Applied Probability*, 8(1998), pp. 131–162.

[13] A. Frieze and M. Karoński, *Introduction to Random Graphs,* Cambridge University Press, 2016.

[14] P. Hall, "On Representatives of Subsets," *J. London Math. Soc.*, 10(1935/1), pp. 26–30.

[15] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs,* Wiley-Interscience, 2000.

[16] M. Jerrum and A. Sinclair, "The Markov Chain Monte Carlo Method: An Approach to Approximate Counting and Integration", In: D.S. Hochbaum (ed.), *Approximation Algorithms for NP-hard Problems*, PWS Publishing Co., Boston, MA, 1997, pp. 482–520.

[17] M. Jerrum and A. Sinclair, "Approximating the Permanent", *SIAM Journal on Computing*, 18(1989), pp. 1149–1178.

[18] M. Jerrum, A. Sinclair, and E. Vigoda, "A Polynomial-Time Approximation Algorithm for the Permanent of a Matrix with Nonnegative Entries," *Journal of the ACM,* 4(2004), pp. 671–697.

[19] M.R. Jerrum, L.G. Valiant, and V.V. Vazirani, "Random Generation of Combinatorial Structures from a Uniform Distribution," *Theoretical Computer Science,* 43(1986), pp. 169–188.

[20] R. Kannan, L. Lovász, and M. Simonovits, "Random Walks and an $O^*(n^5)$ Volume Algorithm for Convex Bodies", *Random Structures and Algorithms*, 11(1997), pp. 1–50.

[21] D.E. Knuth, *Seminumerical Algorithms. The Art of Computer Programming,* 3^{rd} ed., Addison Wesley, 1998.

[22] L. Levin, "Average Case Complete Problems," *SIAM Journal on Computing,* 15(1986), pp. 285–286.

[23] D.A. Levin, Y. Peres, and E.L. Wilmer, *Markov Chains and Mixing Times,* American Mathematical Society, 2009.

[24] L. Lovász, "Random Walks on Graphs: A Survey", In: D. Miklós, V.T. Sós, T. Szőnyi (eds.) *Combinatorics, Paul Erdős is Eighty*, János Bolyai Mathematical Society, Budapest, 1996, Vol. 2, pp. 353–398.

[25] L. Lovász and P. Winkler, "Efficient Stopping Rules for Markov Chains", *STOC'95*, pp. 76–82.

[26] L. Lovász and M. Simonovits, "Random Walks in a Convex Body and an Improved Volume Algorithm," *Random Structures and Algorithms,* 4(1993), pp. 359–412.

[27] L. Lovász and P. Winkler "Exact Mixing in an Unknown Markov Chain" *Electronic Journal of Combinatorics*, 2(1995), Paper #R15.

[28] L. Lovász and P. Winkler, "Mixing Times", In: D. Aldous and J. Propp (Eds.), *Microsurveys in Discrete Probability*, DIMACS Series in Discrete Math. and Theor. Comp. Sci., AMS 1998, pp. 85–133.

[29] M. Luby and E.Vigoda, "Approximately Counting Up To Four," *STOC'97,* pp. 682–687.

[30] C. McDiarmid, "Concentration", In: M. Habib et al. (eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics,* Springer Verlag, 1998, pp. 195–248.

[31] I. Miklós, *Computational Complexity of Counting and Sampling,* CRC Press, 2019.

[32] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis,* Cambridge University Press, 2005.

[33] R. Motwani and P. Raghavan, *Randomized algorithms,* Cambridge University Press, 1995.

[34] J.G. Propp and D.B. Wilson, "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics" *Random Structures and Algorithms*, 9(1996), pp. 223–252.

[35] D. Randall, "Rapidly Mixing Markov Chains with Applications in Computer Science and Physics", *Computing in Science & Engineering,* 8(2006), pp. 30–41.

[36] A. Sinclair, *Algorithms for Random Generation and Counting*, Birkhäuser, Boston, MA, 1993.

[37] A. Sinclair, and M. Jerrum, "Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains," *Information and Computation,* 82(1989), pp. 93–133.

[38] A. Sly, "Computational Transition at the Uniqueness Threshold," *FOCS 2010,*, pp. 287–296.

[39] E. Vigoda, "Markov Chain Monte Carlo Methods – Relationship between Counting and Sampling," Lecture Note, Georgia Tech, 2006,
`https://www.cc.gatech.edu/~vigoda/MCMC_Course/Sampling-Counting.pdf`

[40] S. Zachos, "Probabilistic quantifiers and games", *Journal of Computer and System Sciences*, 36(1988), pp. 433–451.

Appendix A: Properties of the Separation Distance

First we prove a simple lower and upper bound of conditional probabilities and conditional expectations, used at several occasions in the paper.

Lemma 3 *Let A, B be any two events, with $\Pr(B) \geq 1-\varepsilon$ for some $\varepsilon \in [0,1)$. Then*

$$\Pr(A) - \varepsilon \;\leq\; \Pr(A|B) \;\leq\; \frac{1}{1-\varepsilon}\Pr(A). \tag{9}$$

Further, let ν be a random variable, taking its values from the set $\{0,\ldots,T\}$, with some integer $T \geq 0$. Then

$$\mathrm{E}(\nu) - T^2\varepsilon \;\leq\; \mathrm{E}(\nu|B) \;\leq\; \frac{1}{1-\varepsilon}\mathrm{E}(\nu) \tag{10}$$

and

$$(1-\varepsilon)\mathrm{E}(\nu|B) \;\leq\; \mathrm{E}(\nu) \;\leq\; \mathrm{E}(\nu|B) + T^2\varepsilon. \tag{11}$$

Proof. The first inequality in (9) is proved as

$$\Pr(A \,|\, B) = \frac{\Pr(AB)}{\Pr(B)} \geq \Pr(AB) = \Pr(A) - \Pr(A\overline{B}) \geq \Pr(A) - \Pr(\overline{B}) \geq \Pr(A) - \varepsilon,$$

while the second inequality in (9) is shown by

$$\Pr(A \,|\, B) = \frac{\Pr(AB)}{\Pr(B)} \leq \frac{\Pr(AB)}{1-\varepsilon} \leq \frac{1}{1-\varepsilon}\Pr(A).$$

For the expected value we can write, using (9) with $A = \{\nu = t\}$,

$$\mathrm{E}(\nu|B) \;=\; \sum_{t=0}^{T} t\Pr(\nu = t \mid B) \;\geq\; \sum_{t=0}^{T} t\Big(\Pr(\nu = t) - \varepsilon\Big) \;\geq\; \mathrm{E}(\nu) - \varepsilon\sum_{t=0}^{T} t \;\geq\; \mathrm{E}(\nu) - T^2\varepsilon$$

and

$$\mathrm{E}(\nu|B) \;=\; \sum_{t=0}^{T} t\Pr(\nu = t \mid B) \;\leq\; \sum_{t=0}^{T} t\,\frac{1}{1-\varepsilon}\Pr(\nu = t) \;=\; \frac{1}{1-\varepsilon}\mathrm{E}(\nu).$$

Finally, (11) is obtained by rearranging (10).

♠

We are now ready to prove Lemma 1.

Proof of Lemma 1. Statements (i) and (ii) (***Normalized Value*** and ***Separation***) directly follow from the definition. For the ***Triangle Inequality*** in (iii), let $s(\alpha,\beta) = \varepsilon_1$, $s(\beta,\gamma) = \varepsilon_2$. Then $\alpha \geq (1-\varepsilon_1)\beta$ and $\beta \geq (1-\varepsilon_2)\gamma$, implying $\alpha \geq (1-\varepsilon_1)(1-\varepsilon_2)\gamma \geq (1-\varepsilon_1-\varepsilon_2)\gamma$, which yields $s(\alpha,\gamma) \leq \varepsilon_1 + \varepsilon_2 = s(\alpha,\beta) + s(\beta,\gamma)$.

Considering the ***Mixture Representation*** (iv), we first show the equivalence of (a) and (b).

If $\varepsilon = 0$, then $\alpha = \beta$ holds and γ can be arbitrary. If $\varepsilon = 1$, then $s(\alpha,\beta) \leq 1$ is always true and $\gamma = \alpha$ is the unique distribution satisfying $\alpha = (1-\varepsilon)\beta + \varepsilon\gamma$. Thus, the only nontrivial case is $0 < \varepsilon < 1$, which we treat below.

1. Assume $s(\alpha,\beta) \leq \varepsilon$. Let us define a real valued function γ on S by

$$\gamma = \frac{1}{\varepsilon}\alpha + \left(1 - \frac{1}{\varepsilon}\right)\beta. \tag{12}$$

By definition, $\alpha \geq (1 - s(\alpha,\beta))\beta$ always holds. Together with $s(\alpha,\beta) \leq \varepsilon$ they imply

$$\frac{1}{\varepsilon}\alpha \geq \frac{1 - s(\alpha,\beta}{\varepsilon}\beta = \frac{1}{\varepsilon}\beta - \frac{s(\alpha,\beta}{\varepsilon}\beta \geq \frac{1}{\varepsilon}\beta - \beta.$$

Now using the definition of γ we obtain

$$\gamma = \frac{1}{\varepsilon}\alpha + \left(1 - \frac{1}{\varepsilon}\right)\beta \geq \frac{1}{\varepsilon}\beta - \beta + \beta - \frac{1}{\varepsilon}\beta = 0.$$

Thus, γ is nonnegative. Moreover,

$$\sum_{x\in S}\gamma(x) = \frac{1}{\varepsilon}\underbrace{\sum_{x\in S}\alpha(x}_{=1} + \left(1 - \frac{1}{\varepsilon}\right)\underbrace{\sum_{x\in S}\beta(x}_{=1} = 1,$$

so γ is indeed a probability distribution. Expressing α from (12), we get $\alpha = (1-\varepsilon)\beta + \varepsilon\gamma$. It is also clear that whenever $\varepsilon > 0$, the equation $\alpha = (1-\varepsilon)\beta + \varepsilon\gamma$ uniquely determines γ.

2. Conversely, suppose that a probability distribution γ exists on S with $\alpha = (1-\varepsilon)\beta + \varepsilon\gamma$. Expressing γ gives $\gamma = (1/\varepsilon)\alpha + (1 - 1/\varepsilon)\,\beta$. Assume indirectly that $s(\alpha,\beta) > \varepsilon$. Then there is an $x \in S$ with $\alpha(x) < (1-\varepsilon)\beta(x)$. It implies

$$\gamma(x) = \frac{1}{\varepsilon}\alpha(x) + \left(1 - \frac{1}{\varepsilon}\right)\beta(x) < \frac{1}{\varepsilon}(1-\varepsilon)\beta(x) + \left(1 - \frac{1}{\varepsilon}\right)\beta(x) = 0,$$

yielding $\gamma(x) < 0$, which is impossible, since γ is a probability distribution. Thus, $s(\alpha,\beta) \leq \varepsilon$ must hold.

The ***Conditioning Property*** (v) follows from Lemma 3 (see at the beginning of this Appendix). Specifically, let $A = \{\xi = x\}$ and $B = E$, implying $\alpha(x) = \Pr(A)$ and $\alpha^E(x) = \Pr(A|B)$. Then the second inequality in (9) yields

$$\alpha^E(x) \;\leq\; \frac{1}{1-\varepsilon}\Pr(A),$$

After rearranging, it becomes $\alpha(x) \geq (1-\varepsilon)\alpha^E(x)$, which is equivalent to $s(\alpha,\alpha^E) \leq \varepsilon$. To prove $s(\alpha^E,\alpha) \leq \varepsilon/q_0$, we use the first inequality in (9): $\alpha^E(x) \geq \alpha(x) - \varepsilon$. If $\alpha(x) > 0$, then $\alpha(x) \geq q_0 > 0$ holds, so by rearranging the inequality $\alpha^E(x) \geq \alpha(x) - \varepsilon$, we get

$$\alpha^E(x) \;\geq\; \left(1 - \frac{\varepsilon}{\alpha(x}\right)\alpha(x) \;\geq\; \left(1 - \frac{\varepsilon}{q_0}\right)\alpha(x).$$

This was derived for the case $\alpha(x) > 0$. However, if $\alpha(x) = 0$, then $\alpha^E(x) \geq (1-\varepsilon/q_0)\,\alpha(x)$ is trivially true, since the right-hand side is 0. Therefore, we obtain that $\alpha^E(x) \geq (1-\varepsilon/q_0)\,\alpha(x)$ always holds, which yields $s(\alpha^E,\alpha) \leq \varepsilon/q_0$.

To prove the ***Restriction Property*** (vi), let $r(\alpha,\beta) = \varepsilon$. Then we can write $\alpha \geq (1-\varepsilon)\beta$, and

$$\alpha_H = \frac{\alpha}{\alpha(H)} \geq (1-\varepsilon)\frac{\beta}{\alpha(H)} = (1-\varepsilon)\beta_H\frac{\beta(H)}{\alpha(H)}. \tag{13}$$

Let us bound the ratio $\beta(H)/\alpha(H)$ from below. Summing up the probabilities for the complement $\overline{H}$ of H gives

$$\alpha(\overline{H}) \geq (1-\varepsilon)\beta(\overline{H}) = (1-\varepsilon)(1-\beta(H)),$$

yielding

$$\alpha(H) = 1 - \alpha(\overline{H}) \leq 1 - \big((1-\varepsilon)(1-\beta(H)\big).$$

After rearranging we get

$$\alpha(H) \le \varepsilon + \beta(H) - \varepsilon\beta(H) \le \varepsilon + \beta(H) = \beta(H)\left(1 + \frac{\varepsilon}{\beta(H)}\right).$$

This gives us the lower bound

$$\frac{\beta(H)}{\alpha(H} \ge \frac{1}{1 + \frac{\varepsilon}{\beta(H)}} \ge 1 - \frac{\varepsilon}{\beta(H)}.$$

Using it in (13) yields

$$\alpha_H \ge \beta_H(1-\varepsilon)\left(1 - \frac{\varepsilon}{\beta(H)}\right) \ge \beta_H\left(1 - \varepsilon\left(1 + \frac{1}{\beta(H)}\right)\right),$$

which is equivalent to

$$s(\alpha_H, \beta_H) \le s(\alpha, \beta)\left(1 + \frac{1}{\beta(H)}\right),$$

proving the restriction property. The right-hand side in the above inequality may become larger than one, but then it still remains true, since $s(\alpha_H, \beta_H) \le 1$ always holds.

Finally, to prove ***Convexity*** (vii), let $\varepsilon_i = r(\alpha_i, \beta),\ i = 1, \ldots, n$. By definition, we have $\alpha_i \ge (1-\varepsilon_i)\beta$ for every i. Multiplying each inequality by p_i and summing them up, we get

$$\sum_{i=1}^{n} p_i\alpha_i \ge \sum_{i=1}^{n} p_i(1-\varepsilon_i)\beta = \sum_{i=1}^{n} p_i\beta - \sum_{i=1}^{n} p_i\varepsilon_i\beta = \left(1 - \sum_{i=1}^{n} p_i\varepsilon_i\right)\beta,$$

which means

$$s\left(\sum_{i=1}^{n} p_i\alpha_i,\ \beta\right) \le \sum_{i=1}^{n} p_i\varepsilon_i = \sum_{i=1}^{n} p_i s(\alpha_i, \beta).$$

♠

Appendix B: Proof of Theorem 1

In order to prove Theorem 1, we need to develop some tools. First we provide a high level overview of the path we plan to follow.

B.0 Road Map for the Proof of Theorem 1

- In Section B.1 we develop an upper bound on the separation distance between the distributions of the random H-hit and the H-restricted sample in an equidistributed random sequence (see Definition 2). It turns out that based on the concentration of the number of H-hits alone, a meaningful (though quite technical) bound is possible, without assuming independence, or any special kind of dependence, such as the Markov property, in the sequence.
- Unfortunately, in our main application (subset sampling in slowly mixing Markov chains) the number of H-hits is not concentrated enough. Therefore, in Section B.2 we start developing the tools for the "virtual proliferation" approach. The general goal is to handle an array of statistically independent versions of the input sequence, and benefit from its statistical

features. This array is only virtual in the sense that the whole matrix is not actually available as input, we only have access to one run, corresponding to one row.

As a first tool, we introduce the concept of a perfect matching in a square 0–1 matrix. It is quite simple: it corresponds to a perfect matching in the bipartite graph that has the matrix as its bi-adjacency matrix. Then we define the concept of a p-robust 0–1 matrix. Informally, it means that that every entry takes the value 1 with probability at least p, regardless of conditioning on any values of preceding entries. Here the precedence order means reading the matrix row by row, from left to right in each row. Note that the p-robust condition is weaker than independence, but still has useful properties. In particular, we prove that the probability of having a perfect matching in a p-robust $N \times N$ 0–1 matrix is exponentially close to 1, already for moderate values of p.

In order to prove this bound, we need some tools from random graph theory, but we have to adjust the setting from independent edges to a non-independent version that we call p-robust random subgraph of a given graph. This leads to a general result that may be interesting on its own, see Lemma 6.

- In Section B.3 we show a useful property of random perfect matchings in random 0–1 matrices. Under some conditions, we prove that in any fixed column (or row) of the matrix, a 1-entry that belongs to a random perfect matching behaves exactly like a randomly chosen 1-entry of the column (or row).

- Finally, using the developed tools, we provide the proof of Theorem 1 in Section B.4.

B.1 A Bound Based on the Concentration of the Number of H-hits

For convenient reference, let us first repeat the definition of a (ρ, ε)-concentrated random variable:

Definition (Concentration) *A random variable ν is called (ρ, ε)-concentrated, if*

$$\Pr\big((1-\rho)\mathrm{E}(\nu) \le \nu \le (1+\rho)\mathrm{E}(\nu) \big) \ge 1 - \varepsilon$$

holds, where $\rho \ge 0$, $\varepsilon \in [0,1]$, and $\mathrm{E}(\nu)$ denotes the expected value of ν.

Below we prove a technical lemma that bounds the separation distance between the distributions of the random H-hit and the H-restricted sample in an equidistributed random sequence, in terms of the concentration of the number of H-hits, allowing arbitrary dependencies among the sequence entries. Further, the lemma assumes that the sequence is conditioned on an event, which forces the existence of an H-hit.

Lemma 4 *Let $\xi = (\xi_1, \ldots, \xi_T)$ be an equidistributed sequence of random variables, each taking its values from a set S, and having common distribution π, with $q_0 = \min_{x \in S} \pi(x) > 0$. Assume that the number ν of hits in ξ to a given target set $H \subseteq S$ is (ρ, ε)-concentrated, and $\mathrm{E}(\nu) \ge 1$. Let $\pi_H = \pi/\pi(H)$ be the distribution of an H-restricted sample in ξ. Furthermore, let E be an event with $\Pr(\nu \ge 1 | E) = 1$ and $\Pr(E) \ge 1 - p_0$, for some p_0 with $q_0 > p_0 \ge 0$. Let σ be the distribution of a random H-hit, and let σ^E denote its conditional version, given E. Then the separation distance between σ^E and π_H is bounded as*

$$s(\sigma^E, \pi_H) \le (\varepsilon + p_0)\left(1 + \frac{T}{q_0 - p_0}\right) + \rho(1 + T^2 p_0) + T^2 p_0 + \frac{p_0}{q_0}(1+T).$$

Proof. First note that, due to $\Pr(\nu \geq 1 \mid E) = 1$, the random H-hit is well defined under the condition E, as E forces ξ to always hit H. First we prove the special case when $\Pr(E) = 1$, then we extend it to the general case. When $\Pr(E) = 1$, i.e., E is the certain event, we omit marking the conditioning in the distributions, and also simply write $\Pr(\nu \geq 1) = 1$. This case is referred to as the *unconditional case,* which we treat first.

