Noname manuscript No.
(will be inserted by the editor)

Polynomial Time Approximation Schemes for All 1-Center Problems on Metric Rational Set Similarities

Marc Bury · Michele Gentili · Chris Schwiegelshohn · Mara Sorella

Received: date / Accepted: date

Abstract In this paper, we investigate algorithms for finding centers of a given collection $\mathcal{N}$ of sets. In particular, we focus on metric rational set similarities, a broad class of similarity measures including Jaccard and Hamming. A rational set similarity S is called metric if $D = 1 - S$ is a distance function. We study the 1-center problem on these metric spaces. The problem consists of finding a set C that minimizes the maximum distance of C to any set of $\mathcal{N}$. We present a general framework that computes a $(1 + \varepsilon)$ approximation for any metric rational set similarity.

1 Introduction

Clustering algorithms form a fundamental subroutine in any data analysis chain. The aim of clustering is to partition the data set $\mathcal{N}$ such that similar objects are grouped together and dissimilar objects are grouped in distinct clusters. Often, we assume that the objects lie in some metric space, i.e., their similarity (or rather dissimilarity) is characterized by some distance function D. In this case, center-based clustering objectives are particularly popular. For these problems, we aim to find a subset of objects C, such that some function of the distances between the sets of $\mathcal{N}$ and their respectively closest center in C is minimized. Among the most commonly used functions are the sum of distances, which corresponds to the k-median problem, the sum of squared distances, which corresponds to the k-means problem, and the maximum distance, which corresponds to the k-center problem.

In this paper, we focus on the 1-center problem for a large class of metrics defined on sets. Specifically given a collection of sets $\mathcal{N}$ from some universe U and some distance function D, we aim to find a set $C \subset U$ such that $\max_{A\in\mathcal{N}} D(A,C)$ is minimized. In this paper, we assume that the distance function is induced by *rational set similarities*. Given two subsets A and B of some ground set U, the similarity between A and B is defined as a the ratio between linear combinations of the cardinalities of symmetric difference $A \triangle B$, intersection $A \cap B$, and negated union $\overline{A \cup B}$. The induced dissimilarity function is 1 minus the similarity. If the dissimilarity is a distance function, the similarity is known as a metric rational set similarity. Well known examples for similarities include Sokal-Michener's simple matching [17] $\frac{|A\cap B|+|\overline{A\cup B}|}{|U|}$, the Jaccard index [22] $\frac{|A\cap B|}{|A\cup B|}$, the Anderberg similarity [1] $\frac{|A\cap B|}{|A\cup B|+|A\triangle B|}$, and the Rogers-Tanimoto coefficient [30] $\frac{|A\cap B|+|\overline{A\cup B}|}{|U|+|A\triangle B|}$. For further examples, we refer to Gower and Legendre [18]. We are given a collection $\mathcal{N}$ of n subsets of some ground set U. Our aim is to find a center $C \subseteq U$ minimizing the maximum distance to all sets of $\mathcal{N}$. We obtain the following result (see Theorem 2 for a precise statement):

> The 1-center in the metric space induced by sets and the distance function of a metric rational set similarity admits a polynomial time approximation scheme.

M. Gentili and C. Schwiegelshohn are partially supported by ERC Advanced Grant 788893 AMDROMA.

Marc Bury
TU Dortmund. Otto Hahn Strasse 14, Dortmund, Germany
E-mail: marc.bury@tu-dortmund.de

Michele Gentili, Chris Schwiegelshohn, Mara Sorella
Sapienza, University of Rome. Via Ariosto 25, Rome, Italy
E-mail: {surname}@diag.uniroma1.it

Prior to our work, the only metric rational set similarity for which the 1-center problem admits a PTAS was Sokal-Michener's simple matching similarity[1]. The induced distance function is equivalent to the Hamming distance on the Boolean hypercube and the problem itself is more commonly referred to as the Closest String Problem. We refer to more related work in Section 3. Our PTAS runs in time $O(n^{\text{poly}(\varepsilon^{-1})})$, where the the exponent of ε^{-1} depends on the underlying rational set similarity and is never larger than 6. For the Closest String problem, the exponent of ε is 2, which matches the running time of the best previously known algorithms [2, 26] up to polylog ε^{-1} factors in the exponent of n. We note that we require that the coefficients of the linear combination of numerator and denominator are constants. Since all rational set similarities used in practice satisfy this property, we view this as a mild assumption.

Rational set similarities appear in a wide variety of areas, including nearest neighbor searching [6,8], plagiarism detection [4], association rule mining [13], collaborative filtering [15], web compression [9], biogeographical analysis [29], and chemical similarity searching [33]. Most notably, many of them were initially proposed for classification and cluster analysis [1,17,30]. However, a rigorous analysis for classical clustering problems has been mostly constrained to the Closest String Problem and the Jaccard-median problem. Our work significantly expands upon this.

2 Approach and Techniques

The starting point of all known polynomial time approximation schemes for the Closest String problem, as well as the Jaccard-center problem, is a natural LP-formulation [5, 25]. Specifically, the cardinality of symmetric difference $|A \triangle C|$, intersection $|A \cap C|$, and negated union $|\overline{A \cup C}|$ can all be expressed as a linear combination of binary variables $C_i = 1$ if element $i \in C$ or $C_i = 0$ if element $i \notin C$, as long as A is fixed. Then both numerator $Num(A,C)$ and denominator $Den(A,C)$ of the similarity can be expressed as a linear combination. By testing the integer linear program

$$Den(A,C) - Num(A,C) \leq dist \cdot Den(A,C)$$

for feasibility, we know whether a center with maximum distance *dist* exists. For instance, if the similarity is Sokal-Michener's simple matching and the associated distance function is the Hamming norm on the hypercube, the constraints have the form

$$|A \triangle C| = \sum_{i=1}^{|U|} A_i + C_i - 2A_iC_i \leq dist \cdot |U|.$$

The main idea is to compute a feasible fractional solution to the LP and subsequently apply randomized rounding. This simple strategy already provides a high quality center given that the symmetric difference between center and any input set is sufficiently large. This behavior is also observed in real-world instances of these problems [11].

In the case that the rounding fails to provide a good solution, the algorithms switch to a number of enumeration strategies. The first important observation is that using Chernoff bounds [28], one can bound the symmetric difference between an optimal center and any input set by $O(\ln n)$. This already gives rise to a simple quasi-polynomial time algorithm: Pick an arbitrary input set A, and try all sets with symmetric difference $O(\ln n)$ from A. For a ground set U, there are at most $\binom{|U|}{O(\ln n)} \in |U|^{O(\ln n)}$ possibilities.

This type of enumeration may be substantially improved if we simultaneously consider the items of multiple sets $A_1, \ldots, A_m$, all of which have small symmetric distances to the optimum. Here, the number of candidate subsets cannot increase, but may be reduced. For Hamming center, this was achieved via the notion of a generator [27]. Essentially, a generator for an optimal center C is a collection M of sets such that the items either contained in all sets of M are in C and those items not contained in any set of M are not in C. Formally, M is a generator of C if the items $I := \{i \in \{1, \ldots, |U|\} \mid \forall_{A \in M}\ i \in A\}$ and $J := \{i \in \{1, \ldots, |U|\} \mid \forall_{A \in M}\ i \notin A\}$ satisfy $I \subseteq C$ and $J \subseteq U \setminus C$. The conflict set of the generator, consists precisely of the items not included in either I or J, i.e., $U \setminus (I \cup J)$. Given that the conflict size is small, and that we can determine an appropriate M, we can extract C via brute force enumeration in polynomial time. Marx [27] showed how a generator may be efficiently constructed for the Hamming center problem if the distance is small, and Andoni et al. [2] further extended this to

[1] A preliminary version of this paper proving polynomial time approximability of the Jaccard center problem appeared in ICALP 2017 [5]

$(1+\varepsilon)$-approximations. Generators are therefore a natural starting point. The limits of the construction become apparent for the Jaccard-center problem. Items in C and not in C can be treated indiscriminately for the Hamming center problem, i.e., the Hamming center problem of the instance $\mathcal{N}$ is identical to the Hamming center of the instance $\overline{\mathcal{N}} := \{A \subset U \mid U \setminus A \in \mathcal{N}\}$. The same does not hold for arbitrary rational set similarities. For instance the Jaccard distance $\frac{|A \triangle C|}{|A \cup C|}$ is highly sensitive to the support of both A and C.

