Maximum Unique Coverage on Streams: Improved FPT Approximation Scheme and Tighter Space Lower Bound

Philip Cervenjak
School of Computing and Information Systems, The University of Melbourne, Australia

Junhao Gan
School of Computing and Information Systems, The University of Melbourne, Australia

Seeun William Umboh
School of Computing and Information Systems, The University of Melbourne, Australia
ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Melbourne, Australia

Anthony Wirth
School of Computer Science, The University of Sydney, Australia
School of Computing and Information Systems, The University of Melbourne, Australia

Abstract

We consider the Max Unique Coverage problem, including applications to the data stream model. The input is a universe of n elements, a collection of m subsets of this universe, and a cardinality constraint, k. The goal is to select a subcollection of at most k sets that maximizes unique coverage, i.e, the number of elements contained in exactly one of the selected sets. The Max Unique Coverage problem has applications in wireless networks, radio broadcast, and envy-free pricing.

Our first main result is a fixed-parameter tractable approximation scheme (FPT-AS) for Max Unique Coverage, parameterized by k and the maximum element frequency, r, which can be implemented on a data stream. Our FPT-AS finds a $(1-\varepsilon)$-approximation while maintaining a kernel of size $\tilde{O}(kr/\varepsilon)$, which can be combined with subsampling to use $\tilde{O}(k^2r/\varepsilon^3)$ space overall. This significantly improves on the previous-best FPT-AS with the same approximation, but a kernel of size $\tilde{O}(k^2r/\varepsilon^2)$. In order to achieve our first result, we show upper bounds on the ratio of a collection's coverage to the unique coverage of a maximizing subcollection; this is by constructing explicit algorithms that find a subcollection with unique coverage at least a logarithmic ratio of the collection's coverage. We complement our algorithms with our second main result, showing that $\Omega(m/k^2)$ space is necessary to achieve a $(1.5+o(1))/(\ln k - 1)$-approximation in the data stream. This dramatically improves the previous-best lower bound showing that $\Omega(m/k^2)$ is necessary to achieve better than a $e^{-1+1/k}$-approximation.

2012 ACM Subject Classification Theory of computation → Parameterized complexity and exact algorithms; Theory of computation → Approximation algorithms analysis; Theory of computation → Streaming, sublinear and near linear time algorithms

Keywords and phrases Maximum unique coverage, maximum coverage, approximate kernel, data streams

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2024.25

Category APPROX

Related Version *Full Version*: https://arxiv.org/abs/2407.09368

Funding *Philip Cervenjak*: This work was supported by the Elizabeth and Vernon Puzey Scholarship, and by the Faculty of Engineering and Information Technology.

Acknowledgements We thank the anonymous reviewers for their valuable feedback.

1 Introduction

We study the **Max Unique Coverage** problem, where we are given a universe of n elements, a collection of m subsets of the universe, and an integer $k \in \{1, \ldots, m\}$. The goal is to select a collection of at most k subsets that maximizes the number of elements covered by *exactly* one set in the collection. This problem is a natural variant of the classic **Max Coverage** problem, where the goal is to select a collection of k subsets that maximizes the number of elements covered by *at least* one set in the collection.

A weighted version of **Max Unique Coverage** was first formally studied by Demaine et al. [9]. In their motivating scenario, a number of wireless base stations, each with an associated cost, must be placed to maximize the number of mobile clients served. However, due to interference, if covered by more than one base station, a client receives bad service. Demaine et al. point out further applications to radio broadcast and envy-free pricing. They then showed an offline polynomial-time $\Omega(1/\log m)$-approximation algorithm for their problem, which easily translates to a $\Omega(1/\log k)$-approximation for our problem.[1] Under various complexity assumptions, they showed (semi-)logarithmic inapproximability for polynomial-time algorithms; Guruswami and Lee [11] later proved nearly logarithmic inapproximability, assuming NP does not admit quasipolynomial-time algorithms.

Streaming. Our work emphasizes solving **Max Unique Coverage** approximately in the data stream model. All previous works, except McGregor et al. [15], only consider this problem in the offline model. In the data stream model, we focus on *set-streaming*: each set in the stream is fully specified before the next; this setting is assumed in related works [18, 2, 22, 16, 15]. We also constrain the space, measured in bits, to be $o(mn)$, i.e., sublinear in both the number of sets, m, and the size of the universe, n. Thus, we define the **Max Unique Coverage** problem to include the cardinality constraint, k. Previous works often formulate this problem without a cardinality constraint, simply referring to it as the "Unique Coverage" problem; this is equivalent to our formulation when $k = m$.

We are particularly interested in **Max Unique Coverage** when parameterized by k and the *maximum frequency*, r, defined as the maximum number of sets that an element belongs to; we also consider the *maximum set size* d to a lesser extent. This parameterization has received considerable attention in studying fixed-parameter tractable approximation schemes (FPT-AS) for classic coverage problems [21, 4, 20, 14, 15, 13, 19], but not as much for the **Max Unique Coverage** problem [15]. Note that an FPT exact algorithm for this parameterization is unlikely to exist because, when $r = 2$, **Max Unique Coverage** is equivalent to Capacitated Max Cut, which was shown by Misra et al. [17] to be W[1]-hard when parameterized by the capacity constraint.[2]

A central idea in achieving both FPT space and running time bounds is *kernelization*. We transform a problem instance, $\mathcal{I}$, into a smaller problem instance, $\mathcal{I}'$, called the *(approximate) kernel*, such that $|\mathcal{I}'| \leq g(\boldsymbol{\gamma})$, where g is a computable function in terms of problem parameters $\boldsymbol{\gamma}$, while $\mathcal{I}'$ (approximately) preserves the optimal solution value of $\mathcal{I}$; a good solution can be found by brute-force search within $\mathcal{I}'$. Consistent with other works [8, 7, 6, 15], we further require an FPT streaming algorithm to use $g(\boldsymbol{\gamma}) \operatorname{polylog} |\mathcal{I}|$ space.

---

[1] This is by assuming that all sets have unit cost and that the budget is k.

[2] Although Misra et al. [17] prove W[1]-hardness for Budgeted Max Cut when parameterized by the budget, their hardness proof only requires each vertex (corresponding to a set in our formulation) to have unit cost.

1.1 Our Contributions

Our first main result is an FPT-AS for **Max Unique Coverage** with strong space, running time, and approximation bounds, that is applicable to the data stream model. A crucial step in achieving such bounds is showing improved upper bounds on a key parameter of a collection $\mathcal{C}$. The *unique coverage ratio* is the ratio between the coverage of $\mathcal{C}$ and the maximum unique coverage over all subcollections $\mathcal{Q} \subseteq \mathcal{C}$. We let ϕ denote an *upper bound* on the unique coverage ratio: we define performance bounds of our FPT-AS in terms of ϕ.

Main Result 1: FPT Approximation Scheme. We propose the FPT-AS UniqueTopSets, parameterized by the cardinality constraint, k, and the maximum frequency, r, which can be easily implemented in the data stream model. It achieves a $(1-\varepsilon)$-approximation using a kernel of size $\lceil kr(\phi+1)/\varepsilon \rceil$. We formally present this algorithm in Theorem 3.6.

UniqueTopSets is a refined version of the FPT-AS in Theorem 12 of McGregor et al. [15], in that our algorithm achieves a $(1-\varepsilon)$ rather than a $(1/2-\varepsilon)$-approximation using only an extra logarithmic factor of $(\phi+1)$ in the kernel size. Further, our algorithm improves on the FPT-AS in Theorem 10 of McGregor et al. [15] by saving a factor of $O(k/\varepsilon)$ in the kernel size, and therefore a factor of $[O(k/\varepsilon)]^k$ in the running time, while achieving the same approximation factor. See Table 1 for a comparison of our FPT-AS with others.

Table 1 Comparison of FPT-AS for **Max Unique Coverage**, parameterized by cardinality constraint, k, and maximum frequency, r. Note that the running time of each algorithm below is implied by its kernel size. Each finds a solution of size at most k by brute-force search in the kernel. Below, we can assign $\phi = \min(\ln k + 1, 2\ln r + o(\log r), 2\ln d + o(\log d))$.

Reference	Approx.	Kernel Size
[15, Theorem 10]	$1-\varepsilon$	$O\left(k^2 r \log m/\varepsilon^2\right)$
[15, Theorem 12]	$1/2-\varepsilon$	$\lceil kr/\varepsilon \rceil$
Ours, Theorem 3.6 (UniqueTopSets)	$1-\varepsilon$	$\lceil kr(\phi+1)/\varepsilon \rceil$

Unique Coverage Algorithms. In order to show good values for ϕ, we propose a number of offline polynomial-time algorithms that, given an arbitrary $\mathcal{C}$, explicitly return a $\mathcal{B} \subseteq \mathcal{C}$ whose unique coverage is at least a logarithmic ratio of $\mathcal{C}$'s coverage. We refer to them as *unique coverage algorithms*; in fact, they can be thought as approximation algorithms for the unconstrained Unique Coverage problem on an input instance of $\mathcal{C}$.

Our three offline polynomial-time algorithms, UniqueGreedy, UniqueGreedyFreq, and UniqueGreedySize, each take a collection of sets, $\mathcal{C}$, and return a collection, $\mathcal{B} \subseteq \mathcal{C}$, whose unique coverage is at least a $1/(\ln \ell + 1)$, $1/(2\ln r + o(\log r))$, and $1/(2\ln d + o(\log d))$ proportion of $\mathcal{C}$'s coverage respectively; in this context, $\ell = |\mathcal{C}|$, r is the maximum frequency in $\mathcal{C}$, and d is the maximum set size in $\mathcal{C}$. We formally present these algorithms in Theorem 4.1, Theorem 4.4, and in Theorem 4.8, respectively. See Table 2 for a comparison of our algorithms with those of Demaine et al. [9] along with their implied bounds, ϕ, albeit weaker than ours.

Table 2 Polynomial-time algorithms for **Max Unique Coverage**. Compared to others, our methods imply constant-factor improvements in the unique coverage ratio bound, ϕ.

Parameter	Reference	(Implied) ϕ
ℓ = collection size	[9, Theorem 4.1]	$10.66 \ln(\ell+1)$
	Ours, Theorem 4.1 (UniqueGreedy)	$\ln \ell + 1$
r = maximum frequency in a collection	[9, Theorem 4.1]	$10.66 \ln(r+1)$
	Ours, Theorem 4.4 (UniqueGreedyFreq)	$2 \ln r + o(\log r)$
d = maximum set size in a collection	[9, Theorem 4.2]	$21.32 \ln(d+1)$
	Ours, Theorem 4.8 (UniqueGreedySize)	$2 \ln d + o(\log d)$

Implication for FPT Approximation Scheme. The bound on the unique coverage ratio, ϕ, affects the kernel size and therefore the brute-force running time of UniqueTopSets. In particular, when $r = \Omega(\sqrt{k})$, the bound of ϕ implied by UniqueGreedy is 10.66 times smaller than implied by Demaine et al. [9]; whereas when $r = o(\sqrt{k})$, the bound of ϕ implied by UniqueGreedyFreq is almost 5.33 smaller than implied by Demaine et al. This means, by using our implied bounds rather than those implied by Demaine et al., we save a factor of 10.66^k in UniqueTopSets's running-time when $r = \Omega(\sqrt{k})$, and a factor of almost 5.33^k when $r = o(\sqrt{k})$.

Improvements in Polynomial-Time Approximation. As a separate contribution, each of our three unique coverage algorithms finds a logarithmic approximation to **Max Unique Coverage**, both offline and in the data stream. We first find a solution $\mathcal{C}$ to **Max Coverage** in polynomial time, and then run one of our above algorithms on $\mathcal{C}$ to return the subcollection $\mathcal{B} \subseteq \mathcal{C}$. For this purpose, our algorithms UniqueGreedy, UniqueGreedyFreq, and UniqueGreedySize improve the approximation factor due to Demaine et al. [9] by a factor of 10.66, 5.33, and 10.66, respectively. Following the above approach, we propose a single-pass streaming algorithm for **Max Unique Coverage** that achieves a $(1/(2\phi) - \varepsilon)$-approximation using $\tilde{O}(k^2/\varepsilon^3)$ space, where we can assign $\phi = \min(\ln k + 1, 2\ln r + o(\log r), 2\ln d + o(\log d))$. We formally state this in Theorem 3.8.

