On Finding the Jaccard Center*

Marc Bury[1] and Chris Schwiegelshohn[2]

1 **Thyssenkrupp Industrial Solutions AG, Essen, Germany**
`marc.bury@thyssenkrupp.com`
2 **Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Rome, Italy**
`chris.schwiegelshohn@tu-dortmund.de`

Abstract

We initiate the study of finding the Jaccard center of a given collection N of sets. For two sets X, Y, the Jaccard index is defined as $|X \cap Y|/|X \cup Y|$ and the corresponding distance is $1 - |X \cap Y|/|X \cup Y|$. The Jaccard center is a set C minimizing the maximum distance to any set of N.

We show that the problem is NP-hard to solve exactly, and that it admits a PTAS while no FPTAS can exist unless $P = NP$. Furthermore, we show that the problem is fixed parameter tractable in the maximum Hamming norm between Jaccard center and any input set. Our algorithms are based on a compression technique similar in spirit to coresets for the Euclidean 1-center problem.

In addition, we also show that, contrary to the previously studied median problem by Chierichetti et al. (SODA 2010), the continuous version of the Jaccard center problem admits a simple polynomial time algorithm.

1998 ACM Subject Classification F.2.2 Computations on Discrete Structures

Keywords and phrases Clustering, 1-Center, Jaccard

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.23

1 Introduction

The Jaccard index is a widely used similarity measure on item sets. Given two sets X and Y over a base set U, the similarity is defined as $J(X, Y) = |X \cap Y|/|X \cup Y|$ and the distance is $D(X, Y) = 1 - J(X, Y) = |X \triangle Y|/|X \cup Y|$, where $X \triangle Y$ denotes the symmetric difference of X and Y. In this paper we study the problem of finding the center of a given set of item sets under the Jaccard distance, i.e. for a given collection of sets $N = \{X_1, \ldots, X_n\}$ finding a set $C \subset U$ such that $\max_{X \in N} D(X, C)$ is minimized.

The Jaccard index is arguably the oldest [25] and best known similarity measure on binary data. It has found a wide range of applications such as plagiarism detection [7], association rule mining [12], collaborative filtering [13], web compression [9], biogeographical analysis [34], and chemical similarity searching [39]. Most theoretical computer science research dealing with the Jaccard index focuses on hashing algorithms for nearest neighbor problems, which was pioneered by Broder [6], though a number of publications also deal with clustering tasks on the Jaccard metric, see, for instance, Guha et al [22]. Previous research

* This work was supported by the German Research Council (DFG) within the Collaborative Research Center SFB 876, project A2, and the Google Focused Award on Web Algorithmics for Large-scale Data Analysis.

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl;
Article No. 23; pp. 23:1–23:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

most closely related to this paper addresses the Jaccard median problem, i.e. finding an item set that minimizes the sum of Jaccard distances, see [35, 37].Only recently did Chierichetti et al. [10] show that the Jaccard median problem is NP-hard but also admits a PTAS.

From a more general perspective, the task of finding a single center in a metric space has been studied in various forms dating back to the 19th century [36]. In constant Euclidean space, linear time algorithms exist [30, 38]. In higher dimensions, approximate algorithms based on (weak) coresets have been proposed [3, 4, 11, 26, 40]. Hardness for the 1-center problem in certain finite metrics have been established, including permutation metrics such as Kendall tau and Cayley distances [2, 5, 33], the edit distance on strings [14, 32], and the Hamming metric on strings [16, 27]. The latter problem, also known as the *closest string problem*, is one of the most widely studied center problems in computer science with numerous results on fixed parameter algorithms [15, 21, 29] and approximation algorithms [18, 27, 28]. The more general k-center problem admits a tight 2-approximation in any metric space [19, 23], though some improvements are possible in restricted metrics such as Euclidean space [4].

Our Contribution

We show that the problem is NP-hard to solve exactly, even when the input item sets have cardinality 2. Since the Jaccard distance is a metric, any input point is a trivial 2-approximate solution, and it is easy to see that this bound is tight. We propose two algorithms for the problem. The first algorithm is a PTAS with running time $|N|^{O(\varepsilon^{-6})}|U|^2$. The second one is an FPT algorithm with parameter $k = \max_{X \in N} |X \triangle C|$, i.e. the maximum Hamming norm of input points and Jaccard center C and running time $2^{O(k^3)} \cdot |N| \cdot |U|^3$. As a consequence of our hardness result, we show that under the exponential time hypothesis [24] no FPT algorithm with parameter k and running time $2^{o(k)}$ and no PTAS with running time $2^{o(\sqrt{1/\varepsilon})}$ can exist.

