

General Coverage Models: Structure, Monotonicity, and Shotgun Sequencing

Yitzchak Grunbaum, *Student Member, IEEE* Eitan Yaakobi, *Senior Member, IEEE*

Abstract—We study coverage processes in which each draw reveals a subset of $[n]$, and the goal is to determine the expected number of draws until all items are seen at least once. A classical example is the Coupon Collector’s Problem, where each draw reveals exactly one item. Motivated by shotgun DNA sequencing, we introduce a model where each draw is a contiguous window of fixed length, in both cyclic and non-cyclic variants. We develop a unifying combinatorial tool that shifts the task of finding coverage time from probability, to a counting problem over families of subsets of $[n]$ that together contain all items, enabling exact calculation. Using this result, we obtain exact expressions for the window models. We then leverage past results on a continuous analogue of the cyclic window model to analyze the asymptotic behavior of both models. We further study what we call uniform ℓ -regular models, where every draw has size ℓ and every item appears in the same number of admissible draws. We compare these to the batch sampling model, in which all ℓ -subsets are drawn uniformly at random and present upper and lower bounds, which were also obtained independently by Berend and Sher. We conjecture, and prove for special cases, that this model maximizes the coverage time among all uniform ℓ -regular models. Finally, we prove a universal upper bound on the entire class of uniform ℓ -regular models, which illuminates the fact that many sampling models share the same leading asymptotic order, while potentially differing significantly in lower-order terms.

Index Terms—DNA data storage, Shotgun sequencing, The coupon collector’s problem, Group drawings, Coverage problems.

I. INTRODUCTION

COVERAGE PROBLEMS refer to a class of problems in which there are $[n]$ items to draw from, and on each draw a subset is obtained. The question of interest is finding the expected number of draws until all (or some of) the items are collected. One way to describe those problems more rigorously is by considering a hypergraph on $[n]$, with $\{0, 1, \dots, n-1\}$ being the vertices, and all admissible groups we can draw from constituting the hyperedges. Then, the question translates to finding the expected number of draws of hyperedges until all (or a subset of the) vertices are covered, where by being covered we mean that the vertex was contained in one of the drawn hyperedges.

Perhaps the most basic and well-known example to represent this family of problems is the *Coupon Collector’s Problem*. The coupon collector’s problem, introduced in early probability theory ([7], [8]), asks for the expected number of draws needed to collect all n unique coupons under uniform sampling. In hypergraph language, it means

that the hyperedges are singletons, making each $i \in [n]$ have a degree of 1. The solution to this problem is nH_n , where H_n is the n th harmonic number, which grows as $n \log n$ for large n ([9], [10]). Extensions of this problem have considered non-uniform sampling or collecting subsets instead of single elements ([12]–[15]).

Recent works have explored many connections between the coupon collector’s problem and DNA storage systems ([5], [16]–[20]). These include problems such as minimizing coverage depth, ensuring random access, and analyzing fragment overlap structures. In this paper, a connection between *shotgun sequencing* and the coupon collector’s problem is drawn. Shotgun sequencing is a powerful method in DNA sequencing, where long DNA molecules are fragmented into overlapping pieces that are sequenced and computationally assembled ([1]–[3]).

In recent years, shotgun sequencing principles have extended beyond biology into information storage, such as DNA storage systems ([4]). These applications motivate probabilistic models that quantify coverage requirements under structured sampling.

A central question in this setting is: how many fragments must be sampled to ensure full coverage of the original sequence? While practical sequencing involves additional considerations such as assembly heuristics and error correction, the probabilistic aspect of achieving complete coverage is a fundamental challenge with implications for both biology and information theory.

Motivated by the structure of shotgun sequencing, we consider a new model in which each observation corresponds to a *consecutive window* within a sequence. Unlike classical models where subsets are arbitrary, consecutive windows introduce strong dependencies that alter the dynamics of coverage. We consider both cyclic and non-cyclic models, and refer to them correspondingly *the cyclic windows coverage problem* and *the windows coverage problem*.

The rest of the paper is organized as follows. Section II introduces the definitions that are used throughout the paper, and states 4 problems, with two of them being the window models mentioned above (Problems 1 and 2), another problem is a continuous version of Problem 2 (Problem 3), and lastly a problem which is referred to as *the batch coverage problem* (Problem 4) which is the problem of drawing coupons in groups of the same size and any subset is equally likely to be sampled. Inspired by recent work that recasts expectations as combinatorial counts (e.g., [17]), in Section III we develop a compact, unifying tool that reduces the computation of expected coverage time

to counting what we call *recovery sets* in a corresponding hypergraph, which are sets of hyperedges that together contain all vertices. Section IV utilizes this tool to solve Problem 1, and presents more compact solutions for special cases. Section V is the second to make us of this method, and solves Problem 2. Section VI makes a comparison between Problems 2 and 3, allowing to derive bounds and understand the asymptotic behavior of both Problems 1 and 2. Section VII shows lower and upper bounds for the batch coverage model, which were derived essentially at the same time as the independent result of [28]. It introduces a question regarding all coverage problems for which the corresponding hypergraph is uniform ℓ -regular, i.e., all hyperedges are of the same size, and all vertices have the same degree, and suggests that a monotonicity property holds: Restricting the drawings to only a subset of the hyperedges, while keeping uniformity might accelerate coverage in expectation, with a proof provided for some cases, and a conjecture that this is always the case and that among all uniform regular models, the batch coverage model achieves the highest expected coverage time. In this sense, the batch model serves as a universal upper bound for the expected time to full coverage across all such models. Although a lack of formal proof, partial proofs and empirical evidence strongly supports the conjecture that the batch coverage model indeed forms the upper bound for all uniform regular models. Nevertheless, a provable universal upper bound is established, slightly above the batch model's expectation, guaranteeing that all those coverage models remains within the same asymptotic scale. Lastly, Section VIII outlines the remaining questions and a few directions for future research.

II. DEFINITIONS, NOTATIONS AND PROBLEM STATEMENTS

For the problems studied in this paper, the *items* to be collected are natural numbers, i.e. for any $n \in \mathbb{N}^+$, define $[n] = \{0, 1, \dots, n-1\}$. These would be considered *collectable* objects, as well as *items*, *elements* or *positions*. For integers $i < j$, define $[i, j] = \{i, i+1, \dots, j\}$. For the window models, the following notations are set. For any n, ℓ , define $\text{win}_\ell(i) = [i, i+\ell-1] \subseteq [n]$, a length- ℓ window starting at i , and denote the set of windows $\mathcal{W}_{n,\ell} = \{\text{win}_\ell(i) : i \in [n-\ell+1]\}$. Similarly, for the cyclic case, define $\text{win}_\ell^\circ(i) = \{i, i+1, \dots, i+\ell-1\} \bmod n$, a *cyclic-window* starting at i , and define the set of *cyclic-windows* $\mathcal{W}_{n,\ell}^\circ := \{\text{win}_\ell^\circ(i) : i \in [n]\}$. The continuous setting also follows a similar structure. For $\alpha \in [0, 1)$ and $a \in (0, 1)$, define $\text{arc}_a^\circ(\alpha) = \{(\alpha+x) \bmod 1 : x \in [0, a)\}$, the length- a *cyclic-arc* starting at α , and define the set of *cyclic-arcs* $\mathcal{A}_a^\circ := \{\text{arc}_a^\circ(\alpha) : \alpha \in [0, 1)\}$, and define for any $\gamma \in \mathcal{A}_a^\circ$, its starting point as $\text{Pos}(\gamma)$. Set $H_n := \sum_{k=1}^n \frac{1}{k}$ to be the n -th harmonic number. For asymptotic results we denote $a_n \approx b_n$ if $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$ and $a_n \lesssim b_n$ if $\limsup_{n \rightarrow \infty} \frac{a_n}{b_n} \leq 1$ (equivalently, $a_n \leq (1+o(1))b_n$).

We are now ready to formally define the problems studied in this paper.

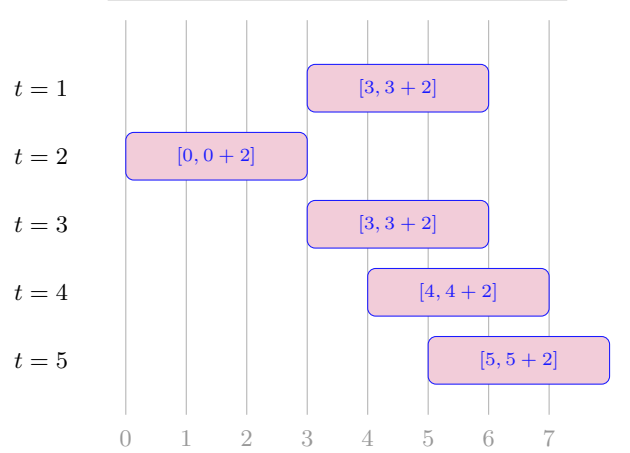


Figure 1: Example run of Problem 1 with $n = 8$ and $\ell = 3$; coverage at $T_{n,\ell} = 5$.

Problem 1. The Windows Coverage Problem. Consider the process of drawing i.i.d. $\omega_1, \omega_2, \dots$ uniformly at random from $\mathcal{W}_{n,\ell}$. The coverage time is the random variable defined as

$$T_{n,\ell} = \min \left\{ t : \bigcup_{r=1}^t \omega_r = [n] \right\}.$$

The problem is to determine the expected coverage time, denoted by $\mathbb{E}[T_{n,\ell}]$.

Example 1. Let $n = 8$ and $\ell = 3$. Consider the i.i.d. draws $\omega_1, \dots, \omega_5 \in \mathcal{W}_{8,3}$ given by their starting positions

$$i_1 = 3, \quad i_2 = 0, \quad i_3 = 3, \quad i_4 = 4, \quad i_5 = 5.$$

The corresponding windows are $\omega_1 = [3, 5]$, $\omega_2 = [0, 2]$, $\omega_3 = [3, 5]$, $\omega_4 = [4, 6]$, and $\omega_5 = [5, 7]$. Their unions after each draw are $\{3, 4, 5\}$, $\{0, 1, 2, 3, 4, 5\}$, $\{0, 1, 2, 3, 4, 5\}$, $\{0, 1, 2, 3, 4, 5, 6\}$, and finally $[n] = \{0, 1, 2, 3, 4, 5, 6, 7\}$, so coverage occurs at time $T_{n,\ell} = 5$. An illustration appears in Figure 1.

The cyclic version of Problem 1 is defined next.

Problem 2. The Cyclic Windows Coverage Problem. Consider the process of drawing i.i.d. $\nu_1, \nu_2, \dots$ uniformly at random from $\mathcal{W}_{n,\ell}^\circ$. The cyclic coverage time is the random variable defined as

$$T_{n,\ell}^\circ = \min \left\{ t : \bigcup_{r=1}^t \nu_r = [n] \right\}.$$

The problem is to determine the expected cyclic coverage time, denoted by $\mathbb{E}[T_{n,\ell}^\circ]$.

