







Learning Partitions Using Rank Queries

Deeparnab Chakrabarty   

Dartmouth College, Hanover, NH, USA

Hang Liao   

Dartmouth College, Hanover, NH, USA

Abstract

We consider the problem of learning an unknown partition of an n element universe using rank queries. Such queries take as input a subset of the universe and return the number of parts of the partition it intersects. We give a simple $O(n)$ -query, efficient, deterministic algorithm for this problem. We also generalize to give an $O(n + k \log r)$ -rank query algorithm for a general partition matroid where k is the number of parts and r is the rank of the matroid.

2012 ACM Subject Classification Theory of computation \rightarrow Streaming, sublinear and near linear time algorithms

Keywords and phrases Query Complexity, Hypergraph Learning, Matroids

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2024.16

Related Version *Extended Version*: <https://arxiv.org/abs/2409.13092>

Funding *Deeparnab Chakrabarty*: Supported by NSF grants 2041920 and 2402571.

Hang Liao: Supported by NSF grant 2041920.

1 Introduction

Let V be a universe of n elements and suppose there is an *unknown* partition $\mathcal{P} = (P_1, \dots, P_k)$ that we want to learn. We have an oracle called **rank** that takes as input any subset $S \subseteq V$ and returns the number of different parts this subset intersects. More precisely $\text{rank}(S) := \sum_{i=1}^k \min(|S \cap P_i|, 1)$. How many queries suffice to learn \mathcal{P} ?

This natural question is a special case of the problem of *learning hypergraphs* under the *additive query* model initially studied by [28]. In this problem, we have an unknown hypergraph on a vertex set V , and an additive query **add**(T) on a subset $T \subseteq V$ returns the *number* of hyperedges completely contained in T . Our unknown partition \mathcal{P} is a special hypergraph whose k hyperedges are disjoint (that is, it is a hypermatching); and for any subset S we observe that $\text{rank}(S)$ is precisely $k - \text{add}(V \setminus S)$. And so, the problem we study can be rephrased as in how few additive queries can a hypermatching be learnt. Although hypermatchings may feel too specialized, the now mature literature on *graph learning* (cf. [22, 17, 15, 16, 23]) began with understanding the case of graph matchings (cf. [28, 4, 3]).

The problem we study is also a special case of a *matroid learning* problem with access to rank oracle queries. Matroids are set systems, whose elements are called *independent* sets, that are defined using certain axioms and these are fundamental objects in combinatorial optimization. It is well known that a partition \mathcal{P} induces the following simple partition matroid: a subset $I \subseteq V$ is independent if $|I \cap P_i| \leq 1$ for all i . The rank of a matroid is the cardinality of the largest independent set of the matroid, and more generally, the rank of subset S is the cardinality of the largest independent set that is a subset of S . A moment's notice shows that for the simple partition matroid this is precisely $\text{rank}(S)$ which explains the name we give to our oracle. So, our problem we study asks: in how few rank queries can a simple partition matroid be learnt?



© Deeparnab Chakrabarty and Hang Liao;

licensed under Creative Commons License CC-BY 4.0

44th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2024).

Editors: Siddharth Barman and Sławomir Lasota; Article No. 16; pp. 16:1–16:14



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

It is rather straightforward¹ to learn the partition using $O(n \log k)$ queries as follows. First, one learns a representative from each part with n -queries; given a set of already learned representatives R , a vertex v is in a new unrepresented part if and only if $\text{rank}(R \cup v) > \text{rank}(R)$. After learning the k representatives, we can learn every other vertex's part by performing a binary search style algorithm. Can one do better? It is instructive to note that the algorithm sketched above does not really utilize the full power of the query model we have. In particular, it would have sufficed if the query took a subset S and said YES if every element in S was in a different part, or NO otherwise. Using the matroid language, an *independence oracle* suffices which only states if a set S is independent or not. Now, an independence oracle answer gives at most 1 bit of information; on the other hand, there are roughly k^n different partitions possible with $\leq k$ parts. Therefore, via an information theoretic argument $\Omega(n \log k)$ independence queries are *necessary* to learn the partition. In contrast, the *rank* oracle gives the *number* of different parts hit by a subset; this is an integer in $\{0, 1, \dots, k\}$ and the information theoretic argument only proves an $\Omega(n)$ lower bound on the number of queries. This naturally leads to the question: can an $O(n)$ -query algorithm exist? The main result of this paper is a simple affirmative answer to this question.

► **Theorem 1.** *There is a deterministic, constructive algorithm that solves unknown partition learning problem using $O(n)$ rank queries.*

► **Remark.** We have not optimized the constant in front of n . We think it can be made less than 10 but don't believe can be made less than 4 using our methods. The best lower bound one can prove using the above information theory argument is n . Figuring out the precise coefficient is left as an open question.

We also consider the generalization of learning a *general* partition matroid using rank queries. In this case, along with the unknown partition \mathcal{P} , we have unknown positive integers r_1, \dots, r_k associated with each part, where $1 \leq r_i < |P_i|$. A subset I is independent in this matroid if $|I \cap P_i| \leq r_i$, for all $1 \leq i \leq k$. When all $r_i = 1$, we have the simple partition matroid. The rank query corresponds to $\text{rank}(S) := \sum_{i=1}^k \min(|S \cap P_i|, r_i)$. In how few rank queries can we learn a general partition matroid?

As in the simple partition matroid case, one can get an $O(n \log k)$ -query algorithm using just an independent set oracle via a more delicate² binary-search-style algorithm. Can we obtain $O(n)$ query algorithm with rank queries? We believe the answer should be yes and take the following first step.

► **Theorem 2.** *There is a deterministic, constructive algorithm that learns a general partition matroid using $O(n + k \log r)$ rank queries where $r := \text{rank}(V) = \sum_i r_i$.*

► **Remark.** When the number of parts $k \leq n / \log n$, we thus get an $O(n)$ -rank query algorithm. However, when $k = \Omega(n)$ we don't do any better than just with independence queries.

Perspective

Our motivation to look at the problem arose from trying to understand the *connectivity* question in hypergraphs using CUT queries. Although, as mentioned earlier, graph learning under query models has been extensively studied, over the last few years, multiple works such as [36, 27, 29, 8, 6, 20, 30] have focused on trying to understand if fewer queries can

¹ Something that can be given in an undergraduate algorithms course when teaching binary search.

