Research Proposal Innovation Case - Research on Data Streaming Algorithms for Bipartite Matching

Ivan Qin wn18676

March 11, 2022

Executive Summary

Since at least two decades, Big Data technology has impacted our daily lives and it requires a new generation of algorithms. Among all researches on algorithms, the Data Streaming algorithm has its own advantages of single pass and only sublinear space complexity, meaning that it has highly relevant for rapid and large data analysis, as well as the great potential for future development.

Matchings in large bipartite graphs are one of the key problems in the Steaming algorithm. Because it can handle a variety of problems that today's Internet needs to be optimized, such as social relationship graphs, order matching, search engines, distributed computing, deep learning, etc. There are many theoretically feasible matching algorithms in academia, but once applied in engineering, they are very difficult to implement.

In this project, we will study the difference in the performance of the algorithm on the real model, and we aim to implement and compare recent Data Streaming algorithms for computing large matchings in massive bipartite graphs. Our focus will lie on a recently published algorithm named Auction Algorithm [2], which makes multiple passes over the input graph and iteratively improves the matching in each pass.

If the project is successfully completed, we will make the following contributions to academia and industry:

- The research and the implementation method on Auction Algorithm.
- A certain number of influential theoretic research papers on the difference between the traditional dataset and the real-world dataset.
- Publish code in the Boost graph library [1], thus more and more algorithm researchers can study and contribute with these open-source Data Streaming algorithms.
- Big companies such as Google, Twitter, Uber, etc. will hugely benefit from these open source tools, and they are able to improve their commercial products.

Evaluation of Potential Opportunity

Data has played a central role in our daily life, and algorithms have been helping us solve the problem of data explosion. For example, Social networks like Twitter and Facebook are made up of billions of monthly active users who are interconnected by trillions of friendships; Uber matches orders with the delivery man, the instantaneous amount of data for this problem is huge, often exceeding 10,000 requests per second; Google's search engine indexes billions of web pages every day and answers more than 100,000 search queries per second on average [11], which amounts to more than a trillion queries per year[10].

With the rapid development of machine learning and data mining, we urgently need to build new algorithms for handling such huge information, especially the algorithm working on dynamic

data, for instance, Data Streaming algorithms. As its name implies, it processes its input in streams, in the other words, sequentially in a single pass or few passes while maintaining a small summary that is sublinear in the input size, thereby addressing the fundamental problem that the RAM of modern computers is much smaller (e.g., a few Gigabytes) than the size of modern massive datasets (Terabytes or even Petabytes).

This research project addresses the processing of massive graphs, such as social network graphs, graph databases, the Internet graph, telephone graphs, models of the brain, and many others. Data Streaming algorithms have been developed for various important graph problems, however, almost all of these algorithms consider static settings, where the underlying data is not subject to change. This assumption, however, contradicts the very nature of real-world massive data sets, which are typically inherently dynamic: For example, social network graphs change structure when new friendships are established or existing ones are ended, and the Internet graph evolves with the creation and deletion of web pages and hyperlinks. In machine learning, additional data suited for learning may become available, and algorithms need to be able to adapt to such insertion-deletion situations on the fly.

In the literature review, the insertion model (static data model) is well known. [3] To cope with evolving data and design insertion-deletion streaming algorithms, one of the ideas is called the Auction Algorithm, which is recently published in the year 2021 [2], it combines the idea from economics. This algorithm is theoretically able to solve approximation streaming matching problem in a deterministic $O(1/\epsilon^2)$-pass $O(n)$-space algorithm in the graph streaming model; this improves the pass complexity of the state-of-the-art algorithms by $O(loglog(1/\epsilon))$ and the space complexity by $O(1/\epsilon)$ to achieve optimal space bounds with no dependence on ϵ.

The Auction Algorithm is a good entry point, with its efficiency, We can discover the feasibility of using the algorithm on real-world data, which will help us to study the problems from the insertion-deletion model.

In the Boost graph library[1], popular research and engineering algorithm library in C++, the only built-in algorithm for bipartite matching is the greedy method. However, this traditional greedy algorithm cannot deal with the large dataset in engineering, and once we publish the code of the Data Streaming algorithm on this library, will greatly promote the research progress in this field.

In general, streaming algorithm research is getting more and more important. It is necessary to study the streaming algorithm in the insertion-deletion model, which can perform better in dynamic data, but the work needs to start now to ensure wide adoption.