Example 2. Let $n = 10$ and $\ell = 4$. Consider the i.i.d. draws $\nu_1, \dots, \nu_5 \in \mathcal{W}_{10,4}^\circ$ given by their starting positions

$$i_1 = 8, \quad i_2 = 1, \quad i_3 = 8, \quad i_4 = 4, \quad i_5 = 6,$$

so that $\nu_1 = \{8, 9, 0, 1\}$, $\nu_2 = \{1, 2, 3, 4\}$, $\nu_3 = \{8, 9, 0, 1\}$, $\nu_4 = \{4, 5, 6, 7\}$, and $\nu_5 = \{6, 7, 8, 9\}$. Their unions after each draw are $\{8, 9, 0, 1\}$, $\{0, 1, 2, 3, 4, 8, 9\}$,

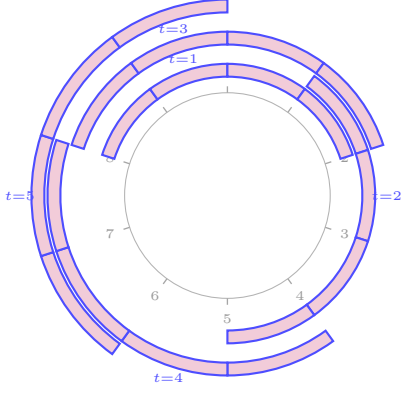


Figure 2: Example run of Problem 2 with $n = 10$ and $\ell = 4$; coverage at $T_{n,\ell}^\circ = 4$.

$\{0, 1, 2, 3, 4, 8, 9\}$, and finally $[n] = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and hence coverage occurs at time $T_{n,\ell}^\circ = 4$ (the fifth draw is redundant). An illustration appears in Figure 2.

A continuous version of Problem 2 is defined as follows.

Problem 3. The Arcs Coverage Problem. Fix a circle with a circumference of length 1. Fix an arc length $a \in (0, 1)$. Consider the process of drawing i.i.d. $\gamma_1, \gamma_2, \dots$ uniformly at random from $\mathcal{A}_a^\circ$, and put each γ_i on the circle's circumference starting at point $\text{Pos}(\gamma_i)$ and going clockwise until the entire circumference, which can be described as $[0, 1) \bmod 1$ cyclically, is covered. The coverage time is the random variable which is defined as

$$\mathcal{T}_a = \min \left\{ t : \bigcup_{r=1}^t \gamma_r = [0, 1) \right\}.$$

The problem is to determine the expected coverage time, denoted by $\mathbb{E}[\mathcal{T}_a]$.

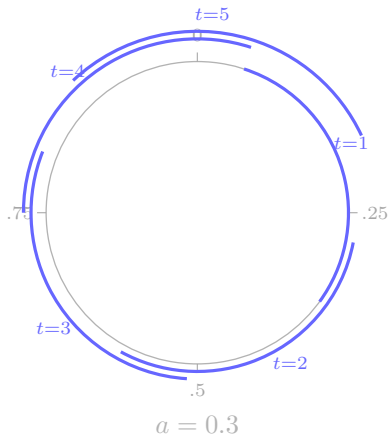


Figure 3: Example run of Problem 3 with $a = 0.3$. On round t , the arc $I_t = [U_t, U_t + a) \bmod 1$ is placed; coverage at $\mathcal{T}_{0.3} = 5$.

Example 3. Fix $a = 0.3$. Consider i.i.d. draws $\gamma_1, \dots, \gamma_4 \in \mathcal{A}_a^\circ$ with starting points

$$x_1 = 0.05, \quad x_2 = 0.28, \quad x_3 = 0.51, \quad x_4 = 0.75,$$

so that, with $\gamma_t = \text{arc}_a^\circ(x_t) = [x_t, x_t + a) \bmod 1$, the arcs are $\gamma_1 = [0.05, 0.35)$, $\gamma_2 = [0.28, 0.58)$, $\gamma_3 = [0.51, 0.81)$, and $\gamma_4 = [0.75, 1) \cup [0, 0.05)$. The unions after each draw are $[0.05, 0.35)$, $[0.05, 0.58)$, $[0.05, 0.81)$, and finally $[0, 1)$. Hence coverage occurs at time $\mathcal{T}_{0.3} = 4$. An illustration appears in Figure 3.

The continuous model was introduced and analyzed in [22]–[25], where the solution for this problem was introduced by

$$\mathbb{E}[\mathcal{T}_a] = 1 + \sum_{k=1}^{\lfloor 1/a \rfloor} (-1)^{k-1} \frac{(1 - ka)^{k-1}}{(ka)^{k+1}},$$

and in [23] and [24], it was shown that

$$\mathbb{E}[\mathcal{T}_a] = \frac{1}{a} \left(\log \frac{1}{a} + \log \log \frac{1}{a} + \gamma + o(1) \right), \text{ as } a \rightarrow 0.$$

Section VI establishes its relationship to Problem 2, as this problem presents a model that can be considered as the continuous version of the cyclic windows coverage problem. The last problem we consider in this paper is defined next.

Problem 4. The Batch Coverage Problem. Fix n and ℓ . Consider the process of drawing subsets $S_1, S_2, \dots$, referred to as batches, of size $|S_i| = \ell$ uniformly from $\{S \subseteq [n] : |S| = \ell\}$. The coverage time is defined as

$$T_{\binom{[n]}{\ell}} = \min \left\{ t : \bigcup_{r=1}^t S_r = [n] \right\}.$$

The problem is to determine the expected coverage time, denoted by $\mathbb{E}[T_{\binom{[n]}{\ell}}]$.

Example 4. Let $n = 7$ and $\ell = 2$. Consider the i.i.d. draws $S_1, \dots, S_7 \subseteq [n]$ with $|S_t| = 2$ given by $S_1 = \{1, 4\}$, $S_2 = \{0, 4\}$, $S_3 = \{0, 2\}$, $S_4 = \{2, 5\}$, $S_5 = \{1, 6\}$, $S_6 = \{2, 4\}$ and $S_7 = \{3, 6\}$. Their unions after each draw are $\{1, 4\}$, $\{0, 1, 4\}$, $\{0, 1, 2, 4\}$, $\{0, 1, 2, 4, 5\}$, $\{0, 1, 2, 4, 5, 6\}$, $\{0, 1, 2, 4, 5, 6\}$ and $\{0, 1, 2, 3, 4, 5, 6\} = [n]$. Thus coverage occurs at time $T_{\binom{[n]}{\ell}} = 7$. An illustration appears in Figure 4.

This model has been studied in [12]–[14], where combinatorial solutions for the moments of $T_{\binom{[n]}{\ell}}$ and particularly for the expectation are given. The first solution for the expectation was given by Pólya in [12], where he proved that the expectation is

$$\mathbb{E}[T_{\binom{[n]}{\ell}}] = \binom{n}{\ell} \sum_{i=0}^{n-1} \frac{(-1)^{n-i+1} \binom{n}{i}}{\binom{n}{\ell} - \binom{i}{\ell}},$$

and analyzed asymptotics for the case $\ell = O(1)$, and obtained for this case that

$$\mathbb{E}[T_{\binom{[n]}{\ell}}] \approx \left(\frac{n+\frac{1}{2}}{\ell} - \frac{1}{2} \right) (\log n + \gamma) + \frac{1}{2} + o(1).$$

A more general setting was solved by [13]. Given $A \subseteq [n]$ with $|A| = m$, let $Z_k(A)$ be the number of draws needed to see at least $k \leq m$ distinct elements of A . The exact formula was obtained

$$\mathbb{E}[Z_k(A)] = \binom{n}{\ell} \sum_{i=0}^{k-1} \frac{(-1)^{k-i+1} \binom{m}{i} \binom{m-i-1}{m-k}}{\binom{n}{\ell} - \binom{i+n-m}{\ell}}.$$

Section VII presents a more transparent expression valid for all $1 \leq \ell \leq n$, and discusses its relationship to all other models of coverage.

III. A GENERAL COMBINATORIAL APPROACH FOR SOLVING COVERAGE PROBLEMS

Before we study the proposed coverage problems in this paper, we first seek to establish a general combinatorial approach which will be used in our analysis throughout the paper. This approach, motivated by the work of [17], produces a tool which converts the problem of calculating the expectation of the coverage time to a combinatorial problem in which the expectation can be expressed in a combinatorial fashion. While the motivation in [17] was to study the so-called *random access problem* in DNA-based storage systems, the approach can be extended in a more general form.

All coverage problems share a common structure in which the expected time to full coverage admits a single, model-independent expression that only depends on the number of *recovery sets* that will be defined next. Thus, the task of computing the expectation reduces to a combinatorial problem of counting recovery sets. First, we show a general framework to present coverage problems using hypergraphs.

Definition 1. A hypergraph is a pair $\mathcal{H} = (V, E)$ where

- V is a finite set of vertices, and
- $E \subseteq 2^V$ consists of non-empty subsets of V , called hyperedges.

A hypergraph is said to be covering if $\bigcup_{e \in E} e = V$.

Definition 2. Let Ω be a set of n coverable objects, i.e., $|\Omega| = n$. A coverage model is specified by a finite family of observable blocks $\mathcal{A} = \{A_1, \dots, A_{N_{\mathcal{H}}}\}$, where each $A_i \subseteq \Omega$ is the set of objects revealed by a single draw. Equivalently, the coverage model can be presented using a hypergraph

$$\mathcal{H} = (V, E), \quad V = \Omega, \quad E = \mathcal{A} \subseteq 2^V,$$

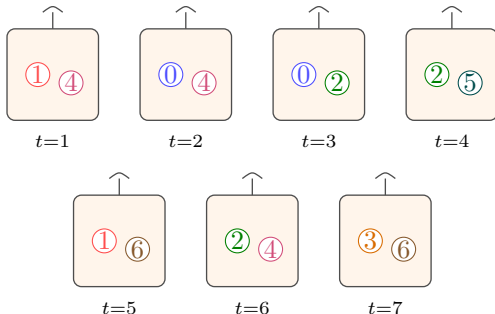


Figure 4: Example run of 4 with $n = 7$ and $\ell = 2$; coverage at $T_{(n)_\ell}^{(n)} = 7$.

where the vertices are the objects and the hyperedges are the observable blocks. The hypergraph $\mathcal{H}$ is assumed to be covering. On each round, an index $I_r \in [N_{\mathcal{H}}]$ is uniformly and independently sampled, revealing the I_r -th hyperedge and collecting its vertices A_{I_r} . The coverage time is defined to be

$$T_{\mathcal{H}} = \min \left\{ t : \bigcup_{r=1}^t A_{I_r} = \Omega \right\}.$$

A recovery set is a subset of the hyperedges $\mathcal{R} \subseteq \mathcal{A}$ such that $\bigcup_{\mathcal{E} \in \mathcal{R}} \mathcal{E} = \Omega$. The set of all recovery sets is defined as $\text{Rec}(\mathcal{A}) := \{ \mathcal{R} \subseteq \mathcal{A} : \mathcal{R} \text{ is a recovery set} \}$. Denote $M_{\mathcal{H}}$ as the maximal size of any non-recovery set, that is, $M_{\mathcal{H}} = \max \{ |\mathcal{R}| : \mathcal{R} \in 2^{\mathcal{A}} \setminus \text{Rec}(\mathcal{A}) \}$. And denote $\alpha_{\mathcal{H}}(s)$ to be the number of s -sized recovery sets. Explicitly $\alpha_{\mathcal{H}}(s) = |\{ \mathcal{R} \in \text{Rec}(\mathcal{A}) : |\mathcal{R}| = s \}|$. When the hypergraph is obvious from context, we omit the $\mathcal{H}$ subscript.

The following examples illustrate how a coverage problem can be presented by a hypergraph.

Example 5. Classical Coupon Collector's Problem. $V = [n]$, $E = \{\{i\} : i \in V\}$, $N_{\mathcal{H}} = n$. Each draw reveals a singleton; coverage means every vertex has been sampled.

For binary vectors $u, v \in \{0, 1\}^d$, their Hamming distance is $\text{dist}_H(u, v) = |\{i \in [d] : u_i \neq v_i\}|$. For a binary vector $v \in \{0, 1\}^d$ and radius $t \in \mathbb{N}$, the Hamming ball of radius t centered at v is $B_t(v) = \{u \in \{0, 1\}^d : \text{dist}_H(u, v) \leq t\}$.