² Maybe a challenging exercise in the aforementioned algorithms course; see Section 3 for this algorithm.

lead to understanding *properties* of graphs. Of particular interest is understanding the *connectivity*/finding spanning forest of a graph using CUT queries. A CUT query takes a subset of vertices as input and returns the number/weight of the cut edges crossing the subset. While graph learning can take $\tilde{\Theta}(m)$ cut queries, a spanning forest of an undirected graph, unweighted or weighted, can be constructed³ in $O(n)$ queries (see [6, 30]). Can such results be generalized⁴ to hypergraphs? To us, the easiest case of a hypergraph was the hypermatching whose only spanning forest is the hypergraph itself. It is not too hard to see that CUT queries and rank queries are intimately related. Formally, after n cut queries, any rank query can be simulated with 2 cut queries. The interesting open question is: *can the connectivity question of an arbitrary hypergraph be solved in $O(n)$ queries?*

The other related problem is *matroid intersection*. Given rank/independence oracle to two matroids over the same universe, the matroid intersection problem asks to find the largest common independent set. It is a classic result in combinatorial optimization due to [26] that this can be solved in polynomially many independence oracle queries. The current state-of-the-art is that $\tilde{O}(n^{1.5})$ -rank queries suffice (see [19]) and $\tilde{O}(n^{7/4})$ -independence oracle queries suffice (see [10]). On the other hand, no *super-linear* lower bounds are known for rank-queries, and only recently, [11] proved an $\Omega(n \log n)$ -lower bound for independence queries. The big open question is: *can matroid intersection be solved in $O(n)$ rank-queries, or can a $\omega(n)$ -lower bound be proved?*

As noted earlier, if we wish to obtain an $o(n \log k)$ -query algorithm, we must exploit the fact that rank-queries output “more” than the independence oracle queries. Our second motivation in writing this paper is to showcase how the techniques that arise from *coin weighing* problems a la [18, 31] exploit this “more”. In the basic coin-weighing problem, one is asked to recover an unknown Boolean vector x with the ability to query any subset S and obtain $\sum_{i \in S} x_i$ (a sum-query). The aforementioned papers showed how to do this making roughly $2n/\log_2 n$ sum-queries. [13] generalized this to learn a Boolean vector with at most d ones in roughly $\frac{2d \log_2 n}{\log_2 d}$ queries. In a different application, [28] showed how to use the coin-weighing result to learn a hidden perfect matching in a bipartite graph using $2n$ CUT queries. These form the backbone of our algorithms. Having said that, there are some big differences between sum-queries and rank-queries since the latter is not “linear” and this underlies the difficulties we’ve faced in generalizing Theorem 2 to obtain a $O(n)$ -query algorithm to the general partition matroid case.

1.1 Related works

There is a vast literature on combinatorial search [1, 25], and we restrict ourselves to the works that are related the most. As mentioned above, our problem can be thought of as learning a *hypermatching* using additive/cut queries. (Hyper)-graph reconstruction questions have been widely studied in the last two decades. A significant body of work has been dedicated to reconstructing graphs using queries, as evidenced by the works of (cf. [28, 4, 3, 35, 23, 34, 15, 17, 22]). These efforts encompass various types of graphs, including unweighted graphs, graphs with positive weights, and graphs with non-zero edge weights, using CUT queries. This has culminated in a result of [22] gives an polynomial time, randomized $O(\frac{m \log n}{\log m})$ -query algorithm for learning graphs on n nodes and m edges with non-zero edge weights, and this query complexity is information theoretic optimal. Concurrently, there has

³ using a randomized Las Vegas algorithm which makes $O(n)$ queries in expectation

⁴ At first glance even a polynomial query algorithm may not be clear; a little thought can lead to an $O(n \log n)$ query algorithm.

been ongoing research on recovering specific structures within graphs without necessarily reconstructing the entire graph, such as figuring its connectivity (see [28, 20, 30]). [5] started the research on learning a hypergraph using edge-detecting queries, that is, whether the input set contains a hyperedge or not; they described algorithms for r -uniform hypergraphs (every hyperedge has exactly r vertices) but the dependence on r was exponential. [14] considered the *additive* model where one gets the number of edge (this was mentioned in the Introduction above) and proved existence of algorithms to learn rank d hypergraphs (every hyperedge has at most d vertices) for constant d using $O_d(m \log(n^d/m)/\log m)$ -queries; the dependence on d is exponential. [9] considered high-rank but low-degree hypergraphs, including hypermatchings. The focus was on edge-detecting queries (indicator whether additive query is zero or non-zero), and they gave $O(n \text{polylog} n)$ -query algorithms which were also “low depth”, that is, with few rounds of adaptivity.

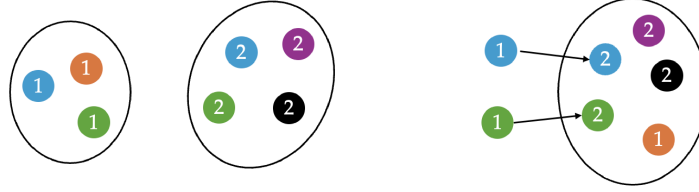
Our problem is also related to the problem of recovering a *clustering* with active queries (see [7, 33, 37, 2, 21, 12, 32]). The setting is the same: there is a universe of n points which we assume is clustered into k unknown parts. The query model, however, is often quite different and much more restrictive usually constraining queries to asking whether a pair or a constant number of elements are in the same cluster/part or not. Such a study was initiated in the works of [24, 7, 33] which prove an $\Omega(nk)$ lower bound, and then provide better upper bounds with extra assumptions. The above-cited works continue on this line.

2 $O(n)$ Query Deterministic Algorithm

Throughout the rest of the paper, unless otherwise mentioned all logarithm base is 2. We begin with an overview of the algorithm. We maintain a collection \mathcal{J} of disjoint independent sets; recall that a subset is independent if it contains at most one element from each part. Initially, \mathcal{J} is the collection of n independent sets each of which is a single element. Let J denote the union of all these independent sets, and so, initially $J = V$. The algorithm will modify this collection \mathcal{J} in iterations, removing some elements from J while doing so. Anytime such an element e is removed, we maintain a map $\text{rep}(e)$ to an element in the current J with the property that e and $\text{rep}(e)$ are in the same partition in \mathcal{P} . We will call two elements in the same parts “friends”, and so, $\text{rep}(e)$ is e ’s friend.

