

Finding Large Set Covers Faster via the Representation Method*

Jesper Nederlof¹

¹ Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands. j.nederlof@tue.nl

Abstract

The worst-case fastest known algorithm for the Set Cover problem on universes with n elements still essentially is the simple $O^*(2^n)$ -time dynamic programming algorithm, and no non-trivial consequences of an $O^*(1.01^n)$ -time algorithm are known. Motivated by this chasm, we study the following natural question: Which instances of Set Cover *can* we solve faster than the simple dynamic programming algorithm? Specifically, we give a Monte Carlo algorithm that determines the existence of a set cover of size σn in $O^*(2^{(1-\Omega(\sigma^4))n})$ time. Our approach is also applicable to Set Cover instances with exponentially many sets: By reducing the task of finding the chromatic number $\chi(G)$ of a given n -vertex graph G to Set Cover in the natural way, we show there is an $O^*(2^{(1-\Omega(\sigma^4))n})$ -time randomized algorithm that given integer $s = \sigma n$, outputs NO if $\chi(G) > s$ and YES with constant probability if $\chi(G) \leq s - 1$.

On a high level, our results are inspired by the ‘representation method’ of Howgrave-Graham and Joux [EUROCRYPT’10] and obtained by only evaluating a randomly sampled subset of the table entries of a dynamic programming algorithm.

1998 ACM Subject Classification G.2.2 Graph Algorithms, Hypergraphs

Keywords and phrases Set Cover, Exact Exponential Algorithms, Fine-Grained Complexity

Digital Object Identifier 10.4230/LIPIcs.ESA.2016.[79]

1 Introduction

The SET COVER problem is, after determining satisfiability of CNF formulas or Boolean circuits, one of the canonical NP-complete problems. It not only directly models many applications in practical settings, but also algorithms for it routinely are used as tools for theoretical algorithmic results (e.g., [17]). It is a problem ‘whose study has led to the development of fundamental techniques for the entire field’ of approximation algorithms.¹ However, the exact exponential time complexity of SET COVER is still somewhat mysterious: We know algorithms need to use super-polynomial time assuming $P \neq NP$ and (denoting n for the universe size) $O^*(2^{\Omega(n)})$ time assuming the Exponential Time Hypothesis, but how large the exponential should be is not clear. In particular, no non-trivial consequences of an $O^*(1.01^n)$ -time algorithm are currently known.

Even though it is one of the canonical NP-complete problems, the amount of studies of exact algorithms for SET COVER pales in comparison with the amount of literature on exact algorithms for CNF-SAT: Many works focus on finding $O^*(c^n)$ -time algorithms for $c < 2$

* Funded by the NWO VENI project 639.021.438. This work was partly done while the author was visiting the Simons Institute for the Theory of Computing during the program ‘Fine-Grained Complexity and Algorithm Design’ in the fall of 2015.

¹ As the Wikipedia page on Set Cover quotes the textbook by Vazirani [34, p15].



for CNF-SAT on n -variable CNF-formulas in special cases such as, among others, bounded clause width [33, 17, 12], bounded clause density [11, 26] or few projections [28, 31, 32]. Improved exponential time algorithms for special cases of problems other than CNF-SAT were also studied for e.g. GRAPH COLORING or TRAVELING SALESMAN on graphs bounded degree/average degree [8, 9, 15, 20].

In this paper we are interested in the exponential time complexity of SET COVER, and study which properties are sufficient to have improved exponential time algorithms. Our interest in finding faster exponential time algorithms for SET COVER does not only stem from it being a canonical NP-complete problem, but also from its unclear relation with CNF-SAT. Intriguingly, on one hand SET COVER has some similarities with the CNF-SAT: **1.** Both problems take an (annotated) hypergraph as input **2.** The improvability of the worst-case complexity of CNF-SAT is essentially equivalent to the improvability of the worst-case complexity of HITTING SET [14], which is just a reparametrization² of SET COVER. But, on the other hand the problems are quite different to our understanding: **1.** Most algorithms for SET COVER use dynamic programming or some variant of inclusion exclusion, while most algorithms for CNF-SAT are based on branching **2.** No connection between the exponential time complexities of both problems is known (see [14]). One hope would be that a better understanding of the exact complexity of SET COVER might shed more light on this unclarity. Moreover, Cygan et al. [14] also show that if we would like to improve the run time $O^*(f(k))$ of several parameterized algorithms to $O^*(f(k)^{1-\Omega(1)})$, we first need to find an $O^*(2^{(1-\Omega(1))n})$ -time algorithm for SET COVER. These parameterized algorithms include the classic algorithm for SUBSET SUM, as well as more recent algorithms for CONNECTED VERTEX COVER and STEINER TREE.

Relevant previous work The algorithmic results on SET COVER that are the most relevant to our work are as follows: The folklore dynamic programming algorithm runs in $O^*(2^n)$ time. A notable special case of SET COVER that can be solved in $O^*(2^{(1-\Omega(1))n})$ time is due to Koivisto [29]: He gives an algorithm that runs in time $O^*(2^{(1-\frac{1}{\sigma(r)})n})$ -time algorithm if all sets are at most of size r . Björklund et al. [10] show that the problem can be solved in $2^n \text{poly}(n)$ time (which is faster if the number of sets is exponentially large in n). Björklund et al. [7] give a randomized algorithm that assumes all sets are of size q and determines whether there exist p pairwise disjoint sets in $O^*(2^{(1-\epsilon)pq})$ time where $\epsilon > 0$ depends on q .

Our Main Results We investigate what are sufficient structural properties of instances of SET COVER, and the closely related SET PARTITION (in which the picked sets need to be disjoint), problems to be solvable in time significantly faster than the currently known algorithms. We will outline our main results now:

► **Theorem 1.1.** *There is a Monte Carlo algorithm that takes an instance of SET COVER on n elements and m sets and an integer s as input and determines whether there exists a set cover of size s in $O(2^{(1-\Omega(\sigma^4))n}m)$ time, where $\sigma = s/n$.*

We remark that this generalizes the result of Koivisto [29] in the sense that it solves a larger class of instances in $O^*(2^{(1-\Omega(1))n})$ time: If all set sizes are bounded by a constant r , a set partition needs to consist of at least n/r sets and Theorem 1.1 applies with $\sigma = 1/r$

² One way of stating HITTING SET in this context, is that we have an instance of the SET COVER problem but aim to find an $O^*(2^{(1-\Omega(1))m})$ time algorithm, where m denotes the number of sets.

(although this gives a slower algorithm than Koivisto's in this special case). Moreover, it seems hard to extend the approach of Koivisto to our more general setting.

The second result demonstrates that our techniques are also applicable to SET COVER instances with exponentially many sets, a canonical example of which being graph coloring:

► **Theorem 1.2.** *There is a randomized algorithm that given graph G and integer $s = \sigma n$, in $O^*(2^{(1-\Omega(\sigma^4))n})$ time outputs **yes** with constant probability, if $\chi(G) < s$, and **no**, if $\chi(G) > s$.*

Representation method for Set Cover We feel the main technique used in this paper is equally interesting as the result, and will therefore elaborate on its origin here. Our technique is on a high level inspired by the following simple observation ingeniously used by Howgrave-Graham and Joux [24]: Suppose $\mathcal{S} \subseteq 2^{[m]}$ is a set of solutions implicitly given and we seek for a solution $X \in \mathcal{S}$ with $|X| = s$ by listing all sets of $\binom{[m]}{s/2}$ and performing pairwise checks to see which two combine to an element of \mathcal{S} . Then we can restrict our search in various ways since there will be as many as $\binom{s}{s/2}$ pairs guiding us to X . In [24] and all subsequent works (including [3, 4, 1, 2]), this idea was used to speed up ‘meet-in-the-middle attacks’ (also called ‘birthday attacks’ [27, Chapter 6]). We will refer to uses of this idea as the ‘representation method’ since it crucially relies on the fact that X has many representations as pairs. To indicate the power of this technique in the context of SET COVER and SET PARTITION we show that without changes it already gives an $O^*(2^{0.3399m})$ -time Monte Carlo algorithm for the SET PARTITION problem with m sets, and even for a more general linear satisfiability problem on m variables. For the latter problem this improves the $O^*(2^{m/2})$ time algorithm based on the meet-in-the-middle attack that was the fastest known before.

At first sight the representation method seemed to be inherently only useful for improving algorithms based on the meet-in-the-middle attack. However, the main conceptual contribution of this work is to show that it is also useful in other settings, or at least for improving the dynamic programming algorithm for the SET COVER and SET PARTITION problems if the solution size is large. On a high level, we show this as follows in the case of SET PARTITION:³ for a subset W of the elements of the SET PARTITION instance, define $T[W]$ to be the minimum number of disjoint sets needed to cover all elements of W . Stated slightly oversimplified, we argue that if a minimal set partition of size s is large, we have that $T[W] + T[[n] \setminus W] = s$ for $\binom{s}{s/2}$ sets W with $|W|$ close to $n/2$. To relate this to later sections, let us remark we refer to such a set W as a *witness halve*. Subsequently, we exploit the presence of many witness halves by using a dynamic programming algorithm that samples a set of the subsets with size close to $n/2$ and only evaluates table entries from this sample plus the table entries required to compute the table entries from the sample.

Organization This paper is organized as follows: In Section 2, we recall preliminaries and introduce notation. In Section 3, we discuss new observations and basic results that we feel are useful for developing a better understanding of the complexity of SET COVER with respect to several structural properties of instances. In Section 4 we formally present the notion of witness halves and prepare tools for exploiting the existence of many witness halves. In Section 5 we prove our main results and in Section 6 we suggest further research.

³ The algorithm for SET COVER actually reduces to SET PARTITION.