We therefore aim to expand the properties of a generator to account for the support of the subsets. This is made more precise via the notion of core-cover. We call a collection of sets M a *core-cover*, if an optimal center C is (mostly) contained in $\bigcup_{A\in M} A$. Specifically, we require that $\bigcup_{A\in M} A$ contains an $(1+\varepsilon)$-approximate solution. An *anchored core-cover* further restricts the possible solutions by always containing the items in the intersection of all sets of the core-cover, i.e., $\bigcap_{A\in M} A \cup \left(C \cap \bigcup_{A\in M} A\right)$ is an $(1+\varepsilon)$-approximate solution. Crucially, we show that the size of an anchored core-cover depends only on ε^{-1}. This allows us to determine by brute force in time $n^{|M|}$ an anchored core-cover and enumerate all possible solutions $\bigcap_{A\in M} A \cup \left(C \cap \bigcup_{A\in M} A\right)$. For more technical remarks comparing core-covers to generators, we refer to Section 6. They are also related in spirit to coresets for the minimum enclosing ball, which corresponds to the 1-center problem in Euclidean space [3,12,24,34].

To extend our analysis to arbitrary rational set similarities, we require a number of additional ideas. First, the denominator of any rational set similarity can be written as linear combination of the denominator of Hamming-distance and Jaccard distance. With this observation, we are able to identify a set of "characteristic" rational set similarities to which any other rational set similarity may be (non-trivially) reduced. These characteristic rational set similarities are sufficiently closely related to either Hamming and Jaccard such that the analysis of the LP rounding as well as the construction of a core-cover can be extended.

3 Related Work

Most center problems for rational set similarities were heuristic, see for instance Guha et al. [20]. In theory, most attention has been shown to the Closest String Problem. The first PTAS was proposed by Li, Ma and Wang [25]. Subsequently, the running time of the PTAS was further improved by Andoni, Indyk and Patrascu [2], and by Ma and Sun [26], with the currently best running time being $n^{O(\varepsilon^{-2})}$. Andoni, Indyk and Patrascu [2] further gave a conditional lower bound showing that any PTAS must have running time $\exp(\varepsilon^{-2})$, assuming the exponential time hypothesis (ETH). Cygan et al. [14] further showed that assuming ETH, any $(1+\varepsilon)$ must require time $O(n^{\varepsilon^{-1}})$. Further, no efficient PTAS (i.e., a PTAS running in time $f(\varepsilon)\cdot \mathrm{poly}(n)$) can exist unless $FPT = W[1]$. The Closest String Problem also has received substantial attention for fixed parameter algorithms, see [7,16,19,21,27] and references therein.

To the best of our knowledge, the only other rational set similarity for which theoretical clustering problems have been analyzed is the Jaccard-median problem, i.e., the task of finding a center C such that the sum of Jaccard distances to C is minimized. Spaeth [31] gave a structural result for continuous Jaccard measures, which proved that even in the Euclidean space, the problem is in NP. Watson [32] gave a vertex descent algorithm without bounds on the running time. Chierichetti et al. [10] proved that the Jaccard-median problem is NP-hard, but admits a PTAS. In previous work [5], we showed that the Jaccard-center problem admits a PTAS.

4 Preliminaries

Let $U = \{u_1, u_2, \ldots, u_d\}$ be a base set containing d elements and let $\mathcal{N} \subseteq \mathcal{P}(U)$ be a collection of n subsets of U. Denote the symmetric difference of two sets by $A \triangle B := (A \setminus B) \cup (B \setminus A)$. We will refer to the complementary set by $\overline{A} := U \setminus A$.

Definition 1 (Rational Set Similarities [18]) Given $x,y \geq 0$ and $z \geq z' \geq 0$, the *rational set similarity* $S_{x,y,z,z'}$ between two non-empty item sets A and B is

$$S_{x,y,z,z'}(A,B) = \frac{x\cdot|A\cap B| + y\cdot|\overline{A\cup B}| + z'\cdot|A\triangle B|}{x\cdot|A\cap B| + y\cdot|\overline{A\cup B}| + z\cdot|A\triangle B|}$$

if it is defined and 1 otherwise. The dissimilarity induced by $S_{x,y,z,z'}$ is defined as

$$D_{x,y,z,z'}(A,B) := 1 - S_{x,y,z,z'}(A,B) = \frac{(z-z')\cdot|A\triangle B|}{x\cdot|A\cap B| + y\cdot|\overline{A\cup B}| + z\cdot|A\triangle B|}$$

if it is defined and 0 otherwise. If $D_{x,y,z,z'}$ is a distance function, we call $S_{x,y,z,z'}$ a metric rational set similarity.

We will assume throughout this paper that x,y,z,z' are either positive constant integers, or 0. Further, without loss of generality, we assume that $x \geq y$, as $D_{x,y,z,z'}(A,B) = D_{y,x,z,z'}(\overline{A},\overline{B})$. The arguably most well-known rational set similarity is the Jaccard similarity $S_{1,0,1,0}(A,B) = \frac{|A\cap B|}{|A\cup B|}$. For distances induced by metric rational set similarities, $D_{1,1,1,0}$ corresponds to the Hamming distance on the d-dimensional hypercube. The precise set of conditions under which a rational set similarity $S_{x,y,z,z'}$ yields a metric $D_{x,y,z,z'}$ can be found in Janssens' thesis [23], see also Chierichetti and Kumar [8].

Proposition 1 (Characterizations of Metric Rational Set Similarities, Janssens [23]) $(\mathcal{P}(U),D_{x,y,z,z'}(A,B))$ *is a metric space if and only if* $z \geq \max(x,y,z')$.

We assume $z > z'$ as otherwise all distances are 0 and the problem is trivial. All rational set similarities considered in this paper will have metric distance functions. To simplify the analysis, we will only consider rational set similarities with certain parameters. In Section 5 we will show that the center problem on the corresponding distance of a general rational set similarity can always be reduced to the center problem on a simpler distance, the *simple rational set distance*, which we define next.

Definition 2 (Simple Rational Set Distance (simple RSD)) Given $1 \geq y \geq 0$ the *simple RSD* D_y between two non-empty item sets A and B is

$$D_y(A,B) := \frac{|A \triangle B|}{|A \cup B| + y \cdot |\overline{A \cup B}|}$$

if it is defined and 0 otherwise.

In the subsequent section, we will establish a strong relationship between $D_{y'}$ and $D_{x,y,z,z'}$, if $\frac{y}{x} = y'$. For now, note that if $y' = y/x$

$$\begin{aligned} D_{y'}(A,B) &= \frac{|A \triangle B|}{|A \cup B| + y' \cdot |\overline{A \cup B}|} = \frac{x \cdot |A \triangle B|}{x \cdot |A \cup B| + y \cdot |\overline{A \cup B}|} \\ &= \frac{x \cdot |A \triangle B|}{x \cdot |A \cap B| + y \cdot |\overline{A \cup B}| + x \cdot |A \triangle B|} = D_{x,y,x,0}(A,B). \end{aligned}$$

We will assume that x,y',z,z' are constants.

Problem 1 (RSD-Center) Given the base set $U = \{u_1, u_2, \ldots, u_d\}$, and a collection $\mathcal{N} \subseteq \mathcal{P}(U)$ of n subsets of U, the RSD-center problem consists of finding a set $C \subseteq U$ such that

$$\max_{A \in \mathcal{N}} D_{x,y,z,z'}(A,C)$$

is minimized.

We denote by OPT the value $\min_{C\subset U} \max_{A\in\mathcal{N}} D_{x,y,z,z'}(A,C)$ throughout this paper. We further assume that $OPT < \frac{1}{1+\varepsilon}$, as any candidate solution $K \subset U$ has distance at most 1 to any other subset of U and therefore would be a $(1+\varepsilon)$ approximation. Lastly, we will frequently use the following easy verifiable facts throughout the paper.

