Approximation Schemes and Sketches for Clustering

David Saulpic,
under the supervision of Christoph Dürr and Vincent Cohen-Addad

Defended September 13rd, 2022

1 Background

Clustering is the task of partitioning a dataset in a meaningful way, such that similar data elements are in the same part and dissimilar ones in different parts. Clustering problems are among the most studied problems in theoretical computer science. Clustering applications range from data analysis [JSB12] to compression [Llo82], and this has given raise to several different ways of modeling the problem. The k-median and k-means problems are among the most prominent clustering objectives: algorithmic solutions to these problems are now part of the basic data analysis toolbox, and used as subroutines for many machine learning procedures [LW09; CNL11; CN12]. Furthermore, k-median and k-means have deep connections with other classical optimization problems, and understanding their complexity has been a fruitful research direction: Their study has led to many interesting algorithmic development (for instance for primal-dual algorithms [JV01; LS13; BPRST15] or local-search based ones [CGHOS22; CGLS23]).

Given a metric space (X, dist), and two sets $P, \mathbb{C} \subseteq X$, the goal of the k-median problem (respectively k-means problem) is to find a set of k centers $\mathcal{S} \subset \mathbb{C}^k$ in order to minimize the sum of distances (resp. distances squared) for each point in P to its closest center in $\mathcal{S}$:

$$\text{cost}(P, \mathcal{S}) = \sum_{p \in P} \min_{s \in \mathcal{S}} \text{dist}^z(p, \mathcal{S}),$$

where $z = 1$ for k-median and $z = 2$ for k-means.

Those two problems are generalizations of the standard median and mean: instead of summarizing the whole dataset with a single point, it allows to have different clusters, each represented by its median or mean. Finding the best clustering should therefore provide a good representation of the data, more precisely than with a single median or mean. The k-median and k-means problems are therefore naturally part of the basic data-analysis toolbox, and are used as building blocks of many machine learning procedures.

However, they are hard to solve in most cases: they are NP-hard, even in very cases. For instance, they are NP-hard when the data merely consists of points lying in the Euclidean plane [MNV12; MS84], or when $k = 2$ [DF09]. Hence, even in low-dimensional Euclidean spaces, one has to resign itself to computing only approximate solutions.

In this thesis, our general goal is to develop algorithms that provably compute good solutions to those problems. We aim at dealing with big data, in complex spaces.

First Part: Fast Algorithms via Embedding into Trees. In the first part of the thesis, we consider the following question:

Question 1. *Is there some metric space where it is possible to solve k-median and k-means in near-linear time?*

Besides the big data motivation, this question is interesting in its own right, as understanding the complexity and the structure of the problems is an attractive mathematical question. Since the problems are NP-hard even in very restricted metric spaces such as the Euclidean plane, we will target approximation schemes, namely, design an algorithm that compute a solution with cost at most $(1+\varepsilon)$ times the optimal cost, for any fixed $\varepsilon > 0$.

We answer this question in low-dimensional Euclidean spaces, and a generalization of those called doubling metrics.[1] Studying the problem with that angle allows us to get insight that can be used in other settings as well. In particular, in the so-called Differential Privacy model.

This is a key application for clustering: as data collection appears everywhere in our lives, people and more generally democracies are concerned with the effect of data analysis can have on privacy. Laws are now enforcing companies to respect some privacy principles while collecting and analyzing data: hence, it becomes necessary to develop data analysis algorithms that respect in some sense the privacy of users. This has been modeled by the notion of Differential Privacy, that we explore via the embedding into ultrametrics. It turns out that the techniques we introduce are particularly suited to that privacy model. We show and experiment a practical private algorithm for clustering that enjoys provable guarantees.

Second Part: Sketching, and Coping with Big Data The other theme of this thesis is to find compression schemes for clustering. Datasets are in many practical cases too large to be processed conventionally, as the data simply does not fit into one computer's memory. Henceforth, sketching, compression, and summarization techniques are at the heart of modern data analysis. This has led to new algorithms operating in other models of computation such as streaming, distributed computing or massively-parallel computation (MPC). For these algorithms, finding good small-size representations – also called *sketches* – of the input data is key.