For any $x \in H$, let γ_x be the random number of times the value x is taken by ξ, that is, $\gamma_x = |\{i \,|\, \xi_i = x\}|$. Given a realization y of the sequence ξ, it uniquely determines the corresponding realizations of γ_x and ν. Therefore, there are deterministic functions f_x, g, such that these realizations are expressed as $\gamma_x = f_x(y)$ and $\nu = g(y)$. Hence, the probability that the random H-hit η takes the value x, given a realization $\xi = y$, is

$$\Pr(\eta = x \,|\, \xi = y) = \frac{f_x(y)}{g(y)}.$$

Note that here $g(y) \geq 1$, due to $\Pr(\nu \geq 1) = 1$. Taking the expected value on both sides over the realizations (using the law of iterated expectation on the left), we get

$$\mathrm{E}(\Pr(\eta = x \,|\, \xi)) = \Pr(\eta = x) = \mathrm{E}\left(\frac{\gamma_x}{\nu}\right).$$

Thus, $\sigma(x) = \mathrm{E}\left(\frac{\gamma_x}{\nu}\right)$, and the denominator never becomes 0, since H is hit at least once, due to $\Pr(\nu \geq 1) = 1$. Now let $R = [(1-\rho)\mathrm{E}\nu,\, (1+\rho)\mathrm{E}\nu]$ be the interval in which ν falls with probability at least $1 - \varepsilon$, by its assumed (ρ, ε)-concentration. We can write

$$\sigma(x) = \mathrm{E}\left(\frac{\gamma_x}{\nu} \,|\, \nu \in R\right) \Pr(\nu \in R) + \mathrm{E}\left(\frac{\gamma_x}{\nu} \,|\, \nu \notin R\right) \Pr(\nu \notin R).$$

Omitting the second summand on the right-hand side, as well as using that $\nu \leq (1+\rho)\mathrm{E}\nu$ holds whenever $\nu \in R$, we obtain

$$\sigma(x) \;\geq\; \frac{\mathrm{E}(\gamma_x \,|\, \nu \in R)}{(1+\rho)\mathrm{E}\nu} \Pr(\nu \in R) \;\geq\; \frac{\mathrm{E}(\gamma_x \,|\, \nu \in R)}{(1+\rho)\mathrm{E}\nu} (1-\varepsilon). \tag{14}$$

Let us bound the conditional expectation $\mathrm{E}(\gamma_x \,|\, \nu \in R)$ from below. By Lemma 3 (see in Appendix A), we get that the concentration condition $\Pr(\nu \in R) \geq 1 - \varepsilon$ yields

$$\mathrm{E}(\gamma_x \,|\, \nu \in R) \;\geq\; \mathrm{E}(\gamma_x) - \varepsilon T^2.$$

Substituting this into (14) gives

$$\sigma(x) \geq \frac{\mathrm{E}(\gamma_x) - \varepsilon T^2}{\mathrm{E}\nu} \cdot \frac{1-\varepsilon}{1+\rho}. \tag{15}$$

Observe now that $\gamma_x = \sum_{t=1}^{T} \chi_{\{\xi_t = x\}}$, where $\chi_{\{\xi_t = x\}}$ is the indicator of the event $\{\xi_t = x\}$. Therefore, by the linearity of expectation, $\mathrm{E}(\gamma_x) = \pi(x)T$ holds, regardless of $\xi_1, \ldots, \xi_T$ being independent or not. We also have $\mathrm{E}(\nu) = \pi(H)T$, due to $\nu = \sum_{x \in H} \gamma_x$. Substituting these into (15), and canceling out a factor of T, gives

$$\sigma(x) \;\geq\; \frac{\pi(x) - \varepsilon T}{\pi(H)} \cdot \frac{1-\varepsilon}{1+\rho} \tag{16}$$

for any $x \in H$. Using $\pi_H(x) = \frac{\pi(x)}{\pi(H)}$ and $\pi(x) \geq q_0$, we can rearrange and bound (16) to obtain

$$\sigma(x) \;\geq\; \pi_H(x) \left(1 - \frac{1}{q_0}\varepsilon T\right) \cdot \frac{1-\varepsilon}{1+\rho}, \tag{17}$$

where $q_0 = \min_{x \in S} \pi(x) > 0$. Bounding the last factor in (17) from below, as

$$\frac{1-\varepsilon}{1+\rho} = 1 - \frac{\rho+\varepsilon}{1+\rho} \geq 1 - (\rho+\varepsilon),$$

we can continue (17) as

$$\sigma(x) \;\geq\; \pi_H(x) \left(1 - \frac{1}{q_0}\varepsilon T\right) (1 - (\rho+\varepsilon)) \;\geq\; \pi_H(x) \left(1 - \left(\frac{1}{q_0}\varepsilon T + \rho + \varepsilon\right)\right).$$

By the definition of the separation distance (see Definition 1), this means

$$s(\sigma, \pi_H) \;\leq\; \frac{1}{q_0}\varepsilon T + \rho + \varepsilon \;=\; \varepsilon\left(1 + \frac{T}{q_0}\right) + \rho, \tag{18}$$

completing the case when $\Pr(E) = 1$, which we refer to as the unconditional case.

Let us now consider the general case, when $\Pr(E) \geq 1 - p_0$, but possibly $\Pr(E) < 1$. Let α be the distribution of ξ, i.e., the joint distribution of $\xi_1, \ldots, \xi_T$. Let α^E denote the conditional distribution of ξ, given E, that is, $\alpha^E(x) = \Pr(\xi = x | E)$. Generally, for every distribution, we distinguish its conditioned version, given E, by adding E as a superscript. With this notation, π^E denotes the conditioned distribution of the entries of ξ, given E, and π_H^E is the distribution of the H-restricted sample, given E. That is, π_H^E is the restriction of π^E to H, so we have $\pi_H^E = \pi^E / \pi^E(H)$. Let $\widetilde{\nu}$ denote the random number of H-hits under the condition E, to distinguish it from the unconditional case. Recall that the random H-hit is well defined under E, as E forces ξ to always hit H, even if $\Pr(E) < 1$.

First we analyze how the concentration of the number of H-hits changes due to the conditioning. Set

$$a = (1-\rho)\mathrm{E}(\nu) \quad \text{and} \quad b = (1+\rho)\mathrm{E}(\nu),$$

where ν is the number of H-hits in the unconditional case. Note that a, b are constants. Using $\Pr(E) \geq 1 - p_0$, and applying Lemma 3 (see in Appendix A), we can write

$$\Pr(a \leq \widetilde{\nu} \leq b) \;=\; \Pr(a \leq \nu \leq b \mid E) \;\geq\; \Pr(a \leq \nu \leq b) - p_0.$$

Observe that $\Pr(a \leq \nu \leq b) \geq 1 - \varepsilon$ holds, due to the assumed (ρ, ε)-concentration of ν, so we obtain

$$\Pr(a \leq \widetilde{\nu} \leq b) \;\geq\; 1 - \varepsilon - p_0 \;=\; 1 - \widetilde{\varepsilon} \tag{19}$$

with $\widetilde{\varepsilon} = \varepsilon + p_0$. Set now

$$\widetilde{a} \;= (1-\widetilde{\rho})\mathrm{E}(\widetilde{\nu}) \quad \text{and} \quad \widetilde{b} \;= (1+\widetilde{\rho})\mathrm{E}(\widetilde{\nu}).$$

We look for a constant $\widetilde{\rho}$, such that $a \geq \widetilde{a}$ and $b \leq \widetilde{b}$ both hold. With such a $\widetilde{\rho}$ we have $[a, b] \subseteq [\widetilde{a}, \widetilde{b}]$, which, taking (19) into account, yields

$$\Pr(\widetilde{a} \leq \widetilde{\nu} \leq \widetilde{b}) \;\geq\; \Pr(a \leq \widetilde{\nu} \leq b) \;\geq\; 1 - \widetilde{\varepsilon},$$

implying

$$\Pr\big((1-\widetilde{\rho})\mathrm{E}(\widetilde{\nu}) \leq \widetilde{\nu} \leq (1+\widetilde{\rho})\mathrm{E}(\widetilde{\nu})\big) \geq 1 - \widetilde{\varepsilon}. \tag{20}$$

That is, we are going to get a $(\widetilde{\rho}, \widetilde{\varepsilon})$-concentration bound for $\widetilde{\nu}$. We still need to find an appropriate $\widetilde{\rho}$ value, i.e., one that satisfies both $a \geq \widetilde{a}$ and $b \leq \widetilde{b}$. Below we show that

$$\widetilde{\rho} \;=\; \rho(1 + T^2 p_0) + T^2 p_0$$

is such a value. In order to prove it, let us first bound $\mathrm{E}(\nu)$ in terms of $\mathrm{E}(\widetilde{\nu}) = \mathrm{E}(\nu|E)$ via Lemma 3 (see in Appendix A):

$$(1-p_0)\mathrm{E}(\widetilde{\nu}) \;\le\; \mathrm{E}(\nu) \;\le\; \mathrm{E}(\widetilde{\nu}) + T^2 p_0.$$

Then we can write

$$(1-\rho)\mathrm{E}(\nu) \;\ge (1-\rho)(1-p_0)\mathrm{E}(\widetilde{\nu}) \;=\; \big(\,1-(\rho(1-p_0)+p_0)\,\big)\,\mathrm{E}(\widetilde{\nu}). \tag{21}$$

Furthermore, also using $\mathrm{E}(\widetilde{\nu}) \ge 1$, which follows from $\Pr(\widetilde{\nu} \ge 1) = \Pr(\nu \ge 1|E) = 1$, we can write

$$\begin{aligned}(1+\rho)\mathrm{E}(\nu) \;&\le (1+\rho)\big(\;\mathrm{E}(\widetilde{\nu}) + T^2 p_0\,\big) \;= (1+\rho)\mathrm{E}(\widetilde{\nu})\left(1+\frac{T^2 p_0}{\mathrm{E}(\widetilde{\nu})}\right) \;\le \\ (1+\rho)\mathrm{E}(\widetilde{\nu})(1+T^2p_0) \;&=\; \big(\,1+\rho(1+T^2p_0)+T^2p_0\,\big)\,\mathrm{E}(\widetilde{\nu}).\end{aligned} \tag{22}$$

Introducing $\widetilde{\rho}_1 = \rho(1-p_0)+p_0$ and $\widetilde{\rho}_2 = \rho(1+T^2p_0)+T^2p_0$, we get from (21) and (22):

$$(1-\rho)\mathrm{E}(\nu) \;\ge (1-\widetilde{\rho}_1)\mathrm{E}(\widetilde{\nu}) \quad \text{ and } \quad (1+\rho)\mathrm{E}(\nu) \;\le (1+\widetilde{\rho}_2)\mathrm{E}(\widetilde{\nu}). \tag{23}$$

Since $\widetilde{\rho}_2 = \rho(1+T^2p_0)+T^2p_0 \;\ge\; \rho(1-p_0)+p_0 = \widetilde{\rho}_1$, therefore, the first inequality in (23) remains valid if we use $\widetilde{\rho}_2$ there, too, to achieve symmetry. Hence, with $\widetilde{\rho} = \rho(1+T^2p_0)+T^2p_0$, the new constants

$$\widetilde{a} = (1-\widetilde{\rho})\mathrm{E}(\widetilde{\nu}) \quad \text{ and } \quad \widetilde{b} = (1+\widetilde{\rho})\mathrm{E}(\widetilde{\nu})$$

indeed satisfy $a \ge \widetilde{a}$ and $b \le \widetilde{b}$, implying (20). Thus, we conclude that $\widetilde{\nu}$ is $(\widetilde{\rho},\widetilde{\varepsilon})$-concentrated, with

$$\widetilde{\rho} = \rho(1+T^2p_0)+T^2p_0 \quad \text{ and } \quad \widetilde{\varepsilon} = \varepsilon + p_0. \tag{24}$$

We can now bound the separation distance $s(\sigma^E, \pi_H)$. Let us first write, using the Triangle Inequality (see (iii) in Lemma 1):

$$s(\sigma^E, \pi_H) \le s(\sigma^E, \pi_H^E) + s(\pi_H^E, \pi_H). \tag{25}$$

The first term on the right-hand side, $s(\sigma^E, \pi_H^E)$, is the separation distance between the conditional distribution of the random H-hit, given E, and of the conditional distribution of the H-restricted sample, given E. That is, now they are *both* conditioned on E. We can view it as a case when *everything* is conditioned on E, so it can be represented as if the whole sequence is drawn from the distribution $\beta = \alpha^E$. Thus, we obtain a case when the sequence is drawn from a distribution β, and it holds that $\Pr(\nu \ge 1) = 1$, and otherwise there is no separate conditioning. But this is exactly what we called the unconditional case, so we fall back to that, regarding the estimation of $s(\sigma^E, \pi_H^E)$.

To use our earlier derived bound (18) for $s(\sigma^E, \pi_H^E)$, we need the concentration parameters for the case when the sequence is drawn from the distribution $\beta = \alpha^E$. We have analyzed that under the condition E (i.e., when the sequence is drawn from $\beta = \alpha^E$) the number of H-hits is $(\widetilde{\rho},\widetilde{\varepsilon})$-concentrated, with $\widetilde{\rho}$ and $\widetilde{\varepsilon}$ given in (24). We also need to replace the value of q_0 by q_0^E, its conditioned value under E, since now everything is conditioned on E. Knowing that $\Pr(E) \ge 1-p_0$, we obtain by Lemma 3 (see in Appendix A) that $q_0^E \ge q_0 - p_0$ holds, with positive right-hand side, as it is assumed that $q_0 > p_0$. Thus, the bound (18) from the unconditional case gives us, with the modified parameters, the following bound:

$$s(\sigma^E, \pi_H^E) \;\le\; \widetilde{\varepsilon}\left(1+\frac{T}{q_0^E}\right) + \widetilde{\rho} \;\le\; (\varepsilon+p_0)\left(1+\frac{T}{q_0-p_0}\right) + \rho(1+T^2p_0) + T^2p_0.$$

Considering now the last term $s(\pi_H^E, \pi_H)$ in (25), we observe that π_H^E and π_H are the H-restricted versions of π^E and π, respectively. Applying the Restriction Property of the separation distance (see (vi) in Lemma 1) to the distributions π^E, π, we obtain

$$s(\pi_H^E, \pi_H) \;\leq\; s(\pi^E, \pi)\left(1 + \frac{1}{\pi(H)}\right).$$

Since $\mathrm{E}(\nu) = \pi(H)T$, and, by assumption, $\mathrm{E}(\nu) \geq 1$, we get $1/\pi(H) = T/\mathrm{E}(\nu) \leq T$, yielding

$$s(\pi_H^E, \pi_H) \;\leq\; s(\pi^E, \pi)(1+T).$$

To bound $s(\pi^E, \pi)$ in this expression, we use the Conditioning Property of the separation distance (see (v) in Lemma 1) to obtain $s(\pi^E, \pi) \leq p_0/q_0$, due to $\Pr(E) \geq 1 - p_0$ and $q_0 = \min_{x \in S} \pi(x) > 0$. Finally, putting the pieces together, according to (25), gives us the desired estimation

$$s(\sigma^E, \pi_H) \;\leq (\varepsilon + p_0)\left(1 + \frac{T}{q_0 - p_0}\right) + \rho(1 + T^2 p_0) + T^2 p_0 + \frac{p_0}{q_0}(1+T),$$

completing the proof. ♠

B.2 Random Perfect Matchings in Random 0–1 Matrices

Definition 10 (Perfect matching in 0–1 matrix) *Let $A = [a_{ij}]$ be a 0–1 matrix of size $N \times N$. We say that A has a perfect matching, if there exist N different row indices $i_1, \ldots, i_N \in [N]$, such that $a_{i_k k} = 1$ holds for every $k \in [N]$.*

The above definition is equivalent to requiring that we can select N matrix entries, such that each has value 1, and no two of them fall in the same row and in the same column. It is also equivalent to having a perfect matching in the bipartite graph, defined by A as its bi-adjacency matrix. This is why we call it perfect matching, even though we need it only for matrices, not graphs.

Definition 11 (Preceding entry) *Let $A = [a_{ij}]$ be a matrix of size $N \times N$. We say that the index pair (i,j) precedes another index pair (k,ℓ), denoted by $(i,j) \prec (k,\ell)$, if either $i < k$, or $i = k$ and $j < \ell$. Then we also say that the entry a_{ij} precedes the entry $a_{k\ell}$, denoted by $a_{ij} \prec a_{k\ell}$.*

In words, for any entry, its preceding entries are those that are located either in a row that is higher up, or in the same row, but to the left of the considered entry.

Definition 12 (p-robust random 0–1 matrix) *Let $X = [\xi_{ij}]$ be a random 0–1 matrix, drawn from an arbitrary distribution, and $p \in (0,1)$. We say that X is* p-robust, *if $\forall i,j:\ \Pr(\xi_{i,j} = 1) \geq p$, and this remains so even if we condition on arbitrary values of any preceding matrix entries. Formally, if $\xi_{i_1 j_1}, \ldots, \xi_{i_m j_m}$ precede ξ_{ij}, then*

$$\Pr(\xi_{ij} = 1 \mid \xi_{i_1 j_1} = a_{i_1 j_1}, \ldots, \xi_{i_m j_m} = a_{i_m j_m}) \;\geq\; p$$

holds for every i,j, and for every choice of the constants $a_{i_1 j_1}, \ldots, a_{i_m j_m} \in \{0,1\}$, whenever the condition has positive probability.

In words, the matrix is p-robust, if any entry takes the value 1 with probability at least p, regardless of fixing the values of any preceding entries. Note that it does not mean the considered entry is independent of the preceding ones, as the probability that it takes the value 1 may depend on them, just this probability never drops below p. Yet, as we are going to see, p-robustness behaves as a weaker, but still very useful, substitute of independence.

The next lemma shows that a p-robust matrix has a high probability of having a perfect matching, whenever p is not too small. This will serve as a handy tool in the proof of Theorem 1.

Lemma 5 *Let $X = [\xi_{ij}]$ be a random 0–1 matrix of size $N \times N$, drawn from an arbitrary distribution, and let $p \in (0,1)$. If X is p-robust* (see Definition 12), *then*

$$\Pr(X \text{ has a perfect matching}) \geq 1 - (N+1)^3(1-p)^{\lfloor N/2 \rfloor}. \tag{26}$$

Remark. The bound (26) is not the tightest possible, but it provides a simple, explicit, non-asymptotic formula, which will be valuable for our purposes in the sequel.

For the proof of Lemma 5 we reformulate the concepts in terms of random graphs, because it is easier to visualize, and allows us to use some graph theoretic tools.

Definition 13 (p-random subgraph) *Let G be a graph and $p \in (0,1)$. Create a random subgraph of G by keeping each edge with probability p, and deleting it with probability $1-p$, doing it independently for each edge. We denote the arising random graph by $G(p)$, and call it a p-random subgraph of G.*

Definition 14 (p-robust random subgraph) *Let $G = (V, E)$ be a graph with $E = \{e_1, \ldots, e_m\}$, and let $p \in (0,1)$. Create a random subgraph of G by randomly keeping/deleting edges of G, but not necessarily independently. We require, however, that the following holds: each edge e_i is kept with probability at least p, given the status (kept/deleted) of $e_{i+1}, \ldots, e_m$, whatever that status is, assuming it has positive joint probability. For $i = m$, when the list $e_{i+1}, \ldots, e_m$ is empty, we require that e_m is kept with probability at least p, unconditionally. The random subgraph arising this way is called a* p-robust random subgraph, *and is denoted by $\widetilde{G}(p)$.*

Remark. Observe that the definitions directly imply the following: a random 0–1 matrix is p-robust if and only if it is the bi-adjacency matrix of a p-robust random subgraph of a complete balanced bipartite graph, assuming that the graph edges are numbered in the following way: if entries $a_{ij}, a_{k\ell}$ of the matrix correspond to graph edges e_q, e_r, respectively, then $a_{ij} \prec a_{k\ell}$ implies $q > r$. If this is the case, we say that the edges are numbered consistently with the matrix precedence relation.

Monotone graph property. Let $\mathcal{Q}$ be a set of graphs. We use it to represent a graph property: a graph has property $\mathcal{Q}$ if and only if $G \in \mathcal{Q}$. Therefore, we identify the property with $\mathcal{Q}$. We say that a graph property $\mathcal{Q}$ is *monotone,* if it is closed with respect to adding new edges. That is, if $G \in \mathcal{Q}$ and $G \subseteq G'$, then $G' \in \mathcal{Q}$.

Below we prove that for any graph G, and for any monotone graph property, the following holds: the probability that a p-random subgraph of G has the property can only increase if we switch to a p-robust random subgraph. It is interesting to note that while this claim is intuitively easy to believe, it takes some effort to prove it formally.

Lemma 6 *Let $\mathcal{Q}$ be a monotone graph property, G be a graph and $p \in (0,1)$. Then the following holds:*

$$\Pr(G(p) \in \mathcal{Q}) \;\leq\; \Pr(\widetilde{G}(p) \in \mathcal{Q})$$

where $G(p)$ and $\widetilde{G}(p)$ are the p-random and p-robust random subgraphs of G, respectively, defined in Definitions 13 *and* 14.

Proof. We are going to generate $\widetilde{G}(p)$ as the union of two random graphs, $G(p)$ and G_2, both on the same vertex set $V(G)$. $G(p)$ is the p-random subgraph of G, the other random graph G_2 will be defined later. The union $G(p) \cup G_2$ is meant with the understanding that if the same edge occurs in both graphs, then we merge them into a single edge. We plan to chose the edge probabilities in G_2, such that $G(p) \cup G_2 \sim \widetilde{G}(p)$, where the "$\sim$" relation between random graphs means that they have the same distribution, i.e., they are statistically indistinguishable. If this can be accomplished, then the claim will directly follow, since then a random graph distributed as $\widetilde{G}(p)$ can be obtained by adding edges to $G(p)$, which cannot destroy a monotone property, once $G(p)$ has it. This will imply the claim of the Lemma.