2 Preliminaries and Notation

For a real number x , $|x|$ denotes the absolute value of x . For a Boolean predicate p , we let $[p]$ denote 1 if p is true and 0 otherwise. On the other hand, if p is an integer we let $[p]$ denote $\{1, \dots, p\}$. As usual, \mathbb{N} denotes all positive integers. Running times of algorithms are often stated using $O^*(\cdot)$ notation which suppresses factors polynomial in the input size. To avoid superscript, we sometimes use $\exp(x)$ to denote e^x . We denote \lg for the base-2 logarithm. If $G = (V, E)$ and $v \in V$ we denote $N(v) = \{w \in V : (v, w) \in E\}$ and for $X \subseteq V$ we extend this notation to $N(X) = \bigcup_{v \in X} N(v)$. For reals $a, b > 0$ we let $a \pm b$ denote the interval $[a - b, a + b]$. A false positive (negative) of an algorithm is an instance on which it incorrectly outputs YES (respectively, NO). In this work we call an algorithm Monte Carlo if it has no false positives and if any instance is a false negative with probability at most $1/4$. We denote vectors with boldface for clarity. For a real number $x \in [0, 1]$, $h(x) = -x \lg x - (1 - x) \lg(1 - x)$ denotes the binary entropy of x , where $0 \lg 0$ should be thought of as 0. It is well known that $\binom{b}{a} \leq 2^{h(a/b)b}$ (and this can for example be proved using Stirling's approximation). It is easy to see from the definition that $h(\cdot)$ is symmetric in the sense that $h(x) = h(1 - x)$.

► **Lemma 2.1.** *The following can be verified using standard calculus:*

1. $h(1/2 - x) = h(1/2 + x) \leq 1 - x^2$ for all $x \in (0, 1/2)$,
2. $h(x) \leq x \lg(4/x)$ for all $x \in (0, 1)$,
3. $(1 - 1/n)^n \leq 1/e$.

► **Lemma 2.2** (Hoeffding bound [22]). *If X_1, \dots, X_s are independent, $Y = \sum_{i=1}^s X_i$ and $a_i \leq X_i \leq b_i$ for $i = 1, \dots, s$ then $\Pr[|Y - \mathbb{E}[Y]| \geq t] \leq 2 \cdot \exp\left(\frac{-2t^2}{\sum_{i=1}^s (b_i - a_i)^2}\right)$.*

Set Cover / Set Partition In the SET COVER problem we are given a bipartite graph $G = (F \cup U, E)$ (where F and U shorthand ‘Family’ and ‘Universe’ respectively), together with an integer s and the task is to determine whether there exists a *solution* $S \subseteq F$ such that $N(S) = U$ and $|S| \leq s$. In the SET PARTITION problem we are given the same input as in the SET COVER problem, but we are set to determine whether there exists $S \subseteq F$ with $N(S) = U$, $|S| = s$ and additionally $N(f) \cap N(f') = \emptyset$ for every $f, f' \in S$ with $f \neq f'$. We will refer to solutions of both problems as set covers and set partitions.

Throughout this paper, we let n, m respectively denote $|U|$ and $|F|$, and refer to instances of SET COVER or SET PARTITION as (n, m, s) -instances to quantify their parameters. Since this work concerns SET COVER or SET PARTITION with large solutions we record the following basic observation that follows by constructing for each⁴ c -tuple $t = (f_1, \dots, f_c) \in F^c$ of sets in the original instance a set f^t with $N(f^t) = \bigcup_{i=1}^t f_i$ in the output instance:

► **Observation 2.3** ([14]). *There is a polynomial time algorithm that takes a constant $c \geq 1$ dividing s , and a (n, m, s) -instance of SET COVER (resp. SET PARTITION) as input and outputs an equivalent $(n, m^c, s/c)$ -instance of SET COVER (resp. SET PARTITION).*

Often it will be useful dispense with linear sized sets. To this end, the following can be achieved by simply iterating over all $f \in F$ with $|N(f)| \geq \epsilon n$ and checking for each such set whether there is a solution containing it using the $2^n \text{poly}(n)$ algorithm for SET COVER [10].

⁴ For SET PARTITION only do this for c -tuples (f_1, \dots, f_c) with $N(f_i)$ disjoint.

► **Observation 2.4.** There is an algorithm that, given a real number $\epsilon > 0$, takes an (n, m, s) -instance of SET COVER as input and outputs an equivalent (n, m', s) -instance with $m' \leq m$ satisfying $|N(f)| \leq \epsilon n$ for every $f \in F$. The algorithm runs in $O(m2^{(1-\epsilon)n} \text{poly}(n))$ time.

As we will see in Theorem 3.4, it makes a difference in the SET PARTITION problem whether empty sets are allowed since we need to find a set partition of size exactly s . To exclude such sets, we will simply say that an instance is ‘without empty sets’.

3 Observations and Basic Results on Set Cover and Set Partition.

To improve our understanding of which properties of instances of SET COVER and SET PARTITION allow faster algorithms, and which techniques are useful for obtaining such faster algorithms, we will record some observations and basic results in this section. To stress that the proof techniques in this section are *not our main technical contribution*, we postpone all proofs to Appendix A.

We prefer to state our results in terms of SET COVER because it is slightly more natural and common, but since SET PARTITION often is easier to deal with for our purposes we will sometimes use the following easy reduction, all of whose steps are contained in [14]:

► **Theorem 3.1.** *There is an algorithm that, given a real $0 < \epsilon < 1/2$, takes an (n, m, s) -instance of SET COVER as input and outputs an equivalent (n, m', s) -instance of SET PARTITION with $m' \leq m2^{\epsilon n}$ sets in time $O(m2^{(1-\epsilon)n})$.*

For completeness, we show that in fact SET COVER and SET PARTITION are equivalent with respect to being solvable in time $O^*(2^{(1-\Omega(1))n})$. This was never stated in print to the best of our knowledge, but the proof uses standard ideas and is found in Appendix A.2.

► **Theorem 3.2.** *For some $\epsilon > 0$ there is an $O^*(2^{(1-\epsilon)n})$ time algorithm for SET COVER if and only if for some $\epsilon' > 0$ there is an $O^*(2^{(1-\epsilon')n})$ time algorithm for SET PARTITION.*

The following natural result is a rather direct consequence of a paper by Koivisto [29]. It reveals some more similarity with the k -CNF-SAT problem: Koivisto shows⁵ that for maximum set size r , SET COVER can be solved in $O^*(2^{(1-\Omega(\frac{1}{r}))n})$ which is analogous to k -CNF-SAT being in $O^*(2^{(1-\Omega(\frac{1}{k}))n})$ time [33, 17, 12], and similarly the following result is the counterpart of $O^*(2^{(1-\Omega(\frac{1}{3}))n})$ -time algorithms for CNF-formula’s of density δ (i.e. at most δn clauses) [11, 26]. Again, this result was never explicitly stated in print to the best of our knowledge, and therefore is proved in Appendix A.3.

► **Theorem 3.3.** *There is an algorithm solving (n, m, s) -instances of SET COVER or SET PARTITION in time $m \cdot \text{poly}(n)2^{n - \frac{n}{O(\lg(m/n))}}$.*

Relevant to our work is the following subtlety on solution sizes in SET PARTITION. It shows that for SET PARTITION with empty sets, finding large solutions is as hard as the general case. The proof is postponed to Appendix A.4.

► **Theorem 3.4.** *Suppose there exist $0 < \epsilon_1, \epsilon_2 < 1/2$ and an algorithm solving $(n, m, \epsilon_1 n)$ -instances of SET PARTITION in time $O^*(2^{(1-\epsilon_2)n})$. Then there exists an $O^*(2^{(1-\epsilon_2/2)n})$ -time algorithm for SET PARTITION.*

⁵ Koivisto only showed this for SET PARTITION, but the straightforward reductions in this section carry this result over to SET COVER.

Finally, it is insightful to see how well the representation method performs on the SET PARTITION problem with few sets (e.g., we consider running times of the $O^*(2^m)$, where m is the number of sets). A straightforward approach of the meet-in-the-middle attack leads directly to an $O^*(2^{m/2})$ time algorithm. We show that the representation method combined with the analysis of [2, 1] in fact solves the more general LINEAR SAT problem. In LINEAR SAT one is given an integer t , matrix $A \in \mathbb{Z}_2^{n \times m}$, and vectors $\mathbf{b} \in \mathbb{Z}_2^n$ and $\boldsymbol{\omega} \in \mathbb{N}^m$. The task is to find $\mathbf{x} \in \mathbb{Z}_2^m$ satisfying $A\mathbf{x} \equiv \mathbf{b}$ and $\boldsymbol{\omega} \cdot \mathbf{x} \leq t$.

► **Theorem 3.5.** *There is an $O^*(2^{0.3399m})$ -time Monte Carlo algorithm solving LINEAR SAT.*

To our best knowledge no $O^*(2^{(0.5-\Omega(1))m})$ -time algorithm for LINEAR SAT was known before. We get as a corollary that, given a bipartite graph $G = (F \dot{\cup} U, E)$, we can determine the smallest size of a set partition in time $O^*(2^{0.3399m})$. We take this as a first signal that the representation method is useful for solving SET PARTITION (and SET COVER) for instances with small universe. To see this consequence, note we can reduce this problem to LINEAR SAT as follows: For every $f \in F$ add the incidence vector of $N(f)$ as a column to A , and set the cost ω_i of picking this column to be $n|N(f)| + 1$. Then the minimum of $\boldsymbol{\omega} \cdot \mathbf{x}$ subject to $A\mathbf{x} \equiv \mathbf{1}$ will be $n^2 + s$ where s is the number of sets in a minimum set partition. Let us remark that [16, Page 130] solves (a counting version) of SET PARTITION in time $O^*(1.2561^m) = O^*(2^{0.329m})$, and Drori and Peleg [18] solve the problem in $O^*(2^{0.3212m})$ time,⁶ so by no means our algorithm is the fastest in this setting. However, both use sophisticated branching and we find it intriguing that the representation method does work quite well even for the seemingly more general LINEAR SAT problem.

