COMBINATORIAL OPTIMIZATION ON MASSIVE DATASETS:
STREAMING, DISTRIBUTED, AND MASSIVELY PARALLEL COMPUTATION

Sepehr Assadi

1 Introduction

With the emergence of massive datasets across different application domains, there is a rapidly growing interest in solving various optimization problems over immense amounts of data. To cope with the sheer size of these big data problems, one needs to design algorithms that are highly efficient in their resource usage: for example, their internal memory in case of streaming algorithms, their communication overhead in case of distributed algorithms, and both memory and communication, as well as as rounds of computation for massively parallel algorithms. This in turn raises the following fundamental question:

> *How well can we solve a large-scale optimization problem on massive datasets in a resource-efficient manner?*

The research presented in this thesis stems from pursuing the above question for two general family of combinatorial optimization problems, namely graph optimization and submodular optimization, in several computational models for processing massive datasets, in particular, streaming, distributed, and massively parallel computation (MPC) models.

A common theme in this thesis is a rigorous theoretical study of problems from both algorithmic perspective and impossibility results. Rather than coming up with an ad hoc solution to each problem at hand, the goal here is to develop general techniques for solving large-scale optimization problems in a unified way in different computational models. This can only be achieved by understanding the powers and limitations of current algorithmic approaches for solving these problems to know when and why these techniques fail and how a new approach can circumvent such limitations.

1.1 Main Contributions

We describe new techniques for developing algorithms and impossibility results for graph optimization and submodular optimization problems on massive datasets, with the following broad consequences:

- the first non-trivial impossibility result for a graph optimization problem in dynamic streams that is tractable in insertion-only streams;
- a general algorithmic approach for graph optimization applicable to all three models discussed above that bypasses the impossibility results for prior popular techniques such as linear sketching;
- a new framework for proving impossibility results for multi-round distributed algorithms and multi-pass dynamic streaming algorithms;
- a host of improved MPC algorithms for fundamental graph problems that improve upon the classical parallel PRAM algorithms even when using a very limited memory per-machine.

Using these techniques, we analyze a variety of central optimization problems in modern computational models for processing massive datasets. In the following, we give a high level overview of these results.

Graph Optimization. Massive graphs abound. As such, there has been an extensive interest in studying graph optimization problems in different models of computation over massive datasets. This thesis contributes to this area by presenting new algorithms and lower bounds for fundamental graph problems of maximum matching, minimum vertex cover, and undirected connectivity in the models discussed above.

Our first main result is a tight bound on space needed for approximating maximum matching in the dynamic streaming model as well as a host of new lower bounds for estimating matching size in insertion-only and dynamic streams. Our results in this part constitute the first non-trivial dynamic streaming lower bound for any graph problem which is tailored to dynamic streams. We next use this result to demonstrate the inefficiency of popular algorithmic approaches such as linear sketching for distributed and MPC computation of matchings and vertex covers. By building on the insights from these impossibility results, we further develop a new approach to design efficient algorithms for graph problems in a unified way in streaming, distributed, and MPC models and use this framework to design algorithms for matching and vertex cover that improve upon the state-of-the-art in all these models simultaneously in some or all parameters involved. We also give the first massively parallel algorithm for connectivity on sparse graphs that improve upon the classical PRAM algorithms for graphs with even a modest expansion.

The materials in this part of the thesis are based on [13] (SODA'16), [12] (SODA'17; invited to Highlights of Algorithms conference HALG'17), [9] (SPAA'17; co-winner of the best paper award and invited to Highlights of Algorithms conference HALG'18), [6] (SODA'19), and [4, 15].

Submodular Optimization. Submodular optimization encompasses a wide variety of problems in combinatorial optimization including set cover, maximum coverage, and minimum/maximum cut, to name a few. These problems have various applications in different areas including in machine learning, operations research, information retrieval, data mining, and web host analysis. As a result, submodular optimization has received quite a lot of attention in models of computation on massive datasets.

The first main contribution of this thesis on this front is proving tight upper and lower bounds on space complexity of set cover in both single-pass and multi-pass streams. We further demonstrate the impossibility of achieving efficient algorithms for maximum coverage and submodular maximization in a small number of passes in dynamic streams, achieving the first multi-pass dynamic streaming lower bound for any optimization problem. In the distributed model, we provide tight bounds on the number of rounds of parallel computation needed to solve maximum coverage and in general submodular maximization. These results are obtained by introducing a new framework for proving lower bounds for multi-round distributed algorithms and multi-pass dynamic streaming algorithms that also unifies many of previous ad hoc lower bounds in this area into a single approach. Incidentally, the insights behind these lower bounds also led to extremely simple distributed and massively parallel algorithms for submodular maximization with performance guarantee that matches the state-of-the-art.