We introduce some notations. Let $e_1, \dots, e_m$ denote the edges of G. For every $i \in [m]$, let h_i be the indicator of the event that the edge e_i is included in the p-robust random graph $\widetilde{G}(p)$. We denote this event by $\{e_i \in \widetilde{G}(p)\}$. Further, let us use the abbreviation $h_i^m = (h_i, \dots, h_m)$. For any $a = (a_1, \dots, a_m) \in \{0,1\}^m$, the event $\{h_1^m = a\}$ means that $\widetilde{G}(p)$ takes a realization in which edge e_i is included if and only $a_i = 1$. Similarly, $\{h_i^m = a_i^m\}$ means $\{h_i = a_i, \dots, h_m = a_m\}$. We also use the abbreviation $a_i^m = (a_i, \dots, a_m)$.

Now let us generate the random graphs $G(p)$ and G_2, as follows.

Step 1. Let $i = m$.

Step 2. If $i = m$, then let $q_m = \Pr(h_m = 1)$. If $i < m$, then set $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m)$, where a_{i+1}^m indicates the already generated edges of $G(p) \cup G_2$.

Step 3. Compute
$$p_i' = \frac{p(1-q_i)}{1-p}. \tag{27}$$

Step 4. Put e_i into $G(p)$ with probability p, and put e_i into G_2 with probability $q_i - p_i'$.

Step 5. If $i > 1$, then decrease i by one, and go to Step 2; else HALT.

Remark: In Step 2 we do not specify a procedure for how to actually compute the value of q_i. This algorithm, however, is only needed for the proof, its *existence* is enough for us. Therefore, it is sufficient to *define* q_i, without providing a specific procedure for its computation. Also observe that the condition in $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m)$ must occur with positive probability, as the condition represents the already generated part. For $i = m$ there is no condition, but Definition 14 implies $q_m = \Pr(h_m = 1) > 0$.

First note that the value $q_i - p_i'$ in Step 4 can indeed be used as a probability. Clearly, $q_i - p_i' \le 1$ holds, as q_i is a probability and $p_i' \ge 0$. To show $q_i - p_i' \ge 0$, observe that

$$p_i' = \frac{p(1-q_i)}{1-p} \le q_i,$$

since the inequality can be rearranged into $p(1-q_i) \le q_i(1-p)$, which simplifies to $p \le q_i$. The latter is indeed true, due to $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m) \ge p$, which follows from the p-robust property.

Next we show that the algorithm generates the random graphs $G(p)$ and G_2 in a way that they satisfy $G(p) \cup G_2 \sim \widetilde{G}(p)$. We prove it by induction, starting from $i = m$ and progressing

downward to $i = 1$. For any $i \in [m]$, let $G^i(p)$, G_2^i denote the already generated parts of $G(p), G_2$, respectively, after executing Step 4 $m-i+1$ times, so they can only contain edges with index $\geq i$. Further, let $\widetilde{G}^i(p)$ be the subgraph of $\widetilde{G}(p)$ in which we only keep the edges with index $\geq i$, that is, $\widetilde{G}^i(p) = \widetilde{G}(p) - \{e_{i-1}, \ldots, e_1\}$. The inductive proof will show that $G^i(p) \cup G_2^i \sim \widetilde{G}^i(p)$ holds for every i. At the end of the induction, having reached $i = 1$, we are going to get $G^1(p) \cup G_2^1 \sim \widetilde{G}^1(p)$, which is the same as $G(p) \cup G_2 \sim \widetilde{G}(p)$.

Let us consider first the base case $i = m$. Then we have $\Pr(e_m \in G(p)) = \Pr(e_m \in G^m(p)) = p$ by Step 4. Then in Step 4, edge e_m is put into G_2 with probability $q_m - p'_m$, yielding $\Pr(e_m \in G_2^m) = q_m - p'_m$. Now observe that the formula (27) is chosen such that p'_i is precisely the solution of the equation

$$p + q_i - p'_i - (q_i - p'_i)p \;=\; q_i \tag{28}$$

for p'_i. For $i = m$ the equation becomes

$$p + q_m - p'_m - (q_m - p'_m)p \;=\; q_m, \tag{29}$$

and $p'_m = \frac{p(1-q_m)}{1-p}$ is the solution of this equation. Since by Step 4 we have $\Pr(e_m \in G^m(p)) = p$ and $\Pr(e_m \in G_2^m) = q_m - p'_m$, therefore, we get that the left-hand side of (29) is precisely the probability of the event $\{e_m \in G^m(p) \cup G_2^m\}$. By (29), this probability is equal to q_m, which is set to $q_m = \Pr(h_m = 1) = \Pr(e_m \in \widetilde{G}^m(p))$ in Step 2. This means that $G^m(p) \cup G_2^m \sim \widetilde{G}^m(p)$, as desired.

For the induction step, assume that the claim is true for $i+1$, i.e., $G^{i+1}(p) \cup G_2^{i+1} \sim \widetilde{G}^{i+1}(p)$ holds. In Step 4, edge e_i is added to $G^{i+1}(p)$ with probability p. It is also added to G_2^{i+1} with probability $q_i - p'_i$. Therefore, just like in the base case, we get that $p + q_i - p'_i - (q_i - p'_i)p = \Pr(e_i \in G^i(p) \cup G_2^i)$. We already know that p'_i satisfies the equation (28), so e_i is added to $\widetilde{G}^{i+1}(p)$ with probability $q_i = \Pr(h_i = 1 \mid h_{i+1}^m = a_{i+1}^m)$, given the already generated part, represented by a_{i+1}^m. By the inductive assumption, h_{i+1}^m is distributed as $\widetilde{G}^{i+1}(p)$, which is the truncated version of $\widetilde{G}(p)$, keeping only the $\geq i+1$ indexed edges. Hence, for h_{i+1}^m, we can write by the chain rule of conditional probabilities:

$$\Pr(h_{i+1}^m = a_{i+1}^m) \;=\; \Pr(h_m = a_m) \prod_{j=i+1}^{m-1} \Pr(h_j = a_j \mid h_{j+1}^m = a_{j+1}^m).$$

After processing e_i (i.e., adding it with probability q_i), we get

$$\begin{aligned}
\Pr(h_i^m = a_i^m) &= \Pr(h_i = a_i \mid h_{i+1}^m = a_{i+1}^m) \Pr(h_{i+1}^m = a_{i+1}^m) \\
&= \Pr(h_i = a_i \mid h_{i+1}^m = a_{i+1}^m) \underbrace{\Pr(h_m = a_m) \prod_{j=i+1}^{m-1} \Pr(h_j = a_j \mid h_{j+1}^m = a_{j+1}^m)}_{\Pr(h_{i+1}^m = a_{i+1}^m)} \\
&= \Pr(h_m = a_m) \prod_{j=i}^{m-1} \Pr(h_j = a_j \mid h_{j+1}^m = a_{j+1}^m),
\end{aligned}$$

which, by the chain rule, is indeed the distribution of $\widetilde{G}^i(p)$, completing the induction.

Thus, at the end, a realization $a = a_1^m \in \{0,1\}^m$ of $\widetilde{G}(p)$ is generated with probability

$$\Pr(h_1^m = a) \;=\; \Pr(h_m = a_m) \prod_{j=1}^{m-1} \Pr(h_j = a_j \mid h_{j+1}^m = a_{j+1}^m),$$

indeed creating $\widetilde{G}(p)$ with its correct probability. Therefore, we get $G(p) \cup G_2 \sim \widetilde{G}(p)$, so $\widetilde{G}(p)$ arises by adding edges to $G(p)$, which cannot destroy a monotone property. This implies the claim of the Lemma. ♠

Now we have the tools to prove Lemma 5, which we repeat here for convenient reference.

Lemma 5. *Let $X = [\xi_{ij}]$ be a random 0–1 matrix of size $N \times N$, drawn from an arbitrary distribution, and let $p \in (0,1)$. If X is p-robust* (see Definition 12), *then*

$$\Pr(X \text{ has a perfect matching}) \geq 1 - (N+1)^3(1-p)^{\lfloor N/2 \rfloor}.$$

Proof. Let $Y = [\eta_{i,j}]$ be another random 0–1 matrix of size $N \times N$, and choose each entry of Y *independently*, with $\Pr(\eta_{ij} = 1) = p$. If Y is viewed as the bi-adjacency matrix of a bipartite graph, then this will be a random bipartite graph with bipartition (V_1, V_2), $|V_1| = |V_2| = N$, in which each edge between V_1, V_2 is included with probability p. This bipartite random graph model is often denoted by $\mathcal{G}(N, N, p)$. In our terminology of the p-random subgraph (see Definition 13), $\mathcal{G}(N, N, p)$ is a p-random subgraph of the complete bipartite graph $K_{N,N}$. Similarly, the p-robust random 0–1 matrix X is the bi-adjacency matrix of a p-robust random subgraph (see Definition 14) of $K_{N,N}$; let us denote this p-robust random subgraph by $\widetilde{K}_{N,N}(p)$. We assume that the edges are numbered consistently with the matrix precedence relation, see the remark after Definition 14. Then Lemma 6 implies that for any monotone graph property $\mathcal{Q}$

$$\Pr\big(\, \mathcal{G}(N,N,p) \in \mathcal{Q} \,\big) \;\leq\; \Pr\big(\, \widetilde{K}_{N,N}(p) \in \mathcal{Q} \,\big) \tag{30}$$

is satisfied.

Now let us estimate the probability that the random bipartite graph $\mathcal{G}(N, N, p)$ has no perfect matching[6]. By Hall's Theorem[7], the graph has no perfect matching if and only if it contains a set $A \subseteq V_1$ or $A \subseteq V_2$, such that $|A| > |\Gamma(A)|$. Let us call such a set A a *Hall violator.*

First we show that if there is a Hall violator A, then there is also one with $|A| \leq \lceil N/2 \rceil$. To prove it, let A_0 be a minimal Hall violator. If $|A_0| = 1$, then we are done, as $1 \leq \lceil N/2 \rceil$. If $|A_0| \geq 2$, then for any $x \in A_0$ it holds that $A_0 - x$ is already not a Hall violator. Therefore, we must have $|\Gamma(A_0)| = |A_0| - 1$, since otherwise we could leave a vertex out of A_0, and still obtain a Hall violator, contradicting to the minimality of A_0.

Assume, without loss of generality, that $A_0 \subseteq V_1$. If $|A_0| \leq \lceil N/2 \rceil$, then we found a Hall violator of size $\leq \lceil N/2 \rceil$. Else, if $|A_0| > \lceil N/2 \rceil$, then set $B = V_2 - \Gamma(A_0)$. Due to $|\Gamma(A_0)| = |A_0| - 1 > \lceil N/2 \rceil - 1$, we have $|\Gamma(A_0)| \geq \lceil N/2 \rceil$, so $|B| \leq N - \lceil N/2 \rceil \leq \lceil N/2 \rceil$. As $B \cap \Gamma(A_0) = \emptyset$, therefore, B has no neighbor in A_0, yielding $\Gamma(B) \cap A_0 = \emptyset$. Then

$$|\Gamma(B)| \;\leq\; N - |A_0| \;=\; N - \underbrace{(|\Gamma(A_0)| + 1)}_{|A_0|} \;=\; N - (\underbrace{N - |B|}_{|\Gamma(A_0)|} + 1) \;\leq\; |B| - 1,$$

giving $|\Gamma(B)| < |B|$, so B is a Hall violator, too. But we have shown that $|A_0| > \lceil N/2 \rceil$ implies $|B| \leq \lceil N/2 \rceil$. Thus, either A_0 or B is a Hall violator of size $\leq \lceil N/2 \rceil$. This proves that if there is a Hall violator, then there is also one with size $\leq \lceil N/2 \rceil$.

[6] This is often analyzed in the random graphs literature, see, e.g., Bollobás [4], Janson, Luczak, and Ruciński [15], Frieze and Karoński [13], etc. However, the end-results are typically presented as asymptotic expressions; we could not find a definite reference to a simple, explixit, *non-asymptotic* formula for finite graphs, so we derive our own.

[7] Originally published in 1935, see Hall [14]; now it is found in almost all textbooks on graph theory.

Let $\mathcal{M}$ denote the event that $\mathcal{G}(N,N,p)$ has a perfect matching, and its negation is denoted by $\overline{\mathcal{M}}$. Further, let $\mathcal{H}$ be the family of all Hall violators of size $\leq \lceil N/2 \rceil$. We already know that if there is a Hall violator, then there is also one of size $\leq \lceil N/2 \rceil$, which implies $\overline{\mathcal{M}} = \{\mathcal{H} \neq \emptyset\}$. Consequently, $\Pr(\overline{\mathcal{M}}) = \Pr(\mathcal{H} \neq \emptyset)$.

Let us now estimate $\Pr(\mathcal{H} \neq \emptyset)$. If $A \subseteq V_1$ with $|A| = k$, then $A \in \mathcal{H}$ if and only if $1 \leq k \leq \lceil N/2 \rceil$, and A has at least $N-k+1$ non-neighbors in V_2. The latter means, there exists a set $B \subseteq V_2$ with $|B| = N-k+1$, such that there is no edge between A and B. With the notation $q = 1-p$, the probability of having no edge between A and B is $q^{k(N-k+1)}$ (Recall that the edges in $\mathcal{G}(N,N,p)$ are independent.)

To estimate $\Pr(\mathcal{H} \neq \emptyset)$, observe that A can be chosen $\binom{N}{k}$ different ways from V_1, with $1 \leq k \leq \lceil N/2 \rceil$. For each such A, the set B can be chosen in $\binom{N}{N-k+1}$ different ways from V_2. Furthermore, we can also symmetrically reverse the positions of A, B, so that they trade sides, meaning that $A \subseteq V_2, B \subseteq V_1$, leading to the same probabilities. Thus, we obtain

$$\Pr(\overline{\mathcal{M}}) \;=\; \Pr(\mathcal{H} \neq \emptyset) \;\leq\; 2 \sum_{k=1}^{\lceil N/2 \rceil} \binom{N}{k} \binom{N}{N-k+1} q^{k(N-k+1)}.$$

To bound the right-hand side from above, we use the symmetry rule $\binom{N}{\ell} = \binom{N}{N-\ell}$, and the (crude) upper bound $\binom{N}{\ell} \leq N^\ell$. Furthermore, $q^{k(N-k+1)} \leq q^{k(N-k)}$. Then we get

$$\Pr(\overline{\mathcal{M}}) \;\leq\; 2 \sum_{k=1}^{\lceil N/2 \rceil} \binom{N}{k} \binom{N}{k-1} q^{k(N-k)} \;\leq\; 2 \sum_{k=1}^{\lceil N/2 \rceil} N^{2k} q^{k(N-k)} \;=\; 2 \sum_{k=1}^{\lceil N/2 \rceil} (N^2 q^{N-k})^k. \tag{31}$$

Since $k \leq \lceil N/2 \rceil$, therefore, $q^{N-k} \leq q^{N-\lceil N/2 \rceil} = q^{\lfloor N/2 \rfloor}$. Hence, we can continue (31) as

$$\Pr(\overline{\mathcal{M}}) \;\leq\; 2 \sum_{k=1}^{\lceil N/2 \rceil} (N^2 q^{N-k})^k \;\leq\; 2 \sum_{k=1}^{\lceil N/2 \rceil} (N^2 q^{\lfloor N/2 \rfloor})^k.$$

If $N^2 q^{\lfloor N/2 \rfloor} < 1$, then the largest summand in the last summation occurs at $k=1$, yielding

$$\Pr(\overline{\mathcal{M}}) \;\leq\; 2\lceil N/2 \rceil N^2 q^{\lfloor N/2 \rfloor} \;\leq (N+1)^3 q^{\lfloor N/2 \rfloor}, \tag{32}$$

which gives us the lower bound

$$\Pr(\mathcal{M}) \;=\; \Pr\big(\,\mathcal{G}(N,N,p) \text{ has a perfect matching}\,\big) \;\geq\; 1 - (N+1)^3 (1-p)^{\lfloor N/2 \rfloor}. \tag{33}$$

In this derivation we used the assumption $N^2 q^{\lfloor N/2 \rfloor} < 1$. However, if $N^2 q^{\lfloor N/2 \rfloor} \geq 1$, then (32) still holds, just the upper bound becomes trivial, being larger than 1. Similarly, (33) becomes trivial with a negative lower bound. Nevertheless, in any case the results (32) and (33) provide valid bounds.

Recall now that having a perfect matching is a monotone property. Therefore, combining (33) with (30), where the latter is based on Lemma 6, gives:

$$\begin{aligned} 1 - (N+1)^3 (1-p)^{\lfloor N/2 \rfloor} \;&\leq\; \Pr\big(\,\mathcal{G}(N,N,p) \text{ has a perfect matching}\,\big) \\ &\leq\; \Pr\big(\,\widetilde{K}_{N,N}(p) \text{ has a perfect matching}\,\big). \end{aligned} \tag{34}$$

By Definition 10, a perfect matching in $\widetilde{K}_{N,N}(p)$ is in 1–1 correspondence with a perfect matching in the 0–1 matrix X. Thus, we get from (34) that

$$\Pr(X \text{ has a perfect matching}) \;\geq\; 1 - (N+1)^3 (1-p)^{\lfloor N/2 \rfloor}$$

holds, completing the proof.

♠

B.3 A Property of Random Perfect Matchings in Random 0–1 Matrices

The next lemma provides a connection between the entries that belong to a random perfect matching in a 0–1 matrix, and a randomly chosen 1-entry in a column or a row. This connection will be very useful in the proof of Theorem 1.

We say that a perfect matching in a constant 0–1 matrix X is *random*, if it is chosen independently and uniformly at random from among all perfect matchings of X. If the matrix X itself is also random, then generating a random perfect matching from X means that we first pick a realization A from the distribution that generates X, and then generate a random perfect matching from A.

Lemma 7 considers a property of random perfect matchings in random 0-1 matrices. We allow that the matrix may come from an arbitrary distribution, but it is assumed to have a perfect matching. Then we average out its row distributions by randomly permuting the set of rows, but keeping the original ordering *within* each row. Then the property we show is this: if a random matching is drawn from this matrix, and we look at any fixed column, then any 1-entry in this column have the same probability of falling into the random matching. In other words, it behaves like a randomly chosen 1-entry from the column (but not independently for different columns). Due to the symmetry resulting from the random permutation of the row set, and from the randomly chosen matching, this claim is intuitively not hard to believe. Nevertheless, it takes some effort to formally prove it.

We use the notation that if Z is a matrix and a permutation α is applied to the set of rows of Z, then the resulting matrix is denoted by $\alpha(Z)$.

Lemma 7 *Let Y be a random 0-1 matrix of size $T \times T$. It can be drawn from an arbitrary distribution, but it is assumed that Y has a perfect matching with probability 1. Create a new matrix $\widetilde{Y} = \alpha(Y)$ by applying an independent random permutation α to the set of rows of Y. Let M be a uniform random perfect matching in $\widetilde{Y}$, and fix a constant $t \in [T]$. Then any 1-entry in column t of $\widetilde{Y}$ has equal probability of falling into M, that is, the 1-entry that falls in M behaves as a randomly chosen 1-entry from the column. Similarly, if α is applied to the set of columns, rather than to the set of rows, and $i \in [T]$ is a constant, then any 1-entry in row i of $\widetilde{Y}$ has equal probability of falling into M, so it behaves as a randomly chosen 1-entry from the row. This latter claim we refer to as the* row version *of the lemma.*

Proof of Lemma 7. First note that by assumption Y has a perfect matching. This is not eliminated by reordering the rows, so $\widetilde{Y} = \alpha(Y)$ also has a perfect matching. Let us represent a uniform random perfect matching in $\widetilde{Y}$ as $M = (e_1, \ldots, e_T)$, with $e_t = (i(t), t)$, $t = 1, \ldots, T$. Here $i(t)$ means the row index of the matching entry in column t; its column index is t.

Fix two integers a, b with $1 \le a \le b \le T$, and consider the a^{th} and b^{th} 1-entries in column t of $\widetilde{Y}$, encountered by scanning the column from top to bottom. Note that the column must contain at least one 1-entry, due to the existence of a perfect matching. (If there is precisely one 1-entry in the column, then $a = b$ must hold.) Assume the a^{th} and b^{th} 1-entries in the column fall in the rows indexed by ν_a and ν_b, respectively (the $\nu_a \le \nu_b$ indices are random variables, but a, b are constants). We are going to show that

$$\Pr(i(t) = \nu_a) = \Pr(i(t) = \nu_b) \tag{35}$$

holds, which means that any two 1-entries in column t of $\widetilde{Y}$ have the same probability of falling into M. This is what the lemma claims.

Note that if $a = b$, then (35) is trivial, so from now on we assume $a < b$, and, consequently, $\nu_a < \nu_b$.