4 Exploiting the Presence of Many Witness β -halves.

For convenience we will work with SET PARTITION in this section; the results straightforwardly extend to SET COVER but we will not need this in the subsequent section.

► **Definition 4.1.** Given an (n, m, s) instance of SET PARTITION, a subset $W \subseteq U$ is said to be a *witness β -halve* if $|W| \in (\frac{1}{2} \pm \beta)n$ and there exist disjoint subsets $S_1, S_2 \subseteq F$ such that $N(S_1 \cup S_2) = U$, $\sum_{f \in S_1 \cup S_2} |N(f)| = n$, $N(S_1) = W$, $N(S_2) = U \setminus W$ and $|S_1| + |S_2| = s$.

Note that this is similar to the intuitive definition outlined in Section 1, except that we require $|W| \in (\frac{1}{2} \pm \beta)|U|$ and we adjusted the definition to the SET PARTITION problem. Since $S_1 \cup S_2$ is a set partition of size s we see that if a witness β -halve exists, we automatically have a yes instance.

In this section we will give randomized algorithms that solve promise-variants of SET PARTITION with the promise that, if the instance is a yes-instance, there will be an exponential number of witness halves that are sufficiently balanced (i.e. of size close to $n/2$). In the first subsection we outline the basic algorithm and in the second subsection we show how tools from the literature can be combined with our approach to also give a faster algorithm if the number of sets is exponential in n .

⁶ We attempted to find any more recent faster algorithm, but did not find this. Though, we would not be surprised if using more recent tools in branching algorithms as [19] one should be able to more significantly outperform our algorithm for SET PARTITION.

4.1 The basic algorithm

► **Theorem 4.2.** *There exists an algorithm **A1** that takes an (n, m, s) -instance of SET PARTITION and real numbers $\beta, \zeta > 0$ satisfying $2\sqrt{\beta} \leq \zeta < 1/4$ as input, runs in time $2^{(1-(\zeta/2)^4)n} \text{poly}(n)m$, and has the following property: If there exist at least $\Omega(2^{\zeta n})$ witness β -halves it returns **yes** with at least constant probability, and if there does not exist a set partition of size s it returns **no**.*

Note that the theorem does not guarantee anything on Algorithm **A1** if a partition of s sets exists and there are only few witness halves, but we will address this later. A high level description of the Algorithm **A1** is given in Figure 1:

<p>Algorithm A1($G = (F \dot{\cup} U, E), s, \zeta, \beta$).</p> <p>Output: An estimate of whether there exists a set partition of size s.</p> <ol style="list-style-type: none"> 1: for integer l satisfying $\lfloor (1/2 - \beta)n \rfloor < l < \lceil (1/2 + \beta)n \rceil$ do 2: Sample $\mathcal{W} \subseteq \binom{U}{l}$ by including every set of $\binom{U}{l}$ with probability $2^{-\zeta n}$. 3: For every $W \in \mathcal{W}$ and $i \in [n]$, compute $c_i(W)$ and $c_i(U \setminus W)$. 4: if $\exists i \in [n] : c_i(W) \wedge c_{s-i}(U \setminus W)$ then return yes. 5: return no. 	<p>Assumes $2\sqrt{\beta} \leq \zeta < 1/4$</p>
--	---

■ **Figure 1** High level description of the Algorithm implementing Theorem 4.2.

Here, we define $c_i(W)$ to be true if and only if there exists $S_1 \subseteq F$ with $|S_1| = i$, $N(S_1) = W$, and for every $f, f' \in S_1$ with $f \neq f'$, $N(f) \cap N(f') = \emptyset$. Given a set family \mathcal{W} , we denote $\downarrow \mathcal{W} = \{X : \exists W \in \mathcal{W} \wedge X \subseteq W\}$ for the *down-closure* of \mathcal{W} . The following lemma concerns the sub-routine invoked in Algorithm 1 and can be proved via known dynamic programming techniques, and is postponed to Appendix A.6.

► **Lemma 4.3.** *There exists an algorithm that given a bipartite graph $G = (F \dot{\cup} U, E)$ and $\mathcal{W} \subseteq 2^U$ with $|U| = n$ and $|F| = m$, computes $c_i(W)$ for all $W \in \mathcal{W}$ and $i \in [n]$ in $O(\text{poly}(n)|\downarrow \mathcal{W}|m)$ time.*

Thus, for further preparation of the proof of Theorem 4.2, we need to analyze the maximum size of the (down/up)-closure of \mathcal{W} in Algorithm **A1** in Figure 1.

► **Lemma 4.4.** *Let $\zeta > 0$, β (which may be negative) be real numbers satisfying $2\sqrt{|\beta|} \leq \zeta < 1/4$ and $|U| = n$. Suppose $\mathcal{W} \subseteq \binom{U}{(1/2+\beta)n}$ with $|\mathcal{W}| \leq 2^{(1-\zeta)n}$. Then $|\downarrow \mathcal{W}| \leq n2^{(1-(\zeta/2)^4)n}$.*

Proof. Let $\lambda \leq \beta$ and $w_\lambda = |\{W \in \downarrow \mathcal{W} : |W| = \lambda n\}|$, so $|\downarrow \mathcal{W}| \leq n \cdot \max_\lambda w_\lambda$. Then we have the following upper bounds:

$$w_\lambda \leq \binom{n}{\lambda n} \leq 2^{h(\lambda)n}, \quad w_\lambda \leq |\mathcal{W}| \binom{(1/2 + \beta)n}{\lambda n} \leq 2^{((1-\zeta) + h(\frac{\lambda}{1/2+\beta})) (1/2 + \beta)n}.$$

To see the second upper bound, note that any set $W \in \mathcal{W}$ can have at most $\binom{(1/2+\beta)n}{\lambda n}$ subsets of size λn . Thus, we see that $|\downarrow \mathcal{W}|/n$ is upper bounded by $2^{f(\zeta, \beta)n}$, where

$$f(\zeta, \beta) = \max_{\lambda \leq 1/2+\beta} \min \left\{ h(\lambda), (1 - \zeta) + h\left(\frac{\lambda}{1/2 + \beta}\right) (1/2 + \beta) \right\}.$$

The remainder of the proof is therefore devoted to upper bounding $f(\zeta, \beta)$. We establish this by evaluating both terms of the minimum, setting λ to be $\lambda' = (1 - \zeta^2)(1/2 + \beta)$. First note that by our assumption

$$\begin{aligned}\lambda' &= (1 - \zeta^2)(1/2 + \beta) = 1/2 - \zeta^2/2 + \beta - \zeta^2\beta \leq 1/2 - \zeta^2/2 + \zeta^2/4 - \zeta^2\beta < 1/2, \\ \lambda'/(1/2 + \beta) &= 1 - \zeta^2 > 1/2.\end{aligned}$$

Therefore, since $h(x)$ is increasing for $x < 1/2$, $h(\lambda) \leq h(\lambda')$ for $\lambda \leq \lambda'$. Similarly, $h\left(\frac{\lambda}{1/2+\beta}\right)$ is at most $h\left(\frac{\lambda'}{1/2+\beta}\right)$ for $\lambda \geq \lambda'$, and we may upper bound $f(\zeta, \beta)$ by the maximum of the two terms of the minimum in $f(\zeta, \beta)$ obtained by setting $\lambda = \lambda'$. For the first term of the minimum, note that by Lemma 2.1, Item 1:

$$\begin{aligned}h(\lambda') &= h((1 - \zeta^2)(1/2 + \beta)) \leq 1 - (1/2 - (1 - \zeta^2)(1/2 + \beta))^2 \\ &= 1 - (\zeta^2/2 - \beta + \beta\zeta^2)^2 \\ &\leq 1 - (\zeta^2/2 - \beta)^2 \\ &\leq 1 - (\zeta^2/4)^2 = 1 - (\zeta/2)^4.\end{aligned}$$

For the second term we have

$$\begin{aligned}1 - \zeta + h\left(\frac{\lambda'}{1/2 + \beta}\right)(1/2 + \beta) &= 1 - \zeta + h(1 - \zeta^2)(1/2 + \beta) && \text{by Lemma 2.1, Item 2} \\ &= 1 - \zeta + h(\zeta^2)(1/2 + \beta) && \beta < \frac{1}{64} \text{ by assumption} \\ &\leq 1 - \zeta + \zeta^2 \lg\left(\frac{4}{\zeta^2}\right) \frac{33}{64} && \zeta \lg\left(\frac{4}{\zeta^2}\right) \leq \frac{3}{2} \\ &\leq 1 - \zeta + \zeta \frac{3}{2} \cdot \frac{33}{64} \\ &\leq 1 - \zeta/10.\end{aligned}$$

note for the penultimate inequality that $\zeta \lg\left(\frac{4}{\zeta^2}\right)$ is monotone increasing for $0 \leq \zeta \leq 1/4$ and substituting $\zeta = 1/4$ in this expression thus upper bounds it with $3/2$. ◀

Now we are ready to wrap up this section with the proof of Theorem 4.2:

Proof of Theorem 4.2. We can implement Line 3 by invoking the algorithm of Lemma 4.3 with both $|\mathcal{W}|$ and $\mathcal{W}' = \{W : [n] \setminus W \in \mathcal{W}\}$. This will take time $O(\text{poly}(n)(|\downarrow W| + |\downarrow \mathcal{W}'|)m)$. This is clearly the bottleneck of the algorithm, so it remains to upper bound (the expectation of) $|\downarrow W| + |\downarrow \mathcal{W}'|$ by applying Lemma 4.4. To do this, note that $\mathcal{W} \subseteq \binom{[n]}{l}$, $\mathcal{W}' \subseteq \binom{[n]}{n-l}$, and we have that $(1/2 - \beta)n \leq l, n - l \leq (1/2 + \beta)n$. Also, $2\sqrt{|\beta|} \leq \zeta$ by assumption so indeed Lemma 4.4 applies. Then on expectation $|\mathcal{W}| \leq \binom{n}{(1/2+\beta)n} 2^{-\zeta n} \leq 2^{(1-\zeta)n}$, and thus the running time⁷ indeed is as claimed.

