Finding the Jaccard Median

Flavio Chierichetti* Ravi Kumar† Sandeep Pandey† Sergei Vassilvitskii†

Abstract

The median problem in the weighted Jaccard metric was analyzed by Späth in 1981. Up until now, only an exponential-time exact algorithm was known. We (a) obtain a PTAS for the weighted Jaccard median problem and (b) show that the problem does not admit a FPTAS (assuming P $\neq$ NP), even when restricted to binary vectors. The PTAS is built on a number of different algorithmic ideas and the hardness result makes use of an especially interesting gadget.

1 Introduction

A widely used set similarity measure is the Jaccard coefficient, introduced more than a century ago [14]. For two sets X, Y, it is defined to be $J(X,Y) = |X \cap Y|/|X \cup Y|$. The Jaccard distance between the sets, defined as $D(X,Y) = 1 - J(X,Y)$, is known to be a metric. A natural generalization of Jaccard similarity, independently proposed several times over many years [7, 11, 12, 15, 16, 18, 24, 25], is to consider n-dimensional non-negative vectors X, Y and define $J(X,Y) = \frac{\sum_{i=1}^{n} \min(X_i,Y_i)}{\sum_{i=1}^{n} \max(X_i,Y_i)}$; the weighted Jaccard distance, $D(X,Y) = 1 - J(X,Y)$, still remains a metric. In this paper we study the computational complexity of the median problem in the Jaccard distance metric, namely, given a family $\mathcal{S}$ of input sets (or vectors), find a set (vector) M^* that minimizes $\sum_{X \in \mathcal{S}} D(M^*, X)$.

The use of the Jaccard metric and Jaccard median is common in many scientific fields: biology [17], botany [13], cognitive sciences [20], ecology [23], geology [24], natural language processing [7, 11, 15, 16], paleontology [22, 24], psychology [12, 26], web sciences [3, 21], and so on. In the field of computer science, Broder et al. [3, 4] introduced "shingles" and min-wise independent permutations for sketching the Jaccard distance; the sets in their case were the web documents, viewed as a bag of words. Charikar [5] gave a way of sketching arbitrary non-negative vectors in a way that preserves their weighted Jaccard distance.

*Work done in part while visiting Yahoo! Research. Supported in part by a grant from Yahoo! Research. Dipartimento di Informatica, Sapienza University of Rome, Italy. Email: chierichetti@di.uniroma.it

†Yahoo! Research, Sunnyvale, CA, USA. Email: {ravikumar,spandey,sergei}@yahoo-inc.com

The Jaccard median problem itself was studied more than two decades ago. Späth [25] showed a "canonical" structural property of the optimal Jaccard median: for each coordinate, its value has to agree with that of some input. This makes the search space finite, albeit exponential ($|\mathcal{S}|^n$). Watson [27] gave a vertex-descent algorithm for Jaccard median and showed that his algorithm terminates and always returns an optimal median. Unfortunately, he did not show any bounds on its running time. Nothing substantial, other than these two pieces of work, is known about the complexity of finding or approximating the Jaccard median.

Our results. In this paper we fully study the computational complexity of the weighted Jaccard median problem. Our main result is a PTAS for the weighted Jaccard median problem. While it is trivial to obtain a two-approximation for the problem (the best of the input vectors achieves this approximation and this bound is tight, see Appendix A), obtaining a $(1+\epsilon)$-approximation turns out to require new ideas, in particular, understanding the structure of the optimal solution.

We first show how to find a $(1+\epsilon)$-approximate median for the binary (i.e., set) version of the Jaccard metric. This is done by combining two algorithms. The first algorithm uses random projections on a carefully selected subspace and outputs an additive approximation; the quality translates to a multiplicative approximation provided the optimum is a large set. The second algorithm focuses on the case when the optimum is a small set and obtains a multiplicative approximation — this algorithm leverages certain structural properties of an optimal solution.

To obtain a PTAS for the weighted Jaccard median problem, we consider three different cases. If the value of the optimum is very small ($O(\epsilon)$), then we show how the Jaccard median problem can be "linearized" and give a PTAS based on linear programming. If the value of the optimum is $\Omega(\epsilon)$, then there are two sub-cases. If the ratio between the maximum and the minimum coordinate values is polynomial, then we map the input instance to a polynomially-sized binary instance, solve it using the PTAS for the binary case, and show how this approximate solution can be mapped back to an

approximate solution to the original instance. If the ratio of the maximum and the minimum coordinate values is super-polynomial, then we show how one can modify the instance so as to guarantee that the ratio becomes polynomial and show that each approximate solution to the modified instance is also an approximate solution to the original instance.

We then show that the binary Jaccard median problem is NP-hard. Interestingly, our proof shows that the problem remains NP-hard even in the following two special cases: (a) when the input sets are not allowed to be repeated (i.e., $\mathcal{S}$ cannot be a multi-set) and (b) when all the input sets consists of exactly two elements (i.e., $|X| = 2, \forall X \in \mathcal{S}$) but the sets themselves are allowed to be repeated (i.e., $\mathcal{S}$ can be a multi-set). Our proofs in fact show that unless P = NP, there can be no FPTAS for finding the Jaccard median.

Related work. The median problem has been actively studied for many different metric spaces. The hardness of finding the best median for a set of points out of a (typically exponentially) large set of candidates strictly depends on the metric in consideration. For instance, the median problem has been shown to be hard for edit distance on strings [8, 19], for the Kendall τ metric on permutations [2, 9], but can be solved in polynomial time for the Hamming distance on sets (and more generally, for the ℓ_1 distance on real vectors), and for the Spearman footrule metric on permutations [9]. The general metric k-median problem has also been studied in the literature; see, for example, [1, 6].

2 Preliminaries

Let $U = \{x_1, \ldots, x_n\}$ be the ground set.

Definition 2.1. (Binary Jaccard measures) *Given $X, Y \subseteq U$, the* Jaccard similarity *is defined as*

$$J(X,Y) = \begin{cases} \frac{|X \cap Y|}{|X \cup Y|} & \text{if } X \cup Y \neq \emptyset, \\ 1 & \text{if } X \cup Y = \emptyset, \end{cases}$$

and the Jaccard distance *is defined as $D(X,Y) = 1 - J(X,Y)$.*

It is known that $D(X,Y)$ is a metric; see, for instance, [5]. Let $S_1, \ldots, S_m$ be (not necessarily distinct) subsets of U, and let $\mathcal{S} = \{S_1, \ldots, S_m\}$; let $X \subseteq U$. We define the Jaccard similarity between X and $\mathcal{S}$ to be $J(X,\mathcal{S}) = \sum_{Y \in \mathcal{S}} J(X,Y)$. If $X \neq \emptyset$, we have

$$J(X,\mathcal{S}) = \sum_{Y \in \mathcal{S}} \frac{|X \cap Y|}{|X \cup Y|} = \sum_{x \in X} \sum_{Y \ni x} \frac{1}{|Y \cup X|}.$$

The Jaccard distance of X to $\mathcal{S}$ is defined to be $D(X,\mathcal{S}) = \sum_{Y \in \mathcal{S}} D(X,Y) = |\mathcal{S}| - J(X,\mathcal{S})$.

Definition 2.2. (Jaccard distance median) *For a given $\mathcal{S}$, $M^* \subseteq U$ is said to be an* optimal Jaccard distance (1-)median *if $D(M^*,\mathcal{S}) = \min_{X \subseteq U} D(X,\mathcal{S})$. For $\alpha \geq 1$, $M \subseteq U$ is said to be an* α-approximate Jaccard distance (1-)median*, if $D(M^*,\mathcal{S}) \leq D(M,\mathcal{S}) \leq \alpha D(M^*,\mathcal{S})$.*

Likewise, a median problem with respect to maximizing the Jaccard similarity can be defined (observe, though, that an approximation to the Jaccard distance median need not be an approximation to the Jaccard similarity median, and vice versa). Unless otherwise specified, we use Jaccard median to denote the Jaccard distance median problem. We assume throughout the paper that $\emptyset \notin \mathcal{S}$. The case $\emptyset \in \mathcal{S}$ is easy — just check the value of $\emptyset$ as a candidate Jaccard median, remove $\emptyset$ from $\mathcal{S}$, solve for the remaining sets, and then return the best solution.

For an element $x \in U$, we will refer to the number of sets in which it is present in $\mathcal{S}$ as its *degree*. Thus, $\deg_{\mathcal{S}}(x) = |\{S \in \mathcal{S} : x \in S\}|$. When it is clear from the context, we will simply write $\deg(x)$.

The Jaccard measures can be generalized to non-negative real vectors (sets being binary vectors); the corresponding Jaccard distance is also known to be a metric [5].

Definition 2.3. (Weighted Jaccard measures) *Given two non-negative n-dimensional real vectors X, Y, their* Jaccard similarity *is defined as*

$$J(X,Y) = \begin{cases} \frac{\sum_{i=1}^n \min(X_i,Y_i)}{\sum_{i=1}^n \max(X_i,Y_i)} & \text{if } \sum_{i=1}^n \max(X_i,Y_i) > 0, \\ 1 & \text{if } \sum_{i=1}^n \max(X_i,Y_i) = 0, \end{cases}$$

and the Jaccard distance *is defined as $D(X,Y) = 1 - J(X,Y)$.*

The weighted Jaccard median problems can be defined as before.

3 A PTAS for the binary Jaccard median

First, we consider the binary Jaccard median problem. Here, we split the analysis based on the quality of the (yet) unknown optimal median. First, suppose the optimal median is large, say, $\Omega(\epsilon m)$. In this case we obtain an algorithm (Section 3.1) that returns an additive $O(\epsilon^2 m)$-approximation to the optimal median; clearly, this additive approximation translates to a $(1+O(\epsilon))$-multiplicative approximation. Next, we obtain an algorithm (Section 3.2) that returns a $(1+O(\epsilon))$-multiplicative approximation, assuming the optimal median has value $O(\epsilon^2 m)$. Thus, by running the two algorithms in tandem, and returning the better solution, we are guaranteed to have a PTAS.

3.1 A PTAS when the optimal median is large

In this section we show how to obtain an additive $O(\epsilon m)$-approximation in time $(nm)^{\frac{1}{\epsilon^{O(1)}}}$. As stated before, when the optimal median is $\Omega(\epsilon m)$, this immediately gives a PTAS.

This algorithm first guesses the number of elements in the optimal median, and then proceeds to "densify" the instance by removing the sets whose sizes are too far away from the size of the optimal median and removing those elements that are not present in too many sets. Intuitively, these steps will be justified since the sets whose sizes are too far away from the optimal will always be far, regardless of the actual choice of median and removing elements that appear in a small number of sets will not affect the solution too much.

If the dense instance has too many elements, we sub-sample further in order to reduce the total number of elements to at most $O(\log(nm)/\epsilon^6)$. At this point we can afford to try all of the possible subsets, to find a solution M_c, which we call the *seed median*, which will be optimal on this restricted space. Finally, we show how to generalize the seed median to the full space of dense elements by solving a linear program and then rounding it randomly.

The flow of the algorithm is presented below.

1. Guess t, the size of the optimal median M^*.
2. Densify the instance by considering only the set family: $\mathcal{S}_t = \{S_j \in \mathcal{S} \mid \epsilon t \leq |S_j| \leq \frac{t}{\epsilon}\}$. Keep only the elements U_t present in at least ϵm sets in $\mathcal{S}_t$.
3. If $|U_t| \leq 9\epsilon^{-6}\ln(nm)$, then try all subsets of U_t, and return its subset M minimizing $D(M, \mathcal{S})$.
4. Otherwise (a) sub-sample elements $\mathcal{P}_t \subseteq U_t$ by selecting each element with probability $\frac{9\ln(nm)}{\epsilon^6|U_t|}$ and (b) for every subset M_c of $\mathcal{P}_t$, generalize this seed median M_c from a solution on $\mathcal{P}_t$ to a solution M on U_t. Finally return M that minimizes $D(M, \mathcal{S})$.

Note that the median returned by the algorithm consists of only the elements in U_t. We first show that restricting only to sets in $\mathcal{S}_t$ adds at most an ϵm to the cost of the solution (Lemma 3.1); then we show that by restricting only to the elements in U_t increases the cost by at most an additional ϵm (Lemma 3.2).

LEMMA 3.1. *Fix t and $\mathcal{S}_t$ as above. Let M^* be the optimal median for $\mathcal{S}$, M_t^* be the optimal median for $\mathcal{S}_t$, and M be such that $D(M, \mathcal{S}_t) \leq D(M_t^*, \mathcal{S}_t) + \alpha$. Then $D(M, \mathcal{S}) \leq D(M^*, \mathcal{S}) + \alpha + \epsilon m$.*

Proof. We can write $D(M, \mathcal{S}) = D(M, \mathcal{S}_t) + D(M, \mathcal{S} \setminus \mathcal{S}_t)$. Consider any set $S \in \mathcal{S} \setminus \mathcal{S}_t$. Suppose $|S| \leq \epsilon t$ (the other case is similar). We have

$$D(M^*, S) = 1 - \frac{|S \cap M^*|}{|S \cup M^*|} \geq 1 - \frac{\epsilon t}{t} = 1 - \epsilon \geq D(M, S) - \epsilon.$$

Therefore,

$$\begin{aligned} D(M, \mathcal{S}) &= D(M, \mathcal{S}_t) + D(M, \mathcal{S} \setminus \mathcal{S}_t) \\ &\leq D(M^*, \mathcal{S}_t) + \alpha + D(M^*, \mathcal{S} \setminus \mathcal{S}_t) + \epsilon|\mathcal{S} \setminus \mathcal{S}_t| \\ &\leq D(M^*, \mathcal{S}) + \alpha + \epsilon m. \end{aligned}$$

□

LEMMA 3.2. *Fix an arbitrary integer k, and let M be any subset of U. If $T \subseteq U$ is the set of elements of degree $\leq k$ then $D(M \setminus T, \mathcal{S}) \leq D(M, \mathcal{S}) + k$.*

Proof. Consider the total similarity of M,

$$\begin{aligned} J(M, \mathcal{S}) &= \sum_{x \in M} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup M|} \\ &= \sum_{x \in M \cap T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup M|} + \sum_{x \in M \setminus T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup M|}. \end{aligned}$$

The first sum can be bounded as

$$\begin{aligned} \sum_{x \in M \cap T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup M|} &\leq \sum_{x \in M \cap T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|M|} \\ &\leq \sum_{x \in M \cap T} \frac{k}{|M|} \leq k. \end{aligned}$$

To bound the total similarity of $M \setminus T$,

$$\begin{aligned} J(M \setminus T, \mathcal{S}) &= \sum_{x \in M \setminus T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup (M \setminus T)|} \\ &\geq \sum_{x \in M \setminus T} \sum_{\substack{S_j \ni x \\ S_j \in \mathcal{S}}} \frac{1}{|S_j \cup M|}. \end{aligned}$$

Thus, if $J(M, \mathcal{S}) \leq k + J(M \setminus T, \mathcal{S})$, then $D(M \setminus T, \mathcal{S}) \leq D(M, \mathcal{S}) + k$. □

So far we have shown that the optimal median on the instance consisting of $\mathcal{S}_t$ with the elements in U_t is an $O(\epsilon m)$-approximate median to the original instance. Now, if $|U_t|$ is sufficiently small, i.e., $|U_t| = O(\frac{\ln nm}{\epsilon^6})$, then we can just enumerate all of the subsets of U_t to find the optimal median.

