Hardness of Set Cover with Intersection 1

V.S.Anil Kumar[1], Sunil Arya[2] and H.Ramesh[3]

[1] MPI für Informatik, Saarbrücken. kumar@mpi-sb.mpg.de
[2] Department of Computer Science, Hong Kong University of Science and Technology. arya@cs.ust.hk
[3] Department of Computer Science and Automation, Indian Institute of Science, Bangalore. ramesh@csa.iisc.ernet.in

Abstract. We consider a restricted version of the general Set Covering problem in which each set in the given set system intersects with any other set in at most 1 element. We show that the Set Covering problem with intersection 1 cannot be approximated within a $o(\log n)$ factor in random polynomial time unless $NP \subseteq ZTIME(n^{O(\log\log n)})$. We also observe that the main challenge in derandomizing this reduction lies in find a hitting set for large volume combinatorial rectangles satisfying certain intersection properties. These properties are not satisfied by current methods of hitting set construction.
An example of a Set Covering problem with the intersection 1 property is the problem of covering a given set of points in two or higher dimensions using straight lines; any two straight lines intersect in at most one point. The best approximation algorithm currently known for this problem has an approximation factor of $\theta(\log n)$, and beating this bound seems hard. We observe that this problem is Max-SNP-Hard.

1 Introduction

The general Set Covering problem requires covering a given base set B of size n using the fewest number of sets from a given collection of subsets of B. This is a classical NP-Complete problem and its instances arise in numerous diverse settings. Thus approximation algorithms which run in polynomial time are of interest.

Johnson[Jo74] showed that the greedy algorithm for Set Cover gives an $O(\log n)$ approximation factor. Much later, following advances in Probabilistically Checkable Proofs [ALMSS92], Lund and Yannakakis [LY93] and Bellare et al. [BGLR93] showed that there exists a positive constant c such that the Set Covering problem cannot be approximated in polynomial time within a $c \log n$ factor unless $NP \subseteq DTIME(n^{O(\log\log n)})$. Feige [F98] improved the approximation threshold to $(1 - o(1)) \log n$, under the same assumption. Raz and Safra[RS97] and Arora and Sudan[AS97] then obtained improved Probabilistically Checkable Proof Systems with sub-constant error probability; their work implied that the Set Covering problem cannot be approximated within a $c \log n$ approximation factor (for some constant c) unless $NP = P$.

Note that all the above hardness results are for general instances of the Set Covering problem and do not hold for instances when the intersection of any pair of sets in the given collection is guaranteed to be at most 1. Our motivation for considering this restriction to intersection 1 arose from the following geometric instance of the Set Covering problem.

Given a collection of points and lines in a plane, consider the problem of covering the points with as few lines as possible. Megiddo and Tamir[MT82] showed that this problem is NP-Hard. Hassin and Megiddo[HM91] showed NP-Hardness even when the lines are axis-parallel but in 3D. The best approximation factor known for this problem is $\Theta(\log n)$. Improving this factor seems to be hard, and this motivated our study of inapproximability for Set Covering with intersection 1. Note that any two lines intersect in at most 1 point.

The problem of covering points with lines was in turn motivated by the problem of covering a rectilinear polygon with holes using rectangles [Le87]. This problem has applications in printing integrated circuits and image compression[CIK88]. This problem is known to be Max-SNP-Hard even when the rectangles are constrained to be axis-parallel. For this case, an $O(\sqrt{\log n})$-factor approximation algorithm was obtained recently by Anil Kumar and Ramesh[AR99]. However, this algorithm does not extend to the case when the rectangles need not be axis-parallel. Getting a $o(\log n)$-factor approximation algorithm for this case seems to require solving the problem of covering points with arbitrary lines, though we are not sure of the exact nature of this relationship.

Our Result. We show that there exists a constant $c > 0$ such that approximating the Set Covering problem with intersection 1 to within a factor of $c \log n$ in random polynomial time is possible only if $NP \subseteq ZTIME(n^{O(\log \log n)})$ (where $ZTIME(t)$ denotes the class of languages that have a probabilistic algorithm running in expected time t with zero error). We also give a sub-exponential derandomization which shows that approximating the Set Covering problem with intersection 1 to within a factor of $c\frac{\log n}{\log \log n}$ in deterministic polynomial time is possible only if $NP \subseteq DTIME(2^{n^{1-\epsilon}})$, where ϵ is any positive constant less than $\frac{1}{2}$.

The starting point for our result above is the Lund-Yannakakis hardness proof[LY93] for the general Set Covering problem. This proof uses an auxiliary set system with certain properties. We show that this auxiliary set system necessarily leads to large intersection. We then replace this auxiliary set system by another carefully chosen set system with additional properties and modify the reduction appropriately to ensure that intersection sizes stay small. The key features of the new set system are partitions of the base set into several sets of smaller size (instead of just 2 sets as in the case of the Lund-Yannakakis system or a constant number of sets as in Feige's system; small sets will lead to small intersection) and several such partitions (so that sets which "access" the same partition in the Lund-Yannakakis system and therefore have large intersection now "access" distinct partitions).

We then show how the new set system above can be constructed in randomized polynomial time and also how this randomized algorithm can be deran-

domized using conditional probabilities and appropriate estimators in $O(2^{n^{1-\epsilon}})$ time, where ϵ is a positive constant. This leads to the two conditions above, namely, $NP \subseteq DTIME(2^{n^{1-\epsilon}})$ (but for a hardness of $O(\frac{\log n}{\log \log n})$) and $NP \subseteq ZTIME(n^{O(\log \log n)})$. A deterministic polynomial time construction of our new set system will lead to the quasi-NP-Hardness of approximating the Set Covering problem with intersection 1 to within a factor of $c \log n$, for some constant $c > 0$.

While the Lund-Yannakakis set system can be constructed in deterministic polynomial time using ϵ-biased limited independence sample spaces, this does not seem to be true of our set system. One of the main bottlenecks in constructing our set system in deterministic polynomial time is the task of obtaining a polynomial size hitting set for *Combinatorial Rectangles*, with the hitting set satisfying additional properties. One of these properties (the most important one) is the following: if a hitting set point has the elements i, j among its coordinates, then no other hitting set point can have both i, j among its coordinates. The only known construction of a polynomial size hitting set for combinatorial rectangles is by Linial, Luby, Saks, and Zuckerman [LL+93] and is based on enumerating walks in a constant degree expander graph. As we show in this paper, the hitting set obtained by [LL+93] does not satisfy the above property for reasons that seem intrinsic to the use of constant degree expander graphs.

We also note that if the proof systems for NP obtained by Raz and Safra[RS97] or Arora and Sudan[AS97] have an additional property then the condition $NP \subseteq ZTIME(n^{O(\log \log n)})$ can be improved to $NP = ZPP$. Similarly, the statement that approximating the Set Covering problem with intersection 1 to within a factor of $c\frac{\log n}{\log \log n}$ in deterministic polynomial time is possible only if $NP \subseteq DTIME(2^{n^{1-\epsilon}})$ can be strengthened to approximation factor $c \log n$ instead of $c\frac{\log n}{\log \log n}$. The property needed of the proof systems is that the *degree*, i.e., the total number of random choices of the verifier for which a particular question is asked of a particular prover, be $O(n^{\delta})$, for some small enough constant value δ. Currently, we are exploring whether this condition can be satisfied by the above proof systems. The degree influences the number of partitions in our auxiliary proof system and therefore needs to be small.

The above proof of hardness for Set Covering with intersection 1 does not apply to the problem of covering points with lines, the original problem which motivated this paper; however, it does indicate that algorithms based on set cardinalities and small pairwise intersection alone are unlikely to give a $o(\log n)$ approximation factor for this problem.

Further, our result shows that constant VC-dimension alone does not help in getting a $o(\log n)$ approximation for the Set Covering problem. This is to be contrasted with the result of Brönnimann and Goodrich[BG94] which shows that if the VC-dimension is a constant and an $O(\frac{1}{\epsilon})$ sized (weighted) ϵ-net can be constructed in polynomial time, then a constant factor approximation can be obtained.

Finally, for the problem of covering points with lines, we observe that the NP-Hardness proof of Megiddo and Tamir[MT82] can be easily extended to a

Max-SNP-Hardness proof. We also show that the obvious linear program for this problem must have an integrality gap of 2. In addition, we give an example which could possibly show an integrality gap of $\Theta(\log n)$; however, we have been unable to prove such a gap. We believe that a good understanding of this example would reveal whether or not the linear program lower bound is strong enough and if not, what other lower bounds one could use.

The paper is organized as follows. Section 2 will give an overview of the Lund-Yannakakis reduction. Section 3 shows why the Lund-Yannakakis proof does not show hardness of Set Covering when the intersection is constrained to be 1. Section 4 describes the reduction to Set Covering with intersection 1. This section describes a new set system we need to obtain in order to perform the reduction. Section 5 will sketch the randomized construction of this set system. Section 6 sketches the sub-exponential time derandomization. Section 7 describes the connection to hitting combinatorial rectangles required to construct the above set system in polynomial time. Section 8 gives a sketch of the Max-SNP-Hardness proof for covering points with lines and shows an example which may have a large integrality gap. Section 9 enumerates several interesting open problems which arise from this paper. Section 13 in the Appendix shows how the condition $NP \subseteq ZTIME(n^{O(\log\log n)})$ can be improved to $NP = ZPP$ if the Raz-Safra[RS97] or the Arora-Sudan[AS97] proof system has a certain property.

2 Preliminaries: The Lund-Yannakakis Reduction

In this section, we sketch the version of the Lund-Yannakakis reduction described by Arora and Lund [AL95]. The reduction starts with a 2-Prover 1-Round proof system for Max-3SAT(5) which has inverse polylogarithmic error probability, uses $O(\log n \log\log n)$ randomness, and has $O(\log\log n)$ answer size. Here n is the size of the Max-3SAT(5) formula $\mathcal{F}$. Arora and Lund[AL95] abstract this proof system into the following *Label Cover* problem.

