Curriculum vitae

Thomas Dybdahl Ahle

May 2021

Education

June 2019 Doctor of Philosophy, IT University of Copenhagen.

2019 Master of Arts in Computer Science, University of Oxford.

2017 Master of Science, IT University of Copenhagen, University of Copenhagen.

2013 Bachelor of Arts in Computer Science, University of Oxford.

Publications

"Minner: Improved Similarity Estimation and Recall onnbsp;MinHashednbsp;Databases". By Thomas Dybdahl Ahle – Submitted, 2021.

"Tiling with Squares and Packing Dominos in Polynomial Time". By A Aamand, M Abrahamsen, Thomas Dybdahl Ahle, P Rasmussen – Submitted, 2020.

"The Power of Hashing with Mersennenbsp;Primes". By Thomas Dybdahl Ahle, J Knudsen, M Thorup – Submitted, 2021.

"Similarity Search with Tensor Core Units". By Thomas Dybdahl Ahle, F Silvestri at International Conference on Similarity Search and Applications (SISAP), 2020.

"On the Problem of p^1 in Locality-Sensitivenbsp;Hashing". By Thomas Dybdahl Ahle at International Conference on Similarity Search and Applications (SISAP), 2020.

"Subsets and Supermajorities: Optimalnbsp;Hashing-based Setnbsp;Similaritynbsp;Search". By Thomas Dybdahl Ahle, J Knudsen at IEEE Symposium on Foundations of Computer Science (FOCS), 2020.

"Oblivious Sketching of High-Degreenbsp;Polynomialnbsp;Kernels". By Thomas Dybdahl Ahle, M Kapralov, J Knudsen, R Pagh, A Velingker, D Woodruff, A Zandieh at ACM-SIAM Symposium on Discrete Algorithms (SODA), 2020.

"Optimal Las Vegas Localitynbsp;Sensitivenbsp;Datanbsp;Structures". By Thomas Dybdahl Ahle at IEEE Symposium on Foundations of Computer Science (FOCS), 2017.

"Parameter-free Locality-Sensitive Hashing for Sphericalnbsp;Rangenbsp;Reporting". By Thomas Dybdahl Ahle, M Aumüller, R Pagh at ACM-SIAM Symposium on Discrete Algorithms (SODA), 2017.

"On the Complexity of Innernbsp;Productnbsp;Similaritynbsp;Join". By Thomas Dybdahl Ahle, R Pagh, I Razenshteyn, F Silvestri at ACM Symposium on Principles of Database Systems (PODS), 2016.

Awards and Scholarships

Research Travel Award, Stibo-Foundation, 2016.
Given to just two Danish students a year, to collaborate in research abroad.

Northwestern Europe Regional Programming Contest, 1st, Association for Computing Machinery, 2014.
With my team Lambdabamserne, becoming the first ever Danish team to qualify for the ACM wold finals.

Danish National Programming Champion, 1st, Netcompany, 2013, 2014.
Algorithm competition known as "DM i Programmering"

Oxford Computer Science Competition, 1st, University of Oxford, 2013.
For my Numberlink solving software, giving the first fixed parameter polynomial algorithm for the problem.

Demyship, Magdalen College, 2010, 2011.
A historic scholarship awarded to the top students each year.

Les Trophées du Libre, 1st, Free Software Foundation Europe, 2007.
For my work on the PyChess free software chess suite.

Industry and Employment

Chief Machine Learning Officer at SupWiz, 2017 - 2018.
I co-founded an NLP start-up with academics from University of Copenhagen. At SupWiz I lead a team of four in developing our chatbot software and putting it into production at 3 of the largest Danish IT companies. (Now many more.) In 2019 the chatbot won the most prestigious prize given by Innovation Fund Denmark. I was also responsible for our hiring efforts, interviewing dozens and employing 4 engineers over a 5 month period.

Teaching at IT University of Copenhagen, 2015 - 2019.
In 2019 I co-designed and taught the Parallel and Concurrent Programming course to 140 master students. Earlier years I assisted in various algorithms design classes.

Teaching at University of Copenhagen, 2014.
I assisted in teaching algorithms to more than 200 bachelor students.

Open Source Projects

Project Owner at PyChess, 2006 - current.
Developed the most used chess client and engine for the Linux desktop. Currently the 7th most used interface on the Free Internet Chess Server. Translated to more than 35 languages. I lead a team of 4-8 developers and designers. In 2009 we won Les Trophées du Libre in Paris. The project is under the Gnu Public License and has been used by people all over the world for research projects and other experiments.

Media

Jon Lund. "En ulv i fåreklæder", Prosa, May 2021. An interview on the use of SimHash in GoogleFLoC system.

"The Stibo-Foundation supports IT-talents", Stibo, August 2016. The announcement of my winning the Stibo Travel grant.

Bidwell, Jonni. "Python: Sunfish chess engine", Linux Format, January 2016. Article about my Sunfish chess software.