Otherwise (i.e., U_t is relatively large), we proceed to sub-sample elements from U_t with probability $p = \frac{9\ln(nm)}{\epsilon^6|U_t|}$. Let $P \subseteq U_t$ be the set of sampled elements. An easy application of the Chernoff bound shows that $|P| =$

$O(\ln(nm)/\epsilon^6)$ with high probability. Furthermore, as the following shows, the size of the intersection between any two sets $A, B \subseteq U_t$ is either small, or is well-preserved.

LEMMA 3.3. *For any $A, B \subseteq U_t$, and let $C = A \cap B$. Then, with probability $\geq 1 - O(nm)^{-3}$, if $|C| \geq \epsilon^4|U_t|$, then $(1-\epsilon)p|C| \leq |C \cap P| \leq (1+\epsilon)p|C|$ and if $|C| < \epsilon^4|U_t|$, then $|C \cap P| \leq 6\epsilon^4 p|U_t|$.*

Proof. By the Chernoff bound, if X is the sum of k independent binary random variables, each with expectation q, it holds that

$$\Pr\left[|X - kq| > \epsilon kq\right] \leq 2\exp\left(-\frac{\epsilon^2}{3}kq\right),$$

and if $u > 2ekq$, then

$$\Pr\left[X > u\right] \leq 2^{-u}.$$

In our case $|C \cap P|$ is the sum of $|C|$ independent binary random variables each with expectation p. When $|C| \geq \epsilon^4|U_t|$, we have

$$\begin{aligned}\Pr[||C \cap P| - p|C|| > \epsilon p|C|] &\leq 2\exp\left(-\frac{\epsilon^2}{3}p|C|\right)\\ &\leq 2\exp\left(-\frac{\epsilon^2}{3}\left(\epsilon^4|U_t|\right)\left(9\epsilon^{-6}|U_t|^{-1}\ln(nm)\right)\right)\\ &= 2\exp\left(-3\ln(nm)\right) \leq O\left(\frac{1}{nm}\right)^3.\end{aligned}$$

If $|C| < \epsilon^4|U_t|$, we have $2ep|C| < 6\epsilon^4 p|U_t| = u$, so the second bound from above can be applied. Observe that $u = 54\epsilon^{-2}\ln(nm) \geq 3\lg(nm)$ and thus

$$\Pr[|C \cap P| > 6\epsilon^4 p|U_t|] \leq \left(\frac{1}{nm}\right)^3. \qquad \square$$

At this point the algorithm proceeds to look at all possible subsets of P as the seed medians, M_c. We now show how to generalize the seed to a median on the full set U_t. Let M_t^* be the optimal median on U_t and let $M_c = M_P^* = M_t^* \cap P$. The condition we require is that the generalization of M_P^* to the ground set U_t happens to be an ϵm (additive) approximate median on $\mathcal{S}_t$.

For a candidate median M_c, let $\mathcal{S}_t(M_c) \subseteq \mathcal{S}_t$ be the sets that have a "large-enough" intersection with M_c. Formally, let $\mathcal{S}_t(M_c) = \{S \in \mathcal{S}_t \mid |M_c \cap S| > \frac{54\ln(nm)}{\epsilon^2}\}$. To generalize M_c, we solve the following system $\mathcal{L}$ of linear inequalities on $(\mathbf{x_1}, \ldots, \mathbf{x}_{|\mathbf{U_t}|})$. We note that while the inequalities contain an irrational number p^{-1}, we can replace it with a sufficiently precise rational approximation without materially affecting the overall answer.

$$\mathcal{L} = \begin{cases} 0 \leq \mathbf{x_i} \leq 1, \quad \forall i, 1 \leq i \leq |U_t| \\ \sum_{x_i \in S \cap U_t} \mathbf{x_i} \leq (1-\epsilon)^{-1} \cdot |S \cap M_c| \cdot p^{-1}, \\ \qquad \forall S \in \mathcal{S}_t(M_c) \\ \sum_{x_i \in S \cap U_t} \mathbf{x_i} \geq (1+\epsilon)^{-1} \cdot |S \cap M_c| \cdot p^{-1}, \\ \qquad \forall S \in \mathcal{S}_t(M_c) \\ \sum_{x_i \in U_t} \mathbf{x_i} \leq (1-\epsilon)^{-1} \cdot |M_c| \cdot p^{-1} \\ \sum_{x_i \in U_t} \mathbf{x_i} \geq (1+\epsilon)^{-1} \cdot |M_c| \cdot p^{-1} \end{cases}$$

If there exists a solution $(\hat{\mathbf{x}}_\mathbf{1}, \ldots, \hat{\mathbf{x}}_{|\mathbf{U_t}|})$, compute M by select each element $x_i \in U_t$ with probability $\hat{\mathbf{x}}_i$ independently.

We begin by showing that unless the optimum solution M^* has a very small intersection with U_t, there will be some solution to the set $\mathcal{L}$ of linear inequalities. We say that some subset $Y \subseteq U_t$, we defined its $\mathcal{L}$-assignment as $\{\mathbf{y}_i\}_{i=1}^{|U_t|}$, where $\mathbf{y}_i = 1$ if $x_i \in Y$ and $\mathbf{y}_i = 0$ otherwise.

LEMMA 3.4. *Let M^* be the optimal median with $|M^*| = t$. Fix U_t, and let $M_t^* = M^* \cap U_t$. Select $P \subseteq U_t$ as above, and let $M_P^* = M^* \cap P$. Then, either $|M_t^*| < \epsilon^2 t$ or with high probability, the $\mathcal{L}$-assignment of M_t^* satisfies $\mathcal{L}$.*

Proof. With high probability, the conditions in Lemma 3.3 hold for every intersection $C = M_t^* \cap S$, with $S \in \mathcal{S}$. Let $\{\mathbf{y}_i\}_{i=1}^{|U_t|}$ be the $\mathcal{L}$-assignment of M_t^*. Fix a set $S \in \mathcal{S}_t(M_P^*)$. The first constraint of $\mathcal{L}$,

$$\sum_{x_i \in S \cap U_t} \mathbf{y}_i \leq \frac{|S \cap M_P^*|}{(1-\epsilon)p},$$

is equivalent to

$$|M_t^* \cap S| \leq \frac{|M_P^* \cap S|}{(1-\epsilon)p} = \frac{|(M_t^* \cap S) \cap P|}{(1-\epsilon)p}.$$

In other words, it states that the intersection $M_t^* \cap S$ is well preserved under the sample P. This is exactly the condition guaranteed by Lemma 3.3, provided that $|M_t^* \cap S| \geq \epsilon^4|U_t|$. Assume to the contrary that $|M_t^* \cap S| < \epsilon^4|U_t|$. Then, the size of $|M_t^* \cap S \cap P| \leq 6\epsilon^4|U_T|p = \frac{54\ln(nm)}{\epsilon^2}$; therefore $S \notin \mathcal{S}_t(M_P)$.

The second constraint is similar. Finally, the remaining constraints say that $|M_t^*| \leq \frac{|M_t^* \cap P|}{(1-\epsilon)p}$. We first derive a bound on $|U_t|$. Since each set in $\mathcal{S}_t$ has at most t/ϵ elements, the multi-set of elements present in some set $S \in \mathcal{S}_t$ is at most $|\mathcal{S}_t|t/\epsilon$. Furthermore, since the elements in U_t have degree at least $\epsilon|\mathcal{S}_t|$, the total number of such elements can be at most $\frac{|\mathcal{S}_t|t/\epsilon}{\epsilon|\mathcal{S}_t|}$. Therefore $|U_t| \leq t\epsilon^{-2}$.

We know by assumption that $|M_t^*| \geq \epsilon^2 t \geq \epsilon^4|U_t|$. Therefore $|M_t^*|$ satisfies the conditions of Lemma 3.3, and $|M_t^*| \leq \frac{|M_t^* \cap P|}{(1-\epsilon)p}$, as we needed to show. $\square$

THEOREM 3.1. *Let M^* be the optimal median, and M be the best median produced by the algorithm above. Then, with high probability $D(M^*, \mathcal{S}) \leq D(M, \mathcal{S}) + O(\epsilon m)$.*

Proof. As before, let $t = |M^*|$, and use U_t and P as above. For ease of notation, denote by $M_t^* = M^* \cap U_t$ and $M_P^* = M^* \cap P$. And suppose the conditions of Lemma 3.4 hold. Let M be the solution reconstructed by the algorithm when $M_c = M_P^*$, or $M = \emptyset$ if $|M_t^*| < \epsilon^2 t$.

Let $\mathcal{S}_N = \{S \in \mathcal{S}_t \mid |S \cap M_t^*| \geq \frac{6\epsilon^2 t}{(1-\epsilon)}\}$. Observe that for every set $S \in \mathcal{S}_t \setminus \mathcal{S}_N$,

$$D(M_t^*, S) \geq 1 - \frac{\frac{6\epsilon^2 t}{1-\epsilon}}{\epsilon t} = 1 - \frac{6\epsilon}{(1-\epsilon)} = 1 - O(\epsilon).$$

Therefore for such sets S any median M, $D(M, S) \leq D(M^*, S) + O(\epsilon)$.

To bound $D(M, \mathcal{S})$, observe that

$$D(M, \mathcal{S}) = D(M, \mathcal{S}_N) + D(M, \mathcal{S}_t \setminus \mathcal{S}_N) + D(M, \mathcal{S} \setminus \mathcal{S}_t).$$

Lemmas 3.1 and 3.2 imply that $D(M, \mathcal{S} \setminus \mathcal{S}_t) \leq D(M^*, \mathcal{S} \setminus \mathcal{S}_t) + O(\epsilon m)$. Therefore what remains to be shown is that the median M is such that $D(M, S_N) \leq D(M_t^*, S_N) + O(\epsilon m)$.

Suppose that $|M_t^*| < \epsilon^2 t$, then $\mathcal{S}_N = \emptyset$ and the proof is complete. Otherwise, for each set $S \in S_N$, notice that $|S \cap M_t^*| \geq 6\epsilon^2 t(1-\epsilon)^{-1} \geq 6\epsilon^4 |U_t|(1-\epsilon)^{-1}$, and therefore $|S \cap M_t^* \cap P| \geq 6\epsilon^4 |U_t| p = \frac{54 \ln(nm)}{\epsilon^2}$. Therefore $\mathcal{S}_N \subseteq \mathcal{S}_t(M_c)$.

Let $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{|U_t|}\}$ be any solution to the system $\mathcal{L}$ when $M_c = M_P^*$. Then for every $S \in \mathcal{S}_N$ we have that

$$\sum_{x_i \in S \cap U_t} \mathbf{y}_i \geq \frac{|S \cap M_P^*|}{(1+\epsilon)p}.$$

Since $|S \cap (M_t^* \cap P)| = |S \cap M_P^*| \geq \epsilon^{-2} 54 \ln(nm)$, an easy application of the Chernoff bound shows that with high probability, a randomized rounding of $\mathbf{y}$ will approximate $|S \cap M_t^*|$ to within a $(1 \pm \epsilon)$ factor. This combined with the fact that $\sum_i \mathbf{y}_i$ is also concentrated with high probability, implies that for any $S \in \mathcal{S}_N$, $J(M, S) \geq J(M_t^*, S) - O(\epsilon)$; thus $D(M, S_N) \leq D(M_t^*, S_N) + O(\epsilon m)$. The proof is complete. □

In the next sections we show a polynomial-time algorithm that produces a $(1 + O(\sqrt{\epsilon}))$-approximate median if the optimal median has value $\leq \epsilon m$. The two algorithms together form a PTAS.

3.2 A PTAS when the optimal median is small

In this section we provide an algorithm that works when the optimal median is very good, and the average distance from a set to the median is ϵ.

DEFINITION 3.1. (ϵ-GOOD INSTANCE) *An instance $\mathcal{S}$ on m sets is* ϵ-good *if the cost of the optimal median is less than ϵm.*

We show an algorithm that achieves a $(1 + O(\sqrt{\epsilon}))$-approximate median to ϵ-good instances in time $O(nm)$.

We begin by proving several structural properties of any ϵ-good instance. First, for the instance $\mathcal{S} = \{S_1, \ldots, S_m\}$, let $\mu = \text{median}\{|S_1|, \ldots, |S_m|\}$.

Any ϵ-good instance has the following properties:

- The size of the best median set, M^*, is $(1 \pm O(\epsilon))\mu$. (Lemma 3.5.)
- There are many – $(1 - O(\sqrt{\epsilon}))\mu$ high-degree elements (elements present in at least $(1 - O(\sqrt{\epsilon}))m$ sets), and all of them are part of each near-optimal median. (Lemma 3.6 and Lemma 3.8.)

This set of properties suggests the following natural linear-time algorithm:

1. Find the set of all high-degree elements, and add them to the optimal median; this adds at least $(1 - O(\sqrt{\epsilon}))\mu$ elements.
2. Greedily select another $O(\sqrt{\epsilon} + \epsilon)\mu$ elements to add to the median. Since we are only adding a small number of extra elements to the set, the denominator does not change by much, but the size of respective intersections is maximized.

We now proceed to formalize these properties.

LEMMA 3.5. *Fix $0 < \epsilon_1 \leq \frac{1}{6}$. If a set $M \subseteq X$ is such that $D(M, \mathcal{S}) \leq \epsilon_1 m$, then*

$$(1 - 3\epsilon_1)\mu \leq |M| \leq (1 + 3\epsilon_1)\mu.$$

Intuitively, consider a median M with $|M| > (1+\epsilon)\mu$. Then, on at least half of the sets (those whose sizes are less than μ), the distance between M and S_i will be at least $\frac{1}{1+\epsilon}$, leading to a contradiction of the goodness of M.