The key routine in the algorithm is a merge operation over independent sets. Given two independent sets I_1, I_2 , define the set of *common nodes* $\text{com}(I_1, I_2) := \{v_1 \in I_1 : \exists v_2 \in I_2, P_i, \{v_1, v_2\} \subseteq P_i\}$ to be the subset of nodes in I_1 which have a friend in I_2 . Note that this friend needs to be unique since I_2 is independent. The MERGE operation takes two independent sets I_1 and I_2 and then (a) finds the set $\text{com}(I_1, I_2)$, (b) for each $e \in \text{com}(I_1, I_2)$, finds its unique neighbor $\text{rep}(e) \in \text{com}(I_2, I_1)$, and (c) returns $\text{com}(I_1, I_2), \text{com}(I_2, I_1)$, and $I_3 := I_1 + I_2 - \text{com}(I_1, I_2)$. See Figure 1 for an illustration.

Given the MERGE routine, the algorithm is very simple: while there exists two independent sets I_1 and I_2 of comparable size (within factor 2), merge them and replace I_1 and I_2 with I_3 returned by the merge. This may remove some elements from J , and in particular this is $\text{com}(I_1, I_2)$, but all these elements will have $\text{rep}(e)$ pointing to their friends who are still in J . When the algorithm can’t do this anymore, there must be at most $\ell = \lceil \log k \rceil$ independent sets remaining in \mathcal{J} . These can be sequentially merged in any order to get one single independent set J , indeed a basis, in \mathcal{J} . To find the partition, consider the directed graph on V where we add the edge $(e, \text{rep}(e))$ for all $e \in V \setminus J$; note this forms a collection of directed in-trees rooted at vertices in J , and the connected components are precisely the parts that



■ **Figure 1** After we merge I_1, I_2 on the left, we get I_3 and a mapping from $\text{com}(I_1, I_2)$ to $\text{com}(I_2, I_1)$.

we desire. In what follows we show how to implement MERGE using existing results from coin-weighing and graph reconstruction, and then argue why the total number of rank queries made by our algorithm is $O(n)$.

2.1 Definitions

We review or introduce several definitions for completeness.

► **Definition 3** (add query). An add query on an unweighted graph $G = (V, E)$: given $S \subseteq V$, obtain $|\{e \in E : e \in S \times S\}|$.

► **Definition 4** (sum query). A sum query on a boolean vector $x \in \{0, 1\}^N$: given $S \subseteq [N]$, obtain $\sum_{i \in S} x_i$.

► **Definition 5** (com of two sets). Given two independent sets I_1, I_2 . The set of common nodes $\text{com}(I_1, I_2) := \{v_1 \in I_1 : \exists v_2 \in I_2, \{v_1, v_2\} \subseteq P_i\}$ is the subset of nodes in I_1 which have a friend in I_2 .

► **Definition 6** (rep(e) of a node). We maintain a map **rep** with the property that a node e and **rep**(e) (representative of e) are in the same partition in \mathcal{P} . **rep** keeps track of the learned partition by mapping the learned node to its friend who is still in J .

2.2 Merging Independent Sets

We begin by introducing some vector/graph reconstruction algorithms from the literature.

► **Lemma 7** ([13]). Let $x \in \{0, 1\}^N$ be an unknown Boolean vector with **sum-query** access. If x has d ones, then there is a polynomial time, adaptive, deterministic algorithm to reconstruct x which makes $O(d \log(N/d)/\log d)$ **sum queries**.

► **Lemma 8** (Paraphrasing Theorem 4 & Section 4.3 [28]). A bipartite graph $G = (V, W, E)$ with $|V| = |W| = m$ where E forms a perfect matching can be learnt in $O(m)$ **add queries**.

Now we are ready to describe MERGE whose properties are encapsulated in the following lemma.

► **Lemma 9.** Let I_1, I_2 be two independent sets and let $k_1 = |I_1|$ and $k_2 = |I_2|$. Suppose $d = |\text{com}(I_1, I_2)| = |\text{com}(I_2, I_1)|$. The procedure MERGE is an adaptive deterministic polynomial time algorithm which returns $I_3 = I_1 + I_2 - \text{com}(I_1, I_2)$ and **rep**(e) $\in \text{com}(I_2, I_1)$ for all $e \in \text{com}(I_1, I_2)$. The procedure makes $O(\frac{d \log(\max(k_1, k_2)/d)}{\log d})$ **rank queries**.

Proof. Given I_1 and I_2 , define the Boolean vector $\mathbf{x} := \mathbf{x}_{(I_1, I_2)} \in \{0, 1\}^{k_1}$ where $\mathbf{x}_e = 1$ if and only if $e \in \text{com}(I_1, I_2)$. We note that a **sum** query can be simulated on \mathbf{x} using a single **rank** query. This is due to the observation that for all $S \subseteq I_1$, $\sum_{e \in S} \mathbf{x}_e = |S| + |I_2| - \text{rank}(S \cup I_2)$. This is because the RHS precisely counts the number of parts of S that are already present in I_2 , or $\text{com}(I_1, I_2) \cap S$. Therefore, we can apply Lemma 7 to learn $\text{com}(I_1, I_2)$ in $O(\frac{d \log(k_1/d)}{\log d})$ many rank queries. Similarly, we can get $\text{com}(I_2, I_1)$ in $O(\frac{d \log(k_2/d)}{\log d})$ queries. Note that the above doesn't give us the friends for $e \in \text{com}(I_1, I_2)$ in $\text{com}(I_2, I_1)$. This pairing can be found as follows. For simplicity, let's use $X := \text{com}(I_1, I_2)$ and $Y := \text{com}(I_2, I_1)$. Consider the bipartite graph $G = (X, Y, E)$ where $e \in X$ has an edge to $f \in Y$ if and only if f is e 's friend. So, G is a perfect matching whose edges are yet unknown. We can now use Lemma 8 to find them. To see why this can be done, note that we can simulate the **add** query because for any $S \subseteq X \cup Y$, simply because $\text{add}(S) = |S| - \text{rank}(S)$. This is because any edge (e, f) with both endpoints in S are precisely the pairs which are counted once in $\text{rank}(S)$ but twice in $|S|$. Thus, finding this matching takes $O(d)$ rank queries. \blacktriangleleft