Lastly, we also briefly remark on the continuous version of the problem. Here the input points are non-negative d-dimensional real vectors and $J_c(X,Y) = \frac{\sum_{i=1}^{d} \min(X_i,Y_i)}{\sum_{i=1}^{d} \max(X_i,Y_i)}$. While the Jaccard median problem remains NP-hard for the continuous setting [10, 35], the center problem becomes solvable in polynomial time.

Our Techniques

Our algorithms are based on the existence of a small subset of input points we call *core-covers*. Informally, the union of items of all sets in the *core-cover* contains the (majority of) items of some optimal center C. Specifically, the intersection of the union of items with C yields an α-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. Crucially, we show that the size of an appropriate (anchored) core-cover is independent of the input when aiming for a $(1+\varepsilon)$-approximation, and dependent only on the parameter k in the context of the FPT algorithm.

In an approximate variant, a core-cover is similar to but weaker than a weak coreset for the Euclidean minimum enclosing ball problem, which requires that the expansion of the minimum enclosing ball computed on the coreset by an $(1+\varepsilon)$-factor contains the entire point set. The existence of constant size weak coresets has been widely studied and utilized [3, 4, 11, 26, 40]. Though the Jaccard distance can be isometrically embedded into (high dimensional) squared Euclidean space, see Gower and Legendre [20], the weak coreset results do not seem to be applicable to the constrained set of solutions corresponding to

embedded item sets. Stronger coreset guarantees extending to arbitrary centers require an exponential dependency on the dimension [1] and therefore also do not seem to be feasible for our purposes.

For the PTAS, we proceed as follows. It turns out that a natural LP relaxation can be efficiently rounded for a large fraction of inputs, namely when for all input sets $X \in N$, we have $\mathrm{OPT} \cdot |X| \in \Omega(\log n/\varepsilon^2)$, where OPT denotes the objective value of the optimum center. If the LP cannot be efficiently rounded, the symmetric difference between any two input sets as well as the optimum set is bounded by $O(\log n/\varepsilon^4)$. A QPTAS now immediately follows by choosing an arbitrary set X, iterating over all subsets S of the base set U with $|S| \in O(\log n/\varepsilon^4)$, and determining the best solution among all candidate centers $X \triangle S$. If we choose multiple sets $X_1, \dots, X_m$ then the number of candidate subsets S will be reduced. In fact, if $X_1, \dots, X_m$ is an anchored core-cover then the dependency on the size of the base set $|U|$ can be replaced by some constant depending only on m and ε. Since there exist anchored core-covers of size $O(1/\varepsilon)$, we obtain a polynomial running time for any fixed ε.

For the FPT-algorithm, the main technical difficulties are to show (1) that the size of an appropriate core-cover can be bounded in terms of the parameter k and (2) that we can efficiently construct an anchored core-cover. As was the case for the PTAS, for a given preliminary anchored core-cover M, we can compute an induced optimum via complete enumeration. If the induced optimum has distance at most OPT to all sets $X \in N$, we are done. Otherwise, any set violating this bound can be added to M. The improvement rate of each added set matches the non-constructive bounds used to show the existence of core-covers, ensuring that the algorithm terminates quickly.

2 Preliminaries

Let $U = \{u_1, \dots u_d\}$ be a base set containing d elements and let $N \subset \mathcal{P}(U)$ be a collection of n subsets of U. Denote the symmetric difference of two sets by $X \triangle Y = (X \setminus Y) \cup (Y \setminus X)$.

▶ **Definition 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 convenient to refer to specific elements of a set X by the characteristic vector $X \in \{0,1\}^d$ where $X_i = 1$ if $u_i \in X$ and $X_i = 0$ otherwise. The extension of the Jaccard measure to vectors with non-negative but otherwise arbitrary entries is as follows.

▶ **Definition 2** (Continuous Jaccard Measures)**.** Given two d dimensional vectors X, Y with non-negative real entries, the *continuous Jaccard similarity* is defined as

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

and the *continuous Jaccard distance* is defined as $D_c(X,Y) = 1 - J_c(X,Y)$.