Main Result 2: Streaming Lower Bound. Our second main result is a significantly improved streaming lower bound for **Max Unique Coverage**. In the data stream model, we prove that any randomized algorithm that achieves a $(1.5 + o(1))/(\ln k - 1)$-approximation for **Max Unique Coverage** w.h.p. requires $\Omega(m/k^2)$ space. We formally state this in Theorem 5.1. Our lower bound improves on the lower bound by McGregor et al. [15], which shows a similar result, but achieves w.h.p. a $e^{-1+1/k} \geq 1/e$-approximation. Interestingly, our approximation threshold is close to 3 times larger than the approximation (in terms of k) achieved by our $\tilde{O}(k^2/\varepsilon^3)$ space algorithm in Theorem 3.8, indicating that a dramatic increase in space is needed to bridge this approximation gap.

Table 3 Comparison of space lower bounds for **Max Unique Coverage** in the data stream. Note that the lower bound by Assadi [1] was shown for **Max Coverage** with constant $k = 2$, but it is not difficult to adapt it for **Max Unique Coverage** because, in the hard instance constructed for the lower bound, the unique coverage of any pair of sets behaves similarly to its coverage.

Reference	Approx.	Space LB
[1, Theorem 4]	$1 - \varepsilon$	$\Omega\left(m/\varepsilon^2\right)$
[15, Theorem 16]	$1/e$	$\Omega\left(m/k^2\right)$
Ours, Theorem 5.1	$(1.5 + o(1))/(\ln k - 1)$	$\Omega\left(m/k^2\right)$

1.2 Technical Overview

FPT Approximation Scheme. UniqueTopSets refines the technique used in the FPT-AS for **Max Unique Coverage** in Theorem 12 of McGregor et al. [15], which is to construct an approximate kernel by storing a number of the largest sets by individual size, and then to find a subcollection of the kernel with maximum unique coverage by brute-force search. Similar techniques have been used in FPT-AS approaches for Max Vertex Cover [14, 13] and **Max Coverage** [21, 20, 15, 19]. Our novelty is providing a stronger analysis of the approximation factor preserved by the kernel, allowing us to achieve a $(1 - \varepsilon)$-approximation while only increasing the kernel size by a logarithmic factor in k, r, or d.

Unique Coverage Algorithms. All of our unique coverage algorithms are combinatorial in design. Our first two, UniqueGreedy and UniqueGreedyFreq, are novel algorithms that each, in some sense, use a greedy approach, noting that UniqueGreedy is used as subroutine of UniqueGreedyFreq. Our third algorithm, UniqueGreedySize, is easily derived by combining UniqueGreedyFreq with the approach by Demaine et al. [9] for sets with maximum cost d (maximum size in our case).

Streaming Lower Bound. Our streaming lower bound relies on a novel reduction from k-player Set Disjointness in the one-way communication model to **Max Unique Coverage** in the data stream. In the hard instance of **Max Unique Coverage** thus constructed, either all collections of $\ell \leq k$ sets have a unique coverage of $ak^2(1.5 + o(1))$ w.h.p. or there exists a single collection of k sets whose unique coverage is at least $ak^2(\ln k - 1)$, where $a = \Omega(k \log m)$. By a standard argument, we show that distinguishing between these instances of **Max Unique Coverage** with a streaming algorithm is as hard as solving Set Disjointness, implying the required space lower bound.

1.3 Paper Structure

After preliminaries in Section 2, Section 3 presents our FPT-AS UniqueTopSets and a polynomial-time algorithm, both applicable to the data stream. In Section 4, we present our component algorithms for bounding the unique coverage ratio. In Section 5, we present a space lower bound for achieving a $(1.5 + o(1))/(\ln k - 1)$-approximation for **Max Unique Coverage**. We conclude in Section 7. Claims whose proofs are found in the full version of this paper are marked thus: (*).

2 Preliminaries

Notation. For convenience, we hence let $[n]$ denote the set of integers $\{1, 2, \ldots, n\}$. Likewise, $U = [n]$ denotes a universe of n elements, while $\mathcal{V}$ denotes a collection of m subsets of U.

Given a collection $\mathcal{C}$ of sets, the *unique cover* of $\mathcal{C}$ is the subset of the universe covered by exactly one set in $\mathcal{C}$, formally, $\tilde{\psi}(\mathcal{C}) := (\bigcup_{S\in\mathcal{C}} S) \setminus (\bigcup_{S\neq T\in\mathcal{C}} S \cap T)$, and the *unique coverage* of $\mathcal{C}$ is $|\tilde{\psi}(\mathcal{C})|$. For convenience, the *cover* of $\mathcal{C}$ is the union of the sets in $\mathcal{C}$, formally $\psi(\mathcal{C}) := \bigcup_{S\in\mathcal{C}} S$, and the *coverage* of $\mathcal{C}$ is $|\psi(\mathcal{C})|$. Further, the *non-unique cover* of $\mathcal{C}$ is the subset of the universe covered by at least two sets from $\mathcal{C}$, formally $\psi_{\geq 2}(\mathcal{C}) = \bigcup_{S\neq T\in\mathcal{C}} S \cap T$ – or equivalently $\psi_{\geq 2}(\mathcal{C}) = \psi(\mathcal{C}) \setminus \tilde{\psi}(\mathcal{C})$ – and the *non-unique coverage* of $\mathcal{C}$ is $|\psi_{\geq 2}(\mathcal{C})|$. The *maximum unique coverage* of $\mathcal{C}$ is the largest unique coverage of a subcollection of $\mathcal{C}$. The *unique coverage ratio* of $\mathcal{C}$ is the ratio between its coverage and maximum unique coverage. In other words, if $\mathcal{Q}$ is the subcollection of $\mathcal{C}$ that has maximum unique coverage, then the unique coverage ratio of $\mathcal{C}$ is $|\psi(\mathcal{C})|/|\tilde{\psi}(\mathcal{Q})|$.

Given an element $x \in U$ and a collection $\mathcal{C}$ of sets, the *frequency* of x in $\mathcal{C}$ is defined as $\mathrm{freq}_{\mathcal{C}}(x) := |\{S \in \mathcal{C} : x \in S\}|$, i.e., the number of sets in $\mathcal{C}$ that contain x; and the *maximum frequency* is defined as $r := \max_{x\in U} \mathrm{freq}_{\mathcal{C}}(x)$. Also, the *maximum set size* is defined as $d := \max_{S\in\mathcal{C}} |S|$. We often use r and d to refer to the maximum frequency and set size, respectively, in $\mathcal{C} = \mathcal{V}$ unless stated otherwise. Note that $r \leq |\mathcal{C}|$ holds for every $\mathcal{C}$. We let $H_z := \sum_{t=1}^{z} 1/t$ denote the z^{th} harmonic number, a term that appears several times.

Formal Problem Definition. An instance of **Max Unique Coverage** consists of an element universe U, a collection $\mathcal{V}$ of m subsets of U, and an integer $k \in [m]$; when the context is clear, we represent an instance with just $\mathcal{V}$ for simplicity. The goal of **Max Unique Coverage** is to return a subcollection $\mathcal{B} \subseteq \mathcal{V}$ (more precisely, a collection of IDs of sets), with $|\mathcal{B}| \leq k$, that maximizes $|\tilde{\psi}(\mathcal{B})|$. We let $\mathcal{O}$ denote an optimal solution to this **Max Unique Coverage** problem, and $\mathrm{OPT} := |\tilde{\psi}(\mathcal{O})|$ as the maximum unique coverage.

Subsampling for the Data Stream Model. The universe subsampling technique has been widely successful in the development of streaming algorithms for coverage problems [10, 12, 3, 16]. In this work, we follow the approach of McGregor and Vu [16], and sample the universe so that each set has size $O(k \log m/\varepsilon^2)$. We assume that $k \in o(mn)$, and also that k is known prior to reading the stream. The main result is given in the following lemma, with a proof sketch of the subsampling approach in Section 6.

▶ **Lemma 2.1** (Subsampling Approach [15]). *Let $\varepsilon \in (0, 1)$ be the subsampling error parameter. Given an instance of **Max Unique Coverage** and an α-approximation streaming algorithm, we can run the algorithm on $\lceil \log_2 n \rceil$ parallel subsampled instances and select one of them such that the algorithm's solution corresponds to a $(\alpha - 2\varepsilon)$-approximation for the original instance with probability $1 - 1/\mathrm{poly}(m)$. Moreover, if the streaming algorithm stores at most s sets in every subsampled instance, then the total space complexity of the subsampling approach is bounded by $\lceil \log_2 n \rceil \cdot s \cdot O\left(k \log m \log n/\varepsilon^2\right)$.*

3 Streaming FPT-AS and Polynomial-Time Algorithms

In Section 3.1, we prove a kernelization lemma. Then, we use it to obtain an FPT-AS and a parameterized streaming algorithm in Section 3.2. Finally, we show how to use a bound on the unique coverage ratio to obtain a polynomial-time streaming algorithm in Section 3.3.

3.1 Kernelization Lemma

Our Kernelization Lemma below, as well as its proof, is a refinement of Lemma 11 by McGregor et al. [15]. We first provide some intuition on why our kernel preserves a $(1-\varepsilon)$-approximation for **Max Unique Coverage**.

Intuition of Kernelization Lemma. For convenience, let ε' be an intermediate error parameter and define the kernel $\mathcal{A}$ as the collection of the $\lceil kr/\varepsilon' \rceil$ largest sets in instance $\mathcal{V}$ by individual size. Given the optimal solution for **Max Unique Coverage**, $\mathcal{O}$, let $\mathcal{O}^{\text{in}}$ and $\mathcal{O}^{\text{out}}$ be the collections of optimal sets found and not found in $\mathcal{A}$ respectively.

One main step in proving our Kernelization Lemma is showing that, in expectation, a collection of $|\mathcal{O}^{\text{out}}|$ sets sampled without replacement from $\mathcal{A}$, denoted by $\mathcal{Z}$, can be appended to $\mathcal{O}^{\text{in}}$ with little overlap in their unique covers. In particular, we can prove that $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})|] \geq (1-\varepsilon')|\tilde{\psi}(\mathcal{O})| - \varepsilon'|\psi(\mathcal{O})|$.

However, due to the $\varepsilon'|\psi(\mathcal{O})|$ term, this is not enough to achieve the required approximation factor. This term reflects the fact that, even if the unique cover of $\mathcal{O}^{\text{in}}$ has little overlap with the unique cover of $\mathcal{Z}$, the *entire* cover of $\mathcal{O}^{\text{in}}$ could be more extensive and, thus, overlap significantly with the unique cover of $\mathcal{Z}$. To address this, in Claim 3.4, we show $\phi|\tilde{\psi}(\mathcal{O})| \geq |\psi(\mathcal{O})|$, where ϕ upper bounds the unique coverage ratio. Substituting this into the lower bound for $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})|]$, and assigning $\varepsilon' = \varepsilon/(\phi+1)$, we obtain $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})|] \geq (1-\varepsilon)|\tilde{\psi}(\mathcal{O})|$, implying the existence of a $(1-\varepsilon)$-approximate subcollection of $\mathcal{A}$. Lastly, the final kernel size of $|\mathcal{A}| = \lceil kr(\phi+1)/\varepsilon \rceil$ follows from the assignment of ε'.

▶ **Lemma 3.1** (Kernelization Lemma). *Suppose that every collection of sets has unique coverage ratio at most ϕ. Let $\mathcal{V}$ denote a collection of sets. Then, for every $\varepsilon \in (0,1)$, the subcollection, $\mathcal{A}$, of the $\lceil kr(\phi+1)/\varepsilon \rceil$-largest sets of $\mathcal{V}$ (by size) contains a subcollection of at most k sets with unique coverage at least $(1-\varepsilon)\mathrm{OPT}$.*

Proof. Assume that $|\mathcal{V}| \geq \lceil kr(\phi+1)/\varepsilon \rceil$: otherwise, $\mathcal{A}$ would contain every set in $\mathcal{V}$ and so would trivially have $\mathcal{O}$ as a subcollection. Let $\mathcal{O}^{\text{in}} = \mathcal{O} \cap \mathcal{A}$ and $\mathcal{O}^{\text{out}} = \mathcal{O} \setminus \mathcal{A}$. Let $\mathcal{Z}$ be a uniform random sample of $|\mathcal{O}^{\text{out}}|$ sets chosen from $\mathcal{A}$ without replacement. The main goal is to prove Claim 3.5, below. Since $\mathcal{O}^{\text{in}}$ and $\mathcal{Z}$ are subsets of $\mathcal{A}$, this implies the existence of subcollection $\mathcal{B} \subseteq \mathcal{A}$ as required by the lemma.

