The Hardness of Approximation of Euclidean k-Means

Pranjal Awasthi[1], Moses Charikar[2], Ravishankar Krishnaswamy[3], and Ali Kemal Sinop[4]

1 **Computer Science Department, Princeton University, USA**
pawasthi@cs.cmu.edu

2 **Computer Science Department, Princeton University, USA**
moses@cs.princeton.edu

3 **Microsoft Research, India**
ravishan@cs.cmu.edu

4 **Simons Institute for the Theory of Computing, USA**
University of California, Berkeley
asinop@cs.cmu.edu

Abstract

The Euclidean k-means problem is a classical problem that has been extensively studied in the theoretical computer science, machine learning and the computational geometry communities. In this problem, we are given a set of n points in Euclidean space $\mathbb{R}^d$, and the goal is to choose k center points in $\mathbb{R}^d$ so that the sum of squared distances of each point to its nearest center is minimized. The best approximation algorithms for this problem include a polynomial time constant factor approximation for general k and a $(1+\epsilon)$-approximation which runs in time poly(n) $\exp(k/\epsilon)$. At the other extreme, the only known computational complexity result for this problem is NP-hardness [1]. The main difficulty in obtaining hardness results stems from the Euclidean nature of the problem, and the fact that any point in $\mathbb{R}^d$ can be a potential center. This gap in understanding left open the intriguing possibility that the problem might admit a PTAS for all k, d.

In this paper we provide the first hardness of approximation for the Euclidean k-means problem. Concretely, we show that there exists a constant $\epsilon > 0$ such that it is NP-hard to approximate the k-means objective to within a factor of $(1+\epsilon)$. We show this via an efficient reduction from the vertex cover problem on *triangle-free graphs*: given a triangle-free graph, the goal is to choose the fewest number of vertices which are incident on all the edges. Additionally, we give a proof that the current best hardness results for vertex cover can be carried over to triangle-free graphs. To show this we transform G, a known hard vertex cover instance, by taking a graph product with a suitably chosen graph H, and showing that the size of the (normalized) maximum independent set is almost exactly preserved in the product graph using a spectral analysis, which might be of independent interest.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems

Keywords and phrases Euclidean k-means, Hardness of Approximation, Vertex Cover

Digital Object Identifier 10.4230/LIPIcs.SOCG.2015.754

1 Introduction

Clustering is the task of partitioning a set of items such as web pages, protein sequences etc. into groups of related items. This is a fundamental task in machine learning, information retrieval, computational geometry, computer vision, data visualization and many other

31st International Symposium on Computational Geometry (SoCG'15).
Editors: Lars Arge and János Pach; pp. 754–767
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

domains. In many applications, clustering is often used as a first step toward other fine grained tasks such as classification. Needless to say, the problem of clustering has received significant attention over the years and there is a large body of work on both the applied and the theoretical aspects of the problem [6, 4, 10, 13, 19, 21, 26, 33, 8, 28, 34]. A common way to approach the task of clustering is to map the set of items into a metric space where distances correspond to how different two items are from each other. Using this distance information, one then tries to optimize an objective function to get the desired clustering. Among the most commonly used objective function used in the clustering literature is the k-means objective function. In the k-means problem, the input is a set S of n data points in Euclidean space $\mathbb{R}^d$, and the goal is to choose k center points $C^* = \{c_1, c_2, \ldots, c_k\}$ from $\mathbb{R}^d$ so as to minimize $\Phi = \sum_{x \in S} \min_i \|x - c(x)\|^2$, where $c(x) \in C^*$ is the center closest to x. Aside from being a natural clustering objective, an important motivation for studying this objective function stems from the fact that a very popular and widely used heuristic (appropriately called the *k-means heuristic* [28]) attempts to minimize this k-means objective function.

While the k-means heuristic is very much tied to the k-means objective function, there are many examples where it converges to a solution which is far away from the optimal k-means solution. This raises the important question of whether there exist provable algorithms for the k-means problem in general Euclidean space, which is the focus problem of our paper. Unfortunately though, the approximability of the problem is not very well understood. From the algorithmic side, there has been much focus on getting $(1+\epsilon)$-approximations that run as efficiently as possible. Indeed, for fixed k, Euclidean k-means admits a PTAS [26, 16]. These algorithms have exponential dependence in k, but only linear dependence in the number of points and the dimensionality of the space. As mentioned above, there is also empirical and theoretical evidence for the effectiveness of very simple heuristics for this problem [33, 28, 25]. For arbitrary k and d, the best known approximation algorithm for k-means achieves a factor of $9+\epsilon$ [21]. In contrast to the above body of work on getting algorithms for k-means, lower bounds for k-means have remained elusive. In fact, until recently, even NP-hardness was not known for the k-means objective [11, 1]. This is perhaps due to the fact that as opposed to many discrete optimization problems, the k-means problem allows one to choose any point in the Euclidean space as a center. The above observations lead to the following intriguing possibility:

"Is there a PTAS for Euclidean k-means for arbitrary k and dimension d?"

In this paper we answer this question in the negative and provide the first hardness of approximation for the Euclidean k-means problem.

▶ **Theorem 1.1.** *There exists a constant $\epsilon > 0$ such that it is NP-hard to approximate the Euclidean k-means to a factor better than $(1+\epsilon)$.*

The starting point for our reduction is the Vertex-Cover problem on triangle-free graphs: here, given a triangle-free graph, the goal is to choose the fewest number of vertices which are incident on all the edges in the graph. This naturally leads us to our other main result in this paper, that of showing hardness of approximation of vertex cover on triangle-free graphs. Kortsarz et al [24] show that if the vertex cover problem is hard to approximate to a factor of $\alpha \geq 3/2$, then it is hard to approximate vertex cover on triangle-free graphs to the same factor of α. While such a hardness (in fact, a factor of $2-\epsilon$ [22]) is known assuming the stronger unique games conjecture, the best known NP-hardness results do not satisfy $\alpha \geq 3/2$. We settle this question by showing NP-hardness results for approximating vertex cover on triangle-free graphs, which match the best known hardness on general graphs.

