Tight Bounds on Subexponential Time Approximation of Set Cover and Related Problems

Marek Cygan
University of Warsaw, Poland
mcygan@mimuw.edu.pl

Magnús M. Halldórsson
ICE-TCS & Department of Computer Science, Reykjavik University, Iceland
mmh@ru.is

Guy Kortsarz
Rutgers University, Camden, NJ, USA
guyk@rutgers.edu

Abstract

We show that Set-Cover on instances with N elements cannot be approximated within $(1-\gamma)\ln N$-factor in time $\exp(N^{\gamma-o(1)})$, for any $0 < \gamma < 1$, assuming the Exponential Time Hypothesis. This essentially matches the best upper bound known by Cygan et al. [7] of $(1-\gamma)\ln N$-factor in time $exp(O(N^{\gamma}))$.

The lower bound is obtained by extracting a standalone reduction from Label-Cover to Set-Cover from the work of Moshkovitz [20], and applying it to a different PCP theorem than done there. We also obtain even tighter lower bounds that are conditional on certain PCP conjectures.

We also treat three problems (Directed Steiner Tree, Submodular Cover, and Connected Polymatroid) that strictly generalize Set-Cover. We give a $(1-\gamma)\ln N$-approximation algorithm for these problems that runs in $exp(\tilde{O}(N^{\gamma}))$ time, for any $1/2 \le \gamma < 1$.

2012 ACM Subject Classification Theory of Computation → Design and analysis of algorithms; Mathematics of computing → Discrete mathematics; Theory of computation → Theory and algorithms for application domains

Keywords and phrases subexponential time algorithms, lower bounds, set cover

Digital Object Identifier 10.4230/LIPIcs.SWAT.2020.11

Funding The work of Marek Cygan is part of a project TOTAL that has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 677651). Magnús Halldórsson is partially supported by Icelandic Research Fund grant 174484-051. Guy Kortsarz is partially supported by NSF grants 1540547 and 1910565.

Acknowledgements We thank Bundit Laekhanukit for helpful comments and discussions.

1 Introduction

We show that Set-Cover on instances with N elements cannot be approximated within $(1-\gamma)\ln N$-factor in time $\exp(N^{\gamma-o(1)})$, assuming the Exponential Time Hypothesis (ETH). This essentially matches the best upper bound known by Cygan et al. [7]. This is obtained by extracting a standalone reduction from Label-Cover to Set-Cover from the work of Moshkovitz [20], and applying it to a different PCP theorem than done there. We also obtain even tighter lower bounds that are conditional on certain PCP conjectures, culminating in a time lower bound of $exp(\tilde{\Omega}(N^{\gamma}))$.

We also treat the Directed Steiner Tree (Dst) problem that strictly generalizes Set-Cover. The input to Dst consists of a directed graph G with costs on edges, a set of terminals, and a designated root r. The goal is to find a subgraph of G that forms an

SWAT 2020.
Editors: Susanne Albers; Article No. 11; pp. 11:1–11:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

arborescence rooted at r containing all the N terminals and minimizing the cost. We give a $(1-\gamma)\ln N$-approximation algorithm for DST that runs in $exp(\tilde{O}(N^{\gamma}))$ time, for any $\gamma \geq 1/2$.

This algorithm also applies to two other generalizations of SET-COVER. In the SUBMODULAR COVER problem, the input is a set system $(U, \mathcal{C})$ with a cost on each element of the universe U. We are given a non-decreasing *submodular* function $f : 2^U \to \mathbb{R}$ satisfying, for every $S \subseteq T \subseteq V$ and for every $x \in U \setminus T$, $f(S+x) - f(S) \geq f(T+x) - f(T)$. The objective is to minimise $c(S)$ subject to $f(S) = f(U)$. In the CONNECTED POLYMATROID problem, which generalizes SUBMODULAR COVER, the elements of U are leaves of a tree and both elements and sets have cost. The goal is to select a set $S \subseteq U$ so that $f(S) = f(U)$ and $c(S) + c(T(S))$ is minimized, where $T(S)$ is the unique tree rooted at r spanning S.

1.1 Related work

Johnson [16] and Lovász [18] showed that a greedy algorithm yields a $1 + \lg N$-approximation of Set Cover, where N is the number of elements. Chvátal [5] extended it to the weighted version. Slavik [24] refined the bound to $\ln n - \ln\ln n + O(1)$.

Lund and Yannakakis [19] showed that logarithmic factor was essentially best possible for polynomial-time algorithms, unless $NP \subseteq DTIME(n^{polylog(n)})$. Feige gave the precise lower bound that SET-COVER admits no $(1-\epsilon)$-approximation, for any $\epsilon > 0$, with a similar complexity assumption. Assuming the stronger ETH, he shows that $\ln N - c\log\log N$-approximation is not possible for polynomial algorithms, for some $c > 0$. The work of Moshkovitz [20] and Dinur and Steurer [9] combined shows that $(1-\epsilon)\ln N$-approximation is hard modulo $P = NP$. All the inapproximability results relate to two prover interactive proofs, via the related LABEL-COVER problem.

Recent years have seen increased interest in subexponential time algorithm, including approximation algorithms. A case in point is the *maximum clique* problem that has a trivial $2^{n/\alpha}\text{poly}(n)$-time algorithm that gives a α-approximation, for any $1 \leq \alpha \leq \sqrt{n}$, and Bansal et al. [1] improved the time to $exp(n/\Omega(\alpha \log^2 \alpha))$. Chalermsook et al. [3] showed that this is nearly tight, as α-approximation requires $exp(n^{1-\epsilon}/\alpha^{1+\epsilon})$ time, for any $\epsilon > 0$, assuming ETH.

For SET-COVER, Cygan, Kowalik and Wykurz[7] gave a $(1-\alpha)\ln N$-approximation algorithm that runs in time $2^{O(n^{\alpha})}$, for any $0 < \alpha < 1$. The results of [20, 9] imply a $exp(N^{\alpha/c})$-time lower bound for $(1-\alpha)\ln N$-approximation, for some constant $c \geq 3$. An unpublished report contains a conditional $exp(\Omega(N^{\alpha}))$-time lower bound for $(1-\alpha)\ln N$-approximation of SET-COVER [6]. In addition to ETH, this also requires the less established Projection Games Conjecture (PGC). Unfortunately, the writeup of [6] defies easy verification. The current paper arose from an effort to make it comprehensible.