"The National Team at the Programming World Cup", Computerworld, June 2015. Coverage of my teams participation in the ICPC World Finals.

Elkær, Mads. "Denmark's Three Greatest Programmers", Computerworld, October 2013.

Contact

Email: thomas@ahle.dk.

Website: thomasahle.com

DBLP List of papers

Google Scholar List of Papers

Linkedin: linkedin.com/in/thomasahle

Github: github.com/thomasahle

Academic Statement

Thomas Dybdahl Ahle

September 2018

Images, sound or genomic sequences are all objects that modern computers can handle due to one key idea: the idea of high dimensional geometry. Even something as diverse as a tweet can be written as a (long) vector where the ith entry denotes how many copies the tweet contains of the ith word in the English language. Looking at data as points in a geometry allows us to use the concept of distance to represent similarity, and of isoperimetry to understand inherent structure in data domains. Different applications lead to different geometries, but only recently did we fully understand the most basic one – the Euclidean geometry.

In my thesis, I worked on multiple problems related to search and data structures. In Ahle et al. (2016) we showed that there are certain limits to how fast you can search in any geometry. We proved that the existence of a very fast algorithm would allow other too fast algorithms in the fundamental computational area of boolean satisfiability, breaking the so-called Strong Exponential Time Hypothesis. The work was extended in Abboud et al. (2017) starting the field of approximate hardness for polynomial algorithms. Another key idea in high dimensional algorithms is randomization. It turns out that even simple high dimensional problems, like packing spheres tightly, are very hard mathematically. Meanwhile simply placing the spheres at random locations works very well, this has problems related to the predictability and fairness of the computation. Today all the most efficient algorithms in the field have some chance of failing without any chance of verifying their results! In Ahle (2017) I found a way to solve this problem for the two most common geometries, and in Wei (2019) the results were extended to cover the Euclidean case as well. My most recent work includes a unified approach to LSH for distances on set/boolean data and explicit feature embeddings of polynomial kernels, giving new state of the art results in linear methods in machine learning.

Research Plan Next, I will tackle the most important problem in search: Edit Distance. The edit distance between two strings like SIMON and SALOON is the minimum number of insertions, deletions or substitutions required to turn one into the other. In this particular case, the answer is 3: delete an O and change AL into SI. Given a database of strings and a query, we would like to quickly retrieve the most similar string in the database. This problem is deeply linked to natural language processing Sidorov et al. (2015) and computational biology McGrane and Charleston (2016), and as of yet a very big computational mystery.

Practical solutions include transforming strings into sets (using so-called k-mers) in which case my recent work Ahle (2019) gives the state of the art algorithm. Another approach embeds the strings into a geometry called L1 which is related to the Euclidean geometry and allows fast search. Unfortunately, both transformations distort the edit distance a great deal, which makes the approach less than ideal.

I believe that a more direct approach will yield much better algorithms, leading to many breakthrough from machine learning to biology. A very recent development of papers Chakraborty et al. (2018); Haeupler et al. (2019) have introduced many new ideas on edit distance, and Andoni et al. (2018) has shown that all geometric search can be viewed through so-called non-linear spectral cuts, giving us a more principled path for research. I will study these quantities from both an upper and lower bound perspective. For upper bounds, new analytical methods developed in my recent papers allow studying candidate algorithms that this far has been a mystery. For lower bounds, new hypercontractive inequalities I have discovered give a new view at spectral cuts.

References

Abboud, A., Rubinstein, A., and Williams, R. (2017). Distributed pcp theorems for hardness of approximation in p. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 25–36. IEEE.

Ahle, T. D. (2017). Optimal las vegas locality sensitive data structures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 938–949. IEEE.

Ahle, T. D. (2019). Subsets and supermajorities: Unifying hashing-based set similarity search. *arXiv preprint arXiv:1904.04045*.

Ahle, T. D., Pagh, R., Razenshteyn, I., and Silvestri, F. (2016). On the complexity of inner product similarity join. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 151–164. ACM.

Andoni, A., Naor, A., Nikolov, A., Razenshteyn, I., and Waingarten, E. (2018). Data-dependent hashing via nonlinear spectral gaps. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 787–800. ACM.

Chakraborty, D., Das, D., Goldenberg, E., Koucky, M., and Saks, M. (2018). Approximating edit distance within constant factor in truly sub-quadratic time. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 979–990. IEEE.

Haeupler, B., Rubinstein, A., and Shahrasbi, A. (2019). Near-linear time insertion-deletion codes and (1+ ε)-approximating edit distance via indexing. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 697–708. ACM.

McGrane, M. and Charleston, M. A. (2016). Biological network edit distance. *Journal of Computational Biology*, 23(9):776–788.

Sidorov, G., Gómez-Adorno, H., Markov, I., Pinto, D., and Loya, N. (2015). Computing text similarity using tree edit distance. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, pages 1–4. IEEE.

Wei, A. (2019). Optimal las vegas approximate near neighbors in l_p. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1794–1813. SIAM.