In both cases the Jaccard distance is a metric. We say that the *Jaccard center* of a collection N is the set $C \subseteq U$ (resp. a non-negative real vector $C \in \mathbb{R}^n_{\geq 0}$ for the continuous case) such that $\max_{X \in N} D(X,C)$ is minimized. Throughout this paper we denote by OPT the

value of $\min_{C \subseteq U} \max_{X \in N} D(X, C)$. We always assume $\emptyset \notin N$, i.e. the empty set is not part of the input, as otherwise $\emptyset$ is a trivial optimal solution with maximum distance 1 if there exists at least one further set in N, and maximum distance 0 if $N = \{\emptyset\}$. Lastly, we will frequently use the following easily verifiable facts throughout the paper.

▶ **Fact 3.** *Let $X, Y \subseteq U$ be two item sets. Then the following statements hold:*
- $|X \cap Y| = (1 - D(X, Y)) \cdot |X \cup Y|$,
- $|X| \geq (1 - D(X, Y)) \cdot |Y|$,
- $|X \setminus Y| \leq D(X, Y) \cdot |X|$.

3 Hardness of Binary Jaccard Center

We reduce the problem of finding the optimum Jaccard center from vertex cover defined as follows.

▶ **Definition 4.** Given a graph $G(V, E)$, a vertex cover is a set $K \subset V$ such that $e \cap K \neq \emptyset$ for any $e \in E$. The minimum vertex cover is the vertex cover of smallest cardinality.

It is well known that computing the minimum vertex cover is NP-hard [17]. We will use instances with a minor constraint added for technical reasons. The minimum vertex cover will always have cardinality at most $\frac{|V|}{2} - 2$. It is easy to see that this does not affect the hardness of the vertex cover problem, for instance by adding an isolated star with one central node and $|V| + 5$ remaining nodes.

▶ **Theorem 5.** *Computing the optimum Jaccard center is NP-hard even if every $X \in N$ has cardinality at most 2.*

Proof. Let K be a minimum vertex cover of cardinality at most $\frac{|V|}{2} - 2$ in a graph $G(V, E)$ with no isolated nodes. Consider now the instance of the Jaccard center problem where the input item sets are E, the base set is V, and the center is some subset of V. We claim that a collection of vertexes C is an optimum Jaccard center if and only if C is a minimum vertex cover.

For every collection of vertices C and any edge $e \in E$, we have the following three cases:

$$D(e, C) = \begin{cases} 1 & \text{if } |C \cap e| = 0 \\ \frac{|C|}{|C|+1} & \text{if } |C \cap e| = 1 \\ \frac{|C|-2}{|C|} & \text{if } |C \cap e| = 2. \end{cases}$$

Note that the distance for some edge is 1 if and only if C is not a vertex cover. Note also that $\frac{|C|}{|C|+1} > \frac{|C|-2}{|C|}$, i.e. if $C \neq V$ then $\max_{e \in E} D(e, C) = \frac{|C|}{|C|+1}$. Now for any collection of vertices C that is a vertex cover with $|C| > |K|$, we have two cases. If $C \neq V$, then

$$\max_{e \in E} D(e, C) = \frac{|C|}{|C|+1} \geq \frac{|K|+1}{|K|+2} > \frac{|K|}{|K|+1} = \max_{e \in E} D(e, K).$$

If $C = V$, then

$$\max_{e \in E} D(e, V) = \frac{|V|-2}{|V|} = \frac{\frac{|V|}{2}-1}{\frac{|V|}{2}} \geq \frac{|K|+1}{|K|+2} > \frac{|K|}{|K|+1} = \max_{e \in E} D(e, K). \qquad \blacktriangleleft$$

▶ **Corollary 6.** *There exists no FPTAS for the binary Jaccard center problem unless P=NP.*

Proof. Two non-equal distances are at least apart by $\frac{1}{d^2}$. If an FPTAS were to exist, we could compute determine a $(1+\frac{1}{d^2})$ approximation in polynomial time. This approximation however would coincide with the optimal solution. ◀

Assuming the exponential time hypothesis (ETH), we can give stronger time bounds for PTAS and FPT. ETH, formulated by Impagliazzio, Paturi and Zane [24] assumes that there exists some positive real number s such that 3-SAT with n variables and m clauses cannot be decided in time $2^{s\cdot n}(n+m)^{O(1)}$.