We start with the following lower bound on the expected unique coverage of $\mathcal{O}^{\text{in}} \cup \mathcal{Z}$, as shown in inequality (1) below. Then we lower bound each of the RHS terms separately and simplify afterwards. By definition,

$$|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})| \geq |\tilde{\psi}(\mathcal{O}^{\text{in}})| + |\tilde{\psi}(\mathcal{Z})| - (|\tilde{\psi}(\mathcal{O}^{\text{in}}) \cap \psi(\mathcal{Z})| + |\psi(\mathcal{O}^{\text{in}}) \cap \tilde{\psi}(\mathcal{Z})|)\,,$$

hence, by linearity of expectation,

$$\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})|] \geq |\tilde{\psi}(\mathcal{O}^{\text{in}})| + \mathbb{E}[|\tilde{\psi}(\mathcal{Z})|] - \mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}}) \cap \psi(\mathcal{Z})|] - \mathbb{E}[|\psi(\mathcal{O}^{\text{in}}) \cap \tilde{\psi}(\mathcal{Z})|]\,. \quad (1)$$

Define an intermediate error parameter, $\varepsilon' = \varepsilon/(\phi+1)$, meaning $|\mathcal{A}| = \lceil kr/\varepsilon' \rceil$. The probability of a set $S \in \mathcal{A}$ being selected in $\mathcal{Z}$ is $p := |\mathcal{O}^{\text{out}}|/|\mathcal{A}| \leq k/(kr/\varepsilon') = \varepsilon'/r$. Now Claim 3.2, below, is easily derived from the proof of Lemma 11 in by McGregor et al. [15].

▷ Claim 3.2. It holds that $\mathbb{E}[|\tilde{\psi}(\mathcal{Z})|] \geq (1-\varepsilon')|\tilde{\psi}(\mathcal{O}^{\text{out}})|\,.$

Proof. Quantity $|\tilde{\psi}(\mathcal{Z})|$ can be lower bounded by summing, over every $S \in \mathcal{Z}$, the number of elements in S not contained in any other $T \in \mathcal{Z} \setminus \{S\}$. From there, we prove inequality (3.2), below. We let $[\mathcal{E}]$ denote the indicator variable for event $\mathcal{E}$.

$$
\begin{aligned}
&|\tilde{\psi}(\mathcal{Z})| \geq \sum_{S\in\mathcal{Z}}\left(|S| - \sum_{T\in\mathcal{Z}\setminus\{S\}} |S\cap T|\right), \quad \text{hence,} \\
&\mathbb{E}[|\tilde{\psi}(\mathcal{Z})|] \\
&\geq \mathbb{E}\left[\sum_{S\in\mathcal{A}}\left(|S|[S\in\mathcal{Z}] - \sum_{T\in\mathcal{A}\setminus\{S\}} |S\cap T|[S\in\mathcal{Z}\wedge T\in\mathcal{Z}]\right)\right] \\
&\geq \sum_{S\in\mathcal{A}}\left(|S|p - \sum_{T\in\mathcal{A}\setminus\{S\}} |S\cap T|p^2\right) && \Pr[S\in\mathcal{Z}\wedge T\in\mathcal{Z}] \leq p^2 \\
&\geq \sum_{S\in\mathcal{A}}\left(|S|p - |S|p^2(r-1)\right) \geq p(1-pr)\sum_{S\in\mathcal{A}}|S| && \text{each } x\in S \text{ intersects } \leq r-1 \text{ other sets} \\
&\geq p(1-\varepsilon')\sum_{S\in\mathcal{A}}|S| && p\leq \frac{\varepsilon'}{r} \\
&\geq p(1-\varepsilon')|\mathcal{A}|\frac{\sum_{Y\in\mathcal{O}^{\text{out}}}|Y|}{|\mathcal{O}^{\text{out}}|} && \text{for all } S\in\mathcal{A} \text{ and all } Y\in\mathcal{O}^{\text{out}}\colon |S|\geq|Y| \\
&\geq p(1-\varepsilon')|\mathcal{A}|\frac{|\tilde{\psi}(\mathcal{O}^{\text{out}})|}{|\mathcal{O}^{\text{out}}|} && \text{subadditivity of } \tilde{\psi} \\
&= p(1-\varepsilon')\frac{|\tilde{\psi}(\mathcal{O}^{\text{out}})|}{p} = (1-\varepsilon')|\tilde{\psi}(\mathcal{O}^{\text{out}})|\,.
\end{aligned}
$$

◁

Claim 3.3 upper bounds the expected size of the overlap between $\tilde{\psi}(\mathcal{O}^{\text{in}})$ and $\psi(\mathcal{Z})$ and the expected size of the overlap between $\tilde{\psi}(\mathcal{Z})$ and $\psi(\mathcal{O}^{\text{in}})$.

▷ Claim 3.3. $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}})\cap\psi(\mathcal{Z})|] \leq \varepsilon'|\tilde{\psi}(\mathcal{O}^{\text{in}})|$ and $\mathbb{E}[|\psi(\mathcal{O}^{\text{in}})\cap\tilde{\psi}(\mathcal{Z})|] \leq \varepsilon'|\psi(\mathcal{O}^{\text{in}})|$.

Proof. To prove the first inequality,

$$
\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}})\cap\psi(\mathcal{Z})|] \leq \sum_{x\in\tilde{\psi}(\mathcal{O}^{\text{in}})}\ \sum_{S\in\mathcal{A}\colon x\in S}\Pr[S\in\mathcal{Z}] \leq \sum_{x\in\tilde{\psi}(\mathcal{O}^{\text{in}})} rp \leq \varepsilon'|\tilde{\psi}(\mathcal{O}^{\text{in}})|\,.
$$

To prove the second inequality, it is clear that $\tilde{\psi}(\mathcal{Z})\subseteq\psi(\mathcal{Z})$ for all $\mathcal{Z}$, so we have $\mathbb{E}[|\psi(\mathcal{O}^{\text{in}})\cap\tilde{\psi}(\mathcal{Z})|] \leq \mathbb{E}[|\psi(\mathcal{O}^{\text{in}})\cap\psi(\mathcal{Z})|]$. Then substituting $\psi(\mathcal{O}^{\text{in}})$ for $\tilde{\psi}(\mathcal{O}^{\text{in}})$ in the argument for the first inequality, we see that $\mathbb{E}[|\psi(\mathcal{O}^{\text{in}})\cap\psi(\mathcal{Z})|] \leq \varepsilon'|\psi(\mathcal{O}^{\text{in}})|$. ◁

We now turn to a property of the optimal solution for **Max Unique Coverage**, $\mathcal{O}$.

▷ Claim 3.4. $\phi|\tilde{\psi}(\mathcal{O})| \geq |\psi(\mathcal{O})|\,.$

Proof. Recall that we assumed that every collection of sets has unique coverage ratio at most ϕ. In particular, $\mathcal{O}$ has a subcollection, $\mathcal{Q}$, of at most k sets with $\phi|\tilde{\psi}(\mathcal{Q})| \geq |\psi(\mathcal{O})|$. By optimality, $\mathcal{O}$'s unique coverage is at least that of $\mathcal{Q}$. Thus, we get the desired inequality. ◁

Starting from Ineq. (1), we can now lower bound $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}}\cup\mathcal{Z})|]$, thus proving the lemma.

▷ Claim 3.5. We have the lower bound $\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}}\cup\mathcal{Z})|] \geq (1-\varepsilon)|\tilde{\psi}(\mathcal{O})|\,.$

Proof.

$$
\begin{aligned}
&\mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}} \cup \mathcal{Z})|] \\
&\geq |\tilde{\psi}(\mathcal{O}^{\text{in}})| + \mathbb{E}[|\tilde{\psi}(\mathcal{Z})|] - \mathbb{E}[|\tilde{\psi}(\mathcal{O}^{\text{in}}) \cap \psi(\mathcal{Z})|] - \mathbb{E}[|\psi(\mathcal{O}^{\text{in}}) \cap \tilde{\psi}(\mathcal{Z})|] && \text{Ineq. (1)} \\
&\geq |\tilde{\psi}(\mathcal{O}^{\text{in}})| + (1-\varepsilon')|\tilde{\psi}(\mathcal{O}^{\text{out}})| - \varepsilon'|\tilde{\psi}(\mathcal{O}^{\text{in}})| - \varepsilon'|\psi(\mathcal{O}^{\text{in}})| && \text{Claims 3.2 and 3.3} \\
&= (1-\varepsilon')\left(|\tilde{\psi}(\mathcal{O}^{\text{in}})| + |\tilde{\psi}(\mathcal{O}^{\text{out}})|\right) - \varepsilon'|\psi(\mathcal{O}^{\text{in}})| \\
&\geq (1-\varepsilon')|\tilde{\psi}(\mathcal{O})| - \varepsilon'|\psi(\mathcal{O}^{\text{in}})| && \text{subadditivity of } \tilde{\psi} \\
&\geq (1-\varepsilon')|\tilde{\psi}(\mathcal{O})| - \varepsilon'|\psi(\mathcal{O})| && \text{monotonicity of } \psi \\
&\geq (1-\varepsilon')|\tilde{\psi}(\mathcal{O})| - \varepsilon'\phi|\tilde{\psi}(\mathcal{O})| && \text{Claim 3.4} \\
&= (1-\varepsilon'(1+\phi))\,|\tilde{\psi}(\mathcal{O})| \\
&= (1-\varepsilon)|\tilde{\psi}(\mathcal{O})|\,. && \triangleleft
\end{aligned}
$$

◀

3.2 Applications of the Kernelization Lemma

We now apply the Kernelization Lemma to prove the following theorem.

▶ **Theorem 3.6.** *Suppose that every collection of sets has unique coverage ratio at most ϕ. Let $\mathcal{V}$ denote a collection of sets, $k \geq 2$ denote the cardinality constraint, $r \geq 2$ denote the maximum frequency in $\mathcal{V}$, and $\varepsilon \in (0,1)$ denote an error parameter. Then, there exists*

1. *an FPT-AS that finds a $(1-\varepsilon)$-approximation for **Max Unique Coverage** and has a running time of $(er(\phi+1)/\varepsilon)^k \operatorname{poly}(m,n,1/\varepsilon)$; and*
2. *a streaming algorithm that finds a $(1-3\varepsilon)$-approximation for **Max Unique Coverage** with probability $1 - 1/\operatorname{poly}(m)$ and uses $\tilde{O}(\phi k^2 r/\varepsilon^3)$ space.*

Our algorithm, UniqueTopSets, takes a collection of sets $\mathcal{V}$ with maximum frequency $r \geq 2$, a cardinality constraint k, and an error parameter $\varepsilon \in (0,1)$, and returns a $(1-\varepsilon)$-approximation for **Max Unique Coverage**. It also takes parameter ϕ, an upper bound on the unique coverage ratio of every subcollection. UniqueTopSets first finds $\mathcal{A}$, the $\lceil kr(\phi+1)/\varepsilon \rceil$-largest sets $S \in \mathcal{V}$ by size $|S|$. Then, it brute-forces over $\mathcal{A}$, i.e., it finds the subcollection of $\mathcal{A}$ containing at most k sets and has the maximum unique coverage.

FPT-AS. Let us first see how UniqueTopSets has the properties of the FPT-AS claimed in Theorem 3.6. The Kernelization Lemma (Lemma 3.1) immediately implies that the solution returned by UniqueTopSets is a $(1-\varepsilon)$-approximation. The running time bound follows by bounding the number of subcollections of $\mathcal{A}$ containing at most k sets.

