

# On the minimum substring cover problem

Stéphane Vialette

LRI, UMR 8623, Univ. Paris-Sud, Orsay, France  
Stephane.Vialette@lri.fr

In a covering problem we are faced with the following situation: We are given two (not necessarily disjoint) sets of elements, the base elements and the covering elements, and the goal is to find a minimum (weight) subset of covering elements that "covers" all the base elements. The exact notion of covering differs from problem to problem, yet this abstract setting is common to many classical combinatorial problems in various application areas. Two famous examples are Minimum Set Cover where the covering elements are subsets of the base elements and the notion of covering corresponds to set inclusion and Minimum Vertex Cover where the setting is graph-theoretic and the notion of covering corresponds to incidence between vertices and edges. Ever since the early days of combinatorial optimization, research on covering problems such as the two examples above proved extremely fruitful in laying down fundamental techniques and ideas. The early work of Johnson and Lovász on Minimum Set Cover pioneered the greedy analysis approach, while Chvátal gave the first analysis based on linear programming (LP) while tackling the same problem. The first LP-rounding algorithm by Hochbaum was also designed for Minimum Set Cover, while Bar-Yehuda and Even gave the first Primal-Dual and Local-Ratio algorithms for Minimum Vertex Cover.

We introduce here a new covering problem which resides in the realm of strings. A string  $c$  is a substring of a string  $s$ , if  $c$  can be obtained by deleting any number of consecutive letters from both ends of  $s$ . In our covering problem, the base elements are strings and the covering elements are their substrings. The notion of covering corresponds to string-factorization. More formally, for a given set of strings  $S$ , let  $C(S)$  denote the set of all substrings of strings in  $S$ . We define a cover of  $S$  to be a subset  $C \subseteq C(S)$  where any string  $s \in S$  can be written as a concatenation of strings in  $C$ . If each string in  $S$  can be written as a concatenation of at most  $\ell$  strings in  $C$ , we say that  $C$  is an  $\ell$ -cover of  $S$ . Given a weight function  $\omega : C(S) \rightarrow \mathbb{Q}^+$ , we are interested in computing an  $\ell$ -cover of  $S$  with minimum possible weight:

**Minimum Substring Cover:**

**Instance:** A set of strings  $S$ , a weight function  $\omega : C(S) \rightarrow \mathbb{Q}^+$ , and an integer  $\ell \geq 2$ .

**Solution:** An  $\ell$ -cover  $C$  of  $S$ .

**Measure:** Total weight of the cover, *i.e.*,  $\omega(C) = \sum_{c \in C} \omega(c)$ .

*Example 1.* Consider the set of strings  $S = \{a, aab, aba\}$ . Then  $C(S) = \{a, b, aa, ab, ba, aab, aba\}$ , and  $C_1 = \{a, b\}$  and  $C_2 = \{a, ab\}$  are covers of  $S$ . The cover  $C_1$  is a 3-cover of  $S$ , while  $C_2$  is a 2-cover.

We use  $n$  to denote the number of strings in  $S$ , and  $m$  to denote the maximum length of any string in  $S$ , *i.e.*,  $n = |S|$  and  $m = \max\{|s| : s \in S\}$ . Note that in case  $\ell \geq m$ , there is no actual bound on the concatenation length of the required cover, and this case is denoted by  $\ell = \infty$ . An  $\infty$ -cover is referred to simply as a cover. Another interesting special case is when  $\ell = 2$ . In this case, we are required to cover  $S$  with a set of prefixes and suffixes in  $S$ . This two extremal cases both give a certain amount of combinatorial leverage, and therefore deserve particular consideration.

We will first present some lower bounds on the approximation factors of polynomial-time algorithms for Minimum Substring Cover. We will show that, in general, the problem is NP-hard to approximate within a factor of  $c \log n$  for some  $c > 0$ , and within  $\lfloor m/2 \rfloor - 1 - \varepsilon$  and all  $\varepsilon > 0$ . We will also show that the problem remains APX-hard even when  $m$  is constant, and the given weight function is either the unitary or the length-weighted function. Next, we will apply the local-ratio technique to obtain three approximation algorithms with performance ratios  $\binom{m+1}{2} - 1$ ,  $m - 1$ , and  $m$ , where the last two are specializations of the first to the cases of  $\ell = 2$  and  $\ell = \infty$  (the latter only applies for restricted types of weight functions). Finally, we will present an algorithm based on rounding the linear programming relaxation of the problem, which achieves a performance ratio of  $O(\log n m^{\frac{(\ell-1)^2}{\ell}})$  with high probability. This algorithm is an extension of an algorithm of Hajiaghayi *et al.* used for solving the Minimum Multicolored Subgraph problem.