Example 6. Hamming Hypercube Coverage. Fix $d, t \in \mathbb{N}$. Let $V = \{0, 1\}^d$ be the set of vertices of the d -dimensional Hamming hypercube, so $|V| = 2^d$. Let

$$E = \{ B_t(x) : x \in V \}$$

be the family of all Hamming balls of radius t . Then $N_{\mathcal{H}} = |V| = 2^d$. Each draw reveals a Hamming ball of radius t ; coverage means that every vertex lies in some revealed ball.

Theorem 1. Given a hypergraph $\mathcal{H}$,

$$\mathbb{E}[T_{\mathcal{H}}] = N_{\mathcal{H}} (H_{N_{\mathcal{H}}} - H_{N_{\mathcal{H}} - M_{\mathcal{H}} - 1}) - \sum_{s=0}^{M_{\mathcal{H}}} \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}-1}{s}}.$$

Proof (following [17]): By the tail-sum formula, $\mathbb{E}[T_{\mathcal{H}}] = \sum_{r \geq 1} \Pr(T_{\mathcal{H}} \geq r)$. Let μ_r denote the number of distinct hyperedges revealed after r rounds. By the law of total probability and since any set of size greater than $M_{\mathcal{H}}$ is a recovery set, $\Pr(T_{\mathcal{H}} \geq r) = \sum_{s=0}^{M_{\mathcal{H}}} \Pr(T_{\mathcal{H}} \geq r \mid \mu_{r-1} = s) \Pr(\mu_{r-1} = s)$. Conditioned on $\mu_{r-1} = s$, $\Pr(T_{\mathcal{H}} \geq r \mid \mu_{r-1} = s) = 1 - \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}}{s}}$, since $\alpha_{\mathcal{H}}(s)$ counts the size- s recovery sets among the $\binom{N_{\mathcal{H}}}{s}$ many s -subsets. Moreover, by the inclusion-exclusion principle,

$$\Pr(\mu_{r-1} = s) = \binom{N_{\mathcal{H}}}{s} \sum_{j=0}^s \binom{s}{j} (-1)^j \left(\frac{s-j}{N_{\mathcal{H}}} \right)^{r-1}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[T_{\mathcal{H}}] &= \sum_{r \geq 1} \sum_{s=0}^{M_{\mathcal{H}}} \left(1 - \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}}{s}}\right) \binom{N_{\mathcal{H}}}{s} \sum_{j=0}^s \binom{s}{j} (-1)^j \left(\frac{s-j}{N_{\mathcal{H}}}\right)^{r-1} \\ &= \sum_{s=0}^{M_{\mathcal{H}}} \left(\binom{N_{\mathcal{H}}}{s} - \alpha_{\mathcal{H}}(s)\right) \sum_{j=0}^s \binom{s}{j} (-1)^j \sum_{r \geq 0} \left(\frac{s-j}{N_{\mathcal{H}}}\right)^r \\ &= \sum_{s=0}^{M_{\mathcal{H}}} \left(\binom{N_{\mathcal{H}}}{s} - \alpha_{\mathcal{H}}(s)\right) \sum_{j=0}^s \binom{s}{j} (-1)^j \frac{N_{\mathcal{H}}}{N_{\mathcal{H}} - s + j}.\end{aligned}$$

Using the identity $\sum_{j=0}^s \binom{s}{j} (-1)^j \frac{N_{\mathcal{H}}}{N_{\mathcal{H}} - s + j} = \frac{1}{\binom{N_{\mathcal{H}}-1}{s}}$, we get

$$\begin{aligned}\mathbb{E}[T_{\mathcal{H}}] &= \sum_{s=0}^{M_{\mathcal{H}}} \frac{\binom{N_{\mathcal{H}}}{s} - \alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}-1}{s}} = \sum_{s=0}^{M_{\mathcal{H}}} \frac{\binom{N_{\mathcal{H}}}{s}}{\binom{N_{\mathcal{H}}-1}{s}} - \sum_{s=0}^{M_{\mathcal{H}}} \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}-1}{s}} \\ &= N_{\mathcal{H}} \sum_{s=0}^{M_{\mathcal{H}}} \frac{1}{N_{\mathcal{H}} - s} - \sum_{s=0}^{M_{\mathcal{H}}} \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}-1}{s}} \\ &= N_{\mathcal{H}} (H_{N_{\mathcal{H}}} - H_{N_{\mathcal{H}}-M_{\mathcal{H}}-1}) - \sum_{s=0}^{M_{\mathcal{H}}} \frac{\alpha_{\mathcal{H}}(s)}{\binom{N_{\mathcal{H}}-1}{s}}.\end{aligned}$$

This approach reduces the task of determining the expected coverage time to computing M and $\alpha_{\mathcal{H}}(s)$. Once these are known, the expectation follows immediately from Theorem 1, turning the probabilistic question into a purely combinatorial calculation.

Theorem 1 is now used to compute the expectation in Examples 5 and 6, illustrating how straightforward this approach is in practice.

Example 5 (continued). *Classical Coupon Collector's Problem.* There are no recovery sets of size $s \leq n-1$, i.e. $\alpha(s) = 0$ for $s \leq n-1$. Hence, by Theorem 1, the expected number of draws is nH_n .

Example 6 (continued). *Hamming Hypercube Coverage.* The maximum size of a non-recovery set is $M = N - |B_t| = 2^d - \sum_{j=0}^t \binom{d}{j}$, since a vertex v remains uncovered if and only if all balls containing v are omitted. Hence, by Theorem 1, the expected number of draws is

$$2^d (H_{2^d} - H_{|B_t|-1}) - \sum_{s=0}^{2^d - |B_t|} \frac{\alpha(s)}{\binom{2^d-1}{s}},$$

where $\alpha(s)$ counts all size- s collections of radius- t balls whose union is the entire cube.

From here on, for every coverage model throughout the paper, we will refer to the possible attainable subsets of collectable items as the hyperedges of the graph of items interchangeably.

IV. PROBLEM 1: THE WINDOWS COVERAGE MODEL

In this section we solve Problem 1, the window coverage problem.

A. General Solution to The Windows Coverage Problem

Using the approach derived in Section III, Problem 1 can be formulated using hypergraphs as follows. Denote

$$\mathcal{H}_{n,\ell} = (V, E), \quad V = [n], \quad E = \mathcal{A} = \mathcal{W}_{n,\ell},$$

so it holds that $N = |E| = n$ and $M = n - \ell$ (omitting the coverage of the rightmost position admits the largest non recovery set). Let $\alpha(s)$ denote the number of size- s recovery sets in $\mathcal{W}_{n,\ell}$. Then, by Theorem 1 it holds that,

$$\mathbb{E}[T_{n,\ell}] = (n - \ell + 1) H_{n-\ell+1} - \sum_{s=0}^{n-\ell} \frac{\alpha(s)}{\binom{n-\ell}{s}}. \quad (\text{IV.1})$$

The task then reduces to determining the values of $\alpha(s)$, yielding the following.

Lemma 1. *It holds that*

$$\begin{aligned}\mathbb{E}[T_{n,\ell}] &= (n - \ell + 1) H_{n-\ell+1} \\ &\quad - \sum_{s=\lceil \frac{n}{\ell} \rceil}^{n-\ell} \sum_{i=0}^{\lfloor \frac{n-\ell-s+1}{\ell} \rfloor} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}.\end{aligned}$$

In particular,

$$\mathbb{E}[T_{n,2}] = (n-1) H_{n-1} - \sum_{s=\lceil \frac{n}{2} \rceil}^{n-2} \frac{\binom{s-1}{n-s-1}}{\binom{n-2}{s}}.$$

Proof: The possible start positions for any window are $\{0, 1, \dots, n - \ell\}$. Any size- s recovery set must include the extreme windows. Denote a recovery set $\mathcal{R}$, and denote its corresponding set of start positions in an increasing order

$$\mathcal{I}_{\mathcal{R}} = \{x_1 = 0 < x_2 < \dots < x_s = n - \ell\}.$$

Let the gaps be $\delta_i = x_{i+1} - x_i \in \{1, \dots, \ell\}$ with $\sum_{i=1}^{s-1} \delta_i = n - \ell$, and set $\delta'_i = \delta_i - 1 \in \{0, \dots, \ell - 1\}$. Then

$$\sum_{i=1}^{s-1} \delta'_i = \underbrace{n - \ell - s + 1}_{\text{telescopic}}.$$

For each $\mathcal{R} \in \text{Rec}(\mathcal{A})$, $\mathcal{I}_{\mathcal{R}}$ corresponds to a unique set of gaps bijectively. Hence, $\alpha(s)$ equals the coefficient of $x^{n-\ell-s+1}$ in the generating function $(1+x+\dots+x^{\ell-1})^{s-1}$. Therefore, by a standard derivation, it can be shown that $\alpha(s) = \sum_{i=0}^{\min(s-1, \lfloor \frac{n-\ell-s+1}{\ell} \rfloor)} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}$. Note that $\alpha(s) = 0$ unless $\lceil \frac{n}{\ell} \rceil \leq s$, and note that for $\ell = 2$, we get a simpler expression since $\delta'_i \in \{0, 1\}$, and thus $\alpha(s) = \binom{s-1}{n-s-1}$. ■

Lemma 1 fully solves Problem 1 and gives an exact expression for the coverage time. Further, for the general case (arbitrary ℓ), algebraic manipulation yields a more informative closed form.

Theorem 2. *For all integers $n > \ell \geq 1$, it holds that*

$$\begin{aligned}\mathbb{E}[T_{n,\ell}] &= \frac{3}{2}(n - \ell + 1) \\ &\quad + \sum_{i=1}^{\lceil n/\ell \rceil - 2} \left(\sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^{i+1} \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \right).\end{aligned}$$

Proof: Denote $c = \lceil \frac{n}{\ell} \rceil$. It holds that $\frac{n}{c} \leq \ell < \frac{n}{c-1}$, and $c \leq s \leq n - \ell$. Thus,
 $\lfloor \frac{n-\ell-s+1}{\ell} \rfloor \leq \frac{n-\ell-s+1}{\ell} \leq \frac{n-\frac{n}{c}-c+1}{\frac{n}{c}} = c-1 - \frac{c(c+1)}{n} < c-1$.

Therefore,

$$\begin{aligned} \mathbb{E}[T_{n,\ell}] &= (n-\ell+1)H_{n-\ell+1} - \sum_{s=c}^{n-\ell} \sum_{i=0}^{c-2} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= (n-\ell+1)H_{n-\ell+1} - \sum_{i=0}^{c-2} \sum_{s=c}^{n-\ell} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}, \end{aligned}$$

where the last equality holds by changing the order of summation.