DST can be approximated within N^{ϵ}-factor, for any $\epsilon > 0$, in polynomial time [17, 4]. In quasi-polynomial time, it can be approximated within $O(\log^2 N/\log\log N)$-factor [13], which is also best possible in that regime [14, 13]. This was recently extended to CONNECTED POLYMATROID [12]. For SUBMODULAR COVER, the greedy algorithm also achieves a $1 + \ln N$-approximation [25].

2 Hardness of Set Cover

We give our technical results in this section. Starting with definitions of ETH and LABEL-COVER in Sec. 2.1, we give a reduction from LABEL-COVER to SET-COVER in Sec. 2.2, and derive specific approximation hardness results in Sec. 2.3. A full proof of the correctness of the reduction is given in the following section.

2.1 Preliminaries

The Exponential Time Hypothesis asserts that the 3-Sat problem on n variables and m clauses cannot be solved in $2^{o(n)}$-time. Impagliazzo, Paturi and Zane [15] showed that any 3-Sat instance can be sparsified in $2^{o(n)}$-time to an instance with $m = O(n)$ clauses. When we refer to Sat input of size n, we mean 3-CNF formula on n variable and $O(n)$ clauses. Thus, ETH together with the sparsification lemma [2] implies the following:

▶ **Conjecture 2.1.** *(*ETH*) Given a boolean SAT input ϕ of size n, there is no $2^{o(n)}$-time algorithm that decides whether ϕ is satisfiable.*

The intermediate problem in all approximation hardness reduction for Set-Cover is the Label-Cover problem.

▶ **Definition 2.2.** *In the* Label-Cover *problem with the projection property (a.k.a., the* Projection Game*), we are given a bipartite graph $G(A, B, E)$, finite alphabets (also called* labels*) Σ_A and Σ_B, and a function $\pi_e : \Sigma_A \to \Sigma_B$ for each edge $e \in E$. A* labeling *is a pair $\varphi_A : A \to \Sigma_A$ and $\varphi_B : B \to \Sigma_B$ of assignments of labels to the vertices of A and B, respectively. An edge $e = (a, b)$ is* covered *(or* satisfied*) by (φ_A, φ_B) if $\pi_e(\varphi_A(a)) = \varphi_B(b)$. The goal in* Label-Cover *is to find a labeling (φ_A, φ_B) that covers as many edges as possible.*

The *size* of a label cover instance $\mathcal{G} = (G = (A, B, E), \Sigma_A, \Sigma_B, \Pi = \{\pi_e\}_e)$ is denoted by $n_{\mathcal{G}} = |A| + |B| + |E|$. The *alphabet size* is $\max(|\Sigma_A|, |\Sigma_B|)$. The Label-Cover instances we deal with will be *bi-regular*, meaning that all nodes of the same bipartition have the same degree. We refer to the degree of nodes in A (B) as the *A-degree* (*B-degree*), respectively.

Label-Cover is a central problem in computational complexity, corresponding to *projection PCPs*, or probabilistically checkable proofs that make 2 queries. A key parameter is the *soundness error*:

▶ **Definition 2.3.** *A* Label-Cover *construction $\mathcal{G}_\phi$, formed from a 3-SAT formula ϕ, has* soundness error ϵ *if: a) whenever ϕ is satisfiable, there is a labeling of $\mathcal{G}_\phi$ that covers all edges, and b) when ϕ is unsatisfiable, then every labeling of $\mathcal{G}$ covers at most an ϵ-fraction of the edges.*

A Label-Cover construction is *almost-linear size* if it is of size $n^{1+o(1)}$, possibly with extra $poly(1/\epsilon)$ factors.

2.2 Set Cover Reduction

We present here a reduction from a generic Label-Cover (a two-prover PCP theorem) to the Set-Cover problem. This is extracted from the work of Moshkovitz [20]. The presentation in [20] was tightly linked with the PCP construction of Dinur and Steurer [9] that was used in order to stay within polynomial time. When allowing superpolynomial time, it turns out to be more frugal to apply the older PCP construction of Moshkovitz and Raz [21].

Our main technical contribution is then to provide a standalone reduction from Label-Cover to Set-Cover that allows specific PCP theorems to be plugged in.

Recall that the $\tilde{O}$-notation hides logarithmic factors. We say that a reduction that originates in Sat achieves *approximation gap ρ* if there is a value a such that Set-Cover instances originating in satisfiable formulas have a set cover of size at most a, while instances originating in unsatisfiable formulas have all set covers of size greater than $\rho \cdot a$.

▶ **Theorem 2.4.** *There is a reduction from* Label-Cover *to* Set-Cover *with the following properties. Let $\mathcal{G}$ be a bi-regular, almost-linear size* Label-Cover *for deciding the satisfiability of a boolean formula ϕ, with size B-degree $poly(1/\epsilon)$. Let n' denote its size, and $\sigma_A(\epsilon)$ the alphabet size as a function of the soundness error ϵ.*

Then the resulting Set-Cover *instance $\mathcal{SC}_\mathcal{G}$ has approximation gap $(1-\gamma)\ln N$, $N = \tilde{O}(n'^{1/\gamma})$ elements, and $M = \tilde{O}(n') \cdot \sigma_A(polylog(n))$ sets. The time of the reduction is linear in the size of $\mathcal{SC}_\mathcal{G}$.*

2.3 Approximation Hardness Results

We use the following PCP theorem of Raz and Moshkovitz [21].

▶ **Theorem 2.5** ([21])**.** *For every $\epsilon \geq 1/polylog(n)$,* Sat *on input of size n can be reduced to* Label-Cover *on a bi-regular graph of degrees* $\mathrm{poly}(1/\epsilon)$*, with soundness error ϵ, size $n^{1+o(1)}\mathrm{poly}(1/\epsilon)$, and alphabet size that is exponential in $poly(1/\epsilon)$. The reduction can be computed in linear time in the size and alphabet size of the Label-Cover.*

We are now ready for our main result. We define the notation $\tilde{o}(f(n)) = \{h(n) : \exists c > 0, h(n) = o(f(n)\log^c n)\}$ to be the collection of functions that grow slower than some polylog factor times the function f.