Proof. Let $\tilde{\epsilon}_1 = 3\epsilon_1$ and consider an arbitrary set $M \subseteq X$ such that $(1 + \tilde{\epsilon}_1)^{-1}\mu \geq (1 - \tilde{\epsilon}_1)\mu \geq |M| \geq (1 + 3\epsilon_1) = (1 + \tilde{\epsilon}_1)\mu$. Let $\mathcal{S}' \subseteq \mathcal{S}$ be such that $S_i \in \mathcal{S}'$ iff $|S_i| \leq \mu$ (resp., $|S_i| \geq \mu$). Note that $|\mathcal{S}'| \geq m/2$.

Note that for each $S_i \in \mathcal{S}'$, it holds that $\frac{|S_i \cap M|}{|S_i \cup M|} < \frac{\mu}{(1+\tilde{\epsilon}_1)\mu} = \frac{1}{1+\tilde{\epsilon}_1}$ because $|S_i \cap M| \leq |S_i| \leq \mu$ and $|S_i \cup M| \geq |M| > (1 + \tilde{\epsilon}_1)\mu$ (resp., $|S_i \cap M| \leq |M| < (1 + \tilde{\epsilon}_1)^{-1}\mu$ and $|S_i \cup M| \geq |S_i| \geq \mu$).

Thus,

$$\begin{aligned} J(M,\mathcal{S}) &= \sum_{S_i\in\mathcal{S}} \frac{|S_i\cap M|}{|S_i\cup M|} \\ &= \sum_{S_i\in\mathcal{S}'} \frac{|S_i\cap M|}{|S_i\cup M|} + \sum_{S_i\in\mathcal{S}\setminus\mathcal{S}'} \frac{|S_i\cap M|}{|S_i\cup M|} \\ &< \sum_{S_i\in\mathcal{S}'} \frac{1}{1+\tilde{\epsilon}_1} + \sum_{S_i\in\mathcal{S}\setminus\mathcal{S}'} 1 \\ &= |\mathcal{S}'|\,\frac{1}{1+\tilde{\epsilon}_1} + (m-|\mathcal{S}'|) \\ &\leq \frac{1}{1+\tilde{\epsilon}_1}\frac{m}{2} + \frac{m}{2} \\ &= \left(1-\frac{\tilde{\epsilon}_1}{2+2\tilde{\epsilon}_1}\right)m. \end{aligned}$$

Thus, the total distance is at least $\frac{\tilde{\epsilon}_1}{2+2\tilde{\epsilon}_1}m \geq \frac{1}{3}\tilde{\epsilon}_1$, for $\tilde{\epsilon}_1 \leq \frac{1}{2}$, i.e., $\epsilon_1 \leq \frac{1}{6}$. □

We next lower bound the number of high-degree elements. Let M^* be the optimal median, and let $D = d(M^*,\mathcal{S})$.

LEMMA 3.6. *Fix some $0 < \epsilon_2 \leq \frac{2-\sqrt{3}}{3}$. We say that an element $j \in X$ has high degree if $\deg(j) = |\{S_i \mid j \in S_i \in \mathcal{S}\}| \geq (1-\sqrt{2\epsilon_2})m$. If $D(M,\mathcal{S}) \leq \epsilon_2 m$, then there exist at least $(1-\sqrt{2\epsilon_2})\mu$ high-degree elements.*

We need one more technical lemma before proving Lemma 3.6. We begin by showing that almost all of the sets have their size in $(1\pm O(\sqrt{\epsilon}))\mu$. Intuitively, if there are many sets whose size is far from the size of the near-optimal median (as bounded in Lemma 3.5), then each of those sets contributes at least an $O(\sqrt{\epsilon})$ to the overall distance, leading to a contradiction.

LEMMA 3.7. *Fix $0 < \epsilon_3 < \frac{1}{6}$. Let $\mathcal{S}' \subseteq \mathcal{S}$ be the class of sets S_i of sizes $(1-4\sqrt{\epsilon_3})\mu \leq |S_i| \leq (1+4\sqrt{\epsilon_3})\mu$. If $\mathcal{S}$ is an ϵ_3-good instance, then $|\mathcal{S}'| \geq \left(1-\sqrt{\epsilon_3}\right)m$.*

Proof. Suppose $|\mathcal{S}\setminus\mathcal{S}'| > \sqrt{\epsilon_3}m$, i.e., suppose that more than $\sqrt{\epsilon_3}m$ sets have size at most $\left(1-4\sqrt{\epsilon_3}\right)\mu$ or at least $\left(1+4\sqrt{\epsilon_3}\right)\mu$. Since $\epsilon_3 \leq \frac{1}{6}$ and by Lemma 3.5, the best median M^* will have size $(1-3\epsilon_3)\mu \leq |M^*| \leq (1+3\epsilon_3)\mu$.

If $|S_i| \leq \left(1-\sqrt{\epsilon_3}\right)\mu$, then

$$J(M^*,S_i) \leq \frac{|M^*\cap S_i|}{|M^*\cup S_i|} \leq \frac{|S_i|}{|M^*|} \leq \frac{1-4\sqrt{\epsilon_3}}{1-3\epsilon_3}.$$

On the other hand, if $|S_i| \geq \left(1+4\sqrt{\epsilon_3}\right)\mu$, then we have

$$J(M^*,S_i) \leq \frac{|M^*\cap S_i|}{|M^*\cup S_i|} \leq \frac{|M^*|}{|S_i|} \leq \frac{1+3\epsilon_3}{1+4\sqrt{\epsilon_3}}.$$

In both cases, $J(M^*,S_i) \leq 1-\sqrt{\epsilon_3}$, since $\epsilon_3 \leq \frac{1}{6}$. Thus,

$$\begin{aligned} J(M^*,\mathcal{S}) &= \sum_{S_i\in\mathcal{S}} J(M^*,S_i) \\ &= \sum_{S_i\in\mathcal{S}'} J(M^*,S_i) + \sum_{S_i\in\mathcal{S}\setminus\mathcal{S}'} J(M^*,S_i) \\ &\leq \sum_{S_i\in\mathcal{S}'} 1 + \sum_{S_i\in\mathcal{S}\setminus\mathcal{S}'} (1-\sqrt{\epsilon_3}) \\ &= |\mathcal{S}'| + |\mathcal{S}\setminus\mathcal{S}'|\,(1-\sqrt{\epsilon_3}) \\ &< (1-\sqrt{\epsilon_3})\,m + \sqrt{\epsilon_3}m\,(1-\sqrt{\epsilon_3}) \\ &= (1-\epsilon_3)m. \end{aligned}$$

Thus, the total distance will be more than $\epsilon_3 m$, a contradiction. □

We are now ready to prove Lemma 3.6.

Proof. Let $X' \subseteq X$ be the set of high-degree elements. Let M^* be the optimal median. By Lemma 3.5, $(1-3\epsilon_2)\mu \leq |M^*| \leq (1+3\epsilon_2)\mu$. Note that the total Jaccard similarity $J(M^*,\mathcal{S})$ can be written as

$$\begin{aligned} J(M^*,\mathcal{S}) &= \sum_{S_i\in\mathcal{S}} \frac{|S_i\cap M^*|}{|S_i\cup M^*|} \\ &= \sum_{S_i\in\mathcal{S}} \sum_{x\in S_i\cap M^*} \frac{1}{|S_i\cup M^*|} \\ &= \sum_{x\in M^*} \sum_{S_i\ni x} \frac{1}{|S_i\cup M^*|} \\ &\leq \sum_{x\in M^*} \sum_{S_i\ni x} \frac{1}{|M^*|} \\ &\leq \frac{1}{|M^*|} \sum_{x\in X'\cap M^*} \left(\sum_{S_i\ni x} 1\right) + \\ &\qquad + \frac{1}{|M^*|} \sum_{x\in(X-X')\cap M^*} \left(\sum_{S_i\ni x} 1\right) \\ &\leq \frac{1}{|M^*|} \sum_{x\in X'\cap M^*} m + \\ &\qquad + \frac{1}{|M^*|} \sum_{x\in(X-X')\cap M^*} \left((1-\sqrt{2\epsilon_2})\cdot m\right). \end{aligned}$$

Now, suppose by contradiction that $|X'| < \left(1-\sqrt{2\epsilon_2}\right)\mu = T$. The overall number of terms in the two sums of the previous expression is at most $|M^*|$; also the higher the number of terms of the first sum, the higher is the value of the expression. Thus,

$$\begin{aligned}
&J(M^*,\mathcal{S})\\
&< \frac{1}{|M^*|}Tm + \frac{1}{|M^*|}\left(|M^*|-T\right)\left(\left(1-\sqrt{2\epsilon_2}\right)m\right)\\
&= \left(1-\sqrt{2\epsilon_2}\right)m + \frac{T}{|M^*|}\sqrt{2\epsilon_2}m\\
&= \left(1-\sqrt{2\epsilon_2}\right)m + \frac{T}{(1-3\epsilon_2)\,\mu}\sqrt{2\epsilon_2}m\\
&= \left(1-\sqrt{2\epsilon_2}\right)m + \frac{\left(1-\sqrt{2\epsilon_2}\right)\mu}{(1-3\epsilon_2)\,\mu}\sqrt{2\epsilon_2}m\\
&= \left(1-\sqrt{2\epsilon_2}\right)m + \left(1-\frac{\sqrt{2\epsilon_2}-3\epsilon_2}{1-3\epsilon_2}\right)\sqrt{2\epsilon_2}m\\
&= m - \frac{\sqrt{2\epsilon_2}-3\epsilon_2}{1-3\epsilon_2}\sqrt{2\epsilon_2}m\\
&= \left(1-\frac{2\epsilon_2-3\sqrt{2\epsilon_2^3}}{1-3\epsilon_2}\right)m.
\end{aligned}$$

This implies $D(M^*,\mathcal{S}) > \frac{2-3\sqrt{2\epsilon_2}}{1-3\epsilon_2}\epsilon_2 m \geq \epsilon_2 m$ (where the last inequality is implied by $\epsilon_2 \leq \frac{2-\sqrt{3}}{3}$). This is a contradiction and hence $|X'| \geq \left(1-\sqrt{2\epsilon_2}\right)\mu$. □

Finally, the next lemma states that each high-degree element is part of any near-optimal median.

LEMMA 3.8. *Fix $0 < \epsilon_4 < \frac{3}{100}$. Let $X^* \subseteq X$ be the set of the elements having degree $\geq \left(1-\sqrt{\epsilon_4}\right)m$. Take any $M \subseteq X$ such that $d(M,\mathcal{S}) \leq \epsilon_4 m$. If $X^* \setminus M \neq \emptyset$, it holds that $d(M \cup X^*,\mathcal{S}) < d(M,\mathcal{S})$.*

Proof. Fix an arbitrary $x^* \in X^* \setminus M$. We will show that $D(M\cup\{x^*\},\mathcal{S}) < D(M,\mathcal{S})$, so that the main statement will be proved.

Note that for any $Y \subseteq X$, it holds that $J(Y,\mathcal{S}) = \sum_{y\in Y}\sum_{S_i \ni y}\frac{1}{|S_i\cup Y|}$.

By Lemma 3.7, there exist at least $(1-\sqrt{\epsilon_4})m$ sets of size $\leq \left(1+4\sqrt{\epsilon_4}\right)\mu$. The element x^* has degree $\geq \left(1-\sqrt{\epsilon_4}\right)m$ so it will be part of at least $(1-2\sqrt{\epsilon_4})m$ sets of size $\leq \left(1+4\sqrt{\epsilon_4}\right)\mu$. Let $\mathcal{S}'_{x^*}$ be the class of these sets.

By Lemma 3.5, the set M will have size $(1-3\epsilon_4)\mu \leq |M| \leq (1+3\epsilon_4)\mu$. So, for $S_i \in \mathcal{S}'_{x^*}$ we can lower bound the term $\frac{1}{|S_i\cup M|}$ (which will be used in the following chain of inequalities) by

$$\begin{aligned}
\frac{1}{|S_i\cup M|} &\geq \frac{1}{|S_i|+|M|}\\
&\geq \frac{1}{\left(1+4\sqrt{\epsilon_4}\right)\mu + (1+3\epsilon_4)\mu}\\
&\geq \frac{1}{\left(2+7\sqrt{\epsilon_4}\right)\mu}.
\end{aligned}$$

Also, we will use the inequality $|M| \geq (1-3\epsilon_4)\mu$.

$$\begin{aligned}
&J(M\cup\{x^*\},\mathcal{S}) - J(M,\mathcal{S}) =\\
&= \sum_{x\in M\cup\{x^*\}}\sum_{\substack{S_i\\ x\in S_i\in\mathcal{S}}}\frac{1}{|S_i\cup M\cup\{x^*\}|} - \sum_{x\in M}\sum_{\substack{S_i\\ x\in S_i\in\mathcal{S}}}\frac{1}{|S_i\cup M|}\\
&= \sum_{\substack{S_i\\ x^*\in S_i\in\mathcal{S}}}\frac{1}{|S_i\cup M\cup\{x^*\}|}\\
&\qquad + \sum_{x\in M}\sum_{\substack{S_i\\ x\in S_i\in\mathcal{S}}}\left(\frac{1}{|S_i\cup M\cup\{x^*\}|} - \frac{1}{|S_i\cup M|}\right)\\
&= \sum_{\substack{S_i\\ x^*\in S_i\in\mathcal{S}}}\frac{1}{|S_i\cup M|} - \sum_{x\in M}\sum_{\substack{S_i\\ x\in S_i\in\mathcal{S}\\ x^*\notin S_i}}\frac{1}{|S_i\cup M|\left(|S_i\cup M|+1\right)}\\
&> \sum_{S_i\in\mathcal{S}'_{x^*}}\frac{1}{|S_i\cup M|} - \sum_{x\in M}\sum_{\substack{S_i\\ x\in S_i\in\mathcal{S}\\ x^*\notin S_i}}\frac{1}{|M|^2}\\
&\geq \sum_{S_i\in\mathcal{S}'_{x^*}}\frac{1}{\left(2+7\sqrt{\epsilon_4}\right)\mu} - \sum_{x\in M}\frac{\sqrt{\epsilon_4}m}{|M|^2}\\
&\geq \frac{\left(1-2\sqrt{\epsilon_4}\right)m}{\left(2+7\sqrt{\epsilon_4}\right)\mu} - \frac{\sqrt{\epsilon_4}m}{|M|}\\
&\geq \frac{\left(1-2\sqrt{\epsilon_4}\right)m}{\left(2+7\sqrt{\epsilon_4}\right)\mu} - \frac{\sqrt{\epsilon_4}m}{(1-3\epsilon_4)\mu}\\
&= \left(\frac{1-2\sqrt{\epsilon_4}}{2+7\sqrt{\epsilon_4}} - \frac{\sqrt{\epsilon_4}}{1-3\epsilon_4}\right)\frac{m}{\mu}.
\end{aligned}$$

Note that the latter is positive for $\epsilon_4 \leq c$, for some positive constant c (in particular for some $c \geq 0.0319\ldots$). Thus if $\epsilon_4 \leq c$ then $J(M\cup\{x^*\},\mathcal{S}) - J(M,\mathcal{S}) > 0$, or equivalently, $D(M\cup\{x^*\},\mathcal{S}) < D(M,\mathcal{S})$. □

At this point we know that every near-optimal median contains no more than $(1+O(\epsilon))\mu$ elements, out of which at least $(1-O(\sqrt{\epsilon}))$ are the easily found dense elements. Thus, we need to chose at most $O(\sqrt{\epsilon}\mu)$ extra elements to include in the solution. The difficulty of finding the optimal median stems from the fact that as we add extra elements to a candidate median set, the total contribution of each set to the overall distance

changes due to changes *both* in the numerator and in the denominator. However, since we have an approximation to the bound on the size of the optimal median, we can effectively freeze the denominators, knowing that we are making at most an $1+\sqrt{\epsilon}$ approximation to the final solution. Once the denominators are frozen, the problem is simpler and the greedy algorithm is optimal.