The materials in this part are based on [11] (STOC'16; invited to SICOMP special issue), [5] (PODS'17; winner of the best student paper award), and [10] (SODA'18).

Resource-Constrained Optimization. Beside large-scale optimization on massive datasets, another related topic addressed in this thesis is optimization in settings which require computation over data which is not particularly large but still imposes restrictions of similar nature. Examples include optimization over data which can only be accessed with limited adaptivity (e.g. in crowdsourcing applications), or corresponds to private information of agents and is not available in an integrated form (e.g., in auction and mechanism design applications).

We use the toolkit developed in the first two parts of the thesis to obtain several improved results for such problems. In particular, this thesis sheds light on the role of interaction between agents participating in a combinatorial auction by determining the exact tradeoff between number of rounds of interaction and welfare of the returned allocation in these markets. Another example is identifying the degree of adaptivity needed for solving several fundamental learning tasks such as exploration in multi-armed bandits or ranking from pairwise comparisons. These results are obtained by exploiting the (loose) connection between these problems and distributed computing and applying the ideas from our framework introduced in the previous

part for proving multi-round distributed algorithms to these problems.

The materials in this part are based on [3] (EC'17; invited to TEAC special issue), and [1] (COLT'17).

2 Graph Optimization

The first contribution of the thesis to this area is to settle the space complexity of approximating matchings in streams that allow both insertion and deletion of edges, a.k.a dynamic streams. Recent years have witnessed a tremendous success in designing dynamic streaming algorithms for nearly all graph problems that have been studied previously only in insertion-only streams including, connectivity, minimum spanning tree, cut and spectral sparsifiers, subgraph counting, and numerous others. However, the prominent problem of maximum matching was notably absent from this list. Indeed, settling the space complexity of approximating matchings in dynamic streams featured prominently in the "List of Open Problems in Sublinear Algorithms" [17]. We resolve this question completely in this thesis and as a corollary obtain that even a very weak approximation ratio of $n^{o(1)}$ for matchings in dynamic streams requires $n^{2-o(1)}$ space, ruling out even mildly efficient algorithms for this problem (here n is the number of vertices in the graph).

We also consider the (algorithmically easier) problem of estimating the size of a maximum matching (as opposed to finding the actual edges) in both insertion-only and dynamic streams. We showed that while this problem provably requires less space than finding approximate matchings in both models, achieving a near optimal solution, i.e., a $(1+\varepsilon)$-approximation, still requires (almost) quadratic in n space. As a corollary of this result, we also obtain that estimating the rank of a given matrix in data streams requires (almost) quadratic space. Our results constitute the first super-linear space (in number of vertices of the graph or rows of the matrix) and the first super-constant approximation lower bounds for both problems, addressing open questions by Li and Woodruff [22] and Kapralov *et al.* [21], respectively. Our proofs in this part rely on different techniques from additive combinatorics (in particular Ruzsa-Szemerédi graphs), Fourier analysis, and information complexity to break the linear-space and constant-approximation barriers in previous work. Our techniques here were also used in the subsequent work of Balcan *et al.* [16] to settle the sketching complexity of property testing of matrix rank.

Our results in the first part can also be used to rule out applicability of existing general-purpose algorithmic approaches for designing streaming, distributed, and MPC algorithms, such as linear sketching, for solving matching and the vertex cover problems. The main insight we borrow from these results is that the intractability of matching and vertex cover is inherently connected to the adversarial ordering/partitioning of the underlying graph in these models. Building on this, we further propose a general approach based on sparsification that can achieve significantly better algorithms for these problems under the assumption that the input is randomly ordered/partitioned. We then use this approach to obtain several improved algorithms for matching and vertex cover in streaming, distributed, and MPC model. This includes a single-pass streaming algorithm for the maximum matching problem in random arrival stream that achieves a $(3/2+\varepsilon)$-approximation using $O(n\sqrt{n})$ space, and a $(3/2+\varepsilon)$-approximation MPC algorithm with only two rounds of computation even on adversarially partitioned inputs. By further extending our approach, we obtain $O(1)$-approximation algorithms for matching and vertex cover in the MPC model that only require $O(\log\log n)$ rounds of computation even when the memory per machine is as small as $O(n/\text{polylog}(n))$. This improves the round complexity of previous algorithms in the literature by a quadratic factor for matching and an exponential factor for vertex cover, and settles multiple open questions in MPC literature (cf. [18]). This high level approach of sparsification has since been used in several papers of the author to address related problems as well, in particular, in [8] for sublinear algorithms for graph coloring, [14] for dynamic algorithms for maximal independent set, and [7] for a general sparsification method for matching problem.