▶ **Theorem 2.6.** *Let $0 < \gamma < 1$. Assuming ETH, there is no $(1-\gamma)\ln N$-approximation of* Set-Cover *with N elements and M sets that runs in time $2^{N^\gamma - o(1)} \cdot poly(M)$.*

More precisely, represent the size of the Label-Cover *construction of [21] is $n' = \tilde{O}(n)h(n)$, for a sublinear function h, and let $g(N) = \tilde{o}(h(n))$. Then obtaining $(1-\gamma)\ln N$-approximation in $2^{N^\gamma/g(N)} \cdot poly(M))$ time is impossible, assuming ETH.*

Proof. We start with the Label-Cover construction $\mathcal{G}'$ of [21] from Thm. 2.5, with $\epsilon = \Theta(1/\log^2 n)$, which has size $n' = n \cdot h(n)\mathrm{poly}(1/\epsilon) = \tilde{O}(n \cdot h(n))$ (where $h(n) = n^{o(1)}$), alphabet size $\sigma_A(\epsilon) = exp(poly(1/\epsilon)) = exp(polylog(n))$, a B-degree $q \doteq \mathrm{poly}(1/\epsilon) = polylog(n)$. We apply Thm. 2.4 to obtain a Set-Cover instance $\mathcal{SC}_\mathcal{G}$ with $N = \tilde{O}(n'^{1/\gamma})$ elements, $M = \tilde{O}(n')\sigma_A(polylog(n)) = exp(polylog(n))$ and approximation gap $(1-\gamma)\ln N$.

Suppose there is a $(1-\gamma)\ln N$-approximation algorithm of Set-Cover running in $exp(N^\gamma/g(n) \cdot polylog(n))poly(M)) = exp(N^\gamma/g(n) \cdot polylog(n))$ time. Since it achieves this approximation, it can decide the satisfiability of ϕ. Now, $N^\gamma = \tilde{O}(n'^{1/\gamma})^\gamma = \tilde{O}(n') = \tilde{O}(n)h(n)$. By the definition of $g(N)$, $g(N) = \tilde{o}(h(n))$. Thus, the running time of the algorithm is

$$exp(N^\gamma/g(N) \cdot polylog(n)) = exp(\tilde{O}(n \cdot h(n)/g(N)) = exp(o(n)) \ .$$

This contradicts ETH. ◀

Note: We could also allow the algorithm greater than polynomial complexity in terms of M, such as $exp(polylog(M))$.

2.3.1 Still tighter bounds under stronger assumptions

Moshkovitz [20] proposed a conjecture on the parameters of possible Label Cover constructions. We require a particular version with almost linear size and low degree.

▶ **Definition 2.7** (The Projection Games Conjecture (PGC))**.** *3-SAT of inputs of size n can be reduced to* Label-Cover *of size $n^{1+o(1)}$, alphabet size $poly(1/\epsilon)$, and bi-regular degrees $poly(1/\epsilon)$, where ϵ is the soundness error parameter.*

The key difference of PGC from known PCP theorems is the alphabet size. PGC is considered quite plausible and has been used to prove conditional hardness results for a number of problems [20, 22].

By assuming PGC, we improve the dependence on M, the number of sets.

▶ **Theorem 2.8.** *Let $0 < \gamma < 1$. Assuming PGC and ETH, there is no $(1-\gamma)\ln N$-approximation, nor a $O(\log M)$-approximation, of* SET-COVER *with N elements and M sets that runs in time $2^{N^{\gamma-o(1)}M^{1-o(1)}}$.*

Namely, both the approximation factor and the time complexity can depend more strongly on the number of sets in the set cover instance. The only result known in terms of M is a folklore $\sqrt{M}$-approximation in polynomial time.

Proof. We can proceed in the same way as in the proof of Thm. 2.6, but starting from the conjectured LABEL-COVER given by PGC, in which the alphabet size is polynomial in ϵ. We then obtain a set cover instance $\mathcal{SC}_{\mathcal{G}}$ that differs only in that M is now $|A||\Sigma_A| = n^{1+o(1)} = N^{\Theta(\gamma)}$. So, a $c \log M$-approximation, with $c > 1$, leads to a $c\gamma \ln N$-approximation, which is smaller than $(1-\gamma)\ln N$-approximation when $c\gamma < (1-\gamma)$ or $\gamma < 1/(2c)$. Also, $2^{M-o(1)} = 2^{o(n)}$, for the right choice of $o(1)$ function. Hence, such a running time again breaks ETH. ◀

A different hypothesis can be made regarding the *size* of the LABEL-COVER's. While the size of the LABEL-COVER construction of [21] is $n^{1+o(1)}$, it is plausible that it could be reduced to $npoly(1/\epsilon) = n \cdot polylog(n)$. In fact, several constructions with different properties indeed have such size bounds (e.g., [8]). This would have the following implication regarding the best possible approximation factors.

▶ **Theorem 2.9.** *Suppose there are* LABEL-COVER *constructions from SAT of size $n \cdot polylog(n)poly(1/\epsilon)$ and degrees $poly(1/\epsilon)$. Then, assuming ETH, there is no $(1-\gamma)\ln N$-approximation of* SET-COVER *with N elements and M sets that runs in time $exp(N^\gamma/polylog(N)) \cdot poly(M)$, for any $\gamma > 0$.*

Alternatively, assuming ETH, there is a constant $c > 0$ such that there is no $(1-\gamma)\ln N - c \log\log N$-approximation algorithm running in time $2^{N^\gamma} \cdot poly(M)$, for any $\gamma > 0$.

Note that alphabet size may remain exponential in $1/\epsilon$.

3 Proof of the Set Cover Reduction

We extract here a sequence of two reductions from the work of Moshkovitz. By untangling them from the Label Cover construction, we can use them for our standalone Set Cover reduction.

Moshkovitz [20] chooses some of the parameters of the lemmas so as to fit her purpose of proving the NP-hardness of approximation. As a result, the size of the intermediate LABEL-COVER instance generated grows to be a polynomial of with a degree larger than 1. This would lead to weaker hardness results for sub-exponential time algorithms than what we desire. We indicate therefore how we can separate a key parameter to maintain nearly-linear size label covers.

A key tool in her argument is the concept of agreement soundness error.

▶ **Definition 3.1** (List-agreement soundness error). *Let $\mathcal{G}$ be a* LABEL-COVER *for deciding the satisfiability of a Boolean formula ϕ. Let φ_A assign each A-vertex ℓ alphabet symbols. We say* *that the A-vertices* totally disagree *on a vertex $b \in B$ if, there are no two neighbors $a_1, a_2 \in A$ of b for which there exist $\sigma_1 \in \varphi_A(a_1), \sigma_2 \in \varphi_A(a_2)$ such that $\pi_{(a_1,b)}(\sigma_1) = \pi_{(a_2,b)}(\sigma_2)$.*

We say that $\mathcal{G}$ has a list-agreement soundness error *if, for unsatisfiable ϕ, for any assignment $\varphi_A : A \to \binom{\sigma_A}{\ell}$, the A-vertices are in total disagreement on at least $1-\epsilon$ fraction of the $b \in B$.*

The reduction of Thm. 2.4 is obtained by stringing together two reductions: from Label-Cover to a modified Label-Cover with a low agreement soundness error, and from that to Set-Cover.