Formally, let M be the set of at least $(1-O(\sqrt{\epsilon}))\mu$ dense elements guaranteed by Lemma 3.8. For an element $x_i \notin M$, let the weight of x be $\sum_{S_j \ni x_i} \frac{1}{|S_j \cup M|} - \sum_{S_j \not\ni x_i} \frac{1}{|S_j \cup M|}$. Set N^* to be the set found by greedily selecting elements in order of decreasing weight, stopping when either (a) the size of N^* is $O(\sqrt{\epsilon})$ or (b) the weight of the element in consideration is non-positive.

THEOREM 3.2. *Let M and N^* as above. Then $D(M^*, \mathcal{S}) \geq \frac{1}{1+O(\sqrt{\epsilon})} \cdot D(M \cup N^*, \mathcal{S})$.*

Proof. For any solution $M \cup N$, we have

$$D(\mathcal{S}, M \cup N) = \sum_{S_j \in \mathcal{S}} \left(1 - \frac{|S_j \cap (M \cup N)|}{|S_j \cup (M \cup N)|}\right)$$
$$= \sum_{S_j \in \mathcal{S}} \frac{|S_j \cup (M \cup N)| - |S_j \cap (M \cup N)|}{|S_j \cup (M \cup N)|}.$$

If we restrict the size of N to be $|N| < O(\sqrt{\epsilon}\mu)$, then for each set S_j,

$$|S_j \cup M| \leq |S_j \cup (M \cup N)| \leq |S_j \cup M|(1 + O(\sqrt{\epsilon})),$$

where the last inequality follows from the lower bound on the size of M. For any set T, let D_T be the distance with each denominator fixed to be $|S_j \cup T|$.

$$D_T(\mathcal{S}, A) = \sum_{S_j \in \mathcal{S}} \frac{|S_j \cup A| - |S_j \cap A|}{|S_j \cup T|}$$
$$= \sum_{S_j \in \mathcal{S}} \sum_{x_i \in S_j \triangle A} \frac{1}{|S_j \cup T|},$$

where for two sets U and V, $U \triangle V$ denotes their symmetric difference. Then we have

(3.1)
$$D_M(\mathcal{S}, M \cup N) \geq D(\mathcal{S}, M \cup N) \geq \frac{D_M(\mathcal{S}, M \cup N)}{1 + O(\sqrt{\epsilon})}.$$

Let N be such that $N \cap M = \emptyset$. It is easy to check that D_M can be rewritten as

$$D_M(\mathcal{S}, M \cup N) = \sum_{S_j \in \mathcal{S}} \sum_{x_i \in S_j \triangle (M \cup N)} \frac{1}{|S_j \cup M|} =$$
$$= \left(\sum_{x_i \in M} \sum_{S_j \not\ni x_i} \frac{1}{|S_j \cup M|} + \sum_{x_i \notin M} \sum_{S_j \ni x_i} \frac{1}{|S_j \cup M|} \right) -$$
$$- \sum_{x_i \in N} \left(\sum_{S_j \ni x_i} \frac{1}{|S_j \cup M|} - \sum_{S_j \not\ni x_i} \frac{1}{|S_j \cup M|} \right).$$

Let N^* be the set that minimizes $D_M(\mathcal{S}, M \cup N^*)$ under the constraints $M \cap N^* = \emptyset$ and $|N^*| < O(\sqrt{\epsilon}\mu)$. If we define the weight of an element $x_i \notin M$ to be $\sum_{S_j \ni x_i} \frac{1}{|S_j \cup M|} - \sum_{S_j \not\ni x_i} \frac{1}{|S_j \cup M|}$, then N^* can be found by greedily selecting elements in order of decreasing weight, stopping when either (a) the size of N^* has reached its limit, or (b) the weight of the element in consideration is non-positive.

Let $M^* = M \cup M'$, $M' \cap M = \emptyset$, be the optimal solution. Recall that $|M'| \leq O(\sqrt{\epsilon})\,|M|$. Then,

$$D(\mathcal{S}, M^*) \geq \frac{1}{1 + O(\sqrt{\epsilon})} D_M(\mathcal{S}, M^*)$$
$$\geq \frac{1}{1 + O(\sqrt{\epsilon})} D_M(\mathcal{S}, M \cup N^*)$$
$$\geq \frac{1}{1 + O(\sqrt{\epsilon})} D(\mathcal{S}, M \cup N^*),$$

where the first and the last inequalities follow from (3.1) and the second from the optimality of N^*. □

Therefore, the solution $M \cup N^*$ found by the algorithm is a $(1+O(\sqrt{\epsilon}))$-approximation to the optimal median.

4 A PTAS for the weighted Jaccard median

In the weighted Jaccard median problem, we are given a (multi-)set of vectors $\mathcal{V} = \{V_1, \ldots, V_m\}$, where the generic V_i is a non-negative real vector on n coordinates, $V_i \in \mathbf{R}^n_{\geq 0}$. In this section we give an algorithm for the weighted Jaccard median problem. We defer the technical details to Appendix C and give a high-level description here.

First, the algorithm of Appendix C.1 returns a $(1 + O(\epsilon))$-multiplicative approximate median M if the value of the optimal median M^* is $O(\epsilon)$, i.e., if the total distance between the median and the input vectors is bounded away from 1. The two algorithms in Appendix C.2.1 and Appendix C.2.2 are guaranteed to return a median M of total distance $D(M, \mathcal{V}) \leq (1 + O(\epsilon^2))D(M^*, \mathcal{V}) + O(\epsilon^2)$, i.e., they incur both a multiplicative error of $(1 + O(\epsilon^2))$ and an additive

error of $O(\epsilon^2)$. Then, if we return the best solution of the three algorithms, we are guaranteed a $(1 + O(\epsilon))$-approximate median. We comment on the latter two algorithms.

The algorithm of Appendix C.2.1 transforms a weighted input instance having "polynomial spread" (i.e., the ratios between the maximum and the minimum non-zero value of each coordinate are at most polynomial) into a set instance such that an approximate solution for the set instance can be mapped to the original instance. The algorithm of Appendix C.2.2 transforms an arbitrary weighted instance into one with polynomial spread such that the solution to the new instance can be mapped to the original instance while preserving the approximation guarantee.

The weighted Jaccard algorithms might return medians that are not "canonical", i.e., the medians might contain coordinate values that are not part of any of the input vectors. However, as shown by [25], each optimal median is in fact canonical. Therefore, limiting the search space to contain only canonical vectors does not affect the optimum. Therefore one might want to define the Jaccard median problem as one having a finite search space (of size at most m^n, spanned by the coordinate values of its input vectors). In Appendix D we show how the "canonical" and the "not-necessarily canonical" problems are essentially the same. We give a polynomial algorithm that transforms a non-canonical median into a canonical one of smaller total distance. This let us give a PTAS for the canonical version of the problem, as well. Further, Appendix D shows that even if we do not require a canonical output, there is still no FPTAS unless P = NP.

5 Hardness of the Jaccard median

In this section we study the hardness of the Jaccard median problems. Since our focus will be on finding the optimum, both Jaccard distance median and Jaccard similarity median can be treated interchangeably, i.e., the optimal Jaccard distance median is the optimal Jaccard similarity median.

First, we describe a gadget that will be central in our reductions; this gadget appears to be "unique" in many aspects. For $t \in \mathbb{Z}^+$, let $B_t = K_{3t,3t-2}$ be the complete bipartite graph; let L denote the set of nodes on the left side, R denote the set of nodes on the right side, and C denote the set of edges in B_t. Let $U = L \cup R$ and each edge $e = (u, v) \in C$ represents the set $\mathcal{S}_e = \{u, v\}$ and let $\mathcal{S}_B = \cup_{e \in C}\{\mathcal{S}_e\}$ be an instance of the Jaccard median problem.

Let $\mathcal{M}_B^*$ denote the set of all subsets of U such that for each $M^* \in \mathcal{M}_B^*$, we have $|L \cap M^*| = t$ and $R \subseteq M^*$, i.e., each $M^* \in \mathcal{M}_B^*$ consists of exactly t nodes from L and all nodes from R. We show that the optimal Jaccard median[1] must come from the set $\mathcal{M}_B^*$ and quantify the gap between any near-optimal solution.

LEMMA 5.1. *For the instance $\mathcal{S}_B$, every $M^* \in \mathcal{M}_B^*$ is an optimal median and $J(M^*, \mathcal{S}_B) > 3t - 2$. Furthermore, $J(M^*, \mathcal{S}_B) - J(M, \mathcal{S}_B) \geq t^{-2}/32$ for $M^* \in M_B^*$ and $M \notin M_B^*$.*

Proof. Consider any $M \subseteq U$ with $|M \cap L| = a$ and $|M \cap R| = b$. We derive the conditions under which M is an optimal Jaccard median. Specifically, we show that for M to be an optimal Jaccard median, $a = t$ and $b = 3t - 2$. First note that we can explicitly write

$$\begin{aligned} J(M, \mathcal{S}_B) &= ab\frac{2}{(a+b)} + a(3t-2-b)\frac{1}{(a+b+1)} \\ &\quad + b(3t-a)\frac{1}{(a+b+1)} \\ &= \frac{3t(a+b)^2 - 2a^2}{(a+b)(a+b+1)}. \end{aligned}$$

From this,

$$\frac{\partial J(M, \mathcal{S}_B)}{\partial b} = \frac{4a^3 + (4b + 3t + 2)a^2 + 6abt + 3b^2t}{((a+b)(a+b+1))^2}.$$

Since $\frac{\partial J(M,\mathcal{S}_B)}{\partial b} > 0$ for all a, b, we have that $J(M, \mathcal{S}_B)$ is monotonically increasing in b and is hence maximized at $b = 3t - 2$, i.e., if M is an optimal Jaccard median, then $R \subseteq M$.

Likewise, we obtain

$$\frac{\partial J(M, \mathcal{S}_B)}{\partial a} = \frac{a^2(3t-2-4b) - 2ab(2b-3t+2) + 3b^2t}{((a+b)(a+b+1))^2},$$

and using the optimality condition $b = 3t - 2$, we calculate $\frac{\partial J(M,\mathcal{S}_B)}{\partial a} \mid_{b=3t-2} =$

$$(3t-2) \cdot \frac{3t(3t-2) - 2a(3t-2) - 3a^2}{((a+3t-2)(a+3t-1))^2}. \tag{5.2}$$

Since $t \geq 1$, setting (5.2) to zero gives us a quadratic equation in a. It is easy to see that the quadratic equation has a positive root at

$$a_r = \left(t - \frac{2}{3}\right) \cdot \left(2\sqrt{1 + \frac{3}{6t-4}} - 1\right).$$

[1] We remark that, in this regard, $K_{3t,3t-2}$ seems crucial — choosing $K_{at,at-b}$, for a not a multiple of 3 or $b \neq 2$, does not seem to give equal or similar guarantees.

We now show that $a_r \in (t-1, t)$. Since $6t-4 \geq 0$, we have $a_r > t - \frac{2}{3} > t-1$. Moreover,

$$\begin{aligned} a_r &= \left(t - \frac{2}{3}\right) \cdot \left(2\sqrt{1 + \frac{3}{6t-4}} - 1\right) \\ &\leq \left(t - \frac{2}{3}\right) \cdot \left(2 + \frac{3}{6t-4} - 1\right) \text{ by Taylor series} \\ &= \left(t - \frac{2}{3}\right) + \frac{1}{2} \\ &< t. \end{aligned}$$

We then note that $\frac{\partial J(M,\mathcal{S}_B)}{\partial a} \mid_{a=t-1,b=3t-2} > 0$ since (5.2) evaluates to $\frac{(3t-2)(10t-7)}{((a+3t-2)(a+3t-1))^2}$ at $a = t-1$, and $\frac{\partial J(M,\mathcal{S}_B)}{\partial a} \mid_{a=t,b=3t-2} < 0$ since (5.2) evaluates to $\frac{(-2t)(3t-2)}{((a+3t-2)(a+3t-1))^2}$ at $a = t$. Moreover, since $a \in \mathbb{Z}$ in our case, this implies that (5.2) attains its maximum value at either $a = t-1$ or $a = t$. It is easy to see that the maximum indeed occurs at $a = t$:

$$\begin{aligned} & J(M,\mathcal{S}_B) \mid_{a=t,b=3t-2} - J(M,\mathcal{S}_B) \mid_{a=t-1,b=3t-2} \\ &= \frac{3t(4t-3) + 4t^2 - 2(2t-1)(4t-1)}{(4t-1)(4t-2)(4t-3)} \\ &\geq \frac{3t-2}{(4t-1)(4t-2)(4t-3)} \geq \frac{t^{-2}}{32}. \end{aligned}$$

Hence, M is optimal if and only if $M \in \mathcal{M}_B^*$, and for each $M \in \mathcal{M}_B^*$, $J(M,\mathcal{S}_B) = \frac{2t(3t-2)}{4t-2} + \frac{2t(3t-2)}{4t-1} > (3t-2)$. And, the second best solution occurs at $a = t-1$ and $b = 3t-2$ and is lower than the optimum value by $t^{-2}/32$. □

COROLLARY 5.1. *For an instance $\mathcal{S}_B$ where each edge has multiplicity ℓ, every $M^* \in \mathcal{M}_B^*$ is an optimal median. Furthermore, $J(M^*,\mathcal{S}_B) - J(M,\mathcal{S}_B) \geq \ell \cdot t^{-2}/32$ for $M^* \in \mathcal{M}_B^*$ and $M \notin \mathcal{M}_B^*$.*

In our reductions we will overlay a graph on L, bijectively mapping nodes to G to nodes in L. There are two competing forces in play for selecting the best Jaccard median. On the one hand, the gadget ensures that we want to select exactly t nodes from L; on the other we would like to pick the densest subset in G. We make sure the gain from selecting exactly t nodes from L is a stronger force, either by duplicating every edge in $\mathcal{S}_B$ as in Section 5.1, or diluting the contribution of edges in G, as in Section 5.2. Given that the optimum median selects exactly t nodes from G, we show that it must select those forming the t-densest subgraph.