At this point we could informally refer to the following symmetry argument: (35) follows from the fact that we applied an independent uniform random permutation to the set of rows, so for every matching with $i(t) = \nu_a$ there is another, equally likely, matching with $i(t) = \nu_b$, resulting from exchanging the two rows. Nevertheless, as ν_a, ν_b point to random rows in a random matrix, in which the entries are not assumed independent, this reasoning may not be entirely obvious, so we detail a more precise argument below.

Let $\widetilde{Y}_0$ represent a fixed realization of the random matrix $\widetilde{Y}$, with $\Pr(\widetilde{Y} = \widetilde{Y}_0) > 0$. Now, by the law of total probability, we can write:

$$\Pr(i(t) = \nu_a) =$$

$$\sum_{p=1}^{T-1} \sum_{q=p+1}^{T} \sum_{\widetilde{Y}_0} \Pr(i(t) = p \,|\, \widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) \Pr(\widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) \tag{36}$$

where p, q run over all possible values in $[T]$ with $p < q$ (due to $\nu_a < \nu_b$), and $\widetilde{Y}_0$ runs over all possible realizations of $\widetilde{Y}$ that satisfy $\Pr(\widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) > 0$. The latter implies that $\widetilde{Y}_0$ must have the property that the a^{th} and b^{th} 1-entries in column t must occur in rows p and q, respectively. The reason is that if the constant matrix $\widetilde{Y}_0$ does not have these properties, then the event $\{\widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q\}$ would be impossible, so it would not have positive probability. In addition, $\widetilde{Y}_0$ must have a perfect matching, being a realization of $\widetilde{Y}$.

For fixed integers a, b, with $1 \leq a < b \leq T$, let $\Psi(p, q)$ denote the set of possible such $\widetilde{Y}_0$ realizations. That is, $\Psi(p, q)$ is the set of all $[T] \times [T]$ sized 0-1 matrices that have a perfect matching, and in which there are at least two 1-entries in column t, and, furthermore, the a^{th} and b^{th} 1-entries among them occur in rows p and q, respectively. It follows that the inequality $\Pr(\widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) > 0$ implies $\widetilde{Y}_0 \in \Psi(p, q)$. In turn, $\widetilde{Y}_0 \in \Psi(p, q)$ yields that the event $\{\widetilde{Y} = \widetilde{Y}_0\}$ implies the event $\{\nu_a = p, \nu_b = q\}$. The reason for the latter implication is this: $\widetilde{Y}_0 \in \Psi(p, q)$ entails that the a^{th} and b^{th} 1-entries in column t of $\widetilde{Y}_0$ occur in rows p and q, respectively, by the definition of $\Psi(p, q)$. Furthermore, by definition, the random variables ν_a, ν_b point to the a^{th} and b^{th} 1-entries in the column, and these are in rows p, q, respectively, in $\widetilde{Y}_0$. Thus $\{\widetilde{Y} = \widetilde{Y}_0\}$ indeed implies $\{\nu_a = p, \nu_b = q\}$. Using this, we can rewrite the probabilities in (36) as follows:

$$\Pr(i(t) = p \mid \widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) \;=\; \Pr(i(t) = p \,|\, \widetilde{Y} = \widetilde{Y}_0),$$
$$\Pr(\widetilde{Y} = \widetilde{Y}_0, \nu_a = p, \nu_b = q) \;=\; \Pr(\widetilde{Y} = \widetilde{Y}_0).$$

Substituting these into (36), we get

$$\Pr(i(t) = \nu_a) = \sum_{p=1}^{T-1} \sum_{q=p+1}^{T} \sum_{\widetilde{Y}_0 \in \Psi(p,q)} \Pr(i(t) = p \,|\, \widetilde{Y} = \widetilde{Y}_0) \Pr(\widetilde{Y} = \widetilde{Y}_0). \tag{37}$$

Note that the last summation, despite its appearance, does not give $\Pr(i(t) = p)$, since $\widetilde{Y}_0 \in \Psi(p, q)$ means that $\widetilde{Y}_0$ runs over only a subset of all possible realizations of $\widetilde{Y}$.

Now for fixed p, q, let us swap the p^{th} and q^{th} rows in the fixed realization $\widetilde{Y}_0$, denoting the the obtained matrix by $\widetilde{Y}_0^{(p,q)}$ (the other rows do not change). This transformation does not alter

the number of 1-entries in column t. Furthermore, the a^{th} and b^{th} 1-entries among them, counting from top to bottom, still occur in rows p and q, respectively. Therefore, $\widetilde{Y}_0 \in \Psi(p,q)$ if and only if $\widetilde{Y}_0^{(p,q)} \in \Psi(p,q)$. In other words, the mapping $\widetilde{Y}_0 \mapsto \widetilde{Y}_0^{(p,q)}$ creates a bijection of $\Psi(p,q)$ to itself. It also induces a bijection on the set of perfect matchings: to every perfect matching M in $\widetilde{Y}_0$ corresponds, in a 1-1 way, a perfect matching $M^{(p,q)}$ in $\widetilde{Y}_0^{(p,q)}$, where the two matchings are mapped into each other by the row exchange. Consequently, the probability that M is obtained from $\widetilde{Y}_0$ is the same as the probability that $M^{(p,q)}$ is obtained from $\widetilde{Y}_0^{(p,q)}$. Furthermore, if $i(t)$, $i'(t)$ mark the row indices of the entries of M and $M^{(p,q)}$ in column t, respectively, then we have that $i(t) = p$ if and only if $i'(t) = q$, and $i(t) = q$ if and only if $i'(t) = p$. These considerations yield

$$\Pr(i(t) = p \,|\, \widetilde{Y} = \widetilde{Y}_0) = \Pr(i(t) = q \,|\, \widetilde{Y} = \widetilde{Y}_0^{(p,q)}).$$

Furthermore, the fact that an independent uniform random permutation has been applied to the set of rows in Y to obtain $\widetilde{Y}$ also provides $\Pr(\widetilde{Y} = \widetilde{Y}_0) = \Pr(\widetilde{Y} = \widetilde{Y}_0^{(p,q)})$. Using these in (37), we obtain that (37) is equal to

$$\sum_{p=1}^{T-1} \sum_{q=p+1}^{T} \sum_{\widetilde{Y}_0^{(p,q)} \in \Psi(p,q)} \Pr(i(t) = q \,|\, \widetilde{Y} = \widetilde{Y}_0^{(p,q)}) \Pr(\widetilde{Y} = \widetilde{Y}_0^{(p,q)}). \tag{38}$$

Now recall that for $\widetilde{Y}_0 \in \Psi(p,q)$ the event $\{\widetilde{Y} = \widetilde{Y}_0\}$ implies the event $\{\nu_a = p, \nu_b = q\}$. Since switching to $\widetilde{Y}_0^{(p,q)}$ does not change the considered column t, this carries over to $\widetilde{Y}_0^{(p,q)}$: the event $\{\widetilde{Y} = \widetilde{Y}_0^{(p,q)}\}$ implies $\{\nu_a = p, \nu_b = q\}$. Therefore, we can write:

$$\Pr(i(t) = q \,|\, \widetilde{Y} = \widetilde{Y}_0^{(p,q)}) \;=\; \Pr(i(t) = q \,|\, \widetilde{Y} = \widetilde{Y}_0^{(p,q)}, \nu_a = p, \nu_b = q), \tag{39}$$

$$\Pr(\widetilde{Y} = \widetilde{Y}_0^{(p,q)}) \;=\; \Pr(\widetilde{Y} = \widetilde{Y}_0^{(p,q)}, \nu_a = p, \nu_b = q) \tag{40}$$

Substituting these into (38), we get the expression

$$\sum_{p=1}^{T-1} \sum_{q=p+1}^{T} \sum_{\widetilde{Y}_0^{(p,q)} \in \Psi(p,q)} \Pr(i(t) = q \,|\, \widetilde{Y} = \widetilde{Y}_0^{(p,q)}, \nu_a = p, \nu_b = q) \times$$

$$\Pr(\widetilde{Y} = \widetilde{Y}_0^{(p,q)}, \nu_a = p, \nu_b = q). \tag{41}$$

From equations (39), (40), we know that the expression (41) is equal to (38). The latter, however, has been shown to be equal to the right-hand side of (37), which, in turn, is equal to $\Pr(i(t) = \nu_a)$. Thus, we obtain that expression (41) is equal to $\Pr(i(t) = \nu_a)$.

At the same time, we can observe that (41) has the same structure as (36), just $i(t) = p$ is replaced by $i(t) = q$, and $\widetilde{Y}_0$ is replaced by $\widetilde{Y}_0^{(p,q)}$. Thus, applying the law of total probability to (41), the same way as in (36), only with these replacements, and noting that $\widetilde{Y}_0^{(p,q)}$ runs over $\Psi(p,q)$, just like $\widetilde{Y}_0$, we get that (41) also equals to $\Pr(i(t) = \nu_b)$. As a result, we obtain the equality claimed in (35):

$$\Pr(i(t) = \nu_a) = \Pr(i(t) = \nu_b)$$

what we wanted to prove. The claim about the rows follows from applying all the above to the transpose of the matrix, completing the proof.

♠

B.4 Random Hits to a Subset

Finally we are ready to prove Theorem 1. For handy reference, we repeat the Theorem below.

Theorem 1. (Subset Sampling Theorem) *Let $X = (X_1, X_2, \ldots, X_T)$ be an equidistributed sequence of random variables, with common range S, and common distribution π. Further, let $H \subseteq S$ be a p-robust subset for X. Pick a sample Z from the sequence X, as follows: if there is an H-hit in X, then select Z as a random H-hit from X. If there is no H-hit in X, then set $Z = z_0$, where $z_0 \in H$ is any fixed element of H. Let α denote the distribution of Z, and let $\pi_H = \pi/\pi(H)$ denote the distribution π conditioned on H, i.e., the distribution of the H-restricted sample. Then*

$$s(\alpha, \pi_H) \leq \frac{2}{n^{k/3}}$$

holds, where n, k are integer parameters that satisfy the following conditions:

$$T = n^k, \quad |S| \leq 2^n, \quad \min_{x \in S} \pi(x) \geq 2^{-n}, \quad p \geq 1/n, \quad k \geq 7, \quad n \geq 3k^{1/5}.$$

Proof. Let us take T independent versions of X. List the versions in an independent random order, but keeping the original ordering of the entries *within* each version. Let the obtained sequences be denoted by $X^i = (X_1^i, \ldots, X_T^i)$, $i = 1, \ldots, T$. Then each X^i becomes an independent, statistically indistinguishable version of X. We arrange them into a matrix, so that X^i becomes the i^{th} row.

Our proof strategy will be that we pick the output sample Z from a randomly selected row, following the sampling rule described in the Theorem. To carry out this proof strategy, we first analyze the case when the sample is selected from a *column* of the matrix, rather than a row. The advantage of considering columns is that they contain independent entries, by the construction. Of course, the columns are not actually available to us, but, as we are going to see, they have strong statistical connections with the random rows. Then we will be able to return to the rows, making the connection via a random perfect matching, and at the end we utilize that the random rows are statistically indistinguishable from the original input. All this takes a somewhat lengthy analysis, using the tools we have developed in Sections B1, B2, B3, but the final result certainly justifies the effort.

To elaborate the details, arrange the T sequences X^i, $i = 1, \ldots, T$, into a $T \times T$ matrix $\mathbf{B} = [X_t^i]$, such that the sequence $(X_1^i, \ldots, X_T^i)$ is put in the i^{th} row. We systematically use the notation that the superscript is the row index, and the subscript is the column index. Let us introduce a transformation Ψ defined by replacing each entry of the matrix by 1 if it falls in H, and 0 otherwise; we call the resulting $T \times T$ sized 0-1 matrix $\mathbf{B}' = \Psi(\mathbf{B})$ the *0-1 skeleton* of the original.

Observe now that the p-robust property of the set H implies that each entry of $\mathbf{B}'$ takes the value 1 with probability at least p, regardless of the values of the preceding entries in its row. Furthermore, the rows are independent by the construction, so this property extends to all entries in other rows. That is, any entry takes the value 1 with probability at least p, no matter what values the preceding entries take in its row, *and* what values any entries take in other rows. These features are not altered by the random ordering of the set of rows. Therefore, $\mathbf{B}'$ is a p-robust 0-1 matrix (see Definition 12). Let $\mathcal{M}$ denote the event that $\mathbf{B}'$ has a perfect matching (see Definition 10). Then Lemma 5 provides, using $T = N$ (as the matrix size in Lemma 5 was $N \times N$):

$$\Pr(\mathcal{M}) \geq 1 - p_0 \quad \text{with} \quad p_0 = (T+1)^3 (1-p)^{\lfloor T/2 \rfloor} \tag{42}$$

With the assumed values $T = n^k$ and $p \geq 1/n$, we get

$$p_0 \leq (T+1)^3 \left(1 - \frac{1}{n}\right)^{\lfloor T/2 \rfloor} \leq (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3} \tag{43}$$

where in the last inequality we applied the bounding $\left(1-\frac{1}{n}\right)^{\lfloor T/2 \rfloor} \leq \left(1-\frac{1}{n}\right)^{T/3} = \left(1-\frac{1}{n}\right)^{n\frac{T}{3n}} \leq \mathrm{e}^{-\frac{T}{3n}}$. (Here we used $\lfloor T/2 \rfloor \geq T/3$, which holds for all $T \geq 2$. The latter is indeed the case due to $T = n^k$, $k \geq 7$, and $n \geq 3k^{1/5}$.)

Let us condition now on the event $\mathcal{M}$, so we assume that $\mathbf{B}'$ has a perfect matching. (By (42) and (43) this holds with probability exponentially close to 1.) Pick a random perfect matching M uniformly at random among all perfect matchings of $\mathbf{B}'$. Note that we use the random perfect matching M only in the proof, we do not actually have to find it algorithmically[8]. Observe that the condition $\mathcal{M}$ guarantees the existence of a 1-entry in each column of $\mathbf{B}'$, which means that an H-hit exists in each column of $\mathbf{B}$.

Next, we would like to apply Lemma 7 to show that in any fixed column t, any 1-entry of column t has the same chance to belong to the random matching. That is, the entry that belongs to M in column t can be viewed as having been chosen uniformly at random among the 1-entries in the column (but, of course, not independently in different columns, as they have to fall in different rows).

To show the above via Lemma 7, we need to check that the conditions of the Lemma are satisfied. Since we condition on $\mathcal{M}$, therefore, the existence of a perfect matching is guaranteed. Note that conditioning may change the probability distribution (in particular, it may destroy independence), but Lemma 7 allows that the matrix is drawn from an arbitrary distribution, just it needs to have a perfect matching with probability 1, which indeed holds if we condition on $\mathcal{M}$. Furthermore, originally the independent copies of X are listed in an independent random order in the construction of the matrix $\mathbf{B}$, so the random permutation condition of the row set in Lemma 7 is also satisfied. Note that the existence of a perfect matching is not affected by the permutation, so it does not matter whether we condition on $\mathcal{M}$ first, and apply the random permutation after, or the other way around. Thus, under the condition $\mathcal{M}$, the 0-1 matrix $\mathbf{B}'$ indeed satisfies the assumptions of Lemma 7.

Therefore, Lemma 7 guarantees that the 1-entries in any given column of the 0-1 skeleton have equal probability of falling into the random perfect matching. It then implies that the corresponding entry of $\mathbf{B}$ behaves the same way as a random H-hit within the same column of $\mathbf{B}$. As a result, the random perfect matching M induces a random H-hit in each column. Due to the definition of perfect matching in a 0-1 matrix (see Definition 10), these random H-hits must fall in different rows, so they are not independent of each other. But *within any fixed column* the entry selected by the random perfect matching is indistinguishable from a random H-hit in the column (under the condition $\mathcal{M}$).

Let ϑ_t denote the row index of the entry that belongs to M in column t of $\mathbf{B}'$. Then the corresponding entry of $\mathbf{B}$ is $X_t^{\vartheta_t}$. Observe that $X_t^{\vartheta_t}$ has constant column index t, but random row index ϑ_t. Let us take a closer look at the entry $X_t^{\vartheta_t}$ which is a random H-hit from column t of $\mathbf{B}$, under the condition $\mathcal{M}$. We want to apply Lemma 4 to show that the distribution of this random H-hit is close, in terms of the separation distance, to the distribution π_H of an H-restricted sample from X, even under the condition $\mathcal{M}$. Note that this applies to a *column*, and none of the columns

[8]Nevertheless, let us mention that there exists an FPRAS to generate an almost uniform random perfect matching in balanced bipartite graphs, see Jerrum, Sinclair and Vigoda [18], but we are not going to need it.

are actually available as input. Therefore, it is quite different from the original sampling rule, but later we will be able to return to the rows.

For the precise analysis, let us examine first how a column of $\mathbf{B}$ behaves in the *unconditional* case, i.e., when the condition $\mathcal{M}$ is not assumed. Recall that in the original sequence each entry has distribution π, and each row of $\mathbf{B}$ is an independent, statistically indistinguishable version of X. Therefore, for any fixed t, the entries in *column* t are i.i.d., each with distribution π, in the unconditional case. This is not influenced by the random permutation. Consequently, the corresponding column t in the 0–1 skeleton $\mathbf{B}'$ is an i.i.d. sequence of Bernoulli random variables, each taking the value 1 with probability $\Pr(X_t^i \in H) = \pi(H)$. If ν denotes their sum, i.e., the number of 1-entries in this column, which is the same as the number of H-hits in the same column of $\mathbf{B}$, then we have $\mathrm{E}(\nu) = \pi(H)T$. Note that in the unconditional case it is possible that $\nu = 0$, i.e., there may be no H-hit in the column at all. Nevertheless, for this unconditional i.i.d. case the following two-sided version of the Chernoff bound[9] gives us a concentration bound for the number of H-hits, with any $\rho \in (0, 1)$:

$$\Pr\big((1-\rho)\mathrm{E}(\nu) \le \nu \le (1+\rho)\mathrm{E}(\nu) \big) \ge 1 - 2\mathrm{e}^{-\mathrm{E}(\nu)\rho^2/3} \tag{44}$$

Using $\mathrm{E}(\nu) = \pi(H)T$ and choosing $\rho = 1/T^{1/3}$, we obtain that ν is (ρ, ε)-concentrated with

$$\rho = \frac{1}{T^{1/3}} \quad \text{and} \quad \varepsilon = 2\mathrm{e}^{-\frac{\pi(H)}{3}T^{1/3}} \le 2\mathrm{e}^{-\frac{1}{3n}T^{1/3}}, \tag{45}$$

where the upper bound on ε is implied by $\pi(H) \ge p \ge 1/n$. Here $\pi(H) \ge p$ follows from the p-robust property of H, and $p \ge 1/n$ was assumed among the conditions of Theorem 1. With $T = n^k$, (45) becomes

$$\rho = \frac{1}{n^{k/3}} \quad \text{and} \quad \varepsilon \le 2\mathrm{e}^{-\frac{1}{3}n^{\frac{k}{3}-1}}, \tag{46}$$

Note that the assumption $k \ge 7$ guarantees that ε tends to 0 exponentially fast. (Here $k > 3$ would already be enough, but we need a larger k later.)

We applied the Chernoff bound (44) to column t of the matrix $\mathbf{B}'$. Returning to the matrix $\mathbf{B}$, there the meaning of ν is the number of H-hits in column t. However, once we condition on $\mathcal{M}$, the independence of the entries may be lost, so we cannot directly use the Chernoff bound.

At this point it is natural to ask: why do we want to condition on $\mathcal{M}$? The reason is that the existence of the perfect matching brings in a certain symmetry, which will be very useful later. Specifically, it guarantees that each column and each row contains precisely one matching entry, which behaves like a random H-hit. This will allow creating a useful connection between the column view (which is being presented now) and the row view, which will follow later.

To make up for the lost independence, due to conditioning on $\mathcal{M}$, our plan is to apply Lemma 4, which does *not* require independence. We are going to use Lemma 4 for the sequence $X_1^t, \dots, X_T^t$, under the condition $\mathcal{M}$. About this sequence we know that in the *unconditional* case it is equidistributed with common distribution is π, and the distribution of an H-restricted sample is $\pi_H = \pi/\pi(H)$, as well as the number ν of H-hits is (ρ, ε)-concentrated, with ρ and ε given in (46). All this applies to the unconditional case.

[9] Variants of the Chernoff bound are extensively discussed in the literature, they are found in essentially every textbook/survey related to the analysis of randomized algorithms, see, e.g., Dubhashi and Panconesi [8], McDiarmid [30], Mitzenmacher and Upfal [32], Motwani and Raghavan [33], etc. As an interesting historical note, Chernoff himself attributes the proof of his bound to Herman Rudin, despite the fact that it is named only after Chernoff, see Bather [3].