The Label Cover Problem. A bipartite graph G having $n' + n'$ vertices and edge set E is given, where $n' = n^{O(\log\log n)}$. All vertices have the same degree deg, which is polylogarithmic in n. For each edge $e \in E$, a partial function $f_e : [d] \to [d']$ is also given, where $d \geq d'$, and d, d' are polylogarithmic in n. The aim is to assign to each vertex on the left, a label in the range $1 \ldots d$, and to each vertex on the right, a label in the range $1 \ldots d'$, so as to maximize the number of edges $e = (u, v)$ satisfying $f_e(label(u)) = label(v)$. Edge $e = (u, v)$ is said to be *satisfied* by a labelling if the labelling satisfies $f_e(label(u)) = label(v)$.

The 2-Prover 1-Round proof system mentioned above ensures that either all the edges in G are satisfied by some labelling or that no labelling satisfies more than a $\frac{1}{\log^3 n}$ fraction of the edges, depending upon whether or not the Max-3SAT(5) formula $\mathcal{F}$ is satisfiable. Next, in time polynomial in the size of G, an instance $\mathcal{SC}$ of the Set Covering problem is obtained from this Label Cover problem $\mathcal{LC}$ with the following properties: if there exists a labelling satisfying all edges in G then there is a set cover of size $2n'$, and if no labelling satisfies

more than a $\frac{1}{\log^3 n}$ fraction of the edges then the smallest set cover has size $\Omega(2n' \log n')$. The base set in $\mathcal{SC}$ will have size polynomial in n'. It follows that the Set Covering problem cannot be approximated to a logarithmic factor of the base set size unless $NP \subseteq DTIME(n^{O(\log \log n)})$.

Improving this condition to $NP = P$ requires using a stronger multi-prover proof system [RS97,AS97] which has a constant number of provers (more than 2), $O(\log n)$ randomness, $O(\log \log n)$ answer sizes, and inverse polylogarithmic error probability. The reduction from such a proof system to the Set Covering problem is similar to the reduction from the Label Cover to the Set Covering problem mentioned above, with a modification needed to handle more than 2 provers (this modification is described in [BGLR93]).

In this abstract, we will only describe the reduction from Label Cover to the Set Covering problem and show how we can modify this reduction to hold for the case of intersection 1. This will show that Set Covering problem with intersection 1 cannot be approximated to a logarithmic factor unless $NP \subseteq ZTIME(n^{O(\log \log n)})$. The multi-prover proof system of the previous paragraph with an additional condition can strengthen the latter condition to $NP = ZPP$; this is described in the appendix.

We now briefly sketch the reduction from an instance $\mathcal{LC}$ of Label Cover to an instance $\mathcal{SC}$ of the Set Covering problem.

2.1 Label Cover to Set Cover

The following auxiliary set system given by a base set $N = \{1 \ldots n'\}$ and its partitions is needed.

The Auxiliary System of Partitions. Consider d' distinct partitions of N into two sets each, with the partitions satisfying the following property: if at most $\frac{\log n'}{2}$ sets in all are chosen from the various partitions with no two sets coming from the same partition, then the union of these sets does not cover N. Partitions with the above properties can be constructed deterministically in polynomial time [AGHP92,NSS95]. Let P_i^1, P_i^2 respectively denote the first and second sets in the ith partition. We describe the construction of $\mathcal{SC}$ next.

Using P_i^js to construct $\mathcal{SC}$. The base set B for $\mathcal{SC}$ is defined to be $\{(e,i) | e \in E, 1 \leq i \leq n'\}$. The collection C of subsets of B contains a set $C(v,a)$, for each vertex v and each possible label a with which v can be labelled. If v is a vertex on the left, then for each a, $1 \leq a \leq d$, $C(v,a)$ is defined as $\{(e,i) | e \text{ incident on } v \wedge i \in P^1_{f_e(a)}\}$. And if v is a vertex on the right, then for each a, $1 \leq a \leq d'$, $C(v,a)$ is defined as $\{(e,i) | e \text{ incident on } v \wedge i \in P^2_a\}$.

That $\mathcal{SC}$ satisfies the required conditions can be seen from the following facts.

1. If there exists a vertex labelling which satisfies all the edges, then B can be covered by just the sets $C(v,a)$ where a is the label given to v. Thus the size of the optimum cover is $2n'$ in this case.
2. If the total number of sets in the optimum set cover is at most some suitable constant times $n' \log n'$, then at least a constant fraction of the edges $e =$

(u, v) have the property that the number of sets of the form $C(u, *)$ plus the number of sets of the form $C(v, *)$ in the optimum set cover is at most $\frac{\log n'}{2}$. Then, for each such edge e, there must exist a label a such that $C(u, a)$ and $C(v, f_e(a))$ are both in this optimum cover. It can be easily seen that choosing a label uniformly at random from these sets for each vertex implies that there exists a labelling of the vertices which satisfies an $\Omega(\frac{1}{\log^2 n'}) \geq \frac{1}{\log^3 n}$ fraction of the edges.

3 $\mathcal{SC}$ has Large Intersection

There are two reasons why sets in the collection C in $\mathcal{SC}$ have large intersections.

Parts in the Partitions are Large. The first and obvious reason is that the sets in each partition in the auxiliary system of partitions are large and could have size $\frac{n'}{2}$; therefore, two sets in distinct partitions could have $\Omega(n')$ intersection. This could lead to sets $C(v, a)$ and $C(v, b)$ having $\Omega(n')$ common elements of the form (e, i), for some e incident on v.

Clearly, the solution to this problem is to work with an auxiliary system of partitions where each partition is a partition into not just 2 large sets, but into several small sets. The problem remains if we form only a constant number of parts, as in [F98]. We choose to partition into $(n')^{1-\epsilon}$ sets, where ϵ is some non-zero constant to be fixed later. This ensures that each set in each partition has size $\theta((n')^{\epsilon}$ polylog$(n))$ and that any two such sets have $O(1)$ intersection. However, smaller set size leads to other problems which we shall describe shortly.

Functions $f_e()$ are not 1-1. Suppose we work with smaller set sizes as above. Then consider the sets $C(v, a)$ and $C(v, b)$, where v is a vertex on the left and a, b are labels with the following property: for some edge e incident on v, $f_e(a) = f_e(b)$. Then each element $(e, *)$ which appears in $C(v, a)$ will also appear in $C(v, b)$, leading to an intersection size of up to $\Omega((n')^{\epsilon} * deg)$, where deg is the degree of v in G. This is a more serious problem. Our solution to this problem is to ensure that sets $C(v, a)$ and $C(v, b)$ are constructed using distinct partitions in the auxiliary system of partitions.

Next, we describe how to modify the auxiliary system of partitions and the construction of $\mathcal{SC}$ in accordance with the above.

4 $\mathcal{LC}$ to $\mathcal{SC}$ with Intersection 1

Our new auxiliary system of partitions $\mathcal{P}$ will have $d' * (deg + 1) * d$ partitions, where deg is the degree of any vertex in G. Each partition has $m = (n')^{1-\epsilon}$ parts, for some $\epsilon > 0$ to be determined. These partitions are organized into d' *groups*, each containing $(deg + 1) * d$ partitions. Each group is further organized into $deg + 1$ *subgroups*, each containing d partitions. The first $m/2$ sets in each partition comprise its *left half* and the last $m/2$ its *right half*.

Let $P_{g,s,p}$ denote the pth partition in the sth subgroup of the gth group and let $P_{g,s,p,k}$ denote the kth set (i.e., part) in this partition. Let B_k denote the set

$\cup_{g,s,p} P_{g,s,p,k}$ if $1 \leq k \leq m/2$, and the set $\cup_{g,s} P_{g,s,1,k}$, if $m/2 < k \leq m$. We also refer to B_k as the kth *column* of $\mathcal{P}$.

We need the following properties to be satisfied by the system of partitions $\mathcal{P}$.

1. The right sides of all partitions within a subgroup are identical, i.e., $P_{g,s,p,k} = P_{g,s,1,k}$, for every $k > m/2$.
2. $P(g,s,p,k) \cap P(g',s',p',k) = \phi$ unless either $g = g', s = s', p = p'$, or, $k > m/2$ and $g = g', s = s'$. In other words, no element appears twice within a column, modulo the fact that the right sides of partitions within a subgroup are identical.
3. $|B_k \cap B_{k'}| \leq 1$ for all k, k', $1 \leq k, k' \leq m$, $k \neq k'$.
4. Suppose N is covered using at most $\beta m \log n'$ sets in all, disallowing sets on the right sides of those partitions which are not the first in their respective subgroups. Then there must be a partition in some subgroup s such that the number of sets chosen from the left side of this partition plus the number of sets chosen from right side of the first partition in s together sum to at least $\frac{3}{4}m$.

ϵ and β are constants which will be fixed later. Let $A_{p,k} = \cup_{g,s} P_{g,s,p,k}$, for each p, k, $1 \leq p \leq d, 1 \leq k \leq m/2$. Let $D_{g,k} = \cup_s P_{g,s,1,k}$, for each g, k, $1 \leq g \leq d'$, $m/2 + 1 \leq k \leq m$. Property 2 above implies that:

5. $|A_{p,k} \cap A_{p',k}| = 0$ for all $p \neq p'$, where $1 \leq p, p' \leq d$ and $k \leq m/2$.
6. $|D_{g,k} \cap D_{g',k}| = 0$ for all $g \neq g'$, where $1 \leq g, g' \leq d'$ and $k > m/2$.

We will describe how to obtain a system of partitions $\mathcal{P}$ satisfying these properties in Section 5, Section 6, and Section 7. First, we show how a set system $\mathcal{SC}$ with intersection 1 can be constructed using $\mathcal{P}$.

4.1 Using $\mathcal{P}$ to construct $\mathcal{SC}$

The base set B for $\mathcal{SC}$ is defined to be $\{(e,i) | e \in E, 1 \leq i \leq n'\}$ as before. This set has size $(n')^2 * deg = O((n')^2 \text{ polylog}(n))$.

The collection C of subsets of B contains $m/2$ sets $C_1(v,a) \ldots C_{m/2}(v,a)$, for each vertex v on the left (in graph G) and each possible label a with which v can be labelled. In addition, it contains $m/2$ sets $C_{m/2+1}(v,a) \ldots C_m(v,a)$, for each vertex v on the right in G and each possible label a with which v can be labelled. These sets are defined as follows.

Let E_v denote the set of edges incident on v in G. We edge-colour G using $deg + 1$ colours. Let $col(e)$ be the colour given to edge e in this edge colouring. For a vertex v on the left side, and any number k between 1 and $m/2$, $C_k(v,a) = \cup_{e \in E_v} \{(e,i) | i \in P_{f_e(a), col(e), a, k}\}$. For a vertex v on the right side, and any number k between $m/2 + 1$ and m, $C_k(v,a) = \cup_{e \in E_v} \{(e,i) | i \in P_{a, col(e), 1, k}\}$.