▶ **Lemma 3.7.** *UniqueTopSets has running time in $(er(\phi+1)/\varepsilon)^k \operatorname{poly}(m,n,1/\varepsilon)$.*

Proof. UniqueTopSets considers every possible collection of $\ell \in [k]$ sets from $\mathcal{A}$ and outputs the one with the best unique coverage. Below, the second inequality holds since replacing ℓ with k makes each binomial coefficient larger, as $\ell \leq k \leq kr/2$ due to $r \geq 2$; the equality holds since $\binom{z+1}{k} = \binom{z}{k}(z+1)/(z+1-k)$; and the final inequality holds since $\binom{z}{k} \leq (ez/k)^k$. Thus, the running time is bounded as follows.

$$
\sum_{\ell=1}^{k} \binom{|\mathcal{A}|}{\ell} \operatorname{poly}(m,n) \leq \operatorname{poly}(m,n) \sum_{\ell=1}^{k} \binom{\frac{kr(\phi+1)}{\varepsilon}+1}{\ell} \leq \operatorname{poly}(m,n)k \binom{\frac{kr(\phi+1)}{\varepsilon}+1}{k}
$$

$$
= \operatorname{poly}(m,n,1/\varepsilon) \binom{\frac{kr(\phi+1)}{\varepsilon}}{k} \leq \operatorname{poly}(m,n,1/\varepsilon) \left(\frac{er(\phi+1)}{\varepsilon}\right)^k . \qquad ◀
$$

Streaming Algorithm. UniqueTopSets can also be run on a data stream using the subsampling approach from Lemma 2.1. UniqueTopSets returns a $(1-\varepsilon)$-approximation in each subsampled instance by Lemma 3.1, implying a $(1-3\varepsilon)$-approximation for the original instance, $\mathcal{V}$, with probability $1-1/\operatorname{poly}(m)$. Further, it stores $|\mathcal{A}| \leq \lceil kr(\phi+1)/\varepsilon \rceil$ sets in each subsampled instance, implying an overall space complexity of $\lceil \log_2 n \rceil \cdot |\mathcal{A}| \cdot \tilde{O}(k/\varepsilon^2) = \tilde{O}(\phi k^2 r/\varepsilon^3)$.

3.3 Polynomial-Time Streaming Algorithm

Here we present a single-pass streaming algorithm that returns a $(1/(2\phi)-\varepsilon)$-approximation for **Max Unique Coverage**, given a bound on the unique coverage ratio, ϕ. We present the algorithm in Theorem 3.8 below.

In the theorem statement, we assume we can use an offline polynomial-time algorithm, Alg, that takes a collection $\mathcal{C}$ and returns a subcollection $\mathcal{B} \subseteq \mathcal{C}$ such that $|\tilde{\psi}(\mathcal{B})| \geq |\psi(\mathcal{C})|/\phi$ for a ratio ϕ depending on $|\mathcal{C}| \leq k$, the maximum frequency r, and the maximum set size d. Alg can be substituted with a procedure that runs all of our unique coverage algorithms from Section 4 on $\mathcal{C}$ and returns the solution with the best unique coverage.

▶ **Theorem 3.8.** *Let $\mathcal{V}$ denote a data stream of m sets, $k \geq 2$ denote a cardinality constraint, $r \geq 2$ denote the maximum frequency in $\mathcal{V}$, $d \geq 2$ denote the maximum set size in $\mathcal{V}$, and $\varepsilon \in (0,1)$ denote an error parameter. Further, assume we have a polynomial-time algorithm Alg with unique coverage ratio ϕ depending on k, r, and d. Then we can find a $(1/(2\phi)-3\varepsilon)$-approximation for **Max Unique Coverage** with probability $1-1/\operatorname{poly}(m)$, using one pass, $\tilde{O}(k^2/\varepsilon^3)$ space, and in polynomial-time.*

Proof. We use the subsampling approach from Lemma 2.1. In each subsampled instance, we use an existing polynomial-time streaming algorithm [16] to find a $(1/2-\varepsilon)$-approximation, $\mathcal{C}$, for **Max Coverage** in one pass while storing $\tilde{O}(k/\varepsilon)$ sets; the sets in $\mathcal{C}$ must be stored explicitly so that we can run Alg on $\mathcal{C}$. Running Alg on $\mathcal{C}$ returns a $\mathcal{B} \subseteq \mathcal{C}$ that is a $(1/(2\phi)-\varepsilon)$-approximation for the subsampled instance of **Max Unique Coverage**. This implies a $(1/(2\phi)-3\varepsilon)$-approximation for the original instance, $\mathcal{V}$, with probability $1-1/\operatorname{poly}(m)$. Further, explicitly storing $\tilde{O}(k/\varepsilon)$ sets in each subsampled instance implies an overall space complexity of $\lceil \log_2 n \rceil \cdot \tilde{O}(k/\varepsilon) \cdot \tilde{O}(k/\varepsilon^2) = \tilde{O}(k^2/\varepsilon^3)$. ◀

4 Algorithms for Bounding the Unique Coverage Ratio

We here present algorithms that run in polynomial time. Given a collection, $\mathcal{C}$, each returns a subcollection $\mathcal{B} \subseteq \mathcal{C}$ such that $\mathcal{B}$'s unique coverage is within a logarithmic ratio of $\mathcal{C}$'s coverage. We hence call this the *unique coverage ratio* of an algorithm. Our algorithms UniqueGreedy (Section 4.1), UniqueGreedyFreq (Section 4.2), and UniqueGreedySize (Section 4.3) have unique coverage ratios that are logarithmic in $\ell = |\mathcal{C}|$, r, and d, respectively.

4.1 UniqueGreedy

We present and analyze our algorithm UniqueGreedy, with pseudocode in Algorithm 1. Its purpose is to take a collection $\mathcal{C}$ of ℓ sets and return a collection $\mathcal{B} \subseteq \mathcal{C}$ whose *unique* coverage is at least a $1/H_\ell$ factor of $\mathcal{C}$'s coverage. We formally state this in Theorem 4.1.

UniqueGreedy Overview. UniqueGreedy first checks whether $\mathcal{C}$'s unique coverage is at least $1/H_\ell$ of its own coverage. If so, then it immediately returns $\mathcal{C}$ as the solution, which of course occurs if $\ell = 1$. If not, then the idea is to discard the set $T \in \mathcal{C}$ with the smallest contribution to $\mathcal{C}$'s unique coverage. It follows that the total loss in coverage from $\mathcal{C}$ to $\mathcal{C} \setminus \{T\}$ is only $1/\ell$ of $\mathcal{C}$'s unique coverage. Observe that T contributes at most $1/\ell$ to $\mathcal{C}$'s unique coverage, and any elements in T that are also in $\mathcal{C}$'s non-unique cover must remain in $\mathcal{C} \setminus \{T\}$'s cover. We then apply UniqueGreedy recursively, to $\mathcal{C} \setminus \{T\}$. As we show in Theorem 4.1, since the performance of UniqueGreedy relates unique coverage to coverage, $\mathcal{C} \setminus \{T\}$ has sufficient coverage so that the recursive solution from UniqueGreedy$(\mathcal{C} \setminus \{T\})$ has a unique coverage of at least $1/H_\ell$ of $\mathcal{C}$'s coverage.

Algorithm 1 UniqueGreedy.

```
   Input: C: collection of ℓ sets.
   Output: B ⊆ C: subcollection satisfying |ψ̃(B)| ≥ |ψ(C)|/H_ℓ.
1  if |ψ̃(C)| ≥ |ψ(C)|/H_ℓ then
2      B ← C
3  else
4      T ← arg min_{S∈C} |S ∩ ψ̃(C)|
5      B ← UniqueGreedy(C \ {T})
6  return B
```

▶ **Theorem 4.1.** *Given a collection of ℓ sets, $\mathcal{C}$, UniqueGreedy returns a collection $\mathcal{B} \subseteq \mathcal{C}$ satisfying*

$$|\tilde{\psi}(\mathcal{B})| \geq \frac{|\psi(\mathcal{C})|}{H_\ell} . \tag{2}$$

Proof. We prove Theorem 4.1 by induction on $\ell = |\mathcal{C}|$.

Base Case. If $\ell = 1$, then $|\tilde{\psi}(\mathcal{C})| = |\psi(\mathcal{C})|$ and $\mathcal{B} = \mathcal{C}$, so we are done.

Inductive Case. Consider the case $\ell \geq 2$, and assume that Theorem 4.1 holds for $\ell - 1$. Then one of two subcases must hold: (i) $|\tilde{\psi}(\mathcal{C})| \geq |\psi(\mathcal{C})|/H_\ell$; or (ii) the negation, $|\tilde{\psi}(\mathcal{C})| < |\psi(\mathcal{C})|/H_\ell$. In subcase (i), the Line 1 condition succeeds and UniqueGreedy returns the subcollection $\mathcal{B} = \mathcal{C}$, which clearly satisfies Ineq. (2).

So, we focus on subcase (ii); since $|\tilde{\psi}(\mathcal{C})| < |\psi(\mathcal{C})|/H_\ell$, the Line 1 condition fails, thus Line 5 assigns to $\mathcal{B}$ the solution from the recursive call on $\mathcal{C} \setminus \{T\}$. Claim 4.3 lower bounds the coverage of this subcollection, $|\psi(\mathcal{C} \setminus \{T\})|$. Prior to that, we prove a handy claim.

▷ Claim 4.2. $|\psi(\mathcal{C} \setminus \{T\})| = |\psi_{\geq 2}(\mathcal{C})| + |\tilde{\psi}(\mathcal{C}) \setminus T|$.

Proof. Observe that $\psi_{\geq 2}(\mathcal{C})$ and $\tilde{\psi}(\mathcal{C}) \setminus T$ are disjoint; so it suffices to show that $\psi(\mathcal{C} \setminus \{T\}) = \psi_{\geq 2}(\mathcal{C}) \cup (\tilde{\psi}(\mathcal{C}) \setminus T)$. We first show that RHS is a subset of LHS. Each element covered at least twice in $\mathcal{C}$ remains covered in $\mathcal{C} \setminus \{T\}$; while each element uniquely covered in $\mathcal{C}$ that is not in T remains covered in $\mathcal{C} \setminus \{T\}$. Going the other way, consider an element that is in neither $\psi_{\geq 2}(\mathcal{C})$ nor $\tilde{\psi}(\mathcal{C}) \setminus T$: then the only set it was in was T, and hence it is not in $\mathcal{C} \setminus \{T\}$. ◁

▷ Claim 4.3. Subcollection $\mathcal{C} \setminus \{T\}$ satisfies

$$|\psi(\mathcal{C} \setminus \{T\})| \geq \left(1 - \frac{1}{\ell H_\ell}\right) |\psi(\mathcal{C})| .$$

Proof. First, observe that the contribution of each $S \in \mathcal{C}$ to $\tilde{\psi}(\mathcal{C})$, i.e., $|S \cap \tilde{\psi}(\mathcal{C})|$, is disjoint from the contributions of all other sets in $\mathcal{C}$: each element in $\tilde{\psi}(\mathcal{C})$ is covered by exactly one set. Therefore, the set $T = \arg\min_{S \in \mathcal{C}} |S \cap \tilde{\psi}(\mathcal{C})|$ in Line 4 satisfies

$$|T \cap \tilde{\psi}(\mathcal{C})| \leq \frac{|\tilde{\psi}(\mathcal{C})|}{\ell}. \tag{3}$$

With Claim 4.2, we now prove Claim 4.3.

$$\begin{aligned}
|\psi(\mathcal{C} \setminus \{T\})| &= |\psi_{\geq 2}(\mathcal{C})| + |\tilde{\psi}(\mathcal{C}) \setminus T| \\
&= |\psi(\mathcal{C})| - |\tilde{\psi}(\mathcal{C})| + |\tilde{\psi}(\mathcal{C})| - |T \cap \tilde{\psi}(\mathcal{C})| \\
&= |\psi(\mathcal{C})| - |T \cap \tilde{\psi}(\mathcal{C})| \\
&\geq |\psi(\mathcal{C})| - \frac{|\tilde{\psi}(\mathcal{C})|}{\ell} && \text{Ineq. (3)} \\
&> |\psi(\mathcal{C})| - \frac{|\psi(\mathcal{C})|}{\ell H_\ell} && \text{subcase (ii)} \\
&= \left(1 - \frac{1}{\ell H_\ell}\right) |\psi(\mathcal{C})|.
\end{aligned}$$

