Superstrings of Uniform-Cardinality Set Systems via the Generalized Traveling Salesperson Problem

Alexander Weissenfels, Enrico Iurlano(✉), and Günther R. Raidl

Algorithms and Complexity Group, TU Wien, Austria

{alexander.weissenfels, eiurlano,raidl}@{student, ac}.tuwien.ac.at

Abstract. We present new approaches to the problem of accommodating (as a contiguous substring) every cardinality-k subset of a finite alphabet $\{1, \ldots, n\}$ in a string of shortest length. The key insight is that we can reduce this problem to the Generalized Traveling Salesperson Problem (GTSP). This allows us to leverage the availability and performance of a powerful established metaheuristic solver for the GTSP. We can use it to compute certain new optimal strings, for which we can make observations on their structure. In addition, we derive a slightly tightened lower bound on the minimum length of such strings. It often is responsible for the derivation of optimality guarantees which are computationally unobtainable when employing several tested different exact solving techniques. A by-product is an intriguing class of instances for the GTSP with a priori known (optimally) tight bounds.

Keywords: Shortest Superstrings · Generalized Traveling Salesperson Problem · Reduction · Lower bound

1 Introduction

In this work we address computational approaches to the *shortest P_n^k-cover* problem, a variant of the *Shortest Superstring Problem* (SSP) [1] by Lipski Jr. [13]. Given the collection of all cardinality-k subsets of a cardinality-n alphabet, one seeks here a shortest string over this alphabet such that, for each subset in the collection, an arbitrary arrangement of its elements appears as contiguous substring of the superstring; we call such a string a P_n^k-cover. In this sense, the string `1234515241352` covers all ten 3-subsets of the alphabet $\{1, \ldots, 5\}$; e.g., at the eighth position the occurring substring `241` is responsible for the coverage of $\{1, 2, 4\}$. For the classical SSP no letter-rearrangements are allowed, instead a pre-specified collection of *strings* (rather than sets) over a finite alphabet is given, and a shortest string has to be found which contains all of them as contiguous substrings [1]. The algorithmic study and the computation of short so-called superstrings is motivated by applications in data compression [15] and computational biology [10]. In the latter field, one frequently faces the problem to reconstruct DNA sequences from a collection of fragments (oligonucleotides) which stem from experimentally obtained samples.

This document is a preprint -- version dated January 24, 2025.

The P_n^k-cover problem has strong connections to the one of finding so-called *universal cycle coverings*, in which the strings are cyclic, i.e., the first element of the superstring is the successor of the last one. Recently, in [5], a necessary divisibility-condition on n and k for the existence of so-called *universal cycles* (where each k-subset is covered by one unique occurrence of a substring of length k) has been shown to be an asymptotically sufficient condition, too. This settled a longstanding conjecture of Chung, Diaconis, and Graham [2] formulated in 1989. Apart from selected values for n and k satisfying the particular divisibility condition, for which this conjecture is valid and yields optimal P_n^k-covers, the optima's structure is not fully understood so far. An upper bound of magnitude $\binom{n}{k} + O(n^{\lfloor k/2 \rfloor})$ has been obtained by a combinatorial construction [14]. Further bounds taking into account the magnitude of k as function of n can be found in [3] for the cyclic problem version. If we more generally demand coverage of any set X in a given subset P of the alphabet's power set, determining the existence of such a length-m superstring becomes NP-complete [11]. Computational approaches for the problem (also due to Lipski Jr. [13]) of superstrings of the entire power set of a finite alphabet are presented in [19], where also some selected results on P_n^k-covers are given for up to $n \leq 7$.

Pursuing an approach relying on the maximum overlap shared by suffixes and prefixes of strings as a distance metric, we design particular instances of the Generalized Traveling Salesperson Problem (GTSP) whose feasible solutions, i.e., generalized Hamiltonian cycles, correspond to feasible superstrings for the P_n^k-cover problem. The vertex clusters correspond to the cardinality-k subsets and their elements to all $k!$ strings representing the subsets. For tractable, smaller values $k \leq n \leq 15$ we approach these derived GTSP instances by `GLNS` [18], an established metaheuristic software for the GTSP being a hybrid of Large Neighborhood Search and Simulated Annealing. The pursued approach, furthermore, allows to tackle a more general interpolation of the SSP and the P_n^k-cover problem by banning custom strings from selected clusters. With the use of intense computational resources, we thereby get insights into the size of (near-)optimal solutions and their structure.

For $k = 3$ and $n \in \{9, 12\}$ as well as $(n, k) = (8, 4)$ we are able to identify, to the best of our knowledge, so-far unknown minimum-length solutions; for other (n, k)-values a small optimality gap remains open, which we believe to be reducible in future by tailored (meta)heuristics that further strengthen the primal bounds. For some claims of optimality, we fall back on a—compared to the literature on P_n^k-covers—slightly strengthened purposefully derived dual bound

$$\beta_n^k := k + (n-k) \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k} \left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil .$$

The paper is structured as follows. In Sect. 2 we provide the required notation. The actual reduction to the GTSP is given in Sect. 3 followed by its computational evaluation in Sect. 4. The validity of the dual bound β_n^k is proved in Sect. 5 which is accompanied with concluding remarks in Sect. 6.

