Brief Announcement: Scalable Diversity Maximization via Small-size Composable Core-sets*

Alessandro Epasto†
Google
New York, NY
aepasto@google.com

Vahab Mirrokni
Google
New York, NY
mirrokni@google.com

Morteza Zadimoghaddam
Google
Zurich, Switzerland
zadim@google.com

ABSTRACT

In this paper, we study the diversity maximization problem (a.k.a. maximum dispersion problem) in which given a set of n objects in a metric space, one wants to find a subset of k distinct objects with the maximum sum of pairwise distances. We address this problem using the distributed framework known as *randomized composable core-sets [3]*. Unlike previous work, we study *small-size* core-set algorithms allowing minimum possible intermediate output size (and hence achieving large speed-up in the computation and increased parallelism), and at the same time, improving significantly over the approximation guarantees of state-of-the-art core-set-based algorithms. In particular, we present a simple distributed algorithm that achieves an almost optimal communication complexity, and asymptotically achieves approximation factor of 1/2, matching the best known global approximation factor for this problem. Our algorithms are scalable and practical as shown by our extensive empirical evaluation with large datasets and they can be easily used in the major distributed computing systems like MapReduce. Furthermore, we show empirically that, in real-life instances, using small-size core-set algorithms allows speed-ups up to > 68 in running time w.r.t. to large-size core-sets while achieving close-to-optimal solutions with approximation factor of > 90%.

ACM Reference Format:
Alessandro Epasto, Vahab Mirrokni, and Morteza Zadimoghaddam. 2019. Brief Announcement: Scalable Diversity Maximization via, Small-size Composable Core-sets. In *Proceedings of Great Lakes Symposium on VLSI 2019 (SPAA '19).* ACM, New York, NY, USA, 2 pages. https://doi.org/https://doi.org/10.1145/3323165.3323172

1 INTRODUCTION

Computing a concise, descriptive, but yet *a diverse*, summary of a dataset is a central problem in machine learning, data mining and information retrieval. In all of these scenarios, the goal is to design efficient methods for searching and summarizing large data sets in a way that preserves the diversity of the data. In one of the most popular forms, also referred to as the *max. dispersion* or the *remote-clique diversity* maximization, the problem is as follows: given a set of n objects in a metric space, one wants to find a subset of k distinct objects with the maximum sum of pairwise distances.

In order to solve such diversity maximization problems for increasingly large data sets in many applications [2], it is desirable to find a scalable distributed algorithm. To achieve this goal, a recent line of research is to apply the distributed framework known as *composable core-sets* [3].

1.1 Preliminaries

Consider a set of n points N with a metric distance $dist$. We denote points with letters P, and Q, and index them with i, and j such as P_i or Q_j. Distance of two points P_i and P_j is denoted by $dist(P_i, P_j)$. We define diversity of a set of points $S \subseteq N$ to be $Div(S) = \sum_{P,Q \in S} dist(P, Q)$.

Diversity Maximization. In an instance of the diversity maximization problem, we are given a parameter k and a set N of n points. The goal is to choose a set $S \subseteq N$ of (at most) k distinct points with maximum diversity. We denote this optimum set by OPT, and define AvgOpt to be the average distance of points in OPT, i.e. $\text{AvgOpt} = Div(\text{OPT})/\binom{k}{2}$. Given a set of elements $U \subseteq N$, let $\text{OPT}(U)$ be the value of optimum solution for the diversity maximization instance over points in U. For example, $\text{OPT}(N)$ corresponds to the value of OPT.

Randomized Composable Core-sets [3]. Assuming that n is large, all points may not fit on one machine, and we need to apply a distributed algorithm to solve the diversity maximization problem. To deal with this issue, we consider distributing all points into m machines with indices $\{1, \ldots, m\}$, where each point goes to one randomly chosen machine. Let $\{T_1, T_2, \ldots, T_m\}$ be subsets of points going to machines $\{1, 2, \ldots, m\}$ respectively. In this case, we say that $\{T_1, T_2, \ldots, T_m\}$ is a *random clustering of* N, i.e., $\{T_1, T_2, \ldots, T_m\}$ is a family of subsets $T_i \subseteq N$, where each point of N is assigned to one randomly chosen subset.