The first one is laid out in the following lemma that combines Lemmas 4.4 and 4.7 of [20]. The proofs of this and the next lemma are given in the upcoming subsection.

▶ **Lemma 3.2.** *Let $D \geq 2$ be a prime power, q be a power of D, $\ell > 1$, and $\epsilon_0 > 0$. There is a polynomial reduction from a* Label-Cover *with soundness error $\epsilon_0^2 D^2$ and B-degree q to a* Label-Cover *with list-agreement soundness error $(\ell, 2\epsilon_0 D^2 \cdot \ell^2)$ and B-degree D. The reduction preserves alphabets, and the size is increased by poly(q)-factor.*

Moshkovitz also gave a reduction from Label-Cover with small agreement soundness error to Set-Cover approximation. We extract a more general parameterization than is stated explicitly around Claim 4.10 in [20]. The proof, with only minor changes from [20], is given in the appendix, for completeness.

▶ **Lemma 3.3** ([20], rephrased). *Let $\mathcal{G}' = (G' = (A', B', E'), \Sigma_A, \Sigma_B, \Pi')$ be a bi-regular* Label-Cover *instance with B'-degree D, for deciding the satisfiability of a boolean formula ϕ. For every $0 < \alpha < 1$, and any $u \geq (D^{c_0 \log D} \log |\Sigma_B|)^{1/\alpha}$ for a certain absolute constant c_0, there is a reduction from $\mathcal{G}$ to a* Set-Cover *instance $\mathcal{SC}_\mathcal{G}$ with the following properties:*

1. Completeness: *If all edges of G can be covered, then $\mathcal{SC}_\mathcal{G}$ has a set cover of size $|A'|$.*
2. Soundness: *If $\mathcal{G}$ has list-agreement soundness error (ℓ, α), where $\ell = D(1-\alpha)\ln u$, then every set cover of $\mathcal{SC}_\mathcal{G}$ is of size more than $|A'|(1-2\alpha)\ln u$.*
3. *The number N of elements of $\mathcal{SC}_\mathcal{G}$ is $|B'| \cdot u$ and the number M of sets is $|A'| \cdot |\Sigma_A|$.*
4. *The time for the reduction is polynomial in $|A'|, |B'|, |\Sigma_A|, |\Sigma_B|$ and u.*

Theorem 2.4 (restated): *T*here is a reduction from Label-Cover to Set-Cover with the following properties. Let $\mathcal{G}$ be a bi-regular, almost-linear size Label-Cover for deciding the satisfiability of a boolean formula ϕ, with size B-degree *poly*$(1/\epsilon)$. Let n' denote its size, and $\sigma_A(\epsilon)$ the alphabet size as a function of the soundness error ϵ.

Then the resulting Set-Cover instance $\mathcal{SC}_\mathcal{G}$ has $N = \tilde{O}(n'^{1/\gamma})$ elements, $M = \tilde{O}(n')\sigma_A(polylog(n))$ and approximation gap $(1-\gamma)\ln N$. The time of the reduction is linear in the size of $\mathcal{SC}_\mathcal{G}$.

Proof of Thm. 2.4. Let D be a prime power, which we may choose as 2. Let $n'' = n' poly(1/\epsilon)$ be the size of the instance that is formed by Lemma 3.2. Let $u = n''^{(1-\gamma)/\gamma}$. Let α be the largest value such that

$$u \geq (D^{\Theta(\log D)} \log |\Sigma_B|)^{1/\alpha} .$$

Let $\ell = D(1-\alpha)\ln u$, $\epsilon = \alpha^2 D^{-2} \ell^{-4}/4$, and $\epsilon_0 = \sqrt{\epsilon}/D = \alpha/(2D\ell^2)$. Let q be the B-degree of $\mathcal{G}$. Observe that $\log u = \Theta(\log n)$, so $\ell = \Theta(\log n)$. The values of $1/\alpha$, $1/\epsilon$, and $1/\epsilon_0$ are polynomially related. Observe that since alphabets are of size at most $exp(1/\epsilon)$, $1/\alpha = \Theta(\log n / \log\log n)$. Then, $1/\epsilon$ and $1/\epsilon$ are $polylog(n)$.

We apply Lemma 3.2 to $\mathcal{G}$ (with D, q, ℓ and ϵ_0) we obtain a Label-Cover instance $\mathcal{G}' = (G' = (A', B', E'), \Sigma_A, \Sigma_B, \Pi)$ with B-degree D and size $n'' = n \cdot h(n) poly(1/\epsilon) =$

$n \cdot h(n) polylog(n)$, with alphabets unchanged. The list-agreement soundness error of $\mathcal{G}'$ is $(\ell, 2\epsilon_0 D^2)$, or (ℓ, α).

Now, $\mathcal{G}'$, α, and D and u satisfy the prerequisites of Lemma 3.3, which yields a Set-Cover instance $\mathcal{SC}_\mathcal{G}$ with $N = |B'| \cdot u = n'' \cdot n''^{(1-\gamma)/\gamma} = n''^{1/\gamma}$ elements and $M = |A'| \cdot |\Sigma_A|$ sets.

If ϕ is satisfiable, then $\mathcal{SC}_\mathcal{G}$ has set cover of size $|A'|$, while if it is unsatisfiable, then every set cover of $\mathcal{SC}_\mathcal{G}$ has size more than $|A'|(1-2\alpha)\ln u$. Note that $\ln u = \ln(N/n'') = (1-\gamma)\ln N$, while $2\alpha = \Theta(\log\log N/\log N)$. Hence, the approximation gap is $(1-2\alpha)\ln u = (1-\Theta(\log\log N/\log N))(1-\gamma)\ln N \geq (1-\gamma-\Theta(\log\log N/\log N))\ln N$.