Fact 1. Let $A,B \subseteq U$ be two item sets. Then the following statements hold:

1. $|A \cap B| = |A \cup B| - |A \triangle B| = |A \cup B| - D_y(A,B) \cdot \left[y \cdot |\overline{A \cup B}| + |A \cup B|\right]$
2. $|A \setminus B| = |A \triangle B| - |B \setminus A| = D_y(A,B) \cdot \left[|A \cup B| + y \cdot |\overline{A \cup B}|\right] - |B \setminus A|$
3. $if \quad y = 0 \Rightarrow |A \setminus B| = D_0(A,B) \cdot |A \cup B| - |B \setminus A| \leq D_0(A,B) \cdot |A|$
4. $if \quad y = 0 \Rightarrow |A| \geq (1 - D_0(A,B)) \cdot |B|$

The remaining paper is now organized as follows. Section 5 contains the reduction from arbitrary metric RSD to simple RSD. Section 6 bounds the size of core-covers for any simple RSD. Section 7 describes the algorithm containing the LP rounding procedure and the core-cover-based enumeration strategy, as well as proving correctness. We conclude with a minor remark showing that for continuous Jaccard measures, the 1-center problem is solvable in polynomial time, whereas the 1-median problem is NP-hard (Section 8).

5 Reduction from RSD Center to Simple RSD Center

Lemma 1 *Let $\mathcal{N} \subset \mathcal{P}(U)$ be a collection of item sets, let y' be a rational number in $[0,1]$ and let x,y,z,z' be non-negative integers satisfying $z \geq \max(x,y,z')$ and $y' = y/x$. Then for any set S, we have*

$$S = \underset{S' \subset U}{\operatorname{argmin}} \ \max_{A \in \mathcal{N}} D_{y'}(A,S) \iff S = \underset{S' \subset U}{\operatorname{argmin}} \ \max_{A \in \mathcal{N}} D_{x,y,z,z'}(A,S).$$

Furthermore, let $\varepsilon \geq 0$ be a parameter. Then if S is an item set satisfying

$$\max_{A \in \mathcal{N}} D_{y'}(A,S) \leq (1+\varepsilon) \min_{C \subset U} \ \max_{B \in \mathcal{N}} D_{y'}(B,C),$$

we have

$$\max_{A \in \mathcal{N}} D_{x,y,z,z'}(A,S) \leq (1+\varepsilon) \min_{C \subset U} \ \max_{B \in \mathcal{N}} D_{x,y,z,z'}(B,C).$$

Proof. In the following, let $B = \underset{A \in \mathcal{N}}{\operatorname{argmax}} \ D_{y'}(A,C)$. We will first show that am optimal solution C for the problem $\min\limits_{C' \subset U} \max\limits_{B \in \mathcal{N}} D_{y'}(B,C')$ is also an optimal solution for the problem $\min\limits_{C' \subset U} \max\limits_{B \in \mathcal{N}} D_{x,y,z,z'}(B,C')$. By optimality of C, we know that for any candidate solution C' there exists some B' with $D_{y'}(B,C) \leq D_{y'}(B',C')$. Hence,

$$\begin{aligned}
& \frac{|B \triangle C|}{|B \cup C| + y'|\overline{B \cup C}|} \leq \frac{|B' \triangle C'|}{|B' \cup C'| + y'|\overline{B' \cup C'}|} \\
\iff & |B \triangle C| \cdot (|B' \cup C'| + y'|\overline{B' \cup C'}|) \leq |B' \triangle C'| \cdot (|B \cup C| + y'|\overline{B \cup C}|) \\
\iff & |B \triangle C| \cdot (|B' \cup C'| + y'|\overline{B' \cup C'}|) + z/x|B \triangle C| \cdot |B' \triangle C'| \\
& \leq |B' \triangle C'| \cdot (|B \cup C| + y'|\overline{B \cup C}|) + z/x|B \triangle C| \cdot |B' \triangle C'| \\
\iff & \frac{|B \triangle C|}{|B \cup C| + y'|\overline{B \cup C}| + z/x|B \triangle C|} \leq \frac{|B' \triangle C'|}{|B' \cup C'| + y'|\overline{B' \cup C'}| + z/x|B' \triangle C'|} \\
\iff & \frac{(z-z')|B \triangle C|}{x|B \cup C| + y|\overline{B \cup C}| + z|B \triangle C|} \leq \frac{(z-z')|B' \triangle C'|}{x|B' \cup C'| + y|\overline{B' \cup C'}| + z|B' \triangle C'|} \\
\iff & D_{x,y,z,z'}(B,C) \leq D_{x,y,z,z'}(B',C'). \qquad (1)
\end{aligned}$$

This proves the first claim. What is left to show is that this still holds for approximations. With B defined as above and $A = \underset{A \in N}{\operatorname{argmax}} \ D_{y'}(A,S)$, we have

$$\begin{aligned}
& \frac{|A \triangle S|}{|A \cup S| + y'|\overline{A \cup S}|} \leq (1+\varepsilon) \frac{|B \triangle C|}{|B \cup C| + y'|\overline{B \cup C}|} \\
\implies & |A \triangle S| \cdot (|B \cup C| + y'|\overline{B \cup C}|) \leq (1+\varepsilon) \cdot |B \triangle C| \cdot (|A \cup S| + y'|\overline{A \cup S}|) \\
& \leq (1+\varepsilon) \cdot |B \triangle C| \cdot (|A \cup S| + y'|\overline{A \cup S}|) + \varepsilon \cdot |B \triangle C| \cdot (z/x)|A \triangle S| \\
\iff & |A \triangle S| \cdot (|B \cup C| + y'|\overline{B \cup C}| + (z/x)|B \triangle C|)) \\
& \leq (1+\varepsilon) \cdot |B \triangle C| \cdot (|A \cup S| + y'|\overline{A \cup S}| + (z/x)|A \triangle S|) \\
\iff & \frac{|A \triangle S|}{|A \cup S| + y'|\overline{A \cup S}| + (z/x)|A \triangle S|} \leq (1+\varepsilon) \frac{|B \triangle C|}{|B \cup C| + y'|\overline{B \cup C}| + (z/x)|B \triangle S|} \\
\iff & \frac{(z-z')|A \triangle S|}{x|A \cup S| + y|\overline{A \cup S}| + z|A \triangle S|} \leq (1+\varepsilon) \frac{(z-z')|B \triangle C|}{x|B \cup C| + y|\overline{B \cup C}| + z|B \triangle S|} \\
\iff & D_{x,y,z,z'}(A,S) \leq (1+\varepsilon) D_{x,y,z,z'}(B,C).
\end{aligned}$$

Combining the final equation with Equation 1 now yields for any $A' \in \mathcal{N}$ and B',C' as defined above

$$D_{x,y,z,z'}(A',S) \leq D_{x,y,z,z'}(A,S) \leq (1+\varepsilon) D_{x,y,z,z'}(B,C) \leq (1+\varepsilon) D_{x,y,z,z'}(B',C').$$

□

We remark that while an optimal solution for the 1-center problem with distance function D_y is equivalent to an optimal solution for the 1-center problem with distance function $D_{x,y,z,z'}$, the same does not hold for approximations. That is, approximating the 1-center problem with distance funciton $D_{x,y,z,z'}$ may be easier than approximation the problem with distance function D_y. In this sense, the problem $\min\limits_{C' \subset U} \max\limits_{B \in \mathcal{N}} D_{y'}(B,C')$ is the canonical hard problem.

Also, as mentioned in the preliminaries, this lemma implies a reduction for the metric spaces of the form $(\mathcal{P}(U), D_{x,y,z,z'})$ with $y > x$, as $D_{x,y,z,z'}(A,B) = D_{y,x,z,z'}(\overline{A},\overline{B})$. As a preprocessing step, we determine $\overline{\mathcal{N}} := \{\overline{A} \mid A \in \mathcal{N}\}$ and compute a $(1+\varepsilon)$ approximation for the appropriate simple *RSD* center problem on $\overline{\mathcal{N}}$.

6 Core Covers

Throughout this section, we consider the metric space $(\mathcal{P}(U), D_y)$. Let $\mathcal{N}$ be a collection of subsets of a base set U, let OPT be the maximum distance of an optimal center to any subset in $\mathcal{N}$.

Our algorithms are based on the existence of a small collection M of input sets such that a high quality center can be extracted from M. Informally, the items of an optimal center are well represented by the items of the sets contained in M.