For the correctness, it is easily checked that the algorithm never returns false positives. Moreover, if there exist at least $\Omega(2^{\zeta n})$ witness β -halves then for some l in the loop of Line 1, there are at least $\Omega(2^{\zeta n}/n)$ witness halves of size l . Thus in this iteration we see by Lemma 2.1, Part 3 that

$$\Pr[\nexists \text{ witness halve } W \in \mathcal{W}] \leq \left(1 - \frac{1}{2^{\zeta n}}\right)^{\Omega(2^{\zeta n}/n)} \leq e^{-1/n}. \quad (4.1)$$

⁷ Due to the sampling in Line 2, we actually only get an upper bound on the expectation of the running time, but by Markov's inequality we can simply ignore iterations where \mathcal{W} exceeds twice the expectation.

and if a witness halve $W \in \mathcal{W}$ exists the algorithm returns **yes** since $c_i(W) \wedge c_{s-i}(U \setminus W)$ holds for some i by the definition of witness halve. Therefore, if we perform n independent trials of Algorithm A_1 it return **yes** with probability at least $1 - 1/e$. ◀

4.2 Improvement in the case with exponentially many input sets.

In this section we show that under some mild conditions, the existence of many witness halves can also be exploited in the presence of exponentially many sets. This largely builds upon machinery developed by Björklund et al. [10, 8]. To state our result as general as possible we assume the sets are given via an oracle so our running can be sublinear in the input if the number of sets is close to 2^n .

► **Theorem 4.5.** *There exists an algorithm that, given oracle access to an (n, m, s) -instance of SET PARTITION and real numbers $\beta, \zeta > 0$ satisfying $2\sqrt{\beta} \leq \zeta < 1/4$, runs in time $2^{(1-(\zeta/2)^4)n} \text{poly}(n)T$ and has the following property: if there exist at least $\Omega(2^{\zeta n})$ witness β -halves, it outputs **yes** with constant probability and if there does not exist a set partition of size s it outputs **no**.*

Here the oracle algorithm accepts $X \subseteq U$ as input, and decides whether there exists $f \in F$ with $N(f) = X$ in time T .

The proof of Theorem 4.5 is identical to the proof of Theorem 4.2 (and therefore omitted), except that here we use the following lemma instead of Lemma 4.3:

► **Lemma 4.6.** *There exists an algorithm that, given $\mathcal{W} \subseteq 2^U$ and oracle access to a bipartite graph $G = (F \cup U, E)$, computes the values $c_i(W)$ for all $W \in \mathcal{W}$ in $O(T|\mathcal{W}| \text{poly}(n))$ time. Here the oracle algorithm accepts $X \subseteq U$ as input, and decides whether there exists $f \in F$ with $N(f) = X$ in time T .*

This lemma mainly reiterates previous work developed by Björklund et al. [10, 8], but since they did not prove this lemma as such we include a proof here in Appendix A.7.

5 Finding Large Set Covers Faster

In this section we will use the tools of the previous sections to prove our main results, Theorems 1.1 and 1.2. We first connect the existence of large solutions to the existence of many witness halves in the following lemma:

► **Lemma 5.1.** *If an (n, m, s) -instance of SET PARTITION has no empty sets and satisfies $s \geq \sigma_0 n$ and $|N(f)| \leq \sigma_0^4 n / 8$ for every $f \in F$, there is a solution if and only if there exist at least $2^{\sigma_0 n} / 4$ witness $(\sigma_0^2 / 4)$ -halves.*

Proof. Note that the backward direction is trivial since by definition the existence of a witness halve implies the existence of a solution.

For the other direction, suppose $S = \{f_1, \dots, f_s\}$ is a set partition, and denote $d_i = |N(f_i)|$. Suppose $S' \subseteq S$ is obtained by including every element of S with probability $1/2$ in S' . since $N(f_i) \cap N(f_j) = \emptyset$ for $i \neq j$, we have that the random variable $|N(S')|$ is a sum of s independent random variables that equal 0 and d_i with probability $1/2$. By the Hoeffding bound (Lemma 2.2) we see that

$$\Pr \left[\left| |N(S')| - \mathbb{E}[|N(S')|] \right| \geq n\sigma_0^2/4 \right] \leq 2 \cdot \exp \left(\frac{-n^2\sigma_0^4/8}{\sum_{e \in S} d_e^2} \right) \leq 2 \cdot \exp \left(\frac{-n^2\sigma_0^4/8}{n^2\sigma_0^4/8} \right) < \frac{3}{4},$$

where the second inequality follows from $d_e \leq \sigma_0^4 n/8$ and $\sum_{e \in S} d_e = n$. So for at least $2^{|S|}/4 \geq 2^{\sigma_0 n}/4$ subsets $S' \subseteq S$ we have that $|N(S')| \in (\frac{1}{2} \pm \sigma_0^2/4)n$. Thus, since for each such S' , $N(S')$ determines S' and thus gives rise to a distinct witness halve, there are at least $2^{\sigma_0 n}/4$ witness $(\sigma_0^2/4)$ -halves. \blacktriangleleft

Now we are ready to prove the first main theorem, which we recall here for convenience.

► **Theorem 1.1 (restated).** There is a Monte Carlo algorithm that takes an instance of SET COVER on n elements and m sets and an integer s as input and determines whether there exists a set cover of size s in $O(2^{(1-\Omega(\sigma^4))n}m)$ time, where $\sigma = s/n$.

Proof. The algorithm implementing Theorem 1.1 is given in Figure 2.

Algorithm $A2(G = (F \dot{\cup} U, E), \sigma)$.

- 1: Ensure $|N(f)| \leq \sigma^4 n/1000$ using Observation 2.4.
- 2: **for** every integer s satisfying $\lfloor \sigma n/2 \rfloor \leq s \leq \sigma n$ **do**
- 3: Create an (n, m', s) -instance $((F' \dot{\cup} U, E), s)$ of SET PARTITION where F' is constructed by adding a vertex f' with $N(f') = X$ for all $f \in F, X \subseteq N(f)$.
- 4: Let $\sigma_0 = s/n$.
- 5: **if** $A1((F' \dot{\cup} U, E), s, \sigma_0, \sigma_0^2/4) = \text{yes}$ **then return yes**.
- 6: Pick an arbitrary subset $X \in \binom{U}{n/2}$.
- 7: Find the sizes l and r of the smallest set covers in the instances induced by elements X and respectively elements $U \setminus X$ in $O(2^{n/2} \text{poly}(n)m)$ time with standard dynamic programming.
- 8: **if** $l + r \leq \sigma n$ **then return yes else return no**.

■ **Figure 2** Algorithm for SET COVER large solutions (implementing Theorem 1.1).

We first focus on the correctness of this algorithm. It is clear that the algorithm never returns false positives on Line 5 since Algorithm A1 also has this property. If **yes** is returned on Line 8 it is also clear there exists a solution.

Now suppose that a set cover S of size at most σn exists. First suppose $\sigma n/2 \leq |S| \leq \sigma n$. We consider $s = |S|$ in some iteration of the loop on Line 2. Notice that now in Line 3 we have reduced the problem to a yes-instance of SET PARTITION without empty sets satisfying $|N(f)| \leq \sigma^4 n/1000$ for every $f \in F$. Therefore Lemma 5.1 applies with $\sigma_0 \geq \sigma/2$ and we see there are at least $2^{\sigma_0 n}/4$ witness $(\sigma_0^2/4)$ -halves. Thus, we can apply Theorem 4.2 with $\zeta = \sigma_0$ and $\beta = \sigma_0^2/4$ to find the set S with constant probability, since $\beta \leq (\zeta/2)^2$.

Now suppose $|S| \leq \sigma n/2$. Then picking every element in S twice is a solution (as a multiset), and it implies that for every $X \subseteq U$ the sizes of the smallest set covers l and r (as defined in the algorithm) satisfy $l + r \leq \sigma n$. Thus Lines 6-8 find such a set and the algorithm returns **yes**.

For the running time, Line 1 takes at most $O(2^{(1-\sigma^4/1000)n} \text{poly}(n)m)$ due to Observation 2.4. For Line 5, due to Theorem 4.2 this runs in time

$$\begin{aligned}
 O(2^{(1-(\zeta/2)^4)n} \text{poly}(n)m') &= O(2^{(1-(\sigma_0/2)^4)n} \text{poly}(n)2^{\sigma^4 n/1000}m) \\
 &\leq O(2^{(1-(\sigma/4)^4)n} \text{poly}(n)2^{\sigma^4 n/1000}m) \\
 &= O(2^{(1+\sigma^4(1/1000-1/4^4))n} \text{poly}(n)m) \\
 &\leq O(2^{(1-\Omega(\sigma^4))n} \text{poly}(n)m).
 \end{aligned}$$

as claimed in the theorem statement. ◀

As a more direct consequence of the tools of the previous section we also get the following result for SET PARTITION:

► **Theorem 5.2.** *There exists a Monte Carlo algorithm for SET PARTITION that, given oracle access to an $(n, m, \sigma n)$ -instance satisfying $0 < |N(f)| \leq \sigma^4 n/8$ for every $f \in F$, runs in $2^{(1-\Omega(\sigma^4))n} \text{poly}(n)T$ time.*

Here the oracle algorithm accepts $X \subseteq U$ as input, and decides whether there exists $f \in F$ with $N(f) = X$ in time T .

