Bar-Ilan University
אוניברסיטת בר-אילן

Department of Computer Science
Professor Amihood Amir

המחלקה למדעי המחשב
פרופסור עמיהוד אמיר

ב"ה ד' כסלו תשפ"ג
Nov. 28th 2022

Professor Andrzej Tarlecki, Chairman
Academic Council of Mathematics and Computer and Information Sciences
University of Warsaw
Poland

Dear Professor Tarlecki,

Re: Review of the Ph.D. thesis of Juliusz Straszyński: Exact Covers and Pattern Matching with Mismatches

Attached is my review of the Ph.D. thesis of Juliusz Straszyński.
The thesis is excellent and I recommend granting him the Ph.D. with an honorary distinction. The reason for my decision appears in the review.

I am at your service if any additional information is required.

Sincerely yours,

פרופ' עמיהוד אמיר
Amihood Amir
Professor of Computer Science

Tel: 03 531 8770 :טל • Fax: 03 7384056 :פקס • amir@cs.biu.ac.il • www.cs.biu.ac.il/~amir
Bar-Ilan University (RA), Ramat Gan 52900, Israel • **www.biu.ac.il** • אוניברסיטת בר-אילן (ע"ר), רמת גן 52900, ישראל

Review of PhD Dissertation:

Exact Covers and Pattern Matching with Mismatches

By Juliusz Straszyński

Context:

The area of this thesis research is *Pattern Matching* and *Word Combinatorics*. The area was historically one of the first in Computer Science - analyzing texts. Recently, however, it became even more important due to the vast amounts of digital data on the internet and on various data repositories. The amounts of data available, and the volume of newly created data are so large that it is impossible to harness them manually. Complex algorithms are required for efficient search and manipulation of the contents. One of the contribution of this thesis is handling *approximation.* Inputing large amounts of data naturally causes errors. The errors may occur in the data gathering stage, in the input stage, and even by natural evolution and amortization. Consider, for example, DNA data. The machinery to decode the sequence may produce slight errors and measuring miscalculations, the scanners or OCR equipment may introduce errors, and the DNA itself may have changed due to evolution or mutation. Thus, the ability to efficiently count and recognize *approximate* occurrences is of utmost importance.

Another aspect of this thesis deals with efficiently discovering special structure in the data. Natural data is not random. Many real world phenomena have a particular type of event that repeats periodically during a certain period of time. Examples of highly periodic events include road traffic peaks, load peaks on web servers, monitoring events in computer networks and many others. Finding periodicity in real-world data often leads to useful insights by shedding light on the structure of the data, and giving a basis to predicting future events. Moreover, in some applications periodic patterns can point out a problem. In a computer network, for example, repeating error messages can indicate a misconfiguration, or even a security intrusion such as a port scan. Finally, the structure of the data can be exploited to enable more efficient algorithms.

This thesis deals with all above aspects. It considers covers of strings, a natural generalization of the period described above, mappability, and efficient algorithms for approximate matching in cyclic strings.

Results:

A *quasiperiod*, is a generalization of a period. For example, *abaabaaba* is clearly a repetition of *aba* three times. Yet, if we consider *abaababababa*, although it is not periodic, we strongly feel that there is a structure there. This is called a *cover*. We can cover the string with copies of *aba*, possibly overlapping. This thesis gives the first efficient algorithm for a 2-cover (where the string is covered by two substrings). In addition it gives an efficient algorithm for all string covers in a tree – another important data structure. Finally, the thesis is the first to give algorithms for *internal* quasiperiod queries. Internal queries are questions on given substrings of a text, rather

than the entire text. For example, we may ask whether a text is coverable, but a more finely-tuned question is whether some given substring of the text is coverable. Of course one can run the cover detection algorithm on the given substring, but that would mean running the algorithm over and over again for the various substrings. The better solution is to find a general preprocessing algorithm that allows subsequent fast answers to local queries. This thesis supplies the first algorithm to provide efficient answers to internal quasiperiodic queries.

The thesis also provides an efficient algorithm for *k-mappability*, essentially discovering all locations in a string where a close approximation of an input substring occurs. Finally, there is an efficient algorithm for finding approximations in circular strings.

Methodology:

The thesis defines some elementary types of query that, if answered efficiently can be used as building blocks for the important internal queries we had described. To solve these queries, the thesis makes sophisticated use of some of the most modern techniques available. Some of these methods involve building tree structures. The height of such trees is logarithmic, which would have introduced a logarithmic factor to many of the proposed solution. It was thus necessary to do more than just ingeniously use sophisticated techniques. It was necessary to create some new techniques, which have the potential of serving future efficient algorithms in the field. An example is the ability to extend a periodic substring to maximal repetition.

Publications:

The results in these thesis appeared as three journal papers, in A-tiered journals, *Algorithmica, J. Comp. Sys. Sci*., and *Theoretical Comp. Sci.* Two additional results appeared in the *International Symposium on String Processing and Information Retrieval* (SPIRE) one of the two best international conferences in the string processing area. The results already achieved international recognition and are sure to promote further research.

Exposition:

The thesis is clear and well-written. Although the material is quite hard technically, the writing is clear and convenient to follow. There is also adequate use of figures and examples.

Conclusion:

This thesis is definitely worthy of awarding a Ph.D. In addition, because of the novelty of the techniques and importance of the problems, as well as the distinguished venues it appeared in, I recommend **awarding the Ph.D. with an honorary distinction**.