We can get the approximation gap into the form $(1-\gamma')\ln N$ by working with a slightly larger value $\gamma' = \gamma + c\log\log N/\ln N$ and defining $u = n''^{1/\gamma'}$. Then, the approximation gap becomes $(1-\gamma'-c\log\log N/\log N)\ln N = (1-\gamma)\ln N$. The effect on the time complexity is minimal, since $N^{\gamma'} = N^{\gamma} exp(c\log\log N) = N^{\gamma} polylog(N)$. Thus, the extra factor folds into the polylog factors. ◀

3.1 Proof of Lemma 3.2

We give here a full proof of Lemma 3.2, based on [20], with minor modification.

When the labeling assigns a single label to each node, i.e., when $\ell = 1$, the list-agreement soundness error reduces to a *agreement soundness error*, which is otherwise defined equivalently. Moshkovitz first showed how to reduce a Label-Cover with small soundness error to one with a small agreement soundness error. The lemma stated here is unchanged from [20] except that Moshkovitz used the parameter name n instead of our parameter q. This was confusing, since n was also used to denote the size of the Label-Cover (like we do here).

▶ **Lemma 3.4** (Lemma 4.4 of [20]). *Let $D \geq 2$ be a prime power and let q be a power of D. Let $\epsilon_0 > 0$. There is a polynomial reduction from a Label-Cover with soundness error $\epsilon_0^2 D^2$ and B-degree q to a Label-Cover with agreement soundness error $2\epsilon_0 D^2$ and B-degree D. The reduction preserves alphabets, and the size is increased by poly(q)-factor.*

We have underlined the parts that changed because of using q as parameter instead of n. The proof is based on the following combinatorial lemma, whose proof we omit. We note that the set U here is different from the one used in Lemma 3.3 (but we retained the notation to remain faithful to [20]).

▶ **Lemma 3.5** (Lemma 4.3 of [20]). *For $0 < \epsilon < 1$, for a prime power D, and q that is a power of D, there is an explicit construction of a regular bipartite graph $H = (U, V, E)$ with $|U| = q$, V-degree D, and $|V| \leq q^{O(1)}$ that satisfies the following. For every partition $U_1, \ldots, U_\ell$ of U into sets such that $|U_i| \leq \epsilon|U|$ for $i = 1, 2, \ldots, \ell$, the fraction of vertices $v \in V$ with more than one neighbor in any single set U_i, is at most ϵD^2.*

Again, we used the parameter name q, rather than n as in [20]. We show how to take a Label-Cover with standard soundness and convert it to a Label-Cover instance with total disagreement soundness, by combining it with the graph from Lemma 3.5

Proof of Lemma 3.4. Let $\mathcal{G} = (G = (A, B, E), \Sigma_A, \Sigma_B, \Pi)$ be the original Label-Cover. Let $H = (U, V, E_H)$ be the graph from Lemma 3.5, where q, D and ϵ are as given in the current lemma. Let us use U to enumerate the neighbors of a B-vertex, i.e., there is a function $E^{\leftarrow} : B \times U \to A$ that given a vertex $b \in B$ and $u \in U$ gives us the A-vertex which is the u neighbor (in G) of b.

We create a new LABEL-COVER $(G = (A, B \times V, E'), \Sigma_A, \Sigma_B, \Pi')$. The intended assignment to every vertex $a \in A$ is the same as its assignment in the original instance. The intended assignment to a vertex $\langle b, v\rangle \in B \times V$ is the same as the assignment to b in the original game. We put an edge $e' = (a, \langle b, v\rangle)$ if $E^{\leftarrow}(b, u) = a$ and $(u, v) \in E_H$. We define $\pi_{e'} \equiv \pi_{(a,b)}$.

If there is an assignment to the original instance that satisfies c fraction of its edges, then the corresponding assignment to the new instance satisfies c fraction of its edges.

Suppose there is an assignment for the new instance $\varphi_A : A \to \Sigma_A$ in which more than $2\epsilon D^2$ fraction of the vertices in $B \times V$ do not have total disagreement.

Let us say that $b \in B$ is "good" if for more than an ϵD^2 fraction of the vertices in $\{b\} \times V$, the A-vertices do not totally disagree. Note that the fraction of good $b \in B$ is at least ϵD^2.

Focus on a good $b \in B$. Consider the partition of U into $|\Sigma_B|$ sets, where the set corresponding to $\sigma \in \Sigma_B$ is:

$$U_\sigma = \{u \in U | a = E^{\leftarrow}(b, u) \wedge e = (a, b) \in E_G \wedge \pi_e(\varphi_A(a)) = \sigma\} .$$

By the goodness of b and the property of H, there must be $\sigma \in \Sigma_B$ such that $|U_\sigma| > \epsilon |U|$. We call σ the "champion" for b.

We define an assignment $\varphi_B : B \to \Sigma_B$ that assigns good vertices b their champions, and other vertices b arbitrary values. The fraction of edges that φ_A, φ_B satisfy in the original instance is at least $\epsilon^2 D^2$. ◀

Moshkovitz then shows that small agreement soundness error translates to the list version. The proof is unchanged from [20] and is given in the appendix.

▶ **Lemma 3.6** (Lemma 4.7 of [20]). *Let $\ell \geq 1$, $0 < \epsilon' < 1$. A* LABEL-COVER *with agreement soundness error ϵ' has list-agreement soundness error $(\ell, \epsilon' \ell^2)$.*

Proof. Assume by the way of contradiction that the LABEL-COVER instance has an assignment $\hat{\varphi}_A : A \to \binom{\sigma_A}{\ell}$ such that on more than $\epsilon' \ell^2$-fraction of the B-vertices, the A-vertices do not totally disagree. Define an assignment $\varphi_A : A \to \Sigma_A$ by assigning every vertex $a \in A$ a symbol picked uniformly at random from the ℓ symbols in $\hat{\varphi}_A(a)$. If a vertex $b \in B$ has two neighbors $a_1, a_2 \in A$ that agree on b under the list assignment $\hat{\varphi}_A$, then the probability that they agree on b under the assignment φ_A is at least $1/\ell^2$. Thus, under φ_A, the expected fraction of the B-vertices that have at least two neighbors that agree on them, is more than ϵ'. In particular, there exists an assignment to the A-vertices, such that more than ϵ' fraction of the B-vertices have two neighbors that agree on them. This contradicts the agreement soundness. ◀

Lemma 3.2 follows directly from combining Lemmas 3.4 and 3.6.