Proof. Lemma 5.1 implies the instance is a YES-instance if and only if there exist $2^{\sigma n}/4$ witness $(\sigma^2/4)$ -halves. Thus Theorem 4.5 implies the theorem statement. ◀

Note that this theorem also implies an $O((m + 2^{(1-\Omega(\sigma^4))n}) \text{poly}(n))$ time algorithm for SET PARTITION where the sets are given explicitly because we can construct a binary search tree after which we can implement the oracle to run in $T = n$ query time. We remark that it would be interesting to see whether the assumption $|N(f)| \leq \sigma^2 n/4$ is needed, but removing this assumption seems to require more ideas than the ones from this work: For example if the solution has three sets of size $3n/10$ there will be no witness halve that is sufficiently balanced, and alternatively using Observation 2.4 seems to be too slow.

However, if we settle for a additive 1-approximation we can deal with this issue in a simple way and have as a particular consequence the second result mentioned in the beginning of this paper:

► **Theorem 1.2 (restated).** *There is a randomized algorithm that given graph G and integer $s = \sigma n$, in $O^*(2^{(1-\Omega(\sigma^4))n})$ time outputs **yes** with constant probability, if $\chi(G) < s$, and **no**, if $\chi(G) > s$.*

Proof. Let $G = (V, E)$ and define a SET PARTITION instance where for every independent set $I \subseteq V$ of G there is an element $f \in F$ with $N(f) = I$. It is easy to see that this instance of SET PARTITION has a solution of size s if and only if $\chi(G) \leq s$.

Check in $\binom{n}{\sigma^4 n/8}$ time whether G has an independent set of size $\sigma^4 n/8$. If such an independent set is found, remove this set from the graph and return **yes** if the obtained graph has a $(k - 1)$ -coloring and **no** otherwise. Using the $O^*(2^n)$ time algorithm by Björklund et al. [10] in the second step, this procedure clearly runs in time $O^*(2^{(1-\Omega(\sigma^4))n})$, and always finds a coloring using at most one more color than the minimum number of colors if a large enough independent set exists.

On the other hand, if the maximum independent set of G is of size at most $\sigma^4 n/8$, we may apply Theorem 5.2 with $T = \text{poly}(n)$ since it can be verified in polynomial time whether a given $X \subseteq V$ is an independent set, and the theorem follows. ◀

6 Directions for Further Research

In this section, we relate the work presented to some notorious open problems. The obvious open question is to determine the exact complexity of the SET COVER problem:

► **Open Problem 1.** Can SET COVER be solved in time $O^*((2 - \Omega(1))^n)$?

This question was already stated at several places. It is known that if a version of SET COVER where the number of solutions modulo 2 is counted can be solved in $(2 - \Omega(1))^n$ the Strong Exponential Time Hypothesis fails. We refer to [14], for more details.

Less ambitiously, it is natural to wonder whether our dependency on σ can be improved. Our algorithm and analysis seem loose, but we feel the gain of a sharpening this analysis does not outweigh the technical effort currently: For a better dependence, we need both a better bound in Lemma 4.4 and to reduce the set sizes more efficiently than in Observation 2.4. As further research we suggest to find a different algorithmic way to deal with the case where many witness halves are unbalanced. But this alone will not suffice to give linear dependence in σ since in Lemma 4.4 we do not expect to get linear dependence on ζ even if $\beta = 0$. It would also be interesting to see which other instances of SET COVER can be solved in $O^*((2 - \Omega(1))^n)$ time. One that might be worthwhile studying is whether this includes instances with optimal set covers in which the sum of the set sizes is at least $(1 + \Omega(1))n$; one may hope to find exponentially many (balanced) witness halves here as well.

In [14], the authors also give a reduction from SUBSET SUM to SET PARTITION. The exact complexity of SUBSET SUM with small integers is also something we explicitly like to state as open problem here, especially since the $O^*(t)$ time algorithm (where t is the target integer) is perhaps one of the most famous exponential time algorithms:

► **Open Problem 2.** Can SUBSET SUM with target t be solved in time $O^*(t^{1-\Omega(1)})$, or can we exclude the existence of such an assuming the Strong Exponential Time Hypothesis?

Note this question was before asked in [25] by the present author. It would be interesting to study the complexity of SUBSET SUM in a similar vein as we did in this paper: are there some special properties allowing a faster algorithm? For example, a faster algorithm for instances of high ‘density’ (e.g., $n/\lg t$) may be used for improving an algorithm of Horowitz&Sahni [23]. Note that here the ‘density’ of a SUBSET SUM instance is the inverse of what one would expect when relating to the definition of density of k -CNF formula.

Another question that has already open for a while is

► **Open Problem 3.** Can GRAPH COLORING be solved in time $O^*(2^{(1-\Omega(1))n})$?

Could the techniques of this paper be used to make progress towards resolving this question? While our algorithm seems to benefit from the existence of many optimal colorings, in an interesting paper Björklund [5] actually shows that the existence of *few* optimal colorings can be exploited in graphs of small pathwidth. Related to this is also the Hamiltonicity problem. In our current understanding this problem becomes easier when there is a promise that there are few Hamiltonian cycles (see [6], but also e.g. [13] allows derandomizations of known probabilistic algorithms in this case), so a natural approach would be to deal explicitly with instances with many solutions by sampling dynamic programming table in a similar vein as done in this paper.

Acknowledgements The author would like to thank Per Austrin, Petteri Kaski, Mikko Koivisto for their collaborations resulting in [1, 2] that mostly inspired this work, Karl Bringmann for discussions on applications of [2] to SET PARTITION, and anonymous reviewers for their useful comments.

References

- 1 Per Austrin, Petteri Kaski, Mikko Koivisto, and Jesper Nederlof. Subset sum in the absence of concentration. In Ernst W. Mayr and Nicolas Ollinger, editors, *32nd International Symposium on Theoretical Aspects of Computer Science, STACS 2015, March 4-7, 2015, Garching, Germany*, volume 30 of *LIPICs*, pages 48–61. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015.

- 2 Per Austrin, Petteri Kaski, Mikko Koivisto, and Jesper Nederlof. Dense subset sum may be the hardest. In Nicolas Ollinger and Heribert Vollmer, editors, *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016, February 17-20, 2016, Orléans, France*, volume 47 of *LIPICs*, pages 13:1–13:14. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2016.
- 3 Anja Becker, Jean-Sébastien Coron, and Antoine Joux. Improved generic algorithms for hard knapsacks. In Kenneth G. Paterson, editor, *Advances in Cryptology - EUROCRYPT 2011 - 30th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tallinn, Estonia, May 15-19, 2011. Proceedings*, volume 6632 of *Lecture Notes in Computer Science*, pages 364–385. Springer, 2011.
- 4 Anja Becker, Antoine Joux, Alexander May, and Alexander Meurer. Decoding random binary linear codes in $2^{n/20}$: How $1 + 1 = 0$ improves information set decoding. In *EUROCRYPT*, volume 7237 of *Lecture Notes in Computer Science*, pages 520–536. Springer, 2012. Talk at <http://www.iacr.org/cryptodb/data/paper.php?pubkey=24271>.
- 5 Andreas Björklund. Uniquely coloring graphs over path decompositions. *CoRR*, abs/1504.03670, 2015.
- 6 Andreas Björklund, Holger Dell, and Thore Husfeldt. The parity of set systems under random restrictions with applications to exponential time problems. In Magnús M. Halldórsson, Kazuo Iwama, Naoki Kobayashi, and Bettina Speckmann, editors, *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, volume 9134 of *Lecture Notes in Computer Science*, pages 231–242. Springer, 2015.
- 7 Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. Narrow sieves for parameterized paths and packings. *CoRR*, abs/1007.1161, 2010.
- 8 Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. Trimmed moebius inversion and graphs of bounded degree. *Theory Comput. Syst.*, 47(3):637–654, 2010.
- 9 Andreas Björklund, Thore Husfeldt, Petteri Kaski, and Mikko Koivisto. The traveling salesman problem in bounded degree graphs. *ACM Transactions on Algorithms*, 8(2):18, 2012.
- 10 Andreas Björklund, Thore Husfeldt, and Mikko Koivisto. Set partitioning via inclusion-exclusion. *SIAM J. Comput.*, 39(2):546–563, 2009.
- 11 Chris Calabro, Russell Impagliazzo, and Ramamohan Paturi. A duality between clause width and clause density for SAT. In *21st Annual IEEE Conference on Computational Complexity (CCC 2006), 16-20 July 2006, Prague, Czech Republic*, pages 252–260. IEEE Computer Society, 2006.
- 12 Timothy M. Chan and Ryan Williams. Deterministic asps, orthogonal vectors, and more: Quickly derandomizing razborov-smolensky. In Robert Krauthgamer, editor, *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1246–1255. SIAM, 2016.
- 13 Suresh Chari, Pankaj Rohatgi, and Aravind Srinivasan. Randomness-optimal unique element isolation with applications to perfect matching and related problems. *SIAM J. Comput.*, 24(5):1036–1050, 1995.
- 14 Marek Cygan, Holger Dell, Daniel Lokshtanov, Dániel Marx, Jesper Nederlof, Yoshio Okamoto, Ramamohan Paturi, Saket Saurabh, and Magnus Wahlström. On problems as hard as CNF-SAT. In *Proceedings of the 27th Conference on Computational Complexity, CCC 2012, Porto, Portugal, June 26-29, 2012*, pages 74–84. IEEE, 2012.
- 15 Marek Cygan and Marcin Pilipczuk. Faster exponential-time algorithms in graphs of bounded average degree. *Inf. Comput.*, 243:75–85, 2015.
- 16 Vilhelm Dahllöf. *Exact Algorithms for Exact Satisfiability Problems*. PhD thesis, Linköping University, TCSLAB, The Institute of Technology, 2006.