■ **Algorithm 1** Merging Independent Sets.

```

1: procedure MERGE( $I_1, I_2$ ):
2:   ▷ Input: Two independent sets
3:   ▷ Output:  $\text{com}(I_1, I_2)$  and  $\text{rep}(e) \in \text{com}(I_2, I_1)$  for  $e \in \text{com}(I_1, I_2)$ .
4:   Learn  $\text{com}(I_1, I_2)$  and  $\text{com}(I_2, I_1)$  as described above using
      $O(d \log(\max(k_1, k_2))/\log d)$  rank queries.
5:   Learn  $\text{rep}(e) \in \text{com}(I_2, I_1)$  for  $e \in \text{com}(I_1, I_2)$  as described above in  $O(d)$  rank
     queries.
6:    $I_3 \leftarrow I_1 \cup I_2 - \text{com}(I_1, I_2)$ . ▷ Note that  $I_3$  is independent and  $\text{rep}(e) \in I_3$  for all
      $e \in \text{com}(I_1, I_2)$ .
7:   return ( $I_3, \text{rep}$ )

```

2.3 The algorithm and analysis

We give the pseudocode of the algorithm in Algorithm 2. We now claim that the algorithm makes $O(n)$ queries. All the queries to **rank** occur in the calls to MERGE in line 7 or line 14. Let's take care of the second ones first since it's straightforward.

▷ **Claim 10.** The total number of rank queries made in MERGE calls in line 14 over the for-loop is $O(n)$.

Proof. There are $\ell = O(\log n)$ merges made; that is the only fact we will use. By Lemma 9, the t th MERGE would make at most $O(d_t \log n / \log d_t)$ many rank queries, where d_t is the size of $\text{com}(I_t, I)$ at that time. All we care for is that $\sum_{t=1}^{\ell} d_t \leq n$. Now we observe (an explicit reference is Claim 3 of [20]) that if $\ell \leq C \log n$, then $\sum_{t=1}^{\ell} \frac{d_t}{\log d_t} = O(n / \log n)$. To see this, note that the contribution to this sum of all the d_t 's which are $\leq \frac{n}{C \log^2 n}$ is at most $\frac{n\ell}{C \log^2 n} < n / \log n$. All the other d_t 's have $\log d_t = \Omega(\log n)$ and so their contribution is $O(\sum_t d_t / \log n) = O(n / \log n)$. Altogether, we see that $O(\sum_{t=1}^{\ell} d_t \log n / \log d_t) = O(n)$. \blacktriangleleft

▷ **Claim 11.** The total number of rank queries made in MERGE calls in line 7 over the while-loop is $O(n)$.

■ **Algorithm 2** Find Partition.

```

1: procedure FINDPARTITION( $V, \text{rank}$ ):
2:   ▷ Input:  $n$  elements with rank query access to hidden partition  $\mathcal{P}$ .
3:   ▷ Output: the partition.
4:   Create  $\mathcal{J} \leftarrow \{\{e_1\}, \{e_2\}, \dots, \{e_n\}\}; J \leftarrow V$ 
5:   Create graph  $G = (V, F)$  with  $F \leftarrow \emptyset$ . ▷ this will be used to find the parts
6:   while  $\exists I_1, I_2 \in \mathcal{J} : |I_1|/|I_2| \in [1/2, 2]$  do:
7:      $(I_3, \text{rep}(e)) \leftarrow \text{MERGE}(I_1, I_2)$ .
8:     For all  $e \in \text{com}(I_1, I_2)$ , add  $(e, \text{rep}(e))$  to the edge-set  $F$ .
9:      $\mathcal{J} \leftarrow \mathcal{J} - \{I_1, I_2\} + I_3; J \leftarrow J \setminus \text{com}(I_1, I_2)$ .
10:  ▷ At this point there can be at most  $\lceil \log n \rceil$  elements in  $\mathcal{J}$ 
11:  ▷ Merge all these sets in any order to get a single set. We provide one below.
12:  Let  $\mathcal{J} = \{I_1, I_2, \dots, I_\ell\}$  with  $\ell \leq \lceil \log n \rceil; I \leftarrow I_1; \mathcal{J} \leftarrow \mathcal{J} \setminus I_1$ 
13:  for  $2 \leq t \leq \ell$  do:
14:     $(I_3, \text{rep}(e)) \leftarrow \text{MERGE}(I_t, I)$ .
15:    For all  $e \in \text{com}(I_t, I)$ , add  $(e, \text{rep}(e))$  to the edge-set  $F$ .
16:     $\mathcal{J} \leftarrow \mathcal{J} - \{I_t, I\} + I_3; I \leftarrow I_3$ .
17:  ▷ At this point  $\mathcal{J}$  has a single independent set  $I$ . Every element in  $e \in V \setminus I$  has a single representative  $\text{rep}(e)$ . So  $G$  is a collection of directed in-trees with roots in  $I$ 
18:  return Connected components of  $G$ .

```

Proof. To argue about the MERGE's in Line 7, we need to partition these into two classes. Note that all such merges take two independent sets I_1 and I_2 which are of similar size k_1 and k_2 respectively; without loss of generality, let $k_1 \leq k_2 \leq 2k_1$. Let $d := |\text{com}(I_1, I_2)|$. We call a merge *thick* if $d \geq \sqrt{k_1}$ and *thin* otherwise. We argue about the thick and thin merges differently.

- Using Lemma 9, we see that a thick merge costs $O(d \log(\max(k_1, k_2))/\log d) = O(d)$ rank queries; we have used here that $k_2 \leq 2k_1$ and $d \geq \sqrt{k_1}$. Thus, we can *charge* these rank queries to the d elements which *leave* J . Thus, the total number of rank queries made across all thick merges is $O(n)$.
- To argue about thin merges, we make a further definition. Let us say that an independent set I is in class t if $|I| \in [2^t, 2^{t+1})$, for $0 \leq t \leq \lceil \log n \rceil$. Fix such a t . A thin merge (I_1, I_2) is called a class t thin-merge if the smaller cardinality set is in class t . An element $e \in V$ participates in a class t thin-merge (I_1, I_2) if it is present in the smaller set. Observe that for a thin class t merge, the resulting independent set I_3 almost doubles in size; in particular, $|I_3| = |I_1| + |I_2| - |\text{com}(I_1, I_2)| \geq 2^{t+1} - 2^{t/2}$. Using this one can argue that the same element cannot participate in more than *two* class t -thin merges; after two merges the set ceases to be class t . In particular, this means the number of thin class t merges is at most $2 \cdot n/2^t$, and each such merge, by Lemma 9, can be done with $O(d \log(2^{t+1})/\log d)$ many rank queries where $d = |\text{com}(I_1, I_2)| < 2^{t/2}$. Since $d/\log d$ is an increasing function of d , we conclude that any class t thin-merge takes at most $O(2^{t/2} \log(2^{t+1})/\log(2^{t/2})) = O(2^{t/2})$ many rank queries. Therefore, the total number of rank queries made within thin merges is at most $\sum_{t=0}^{\log n} \frac{2n}{2^t} \cdot O(2^{t/2}) = O(n)$ ◀

The above two claims imply the proof of Theorem 1.