▶ **Corollary 7.** *Let N be a collection of subsets over a base set U and let $C \subset U$ be the optimal Jaccard center. Assuming ETH, no FPT algorithm with parameter $k = \max_{X\in N} |C \triangle X|$, can run in time $2^{o(k)}poly(N,d)$. Further, no PTAS for the Jaccard Center problem can run in time $2^{o(\sqrt{1/\varepsilon})}poly(N,d)$.*

Proof. Under ETH, no FPT algorithm for vertex cover with parameter $|K|$, the minimal size of the vertex cover, can run in time $2^{o(|K|)}\text{poly}(N)$, see Cai and Juedes [8]. Since $k = \max_{X\in N} |C \triangle X| \in \Theta(|K|)$, the first claim follows. For the second claim, recall any PTAS approximating the Jaccard center problem beyond a factor of $(1+\frac{1}{d^2})$ recovers the optimal solution. ◀

4 Core-Covers

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. The construction is somewhat inspired by coresets for the Euclidean minimum enclosing ball problem, albeit with a weaker guarantee.

▶ **Definition 8** (Core-Covers)**.** Let N be a collection of subsets of a base set U, let OPT be the maximum distance of an optimal Jaccard center to any subset in N, and let $\alpha \geq 1$ be a parameter. A collection $M \subseteq N$ is called an *α-core-cover* if there exists an optimal center C with

$$\max_{X\in N} D\left(X, \left(\bigcup_{X\in M} X\right) \cap C\right) \leq \alpha \cdot \text{OPT}.$$

A collection $M \subseteq N$ with $A_M = \bigcap_{X\in M} X$ and $O_M = \bigcup_{X,Y\in M} X \triangle Y$ is called an *anchored α-core-cover* if there exists an optimal center C with

$$\max_{X\in N} D(X, A_M \cup (O_M \cap C)) \leq \alpha \cdot \text{OPT}.$$

We are especially interested in the size of core-covers with $\alpha = 1$ or $\alpha = 1+\varepsilon$. Core-covers are useful when the supports, i.e. the sets X are small, in which case we can find the solution by enumerating over all possible subsets of $\bigcup_{X\in M} X$. Anchored core-covers are more useful if the supports are large while the optimum value is small. For the remainder of this section, we will give (non-constructive) upper and lower bounds on the number of points required to satisfy both guarantees. Our proofs are essentially based on the following observation.

▶ **Observation 1.** *For any three sets $C, K, X \subseteq U$*

$$D(X,K) \leq D(X, K\cap C) + \frac{|K\setminus C| - 2|(X\cap K)\setminus C)|}{|X\cup K|}.$$

Proof.

$$\begin{aligned} D(X,K) &= \frac{|X \triangle K|}{|X \cup K|} = \frac{|X \triangle (K \cap C)| + |K \setminus C \setminus X| - |X \cap (K \setminus C)|}{|X \cup (K \cap C)| + |K \setminus C \setminus X|} \\ &\leq \frac{|X \triangle (K \cap C)|}{|X \cup (K \cap C)|} + \frac{|K \setminus C \setminus X| - |X \cap (K \setminus C)|}{|X \cup K|} \\ &= D(X, K \cap C) + \frac{|K \setminus C| - 2|X \cap (K \setminus C)|}{|X \cup K|} \end{aligned}$$ ◀

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 $D(X, K \cap C)$ is a good approximation to $D(X, C)$ and $\frac{|K \setminus C| - 2|(X \cap K) \setminus C)|}{|X \cup K|}$ is small or negative.