Definition 3 (Core-covers) Let $\varepsilon > 0$ be a constant. A collection $M \subseteq \mathcal{N}$ with $I_M = \cap_{A \in M} A$ and $O_M = \cup_{A,B \in M} A \triangle B$ is an (ε,y)-*core-cover* if there exists an optimal center C with $K = (I_M \cup O_M) \cap C$ and:

$$\max_{A \in \mathcal{N}} D_y(A,K) \leq (1+\varepsilon) \cdot OPT$$

A collection $M \subseteq \mathcal{N}$ is an (ε,y)-*anchored-core-cover* if there exists an optimal center C with $K = I_M \cup (O_M \cap C)$ and:

$$\max_{A \in \mathcal{N}} D_y(A,K) \leq (1+\varepsilon) \cdot OPT.$$

Marx [27] proposed generator strings for the Hamming center problem on the Boolean hypercube (see Section 3 of the reference). In our terminology, the problem corresponds to the simple RSD D_1 or more generally $D_{1,1,0,1}$. Given a collection of sets M, define $O_M := \{i \in [d], A_i \neq B_i \text{ for some } A,B \in M\}$. By showing that a generator M with constant size O_M exists, he was able to obtain an FPT algorithm for the Hamming center problem, which was later also used by Andoni et al. [2] and Ma and Sun [26] to improve the running time of a PTAS. Hence, anchored core covers are a generalization of generators to arbitrary metrics induced by rational set similarities.

Marx [27] proved a (tight) bound of $O\left(\log \frac{1}{\varepsilon}\right)$ on the size of the generator (or anchored core cover) M for the Hamming center problem. In the following, we will see that for $y > 0$, a bound of $O\left(\frac{\log \frac{1}{\varepsilon y}}{y}\right)$ can also be achieved for any (ε,y)-anchored core cover. For the Jaccard-center problem (and other RSD with $y = 0$), we require a different type of analysis. The resulting bound of $O(\varepsilon^{-1})$ is substantially weaker, but we can also prove that this is necessary. We begin the analysis of the size of a core-cover with the following observation.

Observation 1 Let $A \in \mathcal{N}$ be a set such that $1 \geq D_y(A,K) > (1+\varepsilon) \cdot OPT$ and $K \subseteq C$. Then:

$$|A \cap (C \setminus K)| \geq \begin{cases} OPT \cdot \varepsilon \cdot y \cdot d + y \cdot |(C \setminus K) \setminus A| & if \quad y \neq 0 \\ OPT \cdot \varepsilon \cdot |C| & if \quad y = 0 \end{cases}$$

Proof.

$$\begin{aligned}
|A \cap (C \setminus K)| &\overset{K \subseteq C}{=} |A \cap C| - |A \cap K| \\
&\overset{\text{Fact 1.1}}{=} |A \cup C| - D_y(A,C) \cdot \left(|A \cup C| + y \cdot |\overline{A \cup C}|\right) \\
&\quad -|A \cup K| + D_y(A,K) \cdot \left(y \cdot |\overline{A \cup K}| + |A \cup K|\right) \\
&\geq |A \cup C| - OPT\left(|A \cup C| + y \cdot |\overline{A \cup C}|\right) - |A \cup C| + |(C \setminus K) \setminus A| \\
&\quad +(1+\varepsilon) \cdot OPT\left[|A \cup C| + y \cdot |\overline{A \cup C}| + (y-1) \cdot |(C \setminus K) \setminus A|\right] \\
&= OPT \cdot \varepsilon \cdot (|A \cup C| + y \cdot |\overline{A \cup C}|) + (1+\varepsilon) \cdot OPT \cdot (y-1) \cdot |(C \setminus K) \setminus A| \\
&\quad +|(C \setminus K) \setminus A| \\
&\geq OPT \cdot \varepsilon \cdot (y \cdot d + (1-y)|A \cup C|) + y \cdot |(C \setminus K) \setminus A|,
\end{aligned}$$

where the final inequality follows from $y \leq 1$ and $(1+\varepsilon)OPT < 1$. For $y \neq 0$ we are done. For $y = 0$, we have $|A \cup C| \geq |C|$. □

The following lemma now bounds the size of a core-cover. The main argument is that if K is not a core-cover, Observation 1 guarantees us the existence of a set A containing many elements in $C \setminus K$.

Lemma 2 *For any collection of subsets $\mathcal{N}$ and any simple RSD D_y, there exists an (ε,y)-core-cover M of size*

$$|M| = \begin{cases} O\left(\frac{\log\frac{1}{\varepsilon}}{y}\right) & \textit{if } y > 0 \\ \lceil\frac{1}{\varepsilon}\rceil & \textit{if } y = 0 \end{cases}.$$

Proof. We show the existence of the collection M by proving that we can iteratively add a set to M such that either K already contains a good approximate solution or the added set contains many elements from $C \setminus K$. Thus, we either have C covered by $\cup_{A\in M} A$ or no set violates the approximation guarantee. Let $M^{(0)} = \{A^0\}$ for an arbitrary $A^0 \in \mathcal{N}$. Let $A^{(i)}$ be the set added in the ith iteration. We denote by $K^{(i)} = C \cap (\cup_{A\in M^{(i)}} A)$ our solution after the i-th iteration. Then $|C \setminus K^{(i)}|$ are the components of C that still have to be covered after i iterations. Note that this implies

$$D_y(K^{(i)},C) \leq D_y(K^{(0)},C) \leq OPT \tag{2}$$

for all i.

Moreover, we have the invariant

$$D_y(A^{(i)},K^{(i-1)}) > (1+\varepsilon)\cdot OPT, \tag{3}$$

as otherwise the current collection $M^{(i-1)}$ is already a core-cover. Now we analyze separate cases for simple RSD with either $y = 0$, and $y \neq 0$.

Case $y \neq 0$: We prove the following invariant for the algorithm, assuming that we add a new set to M in every iteration:

$$|C \setminus K^{(i)}| \leq \frac{|C \setminus K^{(0)}|}{(1+y)^i} - \sum_{k=1}^{i} \frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k}. \tag{4}$$

For $i = 0$ this clearly holds. Otherwise, we have by induction

$$\begin{aligned} |C \setminus K^{(i)}| &= |(C \setminus K^{(i-1)}) \setminus A^{(i)}| = |C \setminus K^{(i-1)}| - |A^{(i)} \cap (C \setminus K^{(i-1)})| \\ &\overset{\text{Obs. 1}}{\leq} |C \setminus K^{(i-1)}| - OPT\cdot\varepsilon\cdot y\cdot d - y\cdot|(C \setminus K^{(i-1)}) \setminus A^{(i)}| \\ &= \frac{|C \setminus K^{(i-1)}| - OPT\cdot\varepsilon\cdot y\cdot d}{1+y} \\ &\leq \frac{\left(\frac{|C\setminus K^{(0)}|}{(1+y)^{i-1}} - \sum_{k=1}^{i-1}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k}\right) - OPT\cdot\varepsilon\cdot y\cdot d}{1+y} \\ &= \frac{|C \setminus K^{(0)}|}{(1+y)^i} - \sum_{k=1}^{i}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k} \end{aligned}$$

Then with a bit of calculus, we obtain the following upper bound on the elements that remain to be covered after adding the $A^{(i)}$:

$$\begin{aligned} |C \setminus K^{(i)}| &\overset{Eq.\ 4}{\leq} \frac{|C \setminus K^0|}{(1+y)^i} - \sum_{k=1}^{i}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k} \\ &\overset{\text{Fact 1.2}}{=} \frac{D_y(C,K^{(0)})\cdot[|C\cup K^{(0)}| + y|\overline{C\cup K^{(0)}}|] - |K^{(0)}\setminus C|}{(1+y)^i} - \sum_{k=1}^{i}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k} \\ &\overset{Eq.\ 2}{\leq} \frac{OPT\cdot d}{(1+y)^i} - \sum_{k=1}^{i}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k} = \frac{OPT\cdot d}{(1+y)^i} - \sum_{k=0}^{i}\frac{OPT\cdot\varepsilon\cdot y\cdot d}{(1+y)^k} + OPT\cdot\varepsilon\cdot y\cdot d \\ &= OPT\cdot d\left(\left(\frac{1}{1+y}\right)^i - \varepsilon\cdot y\cdot\frac{1-\left(\frac{1}{1+y}\right)^{i+1}}{1-\frac{1}{1+y}} + \varepsilon\cdot y\right) \\ &= OPT\cdot d\left(\left(\frac{1}{1+y}\right)^i - \varepsilon\cdot\left(1+y-\left(\frac{1}{1+y}\right)^i\right) + \varepsilon\cdot y\right) \end{aligned}$$

$$= OPT \cdot d \left(\left(\frac{1}{1+y} \right)^i (1+\varepsilon) - \varepsilon \right)$$

Suppose the algorithm continues this process until all elements of C are covered, i.e. $C \setminus K^{(i)} = \emptyset$. This happens if

$$OPT \cdot d \left(\left(\frac{1}{1+y} \right)^i (1+\varepsilon) - \varepsilon \right) \leq 0$$
$$\Longrightarrow \left(\frac{1}{1+y} \right)^i \leq \frac{\varepsilon}{1+\varepsilon}$$

Using the Mercator series $\ln(1+y) = \sum_{i=1}^{infty} \frac{y}{i} \cdot (-1)^{i+1}$, we can conclude that $y/2 \leq \ln(1+y) \leq y$ for $y \in (0,1]$. Therefore, after $\log_{1+y} \frac{1+\varepsilon}{\varepsilon} \leq \frac{\ln \frac{2}{\varepsilon}}{\ln(1+y)} \in O\left(\frac{\ln \frac{1}{\varepsilon}}{y}\right)$ many iterations, we have either completely covered C, or the algorithm terminated earlier, meaning that the initial assumption from Equation 3 no longer holds.