5.1 The multi-set, edge case We show that the Jaccard median problem restricted to the case when each set S in the instance $\mathcal{S}$ has exactly two elements from the universe (i.e., each set can be thought of as an "edge" in a graph whose nodes are the elements of the universe) is NP-hard. However, we need to allow $\mathcal{S}$ to be a multi-set.

Our reduction will use the following custom-defined problem called $\frac{1}{3}$-QUASI-REGULAR-CLIQUE: given a graph $G(V,E)$ with maximum degree $\Delta \geq |V|/3$, and with no node $v \in V$ such that $5|V|/18 < \deg(v) < \Delta$, does G contain a clique of size at least $|V|/3$? In Appendix B, we will show that $\frac{1}{3}$-QUASI-REGULAR-CLIQUE is NP-hard.

THEOREM 5.1. *The Jaccard median problem, where each set in the instance has two elements, is NP-hard.*

Proof. We prove the NP-hardness by reducing from $\frac{1}{3}$-QUASI-REGULAR-CLIQUE. Without loss of generality, assume $|V| = 3t$, where $t \in \mathbb{Z}^+$. We consider the bipartite gadget $B_t = (L, R, C)$ described earlier and for each edge in C, replicate it $320t^5$ times in order to obtain the bipartite multi-graph $B = (L, R, C')$. Next we overlay the graph $G(V,E)$ onto L, bijectively mapping nodes in V to nodes in L and adding appropriate edges among the nodes in L according to E; let $B' = (L, R, C' \cup E)$ be the resulting multi-graph.

Each edge $e = (u,v)$ in B' is interpreted as the set $S_e = \{u, v\}$. Let $\mathcal{S}_B = \cup_{e \in C'} S_e$ be the family corresponding to the edges in B and let $\mathcal{S}_G = \cup_{e \in E} S_e$ be the family corresponding to the edges in G. Observe that each set $M \in \mathcal{M}_B^*$ (i.e., each set $M = R \cup L'$, with $L' \subseteq L$ and $|L'| = t$), has the same Jaccard similarity $c_1 = J(M,\mathcal{S}_B)$ to $\mathcal{S}_B$. Define $c_2 = \binom{t}{2}\frac{2}{4t-2} + t\,(\Delta - (t-1))\,\frac{1}{4t-1}$, where Δ is the maximum degree in the $\frac{1}{3}$-QUASI-REGULAR-CLIQUE instance. We ask: does there exist a Jaccard median M^* of total Jaccard similarity $J(M^*,\mathcal{S}) \geq c_1 + c_2$?

First of all, observe that each clique of size t in the original graph contains only nodes of degree Δ. Further, if such a clique H exists then the median $M^* = H \cup R$ is such that $J(M^*,\mathcal{S}) = c_1 + c_2$. Indeed,

$$\begin{aligned} & J(M^*,\mathcal{S}) = J(M^*,\mathcal{S}_B) + J(M^*,\mathcal{S}_G) \\ &= c_1 + \sum_{\substack{S_e \in \mathcal{S}_G \\ |S_e \cap H| = 2}} \frac{2}{t + |R|} + \sum_{\substack{S_e \in \mathcal{S}_G \\ |S_e \cap H| = 1}} \frac{1}{t + |R| + 1} \\ &= c_1 + \binom{t}{2}\frac{2}{4t-2} + t \cdot (\Delta - (t-1)) \cdot \frac{1}{4t-1} \\ &= c_1 + c_2. \end{aligned}$$

Conversely, let $\mathcal{S} = \mathcal{S}_G \cup \mathcal{S}_B$ be the instance of the Jaccard median problem and let M^* be one of its Jaccard medians of value at least $c_1 + c_2$.

Let $L^* = M^* \cap L$. We claim that $M^* \in \mathcal{M}_B^*$ and that the subgraph in G induced by the nodes in L^* is

a clique. Supposing the claim is true, it is easy to see that the reduction from $\frac{1}{3}$-Quasi-Regular-Clique is complete.

We now prove the claim. In particular, first we show that $M^* \in \mathcal{M}_B^*$. We have $J(M^*, \mathcal{S}) = J(M^*, \mathcal{S}_G) + J(M^*, \mathcal{S}_B)$. From Corollary 5.1 we know that $J(M^*, \mathcal{S}_B)$ is maximized when $M^* \in \mathcal{M}_B^*$ (with $J(M^*, \mathcal{S}_B) = c_1$ in that case), and for any $M^* \in \mathcal{M}_B^*$ and $M \notin \mathcal{M}_B^*$, $J(M^*, \mathcal{S}_B) - J(M, \mathcal{S}_B) \geq 320 \cdot t^5 \cdot t^{-2}/32 = 10t^3$. Further, we know that $J(M, \mathcal{S}_G) = \sum_{S_e \in \mathcal{S}_G} J(M, S_e) \leq |\mathcal{S}_G| \leq 9t^2$ for any M. Thus, for any $M \notin \mathcal{M}_B^*$ we have that

$$\begin{aligned} J(M, \mathcal{S}) &= J(M, \mathcal{S}_G) + J(M, \mathcal{S}_B) \\ &\leq 9t^2 + J(M^*, \mathcal{S}_B) - 10t^3 \leq c_1 - t^3, \end{aligned}$$

a contradiction. Hence, $M^* \in \mathcal{M}_B^*$.

Given this, we next claim $J(M^*, \mathcal{S}_G)$ has value c_2 if L^* is a clique, and value at most $c_2 - \frac{2}{(4t-2)(4t-1)}$ otherwise. Suppose $k \leq \binom{t}{2}$ edges of G are completely inside L^*. Then at most $\Delta t - 2k$ edges will have a single endpoint in L^*, since the maximum degree is Δ, and

$$\begin{aligned} J(M^*, \mathcal{S}_G) &= \sum_{\substack{S_e \in \mathcal{S}_G \\ |S_e \cap L^*| = 2}} \frac{2}{t + |R| + \sum_{\substack{S_e \in \mathcal{S}_G \\ |S_e \cap L^*| = 1}} \frac{1}{t + |R| + 1}} \\ &\leq k \cdot \frac{2}{4t-2} + (\Delta t - 2k) \cdot \frac{1}{4t-1} \\ &= k \cdot \frac{2}{(4t-2)(4t-1)} + \frac{\Delta t}{4t-1}. \end{aligned}$$

The latter equals c_2 if $k = \binom{t}{2}$. Also, if L^* is not a clique, then $k < \binom{t}{2}$ and $J(M^*, \mathcal{S}_G) \leq c_2 - \frac{2}{(4t-2)(4t-1)}$. Thus $J(M^*, \mathcal{S}) \leq c_1 + c_2 - \frac{2}{(4t-2)(4t-1)}$, a contradiction. □

Corollary 5.2. *The Jaccard median problem, where each set in the instance has two elements, does not admit an FPTAS if $P \neq NP$.*

Proof. In the proof of Theorem 5.1, we have shown it is NP-hard to approximate the Jaccard median problem to within an additive factor $\frac{2}{(4t-2)(4t-1)}$. In our instances, $m = \Theta(t^7)$ and $n = \Theta(t)$. Note that the number of sets m is an upper bound on the total Jaccard distance of any median. It follows that it is NP-hard to approximate the Jaccard median problem to within a multiplicative factor of $1 + o\left(m^{-9/7}\right)$ or $1 + o\left(n^{-9}\right)$. It follows that no FPTAS exists for the problem if P ≠ NP. □

5.2 The set, hyperedge case We show that the Jaccard median problem restricted to the case when $\mathcal{S}$ is not a multi-set, is NP-hard. However, we need that the sets in the instances have cardinalities more than two, i.e., they are like "hyperedges".

Theorem 5.2. *The Jaccard median problem, where the instance does not contain duplicate sets, is NP-hard.*

Proof. As before, we prove the NP-hardness by reducing from $\frac{1}{3}$-Quasi-Regular-Clique. The steps of the reduction are similar to the earlier case. Let $|V| = 3t$ and we consider $B = B_t = (L, R, C)$. Next we overlay the graph G onto L, bijectively mapping nodes in V to nodes in L and adding appropriate edges among the nodes in L according to E and let $B' = (L \cup R, C \cup E)$ be the resulting graph.

From B', we construct an instance of the Jaccard median problem, whereby for each edge $e = (u, v)$ in B' that came from B, we create the set $S_e = \{u, v\}$ and for each edge $e = (u, v)$ in B' that came from G, we create the set $S_e = \{u, v, \alpha_e^1, \ldots, \alpha_e^k\}$ where $k = t^7$. Since each edge has unique α_e^i's, these α nodes have degree one and we refer to them as *fake nodes* as they belong neither to G nor to B. Let $\mathcal{S}_B = \cup_{e \in C} S_e$ be the family corresponding to the edges in B and let $\mathcal{S}_G = \cup_{e \in E} S_e$ be the family corresponding to the edges in G. Let $\mathcal{S} = \mathcal{S}_G \cup \mathcal{S}_B$ be the instance of the Jaccard median problem and let M^* be its optimal Jaccard median. Lemma 5.4 will complete the reduction from $\frac{1}{3}$-Quasi-Regular-Clique. □

First we prove two simple facts about fake nodes. Let fake(M) denote the set of fake nodes in M.

Lemma 5.2. *For $t \geq 3$, if $|\text{fake}(M)| = O(t^2)$, then $J(M, \mathcal{S}_G) < 0.03$ and otherwise, $J(M, \mathcal{S}_G) < 3/2$.*

Proof. For each $e = (u, v) \in E$, let $T_e = M \cap \{u, v\}$ and let $F_e = (M \cap S_e) \setminus \{u, v\}$, i.e., T_e corresponds to the non-fake nodes and F_e corresponds to the fake nodes from set S_e that are present in M. Let $T = \cup_{e \in E} T_e$ and $F = \text{fake}(M) = \cup_{e \in G} F_e$. Then,

$$\begin{aligned} J(M, \mathcal{S}_G) &= \sum_{e \in E} J(M, S_e) \\ &= \sum_{e \in E} \left(\frac{|T_e \cap S_e|}{|T \cup F \cup S_e|} + \frac{|F_e \cap S_e|}{|T \cup F \cup S_e|} \right) \\ &\leq \sum_{e \in E} \left(\frac{2}{|T \cup F \cup S_e|} + \frac{|F_e|}{|T \cup F \cup S_e|} \right) \\ &\leq \frac{2|E|}{k} + \sum_{e \in E} \frac{|F_e|}{|F \cup S_e|} \\ &\leq \frac{18t^2}{k} + \frac{|F|}{\max\{|F|, k\}}. \end{aligned}$$

If $|F| = O(t^2)$, then since $k = t^7$ and $t \geq 3$, we have $J(M, \mathcal{S}_G) = O(t^{-5}) < 0.03$. Otherwise, $J(M, \mathcal{S}_G) < 18t^{-5} + 1 < 3/2$ for $t \geq 3$. □

LEMMA 5.3. *Let $M \subseteq L \cup R$, such that $J(M, \mathcal{S}_G) \geq 2t$. Let F be any non-empty set of fake nodes. If $|F| \leq 40t$, then $J(M, \mathcal{S}_B) - J(M \cup F, \mathcal{S}_B) \geq 0.035$ and if $|F| > 40t$, then $J(M^*, \mathcal{S}_B) - J(M^* \cup F, \mathcal{S}_B) \geq 1.55$.*

Proof. Let f be the number of edges in B with both endpoints in M and h be the number of edges in B with one endpoint in M. Then, $J(M^*, \mathcal{S}_B) = \frac{f}{4t-2} + \frac{h}{4t-1}$, and the condition on M implies that $f + h \geq 7t^2$.

Since the nodes in F do not have any edges in B, we know that $J(M \cup F, \mathcal{S}_B) = \frac{f}{4t-2+|F|} + \frac{h}{4t-1+|F|}$. Hence,

$$\begin{aligned} & J(M, \mathcal{S}_B) - J(M \cup F, \mathcal{S}_B) \\ = & \frac{f|F|}{(4t-2)(4t-2+|F|)} + \frac{h|F|}{(4t-1)(4t-1+|F|)} \\ \geq & \frac{f|F|}{(4t)(4t+|F|)} + \frac{h|F|}{(4t)(4t+|F|)} \\ = & \frac{|F|\frac{(f+h)}{t}}{4(4t+|F|)} \geq \frac{7t|F|}{4(4t+|F|)}. \end{aligned}$$

The proof is complete by simple calculations. □

LEMMA 5.4. *Given M^* as above, $M^* \in \mathcal{M}_B^*$ and the subgraph in G induced by the nodes in $L^* = M^* \cap L$ is a clique.*

Proof. First, we show that $\text{fake}(M^*) = \emptyset$. We do this by arguing that any M^* must have a high Jaccard similarity score on $\mathcal{S}_B$. Let $M_B^* = M^* \cap (L \cup R)$ denote the non-fake nodes in M^*. Suppose $J(M_B^*, \mathcal{S}_B) < 2t$. Then, using Lemma 5.2, we can conclude that $J(M^*, \mathcal{S}) = J(M^*, \mathcal{S}_B) + J(M^*, \mathcal{S}_G) < J(M_B^*, \mathcal{S}_B) + 1.5 \leq 2t + 1.5$, where $J(M^*, \mathcal{S}_B) < J(M_B^*, \mathcal{S}_B)$ since $\mathcal{S}_B$ does not contain any fake nodes. However, any solution $M' \in \mathcal{M}_B^*$ has $J(M', \mathcal{S}_B) > 2.1t$ for $t > 5$ (from Lemma 5.1), therefore M^* cannot be the optimum. On the other hand, if $J(M_B^*, \mathcal{S}_B) \geq 2t$, then $J(M_B^*, \mathcal{S}) - J(M^*, \mathcal{S}) = (J(M_B^*, \mathcal{S}_B) - J(M^*, \mathcal{S}_B)) + (J(M_B^*, \mathcal{S}_G) - J(M^*, \mathcal{S}_G))$. Lemmas 5.2 and 5.3 together show that $(J(M_B^*, \mathcal{S}) - J(M^*, \mathcal{S})) > 0$, which is a contradiction. Hence, M^* does not contain any fake nodes.