4 Approximation Algorithm for Directed Steiner Tree

Recall that in DST, the input consists of a directed graph G with costs $c(e)$ on edges, a collection X of terminals, and a designated root $r \in V$. The goal is to find a subgraph of G that forms an arborescence $T_{opt} = T(r, X)$ rooted at r containing all the terminals and minimizing the cost $c(T(r, X)) = \sum_{e \in T(r,X)} c(e)$. Let $N = |X|$ denote the number of terminals and n the number of vertices.

Observe that one can model SET-COVER as a special case of DST on a 3-level acyclic digraph, with a universal root on top, nodes representing sets as internal layer, and the elements as leaves. The cost of an edge coming into a node corresponds to the cost of the corresponding element or set.

Our algorithm consists of "guessing" a set C of intermediate nodes of the optimal tree. After computing the optimal tree on top of this set, we use this set as the source of roots for a collection of trees to cover the terminals. This becomes a set cover problem, where we map each set selected to a tree of restricted size with a root in C. Our algorithm then reduces to applying the classic greedy set cover algorithm on this instance induced by the "right" set C. Because of the size restriction, the resulting approximation has a smaller constant factor.

We may assume each terminal is a leaf, by adding a leaf as a child of a non-leaf terminal and transfer the terminal function to that leaf. If a tree contains a terminal, we say that the tree *covers* the terminal.

Let $\ell(T)$ denote the number of terminals in a tree T. Let T_v denote the subtree of tree T rooted at node v. For node v and child w of v, let T_{vw} be the subtree of T formed by $T_v \cup \{vw\}$, i.e., consisting of the subtree of T rooted at v along with the edge to w's parent (v).

▶ **Definition 4.1.** *A set $C \subset V$ is a ϕ-*core *of a tree T if there is a collection of edge-disjoint subtrees $T_1, T_2, \ldots$ of T such that: a) the root of each tree T_i is in C, b) every terminal is contained in exactly one tree T_i, and c) each T_i contains at most ϕ terminals, $\ell(T_i) \leq \phi$.*

▶ **Lemma 4.2.** *Every tree T contains a ϕ-core of size at most $\lceil \ell(T)/\phi \rceil$, for any ϕ.*

Proof. The proof is by induction on the number of terminal in the tree. The root is a core when $\ell(T) \leq \phi$. Let v be a vertex with $\ell(v) \geq \phi$ but whose children fail that inequality. Let C' be a ϕ-core of $T' = T \setminus T_v$ promised by the induction hypothesis, and let $C = C' \cup \{v\}$.

For each child w of v, the subtree T_{vw} contains at most ϕ terminals. Together they cover uniquely the terminals in T_v, and satisfy the other requirements of the definition of a ϕ-core for T_v. Thus C is a ϕ-core for T. Since $\ell(T') \leq \ell(T) - \phi$, the size of C satisfies $|C| = |C'| + 1 \leq \lceil (\ell(T) - \phi)/\phi \rceil + 1 = \lceil \ell(T)/\phi \rceil$. ◀

A core implicitly suggests a set cover instance, with sets of size at most ϕ, formed by the terminals contained in each of the edge-disjoint subtrees. Our algorithm is essentially based on running a greedy set cover algorithm on that instance.

Let $R(v)$ denote the set of nodes reachable from v in G.

▶ **Definition 4.3.** *Let $S \subseteq V$, $U \subset X$, and let ϕ be a parameter. Then S induces a ϕ-*bounded Set Cover *instance $(U, \mathcal{C}^{\phi}_{S,U})$ with $\mathcal{C}^{\phi}_{S,U} = \{Y \subseteq U : |Y| \leq \phi$ and $\exists v \in S, Y \subseteq R(v)\}$. Namely, a subset Y of at most ϕ terminals in U is in $\mathcal{C}^{\phi}_{S,U}$ iff there is a v-rooted subtree containing Y.*

We relate set cover solutions of $\mathcal{C}_{C,X}$ to DST solutions of G with the following lemmas.

For sets F and S, let $T(S, F) = \min_{s \in S} T(s, F)$ be the optimal tree containing F with a node in S as root.

▶ **Lemma 4.4.** *Let $\mathcal{S}$ be a set cover of $\mathcal{C}^{\phi}_{S,U}$ of cost $c(\mathcal{S})$. Then, we can form a valid DST solution $T_{\mathcal{S}}$ by combining $T(r, S)$ with the trees $T(S, F)$, for each $F \in \mathcal{S}$. The cost of $T_{\mathcal{S}}$ is at most $c(T_{\mathcal{S}}) \leq c(T(r, S)) + c(\mathcal{S})$.*

Proof. $T_{\mathcal{S}}$ contains all the terminals since the sets in $\mathcal{S}$ cover X. It contains an r-rooted arborescence since $T(r, S)$ contains a path from r to all nodes in S, and the other subtrees contain a path to each terminal from some node in S. The cost bound follows from the definition of the weights of sets in $\mathcal{C}^{\phi}_{S,U}$. The actual cost could be less, if the trees share edges or have multiple paths, in which case some superfluous edges can be shed. ◀

Let $OPT_{SC}(\mathcal{C})$ be the optimal set cover of a set system $\mathcal{C}$.

▶ **Lemma 4.5.** *Let C be a ϕ-core of G. The cost of an optimal DST of G equals $OPT_{SC}(\mathcal{C}^{\phi}_{C,X})$ plus the cost of an optimal r-rooted tree with C as terminals: $c(T_{Opt}) = OPT_{SC}(\mathcal{C}^{\phi}_{C,X}) + c(T(r,C))$*

Proof. The subtree of T_{Opt} induced by C and the root r has cost $c(T(r,C))$. The rest of the tree consists of the subtrees T_{vw}, for each $v \in C$ and child w of v. T_{vw} contains at most ϕ terminals, so the corresponding set is contained in $\mathcal{C}^{\phi}_{C,X}$. Together, these subtrees contain all the terminals, so the corresponding set collection covers $\mathcal{C}^{\phi}_{C,X}$. Thus, $c(T_{Opt}) \geq c(T(r,C)) + OPT_{SC}(\mathcal{C}_{C,X})^{\phi}$. By Lemma 4.4, the inequality is tight. ◀

The *density* of a tree T' is defined as $c(T')/\ell(T')$.

Given a root and a fixed set S of nodes as leaves, an optimal cost tree $T(r,S)$ can be computed in time $poly(n)2^{|S|}$ by a (non-trivial) algorithm of Dreyfus and Wagner [10].