2 Notation and preliminary considerations

In the following we consider the alphabet $[n] := \{1, \ldots, n\}$ and call its elements letters. A string of length m over $[n]$ is a sequence $s = (s_1, \ldots, s_m) \in [n]^m$, which we often denote by $s_1 \cdots s_m$ for simplicity; if not stated differently, we refer to strings over $[n]$. We say that $t = t_1 \cdots t_k$ is a length-k infix of s, if there is a starting position $i_1 \in \{1, \ldots, m-k+1\}$ for which $s_i s_{i+1} \cdots s_{i+k-1} = t_1 t_2 \cdots t_k$. Among the $N - k + 1$ length-k infixes the one corresponding to position $i = 1$, respectively to $i = N - k + 1$, constitutes the length-k prefix of s, respectively suffix of s. Denote by P_n^k the set of all k-subsets of $[n]$. A string s is a *superstring for* P_n^k, or for short a P_n^k*-cover*, if for each $A \in P_n^k$ there exists an length-k infix t of s satisfying $\{t_1, t_2, \ldots, t_k\} = A$. A k-subset of $[n]$ is said to be covered λ times by x, if there are precisely λ pairwise distinct indices at which, after suitable reordering, the elements of the k-subset appear as length-k infix. We say that x has an associated universal cycle, compare [2], if the length-$(k-1)$ prefix and length-$(k-1)$ suffix of x are coinciding strings and furthermore, each element in P_n^k is covered precisely once by x. A length-k substring (or infix) of x is called *injective* if it consists of k pairwise different letters.

Given two reals L, U $(L \leq U)$, we use the notation $\mathrm{clmp}_L(x) := \max(x, L)$, $\mathrm{clmp}^U(x) := \min(x, U)$, and $\mathrm{clmp}_L^U(x) := \mathrm{clmp}^U(\mathrm{clmp}_L(x))$.

Lemma 1. *For a minimum-length P_n^k-cover x the length-k prefix of x, as well as the length-k suffix x, must both be injective.*

Proof. Non-injective such prefixes and suffixes would give rise to shorter feasible strings by dropping letters left and right, contradicting optimality. □

Sometimes the local injectivity not only applies to the optimal P_n^k-cover's two extremities but to the entire string, too. The following observation states, however, that this is not true in general.

Observation 1. *For a minimum-length P_n^k-cover x precisely one of the two following scenarios occurs.*

(i) $|x| = \binom{n}{k} + k - 1$*, e.g., when x arises from an associated universal cycle, therefore satisfies*
   (a) all length-k infixes of x are injective, and
   (b) each k-subset of $[n]$ is covered exactly once by x.

(ii) $|x| > \binom{n}{k} + k - 1$*, where the excess in length compared to the best-case in (i) arises from meeting the negation of (a) or the negation of (b), i.e., non-injectivity or excess-coverage. Representatives fulfilling both negations cannot be excluded for general (n, k).*

Proof. Concerning (i), an arbitrary string y possesses precisely $|y| - k + 1$ infixes of length k. Therefore, in the best case, when they are all injective, we necessarily have $|x| \geq \binom{n}{k} + k - 1$. To verify that the negations of (a) and (b) can indeed apply jointly, consider for $(n, k) = (12, 3)$ the shortest string

1239b128ac13b75c1b5397ba951785c94b3ac5369723c7a562345ba974
5823579c1837286419a4c2a71c4a8596152a96b6a39c276b32a459315a
38943a148b9c6129b52946acb412c5678926c7b1a26b85143856bc5471
98613746b18a9c874c3642716a7368ab248c64527b4a78b2c83bc8

(represented over the first twelve positive hexadecimal digits) of length coinciding with $\beta_n^k = 228$ (the derived dual bound in Sect. 5) covers the set $\{2, 3, 7\}$ twice—at the 42th and the 69th position. On the other hand, at the 96th position, we locate the non-injective 3-infix 6b6. □

Remark 1. As Observation 1 indicates that in general an optimizer can simultaneously carry local non-injectiveness as well as excess-coverage of the k-sets, this raises the question if in such a situation one is always able to find specific optimal representatives which experience exclusively one of the two phenomena. For $(n, k) = (5, 3)$, on the one hand, the string 1234515241352 covers each 3-subsets exactly once but contains the non-injective infix 515. On the other hand, 1234512413524 is an alternative optimizer doubly covering $\{1, 2, 4\}$, however, consisting exclusively of injective length-3 infixes. If such particular minimizers always exist for more general choices of (n, k) remains here an open question.

3 Reduction to the (Euclidean) Generalized Traveling Salesperson Problem

We see the considerations described in this section as a generalization of the approach proposed in [7]. Assume we have given a vertex set V, a partition of the vertices into *clusters* $C_1, \dots, C_p \subseteq V$ which are pairwise disjoint and cover V, i.e., $\bigcup_{i=1}^{p} C_i = V$, an edge set $E = \{(u, v) \in V \times V : C_p \not\supseteq \{u, v\} \text{ for each } p\}$, and a weight function $w : E \to \mathbb{R}_{\geq 0}$. The *Generalized Traveling Salesperson Problem* (GTSP) asks to minimize $\sum_{i=1}^{p-1} w(z_i, z_{i+1}) + w(z_p, z_1)$ over all directed cycles $(z_1, \dots, z_p) \in V^p$ in $G = (V, E)$ that contain precisely one representative $z_i \in C_i$ per cluster (*generalized Hamiltonian cycles*). If the weight function w is invariant under inversions of the edge direction, we call it a symmetric GTSP. We face an *Euclidean* instance if the triangle inequality applies for w [12].