- 17 Evgeny Dantsin, Andreas Goerdt, Edward A Hirsch, Ravi Kannan, Jon Kleinberg, Christos Papadimitriou, Prabhakar Raghavan, and Uwe Schöning. A deterministic $(2 - 2/(k + 1))^n$ algorithm for k -SAT based on local search. *Theoretical Computer Science*, 289(1):69 – 83, 2002.
- 18 Limor Drori and David Peleg. Faster exact solutions for some NP-hard problems. *Theoretical Computer Science*, 287(2):473 – 499, 2002. Algorithms.
- 19 Fedor V. Fomin, Fabrizio Grandoni, and Dieter Kratsch. A measure & conquer approach for the analysis of exact algorithms. *J. ACM*, 56(5), 2009.
- 20 Alexander Golovnev, Alexander S. Kulikov, and Ivan Mihajlin. Families with infants: A general approach to solve hard partition problems. In Javier Esparza, Pierre Fraignaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, volume 8572 of *Lecture Notes in Computer Science*, pages 551–562. Springer, 2014.
- 21 Godfrey Hardy and Srinivasa Ramanujan. Asymptotic formulæ in combinatory analysis. *Proceedings of the London Mathematical Society*, s2-17(1):75–115, 1918.
- 22 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- 23 Ellis Horowitz and Sartaj Sahni. Computing partitions with applications to the knapsack problem. *J. ACM*, 21(2):277–292, 1974.
- 24 Nick Howgrave-Graham and Antoine Joux. New generic algorithms for hard knapsacks. In Henri Gilbert, editor, *Advances in Cryptology - EUROCRYPT 2010, 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30 - June 3, 2010. Proceedings*, volume 6110 of *Lecture Notes in Computer Science*, pages 235–256. Springer, 2010.
- 25 Thore Husfeldt, Ramamohan Paturi, Gregory B. Sorkin, and Ryan Williams. Exponential Algorithms: Algorithms and Complexity Beyond Polynomial Time (Dagstuhl Seminar 13331). *Dagstuhl Reports*, 3(8):40–72, 2013.
- 26 Russell Impagliazzo, Shachar Lovett, Ramamohan Paturi, and Stefan Schneider. 0-1 integer linear programming with a linear number of constraints. *CoRR*, abs/1401.5512, 2014.
- 27 Antoine Joux. *Algorithmic Cryptanalysis*. Chapman & Hall/CRC, 1st edition, 2009.
- 28 Petteri Kaski, Mikko Koivisto, and Jesper Nederlof. Homomorphic hashing for sparse coefficient extraction. In Dimitrios M. Thilikos and Gerhard J. Woeginger, editors, *Parameterized and Exact Computation - 7th International Symposium, IPEC 2012, Ljubljana, Slovenia, September 12-14, 2012. Proceedings*, volume 7535 of *Lecture Notes in Computer Science*, pages 147–158. Springer, 2012.
- 29 Mikko Koivisto. Partitioning into sets of bounded cardinality. In Jianer Chen and Fedor V. Fomin, editors, *Parameterized and Exact Computation, 4th International Workshop, IWPEC 2009, Copenhagen, Denmark, September 10-11, 2009, Revised Selected Papers*, volume 5917 of *Lecture Notes in Computer Science*, pages 258–263. Springer, 2009.
- 30 R. J. McEliece. A Public-Key Cryptosystem Based On Algebraic Coding Theory. *Deep Space Network Progress Report*, 44:114–116, January 1978.
- 31 Daniel Paulusma, Friedrich Slivovsky, and Stefan Szeider. Model counting for cnf formulas of bounded modular treewidth. *Algorithmica*, pages 1–27, 2015.
- 32 Sigve Hortemo Sæther, Jan Arne Telle, and Martin Vatshelle. Solving maxsat and #sat on structured CNF formulas. In Carsten Sinz and Uwe Egly, editors, *Theory and Applications of Satisfiability Testing - SAT 2014, Vienna, Austria, July 14-17, 2014. Proceedings*, volume 8561 of *Lecture Notes in Computer Science*, pages 16–31. Springer, 2014.
- 33 Uwe Schöning. A probabilistic algorithm for k -SAT based on limited local search and restart. *Algorithmica*, 32(4):615–623, 2002.

34 Vijay V. Vazirani. *Approximation Algorithms*. Springer-Verlag New York, Inc., New York, NY, USA, 2001.

A Missing Proofs

A.1 Proof of Theorem 3.1

Given an (n, m, s) -instance of SET COVER, use Observation 2.4 to assure that for every $f \in F$, $|N(f)| \leq \epsilon n$. It is easy to see that this instance is a YES-instance if and only if the SET PARTITION instance (G', s) is a YES-instance, where G' is the bipartite graph with U on one side and F' on the other with for every $f \in F$ and $X \subseteq N(f)$ a vertex in F' with neighborhood X .

A.2 Proof of Theorem 3.2

For the backward direction, an $O(m^c 2^{(1-\epsilon')n})$ algorithm for SET PARTITION concatenated to the reduction of Theorem 3.1 with $\epsilon = \epsilon'/(2c)$ gives a

$$O\left(\left(m 2^{\epsilon' n/(2c)}\right)^c 2^{(1-\epsilon')n} + m 2^{(1-\epsilon)n}\right) \leq O(m^c 2^{(1-\epsilon'/2)n} + m 2^{(1-\epsilon)n})$$

time algorithm for SET COVER.

For the forward direction, given an (n, m, s) -instance of SET PARTITION first use Observation 2.3 with $c = 2/\epsilon$ to obtain an $(n, m^{2/\epsilon}, s\epsilon/2)$ instance of SET PARTITION (without loss of generality, we may assume s divides $2/\epsilon$ by decreasing ϵ and adding at most $2/\epsilon$ sets and elements). Denoting $s' = s\epsilon/2$, now iterate over all unordered integer partitions $a_1 + \dots + a_{s'} = n$, and for each such partition solve a SET COVER instance $((F' \cup U'), E', s')$ constructed as follows from the SET PARTITION instance with the assumed algorithm:

- For every $i = 1, \dots, s'$ add an element e_i to U'
- For every $f \in F$ and i such that $|N(f)| = a_i$ add a set f' to F' with $N(f') = f \cup \{e_i\}$.

Note that possibly $a_i = a_j$ for $i \neq j$ we make at least two copies of every set $f \in F$ with $|N(f)| = a_i$. It is easy to see that this new SET COVER instance is a YES-instance if and only if in the original SET PARTITION instance there is a set partition of size s' : each set only contains one element from $\{e_1, \dots, e_{s'}\}$ so a Set Cover of size s' needs to correspond with $f_1, \dots, f_{s'} \in F$ satisfying $\sum_{i=1}^{s'} |N(f_i)| = n$.

By a result of Hardy and Ramanujan [21], there are at most $2^{O(\sqrt{n})}$ unordered integer partitions and hence concatenating this reduction with the assumed $O(m^c 2^{(1-\epsilon)n})$ time SET COVER algorithm leads to an

$$O(2^{O(\sqrt{n})} m^{c2/\epsilon} 2^{(1-\epsilon)(n+s')}) \leq O^*(2^{(1-\epsilon)(1+\epsilon/2)n}) \leq O^*(2^{(1-\epsilon/2)n})$$

time algorithm for SET PARTITION.

A.3 Proof of Theorem 3.3

► **Theorem A.1** ([29]). *Given an instance n -element set N , an integer r and a family \mathcal{F} of subsets of N each of cardinality at most r , the partitions of N into a give number of members of \mathcal{F} can be counted in time $|\mathcal{F}| 2^{\lambda_r n} \text{poly}(n)$, where $\lambda_r = (2r - 2)/\sqrt{(2r - 1)^2 - 2 \ln 2}$.*

Note that the above result also immediately implies an algorithm deciding SET COVER in time $2^r |\mathcal{F}| 2^{\lambda_r n} \text{poly}(n)$ since we may reduce this SET COVER instance to a SET PARTITION

instance by adding all subsets of the given sets. It is easily seen that for $r \geq 3$, λ_r is sandwiched as

$$1/2 \leq \lambda_r \leq (2r - 2)/\sqrt{(2r - 1.5)^2} = 1 - \frac{1}{O(r)}.$$

The algorithm implementing the theorem is given in Figure 3.

Algorithm A3 ($G = (F \dot{\cup} U, E), s, r$).

- 1: **if** $\exists f \in F : |N(f)| \geq r$ **then**
- 2: **return** $A3(G[F \setminus \{f\}] \cup U, s, r) \vee A2(G[F \setminus \{f\}] \cup U \setminus N(f), s - 1, r)$.
- 3: **else**
- 4: Use the algorithm of Theorem A.1.

■ **Figure 3** Algorithm for SET COVER or SET PARTITION with few sets (implementing Theorem 3.3).