The main sketch we study in this thesis is called *coreset*. Given $\varepsilon > 0$, a (weighted) set Ω is an ε-coreset for P if for any set of k centers $\mathcal{S}$, $\text{cost}(\Omega, \mathcal{S}) = (1 \pm \varepsilon)\text{cost}(P, \mathcal{S})$. In other words, Ω preserves approximately the cost function.

Computing small coreset has numerous advantages. First, reducing the size of the input may allow to reinstate the "traditional" algorithms, that are already well analyzed and understood. Second, coreset can be used in settings where there are additional constraints on the memory usage of algorithms: for instance, when the input cannot fit in a single machine and is distributed among several of them, those can merely exchange coresets to communicate their data. This has small size compared to the full input and consequently can be stored and processed in a single machine, instead of the full dataset.

Question 2. *What are the best coreset size possible, for k-median and k-means? What particular structure on the metric space is useful to construct small coreset?*

2 State of the Art

Approximation Algorithms fig. 1 summarizes the current state of our knowledge in terms of polynomial-time approximability: we say that a solution $\mathcal{S}$ is an α-approximation if its cost is at most α times the cost of the optimal solution.

[1]The doubling dimension of a metric space is d when any ball of radius R can be covered with 2^d balls of radius $R/2$. Doubling metrics are ones with bounded doubling dimension. This generalizes Euclidean space, as the space $(\mathbb{R}^d, \ell_2)$ has doubling dimension $\Theta(d)$.

	Discrete k-median	Discrete k-means	Euclidean k-median	Euclidean k-means
Lower bound	$1 + 2/e$ [GK99]	$1 + 8/e$ [GK99]	1.06 [CSL22]	1.015 [CSL22]
Upper bound	2.675 [BPRAT15]	6.36 [ANFSW17]	2.41 [CEMN22]	5.96 [CEMN22]

Figure 1: Approximability of k-median and k-means. The lower-bounds are conditioned on the assumption P $\neq$ NP.

Hence, if one wishes to have a polynomial-time algorithm with very good approximation guarantee for (k, z)-clustering, it is necessary to make some assumption on the input data. For instance, there are polynomial approximation schemes in Euclidean spaces of fixed dimensions (see [KR07; CKM19; FRS19]). Here, by polynomial time we mean $|P|^{f(d,\varepsilon)}$ for some function f, as ε and d are considered to be fixed. A standard generalization of the Euclidean dimension that abstracts out a lot of the geometry and allows us to focus on the most crucial properties is called the *doubling dimension*, that we alreay mentioned. In case of bounded doubling dimension, [FRS19] showed how to compute a $(1 + \varepsilon)$-approximation in time $|P|^{f(d,\varepsilon)}$.

Nonetheless, obtaining an efficient approximation scheme (namely a $(1 + \varepsilon)$-approximation algorithm running in time $f(\varepsilon, d)\text{poly}(n)$) for k-Median and k-Means in Euclidean space, or more generally metrics of doubling dimension d has remained a major challenge.

For clustering with privacy constraints, the state-of-the art results have the a multiplicative approximation factor that matches the one of non-private algorithms (See [GKM20]). However, those algorithms suffer from a large additive error, are far from being practical and even hardly implementable. On the other hand, state-of-the-art implementations either have no theoretical guarantees on the quality of the solution obtained, or cannot be implemented in large-scale scenario where the data is distributed. In contrast, we present a parallel implementation of the algorithm, and we show experimentally that its performances are comparable to the best non-private methods.

A Brief History of Coreset for Clustering. The study of coreset for clustering started with the work of [HM04]. They gave a construction based on snapping points to a grid of the space. This is specifically tailored to Euclidean spaces, and has a prohibited exponential dependency in the dimension. The first breakthrough is due to Chen [Che09], who introduced sampling in the coreset toolbox, and managed to show the construction of coresets of size $O(k^2\varepsilon^{-2}\log^2 n)$ for discrete n points metric spaces, and size $O(k^2 d\varepsilon^{-2}\log n)$ in Euclidean space. Notably, the dependency in the dimension is merely linear.