To apply Lemma 4, let us check whether its assumptions hold here for the sequence $X_1^t, \ldots, X_T^t$, i.e., the entries in column t. Observe that $\Pr(\nu \geq 1|\mathcal{M}) = 1$, since the existence of a perfect matching guarantees the existence of an H-hit in every column, so $\mathcal{M}$ can play the role of condition E in Lemma 4. We also have the lower bound $\Pr(\mathcal{M}) \geq 1 - p_0$ by (42), where p_0 is bounded as $0 \leq p_0 \leq (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3}$ in (43). This plays the role of the assumption $\Pr(E) \geq 1 - p_0$ in the Lemma. Another parameter in Lemma 4 is $q_0 = \min_{x \in S} \pi(x)$. It has to be positive, which is provided by $\min_{x \in S} \pi(x) \geq 2^{-n}$, as assumed in Theorem 1. It also needs to satisfy the requirement $q_0 > p_0$ of Lemma 4. In fact, we satisfy a stronger requirement that will be useful later: p_0/q_0 tends to 0 exponentially fast. Specifically, using $p_0 \leq (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3}$ from (42), we can write

$$\frac{p_0}{q_0} \;\leq\; \frac{(n^k+1)^3 \mathrm{e}^{-n^{k-1}/3}}{2^{-n}} \;= (n^k+1)^3 \mathrm{e}^{-\frac{1}{3}n^{k-1}+n\ln 2} \tag{47}$$

where ln stands for logarithm of base e. The bound (47) implies that p_0/q_0 tends to 0 exponentially fast, whenever $k \geq 3$.

Another condition of Lemma 4 is $\mathrm{E}(\nu) \geq 1$. This is satisfied even with $T \geq n$, since $\mathrm{E}(\nu) = \pi(H)T$, and the p-robustness of H implies $\pi(H) \geq p$. Furthermore, $p \geq 1/n$ was assumed, yielding $\mathrm{E}(\nu) \geq T/n$.

Thus, whenever $T = n^k$, with constant $k \geq 7$, all conditions of Lemma 4 are satisfied. Therefore, with $E = \mathcal{M}$, the Lemma gives[10]

$$s(\sigma^{\mathcal{M}}, \pi_H) \;\leq (\varepsilon + p_0)\left(1 + \frac{T}{q_0 - p_0}\right) + \rho(1 + T^2 p_0) + T^2 p_0 + \frac{p_0}{q_0}(1+T), \tag{48}$$

where $\sigma^{\mathcal{M}}$ is the conditional distribution of a random H-hit in column t of $\mathbf{B}$, given $\mathcal{M}$. Let us denote the right-hand side of (48) by ε_1, and bound its size from above. We can rearrange and partition ε_1 as follows:

$$\varepsilon_1 \;=\; \rho + \underbrace{(\varepsilon + p_0)\left(1 + \frac{T}{q_0 - p_0}\right)}_{S_1} + \underbrace{T^2 p_0(\rho + 1)}_{S_2} + \underbrace{\frac{p_0}{q_0}(1+T)}_{S_3}.$$

We claim that the size of ε_1 is determined by $\rho = 1/n^{k/3}$, because the three terms S_1, S_2, S_3 are all exponentially small when $T = n^k$, $k \geq 7$. For S_2 it directly follows from $p_0 \leq (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3}$. For S_3 it is implied by (47). For the term S_1, observe that by (47) we have $p_0/q_0 \leq 1/2$ when $n \geq n_0$, for an appropriate constant n_0 (we determine n_0 at the end of the proof). Then, also using $q_0 \geq 2^{-n}$, we get

$$1 + \frac{T}{q_0 - p_0} = 1 + \frac{T}{q_0\left(1 - \frac{p_0}{q_0}\right)} \;\leq\; 1 + \frac{T}{q_0/2} \leq T 2^n.$$

Therefore, S_1 can be bounded as

$$S_1 \;\leq (\varepsilon + p_0) T 2^n \;\leq\; \left(2\mathrm{e}^{-\frac{1}{3}n^{\frac{k}{3}-1}} + (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3}\right) n^k 2^n$$

[10]Now one can see why it was useful to assume an event E in Lemma 4 with $E \subseteq \{\nu \geq 1\}$: we can now apply it for conditioning on the existence of a perfect matching. Note that $\mathcal{M} \subseteq \{\nu \geq 1\}$, i.e., $\mathcal{M}$ implies the existence of an H-hit in each column, but $\mathcal{M} \neq \{\nu \geq 1\}$.

where we used (46) and (43). The above bound on S_1 shows that S_1 also vanishes exponentially, whenever $T = n^k$, $k \geq 7$, and $n \geq n_0$, with an appropriate constant n_0. (Observe that regarding the value of k, no constant $k \leq 6$ would be enough for S_1 to vanish.) Hence,

$$s(\sigma^{\mathcal{M}}, \pi_H) \;\leq\; \rho + h(n) \;=\; \frac{1}{n^{k/3}} + h(n) \tag{49}$$

holds, where $h(n) = S_1 + S_2 + S_3$ is exponentially small in n, whenever $T = n^k$, $k \geq 7$, and $n \geq n_0$.

So far we have considered samples from columns. Now we move back to the rows. The reason for starting with columns is that in the unconditional case the entries in a column are *independent,* which allowed us to initially estimate the concentration of the number of H-hits. The same would not have worked for a row, because the entries in a row are not assumed independent, and we do not have any good concentration bound for a row.

Let us now apply an independent random permutation to the *set of columns* of $\mathbf{B}$; denote the resulting matrix by $\mathbf{B_1}$, and its 0–1 skeleton by $\mathbf{B_1'}$. Let us pick a row of $\mathbf{B_1}$ randomly. Select the sample Z from this *row* according to the rule described in the Theorem: Z is a random H-hit, if there is an H-hit in the row, otherwise $Z = z_0$. If there is a perfect matching, then the row version of Lemma 7 guarantees that any two 1-entries in any given row of the 0-1 skeleton have equal probability of falling into the perfect matching. Therefore, if there is a perfect matching, then the entry of the row that falls into the matching is distributed as a random H-hit from the row. Let $\delta_i^{\mathcal{M}}$ be this distribution in row i, i.e., it is both the distribution of the matching entry in the row, and, by Lemma 7, is also the distribution of a random H-hit from the row, given $\mathcal{M}$. Similarly, denote this distribution by $\delta^{\mathcal{M}}$ for the randomly selected row. Since we select this row uniformly at random from among all rows, therefore, the distribution $\delta^{\mathcal{M}}$ is the average of the distributions $\delta_1^{\mathcal{M}}, \ldots, \delta_T^{\mathcal{M}}$. That is,

$$\delta^{\mathcal{M}} \;=\; \frac{1}{T}\sum_{i=1}^{T} \delta_i^{\mathcal{M}}.$$

At the same time, if we average out the distributions of the matching entries by column, rather than by row, still under the condition $\mathcal{M}$, we must get the same result, since each row and column contains precisely one entry of the perfect matching. (This is why the perfect matching is useful for us.) Recall that, given $\mathcal{M}$, the distribution of the matching entry in column t was denoted by $\sigma^{\mathcal{M}}$, so we can write

$$\delta^{\mathcal{M}} \;=\; \frac{1}{T}\sum_{i=1}^{T} \delta_i^{\mathcal{M}} \;=\; \frac{1}{T}\sum_{t=1}^{T} \sigma^{\mathcal{M}} \;=\; \frac{1}{T}(T\sigma^{\mathcal{M}}) \;=\; \sigma^{\mathcal{M}}.$$

Next, we get rid of the condition $\mathcal{M}$. If $\mathcal{M}$ does not hold, then let φ denote the (unknown) distribution of Z. Then with probability $\Pr(\mathcal{M})$, the distribution of Z is $\delta^{\mathcal{M}}$, and with probability $1 - \Pr(\mathcal{M})$ it is φ. Finally, let α be the distribution of Z, regardless of whether or not $\mathcal{M}$ holds. Then, α is given by the following mixture:

$$\alpha \;=\; \Pr(\mathcal{M})\delta^{\mathcal{M}} + \big(\,1 - \Pr(\mathcal{M})\,\big)\,\varphi.$$

Then, by the Mixture Representation Property of the separation distance (see (iv) in Lemma 1), it follows that

$$s(\alpha, \sigma^{\mathcal{M}}) \;\leq\; 1 - \Pr(\mathcal{M}) \;\leq\; p_0 \;\leq (n^k + 1)^3 \mathrm{e}^{-n^{k-1}/3}, \tag{50}$$

where we also used the bounds (42) and (43). By the Triangle Inequality of the separation distance (see (ii) in Lemma 1) we can write, also using (50) and (49),

$$s(\alpha, \pi_H) \;\le\; s(\alpha, \sigma^{\mathcal{M}}) + s(\sigma^{\mathcal{M}}, \pi_H) \;\le (n^k+1)^3 \mathrm{e}^{-n^{k-1}/3} + \frac{1}{n^{k/3}} + h(n).$$

Since the terms $(n^k+1)^3\mathrm{e}^{-n^{k-1}/3}$ and $h(n)$ both vanish exponentially, while $1/n^{k/3}$ vanishes only as an inverse polynomial, therefore, we conclude that with $T = n^k$, $k \ge 7$ and $n \ge n_0$, for an appropriate constant n_0, we have

$$s(\alpha, \pi_H) \;\le\; \frac{2}{n^{k/3}}. \tag{51}$$

The role of the factor 2 is only to "swallow" the exponentially vanishing terms (see more about it at the end of the proof). Thus, we have obtained that the distribution α of Z is $\frac{2}{n^{k/3}}$-close to π_H, whenever $T = n^k$, $k \ge 7$ and $n \ge n_0$, whether or not $\mathcal{M}$ holds.

We still have to show that it does not matter whether the sample Z is selected from a random row (what we actually did), or from the original sequence X. To show it, let θ denote the independent random permutation that we applied to the set of columns to obtain $\mathbf{B_1}$. Similarly, let $\theta(X)$ denote the input reordered by the same permutation. In $\mathbf{B_1}$ the rows become $\theta(X^1), \ldots, \theta(X^T)$, which are the versions of $X^1, \ldots, X^T$, reordered by θ. Recall that Z is selected from a row that was randomly chosen from among $\theta(X^1), \ldots, \theta(X^T)$. As mentioned at the beginning of the proof, each X^i is an independent, statistically indistinguishable version of X. After permuting all of them by θ, we get that each of $\theta(X^1), \ldots, \theta(X^T)$ is an independent, statistically indistinguishable version of $\theta(X)$. Hence, Z has the same *distribution* as if it is selected from $\theta(X)$, by the same sampling rule. (The actual sample values may be different, but the distributions are the same.) Furthermore, since $\theta(X)$ is a randomly permuted version of X, therefore, Z is indeed *distributed* as an entry selected by the original selection rule: take a random H-hit from the input X, whenever there is an H-hit, or else take a fixed entry $z_0 \in H$. At the same time, as we have shown, the distribution α of Z satisfies (51), as desired.

Finally, let us estimate the constant n_0. Note that n_0 was only needed to make the sum of exponentially vanishing expressions smaller than $1/n^{k/3}$, via $n \ge n_0$. Specifically, we considered the following expressions:

$$\begin{aligned} S_1 \le (\varepsilon + p_0)T2^n \;&\le\; \left(2\mathrm{e}^{-\frac{1}{3}n^{\frac{k}{3}-1}} + (n^k+1)^3\mathrm{e}^{-n^{k-1}/3}\right) n^k 2^n \\ &=\; 2n^k\mathrm{e}^{-\frac{1}{3}n^{\frac{k}{3}-1}+n\ln 2} + n^k(n^k+1)^3\mathrm{e}^{-\frac{1}{3}n^{k-1}+n\ln 2} \end{aligned} \tag{52}$$

$$S_2 \;=\; T^2 p_0(\rho+1) \;\le\; n^{2k}(n^k+1)^3\mathrm{e}^{-\frac{1}{3}n^{k-1}} \cdot 2 \tag{53}$$

$$S_3 \;=\; \frac{p_0}{q_0}(1+T) \;\le (n^k+1)^4\mathrm{e}^{-\frac{1}{3}n^{k-1}+n\ln 2} \tag{54}$$

$$p_0 \le (n^k+1)^3\mathrm{e}^{-\frac{1}{3}n^{k-1}} \tag{55}$$

We also needed $p_0/q_0 \le 1/2$, but that is implied by $p_0/q_0 \le S_3$, once S_3 becomes small enough. Thus, n_0 needs to be chosen such that for $n \ge n_0$ we have

$$h(n) + p_0 \;=\; S_1 + S_2 + S_3 + p_0 \le \frac{1}{n^{k/3}}. \tag{56}$$

If (56) is satisfied, then (51) is indeed implied via

$$s(\alpha, \pi_H) \;\le\; p_0 + \frac{1}{n^{k/3}} + h(n) \;\le\; \frac{2}{n^{k/3}},$$

whenever $n \geq n_0$. Now observe that the upper bounds on S_1, S_2, S_3, p_0 in (52), (53), (54), (55), respectively, are all bounded from above by $2n^{2k}(n^k+1)^4 \mathrm{e}^{-\frac{1}{3}n^{k-1}+n\ln 2}$. Therefore, if we choose n_0 such that

$$2n_0^{2k}(n_0^k+1)^4 \mathrm{e}^{-\frac{1}{3}n_0^{k-1}+n_0\ln 2} \;\leq\; \frac{1}{4n_0^{k/3}} \tag{57}$$

is satisfied, then we also satisfy (56), as desired. To find a specific value for n_0, we rearrange (57) as

$$8n_0^{\frac{7}{3}k}(n_0^k+1)^4 \mathrm{e}^{-\frac{1}{3}n_0^{k-1}+n_0\ln 2} \;\leq\; 1. \tag{58}$$

Using $8n_0^{\frac{7}{3}k}(n_0^k+1)^4 \leq 8n_0^{3k}(2n_0^k)^4 = 2^7 n_0^{7k}$, $k \geq 7$, and $\ln 2 < 1$, we obtain that the inequality

$$2^7 n_0^{7k} \mathrm{e}^{-\frac{1}{3}n_0^6+n_0} \;\leq\; 1 \tag{59}$$

implies (58), which in turn implies (57). Rearranging (59), and taking the logarithm yields

$$\frac{1}{3}n_0^6 - n_0 \;\geq\; \ln(2^7 n^{7k}) \;=\; 7k\ln(2^{1/k}n_0). \tag{60}$$

Using $2n_0 \geq \ln(2^{1/k}n_0)$, we get that (60) is implied by $\frac{1}{3}n_0^6 - n_0 \;\geq\; 14kn_0$, which is equivalent to

$$n_0 \;\geq (42k+3)^{1/5}. \tag{61}$$

To further simplify the expression, observe that $3k^{1/5} \geq (42k+3)^{1/5}$, so it is enough to satisfy

$$n_0 \;\geq\; 3k^{1/5}$$

to imply (61), completing the proof.

♠

Appendix C: Proof of Theorem 2

For convenient reference, we repeat Algorithm RISL and the statement of Theorem 2 below. The algorithm uses the BIDC Markov chain, introduced in Section 4.1.

Recall the notation: if $G=(V,E)$ is a graph and $v_0 \in V$ is a vertex in it, then $S(v_0)$ stands for the set of edges adjacent to v_0 in G, and $G - S(v_0)$ denotes the graph obtained by removing the edges in $S(v_0)$ from G. The vertex v_0 then remains an isolated vertex in $G - S(v_0)$.

Algorithm *Random Independent Set Lifting (RISL)*

Input: a graph $G=(V,E)$, a specified vertex $v_0 \in V$, and an independent set $X \in \mathcal{I}(G - S(v_0))$.

Output: an independent set $Z \in \mathcal{I}(G)$.

The Algorithm

Step 1 *(Trajectory construction)* Run BIDC on $\mathcal{I}(G-S(v_0))$ from initial state X. Let $Y=(X_1,\ldots,X_T)$ be the obtained trajectory, where $X_1 = X$, $T = n^k$, $n = |V|$.

Step 2 *(Subset sampling)* Set $H = \mathcal{I}(G)$. If Y contains an H-hit, then select an H-hit Z from Y uniformly at random. If there is no H-hit in Y, then set $Z = \emptyset$. Output Z and HALT.

Theorem 2 *Let k be a fixed positive integer constant with $T = n^k$ in the algorithm RISL, and $n = |V(G)|$. Then the algorithm has the following properties:*

(i) *The running time is $O(n^{k+1})$, using $O(n^k \log n)$ random bits.*

(ii) *Assume the input X is a μ-uniform random independent set over $\mathcal{I}(G - S(v_0))$, for some $\mu \in [0, 1)$. Then the output Z is a μ'-uniform random independent set over $\mathcal{I}(G)$, with*

$$\mu' \leq \mu + \frac{2}{n^{k/3}},$$

whenever the parameters satisfy the following conditions: $n \geq 3k^{1/5}$, $k \geq 7$, and $n \geq \Delta + 2$, where Δ is any upper bound on the maximum degree of G.

(iii) *The algorithm receives G, v_0, X as input, and it also uses the constant k. But the algorithm has no access to the parameter Δ, and the running time is also independent of Δ, assuming that the conditions listed in* (ii) *hold.*

Proof.

(i) Regarding the running time, observe that each step of the BIDC can be executed in $O(n)$ time, using $O(\log n)$ random bits, since it only involves drawing a uniform random vertex from $V(G)$, and updating the current independent set. The algorithm generates a trajectory of polynomial length ($T = n^k$), which can be carried out in $O(n^{k+1})$ time, using $O(n^k \log n)$ random bits. Finally, drawing a random H-hit from the generated trajectory, or assigning $Z = \emptyset$ if there is no H-hit, takes $O(n^k)$ time, with $O(k \log n)$ random bits. Thus, overall, the algorithm runs in $O(n^{k+1})$ time, using $O(n^k \log n)$ random bits.

(ii) We apply Theorem 1 to show that the obtained output Z is μ'-uniform, with the stated bound on μ'. First we treat the ideal case when $\mu = 0$, that is, the BIDC starts from a perfectly uniform initial state on $\mathcal{I}(G - S(v_0))$. After that, we take care of the deviation from this ideal case. For short notation, we use $G' = G - S(v_0)$.

In order to apply Theorem 1, let us start by showing that $H = \mathcal{I}(G)$ is a p-robust subset of $\mathcal{I}(G')$ (see Definition 4), with $p \geq 1/n$, for any sequence $(X_1, \ldots, X_T)$ that is a realization of the Markov chain. For any $X_t \in \mathcal{I}(G')$ we can write

$$\Pr(X_t \in H) = \sum_{A \in \mathcal{I}(G')} \Pr(X_t \in H \mid X_{t-1} = A) \Pr(X_{t-1} = A).$$

Let us look at the conditional probability $\Pr(X_t \in H | X_{t-1} = A)$ within the above summation. If $A \in H$, then A can contain at most one of v_0 and any vertex from $\Gamma_G(v_0)$. Therefore, if the random vertex u in the BIDC transition rule does not fall in $\{v_0\} \cup \Gamma_G(v_0)$, then X_t remains in H. The event that the uniformly chosen vertex u is not in $\{v_0\} \cup \Gamma_G(v_0)$ has probability at least $1 - (\Delta + 1)/n$, since $|\{v_0\} \cup \Gamma_G(v_0)| \leq \Delta + 1$ in a graph of maximum degree at most Δ. This yields $\Pr(X_t \in H \mid X_{t-1} = A) \geq 1 - (\Delta + 1)/n$, whenever $A \in H$. If $A \notin H$, then A must contain v_0. If $u = v_0$, then the step that moves X_{t-1} to X_t deletes v_0 from X_{t-1}, bringing X_t into H. Since $\Pr(u = v_0) = 1/n$, we get $\Pr(X_t \in H \mid X_{t-1} = A) \geq 1/n$ for $A \notin H$. Thus, using $1 - (\Delta + 1)/n \geq 1/n$, which holds for $n \geq \Delta + 2$ (and was assumed among the conditions), we obtain that

$$\Pr(X_t \in H \mid X_{t-1} = A) \geq \frac{1}{n} \tag{62}$$

always holds, whether or not $A \in H$. This yields

$$\begin{aligned}\Pr(X_t \in H) &= \sum_{A \in \mathcal{I}(G')} \underbrace{\Pr(X_t \in H \mid X_{t-1} = A)}_{\geq 1/n} \Pr(X_{t-1} = A) \\ &\geq \frac{1}{n} \underbrace{\sum_{A \in \mathcal{I}(G')} \Pr(X_{t-1} = A)}_{=1} = \frac{1}{n}.\end{aligned}$$

Consider now any history $\mathcal{H} = \{X_{t_1} = A_1, \ldots, X_{t_k} = A_k\}$ of X_t, with $t > t_1 > \ldots > t_k$, and $\Pr(\mathcal{H}) > 0$. Set

$$\mathcal{I}^{\mathcal{H}}(G') = \left\{ A \in \mathcal{I}(G') \mid \Pr(X_{t-1} = A, \mathcal{H}) > 0 \right\}.$$

Note that $\mathcal{I}^{\mathcal{H}}(G') \neq \emptyset$. This can be seen, as follows. If $t_1 = t-1$, then $A_1 \in \mathcal{I}^{\mathcal{H}}(G')$. If $t_1 < t-1$, then any set A_0 that is reachable from A_1 in $t-1-t_1$ steps satisfies $A_0 \in \mathcal{I}^{\mathcal{H}}(G')$. Then we can write

$$\Pr(X_t \in H \mid \mathcal{H}) = \sum_{A \in \mathcal{I}^{\mathcal{H}}(G')} \Pr(X_t \in H \mid X_{t-1} = A, \mathcal{H}) \Pr(X_{t-1} = A \mid \mathcal{H}). \tag{63}$$

Since $t-1 \geq t_1 > \ldots > t_k$, therefore, the Markov property implies

$$\Pr(X_t \in H \mid X_{t-1} = A, \mathcal{H}) = \Pr(X_t \in H \mid X_{t-1} = A) \geq \frac{1}{n},$$

where the inequality follows from (62). Using this in (63), we get

$$\Pr(X_t \in H \mid \mathcal{H}) \geq \frac{1}{n} \underbrace{\sum_{A \in \mathcal{I}^{\mathcal{H}}(G')} \Pr(X_{t-1} = A \mid \mathcal{H})}_{=1} = \frac{1}{n},$$

proving that H is indeed a p-robust subset of $\mathcal{I}(G')$, with $p \geq 1/n$, for any sequence that is a realization of the Markov chain.