Let us consider the set of all length-k injective strings over the alphabet $[n]$ as vertices of a GTSP instance. Moreover, let us regard two such vertices $s, t \in [n]^k$ as located in the same cluster if their letters are the same up to reordering, i.e., $\{s_1, \dots, s_k\} = \{t_1, \dots, t_k\}$. We therefore have $\binom{n}{k}$ clusters of uniform size $k!$. Furthermore, consider the following weight function $w(s, t) := k - \max\{r \in \mathbb{N} : \text{the length-}r \text{ suffix of } s \text{ coincides with the length-}r \text{ prefix of } t\}$, i.e., the minimum number of letters that have to be appended to s such that the result contains t as suffix. The larger the overlap between suffixes and prefixes of s and t, respectively, the smaller will be $w(s, t)$. Choosing, e.g., $s =$ 234 and $t =$ 345 with $w(s, t) = 1 \neq 3 = w(t, s)$, we see that w is not symmetric. However, it is immediate that w keeps such instances Euclidean: In fact if $w(s, t) + w(t, u) < w(s, u)$, then, in order to cover u starting from a current suffix s, one would obtain a shorter result by appending the letters responsible for the transition from s to

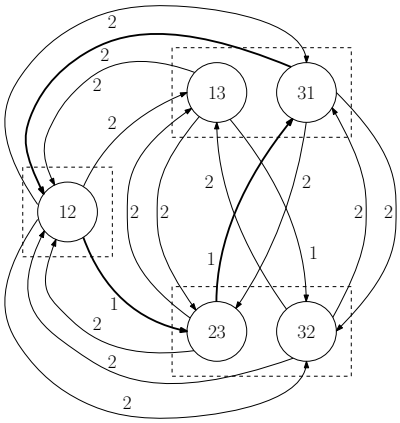

Fig. 1. The constructed GTSP instance for $(n, k) = (3, 2)$. Three clusters are present, one a singleton consisting of $\sigma = \texttt{12}$. The directed edges' labels specify the weight w.

t and afterwards those from t to u, contradicting minimality in the definition of $w(s, u)$.

The following result is based on the observation that each length-k prefix of an optimal P_n^k-cover is injective (see Lemma 1) and P_n^k-covers are closed under letter relabelings (i.e., permutations of the alphabet).

Lemma 2. *Consider the injective string $\sigma = 12\cdots k$. Assume all the $k! - 1$ vertices different from σ have been eliminated from the σ-containing cluster in the aforementioned edge-weighted graph $G = (V, E, w)$, with the difference however, that values $w(s, \sigma)$ have now been overwritten by k, for each $s \in V \setminus \{\sigma\}$; see Fig. 1 for an illustration. Then, an* arbitrary *generalized Hamiltonian cycle naturally corresponds to a P_n^k-cover.*

Proof. Such a directed cycle in fact instructs us on how to transition from a partially constructed string starting with $\sigma = 12\cdots k$: Either the transitional weight $w(z_i, z_{i+1}) = k$ tells us just to append the entire string z_{i+1} to our current partially constructed string, or if $w(z_i, z_{i+1}) < k$, to make use of a respective part of the suffix of z_i in order to save a certain prefix of z_{i+1} in the concatenation. At the same time the weight of the cycle resembles the total string length; the particularity that each $w(s, \sigma) = k$, $s \in V \setminus \{\sigma\}$, is used such that, when the cycle is eventually closed, the tour length finally takes the weight of the starting prefix $12\cdots k$ into account, which earlier never appeared as a successor. □

Lemma 3. *Each* minimum-length *P_n^k-cover is obtainable from a minimum-weight generalized Hamiltonian cycle in the (n, k)-instance of the GTSP.*

Proof. Let $s = s_1 \cdots s_m$ be a minimum-length P_n^k-cover which by Lemma 1 and relabeling of the letters we can assume to start with $12\cdots k$. Assume $\mathcal{F} :=$

$(F_i : i = 1, \dots, \binom{n}{k})$ is any enumeration of P_n^k. Define the first occurrences $i_j = \min\{h : \{s_h, \dots, s_{h+k-1}\} = F_j\}$, $j = 1, \dots, |\mathcal{F}|$. By changing the initial enumeration of F_j we can without loss of generality assume $i_1 < \cdots < i_{|\mathcal{F}|}$. Clearly then $i_1 = 1$ and $i_{|\mathcal{F}|} = m - k + 1$ for the optimality of s. Now if $i_{j+1} - i_j \leq k$ for all $1 \leq j < |\mathcal{F}|$ then s is certainly in the solution space of the the GTSP instance for (n, k). Assume therefore there exists j such that $i_{j+1} - i_j > k$. Then the substring $s_{i_j+k} \cdots s_{i_{j+1}-1}$ does not cover a subset which is not already covered by $s_1 \cdots s_{i_j+k-1} s_{i_{j+1}} \cdots s_m$ in contradiction to the optimality of s. □

It is immediate that a non-optimal P_n^k-cover in general is not always expressible in the fashion of a corresponding generalized Hamiltonian cycle. The proposed GTSP approach can be as well used to address the aforementioned shortest universal cycle covering problem [3,14] by relinquishing the weight-overwriting step in Lemma 2 concerning the σ-ingoing edges.

