



Mont-Saint-Aignan, 23rd December 2022

Prof. Thierry Lecroq
LITIS UR 4108
UFR des Sciences et Techniques
Université de Rouen Normandie
76821 MONT-SAINT-AIGNAN Cedex
FRANCE
Tel: +(33) (0)2 35 14 65 81
Email: Thierry.Lecroq@univ-rouen.fr

Review of the PhD thesis of Juliusz STRASZYŃSKI

The thesis of Juliusz STRASZYŃSKI is entitled “Exact Covers and Pattern Matching with Mismatches”. The work presented in this thesis lies in string algorithmics and deals with two important aspects that are covers and approximate matching. These are important topics since a huge part of the data produced nowadays is constituted by texts or sequences of symbols. It is then crucial to have efficient algorithms to manipulate these data.

The thesis of Juliusz STRASZYŃSKI is an article-based thesis and is composed of four chapters followed by a list of bibliographic references and six articles that he co-authored.

In Chapter 1, entitled “Introduction”, the author briefly presents the context of the work, the main notations and the notions of cover, which is somehow the generalization of period, and of pattern matching with mismatches since its contributions deal with these two notions.

Chapter 2 is entitled “Covers”. In this chapter, the author presents four results concerning:

- the computation of 2-covers of a string,
- the shortest covers of all cyclic shifts of a string,
- internal quasiperiod queries,
- the string covers of a tree.

A 2-cover of a string y of length n is a pair of strings (u, v) of the same length ($|u| = |v|$) that completely covers y . The author designed an algorithm that finds a representation of all 2-covers in $O(n \log n \log \log n)$ expected time. This representation enables to retrieve all 2-covers in $O(n \log n \log \log n + \#cov)$ time, where $\#cov$ is the number of all 2-covers. Thus it allows to find one 2-cover for each length (if it exists) in $O(n \log n \log \log n)$ total time. The problem has two variants whether or not one of the two strings u or v has to be a border of y , in which case it is called a border 2-cover. In the other case it is called a prefix-suffix 2-cover. For the prefix-suffix 2-cover case the author gives an algorithm that computes all of them in $O(n \log \log n)$ expected time (there can be up to n such 2-covers). For the border 2-cover case the author gives an algorithm that computes one such 2-cover per length (if it exists) or all shortest such 2-covers in $O(n \log n \log \log n)$ expected time (there can be up to $\Theta(n^2)$ such 2-covers). All algorithms use linear space. These results are the first known results for these special cases of covers.

The cyclic shifts of a string y is the set of strings vu where $y = uv$ for all possible pairs u, v . The author gives an $O(n \log n)$ -time and $O(n)$ -space algorithm that computes the shortest cover of every cyclic shift of a string of length n and an $O(n)$ -time algorithm that computes the shortest among these covers. These algorithms are the first known for these problems. Then the author also gives results on covers of cyclic shift of the Fibonacci strings.

Internal quasiperiod queries consists in computing covers for the factors of a string. The author shows that the shortest cover and all covers queries can be answered in $O(\log n \log \log n)$ time and $O(\log n (\log \log n)^2)$ time, respectively, with a data structure that uses $O(n \log n)$ space and can be constructed in $O(n \log n)$ time. These results are the first known for this problem.

Then the author considers covering labeled trees by a collection of paths with the same string label, called a (string) cover of a tree. He shows how to compute all covers of a directed rooted labeled tree in $O(n \log n / \log \log n)$ time and all covers of an undirected labeled tree in $O(n^2)$ time and space or in $O(n^2 \log n)$ time and $O(n)$ space. He also shows several essential differences between covers in standard strings and covers in trees. These results were the first for this problem.

The third Chapter is entitled “Pattern matching with mismatches”. The author presents results for the following two problems:

- the efficient computation of sequence mappability,
- the circular pattern matching with k mismatches.

The (k, m) -mappability problem for a string y of length n consists in computing for every position i on y the number of factors $y[j \dots j + m - 1]$ of

length m that are at Hamming distance less or equal than k from the factor of length m starting at position i , namely $y[i \dots i + m - 1]$. The author presents a solution to the problem of (k, m) -mappability in $O(n \times \min\{m^k, \log^k n\})$ time and $O(n)$ space for a fixed k . Moreover he proves that for every $k, m = \Theta(\log n)$ one cannot solve (k, m) -mappability in strongly subquadratic time, unless Strong Exponential Time Hypothesis fails. He also gives an $O(n^2)$ time and $O(n)$ space algorithm finding (k, m) -mappability for each k or for each m .

The circular pattern matching with k mismatches consists in finding the substrings of a text y of length n that are at Hamming distance at most k with any cyclic shift of a pattern x of length m . The authors first show an $O(nk)$ time algorithm and then an $(n + \frac{n}{m}k^4)$ time algorithm. Both algorithms need $O(m)$ space.

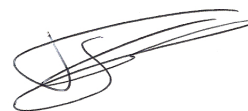
In Chapter 4 entitled “Final words”, the author concludes and gives some perspectives for future works.

All the presented results are original and use a large number of clever and efficient data structures in string algorithms. Thus Juliusz STRASZYŃSKI clearly demonstrated strong knowledge in the field and ability to design efficient solutions to important problems.

These works lead to articles published in the proceedings of important international conferences and in top-level international journals. These articles have been written with many co-authors but Mr STRASZYŃSKI identified his own contributions among these works and there are quite important.

For all these reasons I clearly think that the thesis is largely sufficient to grant a PhD and can be defended.

I also think that this thesis deserves to be awarded an honorary distinction.



Thierry Lecroq
Professor

University of Rouen Normandy, France