Next we show that $M^* \in \mathcal{M}_B^*$. From Lemma 5.1 we know that $J(M^*, \mathcal{S}_B)$ is maximized when $M^* \in \mathcal{M}_B^*$ and for any $M^* \in \mathcal{M}_B^*$ and $M \notin \mathcal{M}_B^*$, $J(M^*, \mathcal{S}_B) - J(M, \mathcal{S}_B) \geq t^{-2}/32$. Also, from Lemma 5.2, $J(M^*, \mathcal{S}_G) \leq 2|E|/k$ for any set M^* with $\text{fake}(M^*) = \emptyset$. Hence, for any $M^* \in \mathcal{M}_B^*$ and $M \notin \mathcal{M}_B^*$, $J(M^*, \mathcal{S}) - J(M, \mathcal{S}) = (J(M^*, \mathcal{S}_B) - J(M, \mathcal{S}_B)) + (J(M^*, \mathcal{S}_G) - J(M, \mathcal{S}_G)) \geq \frac{t^{-2}}{32} - \frac{2|E|}{k} > 0$, since $t \geq 3$ and $k = t^7$. In other words, our choice of parameters guarantees that $M^* \in \mathcal{M}_B^*$, thus, $|L^*| = t$.

Given this, we next claim $J(M^*, \mathcal{S}_G)$ (and therefore $J(M^*, \mathcal{S})$) is maximized when L^* induces a clique in G. In particular, let the induced graph contain f full-edges (i.e., edges with both end points in L^*) and h half-edges (i.e., edges with exactly one end point in L^* and the other end point in $L \setminus L^*$.) Since the degree of each node in G is bounded by Δ, it is easy to see that $h \leq (|L^*| \cdot \Delta) - 2f = t\Delta - 2f$. By definition,

$$\begin{aligned} J(M^*, \mathcal{S}_G) &= \frac{2f}{4t-2+k} + \frac{h}{4t-1+k} \\ &\leq \frac{2f}{4t-2+k} + \frac{t\Delta - 2f}{4t-1+k} \\ &= \frac{2f}{(4t-1+k)(4t-2+k)} + \frac{t\Delta}{4t-1+k} = c. \end{aligned}$$

Since c is increasing in f, it is maximized when $f = \binom{t}{2}$. Observe that $J(M^*, \mathcal{S}_G)$ actually equals this maximum value if L^* induces a clique since in that case $f = \binom{t}{2}$ and each of the nodes of L^* will have degree Δ and $h = t\Delta - 2f$. Hence, L^* is a clique if and only if $J(M^*, \mathcal{S}_G)$ is maximized. □

We note that the no-FPTAS claim also holds here.

6 Conclusions

In this paper we studied the median problem for the weighted Jaccard metric. We gave a PTAS that returns a $(1+\epsilon)$-approximate median in time $(nm)^{\frac{1}{\epsilon^{O(1)}}}$ and showed that the problem does not admit a FPTAS unless P = NP. Two interesting future directions include studying the complexity of the k-median problem for $k > 1$ and obtaining a PTAS for the similarity version of the Jaccard median. For the latter, we can (a) show that the trivial 2-approximation for the distance version is an $\Omega(\sqrt{m})$-approximation for the similarity version and (b) obtain a different 2-approximation algorithm.

References

[1] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

[2] J. Bartholdi, C. A. Tovey, and M. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice Welfare*, 6(2):157–165, 1989.

[3] A. Broder. On the resemblance and containment of documents. In *Proc. Compression and Complexity of Sequences*, pages 21–29, 1997.

[4] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *JCSS*, 60(3):630–659, 2000.

[5] M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th STOC*, pages 380–388, 2002.
[6] M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem. *JCSS*, 65(1):129–149, 2002.
[7] I. Dagan. Contextual word similarity. In R. Dale, H. Moisl, and H. Somers, editors, *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker Inc., 2000.
[8] C. de la Higuera and F. Casacuberta. Topology of strings: Median string is NP-complete. *TCS*, 230:39–48, 2000.
[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. 10th WWW*, pages 613–622, 2001.
[10] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
[11] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.
[12] R. A. M. Gregson. *Psychometrics of Similarity*. Academic Press, 1975.
[13] B. Huntley and H. J. B. Birks. The past and present vegetation of the Morrone Birkwoods national nature reserve. *J. Ecology*, 67(2):417–446, 1979.
[14] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901.
[15] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, 2008.
[16] B. Kendrick. Quantitative characters in computer taxonomy. *Phenetic and Phylogenetic Classification*, pages 105–114, 1964.
[17] H. R. Lasker. Light dependent activity patterns among reef corals: Montastrea cavernosa. *Biological Bulletin*, 156:196–211, 1979.
[18] E. Marczewski and H. Steinhaus. On a certain distance of sets and the corresponding distance of functions. *Colloquium Mathematicum*, 6:319–327, 1958.
[19] F. Nicolas and E. Rivals. Complexities of the centre and median string problems. In *Proc. 14th CPM*, pages 315–327, 2003.
[20] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15:251–270, 1991.
[21] S. Pandey, A. Z. Broder, F. Chierichetti, V. Josifovski, R. Kumar, and S. Vassilvitskii. Nearest-neighbor caching for content-match applications. In *Proc. 18th WWW*, pages 441–450, 2009.
[22] M. E. Patzkowsky and S. M. Holland. Biofacies replacement in a sequence stratigraphic framework: Middle and upper Ordovician of the Nashville dome, Tennessee, USA. *Palaios*, 14:301–323, 1999.
[23] I. C. Prentice. Non-metric ordination methods in ecology. *J. Ecology*, 65:85–94, 1977.
[24] J. J. Sepkoski. Quantified coefficients of association and measurement of similarity. *Math. Geology*, 6(2):131–152, 1974.
[25] H. Späth. The minisum location problem for the Jaccard metric. *OR Spektrum*, 3:91–94, 1981.
[26] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
[27] G. A. Watson. An algorithm for the single facility location problem using the Jaccard metric. *SIAM J. Sci. and Stat. Comput.*, 4:748–756, 1983.

Appendix

A Tightness of the two-approximation

Given a set of points in an arbitrary metric, one of the input points in the set is a $\left(2-\frac{2}{n}\right)$-approximation to the optimal median. Here, we show that this bound is tight for the Jaccard metric. Consider an instance of n sets, $\mathcal{S} = \{S_1, \ldots, S_n\}$, such that $S_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ for $i = 1, \ldots n$. Then, the distance between any two sets in the instance will be $1 - \frac{n-2}{n} = \frac{2}{n}$. Therefore, the optimal point (and in fact, any point) in the instance will act as a median with total distance $(n-1)\frac{2}{n} = 2 - \frac{2}{n}$. Now consider the set $M = \{x_1, \ldots, x_n\}$. The distance of M to an arbitrary $S_i \in \mathcal{S}$ will be $1 - \frac{n-1}{n} = \frac{1}{n}$. Its total distance will thus be 1. The claim follows.

B Hardness of $\frac{1}{3}$-Quasi-Regular-Clique

We prove the NP-hardness by reducing from $\frac{2}{3}$-Clique [10]: given a graph $G = (V, E)$, does G contain a clique of size at least $2\,|V|\,/3$?

We let $n = |V|$. Observe that if G has fewer than $2n/3$ nodes of degree at least $2n/3$, then we can conclude (in polynomial time) that the answer to the problem is no. Therefore, we assume the contrary: there exist at least $2n/3$ nodes of degree at least $2n/3$. Then, the sum of degrees of the nodes in V (denoted $\mathrm{vol}(V)$) is at least $4n^2/9$. If we let Δ denote the maximum degree of G, we have that $2n/3 \leq \Delta < n$. Also, $\mathrm{vol}(V) \leq \Delta n < n^2$.

We create a new graph $G' = (V', E')$ that contains G as a node subgraph (i.e., $V \subseteq V'$ and $E \subseteq E'$). V' will also contain n new nodes and hence $|V'| = 2n$. E' will contain all the edges in E, and possibly some new edges going from the nodes in V to the ones in $V' \setminus V$; these new edges will be added as follows. As long as there exists some node $v \in V$ such that $\deg_{G'}(v) < \Delta$, we choose an arbitrary node $v' \in V' \setminus V$ such that $\deg_{G'}(v') \leq 5n/9$ and add the edge $\{v, v'\}$ to E'. Observe that such a node v' always exists: each time we add an edge, we increase the total degree of V and since $4n^2/9 \leq \mathrm{vol}(V) \leq n^2$, we have that no more

than $5n^2/9$ edges will need to be added. Further, since all n nodes in $V' \setminus V$ had degree 0 in the beginning, it is possible to add at most $5n^2/9$ edges, with each edge having a single endpoint in $V' \setminus V$, in such a way that the maximum degree in $V' \setminus V$ remains upper bounded by $5n/9$.

In the end, for each $v \in V$, we will have $\deg_{G'}(v) = \Delta \geq 2n/3 = |V'|/3$ and for each $v' \in V' \setminus V$, we have $\deg_{G'}(v') \leq 5n/9 = 5\,|V'|/18$.

We claim that G has a clique of size at least $2n/3$ if and only if G' has a clique of size at least $2n/3 = |V'|/3$.

Indeed, if G had such a clique C, then $C \subseteq V$ will also be a clique in G'. On the other hand, suppose there exists a clique $C' \subseteq V'$ in G' of size at least $2n/3$. Then, by the upper bound on the degree of the nodes in $V' \setminus V$, C' must be composed only of nodes in V. But then, by construction, C' will also be a clique in G.

C Algorithms for the weighted Jaccard metric

C.1 A PTAS when the optimal median is small

We present an algorithm that returns a $1 + O(\epsilon)$-approximate median if the optimal median has total distance at most ϵ.

LEMMA C.1. *Let M^* be the optimal median and suppose that $D(M^*, \mathcal{V}) \leq \epsilon$. Then, there exists a polynomial-time algorithm for finding a median M such that $D(M, \mathcal{V}) \leq (1 + \frac{\epsilon}{1-\epsilon}) \cdot D(M^*, \mathcal{V})$.*

Proof. If two generic A, B vectors have Jaccard distance at most δ, it must be that

$$\sum_i \min\left(A(i), B(i)\right) \geq (1-\delta) \sum_i \max\left(A(i), B(i)\right),$$

so that

$$\sum_i A(i) \geq (1-\delta) \cdot \sum_i \max(A(i), B(i)). \tag{C.1}$$

Now, consider two vectors A', B' and suppose

$$D(A', B') = \frac{\sum_i |A'(i) - B'(i)|}{\sum_i \max(A'(i), B'(i))} \leq \epsilon.$$

Then, from (C.1), we have

$$\frac{1}{1-\epsilon} \sum_i A'(i) \geq \sum_i \max(A'(i), B'(i)),$$

and

$$\frac{\sum_i |A'(i) - B'(i)|}{\frac{1}{1-\epsilon} \cdot \sum_i A'(i)} \leq \frac{\sum_i |A'(i) - B'(i)|}{\sum_i \max(A'(i), B'(i))} \leq \epsilon,$$

Thus,

$$\sum_i |A'(i) - B'(i)| \leq \frac{\epsilon}{1-\epsilon} \sum_i A'(i). \tag{C.2}$$

Further, observe that if we have two vectors A'', B'' such that $\sum_i |A''(i) - B''(i)| \leq \frac{\epsilon}{1-\epsilon} \sum_i A''(i)$ then,

$$D(A'', B'') = \frac{\sum_i |A''(i) - B''(i)|}{\sum_i \max(A''(i), B''(i))}$$
$$\leq \frac{\sum_i |A''(i) - B''(i)|}{\sum_i A''(i)} \leq \frac{\frac{\epsilon}{1-\epsilon} \sum_i A''(i)}{\sum_i A''(i)} = \frac{\epsilon}{1-\epsilon}.$$

Now consider the following linear program:

$$\begin{cases} \mathbf{m_i} \geq 0 & \forall \text{ coordinate } i \\ \mathbf{t_i^j} \geq |\mathbf{m_i} - V_j(i)| & \forall V_j \in \mathcal{V}, \forall \text{ coordinate } i \\ \sum_i \mathbf{t_i^j} \leq \frac{\epsilon}{1-\epsilon} \sum_i V_j(i) & \forall V_j \in \mathcal{V} \\ \min \sum_j \frac{1}{\sum_i V_j(i)} \sum_i \mathbf{t_i^j}. & \end{cases}$$

(The inequality $\mathbf{t_i^j} \geq |\mathbf{m_i} - V_j(i)|$ can be replaced by two inequalities $\mathbf{t_i^j} \geq \mathbf{m_i} - V_j(i)$ and $\mathbf{t_i^j} \geq V_j(i) - \mathbf{m_i}$.)

We claim that if an optimal median M^* for $\mathcal{V}$ has total distance $D(M^*, \mathcal{V}) \leq \epsilon$, then the linear program is feasible and each of its optimal solutions are $(1 + \frac{\epsilon}{1-\epsilon})$ approximations to $D(M^*, \mathcal{V})$, i.e., if $\mathbf{M^*} = (\mathbf{m_1^*}, \ldots, \mathbf{m_n^*})$ is an optimal solution to the linear program, then $D(\mathbf{M^*}, \mathcal{V}) \leq (1 + \frac{\epsilon}{1-\epsilon}) D(M^*, \mathcal{V})$.