Remark 2. An asymmetric GTSP instance can be converted to an equivalent symmetric GTSP instance by adapting a folklore reduction (e.g., appearing in [9]) from directed to undirected Hamiltonicity; it doubles the number of vertices.

4 Computational experiments

In this section we study the computational feasibility of the aforementioned GTSP approach. The GTSP instances are created in `Julia 1.12.1`, then are passed to the metaheuristic solver `GLNS` [18] (version) which is launched in solving modality `mode=default`. The experiments were run on a server cluster consisting of AMD EPYC 7402, 2.80GHz CPUs each with 128 GB of RAM.

To report reduced optimality gaps, we fall as well back on the improved novel lower bound β_n^k on minimum-length P_n^k-covers derived in Sect. 5. We reserve for `GLNS` a time budget proportional to the number of vertices in the instance to be solved, such that the largest instance ($(n, k) = (13, 5)$ leading to 154321 vertices) is solved using 48 hours of computation time. The latter is prematurely stopped in case the incumbent's objective function value collapses with the dual bound β_n^k. The results are reported in Table 1 including the optimality gap "$\Delta(\%)$", i.e., the difference of primal and dual bound, afterwards renormalized by dual bound (and stated in percent).

Alternatively to our solver run, we can obtain for $(n, k) = (8, 3)$ a length-m P_n^k-cover with $m = 58 = \binom{8}{3} + 2$ via a known universal cycle [8]—just as for $(n, k) \in \{(10, 3), (11, 13), (13, 3)\}$. Instead for $k = 3$ and $n \in \{9, 12\}$ as well as for $(n, k) = (8, 4)$ the optimal bounds in Table 1 refer to so-far unknown minimum-length solutions we found; Table 2 collects the respective minimizers.

While oftentimes the interplay of β_n^k and the output of `GLNS` yielded convincing results, it might be interesting to mention the following insights which we obtained when assessing the feasibility via exact methods for the GTSP (or via naive encodings of the problem) in this concrete setting: In pivoting experiments we found that a compact Mixed Integer Linear Program (MILP) [16,

Table 1. `GLNS` run on GTSP (n, k)-instances. Bold numbers indicate attained minima, asterisks previously unknown/unproved ones.

n	k	Incumbent	Dual β_n^k	$\Delta(\%)$	Time used (min)	Time limit (min)
8	3	**58**	**58**	0%	1	7
8	4	∗**74**	**74**	0%	31	31
8	5	65	60	8.33%	124	124
8	6	38	34	11.76%	363	363
9	3	∗**90**	**90**	0%	1	10
9	4	136	129	5.43%	57	57
9	5	144	130	10.77%	280	280
9	6	102	91	12.09%	1116	1116
10	3	**122**	**122**	0%	1	14
10	4	229	213	7.51%	94	94
10	5	294	260	13.08%	563	563
10	6	256	215	19.07%	2809	2809
11	3	**167**	**167**	0%	11	19
11	4	359	333	7.81%	148	148
11	5	535	466	14.81%	1033	1033
12	3	∗**228**	**228**	0%	6	25
12	4	549	504	8.93%	222	222
12	5	930	796	16.83%	1772	1772
13	3	296	288	2.78%	32	32
13	4	790	718	10.03%	320	320
13	5	1498	1291	16.03%	2880	2880

pp. 823f] for the GTSP attacked by a state-of-the-art solver (Gurobi[1], version 12.0.3) terminated for $(n, k) = (8, 3)$ with the optimum of 58 within five minutes of computation time. However, termination with the loose interval $[4, 277]$ for the optimum was the result for $(n, k) = (8, 4)$ with a two hours time limit. The latter and similar results lead us to not further pursue such an approach apparently being far from competitive to those in Table 1.

Explaining the decay of performance already for such small alphabet size by the polynomial—but already too large—number of Single Commodity Flow constraints, we also tried a Branch-and-Cut approach: Following Remark 2, we relied on symmetric GTSP instances and fed them to the solver[2] implemented in [6] being a recent adaptation of the framework originally proposed by [4]. Taking for comparison $(n, k) = (8, 4)$, we notice that after two hours and 3898 user cuts, a linear programming relaxation objective of 70.66 was obtained (leading to an almost 18 times better dual bound than the one arisen from the compact MILP); on the downside, no feasible solution was yet found.

[1] `https://www.gurobi.com/` (accessed 2025-11-10)

[2] Itself calling the solver `CPLEX` (in our setup version 12.8), `https://www.ibm.com/products/ilog-cplex-optimization-studio` (accessed 2025-11-10)

Table 2. Optimal P_n^k-covers (over hexadecimal alphabet).