The final contribution of this thesis to this area is an MPC algorithm for connectivity. A captivating algorithmic question in the MPC literature is whether connectivity on sparse undirected graphs can be solved faster in the MPC model with low memory per machine compared to (much weaker) PRAM algorithms. Despite the great attention this problem has received in the last few years, the answer to this question has remained elusive. This thesis makes progress on this question by showing that for a large family of graphs that appear frequently in practice, namely graphs with moderate expansion or more formally spectral gap

of at most $n^{o(1)}$, one can solve connectivity with n^{δ} memory per-machine for any constant $\delta > 0$, in $o(\log n)$ rounds. This constitutes the first improvement on the standard $O(\log n)$-round classical PRAM and MPC algorithms for connectivity when the memory per machine is $n^{\Omega(1)}$ for a general family of input graphs. We also show that when the memory per-machine is slightly sublinear in n, i.e., is $O(n/\text{polylog}(n))$, then this problem can be solved in $O(\log \log n)$ MPC rounds on any graph (regardless of its spectral gap).

3 Submodular Optimization

The second part of this thesis considers submodular optimization and in particular two of its canonical examples, minimum set cover and maximum coverage, on massive datasets. One of the main contributions of this thesis to this area is to fully settle the space complexity of approximating the set cover problem in both single-pass and multi-pass streams. Prior to this thesis, several streaming algorithms were known for the set cover problem that achieve a small approximation ratio of logarithmic, or even a constant in absence of computation time restrictions, using a sublinear space. However, all these algorithms required at least two and typically a much larger number of passes over the stream. As a result, existence of sublinear space algorithms with small approximation ratio in one pass over the stream was still wide open (cf. [20]). This thesis presents a strong negative resolution of this question even for the algorithmically easier problem of estimating the size of an optimal set cover (as opposed to finding the sets) and even on random arrival streams. More formally, we prove that for the set cover problem with m sets and n elements, any single-pass streaming algorithm that outputs an α-approximation to the value of the optimal set cover size requires $\Omega(mn/\alpha^2)$ space. This implies that any truly sublinear space single-pass streaming algorithm for this problem requires approximation ratio of at least n^{ε} for some constant $\varepsilon > 0$.

Furthermore, in this thesis, we present a (slightly) improved multi-pass streaming algorithm for set cover and more importantly prove that the space-approximation tradeoff achieved by this algorithm is optimal. Formally, we prove a tight bound of $\Theta(mn^{1/\alpha})$ on the space complexity of α-approximation multi-pass streaming algorithms for set cover. Both our single-pass and multi-pass lower bounds for set cover are based on proving new direct sum results for set cover using information complexity. These results fully settle the space-approximating tradeoff for set cover in both single-pass and multi-pass streams. Qualitatively similar results for the maximum coverage problem are also presented in this thesis.

We also consider maximum coverage and in general submodular maximization subject to cardinality constraint in the distributed model. Our primary goal here is to understand how many rounds of parallel computations are needed to solve these problems efficiently via distributed algorithms. This question is highly motivated by application to big data analysis such as MapReduce computation as the number of rounds of the computation is the key contributing factor to the overall computation time of the algorithms. We present a tight tradeoff between the three main measures of efficiency in this model: the approximation ratio, the communication cost, and the number of rounds for any algorithm for these problems. To achieve this result, we provide a general framework for proving communication lower bounds for bounded-round protocols, and use this framework to prove our main lower bound for maximum coverage and submodular maximization. This framework also allows us to unify several previous lower bounds in the literature proven using different ad-hoc arguments into a single proof. As a corollary of this result, we also obtain lower bounds on the number of passes needed to solve these two problems in dynamic streams. This is the first multi-pass lower bound for any optimization problem that is specific to dynamic streams, i.e., does not follow from a lower bound in insertion-only streams. Interestingly, these impossibility results also guided us to develop a very simple distributed algorithm for submodular maximization subject to cardinality constraint (which is also implementable in the MPC model) with a performance guarantee that matches the state-of-the-art.

4 Resource Constrained Optimization

We further consider optimization in settings where the goal is to perform computation over data which may not be particularly large but still imposes restrictions of similar nature, the setting which we refer to as resource constrained optimization. We use the toolkit developed in the first two parts of the thesis to obtain several algorithms and impossibility results for problems of this nature, settling multiple open

questions in the literature.

As the first example, we consider the necessity of interaction between n bidders with subadditive valuations in a combinatorial auctions over m items for maximizing the social welfare. The goal is to understand how much interaction, in terms of number of rounds communication between bidders and the central planner, is needed to achieve an almost efficient allocation of items. This question was posed originally by Dobzinski, Nisan, and Oren [19] and subsequently by Alon, Nisan, Raz, and Weinstein [2]. We resolve this fascinating question by providing an almost tight round-approximation tradeoff for this problem, when the players are communicating only polynomially many bits (in n and m). As a corollary, we prove that $\Omega(\frac{\log m}{\log \log m})$ rounds of interaction are necessary for obtaining any efficient allocation with a constant or even a polylog(m, n)-approximation in these markets, matching the upper bound of [19] up to lower order term. Our proof of this result builds on the similarity between this problem and the distributed communication model and our lower bound framework in the previous part.