▶ **Lemma 9.** *For any collection of subsets N, there exists an α-core-cover M of size $\lceil 1/\varepsilon \rceil + 1$ if $\alpha = 1 + \varepsilon$ with $\varepsilon > 0$ and $\min\left\{\frac{\log(OPT \cdot |C|)}{\log(2 - OPT)} + 1, |C|\right\}$ if $\alpha = 1$.*

Proof. We show the existence of the collection M by proving that we can iteratively add a set to M such that either K is already a good approximate solution or the added set contains many elements from $C \setminus K$. Thus, finally we either have C covered by $\bigcup_{X \in M} X$ or no set violates the approximation guarantee. Let $M^{(0)} = \{X\}$ for an arbitrary $X \in N$. We denote by $K^{(i)} = C \cap \left(\bigcup_{X \in M^{(i)}} X\right)$ our solution after the i-th iteration. Note that due to Fact 3, we can assume $|C \setminus K^{(i)}| \leq \mathrm{OPT} \cdot |C|$ as $M^{(i)}$ is non-empty. In the following derivations, we assume that $\alpha \cdot \mathrm{OPT} < 1$, which is always the case for $\alpha = 1$ and always the case for $\alpha = 1 + \varepsilon$ and $\mathrm{OPT} \leq \frac{1}{1+\varepsilon}$. The latter assumption is justified by observing that otherwise any single input point already satisfies the $(1 + \varepsilon)$-core-cover guarantee.

Let $X \in N$ be a set such that $D\left(X, K^{(i)}\right) > \alpha \cdot \mathrm{OPT}$. Then

$$\begin{aligned} |X \cap (C \setminus K^{(i)})| &\overset{K^{(i)} \subseteq C}{=} |X \cap C| - |X \cap K^{(i)}| \\ &\geq (1 - \mathrm{OPT}) \cdot |X \cup C| - (1 - D(X, K^{(i)})) \cdot |X \cup K^{(i)}| \\ &> (1 - \mathrm{OPT}) \cdot |X \cup C| - \\ &\quad (1 - \alpha \cdot \mathrm{OPT}) \cdot (|X \cup C| - |C \setminus K^{(i)}| + |X \cap (C \setminus K^{(i)})|) \\ &\geq (\alpha - 1) \cdot \mathrm{OPT} \cdot |C| + \\ &\quad (1 - \alpha \cdot \mathrm{OPT}) \cdot (|C \setminus K^{(i)}| - |X \cap (C \setminus K^{(i)})|) \end{aligned}$$

For for $\alpha = 1 + \varepsilon$, we have the lower bound $|X \cap (C \setminus K^{(i)})| \geq \varepsilon \cdot \mathrm{OPT} \cdot |C|$. Since $|C \setminus K^{(0)}| \leq \mathrm{OPT} \cdot |C|$, after adding at most $s = \lceil 1/\varepsilon \rceil$ sets to $M^{(0)}$, we have $K^{(s)} = C$, or no set X with $D\left(X, K^{(s)}\right) > (1 + \varepsilon) \cdot \mathrm{OPT}$ exists.

If $\alpha = 1$, we have

$$|X \cap (C \setminus K^{(i)})| \geq \frac{1 - \mathrm{OPT}}{2 - \mathrm{OPT}} \cdot |C \setminus K^{(i)}|$$

which implies that X covers at least $\frac{1-\mathrm{OPT}}{2-\mathrm{OPT}}$ items from $C \setminus K^{(i))}$ in iteration i. Thus, $|C \setminus K^{(i)}| \leq (1 - \frac{1-\mathrm{OPT}}{2-\mathrm{OPT}})^i |C \setminus K^{(0)}| \leq (\frac{1}{2-\mathrm{OPT}})^i \cdot \mathrm{OPT} \cdot |C|$ which is smaller than 1 if $i > \frac{\log(\mathrm{OPT} \cdot |C|)}{\log(2-\mathrm{OPT})}$. Note that $|X \cap (C \setminus K^{(i)})| \geq 1$ if $D(X, K^{(i)}) > OPT$ which concludes the proof. ◀

With the space bound for core-covers, we can prove the main result of this section.

▶ **Lemma 10.** *For any collection of subsets N, there exists an anchored α-core-cover $M \subset N$ of size $O(1/\varepsilon)$ if $\alpha = 1 + \varepsilon$ with $\varepsilon > 0$ and of size $\min\{\frac{\log(OPT \cdot |C|)}{\log(2 - OPT)} + 1, |C|\} + \log \frac{OPT \cdot |C|}{1 - OPT}$ if $\alpha = 1$.*