(n,k)	Minimizer found
$(8,4)$	12345162731587241675834678235682346527438174581245861437521365746281367812
$(9,3)$	123415263156417895679468945793568349264378453768235827918291427581627159237169318425467893
$(10,3)$	123456271563189a7896a859a4679567a56845a39847a38754936a28634a2749263792843725382 5a184a164917681537142619 23a1782159314529a12
$(11,3)$	123456143652789ab78ab679a68b5a7689b569b46857948a5674ab38745b36945a397b4839b2a734b26735928a349264a237b253825a923b182b194a195819271a261b571836a15247168241b7315481a391632791ba
$(12,3)$	See Observation 1.
$(13,3)$	12345623561789abc89bc78ba79c679b6ca968c5ab68a59c4a876b57c4b958b47a567ac84965cb3975864ab3854783c45b36c28639c2b73a94736a2953ac2764b2843b2643a249382735c173c146c192b182a183a1c245a14591a72961b527158293152ab1742163418941bc816a791b32c5

Additionally, for the decision variant of the problem, we tested an approach where we let bitvectors one-hot encode the letters and check satisfiability of a corresponding formula involving sums of k-near letters with the theorem prover `z3`[3] (version, 4.12.2). Here, for $(n,k) = (8,3)$ a 12.07%-suboptimal solution (of length 60) could be found in five minutes computation time, but already a 10.34%-suboptimal solution (of length 59) could not be found in two hours. When trying to verify that there is no solution better than the known optimum 58 in Table 1, the approach failed to do so even within two hours; we conclude that the approach seems not helpful for obtaining (near-)optimal dual bounds.

5 A new lower bound

We derive a new lower bound using ideas of the existing approach in [13] in a refined manner. To aid comprehension of the proof of the subsequent Theorem 1, consider the string s over the alphabet $[6]$ given by

$$1234562.$$

Let $(n,k) = (6,5)$ and assume we want to extend s to a P_n^k-cover $\hat{s}$ which therefore contains s as a prefix. Consider the letter 6. There are currently two length-5

[3] `https://github.com/Z3Prover/z3` (accessed 2025-11-10)

substrings of s containing the letter 6, namely `23456` and `34562` both covering the subset $\{2,3,4,5,6\}$. Since $k=5$ and the letter 6 is in the penultimate position of s we can give rise to new length-k substrings containing 6 by appending up to three letters in $[n]\setminus\{6\}$; e.g., via the following three underlined letters:

`1234562`134

Crucially, appending a fourth letter (non-coincident with 6) will not create a new length-k substring containing 6. We can therefore say that extending s by three letters (non-coincident with 6) can at most cover three new k-subsets containing 6. Hence, together with the one k-subset covered by s, we can cover at most 4 k-subsets containing 6 without appending the letter 6.

To cover the remaining $\binom{5}{4}-4=1$ subsets containing 6 (namely $\{1,3,4,5,6\}$), we consequently will have to place more occurrences of the letter 6 in $\hat{s}$. Since each occurrence of the letter 6 can be contained by at most k surrounding length-k substrings which can cover at most k-subsets containing 6 we need at least $\lceil\left(\binom{5}{4}-4\right)/5\rceil=1$ additional occurrences of 6 and hence at least two occurrences of 6 in $\hat{s}$ and hence $\hat{s}$ must be at least $7+3+1=11$ letters long.

Note that the behavior of the bound on this specific $\hat{s}$ does not apply to all P_6^5-cover as is evidenced by the P_6^5-cover `1234561234` which is only of length 10. This idea of counting additional occurrences of certain letters is the key ingredient for proving Theorem 1.

Theorem 1. *Let s be a string, and $\hat{s}$ be a P_n^k-cover containing s as infix, i.e., $\hat{s}=rst$ for some strings r and t. Then $|\hat{s}|\geq\sum_{i=1}^{n}\mu_{k,i}^{r,t}$ with*

$$\mu_{k,i}^{r,t}:=\left\lceil\frac{a_{k,i}-\operatorname{clmp}_0^{\min(|r|,a_{k,i})}(k-f_i-1)-\operatorname{clmp}_0^{\min(|t|,a_{k,i})}(k-l_i-1)}{k}\right\rceil,\quad(1)$$

where $a_{k,i}$ counts the number of k-subsets containing i that are not yet covered by s, and f_i (respectively l_i) counts the number of letters appearing strictly before (respectively after) the first (respectively last) occurrence of i in s. The latter numbers f_i and l_i are understood as ∞ if letter i does not even occur in s.

Proof. We think of s having been (gradually) extended to $\hat{s}=rst$, eventually guaranteeing feasibility for the P_n^k-cover problem. The proof is similar to Lipski Jr.'s argument [13, p. 254] with the difference of more arduously showing each term $\mu_{k,i}^{r,t}$ to be a lower bound on the number of occurrences of the letter i in rt (i.e., occurrences of letter i in $\hat{s}$ without the occurrences in s) necessarily needed in order to cover all k-subsets of $[n]$ containing i. Summing these lower bounds up for all $i\in[n]$ then yields the desired estimate on the length.

We now show that (1) is a lower bound on the aforementioned number of occurrences. For the reader's convenience the situation shall be thought of as illustrated in Fig. 2 which is meant to address special cases such as $f_i=l_i=\infty$ or $|s|=0$, too. The number of i-entries present in rt is certainly not lower than in the following best-case meeting three aspects: Firstly, left from the f_i entries preceding the first occurrence of i there are $k-1-f_i$ letters leading each to a

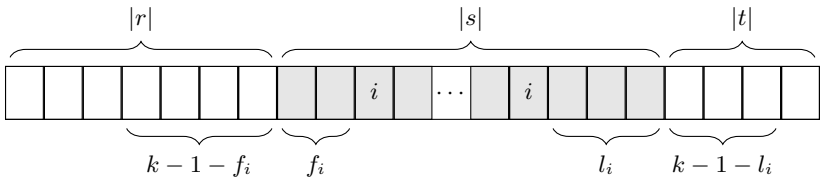

Fig. 2. The role of f_i and l_i for deriving the letter-wise lower bound $\mu_{k,i}^{r,t}$.

novel coverage of one of the so-far uncovered subsets counted by $a_{k,i}$. Secondly, the same applies likewise in rightward extending direction, i.e., for l_i. Thirdly, all potentially still not-yet covered i-containing subsets in P_n^k are novelly covered using a letter-i minimum-density, i.e., at least each k-th entry needs to host the letter i throughout the part of the string rt whose positions are not involved in the aforementioned first two novel-coverage scenarios.