▶ **Theorem 1.2.** *It is NP-hard to approximate Vertex Cover on triangle-free graphs to within any factor smaller than* 1.36*.*

2 Main Technical Contribution

In Section 4, we show a reduction from Vertex-Cover on triangle-free graphs to Euclidean k-means where the vertex cover instances have small cover size if and only if the corresponding k-means instances have a low cost. A crucial ingredient is to relate the cost of the clusters to the structural properties of the original graph, which lets us transition from the Euclidean problem to a completely combinatorial problem. Then in Section 5, we prove that the known hardness of approximation results for Vertex-Cover carry over to triangle-free graphs. This improves over existing hardness results for vertex cover on triangle-free graphs [24]. Furthermore, we believe that our proof techniques are of independent interest. Specifically, our reduction transforms known hard instances G of vertex cover, by taking a graph product with an appropriately chosen graph H. We then show that the size of the vertex cover in the new graph (in proportion to the size of the graph) can be related to spectral properties of H. In fact, by choosing H to have a bounded spectral radius, we show that the vertex covers in G and the product graph are roughly preserved, while also ensuring that the product graph is triangle-free. Combining this with our reduction to k-means completes the proof.

3 Related Work

Arthur and Vassilvitskii [5] proposed k-means++, a random sampling based approximation algorithm for Euclidean k-means which achieves a factor of $O(\log k)$. This was improved by Kanungo et al. [21] who proposed a local search based algorithm which achieves a factor of $(9+\epsilon)$. This is currently the best known approximation algorithm for k-means. For fixed k and d, Matousek [31] gave a PTAS for k-means which runs in time $O(n\epsilon^{-2k^2d}\log^k n))$. Here n is the number of points and d is the dimensionality of the space. This was improved by Badoiu et al. [7] who gave a PTAS for fixed k and any d with run time $O(2^{(k/\epsilon)^{O(1)}}poly(d)n\log^k n)$. Kumar et al. [26] gave an improved PTAS with exponential dependence in k and only linear dependence in n and d. Feldman et al. [16] combined this with efficient coreset constructions to give a PTAS for fixed k with improved dependence on k. The work of Dasgupta [11] and Aloise et al. [1] showed that Euclidean k-means is NP-hard even for $k=2$. Mahajan et al. [30] also show that the k-means problem is NP-hard for points in the plane.

There are also many other clustering objectives related to k-means which are commonly studied. The most relevant to our discussion are the k-median and the k-center objectives. In the first problem, the objective is to pick k centers to minimize the sum of distances of each point to the nearest center (note that the distances are not squared). The problem deviates from k-means in two crucial aspects, both owing to the different contexts in which the two problems are studied: (i) the k-median problem is typically studied in the setting where the centers are one of the data points (or come from a set of possible centers specified in the input), and (ii) the problem is also very widely studied on general metrics, without the Euclidean restriction. The k-median problem has been a testbed of developing new techniques in approximation algorithms, and has constantly seen improvements even until very recently [20, 19, 27]. Currently, the best known approximation for k-median is a factor of $2.611+\epsilon$ due to Bykra et al. [9]. On the other hand, it is also known that the k-median objective (on general metrics) is NP-hard to approximate to a factor better than $(1+1/e)$ [19]. When restricted to Euclidean metrics, Kolliopoulos et al. [23] show a PTAS for k-median on constant dimensional spaces. On the negative side for k-median on Euclidean metrics, it is known that the discrete problem (where centers come from a specified input) cannot have a PTAS under standard complexity assumptions [17]. As mentioned earlier, all these results are for the version when the possible candidate centers is specified in the input. For the problem where any point can be a center, Arora et al. [4] show a PTAS when the points are on a 2-dimensional plane.

In the k-center problem the objective is to pick k center points such that the maximum distance of any data point to the closest center point is minimized. In general metrics, this problem admits a 2-factor approximation which is also optimal assuming $P \neq NP$ [18]. For Euclidean metric when the center could be any point in the space, the upper bound is still 2 and the best hardness of approximation is a factor 1.82 [15].

4 Our Hardness Reduction: From Vertex Cover to Euclidean k-means

In this section, we show a reduction from the Vertex-Cover problem (on triangle-free graphs) to the k-means problem. Formally, the vertex cover problem can be stated as follows: Given an undirected graph $G = (V, E)$, choose a subset S of vertices (with minimum $|S|$) such that S is incident on every edge of the graph. More specifically, our reduction establishes the following theorem.

▶ **Theorem 4.1.** *There is an efficient reduction from instances of Vertex Cover (on triangle-free graphs with m edges) to those of Euclidean k-means that satisfies the following properties:*

(i) *if the Vertex Cover instance has value k, then the k-means instance has cost $\leq m - k$.*
(ii) *if the Vertex Cover instance has value at least $k(1+\epsilon)$, then the optimal k-means cost is $\geq m - (1 - \Omega(\epsilon))k$. Here, ϵ is some fixed constant > 0.*

In Section 5, we show that there exist triangle-free graph instances of vertex cover on $m = \Theta(n)$ edges, and $k = \Omega(n)$ such that it is NP-hard to distinguish if the instance has a vertex cover of size at most k, or all vertex covers have size at least $(1+\epsilon)k$, for some constant $\epsilon > 0$.

Now, let $k = m/\Delta$ where $\Delta = \Omega(1)$ from the hard vertex cover instances. Then, from Theorem 4.1, we get that if the vertex cover has value k, then the k-means cost is at most $m(1 - \frac{1}{\Delta})$, and if the vertex cover is at least $k(1+\epsilon)$, then the optimal k-means cost is at least $m(1 - \frac{1-\Omega(\epsilon)}{\Delta})$. Therefore, the vertex cover hardness says that it is also NP-hard to distinguish if the resulting k-means instance has cost at most $m(1 - \frac{1}{\Delta})$ or cost more than $m(1 - \frac{1-\Omega(\epsilon)}{\Delta})$. Since Δ is a constant, this implies that it is NP-hard to approximate the k-means problem within some factor $(1 + \Omega(\epsilon))$, thereby establishing our main result Theorem 1.1. In what follows, we prove Theorem 4.1.

4.1 Proof of Theorem 4.1

Let $G = (V, E)$ denote the graph in the Vertex Cover instance $\mathcal{I}$, with parameter k denoting the number of vertices we can select. We associate the vertices with natural numbers $[n]$. Therefore, we refer to vertices by natural numbers i, and edges by pairs of natural numbers (i, j).

Construction of k-means Instance $\mathcal{I}_{km}$

For each vertex $i \in [n]$, we have a unit vector $\mathrm{x}_i = (0, 0, \dots, 1, \dots, 0)$ which has a 1 in the i^{th} coordinate and 0 elsewhere. Now, for each edge $e \equiv (i, j)$, we have a vector $\mathrm{x}_e \overset{\text{def}}{=} \mathrm{x}_i + \mathrm{x}_j$. Our data points on which we solve the k-means problem is precisely $\{\mathrm{x}_e : e \in E\}$. This completes the definition of $\mathcal{I}_{km}$.