Value Proposition

The matching algorithm for data streaming was proposed 15 years ago, so far it is not suitable for large dynamic graphs. In the Auction Algorithm paper, the author believes that this technique gives a powerful tool for boosting the approximation ratio of algorithms for the bipartite matching problem from the 2-approximation of maximal matching to $(1-\epsilon)$-approximation in different settings.

However, the author also proposes that the engineering implementation of other auction algorithms will be different, and this algorithm is only a starting point for the study of dynamic graphs. With his idea, we will complete and optimize the algorithm, for example, by modifying the data structure, theoretically analyzing the worst case and trying to avoid it, etc.

In addition, to expand the auction algorithm, we will also explore the similarities and differences between the current ordinary algorithm and the optimized algorithm, so as to design a more efficient algorithm.

Hence, we assume the project is successful, the most direct beneficiaries are academia and industry.

In academia, the data streaming problem will innovate new ideas, and at the same time, there will be progress on distributed computing, machine learning, and other issues.

The value in the industrial is even more pronounced. We will publicly release our optimized algorithm, plenty of data-driven companies like Google can match users with search results more quickly and accurately. The current waiting time for Uber in urban areas is 11 minutes, and once we have our algorithm, a more accurate and fast matching method will bring benefits to users and companies.

As previously noted, the streaming algorithm is only expected to become more integral to our lives in the future. In algorithm engineering, the tiny optimization of this new algorithm increases the efficiency of the matching algorithm, the efficiency of the company's products increases, while storage expenditures will decrease. This fact is good for both the user and the company. Even if the study does not yield general results, the data and research generated during the study will help future projects or research in the field.

Impact Plan

This research will be carried out in the Algorithms and Complexity group at the University of Bristol, [6] under the supervision of Dr Christian Konrad. Dr Konrad's group has substantial knowledge and experience in the area of Data Streaming algorithms. We will build on their expertise, which will allow us to conduct cutting edge research in this area. The objectives of this research proposal build on findings already obtained by Dr Konrad's team's work on their current EPSRC grant "StreamDG: Streaming Algorithms for Massive Dynamic Graphs" [4], their project has received £250,000 of EPSRC funding, and the Engineering and Physical Sciences Research Council (EPSRC) is the main funding body for engineering and physical sciences research in the UK. [12]

The duration of this project is half a year as a master's graduation project, it is a good initial project for algorithm research. It can also be used as a PhD for 4 years of research. Depends on how deep the experiment wants to analyze the streaming algorithm. Based on the data from UKRI [12], it suggests that our project budget £100,000-£250,000 of EPSRC funding for 4 years.

The budget will mainly be used to pay for researchers' salaries, their computer laptop equipment, travel fees for top conferences, and real-world data collection. Since it is research and engineering research, we aim to find four PhD students, one post-doc, and many companies need data streaming algorithms. We cooperate with the company, then they could provide the real-world data, we can provide the algorithm we designed to help these companies adapt faster and earlier to the company's special demand.

To make sure the project is done as flawlessly as possible, I have divided the whole project into 4 phases.

1. **Literature review.** In the first year, we should work on Auction Algorithms analysis. Also, we will analyze and study graph streaming algorithms for Maximum Matching, which includes 1-pass, 2-pass and multi-pass algorithms for matching problems, and many related state-of-art algorithms [9] [8] [5]. This study will require an in-depth study that will span several months. We will particularly focus two-part:

   (a) The differences and similarities between the insertion-only and insertion-deletion models.

   (b) The implementation of insertion-deletion models of real-world data.

   (c) Ideas from other subject areas, such as auction matching algorithms, combine ideas from economics and computer science.

   At this stage, we will have established the current research boundary, and we will target the research questions mentioned above. This will require designing new algorithms and proving lower bounds (impossibility results):

   (a) **Algorithmic Techniques**. A prior all techniques relevant to algorithms design (e.g., Greedy algorithms, divide-and-conquer, charging schemes, dynamic programming, data

structure) may become relevant to this project. However, most insertion-deletion streaming algorithms rely on techniques that are based on linear sketches, such as l_0-sampling and sparse recovery.

(b) **Lower Bound Techniques.** Proving impossibility results is of particular importance to this project since such results have the potential to establish that known algorithms are the best possible, which ultimately concludes a direction of research. Regarding streaming algorithms, the predominant technique is communication complexity, in particular, the one-way two-party communication model [7] for insertion-only streams, and the simultaneous communication model for insertion-deletion streams.