◁

Recall that in subcase (ii), Line 5 assigns to $\mathcal{B}$ the solution from the recursive call on $\mathcal{C} \setminus \{T\}$. Since $|\mathcal{C} \setminus \{T\}| = \ell - 1$, we apply the inductive assumption to prove that $\mathcal{B}$ satisfies Ineq. (2).

$$\begin{aligned}
|\tilde{\psi}(\mathcal{B})| &\geq \frac{|\psi(\mathcal{C} \setminus \{T\})|}{H_{\ell-1}} && \text{inductive assumption} \\
&\geq \frac{1}{H_{\ell-1}} \left(1 - \frac{1}{\ell H_\ell}\right) |\psi(\mathcal{C})| && \text{Claim 4.3} \\
&= \frac{1}{H_{\ell-1}} \frac{\ell H_\ell - 1}{\ell H_\ell} |\psi(\mathcal{C})| \\
&= \frac{1}{H_{\ell-1}} \frac{H_\ell - \frac{1}{\ell}}{H_\ell} |\psi(\mathcal{C})| \\
&= \frac{|\psi(\mathcal{C})|}{H_\ell}. && H_\ell - \frac{1}{\ell} = H_{\ell-1}, \text{ for } \ell \geq 2
\end{aligned}$$

We have proven that $\mathcal{B}$ satisfies Ineq. (2) in the base case and the inductive case, proving Theorem 4.1. ◀

4.2 UniqueGreedyFreq

In this section, we present and analyze our algorithm UniqueGreedyFreq, with pseudocode in Algorithm 2. The purpose of this algorithm is to take a collection $\mathcal{C}$ with maximum frequency $r \leq |\mathcal{C}|$, and an error parameter $\varepsilon_r \in (0, 1)$, and return a collection $\mathcal{B} \subseteq \mathcal{C}$ whose unique coverage is at least a $(1/H_{\lceil r(r-1)/\varepsilon_r \rceil} - \varepsilon_r)$ factor of $\mathcal{C}$'s coverage. By an appropriate choice of ε_r depending on r, this factor can be simplified to $1/(2 \ln r + o(\log r))$.

UniqueGreedyFreq Overview. The idea of UniqueGreedyFreq is to group all of the sets from $\mathcal{C}$ into $\hat{\ell}$ disjoint collections, $\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{\ell}}$, so that the sets must be selected into the solution $\mathcal{B}$ in these groups, i.e., for each $i \in [\hat{\ell}]$, either all of the sets in $\mathcal{G}_i$, or none of the sets in $\mathcal{G}_i$, must be selected into $\mathcal{B}$. Then, letting $\hat{\mathcal{C}}$ be the collection of the covers of $\mathcal{G}_1, \ldots, \mathcal{G}_{\hat{\ell}}$, we can call UniqueGreedy on $\hat{\mathcal{C}}$ to find a selection of these covers, namely $\hat{\mathcal{B}}$. The returned solution, $\mathcal{B}$, is constructed by merging each $\mathcal{G}_i$ whose cover was selected into $\hat{\mathcal{B}}$, which ensures that the sets are selected in groups.

It can be seen that, by calling UniqueGreedy on $\hat{\mathcal{C}}$ and by Theorem 4.1, the unique coverage of $\hat{\mathcal{B}}$ is at least $1/H_{\hat{\ell}}$ of $\hat{\mathcal{C}}$'s coverage, and therefore at least $1/H_{\hat{\ell}}$ of $\mathcal{C}$'s coverage since $\hat{\mathcal{C}}$ and $\mathcal{C}$ have the same cover. The issue now is that sets from the same $\mathcal{G}_i$ can overlap after being selected as a group into $\mathcal{B}$, which would make $\mathcal{B}$'s unique coverage smaller than $\hat{\mathcal{B}}$'s unique coverage. This is addressed by setting the number of groups to be $\hat{\ell} = \lceil r(r-1)/\varepsilon_r \rceil$, and by the way UniqueGreedyFreq allocates the sets into these groups: it allocates each $S \in \mathcal{C}$ to the group $\mathcal{G}_i$ whose unique coverage intersects the least with S. In this way, the total unique coverage that is lost due to overlapping sets in the same $\mathcal{G}_i$ can be bounded by $\varepsilon_r|\psi(\mathcal{C})|$. Thus, $\mathcal{B}$'s unique coverage is at least $(1/H_{\hat{\ell}} - \varepsilon_r) = (1/H_{\lceil r(r-1)/\varepsilon_r \rceil} - \varepsilon_r)$ of $\mathcal{C}$'s coverage. Details are given in the proof of Theorem 4.4.

Algorithm 2 UniqueGreedyFreq.

Input: $\mathcal{C}$: collection with maximum frequency $r \geq 2$, $\varepsilon_r \in (0,1)$: error parameter.
Output: $\mathcal{B} \subseteq \mathcal{C}$: collection satisfying $|\tilde{\psi}(\mathcal{B})| \geq \left(1/H_{\lceil r(r-1)/\varepsilon_r \rceil} - \varepsilon_r\right)|\psi(\mathcal{C})|$.

$\hat{\ell} \leftarrow \lceil r(r-1)/\varepsilon_r \rceil$
for $i \in [\hat{\ell}]$ **do** // Initialize empty groups
  $\mathcal{G}_i \leftarrow \varnothing$
for $S \in \mathcal{C}$ **do** // Allocate sets to groups
  $i \leftarrow \arg\min_{j \in [\hat{\ell}]} |\tilde{\psi}(\mathcal{G}_j) \cap S|$
  $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup \{S\}$
$\hat{\mathcal{C}} \leftarrow \{\psi(\mathcal{G}_1), \ldots, \psi(\mathcal{G}_{\hat{\ell}})\}$ // Define collection of groups' covers
$\hat{\mathcal{B}} \leftarrow$ UniqueGreedy$(\hat{\mathcal{C}})$
$\mathcal{B} \leftarrow \varnothing$
for $\psi(\mathcal{G}_i) \in \hat{\mathcal{B}}$ **do** // Construct returned solution
  $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{G}_i$
return $\mathcal{B}$

▶ **Theorem 4.4.** *Given $\mathcal{C}$ with maximum frequency $r \geq 2$, and error parameter $\varepsilon_r \in (0,1)$, algorithm UniqueGreedyFreq returns a collection $\mathcal{B} \subseteq \mathcal{C}$ satisfying*

$$|\tilde{\psi}(\mathcal{B})| \geq \left(\frac{1}{H_{\lceil r(r-1)/\varepsilon_r \rceil}} - \varepsilon_r\right)|\psi(\mathcal{C})|\,. \tag{4}$$

Moreover, setting $\varepsilon_r = (9.27\ln r)^{-1}(2\ln r + 2\ln\ln r + 5.61)^{-1}$, we obtain

$$|\tilde{\psi}(\mathcal{B})| \geq \left(\frac{1 - 1/(9.27\ln r)}{2\ln r + 2\ln\ln r + 5.61}\right)|\psi(\mathcal{C})| \geq \frac{1}{2\ln r + o(\log r)}|\psi(\mathcal{C})|\,. \tag{5}$$

Proof. We first prove Ineq. (4), starting with the following claim.

▷ Claim 4.5. $\tilde{\psi}(\mathcal{B}) = \tilde{\psi}(\hat{\mathcal{B}}) \setminus \bigcup_{i:\,\psi(\mathcal{G}_i) \in \hat{\mathcal{B}}} \psi_{\geq 2}(\mathcal{G}_i)$.

Proof. Consider an element x that is in exactly one set in $\mathcal{B}$. This means that x is in exactly one set from exactly one group, say $\mathcal{G}_y$, chosen in $\mathcal{B}$. Focusing on $\hat{\mathcal{B}}$, element x is clearly in $\psi(\mathcal{G}_y)$ only, but might occur more than once in $\mathcal{G}_y$. Excluding elements that are in $\psi_{\geq 2}(\mathcal{G}_i)$ for every i, we thus have the claim statement. ◁

With Claim 4.5, we prove Claim 4.6.

▷ Claim 4.6. The solution $\mathcal{B}$ satisfies $|\tilde{\psi}(\mathcal{B})| \geq |\tilde{\psi}(\hat{\mathcal{B}})| - \varepsilon_r|\psi(\mathcal{C})|$.

Proof. Given Claim 4.5, $\mathcal{B}$ satisfies Ineq. (6),

$$|\tilde{\psi}(\mathcal{B})| \geq |\tilde{\psi}(\hat{\mathcal{B}})| - \sum_{i:\ \psi(\mathcal{G}_i)\in\hat{\mathcal{B}}} |\psi_{\geq 2}(\mathcal{G}_i)|$$
$$\geq |\tilde{\psi}(\hat{\mathcal{B}})| - \sum_{i\in[\hat{\ell}]} |\psi_{\geq 2}(\mathcal{G}_i)|\,. \tag{6}$$

We upper bound $\sum_{i\in[\hat{\ell}]} |\psi_{\geq 2}(\mathcal{G}_i)|$ in Ineq. (6). Let S_t be the t^{th} set allocated in the Line 4 loop, let $\mathcal{G}_{i,0} = \varnothing$, and let $\mathcal{G}_{i,t}$ be the subcollection $\mathcal{G}_i$ just after allocating S_t.

Upon inserting S_t into $\mathcal{G}_i$, every element in $\tilde{\psi}(\mathcal{G}_{i,t-1})$ that becomes non-uniquely covered is accounted for by $\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t$. So it holds that $|\psi_{\geq 2}(\mathcal{G}_{i,t})| - |\psi_{\geq 2}(\mathcal{G}_{i,t-1})| = |\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t|$. Thus, $|\psi_{\geq 2}(\mathcal{G}_i)|$ can be expressed by Equation (7) below, observing that for S_t the relevant difference term is zero.

$$|\psi_{\geq 2}(\mathcal{G}_i)| = \sum_{S_t\in\mathcal{G}_i} \left(|\psi_{\geq 2}(\mathcal{G}_{i,t})| - |\psi_{\geq 2}(\mathcal{G}_{i,t-1})|\right) \qquad \text{telescoping series}$$
$$= \sum_{S_t\in\mathcal{G}_i} |\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t|\,. \tag{7}$$

For each $i \in [\hat{\ell}]$ and each $S_t \in \mathcal{G}_i$, we want to show an upper bound of $|\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t| \leq (r-1)|S_t|/\hat{\ell}$. To see this, since the maximum frequency is r, each element $x \in S_t$ is covered by at most $r-1$ other sets, each possibly in a different group. Therefore, we have that

$$\sum_{j\in[\hat{\ell}]} |\tilde{\psi}(\mathcal{G}_{j,t-1}) \cap \{x\}| \leq r-1\,,$$
$$\sum_{x\in S_t}\sum_{j\in[\hat{\ell}]} |\tilde{\psi}(\mathcal{G}_{j,t-1}) \cap \{x\}| \leq \sum_{x\in S_t} (r-1)\,,$$
$$\sum_{j\in[\hat{\ell}]} |\tilde{\psi}(\mathcal{G}_{j,t-1}) \cap S_t| \leq (r-1)|S_t|\,.$$

Recall that S_t was allocated to the group $\mathcal{G}_i = \arg\min_{j\in[\hat{\ell}]} |\tilde{\psi}(\mathcal{G}_{j,t-1}) \cap S_t|$ in Lines 5–6. Therefore, by averaging on the above inequality, we have that for each $i \in [\hat{\ell}]$ and each S_t that ends up in $\mathcal{G}_i$,

$$|\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t| \leq \frac{r-1}{\hat{\ell}}|S_t|\,. \tag{8}$$

Now we upper bound $\sum_{i\in[\hat{\ell}]} |\psi_{\geq 2}(\mathcal{G}_i)|$.