Now let us apply Theorem 1 for the sequence $Y = (X_1, \ldots, X_T)$, obtained in Step 2. Consider first the special case when $\mu = 0$, that is, $X_1 = X'$ is distributed by π, the uniform distribution over $\mathcal{I}(G')$. Then each entry of the sequence Y remains distributed by π, since that is the stationary distribution of the BIDC Markov chain. Consequently, Y is an equidistributed sequence with common distribution π.

To apply Theorem 1, first note that the state space is $S = \mathcal{I}(G')$, satisfying $|S| \leq 2^n$, since in an n-vertex graph there are at most 2^n independent sets. With the uniform distribution π this also implies $\min_{x \in S} \pi(x) \geq 2^{-n}$, as required in Theorem 1. The target set is $H = \mathcal{I}(G)$. Note that $\mathcal{I}(G) \subseteq \mathcal{I}(G')$, since every independent set of G is also an independent set in G', as G' arises by removing the edges from G that are adjacent to v_0.

We have already shown that H is p-robust with $p \geq 1/n$, where $n = |V(G)|$. Taking $z_0 = \emptyset \in H$ in Theorem 1, we can see that the sampling rule in Theorem 1 implements exactly the algorithm RISL with the above parameters. Therefore, Theorem 1 implies that in the considered special case (when $\mu = 0$) the distribution of the output sample Z with $T = n^k$ is $\frac{2}{n^{k/3}}$-uniform over $H = \mathcal{I}(G)$, whenever $n \geq 3k^{1/5}$, $k \geq 7$ (constant), and these conditions are indeed assumed.

Now let us consider the general case when the initial vertex X' is not guaranteed to be perfectly uniform, it is only μ-uniform, with some $\mu \geq 0$. Let α be the distribution of X'. Then, by the

Mixture Representation Property of the separation distance (see (iv) in Lemma 1), there exists a distribution γ, such that

$$\alpha = (1-\mu)\pi + \mu\gamma$$

holds. This can be viewed such that we mix two different initial states, let us call them $X^{(1)}$ and $X^{(2)}$. Here $X^{(1)}$ is drawn from the uniform distribution π on $\mathcal{I}(G')$, and it is chosen with probability $1-\mu$, while $X^{(2)}$ is drawn from γ, and is chosen with probability μ. Let $Z^{(1)}$ be the output that corresponds to initial state $X^{(1)}$, and, similarly, let $Z^{(2)}$ be the output corresponding to $X^{(2)}$. Let $\sigma^{(1)}, \sigma^{(2)}$ be the distributions $Z^{(1)}, Z^{(2)}$, respectively. Finally, let β be the distribution of the eventual output, which is obtained as the mixture of the two cases, that is,

$$\beta = (1-\mu)\sigma^{(1)} + \mu\sigma^{(2)}. \tag{64}$$

We have already analyzed, by means of Theorem 1, that if the initial distribution is uniform, then the output will be $\frac{2}{n^{k/3}}$-close to uniform on $\mathcal{I}(G)$, i.e., to π_H. This applies to initial state $X^{(1)}$, yielding

$$s(\sigma^{(1)}, \pi_H) \;\leq\; \frac{2}{n^{k/3}}. \tag{65}$$

For the mixed output distribution β we can write, using the Triangle Inequality of the separation distance (see (iii) in Lemma 1):

$$s(\beta, \pi_H) \;\leq\; s(\beta, \sigma^{(1)}) + s(\sigma^{(1)}, \pi_H). \tag{66}$$

Using the Mixture Representation Property again (now in the opposite direction), we obtain from (64) that

$$s(\beta, \sigma^{(1)}) \leq \mu.$$

Using this, as well as (65), to bound the right-hand side of (66), yields

$$\mu' \;=\; s(\beta, \pi_H) \;\leq\; \mu + \frac{2}{n^{k/3}}. \tag{67}$$

Thus, we have obtained that if the input X is a μ-uniform random independent set over $\mathcal{I}(G')$, then the output Z is a μ'-uniform random independent set over $\mathcal{I}(G)$, with $\mu' \leq \mu + \frac{2}{n^{k/3}}$.

Note that if $\mu + \frac{2}{n^{k/3}} > 1$ happens to be the case, then it is still an upper bound on μ', just it does not pose any restriction, as the separation distance never exceeds 1, according to the Normalized Value Property (see (i) in Lemma 1).

(iii) The algorithm uses the input G, v_0, X to run the BIDC. It also uses the constant k, to determine the number of steps $T = n^k$. But Δ is not used anywhere in the algorithm, it only plays a role in the analysis. This completes the proof.

♠

Appendix D: Proof of Theorem 3

For convenient reference, we repeat the algorithm and the theorem below.

Algorithm *Random Approximator Extension (RAE)*

Input: a graph $G = (V, E)$, a specified vertex $v_0 \in V$, and a random (μ, δ, h)-approximator $\mathcal{F}_0$ of $\mathcal{I}(G - S(v_0))$, with $|\mathcal{F}_0| = m$.

The algorithm also uses a parameter k, which is a fixed constant, inherited from algorithm RISL (see Section 4.2), which is called here as a subroutine.

Output: a random (μ', δ', h')-approximator $\mathcal{F}$ of $\mathcal{I}(G)$, with $|\mathcal{F}| = m$.

The Algorithm

Step 1 Let $X_1, \ldots, X_m$ be the independent sets in $\mathcal{F}_0$. For every X_i, $i = 1, \ldots, m$, run algorithm RISL with input G, v_0, X_i. Let $Z_i \in \mathcal{I}(G)$ be the corresponding output of RISL.

Step 2 Set $\mathcal{F} = \{Z_1, \ldots, Z_m\}$ (multiset). Output $\mathcal{F}$ and HALT.

Theorem 3 *Let k be a fixed positive integer constant, used in algorithm RISL (see* Section 4.2*), whenever it is called as a subroutine. Let $n = |V(G)|$, $m = |\mathcal{F}_0|$. Then algorithm RAE has the following properties:*

(i) *The algorithm runs in $O(mn^{k+1})$ time, using $O(mn^k \log n)$ random bits.*

(ii) *Assume that $\mathcal{F}_0$ is a (μ, δ, h)-approximator of $\mathcal{I}(G - S(v_0))$. Then the output $\mathcal{F}$ of the algorithm is a (μ', δ', h')-approximator of $\mathcal{I}(G)$, with $|\mathcal{F}| = m$, and*

$$\mu' \le \mu + \frac{2}{n^{k/3}}, \qquad \delta' = \delta, \qquad h' \le \mu + \frac{3}{n^{k/3}},$$

whenever the following conditions are satisfied:

$$\delta \ge m^{-1/3}, \quad m \ge (2^\Delta + 1)^3(6 + k \ln n)^3, \quad k \ge 7, \quad n \ge 3k^{1/5}, \quad n \ge \Delta + 2,$$

where Δ is any upper bound on the maximum degree of G.

(iii) *The algorithm receives $G, \mathcal{F}_0, v_0$ as input, and it also uses the constant k. But the algorithm has no access to the parameters Δ, μ, δ, h, and the running time is also independent of these parameters, assuming that the conditions listed in* (ii) *hold.*

Proof. (i) The running time bound directly follows from statement (i) of Theorem 2, taking into account that algorithm RAE runs algorithm RISL m times.

(ii) From Theorem 2 we know that each Z_i, being the output of RISL, is a μ'-uniform independent set over $\mathcal{I}(G)$, with $\mu' \le \mu + \frac{2}{n^{k/3}}$. (Note that for this bound it is also needed that the relations among the parameters that Theorem 2 requires remain valid. This is indeed the case, since these relations are also required by Theorem 3.) It is also clear that due to running the BIDC Markov chains independently from i.i.d. initial states, for the same number of steps, the resulting output samples of RISL will remain i.i.d. Thus, we obtain that $\mathcal{F}$ is a μ'-uniform i.i.d. family of independent sets over $\mathcal{I}(G)$, with $\mu' \le \mu + \frac{2}{n^{k/3}}$. What we still need to show is that $\mathcal{F}$ is also a (μ', δ, h')-approximator of $\mathcal{I}(G)$, that is, it also satisfies

$$\Pr\big((1-\delta)p(v) \le p(v \,|\, \mathcal{F}) \le (1+\delta)p(v) \big) \ge 1 - h' \tag{68}$$

for every $v \in V(G)$, with $h' \le \mu + \frac{3}{n^{k/3}}$.

To examine the approximation properties of $\mathcal{F}$, let us define, for any family $\mathcal{L}$ of independent sets, and for any given vertex $v \in V(G)$, the event

$$F_v(\mathcal{L}) = \big\{ (1-\delta)p(v) \le p(v \,|\, \mathcal{L}) \le (1+\delta)p(v) \big\}. \tag{69}$$

(For the definition of $p(v \mid \mathcal{L})$ see Definition 5.)

In Step 1, let α be the common distribution of $X_1, \ldots, X_m \in \mathcal{I}(G - S(v_0))$ (recall that they are i.i.d.). Let π be the uniform distribution on $\mathcal{I}(G - S(v_0))$. Since each X_i is μ-uniform, i.e., $s(\alpha, \pi) \leq \mu$, therefore, by the Mixture Representation Property of the separation distance (see (iv) in Lemma 1), there exists a distribution γ on $\mathcal{I}(G - S(v_0))$, such that

$$\alpha = (1 - \mu)\pi + \mu\gamma$$

holds. This can be viewed such that when we run RISL from initial state X_i, then we mix two different initial states, let us call them $X_i^{(1)}$ and $X_i^{(2)}$. It means, $X_i^{(1)}$ is drawn from the uniform distribution π, and it is chosen with probability $1 - \mu$, while $X_i^{(2)}$ is drawn from some other distribution γ, and is chosen with probability μ. As $X_1, \ldots, X_m$ are i.i.d., therefore, we can assume that $X_1^{(1)}, \ldots, X_m^{(1)}$ are i.i.d. with common distribution π, and $X_1^{(2)}, \ldots, X_m^{(2)}$ are i.i.d. with common distribution γ.

Let $Z_i^{(1)}$ be the output that corresponds to initial state $X_i^{(1)}$, and, similarly, let $Z_i^{(2)}$ be the output corresponding to $X_i^{(2)}$. Since we run the same Markov chain independently from i.i.d. initial states, therefore, $Z_1^{(1)}, \ldots, Z_m^{(1)}$ are i.i.d., and $Z_1^{(2)}, \ldots, Z_m^{(2)}$ are also i.i.d. Further, let $\sigma^{(1)}, \sigma^{(2)}$ denote the distributions of $Z_i^{(1)}, Z_i^{(2)}$, respectively, for any $i = 1, \ldots, m$. Note that $\sigma^{(1)}, \sigma^{(2)}$ do not depend on i due to the i.i.d. property. Finally, let β be the distribution of the eventual output Z_i (it is the same for each Z_i, as they are i.i.d.). The distribution β is obtained as the mixture of the two cases, that is,

$$\beta = (1 - \mu)\sigma^{(1)} + \mu\sigma^{(2)}. \tag{70}$$

After running the Markov chains, the initial states $X_i^{(1)}, X_i^{(2)}$ are transformed into the outputs $Z_i^{(1)}, Z_i^{(2)}$, respectively. Let $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$ denote the corresponding families (multisets) of independent sets, consisting of the respective end-states, that is,

$$\mathcal{F}^{(1)} = \{Z_1^{(1)}, \ldots, Z_m^{(1)}\} \quad \text{and} \quad \mathcal{F}^{(2)} = \{Z_1^{(2)}, \ldots, Z_m^{(2)}\}.$$

Then the family $\mathcal{F} = \{Z_1, \ldots, Z_m\}$, consisting of the actual end-states, can be viewed as a mixture of $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$, where $\mathcal{F}^{(1)}$ is taken with probability $1 - \mu$, and $\mathcal{F}^{(2)}$ is taken with probability μ. This means, $\Pr(\mathcal{F} = \mathcal{F}^{(1)}) = 1 - \mu$ and $\Pr(\mathcal{F} = \mathcal{F}^{(2)}) = \mu$.

Now let $\mathcal{E}(\mathcal{F})$ be any event that depends on the family $\mathcal{F}$. Then we can write

$$\begin{aligned}
\Pr(\mathcal{E}(\mathcal{F})) \;=\; & \Pr\Big(\mathcal{E}(\mathcal{F}) \mid \mathcal{F} = \mathcal{F}^{(1)}\Big) \underbrace{\Pr\Big(\mathcal{F} = \mathcal{F}^{(1)}\Big)}_{=1-\mu} + \\
& \Pr\Big(\mathcal{E}(\mathcal{F}) \mid \mathcal{F} = \mathcal{F}^{(2)}\Big) \underbrace{\Pr\Big(\mathcal{F} = \mathcal{F}^{(2)}\Big)}_{=\mu} = \\
& (1 - \mu) \Pr\Big(\mathcal{E}(\mathcal{F}^{(1)})\Big) + \mu \Pr\Big(\mathcal{E}(\mathcal{F}^{(2)})\Big)
\end{aligned} \tag{71}$$

Choosing $\mathcal{E}(\mathcal{F}) = F_v(\mathcal{F})$, where the event $F_v(\mathcal{F})$ was introduced in (69), equation (71) implies for every $v \in V(G)$

$$\Pr\big(F_v(\mathcal{F})\big) = (1 - \mu) \Pr\big(F_v(\mathcal{F}^{(1)})\big) + \mu \Pr\big(F_v(\mathcal{F}^{(2)})\big) \geq (1 - \mu) \Pr\big(F_v(\mathcal{F}^{(1)})\big), \tag{72}$$

where the lower bound follows from simply removing the non-negative term $\mu \Pr\big(F_v(\mathcal{F}^{(2)})\big)$.

Looking now at $\mathcal{F}^{(1)}$, we see that each independent set $Z_i^{(1)} \in \mathcal{F}^{(1)}$ has distribution $\sigma^{(1)}$. Since $Z_i^{(1)}$ is the output of algorithm RISL when run from input $X_i^{(1)}$, as well as $X_i^{(1)}$ is uniform on $\mathcal{I}(G - S(v_0))$ (which means μ-uniform with $\mu = 0$, by the Self-Identity Property of the separation distance), therefore we get from Theorem 2 with $\mu = 0$ that

$$s(\sigma^{(1)}, \pi_H) \;\le\; \frac{2}{n^{k/3}}.$$

The latter means, again by the Mixture Representation Property of the separation distance, that $\sigma^{(1)}$ is a mixture of π_H and some other distribution φ, in the form

$$\sigma^{(1)} \;= (1-\varepsilon)\pi_H + \varepsilon\varphi, \quad \text{with} \quad \varepsilon = \frac{2}{n^{k/3}}, \tag{73}$$

where π_H is the uniform distribution over $H = \mathcal{I}(G)$. Therefore, we can represent $\mathcal{F}^{(1)}$ as a mixture of two families, each containing i.i.d. random independent sets: $\mathcal{F}^{(1a)}$, in which each independent set is distributed by π_H, and $\mathcal{F}^{(1b)}$, in which each independent set is distributed by φ. Then we can write, similarly to (72), but using ε instead of μ:

$$\Pr\big(\, F_v(\mathcal{F}^{(1)})\,\big) = (1-\varepsilon)\Pr\big(\, F_v(\mathcal{F}^{(1a)})\,\big) \;+ \varepsilon \Pr\big(\, F_v(\mathcal{F}^{(1b)})\,\big) \;\ge (1-\varepsilon)\Pr\big(\, F_v(\mathcal{F}^{(1a)})\,\big). \tag{74}$$

Let us look at the family $\mathcal{F}^{(1a)}$; let $Z_1^{(1a)}, \ldots, Z_m^{(1a)}$ denote the independent sets in $\mathcal{F}^{(1a)}$. Then, by the definition of the independence ratio (Definition 5), we can write:

$$p(v \,|\, \mathcal{F}^{(1a)}) \;=\; \frac{1}{m}\sum_{i=1}^{m} \chi\{v \in Z_i^{(1a)}\} \;=\; \frac{1}{m}\sum_{i=1}^{m} \xi_i,$$

where $\xi_i = \chi\{v \in Z_i^{(1a)}\}$ denotes the indicator of the event $\{v \in Z_i^{(1a)}\}$. Observe that $\xi_1, \ldots, \xi_m$ are i.i.d. 0–1 valued random variables. Also recall that $Z_i^{(1a)}$ is distributed by π_H, the uniform distribution over $\mathcal{I}(G)$. Therefore, the expected value of $\xi_i = \chi\{v \in Z_i^{(1a)}\}$ is precisely the probability that v falls into an independent set that is selected uniformly at random from $\mathcal{I}(G)$, which is equal to the independence ratio $p(v)$ of v (see the remark after Definition 5). Therefore,

$$\mathrm{E}\left(p(v \,|\, \mathcal{F}^{(1a)})\right) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left(\chi\{v \in Z_i^{(1a)}\}\right) \;=\; \frac{1}{m}\sum_{i=1}^{m} p(v) \;=\; p(v).$$

Then the deviation between $p(v|\mathcal{F}^{(1a)})$ and its expected value $p(v)$ can be bounded by the two-sided version of the Chernoff bound (see, e.g., [8, 32, 33]), which gives for any $\rho \in (0,1)$

$$\Pr\left((1-\rho)\,\mathrm{E}\big(\textstyle\sum_{i=1}^m \xi_i\big) \;\le\; \displaystyle\sum_{i=1}^{m} \xi_i \;\le (1-\rho)\,\mathrm{E}\big(\textstyle\sum_{i=1}^m \xi_i\big)\right) \;\ge\; 1 - 2\mathrm{e}^{-\mathrm{E}\left(\sum_{i=1}^m \xi_i\right)\rho^2/3}.$$

Using $p(v) = \frac{1}{m}\mathrm{E}(\sum_{i=1}^m \xi_i)$ and $p(v|\mathcal{F}^{(1a)}) = \frac{1}{m}(\sum_{i=1}^m \xi_i)$ in the above formula yields

$$\Pr\big(\,(1-\rho)p(v) \le p(v|\mathcal{F}^{(1a)}) \le (1+\rho)p(v)\,\big) \ge 1 - 2\mathrm{e}^{-mp(v)\rho^2/3}.$$

Let us choose $\rho = 1/m^{1/3}$. Substituting it, we get

$$\Pr\left(\left(1 - 1/m^{1/3}\right)p(v) \le p(v \,|\, \mathcal{F}^{(1a)}) \le \left(1 + 1/m^{1/3}\right)p(v)\right) \;\ge\; 1 - 2\mathrm{e}^{-\frac{p(v)}{3}m^{1/3}}.$$

Since we assumed $\delta \geq 1/m^{1/3}$, as well as we know from Lemma 2 that $p(v) \geq \frac{1}{2^{\Delta}+1}$ holds for a graph with maximum degree $\leq \Delta$, therefore, the above probability bound implies

$$\Pr\Big((1-\delta)\,p(v) \leq p(v\,|\,\mathcal{F}^{(1a)}) \leq (1+\delta)\,p(v)\Big) \;\geq\; 1 - 2\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}}. \tag{75}$$

With the notation introduced in (69) we have

$$F_v(\mathcal{F}^{(1a)})\,\big) \;=\; \Big\{(1-\delta)p(v) \leq p(v\,|\,\mathcal{F}^{(1a)}) \leq (1+\delta)p(v)\Big\}\,,$$

so (75) means

$$\Pr\big(\,F_v(\mathcal{F}^{(1a)})\,\big) \;\geq\; 1 - 2\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}}.$$

Now recall, we know from (72) that $\Pr\big(\,F_v(\mathcal{F})\,\big) \geq (1-\mu)\Pr\big(\,F_v(\mathcal{F}^{(1)})\,\big)$ holds, as well as from (74) that $\Pr\big(\,F_v(\mathcal{F}^{(1)})\,\big) \geq (1-\varepsilon)\Pr\big(\,F_v(\mathcal{F}^{(1a)})\,\big)$ also holds. These inequalities together imply

$$\Pr\big(\,F_v(\mathcal{F})\,\big) \geq (1-\mu)(1-\varepsilon)\Pr\big(\,F_v(\mathcal{F}^{(1a)})\,\big),$$

and then using (75) we obtain

$$\Pr\big(\,F_v(\mathcal{F})\,\big) \geq (1-\mu)(1-\varepsilon)\left(1 - 2\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}}\right) \tag{76}$$

where $\varepsilon = \frac{2}{n^{k/3}}$, see (73). To simplify (76), set $a = 2\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}}$. With this, (76) becomes

$$\Pr\big(\,F(\mathcal{F})\,\big) \geq (1-\mu)(1-\varepsilon)(1-a) \;=\; 1 - (\mu+\varepsilon+a+a\varepsilon\mu-\varepsilon\mu-a\mu-a\varepsilon),$$

yielding

$$\Pr\big(\,F(\mathcal{F})\,\big) \;\geq\; 1 - (\mu+\varepsilon+a+a\varepsilon\mu) \;=\; 1-h'$$

with

$$h' \;=\; \mu+\varepsilon+a+a\varepsilon\mu.$$

To bound h', recall that $\varepsilon = \frac{2}{n^{k/3}}$. If we choose m such that $a + a\varepsilon\mu \leq \frac{1}{n^{k/3}}$ holds, then we obtain

$$h' \;\leq\; \mu + \frac{3}{n^{k/3}}.$$

Thus, we only need to satisfy $a + a\varepsilon\mu \leq \frac{1}{n^{k/3}}$. We can write

$$a + a\varepsilon\mu \;=\; a(1+\varepsilon\mu) \;\leq\; 2a \;=\; 4\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}},$$

so it is enough to choose m such that

$$\mathrm{e}^{-\frac{1}{3(2^{\Delta}+1)}m^{1/3}} \;\leq\; \frac{1}{4n^{k/3}} \tag{77}$$

holds. Taking the logarithm on both sides, rearranging, and using $\ln 4 < 2$, we see that to satisfy (77) it is sufficient to choose

$$m \;\geq (2^{\Delta}+1)^3(6+k\ln n)^3.$$

Thus, collecting the parts, we have obtained that the resulting family $\mathcal{F}$ is indeed a (μ', δ, h')-approximator of $\mathcal{I}(G)$, with

$$\mu' \leq \mu + \frac{2}{n^{k/3}} \quad \text{and} \quad h' \leq \mu + \frac{3}{n^{k/3}},$$

whenever the parameters satisfy the conditions stated in the theorem.