We also consider resource constrained optimization in learning scenarios. In many learning settings, active/adaptive querying is possible, but the number of rounds of adaptivity—the number of rounds of interaction with the feedback generation mechanism—is limited. For example, in crowdsourcing, one can actively request feedback by sending queries to the crowd, but there is typically a waiting time before queries are answered; if the overall task is to be completed within a certain time frame, this effectively limits the number of rounds of interaction.

We study the relationship between query complexity and adaptivity in identifying the k most biased coins among a set of n coins with unknown biases. This problem is a common abstraction of many well-studied problems, including the problem of identifying the k best arms in a stochastic multi-armed bandit, and the problem of top-k ranking from pairwise comparisons. Our main result establishes an optimal lower bound on the number of rounds adaptivity needed to achieve the optimal worst case query complexity for all these problems. In particular, we show that, perhaps surprisingly, no constant number of rounds suffices for this task, and the "correct" number of rounds of adaptivity is in fact $\log^*(n)$. The proof of this result also borrows ideas from our lower bound framework for bounded-round distributed algorithms, using the loose connection between the number of rounds of adaptive querying vs rounds of distributed communication.

References

[1] A. Agarwal, S. Agarwal, S. Assadi, and S. Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 39–75, 2017.

[2] N. Alon, N. Nisan, R. Raz, and O. Weinstein. Welfare maximization with limited interaction. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1499–1512, 2015.

[3] S. Assadi. Combinatorial auctions do need modest interaction. In *Proceedings of the 2017 ACM Conference on Economics and Computation, EC '17, Cambridge, MA, USA, June 26-30, 2017*, pages 145–162, 2017.

[4] S. Assadi. Simple round compression for parallel vertex cover. *CoRR*, abs/1709.04599, 2017.

[5] S. Assadi. Tight space-approximation tradeoff for the multi-pass streaming set cover problem. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 321–335, 2017.

[6] S. Assadi, M. Bateni, A. Bernstein, V. S. Mirrokni, and C. Stein. Coresets meet EDCS: algorithms for matching and vertex cover on massive graphs. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, 2019.

[7] S. Assadi and A. Bernstein. Towards a unified theory of sparsification for matching problems. In *2nd Symposium on Simplicity in Algorithms, SOSA 2019*, 2019.

[8] S. Assadi, Y. Chen, and S. Khanna. Sublinear algorithms for $(\Delta + 1)$ vertex coloring. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, 2019.

[9] S. Assadi and S. Khanna. Randomized composable coresets for matching and vertex cover. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA 2017, Washington DC, USA, July 24-26, 2017*, pages 3–12, 2017.

[10] S. Assadi and S. Khanna. Tight bounds on the round complexity of the distributed maximum coverage problem. In *Proceedings of the Twenty-Nine Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, 2018.

[11] S. Assadi, S. Khanna, and Y. Li. Tight bounds for single-pass streaming complexity of the set cover problem. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 698–711, 2016.

[12] S. Assadi, S. Khanna, and Y. Li. On estimating maximum matching size in graph streams. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1723–1742, 2017.

[13] S. Assadi, S. Khanna, Y. Li, and G. Yaroslavtsev. Maximum matchings in dynamic graph streams and the simultaneous communication model. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1345–1364, 2016.

[14] S. Assadi, K. Onak, B. Schieber, and S. Solomon. Fully dynamic maximal independent set with sublinear in n update time. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, 2019.

[15] S. Assadi, X. Sun, and O. Weinstein. Massively parallel algorithms for finding well-connected components in sparse graphs. *CoRR*, abs/1805.02974, 2018.

[16] M. Balcan, Y. Li, D. P. Woodruff, and H. Zhang. Testing matrix rank, optimally. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, 2019.

[17] Bertinoro workshop 2014, problem 64. http://sublinear.info/index.php?title=Open_Problems:64. Accessed: 2018-4-4.

[18] A. Czumaj, J. Lacki, A. Madry, S. Mitrovic, K. Onak, and P. Sankowski. Round compression for parallel matching algorithms. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, June 25-29, 2018*, pages 471–484, 2018.

[19] S. Dobzinski, N. Nisan, and S. Oren. Economic efficiency requires interaction. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 233–242, 2014.

[20] S. Har-Peled, P. Indyk, S. Mahabadi, and A. Vakilian. Towards tight bounds for the streaming set cover problem. In *Proceedings of the 35th Symposium on Principles of Database Systems, PODS 2016*.

[21] M. Kapralov, S. Khanna, and M. Sudan. Approximating matching size from random streams. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 734–751, 2014.

[22] Y. Li and D. P. Woodruff. On approximating functions of the singular values in a stream. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 726–739, 2016.