We now give the following lemmas which state that the set system $\mathcal{SC}$ has intersection 1 and that it has a set cover of small size if and only if there exists a way to label the vertices of G satisfying several edges simultaneously. The proofs are deferred to Section 10 in the Appendix.

Lemma 1. *The intersection of any two distinct sets $C_k(v,a)$ and $C_{k'}(w,b)$ is at most 1.*

Lemma 2. *If there exists a way of labelling vertices of G satisfying all its edges then there exists a collection of $n'm$ sets in C which covers B.*

Lemma 3. *If the smallest collection C' of sets in C covering the base set B has size at most $\frac{\beta}{2}n'm\log n'$ then there exists a labelling of G which satisfies at least a $\frac{1}{32\beta^2\log^2 n'}$ fraction of the edges. Recall that β was defined in Property 4 of $\mathcal{P}$.*

Corollary 1. *Set Cover with intersection 1 cannot be approximated within a factor of $\frac{\beta\log n'}{2}$ in random polynomial time, for some constant β, $0<\beta\leq\frac{1}{6}$, unless $NP\subseteq ZTIME(n^{O(\log\log n)})$. Further, if the auxiliary system of partitions $\mathcal{P}$ can be constructed in deterministic polynomial (in n') time, then approximating to within a $\frac{\beta\log n'}{2}$ factor is possible only if $NP=DTIME(n^{O(\log\log n)})$.*

5 Randomized Construction of the Auxiliary System $\mathcal{P}$

The obvious randomized construction is the following. Ignore the division into groups and just view $\mathcal{P}$ as a collection of subgroups. For each partition which is the first in its subgroup, throw each element i independently and uniformly at random into one of the m sets in that partition. For each partition P which is not the first in its subgroup, throw each element i which is not present in any of the sets on the right side of the first partition Q in this subgroup, into one of the first $m/2$ sets in P. Property 1 is thus satisfied directly. We need to show that Properties 2,3,4 are together satisfied with non-zero probability.

Property 4 can be shown without much trouble. Slightly weak versions of Properties 2 and 3 (intersection bounds of 2 instead of 1) also follow immediately. This can be improved to 1 using the Lovasz Local Lemma, but this does not give a constant success probability and also leads to problems in derandomization. The details of these calculations appear in the Appendix in Section 11.

To obtain a high probability of success, we need to change the randomized construction above to respect the following additional restriction (we call this Property 7): each set $P_{g,s,p,k}$ has size at most $\frac{d'*(deg+1)*dn'}{m}$, for all g,s,p,k, $1\leq g\leq d', 1\leq s\leq deg+1, 1\leq p\leq d, 1\leq k\leq m$.

The new randomized construction proceeds as in the previous random experiment, fixing partitions in the same order as before, except that any choice of throwing an element $i\in N$ which violates Properties 2,3,7 is precluded. Property 7 enables us to show that not too many choices are precluded for each element, and therefore, this experiment stays close in behaviour to the previous one, except that Properties 2,3,7 are all automatically satisfied. The details of this new construction appear in Section 11.1 in the appendix.

6 Derandomization in $O(2^{n^{1-\epsilon}})$ Time

The main hurdle in derandomizing the above randomized construction in polynomial time is Property 4. There could be up to $O(2^{m\times polylog(n)})=O(2^{(n')^{1-\epsilon'}})$

ways of choosing $\beta m \log n'$ sets from the various partitions in $\mathcal{P}$ for a constant ϵ' slightly smaller than ϵ, and we need that each of these choices fails to cover N for Property 4 to be satisfied.

For the Lund-Yannakakis system of partitions described in Section 2.1, each partition was into 2 sets and the corresponding property could be obtained deterministically using small-bias $\log n$-wise independent sample space constructions. This is no longer true in our case. Feige's [F98] system of partitions, where each partition is into several but still a constant number of parts, can be obtained deterministically using anti-universal sets [NSS95]. However, it is not clear how to apply either Feige's modified proof system or his system of partitions to get intersection 1.

We show in Section 7 that enforcing Property 4 in polynomial time corresponds to hitting combinatorial rectangles with certain restricted kinds of sets. In this paper, we take the slower approach of using Conditional Probabilities and enforcing Property 4 by checking each of the above choices explicitly. However, note that the number of choices is superexponential in n (even though it is sub-exponential in n'). To obtain a derandomization which is sub-exponential in n, we make the following change in $\mathcal{P}$: the base set is taken to be of size n instead of n'. We use an appropriate pessimistic estimator and conditional probabilities to construct $\mathcal{P}$ with parameter n instead of n' (details are given in Section 12 in the Appendix). This will give a gap of $\Theta(\log n)$ (instead of $\Theta(\log n')$) in the set cover instance $\mathcal{SC}$). But since the base set size in $\mathcal{SC}$ is now $O((n' * n)$ polylog$(n))$, we get a hardness of only $\Theta(\log n) = \Theta(\frac{\log n'}{\log \log n'})$ (note that the approximation factor must be with respect to the base set size) unless $NP \subset DTIME(2^{n^{1-\epsilon}})$, for any constant ϵ such that $22\beta < \epsilon < 1/2$.

7 Connection to Hitting Combinatorial Rectangles

First, consider the simpler problem of constructing a system of $d' * (deg + 1) * d$ partitions of $N = \{1 \ldots n'\}$ with the following properties. Each partition has $m = (n')^{1-\epsilon}$ parts. No collection of $\beta m \log n'$ parts from different partitions on the whole should be able to cover N, unless some partition contributes more than $3m/4$ sets. This problem is shown to be equivalent to the problem of hitting combinatorial rectangles as follows.

A combinatorial rectangle is a set $R = R_1 \times R_2 \times \ldots \times R_{d'*(deg+1)*d}$, where $R_i \subseteq [m] = \{1 \ldots m\}$. The volume of R, $vol(R)$, is defined to be $\Pi_k \frac{|R_k|}{m}$. A *hitting set* H is a subset of $[m]^{d'*(deg+1)*d}$ which intersects all *large* rectangles R, i.e., those with volume at least $\frac{1}{4^{\frac{4}{3}\beta \log n'}}$ and for which each R_i has size at least $m/4$.

The desired system of partitions can be obtained using the above hitting set H of size $O(m^{1+\epsilon})$ as follows. Let $H = \{H_1, \ldots, H_n\}$. Let $H_x(i)$ denote the element in the ith coordinate of H_x. The partitions are defined as follows: for each partition i, element $x \in N$ lies in the position $H_x(i)$. That these partitions indeed have the properties described in the first paragraph of this section can

be seen as follows. Consider any collection C of at most $3m/4$ sets from each partition and comprising $\beta m \log n'$ sets on the whole. Let R_i denote the collection of those sets from the ith partition which are not in C. Then C has an associated combinatorial rectangle $R(C)$ given by $R_1 \times R_2 \times \ldots \times R_{d'*(deg+1)*d}$. Each R_i has cardinality at least $m/4$ and the volume of $R(C)$ is at least $\frac{1}{4^{\frac{4}{3}\beta \log n'}}$. Since H hits $R(C)$, there exists an element in N which is not covered by C.

Thus a small hitting set construction for combinatorial rectangles also gives a auxiliary set system with the properties described in the first paragraph of this section. Our problem requires constructing a similar system of partitions but with additional properties, namely Properties 1–4. These properties place the following demands of the hitting set. Property 1 requires each hitting set point to have identical entries in coordinates corresponding to a subgroup, if it has a value more than $m/2$ in the coordinate corresponding to the first partition of the subgroup. Property 2 requires that entries in any hitting set point do not repeat, modulo Property 1. Property 3 requires that no two distinct elements in N are both present among the coordinates in two distinct hitting set points. Property 4 is actually the hitting property itself. But for Property 1 of our set system, we require to hit all large volume rectangles. Property 1 places further restrictions on the nature of the hitting set and also the rectangles to be hit.

The only algorithm known for constructing hitting sets for combinatorial rectangles is due to Linial, Luby, Saks, and Zuckerman [LL+93]. But the hitting set it gives does not satisfy Properties 1–3. Property 3, which is probably the most important of the three properties, does not seem to be satisfied for reasons intrinsic to the algorithm, as described below.

In the above algorithm, the hitting set corresponds to taking all possible walks of length $\Theta(\log m)$ in a constant degree expander graph with m vertices, when $d' * (deg + 1) * d = O(\log m)$. The total number of walks is $O(m^{1+\epsilon})$. There are $\Omega(m^{\epsilon})$ walks starting at any given vertex, and they have to pass through the $O(1)$ neighbours of this vertex. Therefore, there must be $\Omega(m^{\epsilon})$ walks passing through the same pair of vertices. It follows that there could be m^{ϵ} elements in this hitting set, all having the values k and k' in some two consecutive coordinates $i, i+1$, which is a violation of Property 3. Thus the use of a constant degree expander, while facilitating the hitting property, seems to be a fundamental obstruction for Property 3. Further, when $d' * (deg + 1) * d$ is not $O(\log m)$, a dimension reduction procedure is needed, which also leads to several elements being repeated within each hitting set point, violating Property 2.

8 Covering Points with Lines

Max-SNP Hardness. We observe that the NP-Hardness reduction of Megiddo and Tamir[MT82] from 3SAT also gives a Max-SNP hardness proof for this problem, if we start with MAX-3SAT(5) instead of 3SAT. We give a brief sketch of this proof.

For each variable x, there is a 5 by 5 grid with 5 horizontal lines and 5 vertical lines (finally this grid will be oriented arbitrarily in 2D). Choosing the horizontal

lines corresponds to setting x to 1 and choosing the vertical lines corresponds to setting x to 0. Note that either all horizontal lines or all vertical lines have to be chosen to cover the 25 grid points. Next, for each clause, there is a point having 3 lines passing through it. These three lines are chosen from the grid lines in grids associated with the three variables in this clause, one line per variable. This can be done in such a way that any satisfying assignment to the variables will choose 5 lines per variable to cover the variable grids and these lines will also cover all clause points. Further, if no assignment satisfies more than a constant fraction of the clauses, then at least $\Omega(|C|)$ lines in addition to the 5 lines per variable will be needed to cover all points, where $|C|$ is the number of clauses (which is at least a constant fraction of the number of variables). This gives a multiplicative constant gap.