3 General Partition Matroids

We recall the problem. As before, the universe is V and there is a hidden partition $\mathcal{P} = (P_1, \dots, P_k)$. Furthermore, there are integers r_1, \dots, r_k where $0 < r_i < |P_i|$.⁵ This defines a partition matroid where a set I is independent if and only if $|I \cap P_i| \leq r_i$ for $1 \leq i \leq k$. The rank-oracle for this matroid is the following $\text{rank}(S) = \sum_{i=1}^k \min(|S \cap P_i|, r_i)$. We will prove the following theorem in this section.

► **Theorem 2.** *There is a deterministic, constructive algorithm that learns a general partition matroid using $O(n + k \log r)$ rank queries where $r := \text{rank}(V) = \sum_i r_i$.*

Our proof technique will be a *reduction* to the simple partition matroid setting of Section 2. Before we get there, let's first begin with a simple well-known observation.

► **Lemma 12.** *There is an $O(n)$ rank query algorithm that finds a basis B of a partition matroid.*

Proof. This is standard and we give it below for completeness. Note that although described as a “for-loop”, the above algorithm can be implemented in a single round of n many rank queries. ◀

■ **Algorithm 3** Finding a Basis Using Rank Queries.

```

1: procedure FINDBASIS( $V, \text{rank}$ ):
2:   ▷ Input:  $n$  elements in  $V$  with rank query access
3:   ▷ Output: A basis of  $V$ .
4:    $B \leftarrow \{\}$ .
5:   for  $v \in V$  do:
6:     if  $\text{rank}(B + \{v\}) = \text{rank}(B) + 1$  then:
7:        $B \leftarrow B + \{v\}$ .
8:   return  $B$ .
```

To obtain our reduction, what we need apart from this basis B are two sets of *representatives*. A subset $T \subseteq V$ is a set of representative if $|T \cap P_i| = 1$ for each $1 \leq i \leq k$. The reduction will need *two* representative sets: $T_1 \subseteq B$ and $T_2 \cap B = \emptyset$, and the subroutine FINDREPRESENTATIVES(B) will find this. Furthermore, it will also return a map $\phi: T_1 \rightarrow T_2$ where for each $e \in T_1$, $\phi(e)$ belongs to the same part as e . The algorithm does so in $O(n + k \log r)$ queries; in fact, only *independence oracle* queries suffice. This is slightly non-trivial and we described this in Section 3.1. Let us now show how these representatives imply an $O(n)$ -query algorithm to learn the partition \mathcal{P} and the r_i 's.

► **Claim 13.** Algorithm 4 returns the correct partition \mathcal{P} and r_i 's making $O(n)$ many rank queries.

Proof. The main idea is that the representatives allow us to simulate a simple partition matroid rank query on the basis and outside. More precisely, we claim that for any subset $S \subseteq B$, $\text{rank}_1(S) = \sum_{i=1}^k \min(|S \cap P_i|, 1)$. If so, the correctness of Algorithm 4 follows

⁵ Suppose we allow $r_i \geq |P_i|$. Let $M := \{i | r_i \geq |P_i|\}$. Now $\text{rank}(S) = \sum_{i \in M} |S \cap P_i| + \sum_{i \notin M} \min(|S \cap P_i|, r_i)$. This means we get no information for partitions with index in M . To see this, we pick $x_1 \in P_{i_1}, x_2 \in P_{i_2}$ with $i_1, i_2 \in M$ and $i_1 \neq i_2$. If we swap x_1 with x_2 in every set we give to the rank-oracle, the answer it returns is the same. Thus no rank query algorithm can tell x_1, x_2 apart.

■ **Algorithm 4** Using Representatives to Learn Partition.

```

1: procedure LEARNMATROIDWITHREPS( $V, \text{rank}, B, T_1, T_2, \phi : T_1 \rightarrow T_2$ ):
2:    $\triangleright$  Input:  $n$  elements in  $V$  with rank query; basis  $B$ , set of representatives  $T_1 \subseteq B$ ,  $T_2 \cap B = \emptyset$ ,  $\phi(t)$  is a friend of  $t$ .
3:    $\triangleright$  Output: the partition  $\mathcal{P}$ .
4:   For any subset  $S \subseteq B$ , define  $\text{rank}_1(S) := \text{rank}(B - S + T_2) - \text{rank}(B - S)$ .
5:    $\mathcal{P}_1 \leftarrow \text{FINDPARTITION}(B, \text{rank}_1) \triangleright$  Takes  $O(|B|)$  queries.
6:   For each  $1 \leq i \leq k$ ,  $r_i \leftarrow |B \cap P_i^{(1)}|$  where  $\mathcal{P}_1 = (P_1^{(1)}, \dots, P_k^{(1)})$ .
7:   For any subset  $S \subseteq V \setminus B$ , define  $\text{rank}_2(S) := \text{rank}(B + S - T_1) - \text{rank}(B - T_1)$ .
8:    $\mathcal{P}_2 \leftarrow \text{FINDPARTITION}(V \setminus B, \text{rank}_2) \triangleright$  Takes  $O(|V \setminus B|)$  queries.
9:   Use  $\phi$  to merge  $\mathcal{P}_1$  and  $\mathcal{P}_2$  into  $\mathcal{P}$ : for  $t \in T_1$ , merge the part in  $\mathcal{P}_1$  containing  $T_1$ 
      with the part in  $\mathcal{P}_2$  containing  $\phi(t)$ .
10:  return  $(\mathcal{P}, \{r_i\}_{i=1}^k)$ 

```

from Theorem 1. Indeed, $\text{rank}(B - S + T_2) - \text{rank}(B - S)$ gives $+1$ for each part where $B - S$ loses at least one element to which the unique element of T_2 contributes. Similarly, one argues that for any $S \subseteq V \setminus B$, $\text{rank}_2(S) = \sum_{i=1}^k \min(|S \cap P_i|, 1)$. This is also for a similar reason; $B - T_1$ loses exactly one element from each part and so $\text{rank}(B + S - T_1) - \text{rank}(B - T_1)$ counts the parts that S intersects at least once. See Figure 2 for an illustration. \triangleleft