For all $i \in \{0, 1, \dots, c-2\}$, we must ensure that the binomial coefficient $\binom{n-\ell-i\ell-1}{s-2}$ does not vanish. This requires that $s-2 \leq n-\ell-i\ell-1$, or equivalently, $s \leq n-\ell-i\ell+1$. Therefore, we can write

$$\begin{aligned} \mathbb{E}[T_{n,\ell}] &= (n-\ell+1) \cdot H_{n-\ell+1} \\ &\quad - \sum_{i=0}^{c-2} \sum_{s=c}^{\min\{n-\ell, n-\ell-i\ell+1\}} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}. \end{aligned}$$

Consider the sum,

$$\begin{aligned} &\sum_{i=0}^{c-2} \sum_{s=c}^{\min\{n-\ell, n-\ell-i\ell+1\}} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=c}^{n-\ell} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} + \sum_{i=1}^{c-2} \sum_{s=c}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=2}^{n-\ell} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} - \sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &\quad + \sum_{i=1}^{c-2} \sum_{s=c}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}. \end{aligned}$$

It can be shown algebraically that

$$\sum_{s=2}^{n-\ell} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} = (n-\ell+1) \left(H_{n-\ell+1} - \frac{3}{2} \right).$$

Therefore, the sum becomes

$$\begin{aligned} &(n-\ell+1) \left(H_{n-\ell+1} - \frac{3}{2} \right) - \sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &\quad + \sum_{i=1}^{c-2} \sum_{s=c}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}. \end{aligned}$$

Substituting back into the expectation, we get,

$$\begin{aligned} \mathbb{E}[T_{n,\ell}] &= (n-\ell+1) \cdot H_{n-\ell+1} \\ &\quad - \left[(n-\ell+1) \left(H_{n-\ell+1} - \frac{3}{2} \right) - \sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} \right. \\ &\quad \left. + \sum_{i=1}^{c-2} \sum_{s=c}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \right] \\ &= \frac{3}{2}(n-\ell+1) + \sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &\quad - \sum_{i=1}^{c-2} \sum_{s=c}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}. \end{aligned}$$

Further, by adding and subtracting the term

$$\sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}},$$

we get

$$\begin{aligned} \mathbb{E}[T_{n,\ell}] &= \frac{3}{2}(n-\ell+1) + \sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &\quad + \sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &\quad - \sum_{i=1}^{c-2} \sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \frac{3}{2}(n-\ell+1) \\ &\quad + \left[\sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} + \sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \right] \\ &\quad - \sum_{i=1}^{c-2} \sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^i \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}. \end{aligned}$$

Observe that the sum inside the square brackets is equal to zero. Indeed,

$$\begin{aligned} &\sum_{s=2}^{c-1} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} + \sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=2}^{c-1} (-1)^0 \binom{s-1}{0} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} + \sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=1}^{c-1} (-1)^0 \binom{s-1}{0} \frac{\binom{n-\ell-1}{s-2}}{\binom{n-\ell}{s}} + \sum_{i=1}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{i=0}^{c-2} \sum_{s=i+1}^{c-1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=1}^{c-1} \sum_{i=0}^{s-1} (-1)^i \binom{s-1}{i} \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ &= \sum_{s=1}^{c-1} \frac{1}{\binom{n-\ell}{s}} \sum_{i=0}^{s-1} (-1)^i \binom{s-1}{i} \binom{n-\ell-i\ell-1}{s-2}. \end{aligned}$$

For any fixed $s \in \{1, 2, \dots, c-1\}$, the inner sum is an inclusion-exclusion formula for counting the number of binary vectors of length $n-\ell-1$ and weight exactly $s-2$, such that for each $i \in \{1, \dots, s-1\}$, there is at least one bit of value one among the bits in positions $\{(i-1)\ell+1, \dots, i\ell\}$. This would require the vector to have at least $s-1$ nonzero entries (i.e., weight at least s), which contradicts the condition that the weight is $s-2$. Therefore, the number of options is zero. Hence the entire bracketed sum vanishes. This concludes that the expectation is

$$\begin{aligned} \mathbb{E}[T_{n,\ell}] &= \frac{3}{2}(n-\ell+1) \\ &\quad + \sum_{i=1}^{c-2} \left(\sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^{i+1} \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \right). \end{aligned}$$

Recall that $c = \lceil \frac{n}{\ell} \rceil$, can be substituted back and yield the theorem's statement. ■

Thus, a final expression for the expectation is obtained. In Section VI, we show that for $\ell = \omega(\log(n))$, the double

sum $S_{n,\ell}$ satisfies $S_{n,\ell} = o(\frac{3}{2}(n-\ell+1))$, establishing that the expectation is dominated by its linear term (extremes coverage), and for $\ell = o(\log(n))$, $\mathbb{E}[T_{n,\ell}] \approx \frac{n}{\ell} \log(\frac{n}{\ell})$ (middle coverage), leaving only the $\ell = \Theta(\log(n))$ regime's asymptotics ambiguous.

B. Large Windows

This section presents the expected coverage time for special regimes of ℓ relative to n .

Theorem 3. For all integers $\frac{n}{2} \leq \ell < n$, it holds that

$$\mathbb{E}[T_{n,\ell}] = \frac{3}{2}(n-\ell+1).$$

Proof: Write $M = n - \ell + 1$ for the number of windows. When $\ell \geq n/2$, the two extreme windows, $\text{win}_\ell(0) = [0, \ell - 1]$ and $\text{win}_\ell(n - \ell) = [n - \ell, n - 1]$, already cover all positions (each endpoint 0 and $n - 1$ belongs to exactly one of them). Hence full coverage occurs if and only if both extreme windows have appeared. Draws are i.i.d. uniform over the M windows. The waiting time to see either extreme window is geometric with success probability $2/M$, of expectation $M/2$. After the first extreme window is seen, the waiting time to see the other extreme window is geometric with success probability $1/M$, of expectation M . By linearity of expectation,

$$\mathbb{E}[T_{n,\ell}] = \frac{M}{2} + M = \frac{3}{2}M = \frac{3}{2}(n-\ell+1).$$

Theorem 4. For all integers $\frac{n}{3} \leq \ell < \frac{n}{2}$, it holds that

$$\mathbb{E}[T_{n,\ell}] = \frac{3}{2}(n-\ell+1) + (n-\ell+1) \frac{6n-10\ell}{\ell(\ell+1)(\ell+2)(\ell+3)}.$$

Proof: We first observe that full coverage occurs if and only if both extreme windows were sampled and all $n - 2\ell$ middle positions are covered. So only coverage of the middle $n - 2\ell + 2$ positions matters. We model the process using a three-layered Markov chain. States are denoted by S_i^k , where $i \in \{0, 1, \dots, n - 2\ell\}$ counts the number of unseen middle positions, and $k \in \{0, 1, 2\}$ indicates how many extreme windows have been observed. Note that the uncovered middle positions always appear continuously. For each state S_i^k , the number of windows that cover all i uncovered middle positions is $\ell - i + 1$, giving the transition probability $\Pr[S_i^k \rightarrow S_0^k] = \frac{\ell-i+1}{n-\ell+1}$. For each $1 \leq j < i$, the number of windows covering exactly $i - j$ new middle positions is exactly 2, giving a transition probability $\Pr[S_i^k \rightarrow S_j^k] = \frac{2}{n-\ell+1}$. The number of windows that reveal an uncovered extreme window is $2 - k$, giving a transition probability $\Pr[S_i^k \rightarrow S_i^{k+1}] = \frac{2-k}{n-\ell+1}$. Since all probabilities sum to 1, the transition probability of no new covered positions is $\Pr[S_i^k \rightarrow S_i^k] = \frac{n-2\ell-i+k}{n-\ell+1}$. We solve the markov chain in a bottom up, as can be demonstrated visually in Figure 5.

For $k = 2$, using the transition probabilities, we derive the following recursive relation,

$$\mathbb{E}[S_i^2 \rightarrow S_0^2] = 1 + \frac{n-2\ell-i+2}{n-\ell+1} \mathbb{E}[S_i^2 \rightarrow S_0^2] + \frac{2}{n-\ell+1} \sum_{j=1}^{i-1} \mathbb{E}[S_j^2 \rightarrow S_0^2],$$

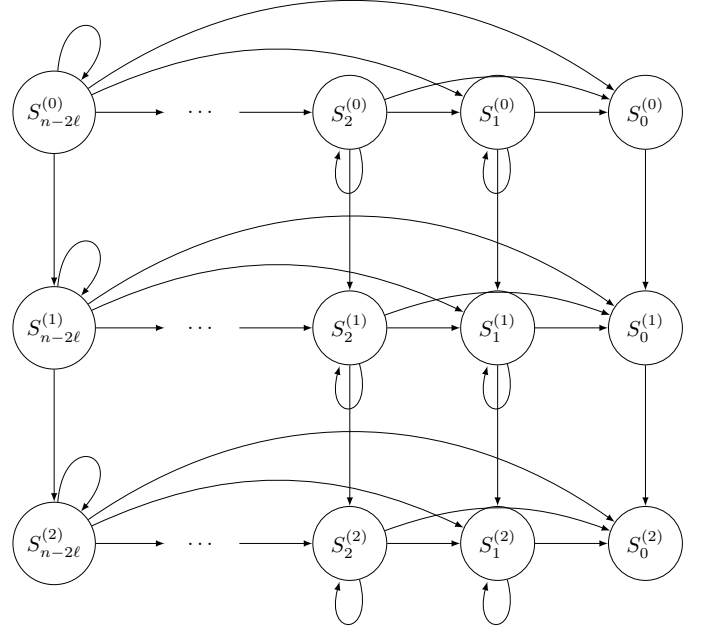


Figure 5: Layered Markov chain, representing the case $\ell \geq \frac{n}{3}$.

and hence

$$\mathbb{E}[S_i^2 \rightarrow S_0^2] = \frac{n-\ell+1}{\ell-1+i} + \frac{2}{\ell-1+i} \sum_{j=1}^{i-1} \mathbb{E}[S_j^2 \rightarrow S_0^2].$$

And it can be shown by induction that

$$E[S_i^2 \rightarrow S_0^2] = \frac{n-\ell+1}{\ell} + \frac{n-\ell+1}{\ell(\ell+1)}(i-1).$$

Similarly, for $k = 1$, it can be derived that

$$\mathbb{E}[S_i^1 \rightarrow S_0^2] = \frac{(\ell-i+2)(n-\ell+1)}{\ell+i} + \frac{n-\ell+1}{\ell(\ell+1)} + \frac{2}{\ell+i} \sum_{j=1}^{i-1} \mathbb{E}[S_j^1 \rightarrow S_0^2],$$

and by induction we get

$$\mathbb{E}[S_i^1 \rightarrow S_0^2] = (n-\ell+1) + \frac{n-\ell+1}{\ell(\ell+1)} + \frac{2(n-\ell+1)}{\ell(\ell+1)(\ell+2)}(i-1).$$

For $k = 0$, the relation becomes

$$\begin{aligned} \mathbb{E}[S_i^0 \rightarrow S_0^2] &= \frac{n-\ell+1}{\ell+1+i} + \frac{\ell-i+1}{\ell+1+i} \cdot \frac{3(n-\ell+1)}{2} \\ &\quad + \frac{2}{\ell+1+i} \left((n-\ell+1) + \frac{n-\ell+1}{\ell(\ell+1)} + \frac{2(n-\ell+1)}{\ell(\ell+1)(\ell+2)}(i-1) \right) \\ &\quad + \frac{2}{\ell+1+i} \sum_{j=1}^{i-1} \mathbb{E}[S_j^0 \rightarrow S_0^2]. \end{aligned}$$

And a final induction argument gives

$$\mathbb{E}[S_i^0 \rightarrow S_0^2] = \frac{3(n-\ell+1)}{2} + \frac{2(n-\ell+1)}{\ell(\ell+1)(\ell+2)} + \frac{6(n-\ell+1)}{\ell(\ell+1)(\ell+2)(\ell+3)}(i-1).$$

Note that $\mathbb{E}[T_{n,\ell}] = \mathbb{E}[S_{n-2\ell}^0 \rightarrow S_0^2]$, therefore, substituting into the closed-form expression yields

$$\mathbb{E}[T_{n,\ell}] = \frac{3(n-\ell+1)}{2} + \frac{2(n-\ell+1)}{\ell(\ell+1)(\ell+2)} + \frac{6(n-\ell+1)(n-2\ell-1)}{\ell(\ell+1)(\ell+2)(\ell+3)}.$$

Rearranging terms yields the expression in the theorem. ■

Note that this result also gives the expected number of steps required for getting from any $i \in [n]$ uncovered positions, to j .