▶ **Lemma 4.6.** *A minimum density set in $\mathcal{C}^{\phi}_{S,X}$ can be found in time $n^{O(\max(\phi,N/\phi))}$.*

Proof. There are at most $2n^{\phi}$ subsets of at most ϕ terminals and at most N/ϕ choices for a root from the set C. Given a potential root r_0 and candidate core S, the algorithm of [10] computes $T(r_0,S)$ in time $poly(n)2^{2N/\phi}$. ◀

Our algorithm for DST is based on guessing the right ϕ-core C, and then computing a greedy set cover of $\mathcal{C}^{\phi}_{S,X}$ by repeatedly applying Lemma 4.6. More precisely, we try all possible subsets $S \subset V$ of size at most $2N/\phi$ as a ϕ-core (of T_{opt}) and for each such set do the following. Set U initially as X, representing the uncovered terminals. Find a min-density set Z of $\mathcal{C}^{\phi}_{S,U}$ and a corresponding optimal cost tree (with some root in S), remove Z from U and repeat until U is empty. We then compute $T(r,S)$ and combine it with all the computed subtrees into a single tree T_S. The solution output, T_{Alg}, is the T_S of smallest total cost, over all the candidate cores S.

▶ **Theorem 4.7.** *Let $\gamma \geq 1/2$ be a parameter, $\phi = N^{1-\gamma}$, and let C be a ϕ-core of T_{opt}. Then the greedy set cover algorithm applied to $\mathcal{C}^{\phi}_{C,X}$ yields a $1+\ln\phi$-approximation of DST. Namely, our algorithm is a $(1-\gamma)\ln n$-approximation of DST. The running time is $n^{O(max(\phi,N/\phi))} = exp(\tilde{O}(N^{\gamma}))$.*

Proof. Let Gr be the size of the greedy set cover of $\mathcal{C}^{\phi}_{C,X}$ and $O = OPT_{SC}(\mathcal{C}^{\phi}_{C,X})$. Let $t_c = c(T(r,C))$ be the cost of optimal tree with root r containing C. Since the cardinality of the largest set in $\mathcal{C}^{\phi}_{C,X}$ is at most ϕ, it follows by the analysis of Chvátal [5] that $Gr \leq (1+\ln\phi)Opt_{SC}(\mathcal{C}^{\phi}_{C,X})$. Thus,

$$c(T_{Alg}) \leq t_C + Gr \leq t_C + (1+\ln\phi)O \leq (1+\ln\phi)(t_C+O) = (1+\ln\phi)c(T_{Opt}) \; .$$

applying Lemma 4.5 in the first inequality and Lemma 4.4 in the last equality. Observe that $\ln\phi = (1-\gamma)\ln n$. For each candidate core S we find a min-density set at most n times. There are $\binom{n}{N/\phi} \leq n^{N/\phi}$ candidate cores and the cost for each is $n \cdot n^{O(\min(\phi,N/\phi))}$, by Lemma 4.6. Hence, the total cost is $n^{N/\phi} \cdot n^{O(\max(\phi,N/\phi))} = n^{O(N/\phi)} = exp(\tilde{O}(N^{\gamma}))$ using that $\phi = N^{1-\gamma} \leq N/\phi$. ◀

Now we observe that the same theorem applies to the Connected Polymatroid problem. Since the function is both submodular and *increasing*, for every collection of pairwise disjoint sets $\{S_i\}_{i=1^k}$, it holds that $\sum_{i=1}^{k} f(S_i) \geq f(\bigcup_{i=1}^{k} S_i)$. Thus, for a given $\gamma \geq 1/2$, at iteration

i there exists a collection S_i of terminals so that $f(S_i)/c(S_i) \geq f(U)/c(U)$. We can guess S_i in time $\exp(N^\gamma \cdot \log n)$ and its set of Steiner vertices X_i in time $O(3^{N^\gamma})$. Using the algorithm of [10], we can find a tree of density at most opt/N^γ. The rest of the proof is identical.

References

1 Nikhil Bansal, Parinya Chalermsook, Bundit Laekhanukit, Danupon Nanongkai, and Jesper Nederlof. New tools and connections for exponential-time approximation. *CoRR*, abs/1708.03515, 2017.

2 Chris Calabro, Russell Impagliazzo, and Ramamohan Paturi. A duality between clause width and clause density for SAT. In *CCC*, pages 252–260, 2006.

3 P Chalermsook, B. Laekhanukit, and D. Nanongkai. Independent set, induced matching, and pricing: Connections and tight (subexponential time) approximation hardnesses. In *FOCS*, 2013.

4 Moses Charikar, Chandra Chekuri, To-Yat Cheung, Zuo Dai, Ashish Goel, Sudipto Guha, and Ming Li. Approximation algorithms for directed Steiner problems. *J. Algorithms*, 33(1):73–91, 1999.

5 Vasek Chvátal. A greedy heuristic for the set-covering problem. *Math. Oper. Res.*, 4(3):233–235, 1979.

6 Marek Cygan, Guy Kortsarz, and Bundit Laekhanukit. On subexponential running times for approximating directed steiner tree and related problems. *CoRR*, abs/1811.00710, 2018.

7 Marek Cygan, Lukasz Kowalik, and Mateusz Wykurz. Exponential-time approximation of weighted set cover. *Inf. Process. Lett.*, 109(16):957–961, 2009.

8 Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3):12, 2007.

9 Irit Dinur and David Steurer. Analytical approach to parallel repetition. *CoRR*, abs/1305.1979, 2013.

10 S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1(3):195–207, 1971.

11 U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

12 Rohan Ghuge and Viswanath Nagarajan. Quasi-polynomial algorithms for submodular tree orienteering and other directed network design problems. In *SODA*, pages 1039–1048. SIAM, 2020.

13 Fabrizio Grandoni, Bundit Laekhanukit, and Shi Li. $O(log^2 k/loglogk)$-approximation algorithm for directed Steiner tree: A tight quasi-polynomial-time algorithm. In *STOC*, pages 253–264, 2019.

14 E. Halperin and R. Krauthgamer. Polylogarithmic inapproximability. In *STOC*, pages 585–594, 2003.

15 Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? In *FOCS*, pages 653–663, 1998.

16 David S. Johnson. Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.*, 9(3):256–278, 1974.

17 G. Kortsarz and D. Peleg. Approximating the weight of shallow Steiner trees. *Discrete Applied Math*, pages 265–285, 1999.

18 L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, pages 383–390, 1975.

19 Carsten Lund and Mihalis Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960–981, 1994.

20 Dana Moshkovitz. The projection games conjecture and the NP-hardness of $\ln n$-approximating set-cover. *Theory of Computing*, 11(7):221–235, 2015.