Here, in all three aforementioned scenarios, carefully taking into consideration the unexceedable (length-)limits given by $|r|$, $|t|$ as well as $a_{k,i}$, non-occurrences ($f_i = l_i = \infty$), and potential impossibilities due to divisibility conditions of "each k-th letter"-densities (permitting to use the ceil function), we end up with a lower bound precisely coinciding with $\mu_{k,i}^{r,t}$. □

Corollary 1. *Let $n \in \mathbb{N}$, $1 \leq k \leq n$ and let s be a string. Let $\hat{s}$ be a P_n^k-cover, that contains s as a factor, i.e., $\hat{s} = rst$ for some strings r and t. Then*

$$|\hat{s}| \geq |s| + \sum_{i=1}^{n} \left\lceil \frac{a_{k,i} - \mathrm{clmp}_0^{a_{k,i}}(k - f_i - 1) - \mathrm{clmp}_0^{a_{k,i}}(k - l_i - 1)}{k} \right\rceil, \quad (2)$$

where $a_{k,i}$, f_i, and l_i are understood as in Theorem 1.

Proof. Use the simple fact that $\min(|r|, a_{k,i}) \leq a_{k,i}$ and $\min(|t|, a_{k,i}) \leq a_{k,i}$. □

Corollary 2. *Let $n \in \mathbb{N}$, $1 \leq k \leq n$ and let s be a string. For each P_n^k-cover $\hat{s}$ that contains s as a prefix, we have*

$$|\hat{s}| \geq |s| + \sum_{i=1}^{n} \left\lceil \frac{a_{k,i} - \mathrm{clmp}_0^{a_{k,i}}(k - l_i - 1)}{k} \right\rceil, \quad (3)$$

where $a_{k,i}$, f_i, and l_i are understood as in Theorem 1.

Proof. Set $r = \varepsilon$ in Theorem 1 and use again $\min(|t|, a_{k,i}) \leq a_{k,i}$. □

The previous results can now be combined to obtain our desired dual bound.

Proposition 1. *Any P_n^k-cover s fulfills*

$$|s| \geq \beta_n^k := k + (n - k) \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k} \left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil. \quad (4)$$

Proof. By Lemma 1 we may assume that s has an injective length-k prefix. Since all P_n^k-covers are invariant under bijective letter relabelings, we can without loss of generality assume $s_1 \cdots s_k = 1 \cdots k$. We now want to apply Corollary 2 on $s_1 \cdots s_k$. Notice that this string, by design, consists of only a single injective length-k substring and thus, in the notation of Corollary 2, $a_{k,i} = \binom{n-1}{k-1} - e_{i,k}$, where $e_{i,k} = 1$ if $i \leq k$, otherwise $e_{i,k} = 0$. Due to the particular ordering chosen, we get that $l_i = k - i$ for $i \leq k$ and since the letters in $\{k+1, \ldots, n\}$ do not appear in $s_1 \cdots s_k$, we have $l_i = \infty$ for $i > k$. For the case $1 \leq i \leq k$ we can now compute $\max(k - l_i - 1, 0) = \max(k - (k-i) - 1, 0) = i - 1$. For $k < i \leq n$ the computation yields vanishing maxima and thus by (3) we have

$$\begin{aligned} |s| &\geq k + \sum_{i=1}^{k} \left\lceil \frac{\binom{n-1}{k-1} - i}{k} \right\rceil + \sum_{i=k+1}^{n} \left\lceil \frac{\binom{n-1}{k-1}}{k} \right\rceil \\ &= k + \sum_{i=1}^{k} \left\lceil \frac{\binom{n-1}{k-1} - i}{k} \right\rceil + (n-k) \left\lceil \frac{\binom{n-1}{k-1}}{k} \right\rceil, \end{aligned} \tag{5}$$

using the identity $\frac{n}{k}\binom{n-1}{k-1} = \binom{n}{k}$ simplifiable to the right-hand side of (4). □

Remark 3. We have

$$\beta_n^k = k - 1 + (n-k+1) \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k-1} \left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil. \tag{6}$$

Corollary 3. *For minimum-length P_n^k-covers s the following assertions hold.*

(i) If $n \mid \binom{n}{k}$, then (4) agrees with the first bound of Lipski Jr. [13, p. 254], i.e.,

$$|s| \geq \beta_n^k = \binom{n}{k} + k - 1. \tag{7}$$

(ii) If $n \nmid \binom{n}{k}$, then (4), using $r := \binom{n}{k} \bmod n$, can be written as

$$|s| \geq \beta_n^k = n \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \left\lceil \frac{kr}{n} \right\rceil - 1. \tag{8}$$

Proof. Let us first address the proof of (i). Since $n \mid \binom{n}{k}$ we have that

$$\left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil = \left\lceil \frac{\binom{n}{k}}{n} \right\rceil = \frac{\binom{n}{k}}{n}$$

for all $i \leq k-1$ and by means of Remark 3 we get that

$$\begin{aligned} |s| &\geq k - 1 + (n-k+1) \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k-1} \left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil \\ &= k - 1 + (n-k+1)\frac{\binom{n}{k}}{n} + (k-1)\frac{\binom{n}{k}}{n} \\ &= \binom{n}{k} + k - 1, \end{aligned}$$

which is exactly the estimate of Lipski Jr. [13, p. 254].