Case $y = 0$: Let us assume an $A^{(i)}$ exists such that $1 \geq D_0(A,K^{i-1}) > (1+\varepsilon) \cdot OPT$. In this case, Observation 1 gives us a different bound on $|A^{(i)} \cap (C \setminus K^{(i-1)})|$. We will show inductively

$$|C \setminus K^{(i)}| = |C \setminus K^0| - i \cdot OPT \cdot \varepsilon \cdot |C|.$$

For $i = 0$, this clearly holds. Otherwise, we have

$$\begin{aligned} |C \setminus K^{(i)}| &= |(C \setminus K^{(i-1)}) \setminus A^{(i)}| = |C \setminus K^{(i-1)}| - |A^{(i)} \cap (C \setminus K^{(i-1)})| \\ &\overset{\text{Obs. 1}}{\leq} |C \setminus K^{(i-1)}| - OPT \cdot \varepsilon \cdot |C| \leq |C \setminus K^{(0)}| - (i-1) \cdot OPT \cdot \varepsilon \cdot |C| - OPT \cdot \varepsilon \cdot |C| \\ &= |C \setminus K^{(0)}| - i \cdot OPT \cdot \varepsilon \cdot |C| \end{aligned}$$

Therefore, after at most $\lceil \frac{1}{\varepsilon} \rceil$ many iterations, C will be completely covered. □

To prove the bounds on the anchored core-cover, we first require the following observation.

Observation 2 For any three sets $C, K, A \subseteq U$ and $y \in [0,1]$

$$D_y(A,K) \leq D_y(A,K \cap C) + \frac{|K \setminus C| - 2|(A \cap K) \setminus C|}{|A \cup K| + y \cdot |\overline{A \cup K}|}$$

Proof.

$$\begin{aligned} D_y(A,K) &= \frac{|A \triangle K|}{|A \cup K| + y \cdot |\overline{A \cup K}|} \\ &= \frac{|A \triangle (K \cap C)| + |(K \setminus C) \setminus A| - |A \cap (K \setminus C)|}{|A \cup K| + y \cdot |\overline{A \cup K}|} \\ &= \frac{|A \triangle (K \cap C)|}{(|A \cup (K \cap C)| + |(K \setminus C) \setminus A|) + y \cdot \left(|\overline{A \cup (K \cap C)}| - |(K \setminus C) \setminus A|\right)} \\ &\quad + \frac{|(K \setminus C) \setminus A| - |A \cap (K \setminus C)|}{|A \cup K| + y \cdot |\overline{A \cup K}|} \\ &\overset{y \leq 1}{\leq} \frac{|A \triangle (K \cap C)|}{|A \cup (K \cap C)| + y \cdot |\overline{A \cup (K \cap C)}|} + \frac{|(K \setminus C) \setminus A| - |A \cap (K \setminus C)|}{|A \cup K| + y \cdot |\overline{A \cup K}|} \\ &= D_y(A, K \cap C) + \frac{|K \setminus C| - 2|A \cap (K \setminus C)|}{|A \cup K| + y \cdot |\overline{A \cup K}|} \end{aligned}$$

□

If X is an arbitrary input point, K is our possible solution, and C is an optimal center, this observation implies that it is sufficient to show that $D_y(X,K \cap C)$ is a good approximation to $D_y(X,C)$ and $\frac{|K \setminus C| - 2|(X \cap K) \setminus C)|}{|X \cup K| + y \cdot |\overline{X \cup K}|}$ is small or negative.

Lemma 3 *For any collection of input subsets $\mathcal{N}$ and any simple RSD D_y, there exists an (ε,y)-anchored-core-cover $M \subseteq \mathcal{N}$ of size*

$$|M| = \begin{cases} O\left(\frac{\log \frac{1}{y\cdot\varepsilon}}{y}\right) & \textit{if } y > 0 \\ O\left(\frac{1}{\varepsilon}\right) & \textit{if } y = 0 \end{cases}.$$

Proof. Let C be an optimal center. Lemma 2 gives a set M such that $K' = C \cap (\cup_{A\in M} A)$ is an $(1+\varepsilon)$-approximate solution. Now let $K = I_M \cup (O_M \cap C)$, which is well defined given M and the optimal C. Note that $K' = K \cap C$. Using Observation 2, the distance between K and some arbitrary set A is:

$$\begin{aligned} D(A,K) &\leq D(A,K\cap C) + \frac{|K\setminus C| - 2\cdot|A\cap(K\setminus C)|}{|A\cup K| + y\cdot|\overline{A\cup K}|} \\ &= D(A,K\cap C) + \frac{|I_M\setminus C| - 2\cdot|A\cap(I_M\setminus C)|}{|A\cup K| + y\cdot|\overline{A\cup K}|} \\ &\leq (1+\varepsilon)\cdot OPT + \frac{|I_M\setminus C| - 2\cdot|A\cap(I_M\setminus C)|}{|X\cup K| + y\cdot|\overline{X\cup K}|} \end{aligned}$$

If for every $A \in N$, we have $2\cdot|A\cap(I_M\setminus C)| \geq |I_M\setminus C|$ then the ratio is non-positive and $D(A,K) \leq D(A,K\cap C) \leq (1+\varepsilon)\cdot OPT$. Otherwise, there exists an A such that $|A\cap(I_M\setminus C)| = |(A\cap I_M)\setminus C| < |I_M\setminus C|/2$. We iteratively augment the collection M with additional sets A. In each iteration, $|I_M\setminus C|$ is halved. We now bound $|I_M\setminus C|$ in terms of OPT. Again we distinguish between two cases.

Case $y \neq 0$: Let $B \in M$ be arbitrary. Then

$$\begin{aligned} |I_M\setminus C| \leq |B\setminus C| &\overset{\text{Fact 1.2}}{=} D(B,C)\cdot\left(|B\cup K| + y\cdot|\overline{B\cup K}|\right) - |C\setminus B| \\ &\leq D(B,C)\cdot\left(|B\cup K| + y\cdot|\overline{B\cup K}|\right) \\ &= D(B,C)\cdot\left(|B\cup K| + |\overline{B\cup K}| - |\overline{B\cup K}| + y\cdot|\overline{B\cup K}|\right) \\ &\leq OPT\cdot[d-(1-y)\cdot|\overline{B\cup K}|] \overset{y\leq 1}{\leq} OPT\cdot d. \end{aligned}$$

After adding $\log(\frac{1}{\varepsilon\cdot y})$ sets such that $|I_M\setminus C|$ is halved with each iteration, we have $|I_M\setminus C| \leq \varepsilon\cdot y\cdot OPT\cdot d$. Therefore,

$$\frac{|I_M\setminus C|}{|A\cup K| + y\cdot|\overline{A\cup K}|} \leq \frac{\varepsilon\cdot y\cdot OPT\cdot d}{|A\cup K| + y\cdot|\overline{A\cup K}|} \leq \frac{\varepsilon\cdot y\cdot OPT\cdot d}{y\cdot d} = \varepsilon\cdot OPT.$$

Case $y = 0$: Let $B \in M$ be arbitrary. We now have

$$\begin{aligned} |I_M\setminus C| \quad &\leq \quad |B\setminus C| \overset{\text{Fact 1.3}}{\leq} D(B,C)\cdot|B| \overset{\text{Fact 1.4}}{\leq} OPT\cdot\frac{|C|}{1-OPT} \\ &\overset{\text{Fact 1.4}}{\leq} OPT\cdot\frac{|A|}{(1-OPT)^2} \overset{OPT\leq\frac{1}{1+\varepsilon}}{\leq} OPT\cdot\frac{(1+\varepsilon)^2\cdot|A|}{\varepsilon^2} \leq OPT\cdot\frac{4}{\varepsilon^2}\cdot|A|, \end{aligned}$$

where the last inequality follows for $\varepsilon \leq 1$. After adding $\log\frac{4}{\varepsilon^3}$ sets such that $|I_M\setminus C|$ is halved with each iteration, we have

$$\frac{|I_M\setminus C|}{|A\cup K|} \leq \frac{\varepsilon\cdot OPT\cdot|A|}{|A\cup K|} \leq \varepsilon\cdot OPT.$$

For both cases, our approximation factor is therefore $(1+2\varepsilon)\cdot OPT$. Rescaling ε by a factor of 2 completes the proof. □

We conclude this section by showing that the bounds for $(\varepsilon,0)$ core-covers are tight, which also implies that the bounds on $(\varepsilon,0)$-anchored-core-covers are asymptotically tight. Hence, the exponential increase cannot be avoided as y tends to 0.