21 Dana Moshkovitz and Ran Raz. Sub-constant error probabilistically checkable proof of almost-linear size. *Computational Complexity*, 19(3):367–422, 2010.

22 Priyanka Mukhopadhyay. The projection games conjecture and the hardness of approximation of SSAT and related problems. *CoRR*, abs/1907.05548, 2019.
23 Moni Naor, Leonard J. Schulman, and Aravind Srinivasan. Splitters and near-optimal derandomization. In *FOCS*, pages 182–191, 1995.
24 Petr Slavík. A tight analysis of the greedy algorithm for set cover. In *STOC*, pages 435–441, 1996.
25 L. A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2:385–393, 1982.

A Proof of Lemma 3.3

We give here a proof of Lemma 3.3, nearly verbatim from [20], for completeness.

We use the following combinatorial construction of Naor, Schulman, and Srinivasan [23] (which as appears as Lemma 4.9 of [20]). They construct a universe together with partitions of it. Each partition covers the universe, but any cover that uses at most one set out of each partition, is necessarily large.

▶ **Lemma A.1** ([23]). *For natural numbers m, D and $0 < \alpha < 1$, for all $u \geq (D^{O(\log D)} \log m)^{1/\alpha}$, there is an explicit construction of a universe U of size u and partitions $\mathcal{P}_1, \ldots, \mathcal{P}_m$ of U into D sets that satisfy the following: there is no cover of U with $\ell = D \ln |U|(1-\alpha)$ sets $S_{i_1}, \ldots, S_{i_\ell}$, $1 \leq i_1 < \cdots < i_\ell \leq m$, such that set S_{i_j} belongs to partition $\mathcal{P}_{i_j}$.*

The following reduction follows Moshkovitz [20], which in turns is along the lines of Feige [11].

Lemma 3.3 (restated). *Let $\mathcal{G} = (G = (A, B, E), \Sigma_A, \Sigma_B, \Pi)$ be a bi-regular* Label-Cover *instance with B-degree D. For every $0 < \alpha < 1$, and every $u \geq (D^{c_0 \log D} \log |\Sigma_B|)^{1/\alpha}$ for a certain absolute constant c_0, there is a reduction from $\mathcal{G}$ to a* Set-Cover *instance $\mathcal{SC}_{\mathcal{G}}$ with the following properties:*

1. Completeness: *If all edges of G can be covered, then $\mathcal{SC}_{\mathcal{G}}$ has a set cover of size $|A|$.*
2. Soundness: *If G has agreement soundness error (ℓ, α), where $\ell = D \ln u(1-\alpha)$, then every set cover of $\mathcal{SC}_{\mathcal{G}}$ is of size more than $|A| \ln u(1-2\alpha)$.*
3. *The number n_{SC} of elements of $\mathcal{SC}_{\mathcal{G}}$ is $|B| \cdot u$ and the number of sets is $|A| \cdot |\Sigma_A|$.*

The time for the reduction is polynomial in $|A|, |B|, |\Sigma_A|, |\Sigma_B|$ and u.

Proof. Let α and u be values satisfying the statement of the theorem. Let $m = |\Sigma_B|$ and let D be the B-degree of $\mathcal{G}$. Apply Lemma A.1 with m, D and u, obtaining a universe U of size u and partitions $\mathcal{P}_{\sigma_1}, \ldots, \mathcal{P}_{\sigma_m}$ of U. We index the partitions by the symbols $\sigma_1, \ldots, \sigma_m$ of Σ_B. The elements of the Set-Cover instances are $B \times U$. Equivalently, each vertex $b \in B$ has a copy of the universe U. Covering this universe corresponds to satisfying the edges that touch b. There are m ways to satisfy the edges that touch b — one for every possible assignment $\sigma \in \Sigma_B$ to b. The different partitions covering u correspond to those different assignments.

For every vertex $a \in A$ and an assignment $\sigma \in \Sigma_A$ to a we have a set $S_{a,\sigma}$ in the Set-Cover instance. Taking $S_{a,\sigma}$ to the cover corresponds to assigning σ to a. Notive that a cover might consist of several sets of the form $S_{a,\cdot}$ for the same $a \in A$, which is the reason we consider list agreement. The set $S_{a,\sigma}$ is a union of subsets, one for every edge $e = (a, b)$ touching a. If e is the i-th edge coming into b ($1 \leq i \leq D$), then the subset associated with e is $\{b\} \times S$, where S is the i-th subset of the partition $P_{\phi_e(\sigma)}$.

Completeness follows from taking the set cover corresponding to each of the A-vertices and its satisfying assignments.

To prove soundness, assume by contradiction that there is a set cover C of $\mathcal{SC}_{\mathcal{G}}$ of size at most $|A| \ln |U|(1-2\alpha)$. For every $a \in A$, let s_a be the number of sets in C of the form $S_{a,\cdot}$. Hence, $\sigma_{a \in A} s_a = |C|$. For every $b \in B$, let s_b be the number of sets in C that participate in covering $\{b\} \times U$. Then, denoting the A-degree of G by D_A,

$$\sum_{b \in B} s_b = \sum_{a \in A} s_a D_A \leq D_A |A| \ln |U|(1-2\alpha) = D|B| \ln |U|(1-2\alpha) \ .$$

In other words, on average over the $b \in B$, the universe $\{b\} \times U$ is covered by at most $D \ln |U|(1-2\alpha)$ sets. Therefore, by Markov's inequality, the fraction of $b \in B$ whose universe $\{b\} \times U$ is covered by at most $D \ln |U|(1-\alpha) = \ell$ sets is at least α. By the contrapositive of Lemma A.1 and our construction, for such $b \in B$, there are two edges $e_1 = (a_1, b), e_2 = (a_2, b) \in E$ with $S_{a_1,\sigma_1}, S_{a_2,\sigma_2} \in C$ where $\pi_{e_1}(\sigma_1) = \pi_{e_2}(\sigma_2)$.

We define assignment $\hat{\varphi}_A : A \to \binom{\sigma_A}{\ell}$ to the A-vertices as follows. For every $a \in A$, pick ℓ different symbols $\sigma \in \Sigma_A$ from those with $S_{a,\sigma} \in C$ (add arbitrary symbols if there are not enough). As we showed, for at least α-fraction of the $b \in B$, the A-vertices will not totally disagree. Hence, the soundness property follows. ◀