Algorithms for Big Data
Project

2024

Project guidelines

Your grade is 60% final written exam and 40% project.

The goal of the project is for you to learn an algorithm not covered in the class and understand its analysis and why it works.

Projects may be completed alone or in a team of at most two people. Both members of the team will get an identical score.

The main deliverable is a 15-minute-long presentation. The date of the presentation will be announced later and will be during the exam session.

You must submit the slides in advance. The due date of the slide will be posted on UV. You will lose 1% per hour of lateness. Project presentations will be during the exam period. Sign up for your presentation time using the link on UV.

The algorithm

You must choose an paper that is

- Relevant to the field of big data.
- Published in a reputable conference or journal in the past 10 years. Examples of reputable conferences where algorithms results appear are can be found here, and examples of non-reputable publishers can be found here. SOSA and ESA Track-S typically have well-written papers that are interesting but not too difficult. There are many garbage conferences which publish anything, part of the reason I ask for approval of your selection is to make sure you pick a paper with real results. Note that in theoretical computer

science, conference publication is typically the main way one gains recognition for a result. Recent papers are also posted on arXiv for convenience, but arXiv has almost no editorial filters. After appearing at a conference a paper may also appear in a journal, but the absence of this step says nothing about the paper's quality. DBLP is the main index of articles in computer science, but it is no guarantee of quality.

- The algorithm must have a rigorous analysis. Typically this would be a proof about its runtime or other measure of its performance. Note that an implementation and plots do not constitute a rigorous analysis! If your paper does not prove in an interesting way something about an algorithm, pick a different paper!
- You have freedom of choice, you can pick something more general or some algorithm developed for a particular domain where big data is prevalent, e.g. bioinformatics. Just be careful to make sure there is a real algorithm with a theoretical analysis, not just heuristics and plots. A bad choice would be a machine learning algorithm, which has some plots showing it works better on real data than other algorithms, but without any formal analysis or proofs.
- I should not be an author of the paper you have chosen.
- If a paper has multiple aspects to it, such as multiple algorithms, you need not present everything. It is better to present something very clearly.
- No duplicates are allowed. Whoever asks me first has the right to submit the project on that paper. You can not choose something that was chosen last year (I will let you know if this happens). Duplicate projects who did not get premission will get a zero score.

Important: You must get approval from me for your paper. Send me a message on teams with the paper you wish to present (or a link if there is a publicly-available PDF). Once approved please add it here.

The presentation

If you do the project in a pair, you must both present at the same time.

You should give with your partner a 15 minute long presentation where you teach me the algorithm and its analysis. In this presentation, you should explain the problem, indicate the model (e.g. RAM, streaming, cache-oblivious, etc), the algorithm, what is the theoretical performance of the algorithm, what was known before this algorithm. It is very crucial that you explain very clearly what problem the algorithm solves. You should usually show how the algorithm works on an example. The goal is that after your presentation, I should understand the algorithm and its analysis.

You should avoid copying verbatim things from the paper. If you do so, please reference it. If you have an example, it should be different from the paper. If you are simply repeating the paper verbatim, I am likely to ask you detailed questions to show that you truly understand what you have copied.

There is no need to present any experimental results found in the paper, and this is just a waste of your time that you could use to better explain the algorithm. Focus on the algorithm and its theoretical analysis.

To hand in

- Email to bigdata24@johniacono.com
- Subject: "Project submission of XXX (and XXX)"
- CC your partner, if a pair
- Attach a pdf of the paper
- Your presentation, in pdf format.
- Any code or other things you would like to share.
- I will confirm submission via email if you don't get a confirmation in a few hours during the working day, contact me via teams.

Questions

I will, of course ask you questions after your presentation. Allow 30 minutes total.

Grading

Your grade will be based on

- The difficulty of the algorithm presented.
- Your ability to explain
  - What problem is being solved
  - How the algorithm works

- Why it works
- Why it has any claimed performance metrics (e.g. running time, competitive ratio, space, etc., as appropriate).
- Why the algorithm is better than previous algorithms for the same problem.

Second session

The project for the second session is due at a time that will be announced on Teams. Presentations will be scheduled during the exam session. If you completed the project for the first session and did not pass the course, you can either re-do the project or keep your project grade from the first session. If you re-do the project you need to pick a new paper.

Academic honesty

All work should be your own or referenced. Reference anything that is not yours. If you use an figure in your presentation that you did not make, it must be referenced. Any breach of academic honestly will result in a zero grade for the project.