$$\sum_{i\in[\hat{\ell}]} |\psi_{\geq 2}(\mathcal{G}_i)| = \sum_{i\in[\hat{\ell}]}\sum_{S_t\in\mathcal{G}_i} |\tilde{\psi}(\mathcal{G}_{i,t-1}) \cap S_t| \qquad \text{Equation (7)}$$
$$\leq \frac{r-1}{\hat{\ell}} \sum_{i\in[\hat{\ell}]}\sum_{S_t\in\mathcal{G}_i} |S_t| \qquad \text{Ineq. (8)}$$
$$= \frac{r-1}{\hat{\ell}} \sum_{S\in\mathcal{C}} |S| \qquad \mathcal{G}_1,\ldots,\mathcal{G}_{\hat{\ell}} \text{ partitions } \mathcal{C}$$
$$\leq \frac{r-1}{\hat{\ell}} r|\psi(\mathcal{C})| \qquad \text{for all } x\in\psi(\mathcal{C})\colon\ \text{freq}_{\mathcal{C}}(x) \leq r$$
$$= \frac{r(r-1)}{\lceil r(r-1)/\varepsilon_r \rceil}|\psi(\mathcal{C})| \qquad \text{value of } \hat{\ell} \text{ (Line 1)}$$

$$\leq \frac{r(r-1)}{r(r-1)/\varepsilon_r}|\psi(\mathcal{C})|$$
$$\leq \varepsilon_r|\psi(\mathcal{C})|\,.$$

Applying the above upper bound to Ineq. (6) completes the proof of the claim. ◁

To prove Ineq. (4), it remains to lower bound $|\tilde{\psi}(\hat{\mathcal{B}})|$, in the inequality of Claim 4.6, in terms of $|\psi(\mathcal{C})|$. Below, $|\psi(\hat{\mathcal{C}})| = |\psi(\mathcal{C})|$ holds since every $S \in \mathcal{C}$ is allocated to some $\mathcal{G}_i \in \hat{\mathcal{C}}$.

$$\begin{aligned}
|\tilde{\psi}(\mathcal{B})| &\geq |\tilde{\psi}(\hat{\mathcal{B}})| - \varepsilon_r|\psi(\mathcal{C})| && \text{Claim 4.6} \\
&\geq \frac{|\psi(\hat{\mathcal{C}})|}{H_{\hat{\ell}}} - \varepsilon_r|\psi(\mathcal{C})| && \text{Line 8 and Theorem 4.1} \\
&= \frac{|\psi(\hat{\mathcal{C}})|}{H_{\lceil r(r-1)/\varepsilon_r\rceil}} - \varepsilon_r|\psi(\mathcal{C})| && \text{value of } \hat{\ell} \text{ (Line 1)} \\
&= \frac{|\psi(\mathcal{C})|}{H_{\lceil r(r-1)/\varepsilon_r\rceil}} - \varepsilon_r|\psi(\mathcal{C})| && |\psi(\hat{\mathcal{C}})| = |\psi(\mathcal{C})| \\
&= \left(\frac{1}{H_{\lceil r(r-1)/\varepsilon_r\rceil}} - \varepsilon_r\right)|\psi(\mathcal{C})|\,.
\end{aligned}$$

Ineq. (5). It remains to show that there exists a choice of ε_r that implies Ineq. (5).

▷ Claim 4.7 (*). Setting $\varepsilon_r = (9.27 \ln r)^{-1}(2\ln r + 2\ln\ln r + 5.61)^{-1}$ implies Ineq. (5).

This completes the proof of Theorem 4.4. ◀

4.3 UniqueGreedySize

In this section, we present UniqueGreedySize, with pseudocode in Algorithm 3, derived by combining UniqueGreedyFreq with the approach in Theorem 4.2 of Demaine et al. [9]. The purpose of this algorithm is to take a collection, $\mathcal{C}$, with maximum set size d, an error parameter, $\varepsilon_d \in (0,1)$, and another error parameter, $\hat{\varepsilon}_d \in (0,1)$, and return a $\mathcal{B} \subseteq \mathcal{C}$ whose unique coverage is at least a logarithmic factor of $\mathcal{C}$'s coverage, where the factor depends on d, ε_d, and $\hat{\varepsilon}_d$. We state this formally in Theorem 4.8 and give the proof for completeness; in fact, our proof slightly generalizes the proof of Theorem 4.2 of Demaine et al. [9], by allowing an arbitrary ε_d rather than fixing $\varepsilon_d = 1/2$.

UniqueGreedySize Overview. UniqueGreedySize first modifies $\mathcal{C}$ into a "minimal" collection by discarding each set T that uniquely covers no element. Then it checks if $\mathcal{C}$'s size is at least an ε_d factor of its own coverage. If so, then it assigns $\mathcal{C}$ to the solution $\mathcal{B}$. Otherwise, it constructs a sub-instance on those elements of frequency at most d and calls UniqueGreedyFreq on the sub-instance with error $\hat{\varepsilon}_d$ to get $\hat{\mathcal{B}}$. Returned solution $\mathcal{B}$ comprises each set $S \in \mathcal{C}$ whose intersection with $\hat{U}$ was selected into $\hat{\mathcal{B}}$.

▶ **Theorem 4.8.** *Let $\mathcal{C}$ denote a collection of sets, d denote the maximum size of a set in $\mathcal{C}$, $\varepsilon_d \in (0,1)$ denote an error parameter, and $\hat{\varepsilon}_d \in (0,1)$ denote an error parameter passed to UniqueGreedyFreq. Then UniqueGreedySize returns a collection $\mathcal{B} \subseteq \mathcal{C}$ satisfying*

$$|\tilde{\psi}(\mathcal{B})| \geq \min(\varepsilon_d, (1-\varepsilon_d)\beta(d,\hat{\varepsilon}_d))|\psi(\mathcal{C})|\,, \tag{9}$$

Algorithm 3 UniqueGreedySize.

Input: $\mathcal{C}$: collection with maximum set size d, $\varepsilon_d \in (0,1)$: error parameter,
$\hat{\varepsilon}_d \in (0,1)$: error parameter used in UniqueGreedyFreq.
Output: $\mathcal{B} \subseteq \mathcal{C}$: subcollection satisfying $|\tilde{\psi}(\mathcal{B})| \geq \min(\varepsilon_d, (1-\varepsilon_d)\beta(d,\hat{\varepsilon}_d))|\psi(\mathcal{C})|$
where $\beta(d,\hat{\varepsilon}_d) = 1/H_{\lceil d(d-1)/\hat{\varepsilon}_d \rceil} - \hat{\varepsilon}_d$.

```
1  while T ← arg min_{S∈C} |S ∩ ψ̃(C)| satisfies |T ∩ ψ̃(C)| = 0 do    // Make C minimal
2      C ← C \ {T}
3  if |C| ≥ ε_d|ψ(C)| then
4      B ← C
5  else                              // Define instance on elements with freq. ≤ d
6      Û ← {x ∈ ψ(C) : freq_C(x) ≤ d}
7      Ĉ ← {S ∩ Û : S ∈ C}
8      B̂ ← UniqueGreedyFreq(Ĉ, ε̂_d)
9      B ← ∅
10     for S ∩ Û ∈ B̂ do                             // Construct returned solution
11         B ← B ∪ {S}
12 return B
```

where $\beta(d,\hat{\varepsilon}_d) = 1/(H_{\lceil d(d-1)/\hat{\varepsilon}_d \rceil}) - \hat{\varepsilon}_d$ *denotes the unique coverage ratio of* UniqueGreedy-Freq. *Moreover, by assigning* $\varepsilon_d = (1/\beta(d,\hat{\varepsilon}_d)+1)^{-1}$, $\hat{\varepsilon}_d = (c_1 \ln d)^{-1}(2\ln d + 2\ln\ln d + c_2)^{-1}$, *and appropriate constants to* c_1 *and* c_2*, we derive from Ineq.* (9) *the simpler inequality below.*

$$|\tilde{\psi}(\mathcal{B})| \geq \frac{1}{2\ln d + o(\log d)} |\psi(\mathcal{C})| \,. \tag{10}$$

Proof. We begin by proving Ineq. (9). Discarding sets from $\mathcal{C}$ that uniquely cover no elements, as in Lines 1–2, does not affect $\psi(\mathcal{C})$. So assume that $\mathcal{C}$ is minimal, i.e., every $S \in \mathcal{C}$ uniquely covers at least one element. This means that $|\tilde{\psi}(\mathcal{C})| \geq |\mathcal{C}|$.

Now one of two cases must hold: (i) $|\mathcal{C}| \geq \varepsilon_d|\psi(\mathcal{C})|$; or (ii) $|\mathcal{C}| < \varepsilon_d|\psi(\mathcal{C})|$. The final ratio in Ineq. (9) is the minimum ratio achieved out of these two cases.

In case (i), UniqueGreedySize returns the solution $\mathcal{B} = \mathcal{C}$, by the success of the condition in Line 3. Further, $|\tilde{\psi}(\mathcal{B})| = |\tilde{\psi}(\mathcal{C})| \geq |\mathcal{C}| \geq \varepsilon_d|\psi(\mathcal{C})|$ holds by the minimality of $\mathcal{C}$. Thus, $\mathcal{B}$ satisfies Ineq. (9) in case (i).

In case (ii), we show that the set $\hat{U}$ of elements $x \in \psi(\mathcal{C})$ with $\mathrm{freq}_{\mathcal{C}}(x) \leq d$, as in Line 6, satisfies $|\hat{U}| \geq (1-\varepsilon_d)|\psi(\mathcal{C})|$. We have

$$\begin{aligned}
|U \setminus \hat{U}| &< \frac{1}{d}\sum_{S\in\mathcal{C}} |S| && \text{for all } x \in U \setminus \hat{U}\colon\ \mathrm{freq}_{\mathcal{C}}(x) > d \\
&\leq |\mathcal{C}| && \text{max. set size is } d \\
&< \varepsilon_d|\psi(\mathcal{C})|\,, && \text{case (ii)} \\
|\hat{U}| &> (1-\varepsilon_d)|\psi(\mathcal{C})|\,.
\end{aligned}$$

By the Line 7 definition, $\psi(\hat{\mathcal{C}}) = \hat{U}$, so $|\psi(\hat{\mathcal{C}})| \geq (1-\varepsilon_d)|\psi(\mathcal{C})|$. Therefore, calling UniqueGreedyFreq on $\hat{\mathcal{C}}$ with maximum frequency d and error $\hat{\varepsilon}_d$, as in Line 8, gives a collection $\hat{\mathcal{B}}$ satisfying $|\tilde{\psi}(\hat{\mathcal{B}})| \geq \beta(d,\hat{\varepsilon}_d)|\psi(\hat{\mathcal{C}})| \geq \beta(d,\hat{\varepsilon}_d)(1-\varepsilon_d)|\psi(\mathcal{C})|$. Likewise, by definition, in Lines 9–11, $|\tilde{\psi}(\mathcal{B})| \geq |\tilde{\psi}(\hat{\mathcal{B}})|$. Thus, $\mathcal{B}$ satisfies Ineq. (9) in case (2), proving Theorem 4.8.

Ineq. (10). We first maximize $\min(\varepsilon_d, (1-\varepsilon_d)\beta(d,\hat{\varepsilon}_d))$ with respect to ε_d by setting the two arguments as equal; this makes the RHS of Ineq. (9) equal to $\varepsilon_d = (1/\beta(d,\hat{\varepsilon}_d)+1)^{-1}$. Then, by assigning $\hat{\varepsilon}_d = (c_1 \ln d)^{-1}(2\ln d + 2\ln\ln d + c_2)^{-1}$ and appropriate constants to c_1 and c_2, UniqueGreedyFreq's unique coverage ratio satisfies $\beta(d,\hat{\varepsilon}_d) \geq 1/(2\ln d + o(\log d))$ as in Theorem 4.4. Further substituting this into the RHS of Ineq. (9) proves Ineq. (10). This completes the proof of Theorem 4.8. ◀

5 Space Lower Bound for a $(1.5+o(1))/(\ln k - 1)$-Approximation

In this section, we prove the following theorem:

▶ **Theorem 5.1.** *Let $e^{2.5} \leq k \leq m$, $a = k\ln m + \ln(k/0.05)$, and assume the universe size to be $n = k(k-1)\sum_{t=1}^{k}\lceil a/t \rceil$. Then every constant-pass randomized streaming algorithm for **Max Unique Coverage** that, with probability at least 0.95, has an approximation factor of $(3/2+3/\sqrt{2k})/(H_k-1)$, requires $\Omega(m/k^2)$ space.*

5.1 High-Level Ideas of the Reduction

Similar to other approaches [16, 15], we prove our space lower bound by reducing the problem of k-player Set Disjointness (with the unique intersection promise) in the one-way communication model, denoted by **Disj**, to **Max Unique Coverage** in the stream model.