The state-of-the-art analysis relies on a VC-dimension type complexity measure: [FL11] presented a way of constructing coresets with a size bounded by this dimension. While this technique provided many strong result in various metric spaces, tighter bounds are often achievable. For instance, in d dimensional Euclidean spaces this would yield coresets of size $O_{\varepsilon,z}(k^2 \cdot d\log^2 k)$, but [HV20] and [BJKW21] showed the existence of a coreset with $O(k \cdot \log^2 k \cdot \varepsilon^{-2z-2})$ points. This VC-dimension based analysis was proven powerful in various metric spaces, such as doubling spaces by [HJLW18], graphs of bounded treewidth by [BBHJKW20] or the shortest-path metric of a graph excluding a fixed minor [BJKW21]. However, range spaces of even heavily constrained metrics do not necessarily have small VC-dimension (e.g. bounded doubling dimension does not imply bounded VC-dimension or vice versa [HJLW18; LL06]), and applying previous techniques requires heavy additional machinery to adapt the VC-dimension approach to them. Moreover, the bounds provided are far from the bound obtained for Euclidean spaces: their dependency in k is at least $\Omega(k^2)$, leaving a significant gap to the best lower bounds of $\Omega(k)$.

3 Contributions

To answer question 1, we present an approximation scheme for (k, z)-Clustering in Euclidean spaces, namely an algorithm that computes a $(1 + \varepsilon)$-approximation to the problem. For any fixed ε and dimension, this algorithm runs in near-linear time – i.e., even faster than assigning naively each point to the closest center, which takes time nk. This is based on a joint work with Vincent Cohen-Addad and Andreas Feldmann [CFS21], that appeared in the Journal of the ACM.

The algorithm we propose is based on embedding the input in a tree-like structure. As mentioned above, this is helpful to get algorithm that respect some form of privacy. Indeed, we manage to apply those techniques to get a private algorithm for k-median and k-means with provable approximation guarantee. This chapter is more oriented towards practice: we present a scalable algorithm, able to run in a distributed setting. In particular, we implemented the algorithm and showed its practical efficiency both in terms of speed and quality of the computed solution. This is based on a collaboration with Vincent Cohen-Addad, Alessandro Epasto, Silvio Lattanzi, Vahab Mirrokni, Andres Munoz, Chris Schwiegelshohn and Sergei Vassilvitskii, that was presented at the conference KDD 22 [CELMMSSV22].

To answer question 2, we present a very generic coreset construction, showing the existence of small coresets under a specific assumption. We then show that this assumption holds in many different metric spaces, resulting in state of the art coreset construction. For discrete metrics, we present coreset of size essentially $O(\varepsilon^{-2}k \log n)$. For metrics of doubling dimension d, we show coreset of size $O(\varepsilon^{-2}kd)$. The Euclidean space $\mathbb{R}^d$ is known to have doubling dimension $\Theta(d)$: the result carries over. It can be further improved using standard dimension reduction techniques: it is possible to replace the dependency in d by $O(\varepsilon^{-2} \log k)$. We also show that metrics induced by graphs with small *separators* have small centroid set: namely, metric induced by graph of bounded treewidth or excluding a minor. In all those case, the dependency in k, the dimension or the treewidth is optimal. Those results are based on an article published at STOC 2021 with Vincent Cohen-Addad and Chris Schwiegelshohn [CSS21a].

We then show that our construction for discrete and doubling metrics are tight: there exist a family of discrete metric space such that any coreset on those must have size at least $\Omega(\varepsilon^{-2}k \log n)$. This completes nicely our understanding of coreset for those spaces: upper and lower bounds are tight. This result is part of a joint work with Vincent Cohen-Addad, Kasper Green Larsen and Chris Schwiegelshohn [CLSS22], that was presented at STOC 2022.