Proposition 2 *There exists a collection of subsets N such that for any $(\varepsilon,0)$-core-cover $M \subseteq N$, we have $|M| \geq 1/\varepsilon$.*

Proof. For a given $\varepsilon > 0$ and assuming $1/\varepsilon$ to be an integer, we consider the set system consisting of $1/\varepsilon$ singleton sets, i.e. each set consists of a single item and all sets are disjoint. The center consisting of all the singleton sets has a Jaccard distance of $\frac{1/\varepsilon-1}{1/\varepsilon} = 1-\varepsilon$ to each singleton set. If the core cover does not contain all singleton sets, the maximum distance of any subset of the core-cover to one of the singleton sets is 1. At the same time $OPT \cdot (1+\varepsilon) = (1-\varepsilon)(1+\varepsilon) = 1-\varepsilon^2 < 1$, hence the core-cover is required to contain all singleton sets. □

Lastly, we briefly compare the bounds $O(\frac{\log\frac{1}{\varepsilon}}{y})$ and $O(\frac{1}{\varepsilon})$ for the core-covers with respect to D_y and $y > 0$ and D_0, respectively. For all rational set similarity defined in literature, the latter bound is better than the former. However, for a sufficiently small y (e.g. $y \leq \varepsilon^2$), the former bound may become larger. This might hint at a gap in our analysis. However, we believe that this may be unavoidable; if y is part of the input a PTAS for the 1-center problem with distance function D_y may not exist. Resolving this conjecture is an interesting open problem.

7 A PTAS for the 1-Center Problem on Metric RSD

We now turn to our main result. Throughout this section, we consider the metric space $(\mathcal{P}(U), D_y)$. Let $\mathcal{N}$ be a collection of subsets of a base set U and let OPT be the maximum distance of an optimal center to any subset in $\mathcal{N}$.

Recall that $C_i = \begin{cases} 0 & \text{if } i \notin C, \\ 1, & \text{if } i \in C \end{cases}$. Observe that, for each set $A \in \mathcal{N}$, we have $|A \triangle C| = \sum_{i=1}^{d} A_i - 2A_i \cdot C_i + C_i$, $|A \cup C| = \sum_{i=1}^{d} A_i - A_i \cdot C_i + C_i$ and $|\overline{A \cup C}| = \sum_{i=1}^{d} 1 - A_i \cdot C_i$.

Hence we obtain a set of linear inequalities of the form

$$|A \triangle C| \leq OPT \cdot (|A \cup C| + y \cdot |\overline{A \cup C}|) \tag{5}$$

which we can test for feasibility by relaxing the integrality constraints on C. Denote a fractional solution $\hat{C}$. We apply randomized rounding to obtain an integer solution S, rounding each C_i to 1 with probability $\hat{C}_i$. We will first characterize the instances where this approach already yields a good solution. To this end, let us first recall the multiplicative Chernoff bounds.

```
Input  : Collection $N$ of subsets, Parameter $\varepsilon > 0$, distance function $D_{x,y,z,z'}$
Output: $(1+\varepsilon)$-approximate RSD center $C$
Let $D = \{\frac{i}{y \cdot d + (1-y) \cdot j} \mid 1 \leq j \leq d$ and $0 \leq i < y \cdot d + (1-y) \cdot j\}$.
Initialize list $Solutions = \emptyset$.
foreach $OPT^* \in D$ do
    if $\exists A \in N :\ OPT^* \cdot (d \cdot y + (1-y) \cdot |A|) > \frac{27\ln(4n)}{\varepsilon^2}$ then
        foreach $M \subseteq N$ with $|M| = \begin{cases} O\left(\frac{\log\frac{1}{\varepsilon}}{y}\right) & \textit{if } y > 0 \\ O\left(\frac{1}{\varepsilon}\right) & \textit{if } y = 0 \end{cases}$ do
            Compute optimal solution $C_{OPT^*} = I_M \cup S$ with $S \subseteq O_M$ (cf. Lemma 3).
            Add $C_{OPT^*}$ to $Solutions$
        end
    else
        Obtain fractional solution $C'_{OPT^*}$ by solving the set of linear equations given by Equation 5
        Obtain $C_{OPT^*}$ by rounding each entry of $C'_{OPT^*}$ to 1 with probability $\left(C'_{OPT^*}\right)_i$
        Add $C_{OPT^*}$ to $C$
    end
end
return $\underset{OPT^* \in D}{\operatorname{argmin}} \{C_{OPT^*} \in Solutions\}$
```

Algorithm 1: PTAS for the RSD-center problem

Lemma 4 (Multiplicative Chernoff-Bounds [28]) *Let $B_1, \ldots B_d$ be independent binary random variables with $\mu = \mathbb{E}[\sum_{i=1}^{d} B_i]$. Then for any $0 < \delta < 1$*

$$\mathbb{P}\left[\sum_{i=1}^{d} B_i > (1+\delta) \cdot \mu\right] \leq \exp\left(-\frac{\delta^2 \cdot \mu}{3}\right) \quad \textit{and} \quad \mathbb{P}\left[\sum_{i=1}^{d} B_i < (1-\delta) \cdot \mu\right] \leq \exp\left(-\frac{\delta^2 \cdot \mu}{2}\right).$$

Lemma 5 *Let S be a random binary vector obtained by rounding a feasible fractional solution of the set of inequalities (5) and let $\epsilon > 0$ be a constant. Assume OPT to satisfy $OPT \cdot (d \cdot y + (1-y) \cdot \min_{A\in\mathcal{N}} |A|) > \frac{27\ln(4n)}{\varepsilon^2}$. Then with probability at least $1/2$, the rounding procedure produces a binary solution S with $\max_{A\in\mathcal{N}} D_y(A,S) \leq (1+\epsilon) \cdot OPT$*

Proof. If the integral solution is feasible, the fractional solution will be feasible as well, which implies that there exists an estimate $\widehat{OPT}$ for which the LP 5 is feasible with $\widehat{OPT} \leq OPT$. Let us denote by $\text{cost}(S) := \max_{A\in\mathcal{N}} D_y(A,S)$ the value of the rounded solution. A rounded vector is not a good center if $OPT \cdot (1+\epsilon) \leq \text{cost}(S)$. We first derive concentration bounds on $|A \triangle S|$, and $|A \cup S| + y \cdot |\overline{A \cup S}|$. To keep the notation concise, we use $Den_y(A,S) = |A \cup S| + y \cdot |\overline{A \cup S}|$ to refer to the denominator of each $A \in \mathcal{N}$. Observe that $\mathbb{E}[S] = \hat{C}$ and

$$\frac{\mathbb{E}[|A \triangle S|]}{\mathbb{E}[Den_y(A,S)]} = \frac{|A \triangle \hat{C}|}{Den_y(A,\hat{C})} \leq \widehat{OPT} \leq OPT. \tag{6}$$

We have, due to the assumption on OPT and independently of the outcome of the rounding procedure,

$$Den_y(A,S) = d \cdot y + (1-y) \cdot |A \cup S| \geq d \cdot y + (1-y)|A| > \frac{27 \ln 4n}{\varepsilon^2 \cdot OPT}. \tag{7}$$