Integrality Gap. The following example shows an integrality gap of 2 for the obvious linear program. Take a collection of points in general position, consider all possible lines defined by pairs of these points, and take the dual of this arrangement. Each point has 2 lines through it in the dual; therefore the linear program optimum equals half the number of lines. But the integer optimum must choose all but one of the lines. This gives a gap of 2.

The following family of examples may give an $\Theta(\log n)$ integrality gap, but we have been unable to obtain a proof to this effect. Consider an $n * n$ grid. Choose $\Theta(\log n)$ directions in this grid so that a line in any of these directions has $\Theta(\frac{n}{\log n})$ points on it. Choose each line in any of these directions with probability $1/2$. This gives a collection of n^2 points and $\Theta(n \log^2 n)$ lines, with each point having $\Theta(\log n)$ lines through it and each line having $\Theta(\frac{n}{\log n})$ points on it. The LP optimum is $O(n \log n)$ (each line can be given a weight of $\frac{O(1)}{\log n}$ for feasibility). If the integer optimum can be shown to be $\Omega(n \log^2 n)$, then an integrality gap of $\Theta(\log n)$ will follow.

9 Open Problems

A significant contribution of this paper is that it leads to several open problems.

1. Is there a polynomial time algorithm for constructing a hitting set for combinatorial rectangles with the properties described in Section 7? Alternatively, can a different proof system be obtained, as in [F98], which will require a set system with weaker hitting properties?

2. Can an integrality gap of $\Theta(\log n)$ be shown for the point-line examples given at the end of Section 8?

3. There are explicit constructions known for the general Set Covering problem in which the integrality gap in $\Theta(\log n)$. Are there such explicit constructions for the the Set Covering problem with intersection 1? Randomized constructions are easy for this but we do not know how to do an explicit construction.

4. Is there a polynomial time algorithm for the problem of covering points with lines which has an $o(\log n)$ approximation factor, or can super-constant hardness (or even a hardness of factor 2) be proved?

References

[AGHP92] N. Alon, O. Goldreich, J. Hastad, R. Perralta. Simple Constructions of Almost k-Wise Independent Random Variables. *Random Structures and Algorithms*, 3, 1992.

[AR99] V.S. Anil Kumar and H. Ramesh. Covering Rectilinear Polygons with Axis-Parallel Rectangles. Proceedings of *31st ACM-SIAM Symposium in Theory of Computing*, 1999.

[AL95] S. Arora, C. Lund. Hardness of Approximation. In *Approximation Algorithms for NP-Hard Problems*, Ed. D. Hochbaum, PWS Publishers, 1995, pp. 399-446.

[ALMSS92] S. Arora, C. Lund, R. Motwani, M. Sudan, M. Szegedy. Proof Verification and Intractability of Approximation Problems. Proceedings of *33rd IEEE Symposium on Foundations of Computer Science*, 1992, pp. 13-22.

[AS97] S. Arora, M. Sudan. Improved Low Degree Testing and Applications. Proceedings of the *ACM Symposium on Theory of Computing*, 1997, pp. 485–495.

[Be91] J. Beck. An Algorithmic Approach to the Lovasz Local Lemma I, *Random Structures and Algorithms*, 2, 1991, pp. 343-365.

[BGLR93] M. Bellare, S. Goldwasser, C. Lund, A. Russell. Efficient Probabilistically Checkable Proofs and Applications to Approximation, Proceedings of *25th ACM Symposium on Theory of Computing*, 1993, pp. 294-303.

[BG94] H. Brönnimann, M. Goodrich. Almost Optimal Set Covers in Finite VC-Dimension. *Discrete Comput. Geom.*, 14, 1995, pp. 263–279.

[CIK88] Y. Cheng, S.S. Iyengar and R.L. Kashyap. A New Method for Image compression using Irreducible Covers of Maximal Rectangles. *IEEE Transactions on Software Engineering*, Vol. 14, 5, 1988, pp. 651–658.

[F98] U. Feige. A threshold of $\ln n$ for Approximating Set Cover. *Journal of the ACM*, 45, 4, 1998, pp. 634–652.

[HM91] R. Hassin and N. Megiddo. Approximation Algorithms for Hitting Objects with Straight Lines. *Discrete Applied Mathematics*, 30, 1991, pp. 29-42.

[Jo74] D.S. Johnson. Approximation Algorithms for Combinatorial Problems. *Journal of Computing and Systems Sciences*, 9, 1974, pp. 256–278.

[Le87] C. Levcopoulos. Improved Bounds for Covering General Polygons by Rectangles. Proceedings of *6th Foundations of Software Tech. and Theoretical Comp. Sc.*, LNCS 287, 1987.

[LL+93] N. Linial, M. Luby, M. Saks, D. Zuckerman. Hitting Sets for Combinatorial Rectangles. Proceedings of *25 ACM Symposium on Theory of Computing*, 1993, pp. 286–293.

[LY93] C. Lund, M. Yannakakis. On the Hardness of Approximating Minimization Problems. Proceedings of *25th ACM Symposium on Theory of Computing*, 1993, pp. 286–293.

[MT82] N. Megiddo and A. Tamir, On the complexity of locating linear facilities in the plane, *Oper. Res. Let*, 1, 1982, pp. 194-197.

[NSS95] M. Naor, L. Schulman, A. Srinivasan. Splitters and Near-Optimal Derandomization. Proceedings of the *36th IEEE Symposium on Foundations of Computer Science*, 1995, pp. 182–191.

[Raz95] R. Raz. A Parallel Repetition Theorem. Proceedings of the *27th ACM Symposium on Theory of Computing*, 1995, pp. 447–456.

[RS97] R. Raz and S. Safra. A Sub-Constant Error-Probability Low-Degree test and a Sub-Constant Error-Probability PCP Characterization of NP. Proceedings of the *ACM Symposium on Theory of Computing*, 1997, pp. 475–484.

Appendix

10 Proofs of Lemmas 1, 2, and 3

Lemma 1 Proof:

Proof. Note that for $|C_k(v,a) \cap C_{k'}(w,b)|$ to exceed 1, either v, w must be identical or there must be an edge between v and w. The reason for this is that each element in $C_k(v,a)$ has the form $(e,*)$ where e is an edge incident at v while each element in $C_{k'}(w,b)$ has the form $(e',*)$, where e' is an edge incident at w. We consider each case in turn.

Case 1. Suppose $v = w$. Then either $k \neq k'$ or $k = k', a \neq b$.

First, consider $C_k(v,a)$ and $C_{k'}(v,b)$ where $k \neq k'$ and v is a vertex in the left side. If $a = b$, observe that $C_k(v,a) \cap C_{k'}(v,a) = \phi$. So assume that $a \neq b$. The elements in the former set are of the form (e,i) where $i \in P_{f_e(a),col(e),a,k}$ and the elements of the latter set are of the form (e,j) where $j \in P_{f_e(b),col(e),b,k'}$. Note that $\cup_{e \in E_v} P_{f_e(a),col(e),a,k} \subseteq B_k$ and $\cup_{e \in E_v} P_{f_e(b),col(e),b,k'} \subseteq B_{k'}$. By Property 3 of $\mathcal{P}$, the intersection $B_k, B_{k'}$ is at most 1. However, this alone does not imply that $C_k(v,a)$ and $C_{k'}(v,b)$ have intersection at most 1, because there could be several tuples in both sets, all having identical second entries. This could happen if there are edges e_1, e_2 incident on v such that $f_{e_1}(a) = f_{e_2}(a), f_{e_1}(b) = f_{e_2}(b)$ and there had been no colouring on edges. Property 2 and the fact that $col(e_1) \neq col(e_2)$ for any two edges e_1, e_2 incident on v rule out this possibility, thus implying that $|C_k(v,a) \cap C_{k'}(v,b)| \leq 1$. The proof for the case where v is a vertex on the right is identical.

Second, consider $C_k(v,a)$ and $C_k(v,b)$, where v is a vertex on the left and $a \neq b$. Elements in the former set are of the form (e,i) where e is an edge incident on v and $i \in P_{f_e(a),col(e),a,k}$. Similarly, elements in the latter set are of the form (e,j) where $j \in P_{f_e(b),col(e),b,k}$. Note that $\cup_{e \in E_v} P_{f_e(a),col(e),a,k} \subseteq A_{a,k}$ and $\cup_{e \in E_v} P_{f_e(b),col(e),b,k} \subseteq A_{b,k}$. The claim follows from Property 5 in this case.

Third, consider $C_k(v,a)$ and $C_k(v,b)$, where v is a vertex on the right, $a \neq b$, and $k > m/2$. Elements in the former set are of the form (e,i) where e is an edge incident on v and $i \in P_{a,col(e),1,k}$. Similarly, elements in the latter set are of the form (e,j) where $j \in P_{b,col(e),1,k}$. Note that $\cup_{e \in E_v} P_{a,col(e),1,k} \subseteq D_{a,k}$ and $\cup_{e \in E_v} P_{b,col(e),1,k} \subseteq D_{b,k}$. The claim follows from Property 6 in this case.

Case 2. Finally consider sets $C_k(v,a)$ and $C_{k'}(w,b)$ where $e = (v,w)$ is an edge, v is on the left side, and w on the right. Then $C_k(v,a)$ contains elements of the form (e',i) where $i \in P_{f_{e'}(a),col(e'),a,k}$. $C_{k'}(w,b)$ contains elements of the form (e',j) where $j \in P_{b,col(e'),1,k'}$. The only possible elements in $C_k(v,a) \cap C_{k'}(w,b)$ are tuples with the first entry equal to e. Since $P_{f_e(a),col(e),a,k} \subseteq B_k$ and $P_{b,col(e),1,k'} \subseteq B_{k'}$ and $k \leq m/2, k' > m/2$, the claim follows from Properties 2 and 3 in this case.

Lemma 2 Proof:

Proof. Let $label(v)$ denote the label given to vertex v by the above labelling. Consider the collection $C' \subset C$ comprising sets $C_1(v, label(v)) \ldots, C_{\frac{m}{2}}(v, label(v))$ for each vertex v on the left and sets $C_{\frac{m}{2}+1}(w, label(w)) \ldots, C_m(w, label(w))$ for each vertex w on the right. We show that these sets cover B. Since there are $m/2$ sets in C' per vertex, $|C'| = 2n' * \frac{m}{2} = n'm$.