We also present a different coreset construction, that allows for *deterministic* coresets – which can be then applied to get deterministic $(1 + \varepsilon)$-approximation. The previous coreset constructions are randomized, and succeed with probability $1 - \delta$. However, it is co-NP hard to verify that the outcome of a randomized coreset construction is indeed a valid coreset [SS22]: hence, determinism may be a desirable property. We present such coresets construction for various metric spaces.

In particular, to achieve deterministic bounds similar to the randomized one in Euclidean spaces, one needs to remove any dependency on the dimension d: one of the technical ingredients of the chapter is to show deterministic dimension reduction for clustering. This is of independent interest, and provides another way of sketching Euclidean input. This is a collaboration with Vincent Cohen-Addad and Chris Schwiegelshohn, currently under submission [CSS].

Finally, we show how to use the coreset knowledge developed previously to construct algorithms running in sublinear time to compute the median, the mean and more generally an approximation to $(1, z)$-clustering. We show that it is enough to consider a *constant* number of input point drawn uniformly at random to compute this solution, and implement our algorithm to show the practical speed-up it allows. This chapter is based on a work that was presented as a spotlight at NeurIPS 2021, with Vincent Cohen-Addad and Chris Schwiegelshohn [CSS21b].

Publications Presented in the Thesis

[CELMMSSV22] Vincent Cohen-Addad, Alessandro Epasto, Silvio Lattanzi, Vahab Mirrokni, Andres Munoz, David Saulpic, Chris Schwiegelshohn, and Sergei Vassilvitskii. "Scalable Differentially Private Clustering via Hierarchically Separated Trees". In: Conference on Knowledge Discovery and Data Mining (KDD). 2022.

[CFS21] Vincent Cohen-Addad, Andreas Emil Feldmann, and David Saulpic. "Near-linear Time Approximation Schemes for Clustering in Doubling Metrics". In: *J. ACM*. Vol. 68. 2021.

[CLSS22] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. "Towards Optimal Lower Bounds for k-median and k-means Coresets". In: *Symposium on Theory of Computing (STOC)*. 2022.

[CSS] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. "On Deterministic Clustering Sketches". In: submitted.

[CSS21a] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. "A new coreset framework for clustering". In: *Symposium on Theory of Computing (STOC)*. 2021.

[CSS21b] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. "Improved Coresets and Sublinear Algorithms for Power Means in Euclidean Spaces". In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

Other Publications of the Author

[BKS17] Amariah Becker, Philip N. Klein, and David Saulpic. "A Quasi-Polynomial-Time Approximation Scheme for Vehicle Routing on Planar and Bounded-Genus Graphs". In: *European Symposium on Algorithms, ESA*. 2017.

[BKS18] Amariah Becker, Philip N Klein, and David Saulpic. "Polynomial-Time Approximation Schemes for k-center, k-median, and Capacitated Vehicle Routing in Bounded Highway Dimension". In: *European Symposium on Algorithms, ESA*. 2018.

[CGHOS22] Vincent Cohen-Addad, Anupam Gupta, Lunjia Hu, Hoon Oh, and David Saulpic. "An Improved Local Search Algorithm for k-Median". In: *Symposium on Discrete Algorithms, SODA*. 2022.

[CHPSS19] Vincent Cohen-Addad, Niklas Hjuler, Nikos Parotsidis, David Saulpic, and Chris Schwiegelshohn. "Fully Dynamic Consistent Facility Location". In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2019.

[CKMS20] Vincent Cohen-Addad, Adrian Kosowski, Frederik Mallmann-Trenn, and David Saulpic. "On the Power of Louvain in the Stochastic Block Model". In: *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2020.

[CMS22a] Vincent Cohen-Addad, Frederik Mallmann-Trenn, and David Saulpic. "A Massively Parallel Modularity-Maximizing Algorithm With Provable Guarantees". In: *Symposium on Principles of Distributed Computing (PODC)*. 2022.