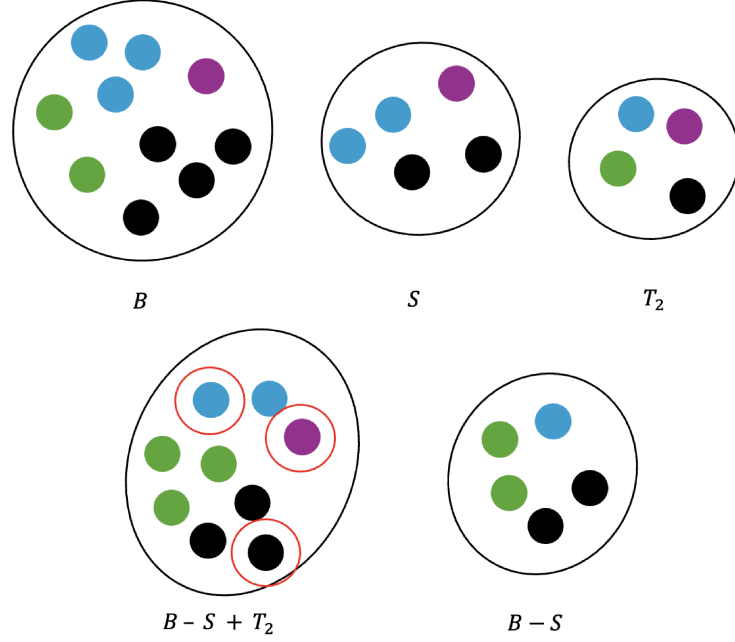
3.1 Finding Representatives via Binary Search

We now describe a procedure which takes a basis B of the partition matroid, and finds two subsets of representatives $T_1 \subseteq B$ and $T_2 \cap B = \emptyset$. The idea behind is a delicate binary search. Fix some $e \in V \setminus B$ and without loss of generality, say $e \in P_1$. We now show how to find one element in $B \cap P_1$ in $O(\log r)$ queries. The way to do it is by halving B to $X_1 \sqcup X_2$ and keep the half with at least P_1 element in it as “search half”. This can be checked by seeing whether $\text{rank}(X_1 + \{e\}) = \text{rank}(X_1)$: if so then X_1 contains all element of P_1 and this is the half we stick with; otherwise X_2 contains some P_1 element and takes precedence. The other half is now added to the new “test half”. We keep searching in our search set until it has exactly 1 class P_1 element left. For instance, say X_2 is further divided to X_{21} and X_{22} . The next query would be to check if $\text{rank}((X_1 + \{e\}) + X_{21}) = \text{rank}(X_1 + X_{21})$. If so, then $X_1 + X_{21}$ contains all class P_1 elements, and so will make X_{21} our new “search set”; otherwise, we continue on X_{22} . We describe the pseudocode in detail in Algorithm 5.

\triangleright **Claim 14.** Algorithm 5 returns (T_1, T_2, ϕ) correctly and makes $O(n + k \log r)$ many rank queries.

Proof. The proof is by induction: we claim that T_1, T_2 contains at most one element from each part and the size $|T_1| = |T_2|$ equals the number of parts spanned by the elements seen by the outer for-loop. And furthermore, the ϕ -relation is correct. This is obviously true before anything occurs, and consider the for-loop for an element e . Now suppose the if-statement in line 6 is *not* true; that is, say $\text{rank}(B - T_1 + e) > \text{rank}(B - T_1)$. This would mean that e contains a friend in T_1 ; the only way the rank could increase is if e filled the “hole” in the part which has exactly one element missing in $B - T_1$. We discard this e . On the other hand if the if-statement holds, then we will discover a new part in B and thus by inductive hypothesis, in $V \setminus B$. We therefore add e to T_2 .

Now consider the invariant in line 10. If that indeed holds true, then when the while-loop terminates, and it does so with $|X| = 1$, the single element $x \in X$ must be in the same part as e . Thus, adding $x \in T_1$ and setting $\phi(x) = e$ is the correct thing to do. To see that the



■ **Figure 2** Illustration of how we simulate a simple partition matroid rank query inside of a basis with representatives outside. We have a basis with b_i s equal to 1 (purple nodes), 2 (green), 3 (blue) and 4 (black) respectively. $\text{rank}(B) = 10$, $\text{rank}(B - S) = 5$. Note $|B - S + T_2| = 10$, yet $\text{rank}(B - S + T_2) = 9$ because the number of green nodes is capped at 2. $\text{rank}(B - S + T_2) - \text{rank}(B - S) = 3$ simulate a simple partition matroid rank query for S . The circled nodes correspond to the 3 partitions included in S .

invariant in line 10 holds, we include the invariant in line 11. This is readily checked in both the “then” and “else” case of the forthcoming if-statement. If line 13 holds true, then akin to the argument above, $Y + X_1$ contains all friends of e in B . So, we focus our search on X_1 since it contains at least one friend of e because, by invariant, X contained at least one friend of e . So setting X to X_1 keeps the invariant satisfied. On the other hand, if line 13 doesn’t hold true, then X_2 must contain at least one friend of e . And so setting X to X_2 keeps the invariant fulfilled.

To find the number of queries is simple. First notice that only line 6 and 13 make any queries. And even then one of them is superfluous. More precisely, since $B - T_1$ and $Y + X_1$ are independent sets, their rank is $|B| - |T_1|$ and $|Y| + |X_1|$ respectively. We make $n - r$ queries in line 13. Of these at most k many satisfy the condition. Each of them leads to a binary-search style argument which takes at most $\lceil \log r \rceil$ many queries. \triangleleft

► **Remark.** Note that line 6 and line 13 can be implemented using only independence oracle queries since they are really asking, respectively, if $B - T_1 + e$ and $Y + X_1 + e$ are independent or not; if the ranks are equal, they are not. This also implies an $O(n \log k)$ algorithm to learn the partition matroid using only independence oracle as alluded to in the Introduction. Let $|T_1| = |T_2| = k$. Once we have the representative sets $T_1 \subseteq B$ and $T_2 \cap B = \emptyset$, for any element $e \notin V \setminus B$, we can use a binary-search style argument on T_1 to find e ’s friend among T_1 in $O(\log k)$ many independence oracle queries. More precisely, we halve T_1 into (X, Y) and check if $B - X + e$ is independent or not. If it is, then X contains e ’s friend; otherwise, Y does. Similarly, for any $e \in B$, we can find e ’s friend in T_2 in $O(\log k)$ many independence oracle queries.