Definition 1. *Consider an algorithm* ALG *that given any subset* $T \subseteq N$ *returns a subset* $\text{ALG}(T) \subseteq T$ *with size at most* k'. *Let* $\{T_1, T_2, \ldots, T_m\}$ *be a random clustering of* N *to* m *subsets. We say that algorithm* ALG *is an algorithm that implements an* α-approximate randomized composable core-set of size k' *for the diversity maximization problem with cardinality constraint* k*, if,*

$$\mathbb{E}\left[\text{OPT}(\text{ALG}(T_1) \cup \ldots \cup \text{ALG}(T_m))\right] \geq \alpha \cdot \text{OPT}(T_1 \cup \ldots \cup T_m),$$

where the expectation is taken over the random choice of $\{T_1, T_2, \ldots, T_m\}$. *For brevity, instead of saying that* ALG *implements a composable core-set, we say that* ALG *is an* α*-approximate randomized composable core-set.*

[*] Please refer to the extended version for full version with all our results.
[†] Work done in part while at Brown University.

SPAA '19, June 22–24, 2019, Phoenix, AZ, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6184-2/19/06.
https://doi.org/https://doi.org/10.1145/3323165.3323172

For ease of notation, when it is clear from the context, we may drop the term composable, and refer to composable core-sets as core-sets. Throughout this paper, we discuss randomized composable core-sets for the diversity maximization problem.

Distributed Approximation Algorithm. Note that we can use a randomized α-approximate composable core-set algorithm ALG to design the following simple distributed approximation algorithm:

1. Based on a random clustering $\{T_1, \ldots, T_m\}$ defined above, allocate points in T_i to machine i.
2. Each machine i computes a randomized composable core-set $S_i \subseteq T_i$ of size k', i.e., $S_i = \mathrm{ALG}(T_i)$ for each $1 \leq i \leq m$.
3. Collect the union of all core-sets, $U = \cup_{1\leq i\leq m} S_i$, on one machine, and apply a *post-processing* algorithm ALG′ to compute a solution S over the set U. Output S.

If ALG′ is a local search or greedy 1/2-approximation for diversity maximization, the above algorithm simply achieves a distributed approximation factor of $\frac{\alpha}{2}$.

1.2 Our Contributions

We obtain both almost minimal core-set sizes $k' << k$ with almost optimal approximation guarantees (assuming data is distributed randomly to the machines). More precisely, we present a simple distributed algorithm that achieves an almost optimal communication complexity, and moreover, it asymptotically achieves approximation factor of 1/2 which is the best possible approximation factor for the global optimization problem under certain complexity theory assumptions. This improves a recent work of Abbasi Zadeh et al. [1] which presented a randomized 8/25 approximation algorithm for diversity maximization which also has a core-set size of at least k as in previous results. We also show hardness results for non-randomized partitioning of the data, proving that random partitioning is necessary to achieve reasonable performance guarantees.

Finally we show that our small-size core-set algorithms are scalable and practical as shown by our extensive empirical evaluation with large datasets, and they can be easily implemented in the major distributed computing systems like MapReduce. Furthermore, we show empirically that reducing the core-set sizes helps achieving major speed-ups while maintaining high diversity in the solutions. In real-life instances, our algorithms reach close-to-optimal solutions with approximation factor of > 90% while allowing speed-ups up to > 68 times w.r.t. the use of distributed large-size core-sets. An approximation factor of > 90% is far exceeding the theoretical approximation barrier for the problem and provides useful output as we show in our evaluation.