Proof. Assume we have some optimal center C. Lemma 9 gives a set M such that $K \cap C$ is an α-approximate solution where we can represent K as $K = A_M \cup (O_M \cap C)$. Using Observation 1, the distance between K and some arbitrary set X is

$$\begin{aligned} D(X,K) &\leq D(X, K\cap C) + \frac{|K \setminus C| - 2\cdot|(X\cap K)\setminus C)|}{|X\cup K|} \\ &= D(X, K\cap C) + \frac{|A_M \setminus C| - 2\cdot|X\cap(A_M\setminus C)|}{|X\cup K|} \\ &\leq \alpha\cdot\mathrm{OPT} + \frac{|A_M \setminus C| - 2\cdot|X\cap(A_M\setminus C)|}{|X\cup K|} \end{aligned}$$

If for every $X \in N$, we have $2\cdot|X\cap(A_M\setminus C)| > |A_M\setminus C|$ then the ratio is negative and $D(X,K) \leq D(X, K\cap C) \leq \alpha\cdot\mathrm{OPT}$. Otherwise, there exists an X such that $|X\cap(A_M\setminus C)| = |(X\cap A_M)\setminus C| \leq |A_M\setminus C|/2$. We iteratively augment the collection M satisfying the space and approximation bounds of Lemma 9 with additional sets X. In each iteration, $|A_M\setminus C|$ is halved.

If $\alpha = 1$ and after adding $i > \log|A_M\setminus C|$ sets, we have $A_M\setminus C = \emptyset$. For a more precise bound on i let $Y \in M$. Then due to Fact 3,

$$|A_M\setminus C| \leq |Y\setminus C| \leq \mathrm{OPT}\cdot|Y\cup C| \leq \frac{\mathrm{OPT}\cdot|C|}{1-\mathrm{OPT}}.$$

For the case $\alpha = 1+\varepsilon$, we assume $\mathrm{OPT} < 1/(1+\varepsilon)$ as otherwise any point is a $(1+\varepsilon)$ approximation. Let $X \in N$. Again due to Fact 3 we have

$$\begin{aligned} |A_M\setminus C| &\leq \mathrm{OPT}\cdot\frac{|C|}{1-\mathrm{OPT}} \leq \mathrm{OPT}\cdot\frac{|X|}{(1-\mathrm{OPT})^2} \\ &\leq \mathrm{OPT}\cdot\frac{(1+\varepsilon)^2\cdot|X|}{\varepsilon^2} \leq \mathrm{OPT}\cdot\frac{4}{\varepsilon^2}\cdot|X|, \end{aligned}$$

where the last inequality follows for $\varepsilon \leq 1$. After adding $\log\frac{4}{\varepsilon^3}$ sets such that $|A_M\setminus C|$ is halved with each sets, we have $|A_M\setminus C|/|X\cup K| \leq \varepsilon\cdot\mathrm{OPT}\cdot|X|/|X\cup K| \leq \varepsilon\cdot\mathrm{OPT}$. Our approximation factor is therefore $\alpha\cdot\mathrm{OPT} + \varepsilon\cdot\mathrm{OPT} = (1+2\varepsilon)\cdot\mathrm{OPT}$. Rescaling ε by a factor of 2 completes the proof. ◀

We would like to remark that the bound on the number of sets required to satisfy the $(1+\varepsilon)$-core-cover guarantee is tight, and that the bound on the number of sets to satisfy the anchored $(1+\varepsilon)$-core-cover guarantee is tight up to constant multiplicative factors. Note that M is constrained to using only input sets. Better bounds are possible when we lift this restriction on M (for instance, if M consists of only an optimum center C then all guarantees are met). It is unclear whether improved guarantees not using input sets can be feasibly used in an algorithm.

▶ **Lemma 11.** *There exists a collection of subsets N such that for any $(1+\varepsilon)$-core-cover $M \subseteq N$, we have $|M| \geq 1/\varepsilon - 1$.*

Proof. For a given $\varepsilon > 0$ and assuming $1/\varepsilon$ to be an integer, we consider the following instance of vertex cover. We are given $1/\varepsilon - 1$ stars, each with at least two leaves. The optimum vertex cover and the optimum Jaccard center consists of the internal nodes, with an optimum objective value for the Jaccard center of $\frac{1/\varepsilon-1}{1/\varepsilon}$. If M does not consist of at least one edge from each star, corresponding to a set containing the element contained in the optimal Jaccard center, any center computed using only the entries of the picked edges will not intersect with at least one star, i.e. have distance 1 to the edges of the omitted star. Since $\frac{1/\varepsilon-1}{1/\varepsilon}\cdot(1+\varepsilon) = 1-\varepsilon^2 < 1$, M has to hit every star. ◀

