Review of Doctoral Dissertation, 2025-01-10

Bartłomiej Dudek's doctoral dissertation "Equivalences between Some Problems on Strings, Trees and Graphs" investigates the computational complexity of three fundamental algorithmic problems and how they are related to various other previously well-studied problems. A number of new theoretical results are presented. The approach taken is a very modern one, based on the concept of *conditional lower bounds*, where it is assumed that some (empirically motivated) hypothesis saying that a particular famous problem X does not admit any exact algorithm faster than some specified bound is true, and one then proves that this implies a similar bound for another problem Y. This line of research belongs to a more general field known as *fine-grained complexity*, which tries to understand how different computational complexity conjectures are connected and to demonstrate that certain algorithmic problems which are already known to be solvable in polynomial time are actually computationally equivalent to each other. My impressions and opinions of the dissertation are stated below.

Strong points

The candidate has managed to make progress on some rather difficult research topics that have been studied by many people over the years, which is impressive. The presented results are nontrivial and require sophisticated mathematical machinery and many novel ideas.

The equivalences (in term of fine-grained complexity and ignoring polylogarithmic factors) between the problems of counting 4-cycles in a graph, computing the quartet distance between two unrooted phylogenetic trees, and counting 4-patterns in permutation established in chapter 6 are particularly outstanding. Furthermore, a new, original randomized algorithm for the Online Context-Free Recognition Problem which works by reducing it to multiple instances of the Online Matrix-Vector Multiplication Problem gives the first improvement on this classic problem since 1995. Also along the way, some open questions posed by Erickson in 1999 related to the 3SUM problem and by Lohrey *et al.* in 2019 related to top trees are resolved, and an open question by Even-Zohar and Leng from 2021 related to permutation patterns is partially resolved.

The research has been presented at some highly competitive international conferences in theoretical computer science. This includes two papers at the top conference STOC (the Annual ACM Symposium on Theory of Computing) in 2019 and 2020 and one paper at ISAAC (the International Symposium on Algorithms and Computation) in 2020 which won a Best Paper Award.

We can also note that the candidate has published several other high-quality papers that were not part of the dissertation (most likely because their topics were different), including a paper at STOC 2018, which indicates that the candidate has a wide range of research interests.

Recommendation

Based on the excellent technical results achieved in the doctoral dissertation, my recommendation is to award the *doktorat* degree to Mr. Bartłomiej Dudek. I would rate it as being among the top 10% of all doctoral dissertations in theoretical computer science that I have seen over the years. Consequently, I recommend it for consideration for a doctoral dissertation award or a scientific prize, as outlined in the document "Information for the guidance of referees".

Minor issues and suggestions for improvement

Here is a list of minor issues that I found during the review. I would recommend the candidate to take them into account when preparing a revised version (if any).

- In the first sentence of the acknowledgments, the name of the Ph.D. advisor seems to have been spelled incorrectly.
- Chapter 1 would benefit from proofreading by a native English speaker, as it contains several grammatical errors. As an example, consider the second sentence of the introduction: "With the recent very rapid growth in the size of data available, it becomes even more pressing need to design efficient algorithms that will be able to handle data from the real-life applications." The meaning is easily understood, but the sentence sounds a bit strange.
- Hypotheses 1.1.1–1.1.4 are all of the form ". . . cannot be solved in . . . time by a randomized algorithm", but it's not clear what this means. A trivial randomized algorithm with error probability 100% could run in $O(1)$ time, for example.
- Chapters 1 and 2 contain a lot of overlap. It would be better to restructure them so that, e.g., chapter 1 gives the general scientific background as well as the detailed overview of the thesis' contributions (including Figure 2.2) while chapter 2 surveys the previous work. Alternatively, in case it is difficult to separate the previous and new work, chapters 1 and 2 could perhaps be combined into one chapter.
- What properties would the hypothetical "surprising algorithm" for triangle listing mentioned at the end of chapter 2.1 have?
- The first paragraph of chapter 2.2 says: "For the more practically relevant case of a sparse undirected graph with m edges, Alon et al. [AYZ97] designed an $O(m^{2\omega/(\omega+1)}) = O(m^{1.41})$ time algorithm for counting triangles (their algorithm is stated for finding a triangle, but can be easily extended)." It is indeed easy to extend the method, and if you would like a reference to somewhere in the literature where it was done explicitly, you could cite [JL14] here.
- In chapter 2.2.1, it would be interesting to know more details about real-world applications of counting occurrences of patterns, and if the theoretical algorithms developed by the TCS community have had any practical relevance.
- More figures in chapter 3 to illustrate Theorem 3.2.1 and its proof would have been helpful. Figures 3.1 and 3.2 are good.
- For completeness, the analysis of the main algorithm in chapter 3 should also prove that it "succeeds with high probability", as claimed. According to my understanding, the only randomized aspect is the part that relies on a randomized algorithm by Larsen and Williams for the Online Matrix-Vector Multiplication Problem that "succeeds with high probability", but it needs to be run very many times (for all of the created intervals). Also note that in the literature, "succeeds with high probability" is often defined slightly differently, that is, $\geq 1 - \frac{1}{n^c}$ with the requirement that $c \geq 1$ rather than $c > 0$. Using that definition, one cannot just directly apply the standard union-bound technique, and so it would be informative to do a more detailed analysis of the success probability to see exactly what it is.