2 ALGORITHMS

We propose algorithms Diameter and Distributed Diameter explained as Algorithms 1 and 2 with much better provable worst-case approximation guarantees in range [0.25, 0.5] (approaching the $\frac{1}{2}$ barrier as number of machines converges to one), and having asymptotically smallest possible core-set size. We note that to have at least k selected distinct points by all machines, the core-set size should be at least $\frac{k}{m}$. Algorithm Diameter has core-set size $\max\{2, 2\lceil\frac{k}{2m}\rceil\} \leq \frac{k}{m} + 2$. We also know that, assuming the Planted Clique Conjecture, 0.5 is the best approximation guarantee even in the single-machine setting. Algorithm Distributed Diameter partitions the input set of n points randomly among m machines giving set T_ℓ to machine ℓ. Each machine runs Algorithm Diameter to find $r = \lceil\frac{k}{2m}\rceil$ disjoint pairs of points in T_ℓ with maximum sum of distances $\sum_{i=1}^{r} dist(P_{2i-1}, P_{2i})$ where $(P_1, P_2), (P_3, P_4), \cdots, (P_{2r-1}, P_{2r})$ are the r pairs of selected points. This can be done using maximum weighted matching algorithms in general graphs by constructing a complete graph with T_ℓ as its set of nodes and putting an edge weight of $dist(P, Q)$ for any $P, Q \in T_\ell$ between nodes representing P and Q. Algorithm Distributed Diameter receives mr pairs of points from the m machines. For simplicity of notation, let $(P_1, P_2), (P_3, P_4), \cdots, (P_{2mr-1}, P_{2mr})$ be these mr pairs. Among them $\lfloor\frac{k}{2}\rfloor$ pairs with maximum distances are chosen as the final output set. In other words, if we denote the $\lfloor k/2\rfloor$ selected pairs by $(P_{2i_1-1}, P_{2i_1}), (P_{2i_2-1}, P_{2i_2}), \cdots, (P_{2i_{\lfloor k/2\rfloor}-1}, P_{2i_{\lfloor k/2\rfloor}})$, we will have $dist(P_{2i_j-1}, P_{2i_j}) \geq dist(P_{2j'-1}, P_{2j'})$ for any $1 \leq j \leq \lfloor k/2\rfloor$ and any $j' \in \{1, 2, \cdots, mr\} \setminus \{i_1, i_2, \cdots, i_{\lfloor k/2\rfloor}\}$.

Algorithm 1 Algorithm Diameter

```
1. $r \leftarrow \lceil\frac{k}{2m}\rceil$
2. Find and return $r$ pairs of disjoint points $(P_1, P_2), (P_3, P_4), \cdots, (P_{2r-1}, P_{2r})$ with maximum sum of distances: $\sum_{i=1}^{r} dist(P_{2i-1}, P_{2i})$
```

Algorithm 2 Algorithm Distributed Diameter

```
1. Start with $m$ empty sets $\{T_\ell\}_{\ell=1}^{m}$.
2. Put each of the $n$ points in one of the $m$ sets $\{T_\ell\}_{\ell=1}^{m}$ independently and uniformly at random.
3. Each machine $1 \leq \ell \leq m$, runs algorithm Diameter on set $T_\ell$, and returns $r$ pairs.
4. Among the $mr$ selected pairs, return $\lfloor k/2\rfloor$ pairs with the maximum distances as the final solution.
```

Theorem 2. *Let $d = \frac{k}{m}$. For any $1 \leq k, m \leq n$, algorithm Diameter is a randomized $(f(d) - O(\frac{1}{m^{1/3}} + \frac{1}{k}))$-approximate composable core-set for the diversity maximization problem where function f is defined as follows for $d \leq 2$: $f(d) = \frac{1-e^{-d}}{d} - \frac{3e^{-d}}{4}$, and it is defined as follows for $d > 2$:*

$$f(d) = \frac{1}{2} - \frac{1}{2}\left(e^{-d}\sum_{r'=2r}^{\infty}\frac{r'+1-2r}{r'+1}\times\frac{d^{r'}}{r'!}\right) - \frac{1}{4}\left(e^{-d}\sum_{r''=0}^{r-1}\frac{1}{(2r''+1)}\times\frac{d^{2r''}}{(2r'')!}\right)$$

where r is $\lceil d/2\rceil$. Moreover, the distributed approximation factor of algorithm Distributed Diameter is also lower bounded by $(f(d) - O(\frac{1}{m^{1/3}} + \frac{1}{k}))$.

REFERENCES

[1] M. Ghadiri, V. Mirrokni, S. A. Zadeh, and M. Zadimoghaddam. Scalable feature selection via distributed diversity maximization. In *to appear in AAAI*, 2017.

[2] P. Indyk, S. Mahabadi, M. Mahdian, and V. Mirrokni. Composable core-sets for diversity and coverage maximization. In *PODS*, 2014.

[3] V. S. Mirrokni and M. Zadimoghaddam. Randomized composable core-sets for distributed submodular maximization. In *STOC*, 2015.