▶ Remark. As stated, the dimensionality of the points we have constructed is n, and we get a hardness factor of $(1 + \epsilon)$. However, by using the dimensionality reduction ideas of Johnson and Lindenstrauss (see, e.g. [12]), without loss of generality, we can assume that the points lie in $O(\log n/\epsilon^2)$ dimensions and our hardness results still hold true. This is because, after the transformation, all pairwise distances (and in particular, the k-means objective function) are preserved upto a factor of $(1 + \epsilon/10)$ of the original values, and so our hardness factor is also (almost) preserved, i.e., we would get hardness of approximation of $(1 + \Omega(\epsilon))$.

However, for simplicity, we stick with the n dimensional vectors as it makes the presentation much cleaner.

4.2 Completeness

Suppose $\mathcal{I}$ is such that there exists a vertex cover $S^* = \{v_1, v_2, \dots, v_k\}$ of k vertices which can cover all the edges. We will now show that we can recover a good clustering of low k-means cost. To this end, let E_{v_ℓ} denote the set of edges which are covered by v_ℓ for $1 \leq \ell \leq k$. If an edge is covered by two vertices, we assume that only one of them covers it. As a result, note that the E_{v_ℓ}'s are pairwise disjoint (and their union is E), and each E_{v_ℓ} is of the form $\{(v_\ell, w_{\ell,1}), (v_\ell, w_{\ell,2}), \dots, (v_\ell, w_{\ell,p_\ell})\}$.

Now, to get our clustering, we do the following: for each $v \in S^*$, form a cluster out of the data points $\mathcal{F}_v := \{\mathrm{x}_e : e \in E_v\}$. We now analyze the average connection cost of this solution. To this end, we begin with some easy observations about the k-means clustering. Indeed, since any cluster is of a set of data points (corresponding to a subset of edges in the graph G), we shall abuse notation and associate any cluster $\mathcal{F}$ also with the corresponding subgraph on V, i.e., $\mathcal{F} \subseteq E$. Moreover, we use $d_{\mathcal{F}}(i)$ to denote the degree of node i in $\mathcal{F}$ and $m_{\mathcal{F}}$ to denote the number of edges in $\mathcal{F}$, $m_{\mathcal{F}} = |\mathcal{F}|$. Finally, we refer by $d_G(i)$ the degree of vertex i in G.

▶ **Claim 4.2.** *For any clustering* $\{\mathcal{F}\}$*:* (a) $\sum_{\mathcal{F}} d_{\mathcal{F}}(i) = d_G(i)$*;* (b) $\sum_i \sum_{\mathcal{F}} d_{\mathcal{F}}(i) = 2m = 2|E|$.

Proof. Immediate, because every edge $e \in E$ belongs to exactly one cluster in $\{\mathcal{F}\}$. ◀

Our next claim relates the connection cost of any cluster $\mathcal{F}$ to the structure of the associated subgraph, which forms the crucial part of the analysis.

▶ **Claim 4.3.** *The total connection cost of any cluster* $\mathcal{F}$ *is* $\sum_i d_{\mathcal{F}}(i)(1 - \frac{1}{m_{\mathcal{F}}} d_{\mathcal{F}}(i))$.

Proof. Firstly, note that $\sum_i d_{\mathcal{F}}(i) = 2m_{\mathcal{F}}$. Now consider the center $\mu_{\mathcal{F}}$ of cluster $\mathcal{F}$. By definition, we have that at coordinate $i \in V$:

$$\mu_{\mathcal{F}}(i) = \frac{1}{m_{\mathcal{F}}} \sum_{S \in \mathcal{F} : i \in S} 1 = \frac{d_{\mathcal{F}}(i)}{m_{\mathcal{F}}}.$$

So $\|\mu_{\mathcal{F}}\|^2 = \frac{1}{{m_{\mathcal{F}}}^2} \sum_i d_{\mathcal{F}}(i)^2$. Hence the total cost of this clustering, $c_{\mathcal{F}}$, is:

$$\sum_{e \in \mathcal{F}} (\|\mathrm{x}_e - \mu_{\mathcal{F}}\|^2) = \sum_{e \in \mathcal{F}} (\|\mathrm{x}_e\|^2 - \|\mu_{\mathcal{F}}\|^2) = 2m_{\mathcal{F}} - \frac{1}{m_{\mathcal{F}}} \sum_{i \in V} d_{\mathcal{F}}(i)^2 = \sum_i d_{\mathcal{F}}(i) - \frac{1}{m_{\mathcal{F}}} d_{\mathcal{F}}(i)^2.$$

Here we used $m_{\mathcal{F}} \mu_{\mathcal{F}} = \sum_{e \in \mathcal{F}} \mathrm{x}_e$ in the first equality and $\|\mathrm{x}_e\|^2 = 2$ in the second one. ◀

▶ **Claim 4.4.** *There exists a clustering of our k-means instance $\mathcal{I}_{km}$ with cost at most $m-k$, where m is the number of edges in the graph $G=(V,E)$ associated with the vertex cover instance $\mathcal{I}$, and k is the size of the optimal vertex cover.*

Proof. Consider a cluster $\mathcal{F}_v$, which consists of data points associated with edges covered by a single vertex v. Then, by Claim 4.3, the connection cost of this cluster is precisely $m_{\mathcal{F}_v}-1$, since the sub-graph associated with a cluster is simply a star rooted at v. Here, $m_{\mathcal{F}_v}$ is the number of edges which v covers in the vertex cover (if an edge is covered by different vertices in the cover, it is included in only one vertex). Then, summing over all clusters, we get the claim. ◀

4.3 Soundness

In this section, we show that if there is a clustering of low k-means cost, then there is a very good vertex cover for the corresponding graph. We begin with some useful notation.