After analyzing two specific cases representing the scenarios of large windows ($\ell \geq \frac{n}{3}$), lastly an analysis of a more general result is given in the next theorem.

Theorem 5. *For any $\ell \geq \frac{n}{c}$ with a constant $c > 1$, it holds that*

$$\mathbb{E}[T_{n,\ell}] = \frac{3}{2}(n - \ell + 1) + O\left(\frac{1}{n^2}\right).$$

Proof: Denote $c = \lceil \frac{n}{\ell} \rceil$. By Theorem 2,

$$\mathbb{E}[T_{n,\ell}] = \frac{3}{2}(n - \ell + 1) + \sum_{i=1}^{c-2} \sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^{i+1} \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}}.$$

The proof proceeds by bounding the absolute sum. It holds that

$$\begin{aligned} & \sum_{i=1}^{c-2} \sum_{s=i+1}^{n-\ell-i\ell+1} (-1)^{i+1} \binom{s-1}{i} \cdot \frac{\binom{n-\ell-i\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ & \stackrel{(a)}{\leq} (c-2) \sum_{s=2}^{n-2\ell+1} s^{c-2} \cdot \frac{\binom{n-2\ell-1}{s-2}}{\binom{n-\ell}{s}} \\ & = (c-2) \sum_{s=2}^{n-2\ell+1} s^{c-2} \cdot \frac{s(s-1)}{(n-\ell)(n-\ell-1)} \prod_{j=1}^{s-2} \left(1 - \frac{\ell-1}{n-\ell-j-1}\right) \\ & \stackrel{(b)}{\leq} \frac{(c-2)}{(n-\ell-1)^2} \sum_{s=2}^{n-2\ell+1} s^c \cdot \prod_{j=1}^{s-2} e^{-\frac{\ell-1}{n-\ell-j-1}} \\ & = \frac{(c-2)}{(n-\ell-1)^2} \sum_{s=2}^{n-2\ell+1} s^c \cdot e^{-\sum_{j=1}^{s-2} \frac{\ell-1}{n-\ell-j-1}} \\ & \stackrel{(c)}{\leq} \frac{(c-2)}{(n-\ell-1)^2} \sum_{s=2}^{n-2\ell+1} s^c e^{-\frac{(\ell-1)(s-2)}{n-\ell}} \\ & = \frac{(c-2)e^{2\alpha}}{(n-\ell-1)^2} \sum_{s=2}^{n-2\ell+1} s^c e^{-\frac{\ell-1}{n-\ell}s} \\ & \stackrel{(d)}{\leq} \frac{(c-2)e^2}{(n-\ell-1)^2} \sum_{s=2}^{\infty} s^c e^{-\frac{1}{2(c-1)}s} = O\left(\frac{1}{n^2}\right). \end{aligned}$$

Where (a) is true since $i \leq c-2$, (b) is true since $1-x \leq e^{-x}$, (c) is true since we took the maximal term and (d) is true since $\frac{\ell-1}{n-\ell} \geq \frac{1}{2(c-1)}$, and since $\sum_{s=1}^{\infty} s^a \cdot e^{-bs}$ converges for any constants $a, b > 0$. ■

V. PROBLEM 2: THE CYCLIC WINDOWS MODEL

This section analyses Problem 2, the cyclic version of the windows coverage problem. First, a general solution is being presented. By applying Section III's approach, we get that in the cyclic case, the hypergraph is

$$\mathcal{H}_{n,\ell}^{\circ} = (V, E), \quad V = [n], \quad E = \mathcal{A} = \mathcal{W}_{n,\ell}^{\circ}.$$

$N = |E| = n$. Every position lies in exactly ℓ cyclic windows, hence the largest non-recovery set has size $M = n - \ell$ (omit all ℓ windows covering one fixed position). Let $\alpha^{\circ}(s)$ be the number of cyclic recovery sets of size s .

As in Problem 1, $\alpha^{\circ}(s) = 0$ unless $s \geq \lceil \frac{n}{\ell} \rceil$. Therefore, by Theorem 1,

$$\mathbb{E}[T_{n,\ell}^{\circ}] = n(H_n - H_{\ell-1}) - \sum_{s=\lceil \frac{n}{\ell} \rceil}^{n-\ell} \frac{\alpha^{\circ}(s)}{\binom{n-1}{s}}. \quad (\text{V.1})$$

In the same manner, the task reduces to determining $\alpha^{\circ}(s)$, yielding the following.

Theorem 6. *For all integers $n > \ell \geq 1$, it holds that*

$$\begin{aligned} \mathbb{E}[T_{n,\ell}^{\circ}] &= n(H_n - H_{\ell-1}) \\ &\quad - n \sum_{s=\lceil \frac{n}{\ell} \rceil}^{n-\ell} \frac{1}{s} \sum_{j=0}^{\lfloor \frac{n-s}{\ell} \rfloor} (-1)^j \binom{s}{j} \cdot \frac{\binom{n-\ell j-1}{s-1}}{\binom{n-1}{s}}. \end{aligned}$$

In particular,

$$\mathbb{E}[T_{n,2}^{\circ}] = n(H_n - 1) - \sum_{s=\lceil \frac{n}{2} \rceil}^{n-2} \frac{\binom{s-1}{n-s-1} + \binom{s}{n-s}}{\binom{n-1}{s}}.$$

Proof: We calculate the value of $\alpha^{\circ}(s)$ for $s \geq \lceil \frac{n}{\ell} \rceil$ and we first consider the case $\ell = 2$. Denote a recovery set $\mathcal{R}$. Let $w^* = \{n-1, 0\}$ be the wrap-around window. We split according to whether $w^* \in \mathcal{R}$. In case $w^* \notin \mathcal{R}$, the problem is exactly as in Problem 1, hence we have $\binom{s-1}{n-s-1}$ recovery sets not containing w^* . In the case where $w^* \in \mathcal{R}$, $\{1, 2, \dots, n-2\}$ must be covered using $s-1$ windows from the remaining $n-1$ edges $\{0, 1\}, \{1, 2\}, \dots, \{n-2, n-1\}$. Denote $\mathcal{R}$'s corresponding set of start positions by

$$\mathcal{I}_{\mathcal{R}} = \{x_1 < x_2 < \dots < x_{s-1}\}.$$

For full coverage, the extremes must satisfy both $x_1 \in \{0, 1\}$ and $x_{s-1} \in \{n-3, n-2\}$, so there are exactly 4 endpoint options. As in Problem 1, set the gaps $\delta_i = x_{i+1} - x_i \in \{1, 2\}$ ($1 \leq i \leq s-2$), and

$$\sum_{i=1}^{s-2} \delta_i = \underbrace{x_{s-1} - x_1}_{\text{telescopic}}.$$

Let $\delta'_i = \delta_i - 1 \in \{0, 1\}$. Then the number of recovery sets is the number of solutions to $\sum_{i=1}^{s-2} \delta'_i = x_{s-1} - x_1 - s + 2$. For the four endpoint choices $(x_1, x_{s-1}) \in \{0, 1\} \times \{n-3, n-2\}$ the right-hand side equals, respectively,

$$n-s, \quad n-s-1, \quad n-s-1, \quad n-s-2.$$

Since $\delta' \in \{0, 1\}$ the four possibilities give $\binom{s-2}{n-s} + 2\binom{s-2}{n-s-1} + \binom{s-2}{n-s-2} = \binom{s}{n-s}$, recovery sets altogether, where the last identity can be proved combinatorially as follows. Consider the number of length- s binary vectors of Hamming weight $n-s$. The right hand side is clearly the solution to this problem. As for the left side, we count the vectors by 4 mutually disjoint groups with the bits in positions 1 and 2 being $(0, 0), (0, 1), (1, 0), (1, 1)$. Combining the two cases, the number of recovery sets is $\alpha^{\circ}(s) = \binom{s-1}{n-s-1} + \binom{s}{n-s}$. Substituting into Equation (V.1) gives the result. This concludes the solution for the case $\ell = 2$.

We now move on to the general case. A size- s recovery set is specified by its s start indices on the circle. Since there is no canonical “first” index, we fix the rotation by anchoring one chosen start index at 0. Denote such recovery set as $\mathcal{R}$, and denote its corresponding set of start positions $\mathcal{I}_{\mathcal{R}} = \{x_1 = 0 < x_2 < \dots < x_s < n\} \subseteq \{0, 1, \dots, n-1\}$. Focusing on only a subset of recovery sets, all those which contain 0 as one of the starting positions. Define the *cyclic gaps* $\delta_1 = x_2 - x_1$, $\delta_2 = x_3 - x_2$, $\dots$, $\delta_{s-1} = x_s - x_{s-1}$, $\delta_s = n - x_s$. As in Problem 1, the number of recovery sets is the number of solutions to $\sum_{i=1}^s \delta_i = \underbrace{(x_2 - x_1) + \dots + (x_s - x_{s-1}) + (n - x_s)}_{\text{telescopic}} = n$,

with $1 \leq \delta_i \leq \ell$. Denote $\delta'_i = \delta_i - 1 \in \{0, 1, \dots, \ell-1\}$. Then $\sum_{i=1}^s \delta'_i = n - s$. Let $\beta^\circ(s)$ be the number of solutions, which is the number of cyclic recovery sets that contain 0. Shortly, like in Problem 1, the solution is derived from the generating function

$$\beta^\circ(s) = [x^{n-s}] (1 + x + \dots + x^{\ell-1})^s = \sum_{j=0}^{\lfloor \frac{n-s}{\ell} \rfloor} (-1)^j \binom{s}{j} \binom{n-\ell j-1}{s-1}.$$

To compute the number of recovery sets, a method similar to the one in the proof of Burnside’s Lemma for finite groups ([21]) is being used. Consider the set of pairs $\mathcal{P}_s = \{(\mathcal{R}, i) : \mathcal{R} \text{ is a size-}s \text{ cyclic recovery set, } i \in \mathcal{I}_{\mathcal{R}}\}$. The size of $\mathcal{P}_s$ can be counted in two different ways. It can be counted looking from the perspective of the sets. Each R contributes exactly s pairs to $\mathcal{P}_s$, so $|\mathcal{P}_s| = s \cdot \alpha^\circ(s)$. It can be also counted, looking from the perspective of the indices. Since the setup is symmetrical with respect to cyclic rotation, for any $i \in [n]$ the number of pairs it appears in is the same. Counting the number of times 0 appears gives us exactly $\beta^\circ(s)$. Hence, $|\mathcal{P}_s| = n \cdot \beta^\circ(s)$. Therefore, we get $\alpha^\circ(s) = \frac{n}{s} \beta^\circ(s) = \frac{n}{s} \sum_{j=0}^{\lfloor \frac{n-s}{\ell} \rfloor} (-1)^j \binom{s}{j} \binom{n-\ell j-1}{s-1}$. Substituting into Equation V.1, the final result is obtained. ■

In Section VI, for $\ell = o(n)$, the asymptotic behavior of $\mathbb{E}[T_{n,\ell}^\circ] \approx \frac{n}{\ell} \log(\frac{n}{\ell})$ is discussed, showing its connection to Problem 3. In Section VII a monotonicity property is presented, regarding a family of sub models that Problem 2 induces and their multi-dimensional extensions.

VI. DERIVING RESULTS AND BOUNDS FROM PROBLEM 3, THE CONTINUOUS CIRCLE

Problems 2 and 3 share a common structure. In both problems, arcs on a circle are being sampled until full coverage of the circumference. The two models can be compared and it is now shown that the discrete model has an expected number of samples that is at most the expected number in the continuous model.