Next, let us prove (ii). Let $r = \binom{n}{k} \bmod n$. Clearly $r > 0$ due to $n \nmid \binom{n}{k}$. Since the index $i \leq k-1$ in the sum of (6) and thus $i/k < 1$, we are interested in when

$$\left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil = \left\lceil \frac{\binom{n}{k}}{n} \right\rceil - 1.$$

Clearly this is the case exactly when $r/n - i/k \leq 0$ which can be equivalently rewritten as $i \geq kr/n$ and since $i \in \mathbb{N}$ this is again equivalent to $i \geq \lceil kr/n \rceil$. Applying our newfound knowledge to (6) we get

$$\begin{aligned} |s| &\geq k - 1 + (n-k+1)\left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k-1} \left\lceil \frac{\binom{n}{k}}{n} - \frac{i}{k} \right\rceil \\ &= k - 1 + (n-k+1)\left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \sum_{i=1}^{k-1} \left\lceil \frac{\binom{n}{k}}{n} \right\rceil - \left(k - \left\lceil \frac{kr}{n} \right\rceil \right) \\ &= n \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \left\lceil \frac{kr}{n} \right\rceil - 1. \end{aligned}$$

□

It is a natural question to ask when exactly (or even if) the lower bound β_n^k compares favorably to the two simpler bounds due to Lipski Jr. [13, p. 254]. In order to answer this question we first need the following intermediate result.

Proposition 2. *If $n \nmid \binom{n}{k}$, then the lower bound in (8) is strictly tighter than the one due to Lipski Jr. [13, p. 254], i.e., with $\binom{n}{k} \bmod n = r > 0$, we have*

$$n \left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \left\lceil \frac{kr}{n} \right\rceil - 1 > \binom{n}{k} + k - 1.$$

Proof. Let $r = \binom{n}{k} \bmod n$. By using the fact that $\lceil \binom{n}{k}/n \rceil = \left(\binom{n}{k} + n - r \right)/n$, we can rewrite (8) slightly differently as

$$|s| \geq \binom{n}{k} + \left\lceil \frac{kr}{n} \right\rceil + n - (r+1). \tag{9}$$

Therefore our proof reduces to showing that

$$\binom{n}{k} + \left\lceil \frac{kr}{n} \right\rceil + n - (r+1) > \binom{n}{k} + k - 1,$$

which holds due to $r < n$ and the chain of inequalities

$$r - \left\lceil \frac{kr}{n} \right\rceil \leq r - \frac{kr}{n} = \frac{r}{n}(n-k) < n - k.$$

□

Theorem 2. *The following inequality*

$$\beta_n^k \geq \max\left(\binom{n}{k} + k - 1, n\left\lceil \frac{\binom{n}{k}}{n} \right\rceil\right)$$

holds, with equality if and only if $kr \leq n$ *where* $r = \binom{n}{k} \bmod n$.

Proof. Assume first that $n \mid \binom{n}{k}$. Then we may apply the case of Corollary 3 (i) and we get that

$$\beta_n^k = \binom{n}{k} + k - 1 = \max\left(\binom{n}{k} + k - 1, n\left\lceil \frac{\binom{n}{k}}{n} \right\rceil\right),$$

indeed with equality since $n\lceil \binom{n}{k}/n \rceil = \binom{n}{k}$ for $r = \binom{n}{k} \bmod n = 0$ and therefore $kr = 0 \leq n$. Next, assume $n \nmid \binom{n}{k}$ and hence $r = \binom{n}{k} \bmod n > 0$. By the case of Corollary 3 (ii) we have that

$$\beta_n^k = n\left\lceil \frac{\binom{n}{k}}{n} \right\rceil + \left\lceil \frac{kr}{n} \right\rceil - 1.$$

Let us further assume that $kr > n$. Then by Proposition 2 we see that $\beta_n^k > \binom{n}{k} + k - 1$, clearly implying $\beta_n^k > n\lceil \binom{n}{k}/n \rceil$. On the other hand if $1 \leq kr \leq n$ then again by means of Proposition 2 and the fact that $\lceil kr/n \rceil - 1 = 0$ we have that

$$\beta_n^k = n\left\lceil \frac{\binom{n}{k}}{n} \right\rceil > \binom{n}{k} + k - 1,$$

completing the proof. □

Remark 4. For the more constrained problem of covering all the subsets of $[n]$ studied by Lipski Jr. [13], the tightest-known bound (see also [17, Seq. A348574]) is automatically strengthened by $\beta_n^{\lceil n/2 \rceil}$.