Consider any edge $e = (v, w)$. It suffices to show that for every i, $1 \leq i \leq n'$, the tuple (e, i) in B is contained in either one of $C_1(v, label(v)) \ldots, C_{\frac{m}{2}}(v, label(v))$ or in one of $C_{\frac{m}{2}+1}(w, label(w)) \ldots, C_m(w, label(w))$. The key property we use is that $f_e(label(v)) = label(w)$.

Consider the partitions $P_{f_e(label(v)),col(e),label(v)}$ and $P_{label(w)),col(e),1}$. Since $f_e(label(v)) = label(w)$, the two partitions belong to the same group and subgroup. Since all partitions in a subgroup have the same right hand side, the element i must be present either in one of the sets $P_{label(w),col(e),label(v),k}$, where $k \leq m/2$, or in one of the sets $P_{label(w),col(e),1,k}$, where $k > m/2$. We consider each case in turn.

First, suppose $i \in P_{label(w),col(e),label(v),k}$, for some $k \leq m/2$. Then, from the definition of $C_k(v, label(v))$, $(e, i) \in C_k(v, label(v))$. Second, suppose $i \in P_{label(w),col(e),1,k}$, for some $k > m/2$. Then, from the definition of $C_k(w, label(w))$, $(e, i) \in C_k(w, label(w))$. The lemma follows.

Lemma 3 Proof:

Proof. Given C', we need to demonstrate a labelling of G with the above property. For each vertex v, define $L(v)$ to be the collection of labels a such that $C_k(v, a) \in C'$ for some k. We think of $L(v)$ as the set of "suggested labels" for v given by C' and this will be a multiset in general. The labelling we obtain will ultimately choose a label for v from this set. It remains to show that there is a way of assigning each vertex v a label from $L(v)$ so as to satisfy sufficiently many edges.

We need some definitions. For an edge $e = (v, w)$, define $\#(e) = |L(v)| + |L(w)|$. Since the sum of the sizes of all $L(v)$s put together is at most $\frac{\beta}{2} n'm \log n'$ and since all vertices in G have identical degrees, the average value of $\#(e)$ is at most $\frac{\beta}{2} m \log n'$. Thus half the edges e have $\#(e) \leq \beta m \log n'$. We call these edges *good*.

We show how to determine a subset $L'(v)$ of $L(v)$ for each vertex v so that the following properties are satisfied. If v has a good edge incident on it then $L'(v)$ has size at most $4\beta \log n'$. Further, for each good edge $e = (v, w)$, there exists a label in $L'(v)$ and one in $L'(w)$ which together satisfy e. Clearly, random independent choices of labels from $L'(v)$ will satisfy a good edge with probability $\frac{1}{16\beta^2 \log^2 n'}$, implying a labelling which will satisfies at least a $\frac{1}{32\beta^2 \log^2 n'}$ fraction of the edges (since the total number of edges is at most twice the number of good edges), as required.

For each label $a \in L(v)$, include it in $L'(v)$ if and only if the number of sets of the form $C_*(v, a)$ in C' is at least $m/4$. Clearly, $|L'(v)| \leq \frac{\beta m \log n'}{m/4} = 4\beta \log n'$, for vertices v on which good edges are incident. It remains to show that for

each good edge $e = (v, w)$, there exists a label in $L'(v)$ and one in $L'(w)$ which together satisfy e.

Consider a good edge $e = (v, w)$. Using Property 4 of $\mathcal{P}$, it follows that there exists a label $a \in L(v)$ and a label $b \in L(w)$ such that the $f_e(a) = b$ and the number of sets of the form $C_*(v, a)$ or $C_*(w, b)$ in C' is at least $3m/4$. The latter implies that the number of sets of the form $C_*(v, a)$ in C' must be at least $m/4$, and likewise for $C_*(w, b)$. Thus $a \in L'(v)$ and $b \in L'(w)$. Since $f_e(a) = b$, the claim follows.

Corollary 1 Proof:

Proof. The second part of the corollary is shown as follows. Lemma 1 ensures that the intersection in $\mathcal{SC}$ is at most 1. Recall from Section 2 that either all the edges in G are satisfied by some labelling or that no labelling satisfies more than a $\frac{1}{\log^3 n}$ fraction. Since $\frac{1}{32\beta^2 \log^2 n'} \geq \frac{1}{\log^2 n'} \geq \frac{1}{\log^3 n}$, we obtain from Lemma 2 and Lemma 3 that either there is a set cover of size $n'm$ for $\mathcal{SC}$ or any set cover for $\mathcal{SC}$ has size more than $\frac{\beta n' m \log n'}{2}$.

Consider the first part next. As will be shown shortly in Section 11.1, there is a randomized algorithm to construct the partition system $\mathcal{P}$ which always satisfies the properties 1, 2, and 3. Further, as we will show in Corollary 2, this partition system will satisfy property 4 with probability at least $\frac{1}{2}$. The corollary then follows as in the previous paragraph. Only the $ZTIME$ assumption needs explanation, as the partition system constructed above is not guaranteed to have property 4.

Consider a set cover instance $\mathcal{SC}$ produced by the reduction and consider any algorithm which approximates the minimum set cover in $\mathcal{SC}$ to a factor of $\frac{\beta \log n'}{2}$. Let C' denote the cover produced by this algorithm. If $|C'| > \frac{\beta n' m \log n'}{2}$ then, irrespective of $\mathcal{P}$ satisfying property 4, no labelling can satisfy more than a $\frac{1}{\log^3 n}$ fraction of the edges in the label cover graph G. But if $|C'| \leq \frac{\beta n' m \log n'}{2}$ then either it is the case that a labelling which satisfies all edges in G exists, or no such labelling exists because $\mathcal{P}$ fails to satisfy property 4. This latter situation can be checked in polynomial time and the experiment can be repeated until this situation does not arise. This checking is done as follows.

The claim is that if $|C'| \leq \frac{\beta n' m \log n'}{2}$ and no labelling which satisfies all edges in G exists, then there must exist a good edge $e = (v, w)$ (as defined in the proof of Lemma 3) with the following property: for each label $a \in L'(v)$, $f_e(a) \notin L'(v)$. This claim is easy to verify from the proof of Lemma 3, and the corresponding check is easily performed in time polynomial in n'.

11 Properties of the Randomized Construction

Recall the randomized construction algorithm from Section 5.

The Covering Property. Consider Property 4. Any collection S of at most $\beta m \log n'$ sets in which the number of sets picked from the left side of any

partition p and the number of sets picked from the right side of the first partition of the subgroup containing p add up to at most $\frac{3m}{4}$, is called a *valid* collection.

We show in the next paragraphs that the probability that a fixed element i is covered by a fixed valid collection S is upper bounded by $1 - \frac{1}{(n')^{22\beta}}$. Then the probability that each element of N is covered by S is at most $(1 - \frac{1}{(n')^{22\beta}})^{n'} \leq \frac{1}{e^{(n')^{1-22\beta}}}$. The number of such sets S is at most $2^{(n')^{(1-\epsilon)}\, d' * (deg+1) * d}$. Since $d' * (deg+1) * d$ is polylogarithmic in n', the total probability of all elements of N being covered by some such S is very small provided $22\beta < \epsilon$.

Consider the collection S and an element $i \in N$ as mentioned above. Let $r_s(S)$ be the number of sets chosen from the right side of the sth subgroup and let $l_{s,p}(S)$ be the number of sets chosen from the left side of the pth partition of the sth subgroup (recall that we are ignoring the division into groups and viewing $\mathcal{P}$ as a collections of subgroups). Then $\sum_s r_s(S) + \sum_{s,p} l_{s,p}(S) \leq \beta m \log n'$ and $r_s(S) + l_{s,p}(S) \leq \frac{3m}{4}$, for all s, p.

The probability that element $i \in N$ is not covered by S in subgroup s is equal to $\frac{1}{2}(1 - \frac{r_s(S)}{m/2}) + \frac{1}{2}\Pi_p(1 - \frac{l_{s,p}(S)}{m/2})$. The first term is the probability that i is in the right side and is not covered by S and the second term is the probability that i lies in the left side and is not covered by S in any of the partitions of this subgroup. The probability that element i is not covered in any subgroup is the product $\Pi_s \left[\frac{1}{2}(1 - \frac{r_s(S)}{m/2}) + \frac{1}{2}\Pi_p(1 - \frac{l_{s,p}(S)}{m/2})\right]$ over all subgroups s. Using $\sum_s r_s(S) + \sum_{s,p} l_{s,p}(S) \leq \beta m \log n'$ and $r_s(S) + l_{s,p}(S) \leq \frac{3m}{4}$, for all s, p, this expression is at least $\frac{1}{(n')^{22\beta}}$.

Intersection Properties: Consider Properties 2 and 3. First, we show an intersection bound of 2 instead of Property 3. Instead of Property 2, we show that no element will occur more than twice in a column, modulo the fact that the right sides of partitions within a subgroup are identical. Subsequently, we will use the Lovasz Local Lemma to get sharper bounds of 1 instead of 2 in each case.

The probability that three fixed elements $h, i, j \in N$ are present in both B_k and $B_{k'}$ is at most $(\frac{d*(deg+1)*d'}{m})^6$. Multiplying this by the number of choices of h, i, j, k, k' gives $(n')^3 m^2 (\frac{d*(deg+1)*d'}{m})^6 = o(1)$, as $m = (n')^{1-\epsilon}$ and $d*(deg+1)*d'$ is polylogarithmic in n. Similarly, the probability that a fixed element i appears thrice in a column k is at most $(\frac{d'*(deg+)*d}{m})^3$. Multiplying this by the number of choices of i, k gives $n'm(\frac{d'*(deg+1)*d}{m})^3 = o(1)$.

To get sharper intersection bounds of 1, we observe that the dependency number is small and use the Lovasz Local Lemma as below.