▶ **Notation 4.5.** *Given a set $E' \subseteq \binom{V}{2}$ of $m_{E'}=|E'|$ edges with corresponding node degrees $(d_1,\ldots,d_n)$, we define $\mathsf{Cost}(E')$ as the following:*

$$\mathsf{Cost}(E') \overset{\text{def}}{=} \sum_{u\in V} d_u\Big(1-\frac{d_u}{m_{E'}}\Big).$$

Note that, by Claim 4.3, the connection cost of a clustering $\Gamma=\{\mathcal{F}_1,\mathcal{F}_2,\ldots,\mathcal{F}_k\}$ of the n points is equal to $\sum_i \mathsf{Cost}(\mathcal{F}_i)$. Recall that we abuse notation slightly and view each cluster $\mathcal{F}_i$ of the data points also as a subset of E. Moreover, because Γ clusters all points, the subgraphs $\mathcal{F}_1,\mathcal{F}_2,\ldots,\mathcal{F}_k$ form a partition of E. Using this analogy, we study the properties of each subgraph and show that if the k-means cost of Γ is small, then most of these subgraphs in fact are stars. This will in turn help us recover a small vertex cover for G. We begin with a simple property of $\mathsf{Cost}(E')$.

▶ **Proposition 4.6.** *For any set of $m_{E'}$ edges E', $m_{E'}-1 \le \mathsf{Cost}(E') \le 2m_{E'}-1$.*

Proof. We have $\mathsf{Cost}(E')=\sum_{u\in V} d_u\Big(1-\frac{d_u}{m_{E'}}\Big)=2m_{E'}-\frac{\sum_{u\in V} d_u^2}{m_{E'}}$. The proof follows from noting that $\frac{\sum_{u\in V} d_u^2}{m_{E'}} \ge \frac{\sum_{u\in V} d_u}{m_{E'}}=2$ and $\frac{\sum_{u\in V} d_u^2}{m_{E'}} \le m_{E'}+1$. The last inequality is due to the fact that $\sum_{u\in V} d_u^2$ is maximized by the degree sequence $(m_{E'},1,1,\ldots,1)$. ◀

▶ **Theorem 4.7.** *If the k-means instance $\mathcal{I}_{km}$ has a clustering $\Gamma=\{\mathcal{F}_1,\ldots,\mathcal{F}_k\}$ with $\sum_{\mathcal{F}\in\Gamma}\mathsf{Cost}(\mathcal{F}) \le m-(1-\delta)k$, then there exists a $(1+O(\delta))k$-vertex cover of G in the instance $\mathcal{I}$.*

Note that this, along with Claim 4.4 will imply the proof of Theorem 4.1.

Proof. For each $i\in[k]$, let $m_i \overset{\text{def}}{=} |\mathcal{F}_i|$ and $\nu_i \overset{\text{def}}{=} \sum_u d_u(\mathcal{F}_i)^2$. Note that $\mathsf{Cost}(\mathcal{F}_i)=2m_i-\frac{\nu_i}{m_i}$. By Proposition 4.6, each $i\in[k]$ satisfies $m_i-1\le \mathsf{Cost}(\mathcal{F}_i)\le 2m_i-1$. Hence if we define δ_i as $\delta_i \overset{\text{def}}{=} \mathsf{Cost}(\mathcal{F}_i)-(m_i-1)$, then $0\le\delta_i\le m_i$. Moreover $\frac{\nu_i}{m_i}=m_i+1-\delta_i$. Thus:

$$m-(1-\delta)k \ge \sum_i \mathsf{Cost}(\mathcal{F}_i)=\sum_i(\delta_i+m_i-1)=\sum_i\delta_i+m-k \implies \delta k\ge\sum_i\delta_i.$$

This means, except $\le 2\delta k$ clusters, the remaining clusters all have $\delta_i \le \frac{1}{2}$. Moreover, Theorem 4.8 implies all these $(1-2\delta)k$ clusters are either stars or triangles and have $\delta_i=0$.

Since the graph is triangle free, they are all stars, and hence the corresponding center vertices cover all the edges in the respective clusters. It now remains to cover the edges in the remaining $2\delta k$ clusters which have larger δ_i values. Indeed, even for these clusters, we can appeal to Theorem 4.8, and choose *two* vertices per cluster to cover all but δ_i edges in each cluster. So the size of our candidate vertex cover is at most $k(1+2\delta)$, and we have covered all but $\sum_i \delta_i$ edges. But now, we notice that $\sum_i \delta_i \leq \delta k$, and so we can simply include one vertex per uncovered edge and would obtain a vertex cover of size at most $k(1+3\delta)$, thus completing the proof. ◀

▶ **Lemma 4.8.** *Given a graph $G_{\mathcal{F}} = (V, \mathcal{F})$ with $m = |\mathcal{F}|$ edges and degrees $(d_1, \ldots, d_n)$; let δ be such that $\frac{1}{m}\sum_u {d_u}^2 = m + 1 - \delta$. Then there always exists an edge $\{u, v\} \in \mathcal{F}$ with $d_u + d_v \geq m + 1 + \delta$. Furthermore, if $\delta < \frac{1}{2}$, then $\delta = 0$ and $G_{\mathcal{F}}$ is either a star or triangle.*

Proof. Since $\sum_u {d_u}^2 = \sum_{u\sim v}(d_u + d_v)$, we can think of $\frac{1}{m}\sum_u {d_u}^2$ as the the expectation of $d_u + d_v$ over a random edge chosen uniformly, $\{u, v\} \in E$:

$$\frac{1}{m}\sum_u {d_u}^2 = \mathbb{E}_{u\sim v}\big[d_u + d_v\big].$$

From this, we can immediately conclude the existence of an edge $\{u, v\}$ with $d_u + d_v \geq m+1-\delta$. Now to complete the second part of the Lemma statement, suppose $d_u \geq d_v$. The number of edges incident to $\{u, v\}$ is:

$$d_u + d_v - 1 \geq m - \delta \overset{\delta<1}{\Longrightarrow} d_u + d_v - 1 = m.$$

So all edges are incident to u or v, and $d_w \leq 2$ if $w \notin \{u, v\}$. If $d_v \leq 1$, then we are done. In the other case, we have $d_v \geq 2 \geq d_w$ for all $w \notin \{u, v\}$. Let $\alpha \overset{\text{def}}{=} d_u$ and $\beta \overset{\text{def}}{=} d_v$. The degree sequence $(d_1, \ldots, d_n)$ is strongly majorized by the following sequence, d':

$$d' \overset{\text{def}}{=} \Big(\alpha, \beta, \underbrace{2, \ldots, 2,}_{\beta-1 \text{ many}} \overbrace{1, \ldots, 1}^{\alpha-\beta \text{ many}}\Big).$$