Set Disjointness in the One-Way Communication Model. In the one-way communication model, players must take turns in some fixed order to send a message to the player next in order, i.e., the j^{th} player can only send a message to the $(j+1)^{\text{th}}$ player. There can be $p \geq 1$ rounds of communication, where a single round is completed once every player has taken their turn. The last player can send a message back to the 1^{st} player at the end of a round if there is a next round.

In an instance of **Disj**, each player $j \in [k]$ is given a set of integers $D_j \subseteq [m]$. Moreover, it is promised that only two kinds of instances can occur:

NO instance. All sets D_j are pairwise disjoint.
YES instance. There is a unique integer $i^* \in [m]$ such that, for all $j \in [k]$, $i^* \in D_j$.

The goal then is for the k^{th} player (in the final round) to correctly answer, with probability at least 0.9, whether the given sets form a YES or NO instance.

The communication complexity of **Disj** in the p-round one-way communication model is $\Omega(m/k)$, even for randomized protocols and even when the players can use public randomness [5]. Thus, as there are $\leq pk$ messages, every (randomized) protocol for **Disj** must have at least one message of size $\Omega(m/(pk^2))$ in the worst case.

Reduction Overview. Given an instance of **Disj**, the main goal of the reduction, with parameter a, is for the players to construct an instance of **Max Unique Coverage** in a stream such that if they were given a NO instance of **Disj**, the optimal unique coverage is less than $ak^2(1.5+o(1))$ (with high probability); whereas if the players were given a YES instance of **Disj**, the optimal unique coverage is at least $ak^2(H_k-1)$. The ratio of these optimal unique coverages is less than $(1.5+o(1))/(H_k-1)$, so the players can use a $(1.5+o(1))/(H_k-1)$-approximation streaming algorithm on the **Max Unique Coverage** instance to distinguish between a NO and YES instance. By a standard argument, this implies a protocol for **Disj** which involves each player sending the memory of the streaming algorithm in a message to the next player. A constant-pass $O(s)$-space streaming algorithm implies a protocol with a maximum message size of $O(s)$ in constant rounds of communication where each pass of the streaming algorithm takes one round. Thus, a $(1.5+o(1))/(H_k-1)$-approximation streaming algorithm for **Max Unique Coverage** requires $\Omega(m/k^2)$ space.

Intuition of Max Unique Coverage Construction. Here, we give the intuition for constructing the streaming instance of **Max Unique Coverage** that achieves the optimal unique coverages above, with details in the proof of Theorem 5.1.

Let the universe of the **Max Unique Coverage** instance be $U = U_1 \cup \cdots \cup U_k$, where $U_1, \ldots, U_k$ are k disjoint sub-universes such that $|U_t| = k(k-1)\lceil a/t \rceil$ (for a sufficiently large a as in Theorem 5.1). Then, for each $i \in [m]$, each player j constructs $S_j^i \subseteq U$ such that $S_1^i, \ldots, S_k^i$ satisfy the following properties:

1. Each set S_j^i covers t/k proportion of U_t for all t.
2. For each $t \in [k]$, the sets $S_1^i, \ldots, S_k^i$ partition a proportion, $q_t \in [0,1]$, of U_t while having a common intersection in the remaining $(1-q_t)$ proportion of U_t. I.e., sets with identical i form a "sunflower", with their overlap concentrated in the sunflower's "kernel".
3. The choice of elements to be covered by S_j^i are independent and uniform random with respect to $i \in [m]$.

The above construction ensures that (with high probability) every collection of $\ell \in [k]$ sets, $S_{j_1}^{i_1}, \ldots, S_{j_\ell}^{i_\ell}$, with distinct $i_1, \ldots, i_\ell$ has a unique coverage less than $ak^2(1.5+o(1))$ (with high probability); whereas a collection of $\ell = k$ sets with identical $i_1, \ldots, i_\ell$ has a unique coverage of at least $ak^2(H_k-1)$. Observe that $k \ge e^{2.5}$ ensures that $H_k - 1 > 1.5 + o(1)$.

Finally, to construct the streaming instance of **Max Unique Coverage**, each player j inserts S_j^i into the stream iff $i \in D_j$. This means that, given a NO instance, every set S_j^i in the stream has a distinct i; whereas given a YES instance, there exists a collection of $\ell = k$ sets in the stream all indexed by i^*, the unique integer contained in all $D_1, \ldots, D_k$. This results in the optimal unique coverages for the NO and YES instances as required.

5.2 Proof of Theorem 5.1

We show a reduction from **Disj** to **Max Unique Coverage**. Assume without loss of generality that the sets D_j are padded so that $|D_1 \cup \cdots \cup D_k| \ge m/4 \ge m/k^2$ holds for $k \ge 2$.

Construction of Max Unique Coverage Instance. First, the players define the **Max Unique Coverage** universe as $U = U_1 \cup \cdots \cup U_k$, where $U_1, \ldots, U_k$ are k disjoint sub-universes such that $|U_t| = k(k-1)\lceil a/t \rceil$. Observe that, as per the assumption in Theorem 5.1, we have $n = |U| = \sum_{t=1}^{k} |U_t| = k(k-1)\sum_{t=1}^{k} \lceil a/t \rceil$.

The players now construct the **Max Unique Coverage** sets so that they satisfy the properties given in the overview. For each $i \in [m]$ and $t \in [k]$, the players define $\tilde{U}_t^i \subseteq U_t$ as an independent and uniformly chosen random subset of size $q_t = (k-t)/(k-1)$ proportion of U_t; they then independently and uniformly-at-random partition $\tilde{U}_t^i$ into k equally sized sets, $P_{t,1}^i, \ldots, P_{t,k}^i$; the players agree on all of these choices using public randomness. For example, the players obtain a common random permutation of U_t and pick the corresponding parts in order. Note that $\tilde{U}_t^i$ can be divided into k equal sets since $|\tilde{U}_t^i|/k$ is an integer, viz.

$$\frac{|\tilde{U}_t^i|}{k} = \frac{q_t|U_t|}{k} = \frac{(k-t)k(k-1)}{k(k-1)} \left\lceil \frac{a}{t} \right\rceil = (k-t)\left\lceil \frac{a}{t} \right\rceil.$$

Then, for each $i \in [m]$, each player j defines their set S_j^i such that, for each $t \in [k]$, it covers the j^{th} set in the partition of $\tilde{U}_t^i$, namely $P_{t,j}^i$; and it covers all of $U_t \setminus \tilde{U}_t^i$. More precisely,

$$S_j^i = \bigcup_{t=1}^{k} \left[P_{t,j}^i \cup (U_t \setminus \tilde{U}_t^i)\right] .$$

Observe Claim 5.2, which we use in Claim 5.4 later.

▷ Claim 5.2. For each $i \in [m]$, $j \in [k]$, and $t \in [k]$, S_j^i covers t/k proportion of U_t.

Proof. The proportion of U_t that S_j^i covers is $|S_j^i \cap U_t|/|U_t|$, which we prove to be t/k below.

$$\begin{aligned}\frac{|S_j^i \cap U_t|}{|U_t|} &= \frac{|P_{t,j}^i|}{|U_t|} + \frac{|U_t \setminus \tilde{U}_t^i|}{|U_t|} = \frac{|\tilde{U}_t^i|}{k|U_t|} + \frac{|U_t \setminus \tilde{U}_t^i|}{|U_t|} = \frac{q_t}{k} + 1 - q_t \\ &= \frac{k-t}{k(k-1)} + 1 - \frac{k-t}{k-1} = \frac{k-t}{k(k-1)} + \frac{t-1}{k-1} \\ &= \frac{k-t+kt-k}{k(k-1)} = \frac{kt-t}{k(k-1)} = \frac{t(k-1)}{k(k-1)} = \frac{t}{k} .\end{aligned}$$

◁

To complete the construction, each player j inserts set S_j^i into the stream iff $i \in D_j$. There are $\Theta(m)$ sets inserted into the stream since $m/4 \leq |D_1 \cup \cdots \cup D_k| \leq m$.

Upper Bound on Optimal Unique Coverage in a NO Instance. Next, we prove Lemma 5.3, which implies the required upper bound on the optimal unique coverage in a NO instance. We say that a collection $\mathcal{L}_{\text{di}} = \{S_{j_1}^{i_1}, \ldots, S_{j_\ell}^{i_\ell}\}$ with distinct $i_1, \ldots, i_\ell$ is a *player-distinct collection*; we also say that $\mathcal{L}_{\text{di}}$ is *feasible* if it contains at most k sets. Note that in the **Max Unique Coverage** instance generated from a NO instance of **Disj**, every feasible solution is a player-distinct collection. Thus, it suffices to upper bound the unique coverage of every feasible player-distinct collection.

▶ **Lemma 5.3.** *With probability at least* 0.95*, every feasible player-distinct collection* $\mathcal{L}_{\text{di}}$ *satisfies* $|\tilde{\psi}(\mathcal{L}_{\text{di}})| < ak^2(3/2 + 3/\sqrt{2k})$.

Proof. First, we upper bound $\mathbb{E}[|\tilde{\psi}(\mathcal{L}_{\text{di}}) \cap U_t|]$ for every feasible player-distinct collection, $\mathcal{L}_{\text{di}}$, and for every sub-universe U_t (Claim 5.4), then we use Hoeffding's inequality to prove an upper bound on $|\tilde{\psi}(\mathcal{L}_{\text{di}}) \cap U_t|$ that with high probability, holds simultaneously for every $\mathcal{L}_{\text{di}}$ and U_t (Claim 5.5). Summing the bound in Claim 5.5 over all k sub-universes suffices.

For a feasible player-distinct collection $\mathcal{L}_{\text{di}}$, let $X_{x,\mathcal{L}_{\text{di}}}$ be the random variable such that $X_{x,\mathcal{L}_{\text{di}}} = 1$ if element $x \in \tilde{\psi}(\mathcal{L}_{\text{di}})$, and $X_{x,\mathcal{L}_{\text{di}}} = 0$ otherwise. This means that for each sub-universe U_t, we have

$$|\tilde{\psi}(\mathcal{L}_{\text{di}}) \cap U_t| = \sum_{x \in U_t} X_{x,\mathcal{L}_{\text{di}}}; \text{ and so } |\tilde{\psi}(\mathcal{L}_{\text{di}})| = \sum_{t=1}^{k} \sum_{x \in U_t} X_{x,\mathcal{L}_{\text{di}}} . \tag{11}$$

▷ Claim 5.4 (*). For each feasible player-distinct $\mathcal{L}_{\text{di}}$ of $\ell \in [k]$ sets and each sub-universe U_t, it holds that $\mathbb{E}\left[|\tilde{\psi}(\mathcal{L}_{\text{di}}) \cap U_t|\right] \leq k(a+t)\ell\,(1-t/k)^{\ell-1}$.

▷ Claim 5.5 (*). With probability at least 0.95, for every feasible player-distinct $\mathcal{L}_{\text{di}}$ of $\ell \in [k]$ sets and every sub-universe U_t, it holds that

$$|\tilde{\psi}(\mathcal{L}_{\text{di}}) \cap U_t| < k(a+t)\ell\left(1 - \frac{t}{k}\right)^{\ell-1} + \frac{k(a+t)}{(2t)^{1/2}} .$$

Finally, summing the inequality of Claim 5.5 over the k sub-universes gives an upper bound on $|\tilde{\psi}(\mathcal{L}_{\text{di}})|$ that holds simultaneously for every feasible player-distinct collection $\mathcal{L}_{\text{di}}$ with high probability. We finalize the proof of Lemma 5.3 in Claim 5.6.

▷ Claim 5.6 (*). With probability at least 0.95, $|\tilde{\psi}(\mathcal{L}_{\text{di}})| < ak^2\left(3/2 + 3/\sqrt{2k}\right)$. ◀

Lower Bound on Optimal Unique Coverage in a YES Instance. Lemma 5.7 supports the required lower bound on the optimal unique coverage in a YES instance.