(iii) The algorithm uses the input $G, \mathcal{F}_0, v_0$, and it also uses the constant k to determine the number of steps $T = n^k$ when it runs RISL, starting from each $X_i \in \mathcal{F}_0$. The claim that it does not use the parameters Δ, μ, δ, h, and that the running time is independent of them, follows from (iii) in Theorem 2, along with the fact that μ, δ, h do not occur in RISL at all. The parameters Δ, μ, δ, h only play a role in the analysis, assuming that they satisfy the relations required in the Theorem. This completes the proof.

♠

Appendix E: Proof of Theorem 4

For convenient reference, we repeat the algorithm and the theorem below.

Algorithm *FPRAS for Independent Sets*

Input: a graph $G = (V, E)$, and a tolerance parameter $\delta > 0$.

Output: A number $\widetilde{I}$, serving as an approximation of $I(G)$.

The Algorithm

Step 1 Let $v_1, \ldots, v_n$ be the vertices of G. Define the graphs H_ℓ, $\ell = 1, \ldots, n$, such that H_ℓ consists of the subgraph of G induced by $v_1, \ldots, v_\ell$, plus adding the remaining $n - \ell$ vertices $v_{\ell+1}, \ldots, v_n$, as isolated vertices.

Step 2 Set $\ell = 1$, $m = \left\lceil 64n^6 \left(1 + \frac{1}{\delta}\right)^3 \right\rceil$.

Step 3 Create a random approximator $\mathcal{F}_\ell$ of $\mathcal{I}(H_\ell)$, as follows:

(a) Set $j = 1$. Pick m random independent sets in H_1 (which consists of n isolated vertices), using an independent random bit to decide for each vertex whether it is put in the random set or not. Call the resulting system $\mathcal{F}_1$.

(b) If $j < \ell$, then create $\mathcal{F}_{j+1}$ from $\mathcal{F}_j$ by running the RAE algorithm (see Section 4.4), with input graph H_j, distinguished vertex $v_0 = v_j$, and random approximator $\mathcal{F}_j$. The output is the random approximator $\mathcal{F}_{j+1}$.

(c) Set $j = j + 1$. If $j < \ell$, then go to (b). Else compute $\widetilde{q}_\ell$ by counting what fraction of independent sets in $\mathcal{F}_\ell$ contain v_ℓ in the graph H_ℓ. Go to Step 4.

Step 4 If $\ell < n$ then set $\ell = \ell + 1$, and repeat from Step 3. Else output the value

$$\widetilde{I} = 2 \prod_{\ell=1}^{n-1} \frac{1}{1 - \widetilde{q}_{\ell+1}}$$

and HALT.

Theorem 4 *The algorithm* FPRAS for Independent Sets *satisfies Definition 9 for input graphs with maximum degree bounded by an arbitrary constant. Specifically,*

(i) *Upon input graph G with $n = |V(G)|$, and tolerance parameter $\delta > 0$, the algorithm produces the output number $\widetilde{I}$ in probabilistic polynomial time in n and $1/\delta$. In particular, the algorithm has the the following complexity:*

$$\text{if } \delta \geq 1 \text{ then it runs in time } O\left(n^{k+9}\right), \text{using } O\left(n^{k+8}\log n\right) \text{ random bits;}$$

$$\text{if } \delta < 1 \text{ then it runs in time } O\left(n^{k+9}(1/\delta)^3\right), \text{using } O\left(n^{k+8}(1/\delta)^3\log n\right) \text{ random bits,}$$

where k is the fixed parameter used in algorithm RAE (see Section 4.4*), which is run here as a subroutine in* Step 3(b).

(ii) *Suppose that*

$$k \in \{7,8,9\}, \quad m \geq (2^{\Delta}+1)^3(6+k\ln n)^3, \quad \text{and} \quad n \geq \Delta+2, \tag{78}$$

where Δ is any constant upper bound on the maximum degree of G. Then the output $\widetilde{I}$ of the algorithm, as required by Definition 9, satisfies

$$\Pr\left(\,(1-\delta)I(G) \leq \widetilde{I} \leq (1+\delta)I(G)v)\,\right) \;\geq\; \frac{3}{4}, \tag{79}$$

assuming that he input graph has at least $n \geq n_0(k)$ vertices, where

$$n_0(k) = \begin{cases} 24 & \textit{if} \quad k=9 \\ 118 & \textit{if} \quad k=8 \\ 13,824 & \textit{if} \quad k=7. \end{cases} \tag{80}$$

Furthermore, the probability in (79) *is meant solely with respect to the internal randomization of the algorithm.*

Proof.

(i) In Step 3(a) it takes $O(n)$ time to generate each independent set, using n random bits. This is carried out $m = \left\lceil 64n^6\left(1+\frac{1}{\delta}\right)^3\right\rceil$ times to obtain $\mathcal{F}_1$. Therefore it takes a running time of $O\left(n^7\left(1+\frac{1}{\delta}\right)^3\right)$ to generate $\mathcal{F}_1$, using $O\left(n^6\left(1+\frac{1}{\delta}\right)^3\right)$ random bits.

In Step 3(b) we run the RAE algorithm, described in Section 4.4. By Theorem 3, this takes $O(mn^{k+1})$ time, using $O(mn^k\log n)$ random bits. Taking into account the value of m, as well as the fact that Step 3(b) is repeated $O(n)$ times to construct $\mathcal{F}_\ell$ for a given ℓ, we obtain that $\mathcal{F}_\ell$ is generated in $O\left(n^7\left(1+\frac{1}{\delta}\right)^3 + n^{k+8}\left(1+\frac{1}{\delta}\right)^3\right) = O\left(n^{k+8}\left(1+\frac{1}{\delta}\right)^3\right)$ time, using $O\left(n^6\left(1+\frac{1}{\delta}\right)^3 + n^{k+7}\left(1+\frac{1}{\delta}\right)^3\log n\right) = O\left(n^{k+7}\left(1+\frac{1}{\delta}\right)^3\log n\right)$ random bits.

The above is repeated for each value of $\ell = 1,\ldots,n$, independently, bringing in another $O(n)$ factor. This leads to a running time of $O\left(n^{k+9}\left(1+\frac{1}{\delta}\right)^3\right)$, using $O\left(n^{k+8}\left(1+\frac{1}{\delta}\right)^3\log n\right)$ random bits. The additional operations, namely the computation of $\widetilde{q}_\ell$ in Step 3(c), and the computation of $\widetilde{I}$ in Step 4 clearly fit in this complexity, without increasing it any further. Finally, taking into account that

$$\left(1+\frac{1}{\delta}\right)^3 = \begin{cases} O(1) & \text{if} \quad \delta \geq 1 \\ O((1/\delta)^3) & \text{if} \quad \delta < 1 \end{cases}$$

holds, we obtain the following overall running time:

$$\text{For } \delta \geq 1: \quad O\left(n^{k+9}\right), \text{ using } O\left(n^{k+8}\log n\right) \text{ random bits;}$$
$$\text{for } \delta < 1: \quad O\left(n^{k+9}(1/\delta)^3\right), \text{ using } O\left(n^{k+8}(1/\delta)^3\log n\right) \text{ random bits.}$$

(ii) Let $G = (V, E)$ be the input graph, with $V = \{v_1, \ldots, v_n\}$. Denote by $V_\ell = \{v_1 \ldots, v_\ell\}$ the set of the first ℓ vertices, and by G_ℓ the subgraph of G induced by V_ℓ. Set $I_\ell = |\mathcal{I}(G_\ell)|$. Since $G_n = G$, therefore, $I_n = |\mathcal{I}(G)|$. Let us simply denote it by I, so $I = I_n$ is what we want to approximate. Then we can express I by a "telescopic product:"

$$I = I_1 \prod_{\ell=1}^{n-1} \frac{I_{\ell+1}}{I_\ell}. \tag{81}$$

As G_1 consists only the single vertex v_1, we have $\mathcal{I}(G_1) = \{\emptyset, \{v_1\}\}$, yielding $I_1 = 2$. To capture the ratio $I_{\ell+1}/I_\ell$, let p_ℓ, $\ell = 1, \ldots, n$, denote the independence ratio of v_ℓ in the graph G_ℓ. Looking at $v_{\ell+1}$ in the graph $G_{\ell+1}$, it follows directly from the definition of the independence ratio (see Definition 5) that

$$\frac{I_{\ell+1} - I_\ell}{I_{\ell+1}} = p_{\ell+1},$$

since $I_{\ell+1} - I_\ell$ is the number of independent sets in $G_{\ell+1}$ that contain $v_{\ell+1}$, divided by the total number of independent sets in $G_{\ell+1}$. After rearranging, it yields

$$\frac{I_{\ell+1}}{I_\ell} = \frac{1}{1 - p_{\ell+1}}.$$

Substituting into (81), along with $I_1 = 2$, we get

$$I = 2 \prod_{\ell=1}^{n-1} \frac{1}{1 - p_{\ell+1}}. \tag{82}$$

Based on (82) a first strategy that comes to mind is to try to find a good enough approximation of each independence ratio p_ℓ, then it will yield the desired approximation of I. Unfortunately, this runs into the following problem. We could use random approximators for approximating p_ℓ, see Definition 8. A natural plan would be to construct random approximators directly for the small initial graphs in the sequence, and then use the Random Approximator Extension (RAE) algorithm to extend the random approximators gradually to larger and larger graphs. A problem, however, is that the parameters μ and h in RAE increase by an inverse polynomial of the number of vertices (see (ii) in Theorem 3) after each extension. When ℓ, which is the number of vertices in G_ℓ, is still small (say, bounded by a constant) at the beginning of the graph sequence, then $1/\ell$ is also constant, and this leads to non-vanishing accumulated error, even if n grows large.

We solve the above problem by introducing a modified graph sequence, in which *each* graph has n vertices, so that we get rid of the small graphs. This is the graph sequence H_ℓ that the algorithm actually uses. We carry out the modification such that the new graphs give rise to the same independence ratios.

Let H_ℓ be the graph obtained by adding $n - \ell$ isolated vertices to G_ℓ. Set $J_\ell = |\mathcal{I}(H_\ell)|$, and we keep the earlier introduced notation $I_\ell = |\mathcal{I}(G_\ell)|$. Observe that $J_\ell = 2^{n-\ell} I_\ell$, because each independent set of H_ℓ arises as the union of an independent set of G_ℓ and another one which is an

arbitrary subset of the added $n-\ell$ isolated vertices (including the empty set). Also note that each H_ℓ is a subgraph of G, so we still have that H_ℓ has maximum degree $\leq \Delta$.

For any vertex v, let I_ℓ^v, J_ℓ^v denote the number of independent sets that contain the vertex v in G_ℓ and H_ℓ, respectively. We have $J_\ell^{v_\ell} = 2^{n-\ell} I_\ell^{v_\ell}$, because each independent set in G_ℓ that contains v_ℓ can be extended into an independent set of H_ℓ by taking its union with any set $B \subseteq V - V_\ell$ (it is possible that $B = \emptyset$), and every independent set in H_ℓ that contains v_ℓ arises this way.

Let p_ℓ, q_ℓ denote the independence ratios (see Definition 5) of v_ℓ in the graphs G_ℓ, H_ℓ, respectively. Then, by the definition of the independence ratio, we can write:

$$p_\ell \;=\; \frac{I_\ell^{v_\ell}}{I_\ell}, \qquad q_\ell \;=\; \frac{J_\ell^{v_\ell}}{J_\ell}.$$

Taking into account $J_\ell^{v_\ell} = 2^{n-\ell} I_\ell^{v_\ell}$ and $J_\ell = 2^{n-\ell} I_\ell$, we obtain

$$q_\ell \;=\; \frac{J_\ell^{v_\ell}}{J_\ell} \;=\; \frac{2^{n-\ell} I_\ell^{v_\ell}}{2^{n-\ell} I_\ell} \;=\; \frac{I_\ell^{v_\ell}}{I_\ell} \;=\; p_\ell.$$

Therefore, we can substitute q_ℓ for p_ℓ in (82) to get

$$I = 2 \prod_{\ell=1}^{n-1} \frac{1}{1 - q_{\ell+1}}. \tag{83}$$

The value of q_ℓ is approximated by a random approximator, see Definition 8. Specifically, in Step 3 of the algorithm a random approximator $\mathcal{F}_\ell$ of $\mathcal{I}(H_\ell)$ is constructed for each $\ell = 1, \ldots, n$, with $|\mathcal{F}_\ell| = m$. We will analyze its parameters $(\mu_\ell, \delta_\ell, h_\ell)$ using Theorem 3. Once $\mathcal{F}_\ell$ is constructed, an estimation $\widetilde{q}_\ell$ of q_ℓ is obtained in Step 3(c), by counting what fraction of sets $F \in \mathcal{F}_\ell$ contain v_ℓ. This is the ratio $p(v_\ell \,|\, \mathcal{F}_\ell)$, according to Definition 5. Since we are going to deal with different graphs, therefore, to avoid confusion, in the proof we always denote the considered graph as a subscript in the independence ratio expressions. That is, $p_H(v)$ denotes the independence ratio of vertex v in the graph H, and $p_H(v \,|\, \mathcal{F})$ denotes the $\mathcal{F}$-independence ratio of vertex v in H. The latter assumes, by definition, that $\mathcal{F} \sqsubseteq \mathcal{I}(H)$.

Let us now detail the construction of $\mathcal{F}_\ell$, as carried out in Step 3 of the algorithm. For the initial graph H_1 we can do the construction directly by picking m random independent sets, since H_1 consists of n isolated vertices. A uniformly random independent set in H_1 is selected in Step 3(a) by including each vertex with probability 1/2 independently (coin flip). Carrying out this set selection m times independently gives us $\mathcal{F}_1$, yielding $|\mathcal{F}_1| = m$.

Let us analyze the parameters of $\mathcal{F}_1$. By the special structure of $\mathcal{F}_1$ (every possible subset of vertices is independent), each vertex v is contained in precisely half of the independent sets. Furthermore, each X_i is perfectly uniform on $\mathcal{I}(H_1)$. Therefore, we have

$$p_{H_1}(v) = \frac{1}{2} \quad \text{and} \quad \mu_1 = 0, \tag{84}$$

the first equality holding for every v. Turning to the approximation, let $X_1, \ldots, X_m$ be the independent sets in $\mathcal{F}_1$. Let $\xi_i = \chi\{v \in X_i\}$ denote the indicator of the event $\{v \in X_i\}$. By the definition of the independence ratio (Definition 5), we can write:

$$p_{H_1}(v \,|\, \mathcal{F}_1) \;=\; \frac{1}{m} \sum_{i=1}^{m} \chi\{v \in X_i\} \;=\; \frac{1}{m} \sum_{i=1}^{m} \xi_i,$$

The $\xi_1, \ldots, \xi_m$ are i.i.d. 0–1 valued random variables, and each X_i is uniform over $\mathcal{I}(H_1)$. Therefore, the expected value of $\xi_i = \chi\{v \in X_i\}$ is precisely the probability that v falls into an independent set that is selected uniformly at random from $\mathcal{I}(H_1)$, which is equal to the independence ratio $p_{H_1}(v)$ (see the remark after Definition 5). Therefore,

$$\mathrm{E}\left(p_{H_1}(v \,|\, \mathcal{F}_1)\right) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left(\chi\{v \in X_i\}\right) \;=\; \frac{1}{m}\sum_{i=1}^{m} p_{H_1}(v) \;=\; p_{H_1}(v).$$

Then the deviation between $p_{H_1}(v|\mathcal{F}_1)$ and its expected value $p_{H_1}(v)$ can be bounded by the two-sided version of the Chernoff bound, which gives for any $\rho \in (0,1)$

$$\Pr\left((1-\rho)\,\mathrm{E}\left(\textstyle\sum_{i=1}^{m}\xi_i\right) \;\le\; \sum_{i=1}^{m}\xi_i \;\le (1+\rho)\,\mathrm{E}\left(\textstyle\sum_{i=1}^{m}\xi_i\right)\right) \;\ge\; 1 - 2\mathrm{e}^{-\mathrm{E}\left(\sum_{i=1}^{m}\xi_i\right)\rho^2/3}.$$

Using $p_{H_1}(v) = \frac{1}{m}\mathrm{E}(\sum_{i=1}^{m}\xi_i)$ and $p_{H_1}(v|\mathcal{F}_1) = \frac{1}{m}(\sum_{i=1}^{m}\xi_i)$ in the above formula yields

$$\Pr\left(\,(1-\rho)p_{H_1}(v) \;\le\; p_{H_1}(v|\mathcal{F}_1) \;\le (1+\rho)p_{H_1}(v)\,\right) \;\ge\; 1 - 2\mathrm{e}^{-mp_{H_1}(v)\rho^2/3}.$$

Let us choose $\rho = 1/m^{1/3}$. Substituting it, we get

$$\Pr\left(\left(1 - 1/m^{1/3}\right)p_{H_1}(v) \;\le\; p_{H_1}(v\,|\,\mathcal{F}_1) \;\le\; \left(1 + 1/m^{1/3}\right)p_{H_1}(v)\right) \;\ge\; 1 - 2\mathrm{e}^{-\frac{p_{H_1}(v)}{3}m^{1/3}}.$$

Set $\delta_1 = 1/m^{1/3}$. Then, using $p_{H_1}(v) = 1/2$ from (84) on the right-hand side, the above probability bound implies

$$\Pr\left((1-\delta_1)\,p_{H_1}(v) \;\le\; p_{H_1}(v\,|\,\mathcal{F}_1) \;\le\; (1+\delta_1)\,p_{H_1}(v)\right) \;\ge\; 1 - 2\mathrm{e}^{-\frac{1}{6}m^{1/3}}. \tag{85}$$

By the definition of a random approximator (see Definition 8), the bound (85) means that $\mathcal{F}_1$ is a (μ_1, δ_1, h_1)-approximator of $\mathcal{I}(H_1)$, with

$$\mu_1 = 0, \qquad \delta_1 = 1/m^{1/3}, \qquad h_1 = 2\mathrm{e}^{-\frac{1}{6}m^{1/3}}, \tag{86}$$

where $\mu_1 = 0$ is already known from (84), and the construction also provides $|\mathcal{F}_1| = m$.