Lemma 2. *For all $1 \leq \ell \leq n$, it holds that*

$$\mathbb{E}[T_{n,\ell}^\circ] \leq \mathbb{E}[\mathcal{T}_{\frac{n}{\ell}}].$$

Proof: The idea is that $T_{n,\ell}$ ’s distribution can be created via $\mathcal{T}_{\frac{n}{\ell}}$ ’s distribution, and it can be shown that for every sampled set of arcs that cover the circle, the

corresponding created windows, cover the length- n circle, thus showing a stochastic dominance of $\mathcal{T}_{\frac{n}{\ell}}$ over $T_{n,\ell}$.

Let $C_i \sim \text{Unif}[0, 1]$, and define $D_i = \lfloor nC_i \rfloor \in \{0, 1, \dots, n-1\}$. We interpret D_i as the starting index of a cyclic window of length ℓ , i.e.,

$$\text{win}_\ell(D_i) = \{D_i, D_i + 1, \dots, D_i + \ell - 1\} \pmod{n}.$$

Thus $\{D_i\}$ is uniformly distributed over all n possible window starts. Suppose the continuous arcs $\{[C_j, C_j + a) \pmod{1}\}$, with $a = \ell/n$ already, cover the entire circle. Then for every index $i \in \{0, 1, \dots, n-1\}$, the point $x_i = i/n$ must lie inside some arc $[C_r, C_r + a)$. Writing $D_r = \lfloor nC_r \rfloor$, this immediately implies $i \in \text{win}_\ell(D_r)$. Hence full coverage in the continuous model implies full coverage in the cyclic-discrete model. ■

This upper bound helps concluding the asymptotic behavior of Problem 2.

Corollary 1. *For $\ell = o(n)$, it holds that*

$$\mathbb{E}[T_{n,\ell}^\circ] \approx \frac{n}{\ell} \log\left(\frac{n}{\ell}\right).$$

Proof: As mentioned in Section II, it was shown that $\mathbb{E}[\mathcal{T}_a] \approx \frac{1}{a} \log\left(\frac{1}{a}\right)$, as $a \rightarrow 0$. According to this and Lemma 2, choosing $a = \frac{\ell}{n}$, concludes that for $\ell = o(n)$, $\mathbb{E}[T_{n,\ell}^\circ] \lesssim \frac{n}{\ell} \log\left(\frac{n}{\ell}\right)$. Let $m = \lfloor \frac{n}{\ell} \rfloor$ and take representatives $S = \{0, \ell, 2\ell, \dots, (m-1)\ell\}$. Any length- ℓ cyclic window covers only one element of S , and for each $i \in S$ the hit probability per draw is $\frac{\ell}{n}$. Thus the problem of covering S reduces to the coupon collector’s problem ([7]–[11]) with $\lfloor \frac{n}{\ell} \rfloor$ coupons. Therefore, $\mathbb{E}[T_{n,\ell}^\circ] \geq \lfloor \frac{n}{\ell} \rfloor H_{\lfloor \frac{n}{\ell} \rfloor} \gtrsim \frac{n}{\ell} \log\left(\frac{n}{\ell}\right)$. ■

Corollary 1 shows the intuitive observation that as n grows and $\ell = o(n)$, the discrete cyclic model shares the same leading asymptotic as the continuous circle covering problem. Now similarly to Section IV, the general asymptotic regime where $\ell = \Omega(n)$ is examined, completing the asymptotic analysis of $\mathbb{E}[T_{n,\ell}^\circ]$ for any $\ell \leq n$.

Corollary 2. *Fix a constant $c \geq 2$. For $\frac{n}{c} \leq \ell < \frac{n}{c-1}$, it holds that*

$$\mathbb{E}[T_{n,\ell}^\circ] = O(1).$$

Furthermore, it holds that

$$\mathbb{E}[T_{n,\ell}^\circ] \leq \mathbb{E}[\mathcal{T}_{\ell/n}] \leq \mathbb{E}[\mathcal{T}_{1/c}] \leq 1 + c^2.$$

Proof: By Lemma 2 (Section VI), $\mathbb{E}[T_{n,\ell}^\circ] \leq \mathbb{E}[\mathcal{T}_a]$ with $a = \frac{\ell}{n}$. If $\ell \geq \frac{n}{c}$ then $a \geq \frac{1}{c}$, and since covering gets easier as a increases, $\mathbb{E}[\mathcal{T}_a] \leq \mathbb{E}[\mathcal{T}_{1/c}]$. Moreover, for t samples of arcs, full coverage not occurs if and only if for every arc, the next one is at distance more than a from it (which happens with probability $1 - a$). By a union bound, the probability that this happens for all t samples is at most $t(1 - a)^{t-1}$. Hence,

$$\begin{aligned} \mathbb{E}[T_a] &= \sum_{t=0}^{\infty} \Pr[T_a > t] = 1 + \sum_{t=1}^{\infty} \Pr[T_a > t] \\ &\leq 1 + \sum_{t=1}^{\infty} t(1 - a)^{t-1} = 1 + \frac{1}{a^2} \leq 1 + c^2, \end{aligned}$$

showing that the bound depends purely on the ratio between n and ℓ . ■

Going back to Problem 1, in order to analyze the asymptotic behavior of the expectation in this problem, we bound it from both sides, which yields the next result.

Theorem 7. *For $\ell = \omega(\log(n))$, it holds that*

$$\mathbb{E}[T_{n,\ell}] = \frac{3}{2}(n - \ell + 1)(1 + o(1)),$$

and for $\ell = o(\log n)$, it holds that

$$\mathbb{E}[T_{n,\ell}] = \frac{n}{\ell} \log\left(\frac{n}{\ell}\right)(1 + o(1)).$$

Proof: The coverage expectation is at most the sum of expectations of covering the extremes $[0, \ell - 1] \cup [n - \ell, n - 1]$, and the complimentary middle regime $[\ell, n - \ell - 1]$ separately. Roughly speaking, $\mathbb{E}[T_{n,\ell}] \leq \mathbb{E}[\text{cover extremes}] + \mathbb{E}[\text{cover middle}]$. As shown in Theorem 3, $\mathbb{E}[\text{cover extremes}] = \frac{3}{2}(n - \ell + 1)$. Covering only the middle is no harder than covering a full cyclic instance with $n - \ell + 1$ positions and window length ℓ , with a probability for each position to be covered that is identical in both cases, being $\frac{\ell}{n - \ell + 1}$. Therefore, $\mathbb{E}[\text{cover middle}] \leq \mathbb{E}[T_{n - \ell + 1, \ell}^\circ] \leq \mathbb{E}[\mathcal{T}_a]$, with $a = \frac{\ell}{n - \ell + 1}$, where the last inequality follows from Lemma 2. In addition, as mentioned in Section II, from [23], [24], $\mathbb{E}[\mathcal{T}_{\frac{\ell}{n - \ell + 1}}] \approx \frac{n - \ell + 1}{\ell} \log\left(\frac{n - \ell + 1}{\ell}\right) \approx \frac{n}{\ell} \log\left(\frac{n}{\ell}\right)$. Hence, the coverage time satisfies $\mathbb{E}[T_{n,\ell}] \lesssim \frac{3}{2}(n - \ell + 1) + \frac{n}{\ell} \log\left(\frac{n}{\ell}\right)$.

Additionally, the coverage time is at least as big as the coverage time to cover the extremes $\frac{3}{2}(n - \ell + 1)$, and it is at least as big as the time to cover the representative points set $S = \{\ell - 1, 2\ell - 1, 3\ell - 1, \dots, \lfloor \frac{n - \ell + 1}{\ell} \rfloor \ell - 1\}$. Each window covers only one representative point from S , with probability $\frac{\ell}{n - \ell + 1}$, therefore, The problem reduces to the coupon collector's problem ([7]–[11]) with $\lfloor \frac{n - \ell + 1}{\ell} \rfloor$ coupons. Hence, $\mathbb{E}[T_{n,\ell}] \geq \lfloor \frac{n - \ell + 1}{\ell} \rfloor H_{\lfloor \frac{n - \ell + 1}{\ell} \rfloor}$. This gives the explicit bounds, $\max\left\{\frac{3}{2}(n - \ell + 1), \lfloor \frac{n - \ell + 1}{\ell} \rfloor H_{\lfloor \frac{n - \ell + 1}{\ell} \rfloor}\right\} \leq \mathbb{E}[T_{n,\ell}]$. Therefore, by a simple analysis we get that for $\ell = \omega(\log(n))$ the linear term dominates, and for $\ell = o(\log(n))$ the coupon collector term dominates, which makes the expectation asymptotically the same as in the cyclic case. ■

Observation 1. *From Theorem 7, we observe that $\log n$ acts as a threshold on the expectation in the following way. Below it, the task is essentially “find all middle positions”, mirroring the cyclic behavior. Above it, the problem becomes essentially retrieving the two extremes.*

VII. THE BATCH SAMPLING MODEL AND COVERAGE MONOTONICITY

Problems 1, 2 and 4 ask for the time to cover n items. Each model can be represented as hypergraph on $[n]$ as mentioned and described in detail in Section III, with the collectable subsets of $[n]$ being the hyperedges of the corresponding Problem's hypergraph. In Problems 2 and 4, each item has the same degree (uniformity), each hyperedge has size ℓ (regularity) and the set of hyperedges is non-empty. We call any hypergraph on $[n]$ that satisfies those

three conditions a *uniform ℓ -regular hypergraph on $[n]$* , or in short, a *uniform ℓ -regular model*. Denote the class of all uniform ℓ -regular hypergraphs on $[n]$ as $\mathcal{U}_{n,\ell}$. For example, $\mathcal{U}_{n,1}$ consists of only one hypergraph denoted by $\mathcal{H}_{n,1}$ which is the hypergraph corresponding to the classical coupon collector's problem, having $\mathbb{E}[T_{\mathcal{H}_{n,1}}] = nH_n$. Note that for any $\mathcal{H} \in \mathcal{U}_{n,\ell}$, the set of hyperedges is a subset of all possible hyperedges admissible in Problem 4 denoted as $\binom{[n]}{\ell}$. Denote the hypergraph of Problem 4 as $\mathcal{H}_{\binom{[n]}{\ell}}$. Note that $T_{\mathcal{H}_{\binom{[n]}{\ell}}}$ is the same as the random variable in Problem 4, $T_{\binom{[n]}{\ell}}$, and will be referred to as such from now on.

In this section a natural question is suggested. Given $\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{U}_{n,\ell}$, with the corresponding sets of hyperedges $\mathcal{F}_1, \mathcal{F}_2$, does it hold that if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, or even if $|\mathcal{F}_1| \leq |\mathcal{F}_2|$, then $\mathbb{E}[T_{\mathcal{H}_1}] \leq \mathbb{E}[T_{\mathcal{H}_2}]$? If so, $\mathcal{H}_{\binom{[n]}{\ell}}$ with $T_{\binom{[n]}{\ell}}$ would have the highest expectation among $\mathcal{U}_{n,\ell}$, and any model with $\frac{n}{\ell}$ disjoint subsets would admit the lowest expectation. That is, for any $\mathcal{H} \in \mathcal{U}_{n,\ell}$, $\lfloor \frac{n}{\ell} \rfloor H_{\lfloor \frac{n}{\ell} \rfloor} \leq \mathbb{E}[T_{\mathcal{H}}] \leq \mathbb{E}[T_{\binom{[n]}{\ell}}]$, providing universal upper and lower bounds on the expectation for the entire class.