Since $\sum_u d_u^2$ is Schur-convex, its value increases under majorization:

$$\begin{aligned}(\alpha+\beta-1)(\alpha+\beta-\delta) =& m(m+1-\delta) = \sum_u d_u^2 \leq \sum_u {d'_u}^2 \\ =& \alpha^2 + \beta^2 + 4(\beta-1) + (\alpha-\beta). \\ \Longrightarrow\ 0 \leq& (\alpha+\beta-1)\delta + 2\alpha + 4\beta - 4 - 2\alpha\beta \\ =& (\alpha+\beta-1)\delta + 2\alpha(1-\beta) + 4(\beta-1).\end{aligned}$$

So we obtain $2\alpha(\beta-1) \leq (\alpha+\beta-1)\delta + 4(\beta-1)$. Since $\beta \geq 2$, we divide both sides by $\beta - 1$:

$$2\alpha \leq \frac{\alpha}{\beta-1}\delta + 4 + \delta \leq \delta\alpha + 4 + \delta.$$

In particular, $(2-\delta)\alpha \leq 4 + \delta \implies \alpha \leq \frac{4+\delta}{2-\delta} < 3$ as $\delta < 1/2$. Hence $\alpha \leq 2$. Consequently, $d_u = d_v = 2$ and $m = d_u + d_v - 1 = 3$. There are two possible cases: The graph is either a 3-cycle or 4-path. In the latter case, the corresponding δ is:

$$\delta = m + 1 - \frac{1}{m}\sum_u {d_u}^2 = 4 - \frac{1}{3}(2^2 + 2^2 + 1 + 1) = 4 - \frac{10}{3} = \frac{2}{3} > \frac{1}{2};$$

which is a contradiction and the graph is a triangle. ◀

Putting the pieces together, we get the proof of Theorem 4.1.

▶ Remark (Unique Games Hardness). Khot and Regev [22] show that approximating Vertex-Cover to factor $(2-\epsilon)$ is hard assuming the Unique Games conjecture. Furthermore, Kortsarz et al. [24] show that any approximation algorithm with ratio $\alpha \geq 1.5$ for Vertex-Cover on 3-cycle-free graphs implies an α approximation algorithm for Vertex-Cover (on general graphs). This result combined with the reduction in this section immediately implies APX hardness for k-means under the unique games conjecture. In the next section we generalize the result of Kortsarz et al. [24] by giving an approximation preserving reduction from Vertex-Cover on general graphs to Vertex-Cover on triangle-free graphs. This would enable us to get APX hardness for the k-means problem.

5 Hardness of Vertex Cover on Triangle-Free Graphs

In this section, we show that the Vertex Cover problem is as hard on triangle-free graphs as it is on general graphs. To this end, for any graph $G = (V, E)$, we define $\mathrm{IS}(G)$ as the size of maximum independent set in G. For convenience, we define rel-$\mathrm{IS}(G)$ as the ratio of $\mathrm{IS}(G)$ to the number of nodes in G:

$$\text{rel-IS}(G) \overset{\text{def}}{=} \frac{\mathrm{IS}(G)}{|V|}.$$

Similarly, let $\mathrm{VC}(G)$ be the size of minimum vertex cover in G and rel-$\mathrm{VC}(G)$ be the ratio $\frac{\mathrm{VC}(G)}{|V|}$. The following is well known, which says independent sets and vertex covers are duals of each other.

▶ **Proposition 5.1.** *Given $G = (V, E)$, $I \subseteq V$ is an independent set if and only if $C = V \setminus I$ is a vertex cover. In particular,* $\mathrm{IS}(G) + \mathrm{VC}(G) = |V|$.

We will prove the following theorem.

▶ **Theorem 5.2.** *For any constant $\varepsilon > 0$, there is a $(1+\varepsilon)$-approximation-preserving reduction for independent set from any graph $G = (V, E)$ with maximum degree Δ to triangle-free graphs with* $\mathrm{poly}(\Delta, \varepsilon^{-1})|V|$ *nodes and degree* $\mathrm{poly}(\Delta, \varepsilon^{-1})$ *in deterministic polynomial time.*

Combining Theorem 5.2 with the best known unconditional hardness result for Vertex Cover, due to Dinur and Safra [14], we obtain the following corollary.

▶ **Corollary 5.3.** *Given any unweighted triangle-free graph G with bounded degrees, it is NP-hard to approximate Vertex Cover within any factor smaller than* 1.36.

Given two simple graphs $G = (V_1, E_1)$ and $H = (V_2, H_2)$, we define the Kronecker product of G and H, $G \otimes H$, as the graph with nodes $V(G \otimes H) = V_1 \times V_2$ and edges:

$$E(G \otimes H) = \Big\{ \{(u, i), (v, j)\} \Big| \{u, v\} \in E(G),\ \{i, j\} \in E(H) \Big\}.$$

Observe that, if A_G and A_H denote the adjacency matrix of G and H, then $A_{G \otimes H} = A_G \otimes A_H$.

Given any symmetric matrix M, we will use $\sigma_i(M)$ to denote the i^{th} largest eigenvalue of M. For any graph G on n-nodes, we define the spectral radius of G, $\rho(G)$, as the following:

$$\rho(G) \overset{\text{def}}{=} \max_{p \perp \mathbf{e}} \frac{|p^T A_G p|}{\|p\|^2} = \max(\sigma_2(A_G), |\sigma_n(A_G)|).$$

Here $\mathbf{e}$ is the all 1's vector of length n.

▶ **Proposition 5.4.** *If H is triangle-free, then so is $G \otimes H$.*

Proof. Suppose $G \otimes H$ has a 3-cycle of the form $((a,i),(b,j),(c,k),(a,i))$. Then (i,j,k,i) is a closed walk in H. H is triangle-free, therefore $i = j$ wlog; a contradiction as H has no loops. ◀

The following Lemma says that as long as H has good spectral properties, the relative size of maximum independent sets in G will be preserved by $G \otimes H$.

▶ **Lemma 5.5.** *Suppose H is a d-regular graph with spectral radius $\leq \rho$. For any graph G with maximum degree Δ,* $\text{rel-IS}(G \otimes H) \geq \text{rel-IS}(G) \geq \left(1 - \frac{\rho\Delta}{2d}\right) \text{rel-IS}(G \otimes H)$.

Proof. Suppose $V(G) = [n]$ and $V(H) = [N]$. Let $A \overset{\text{def}}{=} A_G$ be the adjacency matrix of G and B be the normalized adjacency matrix of H, $B \overset{\text{def}}{=} \frac{1}{d}A_H$.