2. **Academic exchanging and Results Sharing.** In the second year, we should be able to provide at least 3 paper, and these papers should occupy a central position in the algorithm research community and has countless impacts on many areas including distributed computing, online algorithms, communication complexity, property testing, dynamic algorithms, and algorithms for temporal graphs. More often than not, scientific advancements in one of these areas have led to advancements in the others. Our work will thus directly benefit researchers working in all these areas. We will present our research findings in top algorithm conferences (and subsequently in top journals) that have high visibility, such as ICALP, ESA, SODA, and STACS. We will also publish a survey paper on dynamic graph stream processing, potentially in the ACM Computing Surveys journal or as an ACM SIGMOD Record Newsletter.

3. **Simulations and Experiments.** In the third year, we will implement the algorithms designed in the first and second phase, for example, the Auction Algorithm, and compare our new algorithms to the state-of-the-art. Streaming algorithms are efficient by construction and therefore have the potential to outperform non-streaming approaches. Through the current research and experiment, we can roughly see that the Auction Algorithm has better performance in approximate matching problems, such as less space, faster speed, etc. At this stage, we will publish our new streaming algorithms to boost graph library. For algorithm engineering or research, the graph boost library [1] in C++ is by far the most popular. This provides the largest influence beyond academia because it allows any programmer to use the graph library to provide fast, convenient, visual, easy-to-use, and powerful tools for working.

4. **Deployed in Industrial.** In the final year, we are aiming to let our algorithm be deployed into the real-world dataset. For those commercial companies such as Google, Uber, Twitter that need to process large datasets, they can use our algorithms we developed to increase efficiency and reduce costs for themselves. More precisely, we communicated with Uber's employees in our preliminary investigation, we unanimously estimated that this algorithm will greatly reduce the time it takes for taxi riders to get a ride, ensuring that the customer only needs to wait for less than 1 minute to get on the taxi anywhere in any urban area, including residential areas. So, it can give the company at least one million benefits, which will impact their product efficiency. We will provide the support to deploy the algorithm, this cooperation will help us work back on designing a better algorithm.

Through the implementation of this project, we will finally have new streaming algorithms that go beyond ordinary streaming algorithms. In addition to having impacts on academia, We also provide engineering assistance to industrial. Therefore, our research has far-reaching implications.

References

[1] www.boost.org. (n.d.). *The Boost Graph Library - 1.78.0.* URL: https://www.boost.org/doc/libs/1_78_0/libs/graph/doc/index.html. (accessed: 05.03.2022).

[2] Assadi S., Liu S.C. and Tarjan R.E. "An Auction Algorithm for Bipartite Matching in Streaming and Massively Parallel Computation Models." In: *Symposium on Simplicity in Algorithms (SOSA), pp.165–171.* (2021).

[3] Eggert, S., Kliemann, L., Munstermann, P. et al. "Bipartite Matching in the Semi-streaming Model." In: *Algorithmica 63, 490–508 (2012).* (2012).

[4] epsrc. *gow.epsrc.ukri.org.* URL: https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V010611/1. (accessed: 10.03.2022).

[5] Fr´ed´eric Magniez, Christian Konrad. and Claire Mathieu. "Maximum matching in semi-streaming with few passes." In: *In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques- 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012.* (2012).

[6] Bristol algorithm group. *Algorithms and Complexity group in Bristol.* URL: https://bristolalgo.github.io/. (accessed: 10.03.2022).

[7] Christian Konard. *one-way two-party communication model.* URL: http://people.cs.bris.ac.uk/~konrad/courses/2020_2021_COMSM0068/slides/15-lower-bounds-1.pdf. (accessed: 09.03.2022).

[8] Christian Konrad. "A simple augmentation method for matchings with applications to streaming algorithms." In: *In 43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August* (2018).

[9] Christian Konrad. "A simple augmentation method for matchings with applications to streaming algorithms.." In: *In 43rd International Symposium on Mathematical Foundations of Computer Science, MFCS 2018, August* (2018).

[10] mitre. *digital-defense-acquisition-2024.* URL: https://aida.mitre.org/blog/2019/07/16/digital-defense-acquisition-2024/. (accessed: 09.03.2022).

[11] Internet live stats. *Google Search Statistics - Internet Live Stats.* URL: https://www.internetlivestats.com/google-search-statistics/. (accessed: 05.03.2022).

[12] epsrc ukri. *Home - epsrc website 2022.* URL: https://epsrc.ukri.org/. (accessed: 10.03.2022).