As mentioned in Section II, combinatorial solutions for Problem 4 and the generalization, given $A \subseteq [n]$ with $|A| = m$, and $Z_k(A)$ being the number of draws needed to see at least $k \leq m$ distinct elements of A , exact formulas were obtained. While exact, those expressions are numerically delicate for large parameters due to large binomial coefficients and cancellations, and is hard to compute in a regular calculator. The next results gives tight lower and upper bounds on the expectation of this problem, and furthermore for any generalization, as explained below.

While finalizing this paper, it was noticed that a very recent and concurrent work ([28]) independently presented a similar proof to the one presented here. We wish to acknowledge their contribution and give due credit. Nevertheless, the proof is included here, as we believe it serves a broader purpose, demonstrating that the same reasoning naturally extends to generalized versions of the coupon collector's problem, including problems such as the double dixie cup and other variants. We also think and hope that the intuition in this proof can be applied in other coverage problems.

Theorem 8. *For any $1 \leq \ell < n$ and $1 \leq k \leq m \leq n$ and $A \subseteq [n]$ with $|A| = m$, it holds that*

$$\frac{H_m - H_{m-k}}{H_n - H_{n-\ell}} \leq \mathbb{E}[Z_k(A)] \leq \frac{H_m - H_{m-k} - \frac{1}{n}}{H_n - H_{n-\ell}} + 1.$$

Proof: Fix $A \subseteq [n]$ with $|A| = m \leq n$ and $k \leq m$. Let Y be the stopping time of the classical coupon collector's problem with single coupon per draw, when k items of A were obtained. Since each new coupon from A is obtained with probability $\frac{m-j}{n}$, after j distinct ones have already been seen, it holds that

$$\mathbb{E}[Y] = \sum_{j=0}^{k-1} \frac{n}{m-j} = n(H_m - H_{m-k}).$$

Let $X := Z_k(A)$ be the number of ℓ -subset draws until the same coverage. From a single-coupon run $i_1, i_2, \dots, i_Y$, form consecutive *blocks* by cutting as soon as the block contains ℓ distinct coupon labels (within the block). Then, it is possible to partition $i_1, i_2, \dots, i_Y$ as follows.

$$\begin{aligned} i_1, \dots, i_Y = & \underbrace{i_1, \dots, i_{\ell+Z_1}}_{\ell+Z_1}, \underbrace{i_{\ell+Z_1+1}, \dots, i_{2\ell+Z_1+Z_2}}_{\ell+Z_2}, \dots, \\ & \underbrace{i_{(X-2)\ell+\sum_{j=1}^{X-2} Z_j+1}, \dots, i_{(X-1)\ell+\sum_{j=1}^{X-1} Z_j}}_{\ell+Z_{X-1}}, \\ & \underbrace{i_{(X-1)\ell+\sum_{j=1}^{X-1} Z_j+1}, \dots, i_Y}_{\text{partial block}}. \end{aligned}$$

The j -th block then has length $L_j = \ell + Z_j$, where $Z_j \geq 0$ (wasted repeats inside the block). The expected number of draws until ℓ distinct items were drawn is

$$\mathbb{E}[L_j] = \sum_{k=1}^{\ell} \frac{n}{n-k+1} = n(H_n - H_{n-\ell}).$$

Since $Z_j = L_j - \ell$, we get

$$\mathbb{E}[Z_j] = \mathbb{E}[L_j] - \ell = n(H_n - H_{n-\ell}) - \ell.$$

To derive the bounds, consider the following. After $X - 1$ blocks, the coverage process was almost finished, and the last draws until full coverage does not necessarily make a full block. Let $D \geq 1$ be the extra draws needed to reach full coverage after $X - 1$ blocks were drawn (by definition, $X - 1$ draws had not yet covered all items and at most one block is required to complete coverage). Then

$$Y = \sum_{i=1}^{X-1} (\ell + Z_i) + D.$$

Taking expectations gives

$$n(H_m - H_{m-k}) - \mathbb{E}[D] = \ell \mathbb{E}[X - 1] + \mathbb{E}[X - 1] \mathbb{E}[Z],$$

which yields

$$\mathbb{E}[X] = 1 + \frac{H_m - H_{m-k} - \frac{\mathbb{E}[D]}{n}}{H_n - H_{n-\ell}}. \quad (\text{VII.1})$$

By the definition of D , and since in the last D draws the number of different collected items is at most ℓ , it holds that $1 \leq \mathbb{E}[D] \leq \mathbb{E}[L_j]$. Therefore, by substituting into Equation VII.1 we get the final bounds

$$\frac{H_m - H_{m-k}}{H_n - H_{n-\ell}} \leq \mathbb{E}[X] \leq 1 + \frac{H_m - H_{m-k} - \frac{1}{n}}{H_n - H_{n-\ell}}. \quad \blacksquare$$

Using Theorem 8, tight lower and upper bounds for Problem 4 are obtained. The special case of $A = [n]$, and $k = n$, yields that for any $1 \leq \ell \leq n$,

$$\frac{H_n}{H_n - H_{n-\ell}} \leq \mathbb{E}[T_{\binom{[n]}{\ell}}] \leq \frac{H_{n-1}}{H_n - H_{n-\ell}} + 1,$$

which coincides with the result of [28].

Next, more powerful statement is presented, explaining how Theorem 8 can be extended. We define the term

coverage task as getting a collection that satisfies some requirement. The coupons collector's problem has a coverage task that requires to obtain each coupon at least once. Another coverage task can be to obtain any coupon twice, as in the double dixie cups problem ([29]), or maybe to either get each one twice, or to get half of the coupons three time. Let $\mathcal{C}$ be a coverage task. We can consider both $\mathcal{H}_{n,1}$ and $\mathcal{H}_{\binom{[n]}{\ell}}$ to be the hypergraphs from which hyperedges are being sampled, and we can ask about the relation of expected number of steps to complete $\mathcal{C}$ in both models. Denote $Q_{n,1}^{\mathcal{C}}$ as the number of hyperedges drawn from $\mathcal{H}_{n,1}$ until completing $\mathcal{C}$, and denote $Q_{n,\ell}^{\mathcal{C}}$ as the number of draws until completing $\mathcal{C}$ drawing from $\mathcal{H}_{\binom{[n]}{\ell}}$.

Theorem 9. *Let $\mathcal{C}$ be any coverage task. It holds that*

$$\frac{\mathbb{E}[Q_{n,1}^{\mathcal{C}}]}{n(H_n - H_{n-\ell})} \leq \mathbb{E}[Q_{n,\ell}^{\mathcal{C}}] \leq \frac{\mathbb{E}[Q_{n,1}^{\mathcal{C}}] - \frac{1}{n}}{n(H_n - H_{n-\ell})} + 1,$$

Proof: The claim follows directly from the argument in the proof of Theorem 8, substituting $X = Q_{n,\ell}^{\mathcal{C}}$ and $Y = Q_{n,1}^{\mathcal{C}}$. $\blacksquare$

This shows that any coverage task inherits the same normalization factor that governs the transition from using the single-coupon draws of $\mathcal{H}_{n,1}$ to the ℓ -subset draws of $\mathcal{H}_{\binom{[n]}{\ell}}$.

One general framework that encompasses a broad family of coupon collector's problem coverage tasks can be formulated as follows. Let

$$\mathbf{v} = (v_1, v_2, \dots, v_n) \in \mathbb{N}^n$$

be a demand vector, where v_i specifies the required number of times item i must be collected. We let $Q_{n,1}^{\mathbf{v}}$ denote the number of draws $\mathcal{H}_{n,1}$'s hyperedges required for obtaining each $i \in [n]$ at least v_i times, and $Q_{n,\ell}^{\mathbf{v}}$ as number the of $\mathcal{H}_{\binom{[n]}{\ell}}$'s hyperedges needed to complete the same task.

The double dixie cup problem mentioned above is a classical special case, obtained by setting a constant m , and taking $\mathbf{v} = (m, m, \dots, m)$. In this setting ([29]) it is known that

$$\mathbb{E}[Q_{n,1}^{\mathbf{v}}] = n \left(\log n + (m-1) \log \log n + \gamma + o(1) \right).$$

By Theorem 9, it is obtained that

$$\mathbb{E}[Q_{n,\ell}^{\mathbf{v}}] \geq \frac{n \left(\log n + (m-1) \log \log n + \gamma + o(1) \right)}{n(H_n - H_{n-\ell})},$$

and

$$\mathbb{E}[Q_{n,\ell}^{\mathbf{v}}] \leq \frac{n \left(\log n + (m-1) \log \log n + \gamma + o(1) \right) - \frac{1}{n}}{n(H_n - H_{n-\ell})} + 1.$$

Note that the gap between the lower and upper bounds is at most 1, making the bounds tight.

Addressing the monotonicity question, Problem 2 is the parent of a class of problems we call Δ - d window start set. For integers n, ℓ, d with $\ell \mid n$ and $d \mid \ell$, denote $\mathcal{U}_{n,\ell}^{\circ} \subseteq \mathcal{U}_{n,\ell}$ to be the class of hypergraphs with hyperedges that are cyclic windows. Denote the hypergraph of Problem 2 as $\mathcal{H}_{n,\ell}^{\circ}$, and its set of hypersdges $\mathcal{F}_{n,\ell}^{\circ}$. Note that for each

$\mathcal{H}_d \in \mathcal{U}_{n,\ell}^\circ$ with hypersdges set $\mathcal{F}_d \subseteq \mathcal{F}_{n,\ell}^\circ$ which is referred to as Δ - d start set, and the windows start positions are at distance d apart. Note that $\mathcal{H}_d \in \mathcal{U}_{n,\ell}$ is equivalent to $\mathcal{H}_{\frac{n}{d}, \frac{\ell}{d}}^\circ \in \mathcal{U}_{\frac{n}{d}, \frac{\ell}{d}}^\circ$, which is the hypergraph of Problem 2 setting $n' = n/d$ and $\ell' = \ell/d$ (since there are $\frac{n}{d}$ d -lengthed blocks, and on each block start, a window starts, containing $\frac{\ell}{d}$ blocks). As a consequence, proving the monotonicity for the Δ - d window class $\mathcal{U}_{n,\ell}^\circ$ is equivalent to proving the next lemma.¹

Lemma 3. *For any n, ℓ, d with $d \mid \ell \mid n$, it holds that*

$$\mathbb{E}\left[T_{\frac{n}{d}, \frac{\ell}{d}}^\circ\right] \leq \mathbb{E}\left[T_{n, \ell}^\circ\right].$$

Equivalently, within the Δ - d family with fixed n and ℓ , making starts more discrete decreases the expected coverage time.

Proof: The proof is a coupling argument, very similar to the proof of Lemma 2. Fix $d \mid \ell \mid n$ and run the cyclic (n, ℓ) model with starts uniformly on $\{0, 1, \dots, n-1\}$ and windows $\text{win}_\ell^\circ(i)$. Define

$$\phi(i) = d \left\lfloor \frac{i}{d} \right\rfloor \in \{0, d, 2d, \dots, n-d\},$$

which is uniform over the Δ - d start set (each value has d preimages), so using ϕ gives a uniform Δ - d process (which justifies the coupling). In other words, the mapping shifts windows to start at multiples of d (block starts). Let $\mathcal{R} = \{w_1, \dots, w_{|\mathcal{R}|}\}$ be the windows drawn up to $T_{n,\ell}^\circ$ (they cover all n positions). Consider any d -block $[jd, (j+1)d-1]$. If it was covered by a window starting at jd , it remains covered after applying ϕ . If not, some window in $\mathcal{R}$ starts inside this block. So applying ϕ moves the window start to position jd , covering the whole block. Thus all blocks are covered by the ϕ -mapped sequence. Therefore, coverage in the cyclic windows model implies coverage in the Δ - d model. ■

Remark 1. *The same coupling argument extends naturally to higher dimensions. For example, consider a 2-dimensional torus of size $n_1 \times n_2$, where each sample is a rectangular patch of $\ell_1 \times \ell_2$. If $d_1 \mid \ell_1 \mid n_1$ and $d_2 \mid \ell_2 \mid n_2$, we can partition the torus into $d_1 \times d_2$ blocks and define the mapping*

$$\phi(i, j) = (d_1 \lfloor i/d_1 \rfloor, d_2 \lfloor j/d_2 \rfloor),$$

which is uniform over the Δ -(d_1, d_2) start set. The same block-covering argument shows that coverage in the full model implies coverage in the Δ - d model, yielding the same monotonicity property.

