Massachusetts Institute of Technology

Virginia Vassilevska Williams
Associate Professor

MIT CSAIL
32–G640

77 Massachusetts Avenue
Cambridge, Massachusetts
02139–4307

Phone 617 - 715 - 5833
Fax 617 - 258 - 8682
Email virgi@mit.edu
https://people.csail.mit.edu/virgi/

May 21, 2019

**Report on Debarati Das' Ph.D. Thesis**

*Introduction:* I am delighted to report on the Ph.D. thesis of Debarati Das. The thesis contains several stellar results that have excited the theoretical computer science community, and it has received strong recognition already.

*Summary of the thesis:* The thesis consists of two parts. The first part deals with algorithmic problems and the last one is focused on the design of Gray codes.

The first chapter introduces the problems of the first part: (1) approximating the edit distance of two strings, (2) approximating the edit distance between a short pattern text and every prefix of a long text, (3) limitations on certain algorithmic techniques for Boolean Matrix Multiplication and (4) algorithms for computing weight tolerant subgraphs for single source shortest paths.

The second two chapters are concerned with problem (1), approximating the edit distance of two strings. The edit distance of two strings $x$ and $y$ is the number of edit operations (symbol insertion, deletion, substitution) one needs to apply to $x$ to derive $y$. The best algorithms we know of for computing edit distance run in roughly quadratic time and there is significant evidence from fine-grained complexity that this cannot be improved substantially. The thesis presents the first ever algorithm for approximating edit distance within a multiplicative constant-factor that runs in truly subquadratic time $O(n^{2-2/7})$. This is a major result that has impressed the community. Prior to this result, it was not even clear whether such an algorithm can even exist, and there were attempts to show that it does not.

Chapter 2 applies the techniques from Chapter 1 to the approximate pattern matching problem. In approximate pattern matching, one needs to approximate the edit distance between a pattern string $p$ and all prefixes of text $t$. The thesis presents algorithms for this problem in the usual setting and in the online setting with restricted working memory. The algorithms improve on running times over previously known results for the pattern matching problem.

Chapter 4 studies the possibility of designing fast "combinatorial" algorithms for Boolean Matrix Multiplication: given two $n \times n$ Boolean Matrices $A$ and $B$, output their product $C$ defined as $C[i,j]=OR_k A[i,k] AND B[k,j]$. The thesis proposes two notions of "combinatorial" algorithms, one strengthening the other, and proves lower bounds on the time complexity of algorithms for Boolean Matrix Multiplication in those models. These lower bounds are super-quadratic. All known "combinatorial" algorithms for Boolean Matrix Multiplication have roughly cubic time complexity, and there are "non-combinatorial" algorithms running substantially faster, in $O(n^{2.373})$ time.

The fifth chapter deals with maintaining a sparse representation of a graph in the presence of edge faults/weight increments. The goal is to construct a sparse subgraph of a given directed weighted graph with a designated source vertex that preserves the distance from the source to all other vertices as long as the total weight increment of all the edges is bounded by $k$. The thesis presents fixed-parameter algorithms for this problem and shows that the size of the subgraph is close to optimal.

The last part of the thesis deals with construction of Gray codes. It considers Gray codes over binary and non-binary alphabets and it is concerned with the decision tree complexity of determining a successor for a given string in the Gray code. The thesis presents a surprising construction over non-binary alphabets that has logarithmic decision tree complexity. This is in contrast with a recent result for binary alphabet where the complexity must be linear. In the binary case, the thesis also presents a construction of a Gray code with logarithmic decision tree complexity that enumerates all but polynomially many strings of the Boolean cube.

*Evaluation:* As mentioned earlier, the results in this thesis are stellar. Even just the result on edit distance would make a wonderful Ph.D. thesis in itself. Edit distance is an extremely important problem, and obtaining fast algorithms for it has fascinated the theoretical computer science community for a long time. Obtaining exact algorithms has been extremely challenging, and the quadratic time dynamic programming algorithms from

many years ago have stood basically unchallenged. Fine-grained complexity has given us reasons why: due to Backurs and Indyk's result [STOC'15] we know that a truly subquadratic exact algorithm for edit distance would give us faster Boolean Satisfiability algorithms, thus resolving a longstanding open problem and violating the popular Strong Exponential Time Hypothesis. Even stronger results were obtained later on, and almost all hope was lost for obtaining fast edit distance algorithms. The only saving grace was that none of the (conditional) lower bounds techniques seemed to extend to approximation algorithms, so that in principle a fast and very accurate approximation algorithm for edit distance was not known to contradict any hypotheses or conjectures. Nevertheless, the best known approximation algorithm running in subquadratic time by Andoni et al. [STOC'10] was only a polylogarithmic approximation. The algorithm in this thesis is an astounding breakthrough and deservedly received the best paper at FOCS.

The results in the thesis are published in good venues: : one (as mentioned earlier) is a FOCS best paper,  on paper published at ESA, one at STACS, one at SWAT, and one is an arXiv preprint. FOCS is one of the two premier Theoretical Computer Science conferences, and ESA, STACS and SWAT are good conferences in Algorithms.

*Questions, comments for the defense:* There are no correctness questions or issues. My main question is, how far do you think you can push the edit distance techniques? In the thesis you say that 2 seems like a bottleneck for the approximation factor achievable via your techniques. Do you think this is true for all techniques – should we be trying to show that a *2-\epsilon* approximation in *O(n^{2-\delta})* time would contradict some assumption if *\epsilon, \delta>0*, or do you think some other techniques would get us a near-linear time *(1+\epsilon)*-approximation for all *\epsilon>0*?

Another question is, do you think that there can be a strong model of combinatorial Boolean Matrix Multiplication algorithm that would satisfy everyone? That is, it would somehow exclude the computationally costly rank-based methods and would also include all clean practical algorithms? For instance, should such a model consider Strassen's algorithm for matrix multiplication over *GF(2)* combinatorial?

*Conclusion:* In conclusion, this is an outstanding thesis. The candidate has demonstrated her topnotch ability to carry out independent research, and to obtain mathematical breakthroughs where others have failed.

The thesis meets (and exceeds) the usual requirements of a doctoral thesis and I enthusiastically recommend it to be accepted as such

Please do not hesitate to contact me with any questions.

Sincerely,

Virginia V. Williams
Steven and Renee Finn Associate Professor
MIT EECS Dept. and CSAIL