Both $|A \triangle S|$ and $Den_y(A,S)$ are sums of independent (though not identically distributed) Bernoulli random variables. Applying the multiplicative Chernoff bound (Lemma 4), we have for all $A \in \mathcal{N}$:

$$\mathbb{P}\left[Den_y(A,S) < \left(1 - \frac{\epsilon}{3}\right) \cdot \mathbb{E}[Den_y(A,S)]\right] \leq \exp\left(-\frac{\epsilon^2 \cdot \mathbb{E}[Den_y(A,S)]}{18}\right) \overset{Eq.\ 7}{\leq} \exp\left(-\frac{27\ln(4n)}{18 \cdot OPT}\right) \leq \frac{1}{4n}$$

and

$$\begin{aligned}
& \mathbb{P}\left[|A \triangle S| > \mathbb{E}[|A \triangle S|] + \frac{\epsilon}{3} \cdot OPT \cdot \mathbb{E}[Den_y(A,S)]\right] \\
= \; & \mathbb{P}\left[|A \triangle S| > \left(1 + \frac{\epsilon \cdot OPT \cdot \mathbb{E}[Den_y(A,S)]}{3 \cdot \mathbb{E}[|A \triangle S|]}\right) \cdot \mathbb{E}[|A \triangle S|]\right] \\
\leq \; & \exp\left(-\frac{\epsilon^2 \cdot OPT^2 \cdot \mathbb{E}[Den_y(A,S)]^2}{27 \cdot \mathbb{E}[|X \triangle S|]^2} \cdot \mathbb{E}[|X \triangle S|]\right) \\
\overset{Eq.\ 6}{\leq} \; & \exp\left(-\frac{\epsilon^2 \cdot OPT \cdot \mathbb{E}[Den_y(A,S)]}{27}\right) \overset{Eq.\ 7}{\leq} \exp\left(-\ln 4n\right) \leq \frac{1}{4n}
\end{aligned}$$

Combining these two bounds, with probability at least $1 - 1/2n$, we have:

$$\begin{aligned}
D_y(A,S) = \frac{|A \triangle S|}{|A \cup S| + y \cdot |\overline{A \cup S}|} & \leq \frac{\mathbb{E}[|A \triangle S|] + \frac{\varepsilon}{3} \cdot OPT \cdot \mathbb{E}[Den_y(A,S)]}{(1 - \frac{\varepsilon}{3}) \cdot \mathbb{E}[Den_y(A,S)]} \\
& \leq \frac{OPT + \varepsilon/3 \cdot OPT}{1 - \varepsilon/3} \overset{\varepsilon \leq 1}{\leq} (1+\varepsilon) \cdot OPT.
\end{aligned}$$

Applying the union bound we then obtain:

$$\mathbb{P}\left[\text{cost}(S) \leq (1+\epsilon) \cdot OPT\right] = 1 - \mathbb{P}\left[\exists A \in \mathcal{N} : D_y(A,S) > (1+\epsilon) \cdot OPT\right] \geq 1 - \frac{n}{2n} = \frac{1}{2}.$$

□

Our final algorithm (see also Algorithm 1) is now very simple. We try all possible values of OPT. Notice that there exist at most d^2 many distinct values of OPT, as the numerator can only be a number i and the denominator a number $y \cdot d + (1-y) \cdot j$, for $i,j \in \{1, \ldots, d\}$. For each candidate value, we apply the rounding procedure, if the conditions of Lemma 5 are satisfied. Otherwise, we compute an (ε,y)-anchored-core-cover cover (Lemma 3) and enumerate all possible solutions. We prove correctness in the following theorem.

Theorem 1 *Given a collection $\mathcal{N}$ of n subsets from a base set U of cardinality d and any constant $\varepsilon > 0$, there exists an algorithm computing a $(1+\varepsilon)$-approximation to the optimal center problem on D_y with constant probability. The algorithm runs in time $d^2 \cdot \left(n^{O(\varepsilon^{-6})} + LP(n,d)\right)$ for $y = 0$ and in time $d^2 \cdot \left(n^{O\left(\frac{\log^2 \frac{1}{y\varepsilon}}{y^3\varepsilon^2}\right)} + LP(n,d)\right)$ for $y > 0$, where $LP(n,d)$ is the time required to solve a linear program with n constraints and d variables.*

Proof. If the conditions of Lemma 5 are satisfied, i.e. $OPT \cdot (d \cdot y + (1-y) \cdot \min_{A\in\mathcal{N}} |A|) > \frac{27\ln(4n)}{\varepsilon^2}$, the rounding procedure will produce a good solution with constant probability.
Let us assume instead that the conditions are not satisfied.

We know that there exists at least one set $A \in N$ with:

$$OPT \cdot (d \cdot y + (1-y)|A|) \leq \frac{27\ln(4n)}{\varepsilon^2}. \tag{8}$$

Our goal will be to bound the size of O_M of an (ε,y)-anchored-core-cover M. By proving that $|O_M| \in O(\ln n \cdot \text{poly } \varepsilon^{-1})$, a complete enumeration becomes feasible in polynomial time for any fixed ε. As before, we distinguish between the case $y = 0$ and $y > 0$.

Case $y \neq 0$**:** We first bound the size of $|A \triangle B|$, for any $A,B \in \mathcal{N}$. We have

$$|A \triangle B| = D_y(A,B) \cdot (|A \cup B| + y|\overline{A \cup B}|) \leq 2 \cdot OPT \cdot d \leq \frac{27\ln 4n}{y\varepsilon^2}. \tag{9}$$

Now let us consider an (ε,y)-anchored core cover M. From Lemma 3, we have $|M| \in O\left(\frac{\log\frac{1}{y\varepsilon}}{y}\right)$. Finding one requires time $O\left(\binom{n}{|M|}\right) \in n^{O\left(\frac{\log\frac{1}{y\varepsilon}}{y}\right)}$. Combining this with Equation 9 yields

$$|O_M| = \frac{1}{2}\sum_{A\in M}\sum_{B\in M} |A \triangle B| \in O\left(\frac{\log^2\frac{1}{y\varepsilon}}{y^2\varepsilon^2} \cdot \ln n\right).$$

One of the solutions induced by the anchored core-cover is guaranteed to be a $(1+\varepsilon)$-approximation. Trying all of these solutions requires time $2^{O_M} \cdot n^{O\left(\frac{\log\frac{1}{y\varepsilon}}{y}\right)} \in n^{O\left(\frac{\log^2\frac{1}{y\varepsilon}}{y^3\varepsilon^2}\right)}$. Thus, the total running time is in $O\left(d^2 \cdot \left(n^{O\left(\frac{\log^2\frac{1}{y\varepsilon}}{y^3\varepsilon^2}\right)} + LP(n,d)\right)\right)$.

Case $y = 0$**:** We first sharpen the bound given by Equation 8. We have for some A'

$$OPT \cdot |A'| \leq \frac{27\ln 4n}{\varepsilon^2}.$$

This allows us to bound the size of $|A \triangle B|$ as follows

$$\begin{aligned} |A \triangle B| &= D_y(A,B) \cdot |A \cup B| \leq OPT \cdot (|A| + |B|) \overset{\text{Fact 1.4}}{\leq} OPT\frac{2|C|}{1-OPT} \\ &\overset{\text{Fact 1.4}}{\leq} OPT\frac{2|A'|}{(1-OPT)^2} \leq \frac{54\ln 4n}{(1-OPT)^2\varepsilon^2} \overset{OPT\leq\frac{1}{1+\varepsilon}}{\leq} \frac{54\ln 4n}{\varepsilon^4}. \end{aligned} \tag{10}$$

We now bound the size of O_M from an $(\varepsilon,0)$-anchored core cover M. From Lemma 3, we have $|M| \in O(\varepsilon^{-1})$, and we can find one in time $O(n^{O(\varepsilon^{-1})})$. Combining this with Equation 10 yields

$$|O_M| = \sum_{A\in M}\sum_{B\in M} |A \triangle B| \in O\left(\frac{\ln n}{\varepsilon^6}\right).$$

One of the solutions induced by the anchored core-cover is guaranteed to be a $(1+\varepsilon)$-approximation. Trying all of these solutions requires time $2^{O_M} \cdot n^{O(\varepsilon^{-1})} \in n^{O(\varepsilon^{-6})}$. Thus, the total running time is in $O\left(d^2 \cdot \left(n^{O(\varepsilon^{-6})} + LP(n,d)\right)\right)$.