For the lower bound, consider an independent set I in G. It is easy to check that $I \times [N]$ is an independent set in $G \otimes H$, thus $\text{IS}(G \otimes H) \geq N \cdot \text{IS}(G)$ so $\text{rel-IS}(G \otimes H) \geq \text{rel-IS}(G)$.

For the upper bound, consider the indicator vector $f \in \{0,1\}^{[n]\times[N]}$ of an independent set in $G \otimes H$. The corresponding set contains no edges from $G \otimes H$, so $f^T(A \otimes B)f = 0$. Define $p : V \to [0,1]$ as $p_u \overset{\text{def}}{=} \frac{1}{N}\sum_{j\in N} f_{u,j}$. For each $u \in [n]$, pick u with probability p_u. Let $I_0 \subseteq [n]$ be the set of picked nodes. Next, start with $I \leftarrow I_0$. As long as there is an edge of G contained in I, arbitrarily remove one of its endpoints from I. At the end of this process, the remaining set I is an independent set in G, and its size is at least the size of I_0 minus the number of edges contained in I_0. Hence $|I| \geq |I_0| - |E_G(I_0, I_0)|$. Observe that

$$\mathbb{E}\left[|I_0|\right] = \sum_u p_u = \frac{1}{N}\|f\|^2 \quad \text{since } f \text{ is a } \{0,1\} \text{ vector.}$$

The probability of any pair $i \neq j$ being contained in I_0 is given by $\text{Prob}\left[\{i,j\} \subseteq I_0\right] = p_u p_v$. Therefore, the expected number of edges contained in I_0 is:

$$\begin{aligned}\mathbb{E}\left[|E_G(I_0,I_0)|\right] &= \sum_{u<v} A_{uv}p_u p_v = \frac{1}{2}p^T A p \overset{(Theorem\ 5.6)}{=} \frac{1}{2N} f^T(A \otimes \widetilde{J_N})f \\ &\overset{(Theorem\ 5.7)}{\leq} \frac{1}{2N}\left|f^T A \otimes (\widetilde{J_N} - B)f\right| \overset{(Theorem\ 5.8)}{\leq} \frac{\rho\Delta}{2Nd}\|f\|^2.\end{aligned}$$

Putting it all together:

$$\mathbb{E}\left[|I|\right] \geq \mathbb{E}\left[|I_0|\right] - \mathbb{E}\left[|E_G(I_0,I_0)|\right] = \frac{1}{N}\|f\|^2 - \frac{1}{2}p^T A p \geq \frac{\|f\|^2}{N}\left(1 - \frac{\rho\Delta}{2d}\right).$$

Therefore, $\text{IS}(G) \geq \frac{1-\frac{\rho\Delta}{2d}}{N}\,\text{IS}(G \otimes H) \implies \text{rel-IS}(G) \geq \left(1 - \frac{\rho\Delta}{2d}\right)\text{rel-IS}(G \otimes H)$. ◀

In the remaining part, we prove the supporting claims.

▶ **Claim 5.6.** $p^T A p = \frac{1}{N} f^T(A \otimes \widetilde{J_N})f$ *where $\widetilde{J_N}$ is the N-by-N matrix of all $1/N$'s.*

Proof. Let $\mathbf{e}^{u,v} \in \mathbb{R}^{V\times V}$ be the matrix whose entry at u^{th} row and v^{th} column is 1, and all others 0. Notice $A = \sum_{u,v} A_{u,v}\mathbf{e}^{u,v}$. Let J_N be the N-by-N matrix of all 1's. For any pair $(u,v) \in V^2$,

$$\begin{aligned}p_u p_v =& \frac{1}{N^2}\sum_{i,j} f_{u,i}f_{v,j} = \frac{1}{N^2}f^T\left(\mathbf{e}^{u,v} \otimes J_N\right)f = \frac{1}{N}f^T\left(\mathbf{e}^{u,v} \otimes \widetilde{J_N}\right)f. \\ p^T A p =& \frac{1}{N}\sum_{u,v} A_{u,v} f^T\left(\mathbf{e}^{u,v} \otimes \widetilde{J_N}\right)f = \frac{1}{N}f^T\left[\left(\sum_{u,v} A_{u,v}\mathbf{e}^{u,v}\right) \otimes \widetilde{J_N}\right]f \\ =& \frac{1}{N}f^T(A \otimes \widetilde{J_N})f.\end{aligned}$$

The second-to-last identity follows from the bi-linearity of Kronecker product. ◀

▶ **Claim 5.7.** $f^T(A \otimes \widetilde{J_N})f \leq |f^T A \otimes (B - \widetilde{J_N})f|$.

Proof. We have $f^T(A \otimes \widetilde{J_N})f = f^T\Big[A \otimes (\widetilde{J_N} - B)\Big]f + f^T(A \otimes B)f$. As noted above, f being an independent set implies $f^T(A \otimes B)f = 0$:

$$f^T(A \otimes \widetilde{J_N})f = f^T\Big[A \otimes (\widetilde{J_N} - B)\Big]f \leq |f^T A \otimes (\widetilde{J_N} - B)f|.$$ ◀

▶ **Claim 5.8.** $|f^T A \otimes (B - \widetilde{J_N})f| \leq \frac{\Delta\rho}{d}\|f\|^2$.

Proof. Define $C \stackrel{\text{def}}{=} B - \widetilde{J_N}$. For any symmetric matrix M, let $\rho(M)$ be its spectral radius, $\rho(M) \stackrel{\text{def}}{=} \max_p \frac{|p^T M p|}{\|p\|^2}$. Observe that $\rho(M) = \max(|\sigma_i(M)|)$. We have:

$$|f^T A \otimes (B - \widetilde{J_N})f| = |f^T A \otimes Cf| \leq \rho(A \otimes B)\|f\|_2^2.$$

We know that the spectrum of the Kronecker product of two symmetric matrices correspond to the pairwise product of the spectrum of corresponding matrices, i.e., all eigenvalues of $A \otimes C$ are of the form $\sigma_i(A) \cdot \sigma_j(C)$ for each i and j. Therefore,

$$\rho(A \otimes C) = \max(|\sigma_i(A)\sigma_j(C)|) = \max(|\sigma_i(A)|)\max(|\sigma_j(C)|) = \rho(A) \cdot \rho(C).$$

Observe that $\rho(A) \leq \Delta$, since A is the adjacency matrix of a graph with degree $\leq \Delta$. Now we will upper bound $\rho(C)$. Since H is a regular graph and B is its normalized adjacency matrix, the largest eigenvector of B is all 1's and the corresponding eigenvalue is 1. Therefore C has the same eigenspace with B. Moreover $C\mathbf{e} = 0$, thus:

$$\begin{aligned}\rho(C) &= \max(|\sigma_i(C)| \ : \ 1 \leq i \leq n) = \max(|\sigma_i(B)| \ : \ 2 \leq i \leq n)\\ &= \max(\sigma_2(B), |\sigma_n(B)|) = \frac{1}{d}\rho(G).\end{aligned}$$ ◀

We now prove the main theorem needed for our reduction.