- Chapter 4 introduces numerous generalizations of the 3SUM problem and presents reductions between them to prove that they are subquadratic-equivalent. However, many of the reductions are very similar, so I wonder if some of them could be unified to make the proofs more general and the text more compact. Could the constructions in some of the reductions be summarized in a table, for example?

- Chapter 4.1 points out that instances of 3LDT with $c \geq 2$ and instances of Conv3LDT with $c \geq 1$ can be solved in subquadratic time using the fast Fourier transform. A short explanation of how this works and why the bound on c is different for 3LDT and Conv3LDT would be convenient for the reader.

- Chapters 2.2 and 4.1 talk about a "folklore" reduction, but they seem to refer to two totally different reductions. One is from 1-partite 3SUM to 3-partite 3SUM, which uses Alon *et al.*'s color-coding technique, while the other one is from Conv3LDT to 3LDT, which uses a simple trick to increase the size of the universe by a factor of n. Instead of just calling them "folklore", can you find any previous work in the literature that have used these two respective reduction techniques before?

- Some parts of chapter 4 refer to "LDT". Should this in fact be "3LDT"? Also, for consistency, change "3-SUM" to "3SUM" in two places on p. 53.

- When introducing top trees in chapter 5.1, also give credit to its inventors (Alstrup *et al.*).

- To illustrate the concept of a top tree clearly, a more detailed example than Figure 5.1 is needed.

- Lemma 6.2.4 states: "There exists a deterministic data structure that preprocesses a set of n weighted points in $O(n \log n)$ time and answers queries about the number or the sum of weights of points inside rectilinear rectangles in $O(\log n)$ time.". I think the query time can be improved to $O(n \log n / \log \log n)$ by instead using Theorem 12 in T. Chan, Q. He, Y. Nekrich: "Further Results on Colored Range Searching", Proceedings of SoCG 2020 (this corresponds to Theorem 20 in the full version of the paper on arXiv) with $k = 1$ "colors", assuming that the sum of the points' weights is not too large. Would using Chan *et al.*'s data structure in Lemma 6.2.4 have any effect on any of the results in chapter 6.2?

- Clarify if the tilting technique used in the reduction from counting 4-cycles in 4-circle-layered graphs to counting non-trivial patterns in chapter 6.3 is the same as the one used by Berendsohn *et al.* [BKM19] or not.

- It's difficult to understand why the stars have to be classified into type I and type II and treated separately in chapter 6.4. Why can't stars of type II be handled by the same method as the one for stars of type I, using the technique for counting "missing" stars of type I?

- The beginning of chapter 6.5.2 is a duplicate of the beginning of chapter 5.1, and should be replaced by some text referring to that part of the dissertation.

- The time complexity obtained in the final theorem when putting everything together is $\widetilde{O}(\min\{n^{1.48}, n^{1.16}d^{0.43}, nd^{0.69}\})$, which does not seem like a very natural bound. Is there a way to unify the three algorithms in order to express the trade-off between n and d more elegantly?

- It would be nice to have a summary of the remaining related open problems somewhere at the end of the dissertation, along with some new challenging open problems or conjectures posed by the candidate himself.

- Although the candidate has published many strong conference papers, he only has a single journal publication. Moreover, this publication is in *Theoretical Computer Science* which is a decent, but not really a "top", journal. For this research to have a higher future impact, I suggest spending more effort on trying to publish it in some prestigious journals, where it belongs (in my personal opinion).

Sincerely,

Jesper Jansson, Associate Professor
Graduate School of Informatics, Kyoto University
Kyoto 606-8501, Japan

E-mail address: `jj@i.kyoto-u.ac.jp`