Consider Property 3. Let $F(x, y, B_k, B_{k'})$ be the event that elements $x, y \in N$ both occur in B_k and $B_{k'}$. For Property 3 to hold, no such event must occur. The probability of occurrence of any event $F(x, y, B_k, B_{k'})$ is at most $(\frac{(d'*(deg+1)*d)}{m})^4$. The number of events $F(*, *, *, *)$ is $(n')^2 m^2$ but the number of events on which a particular event $F(x, y, B_k, B_{k'})$ depends is at most $n'm^2$ because $F(x, y, *, *)$ and $F(x', y', *, *)$ are independent if x, y, x', y' are all dis-

tinct. Since $n'm^2(\frac{(d'*(deg+1)*d)}{m})^4 \leq \frac{1}{4}$, the condition for the Lovasz Local Lemma is satisfied.

Consider Property 2 next. An event violating property 2 involves some element i occurring twice in column k. The number of such events equals the number of choices of i, k, which is nm. Each event depends on only m events, as an event involving i is independent of one involving j. Since $m(\frac{d*(deg+1)*d'}{m})^2 \leq \frac{1}{4}$, the condition for the Lovasz Local Lemma is satisfied for this property as well.

Using a version of the Lovasz Local Lemma, we get that Properties 2 and 3 together hold with probability at least $(1-2\frac{(d'*(deg+1)*d)^4}{m^4})^{(n')^2m^2}(1-2\frac{(d'*(deg+1)*d)^2}{m^2})^{n'm} \geq (\frac{1}{e})^{(n')^{4\epsilon}}$.

The above use of the Lovasz Local Lemma poses some problems in derandomization. Typical derandomization of this lemma[Be91] requires $e\text{poly}(\Delta)p < 1$ as opposed to $e\Delta p < 1$, where Δ is the degree of dependency. This slack is too much for our situation. Instead, we first obtain a slightly different random experiment which does not require the Lovasz Local Lemma and then use appropriate pessimistic estimators to do the derandomization.

11.1 The New Randomized Experiment

In order to bypass the Lovasz Local Lemma, we will impose another restriction on the system of partitions $\mathcal{P}$, namely, that each set $P_{g,s,p,k}$ has size at most $\frac{d'*(deg+1)*dn'}{m}$, for all g, s, p, k, $1 \leq g \leq d', 1 \leq s \leq deg+1, 1 \leq p \leq d, 1 \leq k \leq m$. We call this Property 7.

Then we proceed as in the previous random experiment, fixing partitions in the same order as before, except that any choice of throwing an element $i \in N$ which violates Properties 2,3,7 is precluded. Property 7 enables us to show that not too many choices are precluded for each element, and therefore, this experiment stays close in behaviour to the previous one, except that Properties 2,3,7 are all automatically satisfied.

Suppose the partition that is being fixed currently is $P_{g,s,p}$. A position k for i would cause a violation of Property 2 if i occurs in some set $P_{g',s',p',k}$ which has already been fixed. A position k for i would cause a violation of Property 3 if there exist $k' \leq m$ and $j \in N$ such that i and j both already occur in $B_{k'}$ and j already occurs in B_k. A position k for i causes a violation of Property 7 if the size of the set $P_{g,s,p,k}$ exceeds $\frac{d'*(deg+1)*dn'}{m}$ after i is put in that position. All such positions above are said to be *bad* for i in the current partition. The following lemma shows that very few positions are bad for i in any given partition, and therefore, this new experiment behaves similar to the previous one; therefore, Property 4 will continue to hold, but with a slightly modified proof. This proof appears as part of the derandomization in Section 12 (see Corollary 2).

Lemma 4. *The total number of bad positions for i when processing partition $P_{g,s,p}$ is at most $\frac{3m}{d'*(deg+1)*d}$. Therefore, each element i is distributed uniformly over a range of at least $m(1-\frac{3}{d'*(deg+1)*d})$ sets in the first partition of any sub-*

group and over a range of at least $\frac{m}{2}(1-\frac{6}{d'*(deg+1)*d})$ *sets in the other partitions of any subgroup in the above random experiment.*

Proof. Since all previous partitions (suppose there are x of these) have been fixed so far satisfying Properties 2, 3 and 7, the size of the largest column is at most $x\frac{n'd'*(deg+1)*d}{m}$.

The number of bad positions for i violating Property 2 is at most x. The number of bad positions k for i violating Property 3 is at most $x\frac{n'd'*(deg+1)*d}{m}*x^2$. This is because k is bad if there exist $j \in N$ and $k' \leq m$ such that i and j both already occur in $B_{k'}$ and j already occurs in B_k; the number of js can be at most $x\frac{n'd'*(deg+1)*d}{m}*x$ (all elements in the at most x columns already containing i are candidates) and the number of k's for each such j is at most x. The number of bad positions for i violating Property 7 is at most $\frac{m}{d'*(deg+1)*d}$. The total number of bad positions is thus at most $3\frac{m}{d'*(deg+1)*d}$ since $x^3\frac{n'd'*(deg+1)*d}{m} \leq \frac{m}{d'*(deg+1)*d}$. The last statement follows if $n' < m^2$, ie, $\epsilon < 1/2$.

12 Derandomization using Conditional Probabilities

First, we describe our pessimistic estimator. Subsequently, we show how to use it for derandomization.

12.1 The Pessimistic Estimator for Conditional Probabilities

We order all the subgroups globally. At any instant in our new randomized experiment, we will be processing a particular partition in some subgroup. All previous partitions would have been fixed and all subsequent partitions are currently untouched. Further, the positions of some elements in the current partition would also have been fixed. Before defining the estimator, we need the following definitions.

Definitions. We classify all subgroups in a partly fixed set system H into 3 classes: *completely fixed*, *partly fixed* and *untouched*.

Let $U = \frac{3m}{d'*(deg+1)*d}$. By Lemma 4, U is an upper bound on the number of bad locations in any partition.

Consider a particular subgroup s and a valid collection S. Let $r_s(S)$ be the number of sets in S in the right side of subgroup s and $r'_s(S) = max(m/2 - r_s(S) - U, 0)$. Let $l_{s,p}(S)$ be the number of sets in S in the left side of the pth partition of the sth subgroup and $l'_{s,p}(S) = max(m/2 - l_{s,p}(S) - U, 0)$.

For each subgroup s, $i \in N$, valid collection S, define $h(s, i, S, H)$ as follows. $h(s, i, S, H)$ will be a lower bound on the probability that element i is not covered by S in subgroup s. There are several cases, and the definition of $h(s, i, S, H)$ is different in each case.

1. If i is already covered by S in any of the subgroups fixed in the partially fixed set system H, $h(s, i, S, H) = 0$. Otherwise, one of the following cases holds.

2. Subgroup s has already been fixed and element i is not covered by S in H. Then $h(s,i,S,H) = 1$.
3. Element i has not yet been fixed in the first partition of the subgroup s and is not covered by S in H. $h(s,i,S,H) = (\frac{1}{2} - \frac{U}{m})(\frac{r'_s(S)}{m/2} + \Pi_p \frac{l'_{s,p}(S)}{m/2})$.
4. Element i has been fixed in the first partition of subgroup s but not in all the partitions of the subgroup. Suppose partition $p > 1$ of subgroup s is being fixed currently. If i lies the right side of the first partition of the subgroup s, then $h(s,i,S,H) = 1$ and if i lies in the left side of the first partition of the subgroup s, $h(s,i,S,H) = \Pi_{p' \geq p} \frac{l'_{s,p'}(S)}{m/2}$.

Now define another quantity $g(i,S,H)$, which will be shown to be an upper bound on the probability that i is covered over the remaining choices, as $1 - \Pi_s h(s,i,S,H)$, the product being over all subgroups s which haven't been fixed completely. Finally, define $f(S,H) = \Pi_i g(i,S,H)$, which will be an upper bound on the probability that every element i is covered by S.

The pessimistic estimator $F(H)$ is defined as $\sum_S f(S,H)$, where the sum is over all valid collections S. This will turn out to be an upper bound on the the expected number of valid collections S which cover N.

Lemma 5. *For any partial set system H, $F(H)$ is an upper bound on the expected number of valid collections that cover N, the expectation being over all random set systems that contain H.*

Proof. Consider any valid collection S. Recall that there is a global ordering on all subgroups, without any partition into groups. Let $l_{s,p}(S), l'_{s,p}(S), r_s(S), r'_s(S)$ be defined as above for each subgroup s and partition p. We prove below that the probability that element i is not covered in subgroup s, conditioned on element i not being covered in all earlier subgroups and on any subset $N' \subset N$ of elements being covered, is at least $h(s,i,S,H)$. Then the probability that i is not covered, conditioned on the set N' of elements being covered, is at least $\Pi_s h(s,i,S,H)$, the product being over all subgroups s that have not been completely fixed. Therefore, the probability that element i is covered, conditioned on the above event, is at most $1 - \Pi_s h(s,i,S,H) = g(i,S,H)$. The probability that all elements are covered is consequently at most $f(S,H) = \Pi_i g(i,S,H)$ and the expected number of valid collections S that cover all elements is at most $F(H) = \sum_S f(S,H)$.

It remains to be shown that the probability that element i is not covered in subgroup s, conditioned on it not being covered in earlier subgroups and on any subset $N' \subset N$ of elements being covered, is at least $h(s,i,S,H)$. We consider various cases below which are the same as in the earlier definition of $h(s,i,S,H)$.

1. i is covered in some subgroup by S. $h(s,i,S,H) = 0$ is clearly a lower bound.
2. Subgroup s is fixed and i is not covered anywhere in H by S. Then $h(s,i,S,H) = 1$ is exactly the probability that i is not covered in s.
3. Element i is not yet fixed in the first partition of s, and it is not yet covered by S. Then the probability that i is not covered by S in this subgroup

is the sum of the probabilities of not being covered in the left and in the right. The probability that i is placed in the right is the ratio of the number of good locations in the right to the total number of good locations. This is at least $\frac{m/2-U}{m}$. This is smaller than the actual value, because some positions could become bad after the conditioning (recall that we want to show that $h(s,i,S,H)$ is a lower bound on the probability of not getting covered in s, *conditioned* on not getting covered in earlier subgroups). The probability that element i is not covered if placed in the right is at least $\frac{max(m/2-r_s(S)-U,0)}{\#\text{ good locations in the right}} \geq \frac{r'_s(S)}{m/2}$. Therefore the probability that i is not covered in the right is at least $(\frac{1}{2}-\frac{U}{m})\frac{r'_s(S)}{m/2}$. Similarly, given that i is placed on the left side, the probability that i is not covered in partition p is at least $\frac{max(m/2-l_{s,p}(S)-U,0)}{\#\text{ good locations in the left}} \geq \frac{l'_{s,p}(S)}{m/2}$. Therefore the probability that i is not covered in the left is at least $(\frac{1}{2}-\frac{U}{m})\Pi_p(\frac{l'_{s,p}(S)}{m/2})$. The sum of the above quantities is equal to $h(s,i,S,H)$.