■ **Algorithm 5** Finding Representatives.

```

1: procedure FINDREPRESENTATIVES( $V, \text{rank}, B$ ):
2:   ▷ Input:  $n$  elements in  $V$  with rank query; basis  $B$ 
3:   ▷ Output: Sets of Representatives  $T_1 \subseteq B$ ,  $T_2 \cap B = \emptyset$  and map  $\phi : T_1 \rightarrow T_2$ .
4:    $T_1, T_2 \leftarrow \emptyset$ .
5:   for  $e \in V \setminus B$  do:
6:     if  $\text{rank}(B - T_1 + \{e\}) = \text{rank}(B - T_1)$  then: ▷  $e$  is an element with no friends in  $T_1$ 
       and  $T_2$ :
7:        $T_2 \leftarrow T_2 + e$ .
8:        $X \leftarrow B$ ;  $Y \leftarrow \emptyset$ .
9:       while  $|X| > 1$  do:
10:        ▷ Invariant:  $X$  has at least one element in same part as  $e$ 
11:        ▷ Invariant:  $X \cup Y = B$ 
12:         $(X_1, X_2) \leftarrow$  arbitrary equipartition of  $X$ .
13:        if  $\text{rank}(Y + X_1 + e) = \text{rank}(Y + X_1)$  then: ▷  $X_2$  contains no friends of  $e$ 
14:           $Y \leftarrow Y + X_2$ ;  $X \leftarrow X_1$ .
15:        else: ▷  $X_2$  contains at least one friend of  $e$ 
16:           $Y \leftarrow Y + X_1$ ;  $X \leftarrow X_2$ .
17:        ▷  $X$  is a singleton element of  $B$ ; let  $X = \{x\}$ 
18:         $T_1 \leftarrow T_1 + x$ ; Set  $\phi(x) = e$ .
19:   return  $(T_1, T_2, \phi)$ .

```

■ **Algorithm 6** Learning a Partition Matroid.

```

1: procedure LEARNPARTITION( $V, \text{rank}$ ):
2:   ▷ Input: partition matroid on  $n$  elements in  $V$  with rank query
3:   ▷ Output: the partition  $\mathcal{P}$  and  $r_i$ 's
4:   Learn a basis  $B$  using FINDBASIS( $V, \text{rank}$ ) a la Algorithm 3.
5:    $(T_1, T_2, \phi) \leftarrow$  FINDREPRESENTATIVES ( $V, B, \text{rank}$ ) a la Algorithm 5.
6:   return  $(\mathcal{P}, \{r_i\}) \leftarrow$  LEARNMATROIDWITHREPS( $V, \text{rank}, B, T_1, T_2, \phi$ ) a la Al-
     gorithm 4.

```

For completeness, we end the section by giving the pseudocode for the final algorithm in Algorithm 6. Lemma 12 establishes that Algorithm 6 makes n rank queries, Claim 14 establishes that Algorithm 6 makes $n + k \log r$ rank (in fact independence oracle) queries, and Claim 13 establishes that Algorithm 6 makes $O(n)$ rank queries. This completes the proof of Theorem 2.

4 Conclusion

In this paper we looked at the question of learning a hidden partition using rank queries which given a subset tells how many different parts it hits. We gave a simple but non-trivial, deterministic, and efficient algorithm which makes $O(n)$ -rank queries. This is optimal up to constant factors. The main non-triviality arises in the use of techniques devised in coin-weighing algorithms a la [18, 31], and our work falls in a growing line of such results [28, 23, 14, 6, 20, 30] which explores the use of these techniques to solve combinatorial search problems.

The obvious question left open by our paper is whether there are $O(n)$ algorithms to learn general partition matroids especially when $k = \Theta(n)$. We have not been able to directly port the coin-weighting techniques to solve this problem even in the case of $r_i = 2$ for all i . The main technical challenge that the rank query, ultimately, is not a linear query and in Section 3 we could make it “behave linear” with the help of representatives. Our algorithm to find representatives, however, didn’t utilize the “more information” given by rank-queries over independence oracle queries. Investigating this may lead to new algorithmic primitives. On the other hand, perhaps there is a $\omega(n)$ lower bound for this problem when $k = \Theta(n)$.

References

- 1 Martin Aigner. *Combinatorial search*. John Wiley & Sons, Inc., 1988.
- 2 Nir Ailon, Anup Bhattacharya, and Ragesh Jaiswal. Approximate correlation clustering using same-cluster queries. In *Proc., Latin American Theoretical Informatics Symposium*, pages 14–27, 2018. doi:10.1007/978-3-319-77404-6_2.
- 3 Noga Alon and Vera Asodi. Learning a hidden subgraph. *SIAM Journal on Discrete Mathematics (SIDMA)*, 18(4):697–712, 2005. doi:10.1137/S0895480103431071.
- 4 Noga Alon, Richard Beigel, Simon Kasif, Steven Rudich, and Benny Sudakov. Learning a hidden matching. *SIAM Journal on Computing (SICOMP)*, 33(2):487–501, 2004. doi:10.1137/S0097539702420139.
- 5 Dana Angluin and Jiang Chen. Learning a hidden hypergraph. In *Proc., Conf. on Learning Theory (COLT)*, pages 561–575. Springer, 2005. doi:10.1007/11503415_38.
- 6 Simon Apers, Yuval Efron, Pawel Gawrychowski, Troy Lee, Sagnik Mukhopadhyay, and Danupon Nanongkai. Cut query algorithms with star contraction. *Proc., IEEE Conference on the Foundations of Computer Science (FOCS)*, 2022.
- 7 Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. *Adv. in Neu. Inf. Proc. Sys. (NeurIPS)*, 29, 2016.
- 8 Arinta Auza and Troy Lee. On the query complexity of connectivity with global queries. *arXiv preprint arXiv:2109.02115*, 2021. arXiv:2109.02115.
- 9 Eric Balkanski, Oussama Hanguir, and Shatian Wang. Learning low degree hypergraphs. In *Proc., Conf. on Learning Theory (COLT)*, pages 419–420. PMLR, 2022. URL: <https://proceedings.mlr.press/v178/balkanski22a.html>.
- 10 Joakim Blikstad. Breaking $O(nr)$ for Matroid Intersection. In *Proc., International Conference on Algorithms, Logic, and Programming (ICALP)*, pages 31:1–31:17, 2021. doi:10.4230/LIPICS.ICALP.2021.31.
- 11 Joakim Blikstad, Sagnik Mukhopadhyay, Danupon Nanongkai, and Ta-Wei Tu. Fast algorithms via dynamic-oracle matroids. In *Proc., ACM Symposium on the Theory of Computing (STOC)*, pages 1229–1242, 2023. doi:10.1145/3564246.3585219.
- 12 Marco Bressan, Nicolò Cesa-Bianchi, Silvio Lattanzi, and Andrea Paudice. On margin-based cluster recovery with oracle queries. *Adv. in Neu. Inf. Proc. Sys. (NeurIPS)*, pages 25231–25243, 2021.
- 13 Nader H. Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. In *Proc., Conf. on Learning Theory (COLT)*, 2009.
- 14 Nader H. Bshouty and Hanna Mazzawi. Optimal Query Complexity for Reconstructing Hypergraphs. In *Proc., Symposium on the Theoretical Aspects of Computer Science (STACS)*, pages 143–154, 2010. doi:10.4230/LIPICS.STACS.2010.2496.
- 15 Nader H. Bshouty and Hanna Mazzawi. Algorithms for the coin weighing problems with the presence of noise. *Electron. Colloquium Comput. Complex.*, page 124, 2011. URL: <https://eccc.weizmann.ac.il/report/2011/124>, arXiv:TR11-124.
- 16 Nader H. Bshouty and Hanna Mazzawi. On parity check $(0, 1)$ -matrix over \mathbb{Z}_p . In *Proc., ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1383–1394, 2011.