To show that the linear program is feasible, take $\mathbf{m_i} = M^*(i)$ for each i, and $\mathbf{t_i^j} = |M^*(i) - V_j(i)|$ for each i, j. Since $D(M^*, \mathcal{V}) \leq \epsilon$ it must be that, for each $V_j \in \mathcal{V}$, $D(M^*, V_j) \leq \epsilon$. Then, setting $A' := V_j$ and $B' := M^*$ in (C.2), we obtain

$$\sum_i \mathbf{t_i^j} = \sum_i |M^*(i) - V_j(i)| \leq \frac{\epsilon}{1-\epsilon} \sum_i V_j(i),$$

so all of the constraints are satisfied. The value of the objective function is:

$$f^* = \sum_j \frac{\sum_i |\mathbf{m_i} - V_j(i)|}{\sum_i V_j(i)} = \sum_j \frac{\sum_i |M^*(i) - V_j(i)|}{\sum_i V_j(i)}.$$

For each j we apply (C.1) with $A := V_j$, $B := M^*$, and $\delta = \epsilon$, obtaining $\sum_i V_j(i) \geq (1 - \epsilon) \sum_i \max(V_j(i), M^*(i))$. Then,

$$f^* \leq \frac{1}{1-\epsilon} \sum_j \frac{\sum_i |M^*(i) - V_j(i)|}{\sum_i \max(V_j(i), M^*(i))} = \frac{1}{1-\epsilon} D(M^*, \mathcal{V}).$$

Now take any optimal solution to the linear program: $\mathbf{M^*} = (\mathbf{m_1^*}, \ldots, \mathbf{m_n^*})$. Consider the function that the linear program is minimizing,

$$f = \sum_j \frac{\sum_i \mathbf{t_i^j}}{\sum_i V_j(i)}.$$

Since $\mathbf{M^*}$ is optimal we have $\mathbf{t_i^j} = |\mathbf{m_i^*} - V_j(i)|$, for each i, j, and

$$f = \sum_j \frac{\sum_i |\mathbf{m_i^*} - V_j(i)|}{\sum_i V_j(i)}.$$

Observe that if we were to use the vector $\mathbf{M}^*$ as a median, we would have total distance

$$D(\mathbf{M}^*, \mathcal{V}) = \sum_j \frac{\sum_i |\mathbf{m_i^*} - V_j(i)|}{\sum_i \max(\mathbf{m_i^*}, V_j(i))} \leq \sum_j \frac{\sum_i |\mathbf{m_i^*} - V_j(i)|}{\sum_i V_j(i)} = f.$$

Further, since f is optimal, and f^* is feasible, we will have $f \leq f^*$, and

$$D(\mathbf{M}^*, \mathcal{V}) \leq f \leq f^* \leq \frac{1}{1-\epsilon} D(M^*, \mathcal{V}),$$

so $\mathbf{M}^*$ is an $\frac{1}{1-\epsilon}$-approximate median. □

C.2 A PTAS when the optimal median is large

When the optimal median is large, we consider two different approaches, depending on the *s*pread of the instance.

Given an input set $\mathcal{V}$, not all null, let α be the minimum non-zero coordinate value, $\alpha = \alpha_{\mathcal{V}} = \min_{V \in \mathcal{V}, 1 \leq i \leq n, V(i) > 0} V(i)$, and let β be their maximum coordinate value, $\beta = \beta_{\mathcal{V}} = \max_{V \in \mathcal{V}, 1 \leq i \leq n} V(i)$.

Observe that if all the input vectors are zero vectors, then the input is a set instance, and then the optimal median is trivially the all-zeros vector. Otherwise both α and β are well-defined, and we define the *spread* of $\mathcal{V}$ as $\sigma = \beta/\alpha$.

C.2.1 Instances with polynomial spread.

Suppose that the spread is polynomial, $\sigma \leq (nm)^{O(1)}$.

Scale the vectors by α^{-1} and obtain the multi-set of vectors $\mathcal{V}_\alpha = \{\alpha^{-1} \cdot V \mid V \in \mathcal{V}\}$. Then, the minimum non-zero coordinate in $\mathcal{V}_\alpha$ is 1, and the maximum is σ. Let $\xi > 0$ be sufficiently small and define $k = \lceil \xi^{-1} \rceil$. Observe that $k^{-1} \leq \xi$. Given a vector V on n coordinates, with each coordinate value $\leq \sigma$, we define its *expansion* $e_{k,\sigma}(V) = e(V)$ as a binary vector on $nk\lceil\sigma\rceil$ coordinates, as follows:

$$e(V) = (\underbrace{1, 1, \ldots, 1}_{\substack{t_1 = \lceil kV(1) \rceil \\ \text{times}}}, \underbrace{0, 0, \ldots, 0}_{\substack{\lceil k\sigma \rceil - t_1 \\ \text{times}}}, \ldots, \underbrace{1, 1, \ldots, 1}_{\substack{t_n = \lceil kV(n) \rceil \\ \text{times}}}, \underbrace{0, 0, \ldots, 0}_{\substack{\lceil k\sigma \rceil - t_n \\ \text{times}}}).$$

We then use the PTAS for binary instances (Section 3) to obtain a $(1+\epsilon)$-approximation of the following binary instance:

$$\mathcal{V}_S = \{e_{k,\sigma}(V) \mid V \in \mathcal{V}_\alpha\} = \{e_{k,\sigma}(\alpha^{-1}V) \mid V \in \mathcal{V}\}.$$

We show that distances are preserved by this expansion.

LEMMA C.2. *Let V, W be any two non-negative real vectors, having minimum coordinate value $\geq \alpha$ and maximum coordinate value $\leq \beta$. Let $\xi > 0$ be sufficiently small. Let $\sigma = \beta/\alpha$, and $k = \lceil \xi^{-1} \rceil$. Then,*

$$D(V, W) - \xi \leq D(e_{k,\sigma}(\alpha^{-1}V), e_{k,\sigma}(\alpha^{-1}W)) \leq D(V, W) + \xi.$$

Proof. If $V = W$, then the claim is trivial as they will both be mapped to the same vector. Otherwise, $D(V, W) > 0$, and it is easy to see that $D(V, W) = D(\alpha^{-1}V, \alpha^{-1}W)$.

Now, let $V' = e_{k,\sigma}(\alpha^{-1}V)$. For any $i = 1, \ldots, n$, consider

$$V'_i = \sum_{j=(i-1)\lceil k\sigma \rceil + 1}^{i\lceil k\sigma \rceil} V'(j).$$

Then, $\frac{1}{k} V'_i = \frac{1}{k} \lceil k\alpha^{-1}V(i) \rceil \leq \alpha^{-1}V(i) + k^{-1} \leq \alpha^{-1}V(i) + \xi$, and $\frac{1}{k} V'_i \geq \alpha^{-1}V(i)$. As $\alpha \leq V(i)$ by definition, we have that $\frac{\xi}{\alpha^{-1}V(i)} \leq \xi$. Thus,

$$\alpha^{-1}V(i) \leq V'_i \leq (1+\xi)\alpha^{-1}V(i).$$

Analogously, if $W' = e_{k,\sigma}(\alpha^{-1}W)$, then we have

$$\alpha^{-1}W(i) \leq W'_i \leq (1+\xi)\alpha^{-1}W(i).$$

Then, $D(V', W') =$

$$\begin{aligned}
&= 1 - \frac{\sum_{i=1}^n \min\{V'(i), W'(i)\}}{\sum_{i=1}^n \max\{V'(i), W'(i)\}} \\
&\leq 1 - \frac{\sum_{i=1}^n \min\{\alpha^{-1}V(i), \alpha^{-1}W(i)\}}{\sum_{i=1}^n \max\{(1+\xi)\alpha^{-1}V(i), (1+\xi)\alpha^{-1}W(i)\}} \\
&= 1 - \frac{1}{1+\xi} \frac{\sum_{i=1}^n \min\{V(i), W(i)\}}{\sum_{i=1}^n \max\{V(i), W(i)\}} \\
&= 1 - \frac{1}{1+\xi}(1 - D(V, W)) \\
&= \frac{1}{1+\xi} D(V, W) + \frac{\xi}{1+\xi} \\
&\leq D(V, W) + \xi.
\end{aligned}$$

And,

$$\begin{aligned}
D(V', W') &= 1 - \frac{\sum_{i=1}^n \min\{V'(i), W'(i)\}}{\sum_{i=1}^n \max\{V'(i), W'(i)\}} \\
&\geq 1 - (1+\xi) \frac{\sum_{i=1}^n \min\{V(i), W(i)\}}{\sum_{i=1}^n \max\{V(i), W(i)\}} \\
&= -\xi + (1+\xi) D(V, W) \geq D(V, W) - \xi.
\end{aligned}$$

□

To complete the proof, fix $\xi = \epsilon^2/m$. Then the approximate median M of the binary instance returned by the algorithm in Section 3 will be a binary vector with $(n \lceil \xi^{-1} \rceil \lceil \sigma \rceil)$ coordinates. Lemma C.3 shows how one can round M to find a a $(1+\epsilon)$-approximation in polynomial time.

LEMMA C.3. *Fix $\mathcal{V}$, and let M be a $(1+\epsilon)$-approximate median for $\mathcal{V}_S$. A real vector M' such that $D(M', \mathcal{V}) \leq D(M, \mathcal{V}_S) + \xi m$ can be found in time $(nm\xi^{-1}\sigma)^{O(1)}$.*

Proof. Let $w_i = \sum_{j=(i-1)\lceil k\sigma\rceil+1}^{i\lceil k\sigma\rceil} M(j)$, be the number of 1's in the block of coordinates of M corresponding to the ith coordinate of the original real vector space. Set $w'_i = \max(w_i, k)$.

We create a binary vector A from M, by pushing all the 1's of M on the left side of their respective blocks:

$$A = (\underbrace{1,1,\ldots,1}_{\substack{w'_1\\ \text{times}}},\underbrace{0,0,\ldots,0}_{\substack{\lceil k\sigma\rceil - w'_1\\ \text{times}}},\ldots,\underbrace{1,1,\ldots,1}_{\substack{w'_n\\ \text{times}}},\underbrace{0,0,\ldots,0}_{\substack{\lceil k\sigma\rceil - w'_n\\ \text{times}}}).$$

Observe that for each $V_S \in \mathcal{V}_S$, we will have $D(A, V_S) \leq D(M, V_S)$. This follows from the fact that each such V_S has all of its 1 coordinates on the left sides of its blocks, and that each V_S has at least k many 1's in each block.

Further, observe that A is the $e_{k,\sigma}$ expansion of the vector $\frac{1}{k}(w'_1, \ldots, w'_n)$. Let $M' = \alpha\frac{1}{k}(w'_1, \ldots, w'_n)$. By Lemma C.2, we have that, for each real vector $V \in \mathcal{V}$,

$$D(M', V) - \xi \leq D(A, e_{k,\sigma}(\alpha^{-1}V)),$$

or, equivalently,

$$D(M', V) \leq D(A, e_{k,\sigma}(\alpha^{-1}V)) + \xi.$$

Thus,

$$\begin{aligned} D(M', \mathcal{V}) &= \sum_{V\in\mathcal{V}} D(M', V) \\ &\leq \sum_{V\in\mathcal{V}} \left(D(A, e_{k,\sigma}(\alpha^{-1}V)) + \xi\right) \\ &= \sum_{V_S\in\mathcal{V}_S} (D(A, V_S) + \xi) \\ &\leq \sum_{V_S\in\mathcal{V}_S} (D(M, V_S) + \xi) \\ &= D(M, \mathcal{V}_S) + \xi m. \qquad \square \end{aligned}$$

C.2.2 Instances with arbitrary spread. Let $\mathcal{V}$ be an arbitrary Jaccard median instance. To compute the median of $\mathcal{V}$, we start by guessing the largest coordinate value of one of its optimal medians (observe that by [25], and Lemma D.1, this coordinate value will be shared with the median by at least one input vector). First, we remove all the sets that would be too far to a median having such a (large) coordinate value (these would be the sets having too small coordinate values). Next, we set to zero those coordinate values that were much smaller than our guess (by doing this, we do not distort distances by much). This way, we obtain an instance having polynomial spread and apply the algorithm from Section C.2.1.

More precisely, for each input coordinate value α (there are at most nm such values), we

- Remove all sets having a coordinate value larger than $\alpha\frac{n}{\epsilon}$, or having total weight less than $\epsilon\alpha$ obtaining the class $\mathcal{V}_\alpha$,
$$\mathcal{V}_\alpha = \left\{V_j \in \mathcal{V} \mid \wedge \max_i V_j(i) \leq \alpha\frac{n}{\epsilon} \wedge \sum_i V_j(i) \geq \epsilon\alpha\right\}.$$
- For each vector $V_j \in \mathcal{V}_\alpha$, set to zero all of its coordinates having value at most $\alpha\frac{\epsilon^2}{nm}$, obtaining a vector V'_j,
$$V'_j(i) = \begin{cases} 0 & \text{if } V_j(i) \leq \frac{\alpha\epsilon^2}{nm} \\ V_j(i) & \text{otherwise} \end{cases}$$
- Finally, let $\mathcal{V}'_\alpha$ be the resulting instance,
$$\mathcal{V}'_\alpha = \left\{V'_j \mid V_j \in \mathcal{V}_\alpha\right\}.$$

The spread of $\mathcal{V}'_\alpha$ will be at most $\frac{n^2m^2}{\epsilon^3}$. We then apply the polynomial spread algorithm (Section C.2.1) to obtain a $(1 + O(\epsilon))$-approximate median M for $\mathcal{V}_\alpha$. We now show that, given the appropriate choice of α, M will be an approximately optimal median for $\mathcal{V}$.

LEMMA C.4. *Let M^* be an optimal median for $\mathcal{V}$, and let $\alpha = \max_i M^*(i)$. If M is a $(1 + O(\epsilon))$-approximate median for $\mathcal{V}'_\alpha$, then*

$$D(M, \mathcal{V}) \leq (1 + O(\epsilon))D(M^*, \mathcal{V}) + O(\epsilon).$$

Proof. We start by showing that M is an approximate median for $\mathcal{V}_\alpha$. The observation that M^* is at distance at least $1 - \epsilon$ to each vector in $\mathcal{V} - \mathcal{V}_\alpha$ will complete the proof, since any median is at distance at most 1 from each vector in $\mathcal{V} - \mathcal{V}_\alpha$.