▶ **Theorem 5.9.** *Given a graph $G = (V, E)$ with maximum degree Δ, for any $\varepsilon > 0$, we can construct in polynomial time, a triangle-free graph $\widehat{G} = (\widehat{V}, \widehat{E})$ with:*

$$\text{rel-IS}(G) \leq \text{rel-IS}(\widehat{G}) \leq (1 + \varepsilon)\,\text{rel-IS}(G).$$

Moreover $\widehat{G}$ has (a) $\text{poly}(\Delta, \varepsilon^{-1})|V|$ *nodes,* (b) *degree* $O\big(\Delta^3\varepsilon^{-2}\big)$.

Proof. For any d and N, it is known how to construct [29, 32] in deterministic polynomial time, a $O(d)$-regular Ramanujan graph H with girth $\Omega(\log_d N)$ and spectral radius at most $\rho \leq O(\sqrt{d})$. Thus for some choice of $d = O(\Delta^2\varepsilon^{-2})$ and $N = d^{O(1)} = \text{poly}(\Delta, \varepsilon^{-1})$, we can find a d-regular graph H with girth at least $\Omega(1)$ and spectral radius $\rho \leq d\varepsilon/\Delta$. For such H, let $\widehat{G} \leftarrow G \otimes H$. We have $\left(1 - \frac{\rho\Delta}{2d}\right)^{-1} \leq \big(1 - \varepsilon/2\big)^{-1} \leq 1 + \varepsilon$. Proposition 5.4 implies $G \otimes H$ is triangle free. By Theorem 5.5:

$$\text{rel-IS}(G) \leq \text{rel-IS}(G \otimes H) \leq \left(1 - \frac{\rho\Delta}{2d}\right)^{-1}\text{rel-IS}(G) \leq \Big(1 + \varepsilon\Big)\,\text{rel-IS}(G).$$

Now we prove the remaining properties:
(a) $|V(G \otimes H)| = |V(G)| \cdot |V(H)| \leq |V| \cdot \text{poly}(\Delta, \varepsilon^{-1})$.
(b) $d_{\max}(G \otimes H) \leq d_{\max}(G) \times d_{\max}(H) \leq O(\Delta d) = O(\Delta^3\varepsilon^{-2})$. ◀

▶ **Note.** *Noga Alon has provided an alternate construction where one can obtain a triangle free graph $\hat{G}$ such that* $\text{rel-IS}(\hat{G}) = \text{rel-IS}(G)$. *This however, does not lead to improved constant in our analysis. For the sake of completeness, we include the alternate theorem in the Appendix (see Theorem A.1).*

Before we end the section with the proof of Theorem 5.3, we need the following hardness result from [14], which follows from Corollary 2.3 and Appendix 8 (weighted to unweighted reduction) of [14]. As noted in [14], the construction produces bounded degree graphs.

▶ **Theorem 5.10** (Dinur, Safra [14]). *For any constant $\varepsilon > 0$, given any unweighted graph G with bounded degrees, it is NP-hard to distinguish between:*

- (Yes) $\text{rel-IS}(G) > c - \varepsilon$,
- (No) $\text{rel-IS}(G) < s + \varepsilon$;

where c and s are constants such that $\frac{1-s}{1-c} \approx 1.36$.

Proof of Theorem 5.3. Given a bounded degree graph G, consider the graph $\widehat{G}$ given by Theorem 5.9 for some small constant $\varepsilon_0 < \varepsilon$. Since G is bounded degree and ε_0 is constant, $\widehat{G}$ is also bounded degree. Furthermore, $\widehat{G}$ satisfies $\text{rel-IS}(G) \leq \text{rel-IS}(\widehat{G}) \leq (1+\varepsilon_0)\,\text{rel-IS}(G)$. Completeness follows immediately: $\text{rel-IS}(\widehat{G}) > c-\varepsilon$. For the soundness, suppose $\text{rel-IS}(\widehat{G}) > s+\varepsilon$. Then $\text{rel-IS}(G) \geq \frac{s+\varepsilon}{1+\varepsilon_0} \geq s+\varepsilon$ for suitable ε_0. The hardness of Vertex Cover follows from Proposition 5.1. ◀

6 Conclusions

In this paper we provide the first hardness of approximation for the fundamental Euclidean k-means problem. Although our work clears a major hurdle of going beyond NP-hardness for this problem, there is still a big gap in our understanding with the best upper bound being a factor $(9+\epsilon)$. We believe that our result and techniques will pave way for further work in closing this gap. Our reduction from vertex cover produces high dimensional instances ($d = \Omega(n)$) of k-means. However, by using the Johnson-Lindenstrauss transform [12], we can project the instance onto $O(\log n/\epsilon^2)$ dimensions and still preserve pairwise distances by a factor $(1+\epsilon)$ and the k-means cost by a factor of $(1+\epsilon)^2$. We leave it as an open question to investigate inapproximability results for k-means in constant dimensions. It would also be interesting to study whether our techniques give hardness of approximation results for the Euclidean k-median problem. Finally, our hardness reduction in Section 5 provides a novel analysis by using the spectral properties of the underlying graph to argue about independent sets in graph products – this connection could have applications beyond the present paper.

Acknowledgments. We would like to thank Noga Alon and Oded Regev for valuable feedback on the results in Section 5, in particular for suggesting alternate proofs of Proposition 5.4 and Theorem 5.5. We would also like to thank Noga for pointing out that the graph product construction in Section 5 does not eliminate even cycles. Finally we thank the anonymous reviewers for their comments.

References

1 Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

2 Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M. Roth. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Transactions on Information Theory*, 38(2):509–516, 1992.

3 Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley, 1992.

4 Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for Euclidean k-medians and related problems. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*, pages 106–113, 1998.

5 David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035, 2007.

6 Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

7 Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 250–257, 2002.

8 Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2009, New York, NY, USA, January 4–6, 2009*, pages 1068–1077, 2009.