4. i is fixed in the first partition of s, but not in all the partitions of s. Suppose partition $p>1$ is being fixed. If i lies in the right side of partition 1 of s, $h(s,i,S,H)=1$. If i lies in the left side of partition 1 of s, the probability that i is not covered is at least $h(s,i,S,H)=\Pi_{p'\geq p}\frac{l'_{s,p}(S)}{m/2}$, by an argument similar to the one in case 1.

Let $H=\phi$ denote the set system at the beginning of the experiment, when nothing has been fixed.

Lemma 6. $F(\phi) \leq (\frac{1}{e})^{n'^{1-23\beta}}$.

Proof. Consider any valid collection S. Let $l_{s,p}(S), l'_{s,p}(S), r_s(S), r'_s(S)$ be defined as above for each subgroup s and partition p. Then $r_s(S)+l_{s,p}(S) \leq \frac{3m}{4}$ for each s,p. In addition, $\sum_s r_s(S)+\sum_{s,p} l_{s,p}(S) \leq \beta m \log n'$. We prove below that the probability that element i is not covered, $\Pi_s h(s,i,S,H) \geq \frac{1}{n'^{22\beta}}$. Then the probability that i is covered, $g(i,S,H)=1-\Pi_s h(s,i,S,H) \leq 1-\frac{1}{n'^{22\beta}}$. From this, it follows that the probability that all elements in N are covered is at most $f(S,H)=\Pi_i g(i,S,H) \leq (1-\frac{1}{n'^{22\beta}})^{n'} \leq (\frac{1}{e})^{n'^{1-22\beta}}$.

The above quantity, $f(S,H)$ is an upper bound on the probability that a fixed valid collection S covers N. So the expected number of such collections that cover N is $F(H)=\sum_S f(S,H)$, where the sum is over all valid collections S. Since there are at most $2^{n'^{1-\epsilon}d'*(deg+1)*d}$ such collections, the expected number of such collections is at most $2^{n'^{1-\epsilon}d'*(deg+1)*d}(\frac{1}{e})^{n'^{1-22\beta}}$. The above quantity is at most $(\frac{1}{e})^{n'^{1-23\beta}}$, thus implying the lemma.

Finally, we show that $\Pi_s h(s,i,S,H) \geq \frac{1}{n'^{22\beta}}$. The following simple fact will be useful. We state it without proof.

Fact 1 *$\Pi_{i=1\ldots k}(1-\frac{a_i}{m}) \geq (1-r)^{\frac{x}{mr}}$ where $\sum_i a_i \leq x$ and $a_i \leq mr$ for all i and $r<1$.*

We consider three cases. Let $x_s=\sum_p(l_{s,p}(S)+U)$ in the following cases.

1. $\frac{2(r_s(S)+U)}{m} \geq \frac{3}{4}$
   Since $\frac{l_{s,p}(S)+U}{m} + \frac{r_s(S)+U}{m} \leq \frac{3}{4} + \frac{2U}{m}$ for each p, it follows that $\frac{2(l_{s,p}(S)+U)}{m} \leq \frac{3}{4} + \frac{4U}{m}$. $h(s,i,S,H) \geq \frac{1}{2}(1-\frac{U}{2m})\Pi_p \frac{l'_{s,p}(S)}{m/2} \geq \frac{1}{2}(1-\frac{U}{2m})(1-\frac{3}{4}-\frac{4U}{m})^{\frac{4x_s}{3m}}$. The second inequality follows from the above fact. The number of such subgroups is at most $4\beta \log n'$. Hence the product of $h(s,i,S,H)$ over all such subgroups s is at least $\frac{1}{n'^{4\beta}}\frac{1}{2}(\frac{1}{8})^{\frac{4\beta \log_2 n'}{3}} \geq \frac{1}{n'^{8\beta}}$. This is because the product of $\frac{1}{2}(1-\frac{U}{2m})$ over all such subgroups is at least $\frac{1}{2}\frac{1}{n'^{4\beta}}$ and the product $(1-\frac{3}{4}-\frac{4U}{m})^{\frac{4x_s}{3m}}$ over all such subgroups is at least $(\frac{1}{8})^{\frac{4\beta \log n'}{3}}$ (since $\sum_s x_s \leq \beta m \log n'$).
2. $\frac{2(r_s(S)+U)}{m} \leq \frac{3}{4}$ and $x_s \leq \frac{m}{4}$
   Each $(l_{s,p}(S)+U) \leq \frac{m}{4}$ since $x_s = \sum_p (l_{s,p}(S)+U) \leq \frac{m}{4}$. Therefore $\frac{2(l_{s,p}(S)+U)}{m} \leq \frac{1}{2}$. Let the number of subgroups that satisfy the above conditions be y. Since $h(s,i,S,H) = (\frac{1}{2}-\frac{U}{m})(\frac{r'_s(S)}{m/2} + \Pi_p \frac{l'_{s,p}(C)}{m/2}) = (\frac{1}{2}\frac{r'_s(S)}{m/2} + \frac{1}{2}\Pi_p \frac{l'_{s,p}(S)}{m/2})(1-\frac{U}{2m})$, the product $\Pi_s h(s,i,S,H) = (1-\frac{U}{2m})^y \Pi_s (\frac{1}{2}\frac{r'_s(S)}{m/2} + \frac{1}{2}\Pi_p \frac{l'_{s,p}(S)}{m/2})$. The above product is only over subgroups s that satisfy the above conditions. The first term, $(1-\frac{U}{2m})^y$ is at least $\frac{1}{2}$. The second term, $\Pi_s(\frac{1}{2}\frac{r'_s(S)}{m/2} + \frac{1}{2}\Pi_p \frac{l'_{s,p}(S)}{m/2})$ can be expanded as a sum of 2^y terms, each corresponding to a set $A \subset \{1,\ldots,y\}$. The term corresponding to one such set A is $\frac{1}{2^y}\Pi_{s\in A}\frac{r'_s(S)}{m/2}\Pi_{s\notin A}\Pi_p \frac{l'_{s,p}(S)}{m/2}$. We show that each such term is at least $\frac{1}{2^y}\frac{1}{n'^{16\beta/3}}$ For each $s \in A$, $\frac{r'_s(S)}{m/2} = 1 - \frac{r_s(S)+U}{m/2}$ with $2(r_s(S)+U) \leq \frac{3}{4}m$. For each $s \notin A$ and for all p, $\frac{l'_{s,p}(S)}{m/2} = 1 - \frac{l_{s,p}(S)+U}{m/2}$ with $2(l_{s,p}(S)+U) \leq \frac{1}{2}m \leq \frac{3}{4}m$. In addition, $\sum_{s\in A}(r_s(S)+U) + \sum_{s\notin A, p\leq d}(l_{s,p}(S)+U) \leq \beta m \log n' + 2d' * (deg+1) * dU$ because $\beta m \log n'$ is an upper bound on the number of sets chosen in S. Using the above fact, the term for A is at least $\frac{1}{2^y}(1-\frac{3}{4})^{\frac{8\beta m \log n'}{3m}}$ which is at least $\frac{1}{2^y}\frac{1}{n'^{16\beta/3}}$. Therefore the product over all subgroups satisfying this case is at least $\frac{1}{n'^{6\beta}}$.
3. $\frac{2(r_s(S)+U)}{m} \leq \frac{3}{4}$ and $x_s \geq \frac{m}{4}$
   Then $\frac{1}{2}(1-\frac{2(r_s(S)+U)}{m}) \geq \frac{1}{2}\frac{1}{4}$. But there are at most $4\beta \log n'$ such subgroups s with $x_s \geq \frac{m}{4}$. Since $h(s,i,S,H) \geq \frac{1}{2}(1-\frac{U}{2m})(1-\frac{2(r_s(S)+U)}{m})$, the product of $h(s,i,S,H)$ over all such subgroups is at least $\frac{1}{2}(\frac{1}{4})^{4\beta \log_2 n'}$ which is at least $\frac{1}{n'^{8\beta}}$.

Therefore the product of $h(s,i,S,H)$ over all subgroups gives a lower bound of $\frac{1}{n'^{22\beta}}$.

Corollary 2. *The randomized experiment succeeds in giving the required partition system $\mathcal{P}$ with probability at least $\frac{1}{2}$.*

12.2 The Derandomization

We now use the method of conditional probabilities to find such a set system. At each step of the experiment, the position of some element i is being fixed. This

position is chosen from only those possibilities that do not cause a violation of properties 2,3,7. From Lemma 4, there is a large set of choices which do not cause a violation of properties 2,3,7. We need one that would not cause a violation of property 4. Suppose the current partial configuration is H. For each possible configuration H_k resulting from the choice of position k for element i, we compute the value of the estimator $F(H_k)$. The lemma below shows that $F(H)$ is at least as much as the average of $F(H_k)$ over all k that can be chosen (recall that a choice is made from the set of positions that are not bad). If $F(H) < 1$, there exists a k such that $F(H_k) < 1$. By Lemma 6, $F(\phi) < 1$. So at each step of the experiment one can find a choice that does not increase the value of the estimator. Since $F(H)$ is an upper bound on the expected number of valid collections S that cover N, we eventually get a set system with the desired properties.

Lemma 7. *Let H be a partial set system at instant t for the random experiment and suppose the position of element i is being fixed at the present instant. Let H_k denote the configuration corresponding to the choice of k as the position for i. Then $F(H)$ is at least as much as the average of $F(H_k)$ over all possible good choices k.*

Proof. We show that the term for each valid collection S, $f(S,H)$ is at least as much as the average of $f(S,H_k)$ over all possible choices k for i at instant $t+1$ in the random experiment. Since $F(H) = \sum_S f(S,H)$, the lemma then follows. For $i' \neq i, g(i',S,H_k) = g(i',S,H)$. Since $f(S,H) = \Pi_z g(z,S,H)$, it suffices to show that $g(i,S,H) \geq \frac{1}{\#\text{ choices of k}} \sum_k (g(i,S,H_k))$. Recall that $g(i,S,H) = 1 - \Pi_s h(s,i,S,H)$, the product being over subgroups s that have not been completely fixed.