This monotonicity extends inductively to any dimension. Moreover, there is always a corresponding continuous version obtained by letting $n_i \rightarrow \infty$ while fixing the ratios ℓ_i/n_i , which dominates all discrete versions in expectation (by the same argument proving Lemma 2).

Remark 2. *For Problem 2, a lower bound of $\lfloor \frac{n}{\ell} \rfloor H_{\lfloor \frac{n}{\ell} \rfloor}$ had been derived in the proof of Lemma 1. In the special case*

¹one may ask why wasn't the expression in Theorem 6 used in the proof. It is due to the complex nature of the inclusion-exclusion terms that makes it hard to compare.

where $\ell \mid n$ this matches with a general lower bound that is true for all models, established in [26] for $n = 2$, and as stated in [28], can be extended to any $\ell \mid n$.

These monotonicity results suggest three universal principles which we believe are true.

Conjecture 1. *Given $\mathcal{H}_1, \mathcal{H}_2 \in \mathcal{U}_{n,\ell}$, with the corresponding sets of hyperedges $\mathcal{F}_1, \mathcal{F}_2$ such that $\mathcal{F}_1 \subseteq \mathcal{F}_2$, it holds that*

$$\mathbb{E}[T_{\mathcal{H}_1}] \leq \mathbb{E}[T_{\mathcal{H}_2}].$$

Conjecture 2. *Given $\mathcal{H} \in \mathcal{U}_{n,\ell}$, it holds that*

$$\mathbb{E}[T_{\mathcal{H}}] \leq \mathbb{E}[T_{\binom{[n]}{\ell}}].$$

Conjecture 3. *Given $\mathcal{H} \in \mathcal{U}_{n,\ell}$, it holds that*

$$\lfloor \frac{n}{\ell} \rfloor H_{\lfloor \frac{n}{\ell} \rfloor} \leq \mathbb{E}[T_{\mathcal{H}}].$$

The conjectures suggest a universal upper bound on the uniform ℓ -regular models. Furthermore, a conjecture by [27], later proved by [28], stated that the uniform distribution, corresponding to the batch coverage model in our terminology, does not achieve the minimal expected time. Our conjecture broadens the horizon of this idea and claims that not only is this model not optimal, but may even be the worst within the entire class $\mathcal{U}_{n,\ell}$.

Although the conjectures, and in particular Conjecture 2 remains currently unproved, a slightly weaker bound can already be established. The following theorem states this bound explicitly.

Theorem 10. *Given $\mathcal{H} \in \mathcal{U}_{n,\ell}$, it holds that*

$$\mathbb{E}[T_{\mathcal{H}}] \leq \frac{n}{\ell} (\log n + 1).$$

Proof: Let U_t be the number of items not yet seen after t draws. Due to uniformity, the probability for a position to not be sampled after t rounds is $(1 - \frac{\ell}{n})^t$. Thus, $\mathbb{E}[U_t] = n(1 - \ell/n)^t$. Using the tail-sum formula and Markov's inequality, $\mathbb{E}[T_{\mathcal{H}}] = \sum_{r=1}^{\infty} \Pr[T_{\mathcal{H}} \geq r] = \sum_{r=1}^{\infty} \Pr[U_{r-1} \geq 1] \leq \sum_{r=1}^{t_0} 1 + \sum_{r=t_0+1}^{\infty} \mathbb{E}[U_{r-1}]$, where we choose $t_0 := \left\lceil \frac{n \log n}{\ell} \right\rceil$. Then $\sum_{r=t_0+1}^{\infty} \mathbb{E}[U_{r-1}] = n \sum_{r=t_0}^{\infty} (1 - \ell/n)^r = n \cdot \frac{(1 - \ell/n)^{t_0}}{1 - (1 - \ell/n)} = \frac{n^2}{\ell} (1 - \ell/n)^{t_0}$. Using $(1 - \ell/n)^{t_0} \leq e^{-(\ell/n)t_0} \leq e^{-(\ell/n) \cdot (n \log n / \ell)} = e^{-\log n} = 1/n$, we get

$$\sum_{r=t_0+1}^{\infty} \mathbb{E}[U_{r-1}] \leq \frac{n^2}{\ell} \cdot \frac{1}{n} = \frac{n}{\ell}.$$

Therefore,

$$\mathbb{E}[T_{\mathcal{H}}] \leq t_0 + \frac{n}{\ell} \leq \frac{n}{\ell} \log n + \frac{n}{\ell} = \frac{n}{\ell} (\log n + 1). \quad \blacksquare$$

VIII. FUTURE WORK

A natural direction for future research is obtaining a more accurate asymptotic expression for the expectation in Problem 1. Unlike the cyclic or continuous versions, Problem 1 is harder to analyze, resulting in looser asymptotics which remain relatively crude.

Another important direction concerns the precision gap between the asymptotics of different models. For structured models such as the cyclic window (Problem 2), and its higher dimensional generalizations in $\mathcal{U}_{n,\ell}$, a lower bound of $\lfloor \frac{n}{\ell} \rfloor H_{\lfloor \frac{n}{\ell} \rfloor}$ was established as mentioned in Remark 2, and is believed to be universal for $\mathcal{U}_{n,\ell}$ as stated in Conjecture 3. A universal upper bound of $\frac{n}{\ell}(\log n + 1)$ was established for $\mathcal{U}_{n,\ell}$.

The two bounds share the same leading term $\frac{n}{\ell} \log n$ when $\ell = o(n)$. On one hand, this is encouraging since it means that all such models are asymptotically solved. On the other hand, it highlights an important ambiguity: describing the expectation of a coverage problem as “asymptotically $\frac{n}{\ell} \log n$ ” is not very informative. Different models may yield the same asymptotic form, yet differ substantially in their non-leading terms.

Although the lower and upper bounds behave asymptotically the same with respect to the dominant term (in particular, their ratio approaches one as n increases), there is an $O(\frac{n}{\ell} \log \ell)$ gap in their difference. Closing this gap or analyzing the expectation for each model separately is left for future work. This suggests that further study of sub-leading terms and structural effects between models is essential to understand the fine behavior inside the regime where $\frac{n}{\ell} \log(\frac{n}{\ell}) \approx \frac{n}{\ell} \log n$.

Finally, proving or disproving either Conjecture 1 or Conjecture 2 could reveal deeper structural insights into coverage problems with uniform ℓ -regular hypergraphs. Even a single counterexample or partial proof could illuminate the principles governing how geometry, overlap, and locality influence coverage time.

REFERENCES

- [1] F. Sanger and A. R. Coulson, “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase,” *Journal of Molecular Biology*, 94(3), 441–448, 1975.
- [2] E. S. Lander et al, “Initial sequencing and analysis of the human genome,” *Nature*, 409, 860–921, 2001.
- [3] J. C. Venter et al, “The sequence of the human genome,” *Science*, 291, 1304–1351, 2001.
- [4] J. Bornholt, R. Lopez, D. Carmean, L. Ceze, G. Seelig, and K. Strauss, “A DNA-based archival storage System,” *ASPLOS*, 637–649, 2016.
- [5] A. Narayan Ravi, A. Vahid, and I. Shomorony, “Coded shotgun sequencing,” *IEEE Transactions on Information Theory*, vol. 68, no. 7, pp. 4241–4253, 2022.
- [6] J. E. Kobza, S. H. Jacobson, and D. E. Vaughan, “A survey of the Coupon Collector’s Problem with random sample sizes,” *Springer*, 2007.
- [7] A. de Moivre, “De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus,” *Philosophical Transactions of the Royal Society of London A*, 27, 213–264, 1711.
- [8] P. S. Laplace, “Théorie analytique des probabilités,” Paris, 1812.
- [9] S. Bernstein, “Teoria veroiatnostei,” Gostechizdat, Moscow, 1945.
- [10] W. Feller, “An introduction to probability theory and its applications, Vol. 1,” Wiley, 1950.
- [11] P. Erdős and A. Rényi, “On a classical problem of probability theory,” *Magyar Tud. Akad. Mat. Kut. Int. Közl.*, 6, 215–220, 1961.
- [12] G. Pólya, “Eine wahrscheinlichkeitsaufgabe zur kundenwerbung,” *Zeitschrift für angewandte Mathematik und Mechanik*, 10, 96–97, 1930.
- [13] W. Stadje, “The collector’s problem with group drawings,” *Advances in Applied Probability*, 21(4), 866–882, 1990.
- [14] I. Adler and S. M. Ross, “The coupon subset collection problem,” *Journal of Applied Probability*, 37(3), 737–746, 2001.
- [15] P. Flajolet, X. Gourdon, and P. Dumas, “Mellin transforms and asymptotics: harmonic sums,” *Theoretical Computer Science*, 144(1-2), 3–58, 1995.
- [16] D. Bar-Lev, O. Sabary, R. Gabrys and E. Yaakobi, “Cover your bases: how to minimize the sequencing coverage in DNA storage systems,” *ISIT*, 2023.
- [17] A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, “A combinatorial perspective on random access efficiency for DNA storage,” *arXiv*, 2024.
- [18] H. Abraham, R. Gabrys, and E. Yaakobi, “Covering all bases: the next inning in DNA sequencing efficiency,” *arXiv:2401.17649*, 2024.
- [19] I. Preuss, B. Galili, Z. Yakhini, and L. Anavy, “Sequencing coverage analysis for combinatorial DNA-based storage systems,” *IEEE Trans. Mol., Biol., Multi-Scale Commun.*, 10(2), 297–316, 2024.
- [20] T. Cohen and E. Yaakobi, “Optimizing the decoding probability and coverage ratio of composite DNA,” *ISIT*, pp. 1949–1954, 2024.
- [21] W. Burnside, “Theory of groups of finite order,” *Cambridge University Press*, 1897.
- [22] W. L. Stevens, “Solution to a geometrical problem in probability,” *Annals of Eugenics*, 9, 315–320, 1939.
- [23] L. Flatto and A. G. Konheim, “The random division of an interval and the random covering of a circle,” *SIAM Review*, 4, 211, 1962.
- [24] F. W. Steutel, “Random division of an interval,” *Statistica Neerlandica*, vol. 21, pp. 231–244, 1967.
- [25] H. Solomon, “Geometric probability,” *SIAM*, Philadelphia, 1978.
- [26] C. Fang and E. Chang, “Optimal strategy of coupon subset collection when each package contains half of the coupons,” *Information Processing Letters*, vol. 114, no. 12, pp. 703–705, 2014.
- [27] J. Schilling, “Results and conjectures on the role of the uniform distribution in the coupon collector’s problem with group drawings,” *Information Processing Letters*, vol. 173, pp. 106–112, 2021.
- [28] D. Berend and T. Sher, “On a conjecture of schilling regarding the Coupon Collector’s Problem,” *arXiv preprint arXiv:2509.17201*, 2025.
- [29] D. J. Newman and L. Shepp, “The double Dixie cup problem,” *Amer. Math. Monthly*, 67, 58–61, 1960.