[CMS22b] Vincent Cohen-Addad, Frederik Mallmann-Trenn, and David Saulpic. "Community Recovery in the Degree-Heterogeneous Stochastic Block Model". In: *Conference on Learning Theory (COLT)*. 2022.

[FS21] Andreas Emil Feldmann and David Saulpic. "Polynomial time approximation schemes for clustering in low highway dimension graphs". In: *J. Comput. Syst. Sci.* Vol. 122. 2021.

[HIPS19] Niklas Hjuler, Giuseppe F. Italiano, Nikos Parotsidis, and David Saulpic. "Dominating Sets and Connected Dominating Sets in Dynamic Graphs". In: *International Symposium on Theoretical Aspects of Computer Science, STACS*. 2019.

Bibliography

[ANFSW17] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. “Better guarantees for k-means and euclidean k-median by primal-dual algorithms”. In: *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on.* Ieee. 2017, pp. 61–72.

[BBHJKW20] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H. C. Jiang, Robert Krauthgamer, and Xuan Wu. *Coresets for Clustering in Graphs of Bounded Treewidth.* 2020.

[BJKW21] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. “Coresets for Clustering in Excluded-minor Graphs and Beyond”. In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021.* Ed. by Dániel Marx. Consulted on arxiv on May 2022. SIAM, 2021, pp. 2679–2696. URL: `https://doi.org/10.1137/1.9781611976465.159`.

[BPRAT15] Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Srinivasan Aravind, and Khoa Trinh. “An Improved Approximation for k-median, and Positive Correlation in Budgeted Optimization”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015.* 2015, pp. 737–756.

[BPRST15] Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. “An Improved Approximation for k-median, and Positive Correlation in Budgeted Optimization”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015.* 2015, pp. 737–756. DOI: `10.1137/1.9781611973730.50`. URL: `http://dx.doi.org/10.1137/1.9781611973730.50`.

[CEMN22] Vincent Cohen-Addad, Hossein Esfandiari, Vahab S. Mirrokni, and Shyam Narayanan. “Improved Approximations for Euclidean k-means and k-median, via Nested Quasi-Independent Sets”. In: (2022).

[CGLS23] Vincent Cohen-Addad, Fabrizio Grandoni, Euiwoong Lee, and Chris Schwiegelshohn. “Breaching the 2 LMP Approximation Barrier for Facility Location with Applications to k-Median”. In: *Accepted at SODA 23* (2023).

[Che09] Ke Chen. “On Coresets for k-Median and k-Means Clustering in Metric and Euclidean Spaces and Their Applications”. In: *SIAM J. Comput.* 39.3 (2009), pp. 923–947.

[CKM19] Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. “Local Search Yields Approximation Schemes for k-Means and k-Median in Euclidean and Minor-Free Metrics”. In: *SIAM J. Comput.* 48.2 (2019), pp. 644–667. DOI: `10.1137/17M112717X`. URL: `https://doi.org/10.1137/17M112717X`.

[CN12] Adam Coates and Andrew Y Ng. “Learning feature representations with k-means”. In: *Neural networks: Tricks of the trade.* Springer, 2012, pp. 561–580.

[CNL11] Adam Coates, Andrew Ng, and Honglak Lee. “An analysis of single-layer networks in unsupervised feature learning”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics.* JMLR Workshop and Conference Proceedings. 2011, pp. 215–223.

[CSL22] Vincent Cohen-Addad, Karthik C. S., and Euiwoong Lee. “Johnson Coverage Hypothesis: Inapproximability of k-means and k-median in ℓ_p-metrics”. In: *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022.* Ed. by Joseph (Seffi) Naor and Niv Buchbinder. SIAM, 2022, pp. 1493–1530. DOI: `10.1137/1.9781611977073.63`. URL: `https://doi.org/10.1137/1.9781611977073.63`.

[DF09] Sanjoy Dasgupta and Yoav Freund. “Random projection trees for vector quantization”. In: *IEEE Transactions on Information Theory* 55.7 (2009), pp. 3229–3242.