6 Conclusion

We adopted a GTSP-based viewpoint on P_n^k-covers turning out to be an equivalent characterization for optimal solutions. This approach would also allow to model more flexible versions of the problem where only a custom selection of strings representing the set is needed to be covered (in the most restrictive case we recover instances of the shortest superstring problem), and more general designs of the clusters would be conceivable. The derived lower bound β_n^k appears to be near-optimal, making our class of GTSP instances particularly appealing for the performance-study of exact solvers; additionally one might want to fall back on the subclass whose known optimality results from the literature on universal cycles [8]. Our results lead to several remaining open problems, in particular the following ones.

- Can we always find optima exhibiting exclusively excess-coverage or exclusively non-injectivity (see Remark 1), when at least one of them must appear (see Observation 1)?
- Can the bound in Theorem 1 be exploited for an efficient Branch-and-Bound approach; and is the bound β_n^k an optimal one for $k = 3$? (By [8] and Theorem 2 this already applies for all $n \geq 8$ which are nondivisible by 3.)
- Can we strategically partition the k-subsets such that we can separately use the GTSP approach on smaller instances and merge the arising cycles?

Disclosure of Interests. The authors have no competing interests.

References

1. Armen, C., Stein, C.: Short superstrings and the structure of overlapping strings. Journal of Computational Biology **2**(2), 307–332 (1995). https://doi.org/10.1089/CMB.1995.2.307
2. Chung, F.R.K., Diaconis, P., Graham, R.L.: Universal cycles for combinatorial structures. Discrete Mathematics **110**(1-3), 43–59 (1992). https://doi.org/10.1016/0012-365X(92)90699-G
3. Dębski, M., Lonc, Z.: Universal cycle packings and coverings for k-subsets of an n-set. Graphs and Combinatorics **32**(6), 2323–2337 (2016). https://doi.org/10.1007/S00373-016-1727-6
4. Fischetti, M., Salazar-Gonzalez, J.J., Toth, P.: The generalized traveling salesman and orienteering problems, pp. 609–662. Springer (2007). https://doi.org/10.1007/0-306-48213-4_13
5. Glock, S., Joos, F., Kühn, D., Osthus, D.: Euler tours in hypergraphs. Combinatorica **40**(5), 679–690 (2020). https://doi.org/10.1007/S00493-020-4046-8
6. Heßler, K., Irnich, S.: Exact solution of the single-picker routing problem with scattered storage. INFORMS Journal on Computing **36**(6), 1417–1435 (2024). https://doi.org/10.1287/IJOC.2023.0075
7. Houston, R.: Tackling the minimal superpermutation problem. arXiv preprint [math.CO] (2014). https://doi.org/10.48550/arXiv.1408.5108
8. Jackson, B.W.: Universal cycles of k-subsets and k-permutations. Discrete Mathematics **117**(1-3), 141–150 (1993). https://doi.org/10.1016/0012-365X(93)90330-V
9. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W. (eds.) Proceedings of a symposium on the Complexity of Computer Computations, held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. pp. 85–103. The IBM Research Symposia Series, Plenum Press, New York (1972). https://doi.org/10.1007/978-1-4684-2001-2_9
10. Kasprzak, M.: On the link between DNA sequencing and graph theory. Computational Methods in Science and Technology **10**(1), 39–47 (2004), https://cmst.eu/wp-content/uploads/files/10.12921_cmst.2004.10.01.39-46_Kasprzak.pdf
11. Kou, L.T.: Polynomial complete consecutive information retrieval problems. SIAM Journal on Computing **6**(1), 67–75 (1977). https://doi.org/10.1137/0206004, https://doi.org/10.1137/0206004

12. Laporte, G., Nobert, Y.: Generalized travelling salesman problem through n sets of nodes: An integer programming approach. INFOR: Information Systems and Operational Research **21**(1), 61–75 (1983). https://doi.org/10.1080/03155986.1983.11731885
13. Lipski Jr., W.: On strings containing all subsets as substrings. Discrete Mathematics **21**(3), 253–259 (1978). https://doi.org/10.1016/0012-365X(78)90157-7
14. Lonc, Z., Traczyk, T., Truszczynski, M.: Optimal f-graphs for the family of all k-subsets of an n-set. In: Ghosh, S.P., Kambayashi, Y., Lipski Jr., W. (eds.) Data base file organization. pp. 247–270. Academic Press (1983)
15. Mayne, A., James, E.B.: Information compression by factorising common strings. The Computer Journal **18**(2), 157–160 (1975). https://doi.org/10.1093/COMJNL/18.2.157
16. Pop, P.C., Cosma, O., Sabo, C., Sitar, C.P.: A comprehensive survey on the generalized traveling salesman problem. European Journal of Operational Research **314**(3), 819–835 (2024). https://doi.org/10.1016/J.EJOR.2023.07.022
17. Sloane, N. J. A. (editor): The On-Line Encyclopedia of Integer Sequences (2025), published electronically at https://oeis.org/, accessed: 2025-11-10
18. Smith, S.L., Imeson, F.: GLNS: An effective large neighborhood search heuristic for the generalized traveling salesman problem. Computers & Operations Research **87**, 1–19 (2017). https://doi.org/10.1016/J.COR.2017.05.010
19. Tabatabai, P.: On sequences covering subsets of a finite set. Master's thesis, TU Graz (2018), https://diglib.tugraz.at/on-sequences-covering-subsets-of-a-finite-set-2018