□

Finally, combining Theorem 1 with Lemma 1 gives us our main result.

Theorem 2 *Let $\mathcal{N}$ be a collection of n subsets from a base set U of cardinality d, let x,y,z,z' be either positive constant integers or 0 and $z \geq \max(x,y,z')$ and let $\varepsilon > 0$ be a constant. Then there exists an algorithm computing a $(1+\varepsilon)$-approximation to the RSD-center problem with distance function $D_{x,y,z,z'}$ with constant probability. The algorithm runs in time $d^2 \cdot \left(n^{O(\varepsilon^{-6})} + LP(n,d)\right)$ for $y = 0$ and in time $d^2 \cdot \left(n^{O\left(\frac{\log^2 \frac{1}{y\varepsilon}}{y^3\varepsilon^2}\right)} + LP(n,d)\right)$ for $y > 0$, where $LP(n,d)$ is the time required to solve a linear program with n constraints and d variables.*

8 A Note on Continuous Jaccard Center

We conclude by briefly describing how to find the continuous Jaccard center of set $\mathcal{N}$ of n points in d-dimensional Euclidean space in polynomial time. We consider this fact to be notable as the continuous Jaccard median problem is NP-hard [10]. To the best of our knowledge, this is the only distance measure we are aware of where the 1-median problem is hard while the 1-center problem is easy. For instance, both 1-center and 1-median problem for the ℓ_1 norm (the continuous variant of the Hamming norm on the hypercube) are solvable in polynomial time using convex optimization.

The Jaccard measure in d-dimensional Euclidean space with non-negative coordinates is defined as

$$J_{cont}(A,B) := \begin{cases} \frac{\sum_{i=1}^{d} \min(A_i,B_i)}{\sum_{i=1}^{d} \max(A_i,B_i)} & \text{if } \sum_{i=1}^{d} \max(A_i,B_i) \neq 0 \\ 1 & \text{otherwise} \end{cases}.$$

The corresponding distance $D_{cont}(A,B)$ is again $1 - J_{cont}(A,B)$. We will formulate the decision problem of finding a center with distance at most *dist* as an LP. The optimum center can thereafter be determined in polynomial time using binary search over the possible values of *dist*. In the following let $A^j \in \mathcal{N}$ be the jth point of N w.r.t. some arbitrary ordering. We use the variable $c_i \geq 0$ to denote the ith entry of the Jaccard center C. We further use the variables $\max_{i,j}$ and $\min_{i,j}$ for all $i \in \{1,\ldots,d\}$ and $j \in \{1,\ldots n\}$ to denote the maximum and minimum of A_i^j and c_i. We then use the constraints

$$\begin{aligned} \sum_{i=1}^{d} \min_{i,j} \geq (1 - \mathit{dist}) \cdot \sum_{i=1}^{d} \max_{i,j} \quad & \text{for all } j \in \{1,\ldots n\} \\ \min_{i,j} \leq c_i, A_i^j \leq \max_{i,j} \quad & \text{for all } j \in \{1,\ldots n\}, i \in \{1,\ldots d\} \\ c_i \geq 0 \quad & i \in \{1,\ldots d\}. \end{aligned}$$

Note that the top most equation corresponds to $\sum_{i=1}^{d} \min(A_i^j,c_i) \geq (1 - \mathit{dist}) \cdot \sum_{i=1}^{d} \max(A_i^j,c_i)$ which is equal to $D_{cont}(A,C) = 1 - \frac{\sum_{i=1}^{d} \min(A_i^j,c_i)}{\sum_{i=1}^{d} \max(A_i^j,c_i)} \leq \mathit{dist}$.

9 Conclusions and Open Problems

We have presented a polynomial time approximation scheme for the 1-center problem on metric rational set similarities, which are a large class similarity measures on sets. Except for the simple matching similarity, for which the corresponding distance is the Hamming distance [25] and the Jaccard distance, for which we had shown a PTAS in a preliminary version of this paper [5], our work is the first polynomial time approximation scheme for the center problem on these distances. Though we are not able to exactly match the running time of the state-of-the-art Hamming center PTAS [26], our algorithm is competitive up to polylog ε^{-1} factors in the exponent.

For the Jaccard-center problem on n sets, our algorithm runs in time $n^{O(\varepsilon^{-6})}$. It seems unlikely that the running time can be reduced beyond $n^{O(\varepsilon^{-3})}$ using our approach. Showing either a conditional lower bound or devising an altogether new approach that achieves $n^{O(\varepsilon^{-2})}$ running time is a challenging open problem.

References

1. M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, 1973.
2. A. Andoni, P. Indyk, and M. Patrascu. On the optimality of the dimensionality reduction method. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 449–458, 2006.
3. M. Badoiu and K. L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008.
4. A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. *Computer Networks*, 29(8-13):1157–1166, 1997.
5. M. Bury and C. Schwiegelshohn. On Finding the Jaccard Center. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, volume 80 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 23:1–23:14, 2017.
6. M. Bury, C. Schwiegelshohn, and M. Sorella. Sketch 'em all: Fast approximate similarity search for dynamic data streams. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 72–80, 2018.
7. Z.-Z. Chen, B. Ma, and L. Wang. Randomized fixed-parameter algorithms for the closest string problem. *Algorithmica*, 74(1):466–484, 2016.
8. F. Chierichetti and R. Kumar. LSH-preserving functions and their applications. *J. ACM*, 62(5):33, 2015.
9. F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 219–228, 2009.
10. F. Chierichetti, R. Kumar, S. Pandey, and S. Vassilvitskii. Finding the Jaccard median. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 293–311, 2010.
11. M. Chimani, M. Woste, and S. Böcker. A closer look at the closest string and closest substring problem. In *Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments, ALENEX 2011, Holiday Inn San Francisco Golden Gateway, San Francisco, California, USA, January 22, 2011*, pages 13–24, 2011.
12. K. L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms*, 6(4), 2010.
13. E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, and C. Yang. Finding interesting associations without support pruning. *IEEE Trans. Knowl. Data Eng.*, 13(1):64–78, 2001.
14. M. Cygan, D. Lokshtanov, M. Pilipczuk, M. Pilipczuk, and S. Saurabh. Lower bounds for approximation schemes for closest string. In *15th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2016, June 22-24, 2016, Reykjavik, Iceland*, pages 12:1–12:10, 2016.
15. A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 271–280, 2007.
16. M. R. Fellows, J. Gramm, and R. Niedermeier. On the parameterized intractability of CLOSEST substringsize and related problems. In *STACS 2002, 19th Annual Symposium on Theoretical Aspects of Computer Science, Antibes - Juan les Pins, France, March 14-16, 2002, Proceedings*, pages 262–273, 2002.
17. J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.
18. J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48, 1986.
19. J. Gramm, R. Niedermeier, and P. Rossmanith. Fixed-parameter algorithms for CLOSEST STRING and related problems. *Algorithmica*, 37(1):25–42, 2003.
20. S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Inf. Syst.*, 25(5):345–366, 2000.
21. J. Guo, D. Hermelin, and C. Komusiewicz. Local search for string problems: Brute-force is essentially optimal. *Theor. Comput. Sci.*, 525:30–41, 2014.
22. P. Jaccard. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901.
23. S. Janssens. *Bell inequalities in cardinality-based similarity measurement*. PhD thesis, Ghent University, 2006.

24. P. Kumar, J. S. B. Mitchell, and E. A. Yildirim. Approximate minimum enclosing balls in high dimensions using core-sets. *ACM Journal of Experimental Algorithmics*, 8, 2003.
25. M. Li, B. Ma, and L. Wang. On the closest string and substring problems. *J. ACM*, 49(2):157–171, 2002.
26. B. Ma and X. Sun. More efficient algorithms for closest string and substring problems. *SIAM J. Comput.*, 39(4):1432–1443, 2009.
27. D. Marx. Closest substring problems with small distances. *SIAM J. Comput.*, 38(4):1382–1410, 2008.
28. M. Mitzenmacher and E. Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.
29. R. Real and J. M. Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.
30. D.J. Rogers and T.T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.
31. H. Späth. The minisum location problem for the Jaccard metric. *Operations-Research-Spektrum*, 3(2):91–94, 1981.
32. G. A. Watson. An algorithm for the single facility location problem using the Jaccard metric. *SIAM Journal on Scientific and Statistical Computing*, 4(4):748–756, 1983.
33. P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.
34. E. A. Yildirim. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.