We may bound the running time of Algorithm A3 by analyzing the branching tree of recursive calls where an execution reaching Line 4 represents a leaf. Note that the depth of the branching tree is at most m and if we refer to the second recursive call of Line 2 as the right branch, on any path from the root to a leaf there are at most n/r right branches since in each such recursion step we decrease the size of U in one call with at least r . Thus the number of paths from the root to a leaf in this tree with i right branches is at most $\binom{m}{i}$. For each such a branch the size of U has become $n - ir$ and since Line 4 is reached all sets are of size at most r , Theorem 3.3 applies and thus we spend $m \cdot \text{poly}(n) 2^{\lambda_r(n-ir)}$ per leaf of this type.

Thus, denoting $\mu = m/n$, the total running time can be written as

$$\begin{aligned} m \cdot \text{poly}(n) \sum_{i=1}^m \binom{m}{i} 2^{\lambda_r(n-ir)} &= m \cdot \text{poly}(n) 2^{\lambda_r n} \left(1 + \frac{1}{2^{\lambda_r r}}\right)^m \\ &\leq m \cdot \text{poly}(n) \left(2^{\lambda_r} \exp\left(\mu/2^{r/2}\right)\right)^n, \end{aligned}$$

where the equality follows from the binomial theorem and the inequality uses $(1 + \frac{1}{n})^n \leq e$. Denoting $\mu = m/n$ and setting $r = \lceil 4 \lg(\mu) \rceil$, we see that when taking asymptotics for growing μ (which is allowed since we may assume the running time complexity is monotone increasing with μ), the running time becomes

$$m \cdot \text{poly}(n) \left(2^{1 - \frac{1}{O(\lg \mu)} + \lg(e)/\mu}\right)^n = m \cdot \text{poly}(n) 2^{n - \frac{n}{O(\lg(m/n))}}.$$

A.4 Proof of Theorem 3.4

Given an arbitrary instance of SET PARTITION consisting of $G = (F \dot{\cup} U, E)$ and integer s , first construct an equivalent (n, m', s') instance using Observation 2.3 with $s' \leq \min\{\epsilon_1, \epsilon_2/2\}n$. Then add vertices $e_1, \dots, e_{s'}$ to U and replace every $f \in F$ with s' copies $f_1, \dots, f_{s'}$ where $N(f_i) = N(f) \cup e_i$. It is easy to see that in this SET PARTITION instance all set partitions are of size s' and it has a set partition of size s' if and only if the original instance has one of size s . Thus, to transform the new instance into an equivalent $(n + s', m', \epsilon_1(n + s'))$ -instance, simply add sufficiently many isolated vertices to F . Running the assumed algorithm on this instance results thus in an $O^*(2^{(1-\epsilon_2)(n+s')}) = O^*(2^{(1-\epsilon_2)(1+\epsilon_2/2)n}) = O^*(2^{(1-\epsilon_2/2)n})$ time algorithm.

A.5 Proof of Theorem 3.5

Recall from the main text that the LINEAR SAT problem is defined as follows: given an integer t , matrix $A \in \mathbb{Z}_2^{n \times m}$ and vectors $\mathbf{b} \in \mathbb{Z}_2^n$ and $\boldsymbol{\omega} \in \mathbb{N}^m$ the task is to find $\mathbf{x} \in \mathbb{Z}_2^m$ satisfying $A\mathbf{x} \equiv \mathbf{b}$ and $\boldsymbol{\omega} \cdot \mathbf{x} \leq t$.

We let $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{Z}_2^n$ denote the column vectors of A , and denote $\text{wt}(\mathbf{x})$ for the Hamming weight of a vector $\mathbf{x} \in \mathbb{Z}_2^m$.

The algorithm implementing the theorem is outlined in Algorithm 4. The algorithm assumes the Hamming weight of the vector \mathbf{x} minimizing minimum $\boldsymbol{\omega}\mathbf{x}$ is at most $2m/3$ (which is in order to facilitate the running time analysis). We can assume this since if the minimum set of columns summing to \mathbf{b} consists of more than $2m/3$ columns, the rank of A is at least $2m/3$ (as the solution needs to consist of linearly independent columns) and we can solve the problem in $O^*(2^{m/3})$ time by finding a column basis, iterating over all subsets of columns not in the basis and for each such compute the unique way to extend it to a set of columns summing to \mathbf{b} if it exists (this is a standard technique called ‘Information Set Decoding’ introduced in [30]). Note, to ensure the Hamming weight is at most $2m/3$, we cannot simply assume the solution is at most size at most $m/2$ by looking for the complement otherwise since this gives a maximization problem.

Running Time First note that Line 16 and Line 17 can be implemented in time $2^{s_2}m$. Line 18 can be implemented with linear delay by elementary methods, and thus `list2` runs in $O(m(2^{s_2} + |\text{list2}(A, \mathbf{b}, s_2)|))$ time. For Algorithm `list1`, Line 11 and Line 12, we see that $\mathbb{E}[|\text{list2}(A, \mathbf{b}, s_1/2)|] = \binom{m}{s_1/2} 2^{-s_1}$, because any vector from \mathbb{Z}_2^m of weight $s_1/2$ will be included with probability 2^{-s_1} in the output. Thus these lines run in expected time $O(m(2^{s_1/2} + \binom{m}{s_1/2} 2^{-s_1}))$.

Similarly, the for-loop at Line 13 will take on expectation $|\mathcal{P}|2^{-s_1}$ iterations, where \mathcal{P} is the set of pairs $(\mathbf{x}, \mathbf{y}) \in (\mathbb{Z}_2^m)^2$ such that $\text{wt}(\mathbf{x}) = \text{wt}(\mathbf{y}) = s_1/2$ and $A\mathbf{x} + A\mathbf{y} \equiv \mathbf{b}$. Here we divide by 2^{s_1} since this is the probability that such pair is in $\mathcal{L} \times \mathcal{R}$ because for this it needs to satisfy $H A \mathbf{x} \equiv \mathbf{b}_{\mathcal{L}}$.

For algorithm `A4` at iteration s , Line 3 and Line 5 thus take expected time

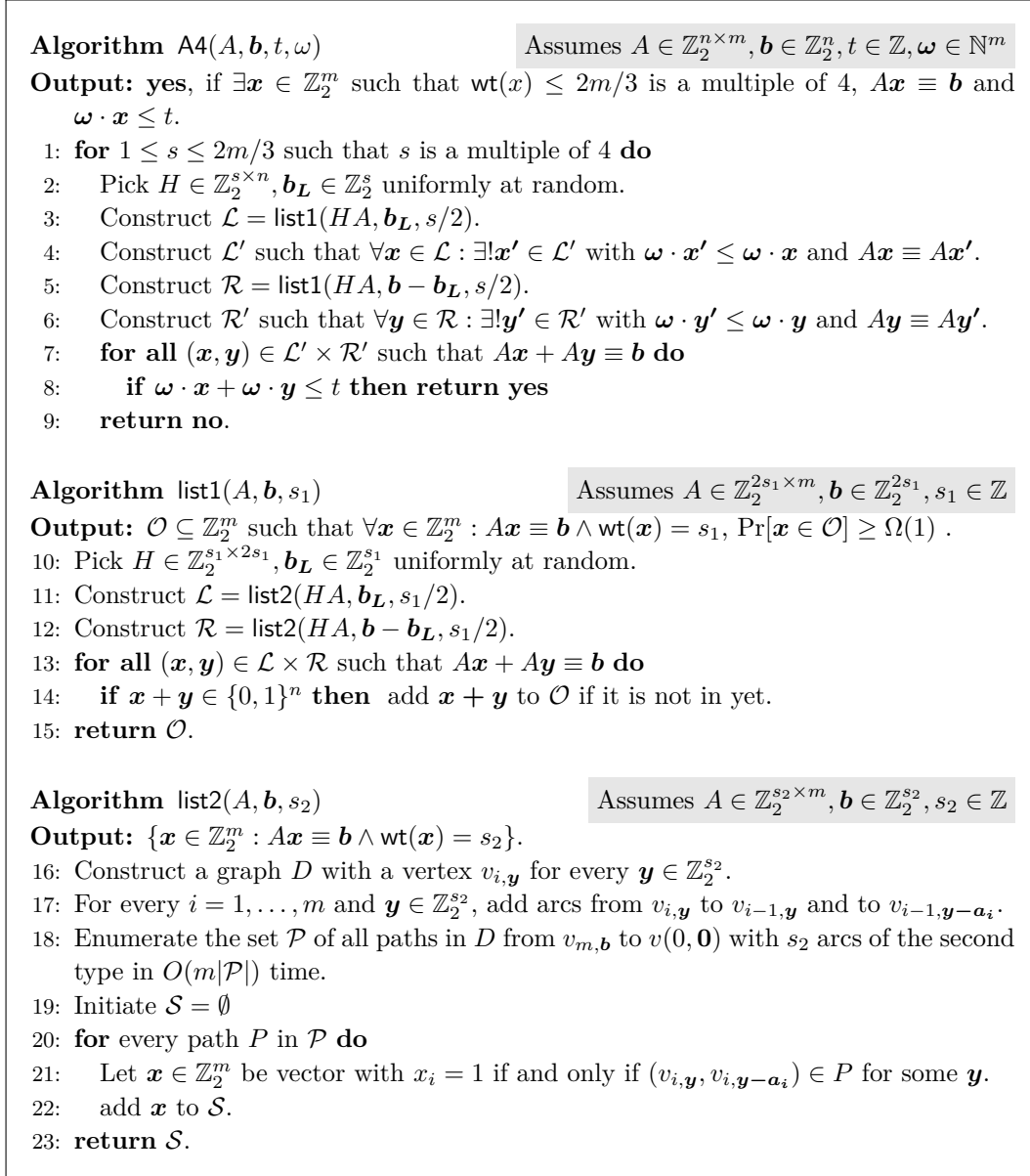
$$O\left(m\left(2^{s_1/2} + \binom{m}{s_1/2} 2^{-s_1} + \mathbb{E}[|\mathcal{P}|2^{-s_1}]\right)\right),$$

where s_1 denotes $s/2$. Note that $\mathbb{E}[|\mathcal{P}|] \leq \binom{m}{s_1/2}^2 / 2^s$ since any pair (\mathbf{x}, \mathbf{y}) in \mathcal{P} also needs to satisfy $A\mathbf{x} + A\mathbf{y} \equiv \mathbf{b}$. Denoting $\sigma = s/m$ and using $s_2 = s_1/2 = s/4$ we can thus upper bound the running time with

$$O\left(m2^m \max_{\sigma \leq 2/3} \{\sigma/4, h(\sigma/4) - \sigma/2, 2h(\sigma/4) - 3\sigma/2\}\right).$$