Suppose that the pth partition of the sth subgroup is being fixed currently. Then there are two cases : $p = 1$ and $p > 1$. First, consider the case $p = 1$. Then $g(i,S,H) = 1 - h(s,i,S,H)\Pi_{s'>s} h(s',i,S,H) = 1 - (\frac{1}{2} - \frac{U}{m})(\frac{r'_s(S)}{m/2} + \frac{l'_{s,1}(S)}{m/2}\alpha)\beta$, where $\alpha = \Pi_{p'>1}(\frac{l'_{s,p'}(S)}{m/2})$ and $\beta = \Pi_{s'>s} h(s',i,S,H)$. $g(i,S,H_k) = 1$ if i gets covered in H_k. If i is not covered but is placed in right side, $g(i,S,H_k) = 1 - \beta$. If i is not covered but is placed in the left side, $g(i,S,H_k) = 1 - \alpha\beta$.

Let b_1 and b_2 denote the number of bad locations in the left and right sides respectively and $b = b_1 + b_2$. Let g_1 denote the number of bad positions in the left which are in S and g'_1 denote the number of bad positions in the left which are not in S. Similarly g_2 denotes the number of bad positions in the right which are in S and g'_2 denotes the number of bad positions in the right not in S. Then $g_1 + g'_1 = b_1$ and $g_2 + g'_2 = b_2$. Let $g = g_1 + g_2$. Then $l_{s,1}(S) + r_s(S) - g \geq 0$, $m/2 + g_1 - l_{s,1}(S) - b_1 \geq 0$ and $m/2 + g_2 - r_s(S) - b_2 \geq 0$. The average of $g(i,S,H_k)$ over all k that are good in the random experiment for i is exactly $\frac{l_{s,1}(S)+r_s(S)-g}{m-b} + \frac{m/2+g_1-l_{s,1}(S)-b_1}{m-b}(1-\alpha\beta) + \frac{m/2+g_2-r_s(S)-b_2}{m-b}(1-\beta) = 1 - \frac{m/2+g_1-l_{s,1}(S)-b_1}{m-b}\alpha\beta - \frac{m/2+g_2-r_s(S)-b_2}{m-b}\beta$. Since $r'_s(S) = max(m/2 - r_s(S) - U, 0) \leq m/2 - r_s(S) - b_2 + g_2$ and $l'_{s,1}(S) = max(m/2 - l_{s,1}(S) - U, 0) \leq m/2 - l_{s,1}(S) - b_1 + g_1$, the lemma follows.

Next consider the case $p > 1$. Then $g(i,S,H) = 1 - h(s,i,S,H)\Pi_{s'>s}h(s',i,S,H) = 1 - (\frac{l'_{s,p}(S)}{m/2})\gamma\beta$ where $\beta = \Pi_{s'>s}h(s',i,S,H)$ and $\gamma = 1$ if $p = d$, the last partition in the subgroup, and is $\Pi_{p'>p}(\frac{l'_{s,p'}(S)}{m/2})$ otherwise. $g(i,S,H_k) = 1$ if i gets covered and it is $1 - \gamma\beta$ if it does not get covered. Let b_1, g_1, g'_1 be as defined above. Then the average of $g(i,S,H_k)$ over all valid k at this instant is $\frac{l_{s,p}(S)-g_1}{m/2-b_1} + \frac{m/2+g_1-l_{s,p}(S)-b_1}{m/2-b_1}(1-\gamma\beta) = 1 - \frac{m/2+g_1-l_{s,p}(S)-b_1}{m/2-b_1}\gamma\beta$. Since $l'_{s,p}(S) = max(m/2 - l_{s,p}(S) - U, 0) \leq m/2 + g_1 - l_{s,p}(S) - b_1$, the lemma follows.

13 Showing Hardness under $NP = ZPP$

This requires starting with a proof system for NP satisfying the following properties: all but (5) below are satisfied by [RS97,AS97]. We are currently exploring whether (5) can be satisfied.

1. A constant number of provers, say p.
2. $O(\log n)$ bit of randomness.
3. Error probability which is $O(\frac{1}{\log^k n})$.
4. $O(\log\log n)$ answer sizes.
5. $O(n^\delta)$ *degree* for some small enough constant δ; the degree is the number of random strings for which a particular question is asked of a particular prover.
6. The questions asked of a particular prover are uniformly distributed over a set of all possible questions asked of this prover (this is the uniformity property).
7. The cardinalities of the sets of all possible questions asked of each prover are the same (this is the equality property).
8. For each random string generated by the verifier and for each answer by the first prover to the question generated by this random string, there is at most one combination of answers for the remaining provers for which the verifier accepts (this is the uniqueness property).
9. The set of answers returned by a prover is disjoint from the set of answers returned by any other prover (this is the disjointness property).

The corresponding label cover abstraction for this would be a Label Cover in a multi-layered hypergraph. The number of layers equals the number of provers. A hyperedge is a collection of vertices, exactly one from each layer, and corresponds to one question asked by the verifier to each of the provers. Since the verifier uses only $O(\log n)$ random bits, there are a polynomial number of vertices and hyperedges in this hypergraph. Further, the uniformity property ensures that the ith component of a random hyperedge is uniformly distributed over the ith layer of vertices (this will be useful in the counterpart of Lemma 3 which we will need now).

The new Label Cover problem now involves giving labels to the vertices so that as many hyperedges as possible are made consistent. These labels correspond to the answers returned by the provers. It is guaranteed now that either

all edges can be made consistent or at most $O(\frac{1}{\log^p n})$ edges can be made consistent (here p is the number of provers).

The uniqueness property ensures that for a particular hyperedge and for any way of labelling the first vertex in this hyperedge, there is a unique labelling to the other vertices in this hyperedge which will lead to consistency for this hyperedge. Since the answer sizes are $O(\log \log n)$, the pool of labels is $O(\text{polylog}(n))$ in size. Since the degree is $O(n^\delta)$, the number of hyperedges incident on a vertex is $O(n^\delta)$ (this is the value of deg now). With these observations, it can be seen that only the following changes need to be made in our construction of $\mathcal{SC}$.

The edges of the hypergraph are coloured using $O(p * deg)$ colours. As before, the sets corresponding to vertices in the first layer will be associated with the left sides of a partition. The sets corresponding to vertices in the remaining layers will be associated with the right sides of a partition. These are defined as follows.

We will now have a partition system $\mathcal{P}$ but with $n' = n^y$, for some constant y. For a vertex v which is not in the first layer and for a label a given to this vertex, the set $C_k(v,a) = \cup_{e \in E_v} \{(e,i) | i \in P_{a,col(e),1,k}\}$, with $m/2 < k \leq m$, as before. For a vertex v which is in the first layer and for a label a given to this vertex, the set $C_k(v,a)$ is now defined as follows. Let $f_e(a) = (b_2, \ldots, b_k)$ be now defined as the unique labelling to the other vertices in e which makes hyperedge e consistent, given that the first vertex has label a. For each $k = 1 \ldots m/2$, $C_k(v,a) = \cup_{e=(v,v_2,\ldots,v_k) \in E_v} \{(e,i) | i \in P_{b_2,col(e),a,k} \wedge \forall k' = m/2+1 \ldots m, \forall j, 2 \leq j \leq p, (e,i) \notin C_{k'}(v_j, b_j),$ where $f_e(a) = (b_2, \ldots, b_k)\}$.

Intersection properties follow as before (but using the disjointness property as well); note that deg has gone up but the earlier argument (see Lemma 4) works as long as $deg = O(n^\delta)$ for some small enough δ. Next, if all hyperedges are satisfied by some label cover, then the minimum set cover has size at most $\frac{m}{2}$ times the total number of vertices. We now prove a statement analogous to Lemma 3 that a small set cover leads to several hyperedges being satisfied.

If the optimum set cover size is a suitable constant times $\beta m \log n^y$ times the number of vertices, then by the uniformity property, several hyperedges have a total of at most $M = \frac{\beta m \log n^y}{2}$ labels each. Call a label a for vertex v *heavy* if at least $m/4$ sets of the form $C_*(v,a)$ are picked in the optimum set cover. Consider one such hyperedge $e = (v, v_2, \ldots, v_p)$. It now suffices to show that there exists a consistent labelling $(a, b_2, \ldots, b_p)$ of vertices in this hyperedge e comprising only heavy labels. Then, a random choice at each vertex now satisfies a $\Theta(\frac{1}{\log^p n})$ fraction of the edges.

Suppose this is not true, i.e., each consistent labelling of the vertices of e has a light label. Then we derive a contradiction by showing that all tuples of the form (e,i), $1 \leq i \leq n^y$, could not possibly have been covered by the sets in the (claimed) optimum set cover being considered.

Since we have fixed e, we will just talk in terms of covering all i, $1 \leq i \leq n^y$, instead of (e,i). We say that set $C_*(*,*)$ contains i if (e,i) is in this set. Thus we can talk of each set in the claimed optimum set cover containing or not containing some of the i's in the range $1 \ldots n^y$.

We now derive a contradiction by exhibiting a new collection of sets whose union contains all the is covered by the sets in the claimed optimum set cover, and in which the following two properties hold: i) this collection has at most $2M$ sets and, ii) no more than $3m/4$ sets come from any single partition. By Property 4, this new collection could not have covered all the is. The contradiction follows.

The new collection of sets is obtained as follows. Start with the claimed optimum cover and do the following for each consistent labelling $(a, b_2, \ldots, b_p)$ of the vertices of e. If a is light for v then do nothing. Otherwise, there is a light label b_j; then discard sets of the form $C_*(v, a)$ in the claimed optimum set cover and include all $m/2$ sets of the form $\{i | i \in P_{b_j, col(e), a, k}\}$ (one set for each k from $1 \ldots m/2$). Performing the above for all consistent labellings of the vertices of e gives the new collection of sets.

It remains to show that this collection has the required properties. Note that $\cup_{k=1}^{m/2} P_{b_j, col(e), a, k}$ contains all the is that are contained in sets of the form $C_*(v, a)$ in the claimed optimum set cover. Clearly, since b_j is light, the new collection does not contain more than $3m/4$ sets from any partition. Further, the total number of sets in the new collection is at most $2M$ (at least $m/4$ sets must be discarded for $m/2$ new sets to be added). This completes the proof.