Let W be any non-negative vector on n coordinates. First of all, observe that, for each $V'_j \in \mathcal{V}'_\alpha$, we have

$$\begin{aligned} \sum_i \max(V_j(i), W(i)) &\geq \sum_i \max(V'_j(i), W(i)) \\ &\geq \sum_i \max(V_j(i), W(i)) - \frac{\alpha\epsilon^2}{m}, \end{aligned}$$

and $\sum_i \max(V_j(i), W(i)) \geq \sum_i V_j(i) \geq \epsilon\alpha$. Therefore,

$$\begin{aligned} \sum_i \max(V_j(i), W(i)) &\geq \sum_i \max(V'_j(i), W(i)) \\ &\geq \left(1 - \frac{\epsilon}{m}\right) \sum_i \max(V_j(i), W(i)). \end{aligned}$$

Further,

$$\sum_i |V_j(i) - W(i)| + \frac{\alpha\epsilon^2}{m} \geq \sum_i \left|V'_j(i) - W(i)\right| \geq \sum_i |V_j(i) - W(i)| - \frac{\alpha\epsilon^2}{m}.$$

We now show that the values of a median M' for $\mathcal{V}_\alpha$ and $\mathcal{V}'_\alpha$ are very close to each other. We start with an upper bound.

$$\begin{aligned}
&D(M', \mathcal{V}'_\alpha) \\
&= \sum_{V'_j \in \mathcal{V}'_\alpha} D(M', V'_j) \\
&= \sum_{V'_j \in \mathcal{V}'_\alpha} \frac{\sum_i \left|V'_j(i) - M'(i)\right|}{\sum_i \max(V'_j(i), M'(i))} \\
&\leq \frac{1}{1-\frac{\epsilon}{m}} \sum_{V'_j \in \mathcal{V}'_\alpha} \frac{\sum_i |V_j(i) - M'(i)| + \frac{\alpha\epsilon^2}{m}}{\sum_i \max(V_j(i), M'(i))} \\
&= \frac{1}{1-\frac{\epsilon}{m}} \sum_{V'_j \in \mathcal{V}'_\alpha} \left(D(M', V_j) + \frac{\frac{\alpha\epsilon^2}{m}}{\sum_i \max(V_j(i), M'(i))} \right) \\
&\leq \frac{1}{1-\frac{\epsilon}{m}} \sum_{V'_j \in \mathcal{V}'_\alpha} \left(D(M', V_j) + \frac{\frac{\alpha\epsilon^2}{m}}{\epsilon\alpha} \right) \\
&\leq \frac{1}{1-\frac{\epsilon}{m}} \sum_{V'_j \in \mathcal{V}'_\alpha} D(M', V_j) + \frac{\epsilon}{1-\frac{\epsilon}{m}} \\
&\leq \frac{1}{1-\epsilon} D(M', \mathcal{V}_\alpha) + \frac{\epsilon}{1-\epsilon}.
\end{aligned}$$

For the lower bound,

$$\begin{aligned}
&D(M', \mathcal{V}'_\alpha) = \\
&= \sum_{V'_j \in \mathcal{V}'_\alpha} \frac{\sum_i \left|V'_j(i) - M'(i)\right|}{\sum_i \max(V'_j(i), M'(i))} \\
&\geq \sum_{V'_j \in \mathcal{V}'_\alpha} \frac{\sum_i |V_j(i) - M'(i)| - \frac{\alpha\epsilon^2}{m}}{\sum_i \max(V_j(i), M'(i))} \\
&\geq D(M', \mathcal{V}_\alpha) - \sum_{V'_j \in \mathcal{V}'_\alpha} \frac{\frac{\alpha\epsilon^2}{m}}{\sum_i \max(V_j(i), M'(i))} \\
&\geq D(M', \mathcal{V}_\alpha) - \epsilon.
\end{aligned}$$

Now, consider an optimal median M^* for $\mathcal{V}$. Using the upper bound with $M' := M^*$, we obtain

$$D(M^*, \mathcal{V}'_\alpha) \leq (1 + O(\epsilon)) D(M^*, \mathcal{V}_\alpha) + O(\epsilon).$$

Since M is a $(1 + O(\epsilon))$-approximate median for $\mathcal{V}'_\alpha$, we will also have $D(M, \mathcal{V}'_\alpha) \leq (1 + O(\epsilon)) D(M^*, \mathcal{V}'_\alpha)$. Thus,

$$D(M, \mathcal{V}'_\alpha) \leq (1 + O(\epsilon)) D(M^*, \mathcal{V}_\alpha) + O(\epsilon).$$

Finally, since $D(M, \mathcal{V}'_\alpha) \geq D(M, \mathcal{V}_\alpha) - \epsilon$, we obtain that

$$D(M, \mathcal{V}_\alpha) \leq (1 + O(\epsilon)) D(M^*, \mathcal{V}_\alpha) + O(\epsilon).$$

Now consider the vectors in $\mathcal{V} - \mathcal{V}_\alpha$. Let $V_j \in \mathcal{V} - \mathcal{V}_\alpha$. If $\max_i V_j(i) \geq \alpha \frac{n}{\epsilon}$, then

$$\sum_i \max(M^*(i), V_j(i)) \geq \sum_i V_j(i) \geq \max_i V_j(i) \geq \alpha n \epsilon^{-1}.$$

Further,

$$\sum_i \min(M^*(i), V_j(i)) \leq \sum_i M^*(i) \leq \sum_i \alpha = n\alpha.$$

Then,

$$D(M^*, V_j) = 1 - \frac{\sum_i \min(M^*(i), V_j(i))}{\sum_i \max(M^*(i), V_j(i))} \geq 1 - \epsilon.$$

If we have $\sum_i V_j(i) < \epsilon\alpha$, then

$$\sum_i \max(M^*(i), V_j(i)) \geq \sum_i M^*(i) \geq \max_i M^*(i) = \alpha.$$

On the other hand,

$$\sum_i \min(M^*(i), V_j(i)) \leq \sum_i V_j(i) \leq \epsilon\alpha.$$

Again, these imply $D(M^*, \mathcal{V} - \mathcal{V}_\alpha) \geq 1 - \epsilon$.

The maximum Jaccard distance is 1. Therefore, $D(X, \mathcal{V} - \mathcal{V}_\alpha) \leq (1 + O(\epsilon)) D(M^*, \mathcal{V} - \mathcal{V}_\alpha)$ for each vector X, in particular for $X = M$. Putting everything together, we get

$$\begin{aligned}
D(M, \mathcal{V}) &= D(M, \mathcal{V}_\alpha) + D(M, \mathcal{V} - \mathcal{V}_\alpha) \\
&\leq ((1 + O(\epsilon)) D(M^*, \mathcal{V}_\alpha) + O(\epsilon)) + \\
&\quad + ((1 + O(\epsilon)) D(M^*, \mathcal{V} - \mathcal{V}_\alpha)) \\
&= (1 + O(\epsilon)) D(M^*, \mathcal{V}) + O(\epsilon). \qquad \square
\end{aligned}$$

D Canonical medians

As first observed by Späth [25] every value in the optimal median is present in one of the input vectors. We call such medians *canonical*, and in this section we give a simple polynomial rounding technique that transforms non-canonical medians to canonical medians without decreasing their value. Formally, we say that a median M for $\mathcal{V}$ is *canonical* if, for each i, $M(i) = V(i)$ for some $V \in \mathcal{V}$. Notice that the median obtained by our PTAS may not be canonical.

We will show in Lemma D.1 that every non-canonical median can be transformed into a canonical one of smaller or equal total distance. Thus, we can conclude that the value of the optimal medians is the same, whether or not we require the output to be canonical. Moreover, if P $\neq$ NP, no FPTAS exists for the not-necessarily canonical median problem, either.

The main argument here is quite similar to that of Späth [25], who shows how each optimal Jaccard median is canonical. We present this argument for completeness.

LEMMA D.1. *Let M be a median for $\mathcal{V}$. Suppose there exists a coordinate j such that $M(j) \notin \{V(j) \mid V \in \mathcal{V}\}$.*

i. If $M(j) > \max_{V\in\mathcal{V}} V(j)$, then

$$M_j^- = \Big(M(1), \ldots, M(j-1), \max_{\substack{V\in\mathcal{V}\\ V(j)<M(j)}} V(j), M(j+1), \ldots, M(n)\Big)$$

is a better median than M.

ii. If $M(j) < \min_{V\in\mathcal{V}} V(j)$, then

$$M_j^+ = \Big(M(1), \ldots, M(j-1), \min_{\substack{V\in\mathcal{V}\\ V(j)>M(j)}} V(j), M(j+1), \ldots, M(n)\Big)$$

is a better median than M.

iii. otherwise, either M_j^- or M_j^+ is a better median than M.

Proof. The first two cases are easy. If $M(j) > \max_{V\in\mathcal{V}} V(j)$, then for each $V \in \mathcal{V}$, it holds that $\max(V(j), M(j)) = M(j) > M_j^-(j) = \max(V(j), M_j^-(j))$ and $\min(V(j), M(j)) = V(j) = \min(V(j), M_j^-(j))$. I.e.,

$$D(M,V) = 1 - \frac{\sum_j \min(V(j), M(j))}{\sum_j \max(V(j), M(j))} \geq 1 - \frac{\sum_j \min(V(j), M_j^-(j))}{\sum_j \max(V(j), M_j^-(j))} = D(M_j^-, V).$$

For the second case observe that if $M(j) < \min_{V\in\mathcal{V}} V(j)$, then for each $V \in \mathcal{V}$, it holds that $\max(V(j), M(j)) = V(j) = \max(V(j), M_j^+(j))$ and $\min(V(j), M(j)) \leq M(j) \leq M_j^+(j) = \min(V(j), M_j^+(j))$. Again, we obtain $D(M,V) \geq D(M_i^+, V)$.

Consider the third case. Let M_i' be such that $M_i'(j) = M(j)$, for each $j \neq i$. We define $s_{V,i} = \sum_{j\neq i} \min(V(j), M(j))$ and $S_{V,i} = \sum_{j\neq i} \max(V(j), M(j))$. Then,

$$D(V,M) = 1 - \frac{\min(V(i), M(i)) + s_{V,i}}{\max(V(i), M(i)) + S_{V,i}}, \text{ and}$$

$$D(V,M_i') = 1 - \frac{\min(V(i), M_i'(i)) + s_{V,i}}{\max(V(i), M_i'(i)) + S_{V,i}}.$$

Further, let $\mathcal{V}_< = \{V \mid V \in \mathcal{V} \wedge V(i) < M(i)\}$ and $\mathcal{V}_> = \{V \mid V \in \mathcal{V} \wedge V(i) > M(i)\}$. Observe that $\mathcal{V}_< \cup \mathcal{V}_> = \mathcal{V}$. Define $\delta = D(M,\mathcal{V}) - D(M_i',\mathcal{V})$. Then,

$$\begin{aligned}
\delta &= D(M,\mathcal{V}) - D(M_i',\mathcal{V}) \\
&= \sum_{V\in\mathcal{V}} \left(\frac{\min(V(i), M_i'(i)) + s_{V,i}}{\max(V(i), M_i'(i)) + S_{V,i}} - \frac{\min(V(i), M(i)) + s_{V,i}}{\max(V(i), M(i)) + S_{V,i}} \right) \\
&= \sum_{V\in\mathcal{V}_<} \left(\frac{\min(V(i), M_i'(i)) + s_{V,i}}{\max(V(i), M_i'(i)) + S_{V,i}} - \frac{\min(V(i), M(i)) + s_{V,i}}{\max(V(i), M(i)) + S_{V,i}} \right) + \\
&\quad + \sum_{V\in\mathcal{V}_>} \left(\frac{\min(V(i), M_i'(i)) + s_{V,i}}{\max(V(i), M_i'(i)) + S_{V,i}} - \frac{\min(V(i), M(i)) + s_{V,i}}{\max(V(i), M(i)) + S_{V,i}} \right) \\
&= \sum_{V\in\mathcal{V}_<} \left(\frac{V(i) + s_{V,i}}{M_i'(i) + S_{V,i}} - \frac{V(i) + s_{V,i}}{M(i) + S_{V,i}} \right) + \\
&\quad + \sum_{V\in\mathcal{V}_>} \left(\frac{M_i'(i) + s_{V,i}}{V(i) + S_{V,i}} - \frac{M(i) + s_{V,i}}{V(i) + S_{V,i}} \right) \\
&= \sum_{V\in\mathcal{V}_<} \left((V(i) + s_{V,i}) \left(\frac{1}{M_i'(i) + S_{V,i}} - \frac{1}{M(i) + S_{V,i}} \right) \right) + \\
&\quad + \sum_{V\in\mathcal{V}_>} \frac{M_i'(i) - M(i)}{V(i) + S_{V,i}} \\
&= \sum_{V\in\mathcal{V}_<} \left((V(i) + s_{V,i}) \frac{M(i) - M_i'(i)}{(M_i'(i) + S_{V,i})\,(M(i) + S_{V,i})} \right) + \\
&\quad + \sum_{V\in\mathcal{V}_>} \frac{M_i'(i) - M(i)}{V(i) + S_{V,i}} \\
&= (M_i'(i) - M(i)) \Big(\sum_{V\in\mathcal{V}_>} \frac{1}{V(i) + S_{V,i}} - \\
&\quad + \sum_{V\in\mathcal{V}_<} \frac{V(i) + s_{V,i}}{(M_i'(i) + S_{V,i})\,(M(i) + S_{V,i})} \Big).
\end{aligned}$$

Let $A = \sum_{V\in\mathcal{V}_>} \frac{1}{V(i)+S_{V,i}}$ and $B(x) = \sum_{V\in\mathcal{V}_<} \frac{V(i)+s_{V,i}}{(x+S_{V,i})(M(i)+S_{V,i})}$. Observe that $0 < x_1 < x_2$

implies $B(x_1) > B(x_2)$. Then,

$$\delta \quad = \quad (M_i'(i) - M(i)) \left(A - B_{M_i'(i)}\right).$$

We will either choose $M_i' = M_i^+$ or $M_i' = M_i^-$. Suppose that $A - B_{M_i^+} > 0$. Then, choosing $M_i' = M_i^+$ will guarantee that $\delta > 0$ (as $M_i^+(i) - M(i) > 0$) and therefore $D(M, \mathcal{V}) > D(M_i^+, \mathcal{V})$.

On the other hand, if $A - B_{M_i^+} < 0$, then we will also have $A - B_{M_i^-} < 0$, by $B_{M^-(i)} \geq B_{M^+(i)}$. Thus, choosing $M_i' = M_i^-$ will give $D(M, \mathcal{V}) > D(M_i^-, \mathcal{V})$ (as $\delta > 0$, by $M_i^-(i) - M(i) < 0$). □

The proof gives an easy rounding algorithm to transform non-canonical medians into canonical medians. Suppose M is non-canonical on some coordinate j. Then either M_j^+ or M_j^- are better medians than M, in which case update M to be the optimal between M_j^+ and M_j^-. After iterating over all non-canonical coordinates completes the proof we obtain a canonical median no worse than the original median M.