9 Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. *CoRR*, abs/1406.2951, 2014.

10 Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem. *J. Comput. Syst. Sci.*, 65(1):129–149, 2002.

11 Sanjoy Dasgupta. The hardness of k-means clustering. Technical report, University of California, San Diego, 2008.

12 Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

13 Wenceslas Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9–11, 2003, San Diego, CA, USA*, pages 50–58, 2003.

14 Irit Dinur and Samuel Safra. On the hardness of approximating minimum vertex cover. *Annals of Mathematics*, 162(1):439–485, 2005.

15 Tomás Feder and Daniel H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing, May 2–4, 1988, Chicago, Illinois, USA*, pages 434–444, 1988.

16 Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for k-means clustering based on weak coresets. In *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6–8, 2007*, pages 11–18, 2007.

17 Venkatesan Guruswami and Piotr Indyk. Embeddings and non-approximability of geometric problems. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12–14, 2003, Baltimore, Maryland, USA.*, pages 537–538, 2003.

18 Dorit S. Hochbaum and David B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. *J. ACM*, 33(3):533–550, 1986.

19 Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19–21, 2002, Montréal, Québec, Canada*, pages 731–740, 2002.

20 Kamal Jain and Vijay V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *J. ACM*, 48(2):274–296, 2001.

21 Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Comput. Geom.*, 28(2-3):89–112, 2004.

22 Subhash Khot and Oded Regev. Vertex cover might be hard to approximate to within $2-\varepsilon$. *Journal of Computer and System Sciences*, 74(3):335–349, 2008.

23 Stavros G. Kolliopoulos and Satish Rao. A nearly linear-time approximation scheme for the Euclidean k-median problem. *SIAM J. Comput.*, 37(3):757–782, 2007.

24 Guy Kortsarz, Michael Langberg, and Zeev Nutov. Approximating maximum subgraphs without short cycles. *SIAM J. Discrete Math.*, 24(1):255–269, 2010.

25 Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23–26, 2010, Las Vegas, Nevada, USA*, pages 299–308, 2010.

26 Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Symposium on Foundations of Computer Science (FOCS 2004), 17–19 October 2004, Rome, Italy, Proceedings*, pages 454–462, 2004.

27 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*, pages 901–910, 2013.

28 Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982.

29 Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

30 Meena Mahajan, Prajakta Nimbhorkar, and Kasturi R. Varadarajan. The planar k-means problem is NP-hard. *Theor. Comput. Sci.*, 442:13–21, 2012.

31 Jiri Matoušek. On approximate geometric k-clustering. *Discrete and Computational Geometry*, 24(1), 2000.

32 Moshe Morgenstern. Existence and explicit constructions of $q+1$ regular ramanujan graphs for every prime power q. *J. Comb. Theory, Ser. B*, 62(1):44–62, 1994.

33 Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28, 2012.

34 Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus F. M. Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37, 2008.

A An Alternative Proof of Theorem 5.2

The following were suggested by Noga Alon as an alternative to Theorem 5.2.

▶ **Theorem A.1.** *Let $G=(V,E)$ be an arbitrary graph with maximum degree Δ. It is possible to construct in polynomial time a triangle free graph $\hat{G}$ such that* $\text{rel-IS}(\hat{G}) = \text{rel-IS}(G)$.

Before proving the theorem, we need the following standard facts about (n,d,λ) graphs.

▶ **Lemma A.2.** *Let $H=(U,F)$ be an (n,d,λ) graph, assume $\lambda < d/4$ and let B be a set of vertices of H. Let $N(B)$ denote the set of all neighbors of B in H. Then:*

1. *If $|B| > \frac{\lambda}{d}n$, then $|N(B)| > n - \frac{\lambda}{d}n$.*
2. *If $|B| \leq \frac{\lambda}{d}n$, then $|N(B)| \geq \frac{\lambda}{2d}n$.*

Proof. (1) is proved in Corollary 1 in [2]. We will prove (2). When $\frac{2\lambda^2}{d^2}n \leq |B| \leq \frac{\lambda}{d}n$, it follows from the same corollary again, which implies that in this range $|N(B)| \geq \frac{n}{2}$. For $|B| \leq \frac{2\lambda^2}{d^2}n$, the result follows from the expander mixing lemma (see [3], corollary 9.2.5), as there are $d|B|$ edges between B and $N(B)$. ◀

Proof of Theorem A.1. Let $H = (U, F)$ be a (n, d, λ)-expander with $\lambda \leq 2\sqrt{d-1}$[1] Let $\hat{G} = G \otimes H$. Further, let $\frac{d}{2\lambda} \geq \Delta$. It is well known that such graphs exist. It is easy to see that any rel-IS$(G \otimes H) \geq$ rel-IS(G), since any independent set S in G leads to an independent set $S \otimes U$ in $G \otimes H$.

For the other direction, let $S \subset V \times U$ be an independent set in $G \otimes H$. Define T as:

$$T \overset{\text{def}}{=} \{v \in V : |\{u \in U : (v, u) \in S\}| \geq \frac{\lambda}{d}n\}.$$

By Lemma A.2 (1), T is an independent set in G. Let T' be a maximal (with respect to containment) independent set in G that contains T. By maximality, every vertex in $V \setminus T'$ has at least one neighbor in T'. Thus T' is a dominating set in G and there is a collection of stars $\{S_v : v \in T'\}$, covering all the vertices of G. As T' is an independent set, $|T'| \leq$ rel-IS$(G)|V|$. To complete the proof it suffices to show that for each of the stars S_v in our collection, whose set of vertices in G is V_v, we have:

$$|\{(v', u) : (v', u) \in S, v' \in V_v\}| \leq |U| = n \tag{1}$$

The number of leaves of the star S_v is at most Δ. For each such leaf v', the set of vertices of H given by $B_{v'} \overset{\text{def}}{=} \{u \in U : (v', u) \in S\}$ is of cardinality smaller than $\frac{\lambda}{d}n$. Moreover, all its neighbors in H cannot belong to the set $B_v = \{u \in U : (v, u) \in S\}$ where v is the center of the star S_v. By Lemma A.2 (2), the number of these neighbors is at least $\frac{d}{2\lambda} \geq \Delta$ times the cardinality of $B_{v'}$. This implies that the total size of all sets $B_{v'}$ where the sum ranges over all leaves v' of S_v is at most the number of vertices in $U - B_v$, implying (1) and completing the proof. ◀

[1] This means that all non-trivial eigenvalues of H are bounded by λ.