Once we have constructed $\mathcal{F}_1$, we can extend the construction to $\mathcal{F}_j$ with $j > 1$. This is carried out in Step 3 via repeatedly increasing j, starting from $j = 1$, until ℓ is reached. In the move from $\mathcal{F}_j$ to $\mathcal{F}_{j+1}$ in Step 3(b) we utilize that for every j, the relationship between the graphs H_{j+1} and H_j is

$$H_j = H_{j+1} - S(v_{j+1}),$$

where $S(v_{j+1})$ is the set of edges adjacent to v_{j+1} in H_{j+1}. That is, H_j is obtained from H_{j+1} by removing the edges adjacent to v_{j+1} in H_{j+1}. This is precisely the setting we have considered in the Random Approximator Extension (RAE) algorithm (see Section 4.4). Thus, once we have $\mathcal{F}_j$, which is a random (μ_j, δ_j, h_j)-approximator of $\mathcal{I}(H_j)$, then applying the RAE algorithm, $\mathcal{F}_j$ is transformed into $\mathcal{F}_{j+1}$, which is a random $(\mu_{j+1}, \delta_{j+1}, h_{j+1})$-approximator of $\mathcal{I}(H_{j+1})$. Recall that the RAE algorithm also contains a parameter k, of which $k \ge 7$ is assumed. Then by claim (ii) of Theorem 3, assuming that the requirements on the parameters are satisfied, we get that $(\mu_{j+1}, \delta_{j+1}, h_{j+1})$ fulfills

$$\mu_{j+1} \;\le\; \mu_j + \frac{2}{n^{k/3}}, \qquad \delta_{j+1} \;=\; \delta_j \;=\; \frac{1}{m^{1/3}}, \qquad h_{j+1} \;\le\; h_j + \frac{3}{n^{k/3}}.$$

The conditions in (ii) of Theorem 3 that need to be satisfied are

$$(1)\ \delta \geq m^{-1/3}, \quad (2)\ m \geq (2^{\Delta}+1)^3(6+k\ln n)^3, \quad (3)\ k \geq 7, \quad (4)\ n \geq 3k^{1/5}, \quad (5)\ n \geq \Delta+2,$$

where Δ is any upper bound on the maximum degree of G. Let us show that all these conditions hold, either by assumption, or they are implied by some others. The first condition $\delta \geq m^{-1/3}$ is equivalent to $m \geq (1/\delta)^3$, which is implied by the value of m set in Step 2 of the algorithm. The conditions (2) and (5) are assumed in (78). Condition (3) follows from $k \in \{7,8,9\}$, which is assumed in (78). Finally, condition (4), requiring $n \geq 3k^{1/5}$, follows from the assumption $n \geq n_0(k)$, where $n_0(k)$ is given in (80). Then by straightforward induction, starting with (86), we obtain for every ℓ

$$\mu_\ell \leq (\ell-1)\frac{2}{n^{k/3}} \leq \frac{2}{n^{k/3-1}}, \qquad \delta_\ell = \frac{1}{m^{1/3}}, \qquad h_\ell \leq h_1 + (\ell-1)\frac{3}{n^{k/3}} \leq 2\mathrm{e}^{-\frac{1}{6}m^{1/3}} + \frac{3}{n^{k/3-1}}.$$

This way we can create a $(2/n^{k/3-1}, 1/m^{1/3}, h_1 + 3/n^{k/3-1})$-approximator for each $\mathcal{I}(H_\ell)$, with $|\mathcal{F}_\ell| = m$. This is carried out in Step 3 for each $\ell = 1,\dots,n$ independently, that is, starting over each time with an independent version of $\mathcal{F}_1$. Then we get n stochastically independent random approximators $\mathcal{F}_\ell$, $\ell = 1,\dots,n$, such that each $\mathcal{F}_\ell$ is a random $(2/n^{k/3-1}, 1/m^{1/3}, h_1 + 3/n^{k/3-1})$-approximator of $\mathcal{I}(H_\ell)$, with $|\mathcal{F}_\ell| = m$.

Let us now use each $\mathcal{F}_\ell$ to estimate the value of q_ℓ, needed in the formula (83). Recall that q_ℓ is the independence ratio of the vertex v_ℓ in the graph H_ℓ, i.e., $q_\ell = p_{H_\ell}(v_\ell)$. We can approximate q_ℓ by the $\mathcal{F}_\ell$-independence ratio, via simply counting that what fraction of independent sets in $\mathcal{F}_\ell$ contains v_ℓ. This is done in Step 3(c) of the algorithm. Let the obtained ratio be denoted by $\widetilde{q}_\ell$, that is, $\widetilde{q}_\ell = p_{H_\ell}(v_\ell \mid \mathcal{F}_\ell)$. Recall that the $\mathcal{F}_\ell$, $\ell = 1,\dots,n$, are constructed such that they are stochastically independent, therefore, $\widetilde{q}_\ell$, $\ell = 1,\dots,n$, are independent random variables.

Using the notations

$$\delta_1 = 1/m^{1/3} \quad \text{and} \quad \varepsilon = 2\mathrm{e}^{-\frac{1}{6}m^{1/3}} + 3/n^{k/3-1}, \tag{87}$$

the approximation properties of $\mathcal{F}_\ell$ (by definition) provide us with

$$\Pr\big(\,(1-\delta_1)q_\ell \leq \widetilde{q}_\ell \leq (1+\delta_1)q_\ell\,\big) \geq 1-\varepsilon. \tag{88}$$

Rearranging within the probability expression gives

$$\Pr\big(\,1-(1-\delta_1)q_\ell \geq 1-\widetilde{q}_\ell \geq 1-(1+\delta_1)q_\ell\,\big) \geq 1-\varepsilon.$$

Let us now rearrange the bounding expressions of $1-\widetilde{q}_\ell$ as

$$1-(1-\delta_1)q_\ell = (1-q_\ell)\left(1+\delta_1\frac{q_\ell}{1-q_\ell}\right) \quad \text{and} \quad 1-(1+\delta_1)q_\ell = (1-q_\ell)\left(1-\delta_1\frac{q_\ell}{1-q_\ell}\right).$$

Then, via introducing $a_\ell = q_\ell/(1-q_\ell)$, the inequality (88) can be reformulated as

$$\Pr\big((1-q_\ell)(1+\delta_1 a_\ell) \geq 1-\widetilde{q}_\ell \geq (1-q_\ell)(1-\delta_1 a_\ell)\,\big) \geq 1-\varepsilon. \tag{89}$$

Let us use that the $\widetilde{q}_1,\dots,\widetilde{q}_n$ random variables are independent, and they are the only random quantities in (89). Then writing down (89) with $\ell+1$, and taking the product over $\ell = 1,\dots,n-1$ yields

$$\Pr\left(\prod_{\ell=1}^{n-1}\big((1-q_{\ell+1})(1+\delta_1 a_{\ell+1})\,\big) \geq \prod_{\ell=1}^{n-1}(1-\widetilde{q}_{\ell+1}) \geq \prod_{\ell=1}^{n-1}\big((1-q_{\ell+1})(1-\delta_1 a_{\ell+1})\,\big)\right) \geq$$

$$(1-\varepsilon)^{n-1}. \tag{90}$$

Now let us introduce the constants

$$M = \max_{\ell=1,\ldots,n-1}(1+\delta_1 a_{\ell+1}) \quad \text{and} \quad m = \min_{\ell=1,\ldots,n-1}(1-\delta_1 a_{\ell+1}). \tag{91}$$

Then we have

$$M^{n-1} \geq \prod_{\ell=1}^{n-1}(1+\delta_1 a_{\ell+1}) \quad \text{and} \quad m^{n-1} \leq \prod_{\ell=1}^{n-1}(1-\delta_1 a_{\ell+1}). \tag{92}$$

Since extending an interval can only increase the probability that a random variable falls into it, therefore, the probability bound (90), using (92), implies

$$\Pr\left(M^{n-1}\prod_{\ell=1}^{n-1}(1-q_{\ell+1}) \geq \prod_{\ell=1}^{n-1}(1-\widetilde{q}_{\ell+1}) \geq m^{n-1}\prod_{\ell=1}^{n-1}(1-q_{\ell+1})\right) \geq (1-\varepsilon)^{n-1}. \tag{93}$$

Why is this useful? Because we have by (83) that the true number I of independent sets satisfies

$$I = 2\prod_{\ell=1}^{n-1}\frac{1}{1-q_{\ell+1}}.$$

We approximate it by

$$\widetilde{I} = 2\prod_{\ell=1}^{n-1}\frac{1}{1-\widetilde{q}_{\ell+1}}.$$

From these expressions one can obtain

$$2/I = \prod_{\ell=1}^{n-1}(1-q_\ell) \quad \text{and} \quad 2/\widetilde{I} = \prod_{\ell=1}^{n-1}(1-\widetilde{q}_\ell).$$

Substituting them into (93) yields

$$\Pr\left(M^{n-1}(2/I) \geq 2/\widetilde{I} \geq m^{n-1}(2/I)\right) \geq (1-\varepsilon)^{n-1}. \tag{94}$$

Dividing each side by 2 within the probability expression, and taking the reciprocal, which reverses the inequalities, we obtain

$$\Pr\left(I\,\frac{1}{M^{n-1}} \leq \widetilde{I} \leq I\,\frac{1}{m^{n-1}}\right) \geq (1-\varepsilon)^{n-1}. \tag{95}$$

Let us now analyze the expressions $1/M^{n-1}$ and $1/m^{n-1}$. Since each q_ℓ is the independence ratio of a vertex in a graph of maximum degree $\leq \Delta$, therefore, Lemma 2 gives us

$$\frac{1}{2^\Delta+1} \leq q_\ell \leq \frac{1}{2}.$$

At this point we only need the upper bound $q_\ell \leq \frac{1}{2}$. As $a_\ell = \frac{q_\ell}{1-q_\ell}$ is an increasing function of q_ℓ, we get that $a_\ell \leq \frac{1/2}{1-1/2} = 1$. Using this, it follows from (91) that

$$M \leq 1+\delta_1 \quad \text{and} \quad m \geq 1-\delta_1,$$

yielding

$$\frac{1}{M^{n-1}} \geq \frac{1}{(1+\delta_1)^{n-1}} \quad \text{and} \quad \frac{1}{m^{n-1}} \leq \frac{1}{(1-\delta_1)^{n-1}}. \tag{96}$$

If in (95) we replace $\frac{1}{M^{n-1}}$ by a smaller value, and, similarly, we replace $\frac{1}{m^{n-1}}$ by a larger value, then we extend the interval within the probability expression, so the probability can only grow. Therefore, (95) and (96) together imply

$$\Pr\left(I \frac{1}{(1+\delta_1)^{n-1}} \leq \widetilde{I} \leq I \frac{1}{(1-\delta_1)^{n-1}} \right) \geq (1-\varepsilon)^{n-1}. \tag{97}$$

To simplify the expression, we bound $\frac{1}{(1+\delta_1)^{n-1}}$ from below, as well as $\frac{1}{(1-\delta_1)^{n-1}}$ from above, by simpler formulas. Let us use the following well known inequality about the exponential function:

$$\mathrm{e}^x \geq 1 + x, \tag{98}$$

which holds for every x, and follows directly from the convexity of e^x (as $1+x$ is a tangent line to e^x at $x=0$). By substituting $x/(n-1)$ for x in (98), we get $\mathrm{e}^{x/(n-1)} \geq 1 + \frac{x}{n-1}$. Raising both sides to the power of $n-1$ provides

$$\mathrm{e}^x \geq \left(1 + \frac{x}{n-1}\right)^{n-1}.$$

Using this with $x = \delta_1(n-1)$ gives

$$\mathrm{e}^{\delta_1(n-1)} \geq \left(1 + \frac{\delta_1(n-1)}{n-1}\right)^{n-1} = (1+\delta_1)^{n-1},$$

and then taking reciprocals yields

$$\mathrm{e}^{-\delta_1(n-1)} \leq \frac{1}{(1+\delta_1)^{n-1}}. \tag{99}$$

Finally, applying (98) again to the left-hand side with $x = -\delta_1(n-1)$ provides a simple lower bound for $\frac{1}{(1+\delta_1)^{n-1}}$, as follows:

$$1 - \delta_1(n-1) \leq \frac{1}{(1+\delta_1)^{n-1}}. \tag{100}$$

To bound $\frac{1}{(1-\delta_1)^{n-1}}$ from above, we can use the Bernoulli inequality to obtain

$$(1-\delta_1)^{n-1} \geq 1 - \delta_1(n-1)$$

which is valid for $\delta_1 \leq 1$. Taking reciprocals (assuming $\delta_1(n-1) < 1$), gives a simple upper bound on $\frac{1}{(1-\delta_1)^{n-1}}$, as follows:

$$\frac{1}{(1-\delta_1)^{n-1}} \leq \frac{1}{1-\delta_1(n-1)}. \tag{101}$$

Note that the assumption $\delta_1(n-1) < 1$ is indeed valid, see the Remark after expression (104).

If we replace $\frac{1}{(1+\delta_1)^{n-1}}$ by its lower bound (100), and $\frac{1}{(1-\delta_1)^{n-1}}$ by its upper bound (101), then we extend the interval within the probability expression of (97), so the probability can only grow. Therefore, we obtain

$$\Pr\left(I\big(1 - \delta_1(n-1)\big) \leq \widetilde{I} \leq I \frac{1}{1-\delta_1(n-1)} \right) \geq (1-\varepsilon)^{n-1}. \tag{102}$$

Now let us look into the relationship between δ_1 and the input tolerance parameter δ (see Definition 9). In order to guarantee (6) in Definition 9, we look for a δ_1 value, such that

$$1-\delta \;\le\; 1-\delta_1(n-1) \quad \text{and} \quad 1+\delta \;\ge\; \frac{1}{1-\delta_1(n-1)} \tag{103}$$

both hold. Indeed, if (103) is satisfied, then the interval is extended by moving to $[I(1-\delta), I(1+\delta)]$, so the probability can only grow. Therefore, we get

$$\begin{aligned}&\Pr\Big(I(1-\delta) \;\le\; \widetilde{I} \;\le\; I(1+\delta)\Big) \;\ge \\ &\Pr\left(I\Big(1-\delta_1(n-1)\Big) \;\le\; \widetilde{I} \;\le\; I\,\frac{1}{1-\delta_1(n-1)}\right) \;\ge (1-\varepsilon)^{n-1}.\end{aligned} \tag{104}$$

To satisfy (103), let us rearrange the two inequalities in (103). From the first we get $\delta_1 \le \frac{1}{n-1}\delta$. From the second, assuming $\delta_1(n-1) < 1$, we get $\delta_1 \le \frac{1}{n-1}\frac{\delta}{1+\delta}$.

Remark: The assumption $\delta_1(n-1) < 1$ is satisfied, whenever $m > (n-1)^3$ holds. This follows from $\delta_1 = 1/m^{1/3}$, which we know from (87). Our choice for m will guarantee $m > (n-1)^3$, so $\delta_1(n-1) < 1$ indeed holds.

Since the second upper bound on δ_1 provides smaller value, it satisfies both inequalities. Therefore, we have that

$$\delta_1 \;\le\; \frac{1}{n-1}\cdot\frac{\delta}{1+\delta} \tag{105}$$

satisfies both inequalities in (103), and, consequently, it leads to

$$\Pr\Big(I(1-\delta) \;\le\; \widetilde{I} \;\le\; I(1+\delta)\Big) \;\ge (1-\varepsilon)^{n-1},$$

via (104). Taking into account $\delta_1 = 1/m^{1/3}$, we obtain from (105) that m needs to be chosen to satisfy

$$\frac{1}{m^{1/3}} \;\le\; \frac{1}{n-1}\cdot\frac{\delta}{1+\delta}.$$

Rearranging, we get

$$m \;\ge (n-1)^3\left(1+\frac{1}{\delta}\right)^3. \tag{106}$$

If we choose

$$m \;=\; \left\lceil n^3\left(1+\frac{1}{\delta}\right)^3\right\rceil, \tag{107}$$

then it certainly satisfies (106), and also $m > (n-1)^3$, to get $\delta_1(n-1) < 1$. Since the running time of the algorithm is polynomial in n and m, therefore, this choice of m will keep it polynomial in n, m and $1/\delta$, as required by Definition 9. (Note: later we are going to replace n^3 in (107) by $64n^6$, to allow us obtaining a bound on $n_0(k)$, but this alteration still keeps the running time polynomial in n, m and $1/\delta$.)

Finally, we still have to account for the confidence $(1-\varepsilon)^{n-1}$. Specifically, to satisfy Definition 9, we need to make sure $(1-\varepsilon)^{n-1} \ge 3/4$. Recall that the value of ε was defined in (87) as

$$\varepsilon = 2\mathrm{e}^{-\frac{1}{6}m^{1/3}} + \frac{3}{n^{k/3-1}}.$$

By the Bernoulli inequality, we can write

$$(1-\varepsilon)^{n-1} \;\geq\; 1-(n-1)\varepsilon \;\geq\; 1-(n-1)\left(2\mathrm{e}^{-\frac{1}{6}m^{1/3}}+\frac{3}{n^{k/3-1}}\right) \;\geq\; 1-2n\mathrm{e}^{-\frac{1}{6}m^{1/3}}-\frac{3}{n^{k/3-2}}. \quad (108)$$

To simplify the expression, let us apply again the basic inequality $\mathrm{e}^x \geq 1+x$. Substituting $-x$ into it, we get $\mathrm{e}^{-x} \geq 1-x$. Assuming $x<1$, taking reciprocals gives $\mathrm{e}^x \leq \frac{1}{1-x}$. Using it with $x=-\frac{1}{6}m^{1/3}$ gives

$$1-2n\mathrm{e}^{-\frac{1}{6}m^{1/3}} \;\geq\; 1-2n\frac{1}{1+\frac{1}{6}m^{1/3}} \;\geq\; 1-\frac{2n}{\frac{1}{6}m^{1/3}} \;=\; 1-\frac{12n}{m^{1/3}}.$$

Therefore, we can continue (108) as

$$(1-\varepsilon)^{n-1} \;\geq\; 1-\frac{12n}{m^{1/3}}-\frac{3}{n^{k/3-2}}.$$

To achieve the needed $(1-\varepsilon)^{n-1} \geq 3/4$, we have to make sure that

$$\frac{12n}{m^{1/3}}+\frac{3}{n^{k/3-2}} \;\leq\; \frac{1}{4} \quad (109)$$

holds. Recall that by (107) we already have $m \geq n^3$. Let us increase m further to $m \geq 64n^6$, so we replace (107) by

$$m \;=\; \left\lceil 64n^6\left(1+\frac{1}{\delta}\right)^3\right\rceil \;\geq\; 64n^6. \quad (110)$$

Then we get

$$\frac{12n}{m^{1/3}} \;\leq\; \frac{12n}{(64n^6)^{1/3}} \;=\; \frac{12n}{4n^2} \;=\; \frac{3}{n}.$$

Hence, the inequality (109) is implied whenever we satisfy

$$\frac{3}{n}+\frac{3}{n^{k/3-2}} \;\leq\; \frac{1}{4},$$

or, equivalently,

$$\frac{1}{n}+\frac{1}{n^{k/3-2}} \;\leq\; \frac{1}{12}. \quad (111)$$

If $k \geq 9$, then $n \leq n^{k/3-2}$, implying $\frac{1}{n}+\frac{1}{n^{k/3-2}} \leq \frac{2}{n}$, so we can satisfy (111) by $\frac{2}{n} \leq \frac{1}{12}$, which gives $n \geq 24$. If $k<9$, which is only possible via $k=8$ or $k=7$, then we get $\frac{1}{n}+\frac{1}{n^{k/3-2}} \leq \frac{2}{n^{k/3-2}}$. If $k=8$, this yields $\frac{2}{n^{2/3}} \leq \frac{1}{12}$, or $n \geq 24^{2/3} \cong 117.58$, if $k=7$, we get $\frac{2}{n^{1/3}} \leq \frac{1}{12}$, or $n \geq 24^3 = 13,824$. Thus, an upper bound on the minimum required value of n is given by

$$n_0(k) = \begin{cases} 24 & \text{if} \quad k \geq 9 \\ 118 & \text{if} \quad k=8 \\ 13,824 & \text{if} \quad k=7. \end{cases}$$

Since by this calculation $k>9$ does not give further significant advantage over $k=9$, therefore, k is restricted to $\{7,8,9\}$ in the Theorem, leading to the value of $n_0(k)$ stated in (80). (Note that the requirement $k \geq 7$ originates in Theorem 1.) This completes the proof.

♠