- 17 Nader H. Bshouty and Hanna Mazzawi. Toward a deterministic polynomial time algorithm with optimal additive query complexity. *Theoretical Computer Science*, 417:23–35, 2012. doi:10.1016/J.TCS.2011.09.005.
- 18 David G. Cantor and W. H. Mills. Determination of a subset from certain combinatorial properties. *Canadian Journal of Mathematics*, 18:42–48, 1966.
- 19 Deeparnab Chakrabarty, Yin Tat Lee, Aaron Sidford, Sahil Singla, and Sam Chiu-wai Wong. Faster matroid intersection. In *Proc., IEEE Conference on the Foundations of Computer Science (FOCS)*, pages 1146–1168, 2019. doi:10.1109/FOCS.2019.00072.
- 20 Deeparnab Chakrabarty and Hang Liao. A query algorithm for learning a spanning forest in weighted undirected graphs. In *Proc., International Conference on Algorithmic Learning Theory (ALT)*, pages 259–274, 2023. URL: <https://proceedings.mlr.press/v201/chakrabarty23a.html>.
- 21 I Eli Chien, Huozhi Zhou, and Pan Li. hs^2 : Active learning over hypergraphs with pointwise and pairwise queries. In *Proc., International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2466–2475, 2019. URL: <http://proceedings.mlr.press/v89/chien19a.html>.
- 22 Sung-Soon Choi. Polynomial time optimal query algorithms for finding graphs with arbitrary real weights. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proc., Conf. on Learning Theory (COLT)*, volume 30, pages 797–818, 2013. URL: <http://proceedings.mlr.press/v30/Choi13.html>.
- 23 Sung-Soon Choi and Jeong Han Kim. Optimal query complexity bounds for finding graphs. *Artif. Intell.*, 174(9-10):551–569, 2010. doi:10.1016/J.ARTINT.2010.02.003.
- 24 Susan Davidson, Sanjeev Khanna, Tova Milo, and Sudeepa Roy. Top-k and clustering with noisy comparisons. *ACM Transactions on Database Systems (TODS)*, 39(4):1–39, 2014. doi:10.1145/2684066.
- 25 Dingzhu Du and Frank K Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 2000.
- 26 Jack Edmonds. Submodular functions, matroids, and certain polyhedra. *Combinatorial Structures and their Applications*, 18:69–87, 1970.
- 27 Andrei Graur, Tristan Pollner, Vidhya Ramaswamy, and S Matthew Weinberg. New query lower bounds for submodular function minimization. *Proc., Innovations in Theoretical Computer Science (ITCS)*, page 64, 2020.
- 28 Vladimir Grebinski and Gregory Kucherov. Optimal reconstruction of graphs under the additive model. *Algorithmica*, 28(1):104–124, 2000. doi:10.1007/S004530010033.
- 29 Troy Lee, Miklos Santha, and Shengyu Zhang. Quantum algorithms for graph problems with cut queries. In *Proc., ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 939–958, 2021. doi:10.1137/1.9781611976465.59.
- 30 Hang Liao and Deeparnab Chakrabarty. Learning spanning forests optimally in weighted undirected graphs with cut queries. In *Proc., International Conference on Algorithmic Learning Theory (ALT)*, 2024.
- 31 Bernt Lindström. On a combinatorial problem in number theory. *Canadian Mathematical Bulletin*, 8(4):477–490, 1965.
- 32 Xizhi Liu and Sayan Mukherjee. Tight query complexity bounds for learning graph partitions. In *Proc., Conf. on Learning Theory (COLT)*, pages 167–181. PMLR, 2022. URL: <https://proceedings.mlr.press/v178/liu22a.html>.
- 33 Arya Mazumdar and Barna Saha. Query complexity of clustering with side information. *Adv. in Neu. Inf. Proc. Sys. (NeurIPS)*, 2017.
- 34 Hanna Mazzawi. Optimally reconstructing weighted graphs using queries. In *Proc., ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 608–615, 2010. doi:10.1137/1.9781611973075.51.

- 35 Lev Reyzin and Nikhil Srivastava. Learning and verifying graphs using queries with a focus on edge counting. In *Proc., International Conference on Algorithmic Learning Theory (ALT)*, pages 285–297. Springer, 2007. doi:10.1007/978-3-540-75225-7_24.
- 36 Aviad Rubinfeld, Tselil Schramm, and S. Matthew Weinberg. Computing exact minimum cuts without knowing the graph. In *Proc., Innovations in Theoretical Computer Science (ITCS)*, pages 39:1–39:16, 2018. doi:10.4230/LIPICS.ITCS.2018.39.
- 37 Barna Saha and Sanjay Subramanian. Correlation clustering with same-cluster queries bounded by optimal cost. In *Proc., European Symposium on Algorithms*, pages 81:1–81:17, 2019. doi:10.4230/LIPICS.ESA.2019.81.