[FL11] Dan Feldman and Michael Langberg. “A unified framework for approximating and clustering data”. In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011.* 2011, pp. 569–578.

[FRS19] Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. "Local Search Yields a PTAS for k-Means in Doubling Metrics". In: *SIAM J. Comput.* 48.2 (2019), pp. 452–480. DOI: 10.1137/17M1127181. URL: https://doi.org/10.1137/17M1127181.

[GK99] Sudipto Guha and Samir Khuller. "Greedy Strikes Back: Improved Facility Location Algorithms". In: *J. Algorithms* 31.1 (1999), pp. 228–248. DOI: 10.1006/jagm.1998.0993. URL: http://dx.doi.org/10.1006/jagm.1998.0993.

[GKM20] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. "Differentially Private Clustering: Tight Approximation Ratios". In: *Advances in Neural Information Processing Systems*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020.

[HJLW18] Lingxiao Huang, Shaofeng H.-C. Jiang, Jian Li, and Xuan Wu. "Epsilon-Coresets for Clustering (with Outliers) in Doubling Metrics". In: *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*. 2018, pp. 814–825. DOI: 10.1109/FOCS.2018.00082. URL: https://doi.org/10.1109/FOCS.2018.00082.

[HM04] Sariel Har-Peled and Soham Mazumdar. "On coresets for k-means and k-median clustering". In: *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*. 2004, pp. 291–300.

[HV20] Lingxiao Huang and Nisheeth K. Vishnoi. "Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal". In: *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*. Ed. by Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy. ACM, 2020, pp. 1416–1429. DOI: 10.1145/3357713.3384296. URL: https://doi.org/10.1145/3357713.3384296.

[JSB12] Krzystof Jajuga, Andrzej Sokolowski, and Hans-Hermann Bock. "Classification, clustering, and data analysis: recent advances and applications". In: (2012).

[JV01] Kamal Jain and Vijay V. Vazirani. "Approximation algorithms for metric facility location and k-Median problems using the primal-dual schema and Lagrangian relaxation". In: *J. ACM* 48.2 (2001), pp. 274–296. DOI: 10.1145/375827.375845. URL: http://doi.acm.org/10.1145/375827.375845.

[KR07] Stavros G Kolliopoulos and Satish Rao. "A nearly linear-time approximation scheme for the Euclidean k-median problem". In: *SIAM Journal on Computing* 37.3 (2007), pp. 757–782.

[LL06] Yi Li and Philip M. Long. "Learnability and the doubling dimension". In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. 2006, pp. 889–896. URL: http://papers.nips.cc/paper/3068-learnability-and-the-doubling-dimension.

[Llo82] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[LS13] Shi Li and Ola Svensson. "Approximating k-median via pseudo-approximation". In: *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*. 2013, pp. 901–910. DOI: 10.1145/2488608.2488723. URL: http://doi.acm.org/10.1145/2488608.2488723.

[LW09] Dekang Lin and Xiaoyun Wu. "Phrase clustering for discriminative learning". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009, pp. 1030–1038.

[MNV12] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi R. Varadarajan. "The planar k-means problem is NP-hard". In: *Theor. Comput. Sci.* 442 (2012), pp. 13–21.

[MS84] Nimrod Megiddo and Kenneth J Supowit. "On the complexity of some common geometric location problems". In: *SIAM journal on computing* 13.1 (1984), pp. 182–196.

[SS22] Chris Schwiegelshohn and Omar Ali Sheikh-Omar. “An Empirical Evaluation of k-Means Coresets”. In: *30th Annual European Symposium on Algorithms, ESA 2022, September 5-9, 2022, Berlin/Potsdam, Germany*. Ed. by Shiri Chechik, Gonzalo Navarro, Eva Rotenberg, and Grzegorz Herman. Vol. 244. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022, 84:1–84:17. DOI: 10.4230/LIPIcs.ESA.2022.84. URL: https://doi.org/10.4230/LIPIcs.ESA.2022.84.