It is easily verified by standard calculus or a Mathematica computation that the exponent in this expression is maximized for $\sigma = 0.4444$ where it is at most $O(m2^{0.3399m})$. Note that Line 4 and Line 6 can be implemented in time $O(m|\mathcal{L}|)$ and $O(m|\mathcal{R}|)$ with standard data structures and similarly we can efficiently enumerate all pairs in the for-loop at Line 7. Also note that in this for-loop the number of enumerated pairs is at most $|\mathcal{L}'| + |\mathcal{R}'|$ since all vectors $A\mathbf{x}$ for $\mathbf{x} \in \mathcal{L}'$ are different. Thus the algorithm indeed runs in the claimed running time.

Correctness For correctness, first note that whenever `yes` is returned on Line 8 this is clearly correct. Now suppose the instance of LINEAR SAT is a YES-instance. Then there



■ **Figure 4** An $O^*(2^{0.3399n})$ time algorithm for LINEAR SAT.

exists a minimum weight $\mathbf{x} \in \mathbb{Z}_2^m$ satisfying $A\mathbf{x} \equiv \mathbf{b}$ and $\boldsymbol{\omega} \cdot \mathbf{x} \leq t$. Let us consider the iteration of the for-loop on Line 1 where $s = \text{wt}(\mathbf{x})$. Since \mathbf{x} has minimum weight, the columns i of A for which $x_i = 1$ need to be linearly independent since otherwise leaving out a linear combination would result in a smaller weight vector \mathbf{x} . This means that if we denote $\mathcal{Z} = \{A\mathbf{y} : \mathbf{y} \subseteq \mathbf{x} \wedge \text{wt}(\mathbf{y}) = s/2\}$, then $|\mathcal{Z}| = \binom{s}{s/2}$, where $\mathbf{y} \subseteq \mathbf{x}$ denotes $y_i \leq x_i$ for every coordinate i .

For $\mathbf{v} \in \mathbb{Z}_2^s$, let $f(\mathbf{v}) = |\{\mathbf{z} \in \mathcal{Z} : H\mathbf{z} = \mathbf{v}\}|$ be an indicator function. We see that

$$\mathbb{E} \left[\sum_{\mathbf{v} \in \mathbb{Z}_2^s} f(\mathbf{v})^2 \right] = \sum_{\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}^2} \Pr[H(\mathbf{z}_1 - \mathbf{z}_2) = 0] \leq |\mathcal{Z}| + |\mathcal{Z}|^2 2^{-s} \leq 2|\mathcal{Z}|.$$

By Markov's inequality we thus see that $\Pr[\sum_{\mathbf{x} \in \mathbb{Z}_2^s} f(\mathbf{x})^2 \leq 4|\mathcal{Z}|] \geq 1/2$ over the choice of H . Conditioned on this, the Cauchy-Schwarz inequality implies that the number of $\mathbf{x} \in \mathbb{Z}_2^s$ such that $f(\mathbf{x}) > 0$ is at least $|\mathcal{Z}|^2 / (2|\mathcal{Z}|) = |\mathcal{Z}|/2$. When this happens, we have with probability at least $\binom{s}{s/2} / 2^s = \Omega(1/\sqrt{s})$ that $\mathbf{b}_L \in \mathcal{Z}$. If $\mathbf{b}_L \in \mathcal{Z}$, we can apply the same reasoning to show that $\text{list1}(HA, \mathbf{b}_L, s/2)$ contains \mathbf{y} with probability $\Omega(1/\sqrt{s})$ and $\text{list1}(HA, \mathbf{b} - \mathbf{b}_L, s/2)$ contains $\mathbf{x} - \mathbf{y}$ with probability $\Omega(1/\sqrt{s})$, and the pair will be considered in the loop on Line 7. Since the latter two probabilities are independent we see that if a solution exists we find it with probability at least $\Omega(n^{-1.5})$, and thus we may repeat the algorithm $O(n^{1.5})$ times to ensure it finds a solution with constant probability if it exists.

A.6 Proof of Lemma 4.3

Let $F = \{f_1, \dots, f_m\}$ be arbitrarily ordered. For integers $i \in [n]$, $j \in [m]$ and $X \subseteq U$ define $c_j^i[X]$ to be true if and only if there exists $S_1 \subseteq \{f_1, \dots, f_j\}$ such that $|S_1| = i$, $N(S_1) = X$ and for every $f, f' \in S_1$ with $f \neq f'$, $N(f) \cap N(f') = \emptyset$. We see that $c_0^i[X]$ is true if and only if $i = 0$ and $X = \emptyset$, and for $j > 0$ we have

$$c_j^i[X] = (c_{j-1}^{i-1}[X \setminus N(f_j)] \wedge N(f_j) \subseteq X) \vee c_{j-1}^i[X].$$

The values $c_i(W)$ for $W \in \mathcal{W}$ and $i \in [n]$ can be read off from $c_i^j(W)$, and because for computing $c_j^i[X]$ we only need the entries $c_j^i[Y]$ where $Y \subseteq X$, we can restrict our computation to computing $c_j^i(X)$ for $X \in \downarrow\mathcal{W}$, and the runtime bound follows.

A.7 Proof of Lemma 4.6

Define $f_x(X)$ to be true if and only if $\exists f \in F : N(f) = X$ and $|X| = x$. Note we can, within the claimed time bound, create a table storing the values $f_x(X)$ for all $X \in \downarrow\mathcal{W}$ using the oracle. Now define $g_x(Y) = \sum_{Y \subseteq X} f_x(X)$. Let $U = \{u_1, \dots, u_n\}$ and define

$$g_x^j(X) = \sum_{X \cap \{u_1, \dots, u_j\} \subseteq Y \subseteq X} f_x(Y).$$

Then we see that $g_x^n(X) = f_x(X)$, and for $j < n$ we can compute g_x^j using

$$g_x^j(X) = [u_j \in X]g_x^{j+1}(X \setminus \{u_j\}) + g_x^{j+1}(X). \tag{A.1}$$

Thus, by straightforward dynamic programming using (A.1) we can compute $g_x(X) = g_x^0(X)$ for every x and $X \in \downarrow\mathcal{W}$ in $O(n^2|\downarrow\mathcal{W}|)$ time. Next, define

$$h_{x,i}(X) = \sum_{x_1+x_2+\dots+x_i=x} g_{x_1}(X)g_{x_2}(X) \cdots g_{x_i}(X),$$

or equivalently, $h_{x,i}$ is the number of tuples f_1, \dots, f_i such that $N(f_i) \subseteq X$ and $\sum_{l=1}^i |N(f_l)| = x$. Note that for a fixed X , given the entries $g_x(X)$ for every $x \leq n$, we can compute $h_{x,i}(X)$ in time $\text{poly}(n)$ using standard dynamic programming or the Fast Fourier Transformation. Denote

$$c'_{x,i}(X) = \left| \left\{ (f_1, \dots, f_i) \in F^i : N(\{f_1, \dots, f_i\}) = X \wedge \sum_{l=1}^i |N(f_l)| = x \right\} \right|,$$

which is easily seen to be the number i -tuples of disjoint sets whose union is exactly X if $x = |X|$. Note that $h_{x,i}(X) = \sum_{Y \subseteq X} c'_{x,i}(Y)$ since every i -tuple counted in $h_{x,i}(X)$ is counted once in a $c'_{x,i}(Y)$ where Y is the union of the i -tuple. Then, since any non-empty set has equally many even-sized as odd-sized subsets (e.g., by inclusion exclusion), we see that

$$\begin{aligned} c'_{x,i}(X) &= \sum_{Y \subseteq X} \left(\sum_{Z \subseteq X \setminus Y} (-1)^{|Z|} \right) c'_{x,i}(Y) \\ &= \sum_{Z \subseteq X} (-1)^{|Z|} \left(\sum_{Y \subseteq X \setminus Z} c'_{x,i}(Y) \right) = \sum_{Z \subseteq X} (-1)^{|Z|} h_{x,i}(X \setminus Z). \end{aligned} \tag{A.2}$$

Then, define

$$h_{x,i}^j(X) = \sum_{X \cap \{u_1, \dots, u_j\} \subseteq Z \subseteq X} (-1)^{|Z|} h_{x,i}(X \setminus Z),$$

and similarly to (A.1), we see that $h_{x,i}^n(X) = h_{x,i}(X)$ and for $j < n$ we can compute $h_{x,i}^j$ using

$$h_{x,i}^j[X] = -[u_i \in X] h_{x,i}^{j+1}(X \setminus \{u_i\}) + h_{x,i}^{j+1}(X). \tag{A.3}$$

Consequently, we can use Equation A.3 for a straightforward dynamic programming algorithm to determine $h_{|X|,i}^0[X]$ which equals $c'_{x,i}(X)$ by (A.2), and it is easy to see that $c_i(X)$ is true if and only if $c_{|X|,i}(X) > 0$.