▶ **Lemma 5.7.** *For all i, collection $\mathcal{L}_{\text{id}} = \{S_1^i, \ldots, S_k^i\}$ satisfies $|\tilde{\psi}(\mathcal{L}_{\text{id}})| \geq ak^2(H_k - 1)$.*

Proof. For each $t \in [k]$, $\mathcal{L}_{\text{id}}$ uniquely covers $|\tilde{U}_t^i|$ by construction. Below, the inequality holds since $|U_t| = k(k-1)\lceil a/t \rceil \geq ak(k-1)/t$.

$$|\tilde{\psi}(\mathcal{L}_{\text{id}})| = \sum_{t=1}^{k} |\tilde{U}_t^i| = \sum_{t=1}^{k} q_t |U_t| \geq \sum_{t=1}^{k} \frac{k-t}{k-1}\frac{ak(k-1)}{t} = \sum_{t=1}^{k} \frac{k-t}{t} ak$$
$$= ak \sum_{t=1}^{k}\left(\frac{k}{t} - 1\right) = ak\left(k\sum_{t=1}^{k}\frac{1}{t} - k\right) = ak^2\left(H_k - 1\right). \quad ◀$$

To conclude, when the players reduce from a NO instance of **Disj**, with probability at least 0.95, the optimal unique coverage is less than $ak^2(3/2 + 3/\sqrt{2k})$, since the streamed sets are player distinct and by Lemma 5.3; whereas when they reduce from a YES instance, the optimal unique coverage is at least $ak^2(H_k - 1)$ since the sets $S_1^{i^*}, \ldots, S_k^{i^*}$ are in the stream and by Lemma 5.7. The required optimal unique coverage in a NO instance fails with probability at most 0.05. Let $\alpha = (3/2 + 3/\sqrt{2k})/(H_k - 1)$. Given a randomized $O(s)$-space α-approximation streaming algorithm with failure probability at most 0.05, the players can run this algorithm on the **Max Unique Coverage** instance to distinguish between a NO or YES instance with failure probability at most 0.1. This implies a protocol for **Disj** with maximum message size $O(s)$. Thus, a constant-pass randomized α-approximation streaming algorithm with success probability at least 0.95 requires $\Omega(m/k^2)$ space.

6 Subsampling for the Data Stream

Here we outline the subsampling approach from [15]. Given a data stream instance of **Max Unique Coverage**, it is possible to construct a number of *subsampled* instances by sampling the universe U at varying rates. By running an algorithm on these subsampled instances in parallel, we lose only a small error in approximation w.h.p. while only needing to store sets of size $O(k \log m/\varepsilon^2)$. We summarize the overall approach in Lemma 2.1 and give a proof sketch.

Proof Sketch of Lemma 2.1. Given an instance of **Max Unique Coverage** with universe U and collection o sets $\mathcal{V}$, let v be a guess of the optimal solution value; each subsampled instance corresponds to some value of v (we calculate these guesses shortly). Let $h : U \to \{0, 1\}$ be a hash function that is $\Omega(k \log m/\varepsilon^2)$-wise independent such that

$$\Pr[h(x) = 1] = p = \frac{ck \log m}{\varepsilon^2 v},$$

where c is a sufficiently large constant. Let $U' = \{x \in U : h(x) = 1\}$ be the subsampled universe, $S' = S \cap U'$, $\mathcal{V}' = \{S' : S \in \mathcal{V}\}$ be the subsampled subsets, and OPT' be the optimal unique coverage in the subsampled instance. Further, let $\mathcal{B}'$ be a solution from $\mathcal{V}'$

and $\mathcal{B}$ be the corresponding solution from the original collection $\mathcal{V}$. Then Lemma 6.1 below (a restatement of [15, Lemma 23]) shows that, in a subsampled instance where $v \leq \mathrm{OPT}$, w.h.p., the loss in approximation is at most 2ε.

▶ **Lemma 6.1** ([15], Lemma 23). *If $v \leq \mathrm{OPT}$, then with probability at least $1 - 1/\operatorname{poly}(m)$, we have that*

$$(1+\varepsilon)p\mathrm{OPT} \geq \mathrm{OPT}' \geq (1-\varepsilon)p\mathrm{OPT}\,.$$

Furthermore, for some $\alpha \in (0,1)$, if $\mathcal{B}' \subseteq \mathcal{V}'$ satisfies $|\tilde{\psi}(\mathcal{B}')| \geq \alpha(1-\varepsilon)p\mathrm{OPT}$, then $|\tilde{\psi}(\mathcal{B})| \geq (\alpha - 2\varepsilon)\mathrm{OPT}$.

We guess $v = 2^i$ for each $i \in [\lceil \log_2 n \rceil]$ and construct a subsampled instance for each v in parallel. Then, in the particular subsampled instance where $\mathrm{OPT}/2 \leq v \leq \mathrm{OPT}$, Lemma 6.1 implies the following upper bound on every set size $|S'|$ with probability $1 - 1/\operatorname{poly}(m)$.

$$|S'| \leq \mathrm{OPT}' \leq (1+\varepsilon)p\mathrm{OPT} = (1+\varepsilon)\frac{ck \log m}{\varepsilon^2 v}\mathrm{OPT} \leq (1+\varepsilon)\frac{2ck\log m}{\varepsilon^2} = O\left(\frac{k \log m}{\varepsilon^2}\right).$$

To ensure that we only ever store sets of size $O(k \log m/\varepsilon^2)$, we terminate every subsampled instance that contains a set S' with $|S'| > (2ck \log m/\varepsilon^2)(1+\varepsilon)$. W.h.p., this does not terminate the subsampled instance where $\mathrm{OPT}/2 \leq v \leq \mathrm{OPT}$ by the above upper bound on $|S'|$ for every S' in this particular instance.

This means that, out of the nonterminated subsampled instances, we should select the one with the smallest v and return the corresponding solution, giving an $(\alpha - 2\varepsilon)$-approximation for the original instance w.h.p. (this works even if the smallest nonterminated guess satisfies $v < \mathrm{OPT}/2$ since Lemma 6.1 holds for all $v \leq \mathrm{OPT}$).

The overall space complexity, $\lceil \log_2 n \rceil \cdot s \cdot O(k \log m \log n/\varepsilon^2)$, follows from the number of guesses of v and, for each guess, the algorithm storing at most s sets of size $O(k \log m/\varepsilon^2)$ and using $O(\log n)$ bits to store each element. ◀

7 Conclusions

We are pleased to present a suite of algorithms, and a streaming lower bound, for **Max Unique Coverage**. The component algorithms that build a solution to **Max Unique Coverage** from a solution to **Max Coverage** serve to support a fixed-parameter tractable approximation scheme (FPT-AS). The lower bound shows that $\Omega(m/k^2)$ space is required even to get within a $(1.5 + o(1))/(\ln k - 1)$ factor of optimal.

A plausible future direction would be to reduce, or indeed eliminate, the role of the upper bound on the unique coverage ratio, ϕ, in the kernel size in a FPT-AS. This would match the kernel size used in existing FPT-ASs for **Max Coverage**, but may not be possible due to the inherent hardness of **Max Unique Coverage**. Another direction would be proving a streaming lower bound with a tighter approximation threshold. This may require a reduction from a different communication problem, rather than the renowned k-player Set Disjointess.

References

1 Sepehr Assadi. Tight space-approximation tradeoff for the multi-pass streaming set cover problem. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 321–335, 2017.

2 Giorgio Ausiello, Nicolas Boria, Aristotelis Giannakos, Giorgio Lucarelli, and V.T̃h Paschos. Online maximum k-coverage. *Discrete Applied Mathematics*, 160(13-14):1901–1913, 2012.

3 MohammadHossein Bateni, Hossein Esfandiari, and Vahab S. Mirrokni. Almost optimal streaming algorithms for coverage problems. In Christian Scheideler and MohammadTaghi Hajiaghayi, editors, *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA 2017, Washington DC, USA, July 24-26, 2017*, pages 13–23. ACM, 2017. doi:10.1145/3087556.3087585.

4 Édouard Bonnet, Vangelis Th Paschos, and Florian Sikora. Parameterized exact and approximation algorithms for maximum k-set cover and related satisfiability problems. *RAIRO-Theoretical Informatics and Applications-Informatique Théorique et Applications*, 50(3):227–240, 2016.

5 Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *18th IEEE Annual Conference on Computational Complexity, 2003. Proceedings.*, pages 107–117. IEEE, 2003.

6 Rajesh Chitnis and Graham Cormode. Towards a theory of parameterized streaming algorithms. In *14th International Symposium on Parameterized and Exact Computation*, 2019.

7 Rajesh Chitnis, Graham Cormode, Hossein Esfandiari, MohammadTaghi Hajiaghayi, Andrew McGregor, Morteza Monemizadeh, and Sofya Vorotnikova. Kernelization via sampling with applications to finding matchings and related problems in dynamic graph streams. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 1326–1344. SIAM, 2016.

8 Rajesh Chitnis, Graham Cormode, Hossein Esfandiari, MohammadTaghi Hajiaghayi, and Morteza Monemizadeh. New streaming algorithms for parameterized maximal matching & beyond. In *Proceedings of the 27th ACM symposium on Parallelism in Algorithms and Architectures*, pages 56–58, 2015.

9 Erik D. Demaine, Uriel Feige, MohammadTaghi Hajiaghayi, and Mohammad R. Salavatipour. Combination can be hard: Approximability of the unique coverage problem. *SIAM Journal on Computing*, 38(4):1464–1483, 2008.

10 Erik D. Demaine, Piotr Indyk, Sepideh Mahabadi, and Ali Vakilian. On streaming and communication complexity of the set cover problem. In Fabian Kuhn, editor, *Distributed Computing - 28th International Symposium, DISC 2014, Austin, TX, USA, October 12-15, 2014. Proceedings*, volume 8784 of *Lecture Notes in Computer Science*, pages 484–498. Springer, 2014. doi:10.1007/978-3-662-45174-8_33.

11 Venkatesan Guruswami and Euiwoong Lee. Nearly optimal NP-hardness of unique coverage. *SIAM Journal on Computing*, 46(3):1018–1028, 2017.

12 Sariel Har-Peled, Piotr Indyk, Sepideh Mahabadi, and Ali Vakilian. Towards tight bounds for the streaming set cover problem. In *PODS*, pages 371–383. ACM, 2016.

13 Chien-Chung Huang and François Sellier. Matroid-constrained maximum vertex cover: Approximate kernels and streaming algorithms. In *SWAT 2022*, 2022.

14 Pasin Manurangsi. A note on max k-vertex cover: Faster FPT-AS, smaller approximate kernel and improved approximation. In *2nd Symposium on Simplicity in Algorithms (SOSA 2019)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

15 Andrew McGregor, David Tench, and Hoa T. Vu. Maximum Coverage in the Data Stream Model: Parameterized and Generalized. In *24th International Conference on Database Theory*, 2021.

16 Andrew McGregor and Hoa T. Vu. Better streaming algorithms for the maximum coverage problem. *Theory of Computing Systems*, 63(7):1595–1619, 2019.

17 Neeldhara Misra, Hannes Moser, Venkatesh Raman, Saket Saurabh, and Somnath Sikdar. The parameterized complexity of unique coverage and its variants. *Algorithmica*, 65:517–544, 2013.

18 Barna Saha and Lise Getoor. On maximum coverage in the streaming model & application to multi-topic blog-watch. In *Proceedings of the 2009 siam international conference on data mining*, pages 697–708. SIAM, 2009.

19 François Sellier. Parameterized matroid-constrained maximum coverage. In *31st Annual European Symposium on Algorithms (ESA 2023)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

20 Piotr Skowron. FPT approximation schemes for maximizing submodular functions. *Information and Computation*, 257:65–78, 2017. doi:10.1016/j.ic.2017.10.002.

21 Piotr Skowron and Piotr Faliszewski. Fully proportional representation with approval ballots: Approximating the MaxCover problem with bounded frequencies in FPT time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2124–2130, 2015.

22 Huiwen Yu and Dayu Yuan. Set coverage problems in a one-pass data stream. In *Proceedings of the 2013 SIAM international conference on data mining*, pages 758–766. SIAM, 2013.