Algorithm 1: PTAS for the Jaccard center problem

Input : Collection N of subsets, Parameter $\varepsilon > 0$
Output : $(1+\varepsilon)$-approximate Jaccard center C

1 Let $D = \{\frac{i}{j} \mid 1 \leq j \leq d \text{ and } 0 \leq i < j\}$.
2 Initialize list $C = \emptyset$.
3 **foreach** $\widehat{OPT} \in D$ **do**
4 **if** $\exists X \in N : \ \widehat{OPT} \cdot |X| < \frac{27 \ln(4n)}{\varepsilon^2}$ **then**
5 **foreach** $M \subseteq N$ with $|M| = \lceil \frac{5}{\varepsilon} + 5 \rceil$ **do**
6 Compute optimal solution $K_{\widehat{\text{OPT}}} = A_M \cup S$ with $S \subseteq O_M$ (cf. Lemma 10).
7 Add $K_{\widehat{\text{OPT}}}$ to C
8 **else**
9 Obtain non-integral solution $K'_{\widehat{\text{OPT}}}$ by solving the set of linear equations given by Equation 1
10 Obtain $K_{\widehat{\text{OPT}}}$ by rounding each entry of $K'_{\widehat{\text{OPT}}}$
11 Add $K_{\widehat{\text{OPT}}}$ to C
12 **return** $\underset{\widehat{\text{OPT}} \in D}{\text{argmin}} \{K_{\widehat{\text{OPT}}} \in C\}$

5 A PTAS for Binary Jaccard Center

This section mainly consists of the proof of the following theorem.

▶ **Theorem 12.** *Given a collection N of n subsets from a base set U of cardinality d and any $\varepsilon > 0$, there exists an algorithm computing a $(1+\varepsilon)$-approximation to the optimal Jaccard center. The algorithm runs in time $d^2 \cdot (n^{O(\varepsilon^{-6})} + LP(n,d))$, where $LP(n,d)$ is the time required to solve a linear program with n constraints and d variables.*

The algorithm (see also Algorithm 1) consists of two main steps. Let OPT be the optimal objective value. Since there are $O(d^2)$ distinct objective values for the Jaccard center problem with a base set of size d, we can try to a find solution for each value (cf. line 3 Algorithm 1). Recall that $C_i = \begin{cases} 0 & \text{if } i \notin C \\ 1 & \text{if } i \in C \end{cases}$ and that $D(X,C) \leq \text{OPT}$ holds for all $X \in N$. By multiplying both sides of the inequality with $|X \cup C|$, we obtain

$$|X \triangle C| \leq \widehat{\text{OPT}} \cdot |X \cup C|. \tag{1}$$

Observe that $|X \triangle C| = \sum_{i=1}^{d} X_i - 2X_iC_i + C_i$ and $|X \cup C| = \sum_{i=1}^{d} X_i - X_iC_i + C_i$. Hence, we obtain a set of linear inequalities which we can test for feasibility by relaxing the integrality constraints on C. Denote a feasible non-integral solution by C'. The existence of a feasible integral solution of Equation 1 implies a feasible relaxed solution C'. We interpret the C'_i as probabilities, i.e. we obtain a binary vector C by rounding each C'_i to 1 with probability C'_i. Using Chernoff bounds, this approach yields a good solution if $\text{OPT} \cdot |X| > s \cdot \log n/\varepsilon^2$ for all X and some constant s (cf. lines 4–7 of Algorithm 1).

If $\text{OPT} \cdot |Y|$ is smaller than this threshold for at least one $Y \in N$ then we could employ a naive brute force algorithm by iterating over all $\binom{d}{s \cdot \log n/\varepsilon} \in O(d^{s \cdot \log n/\varepsilon})$ subsets S and outputting the best $Y \triangle S$. To eliminate the dependency on d, we first show that a bound on $\text{OPT} \cdot |Y|$ implies that $|X_1 \triangle X_2|$ for any two sets $X_1, X_2 \in N$ is bounded. Then we compute an anchored core-cover M by enumerating all collections of $O(1/\varepsilon)$ input sets.

Having determined M, computing the optimum $A_M \cup S$ with $S \subseteq O_M$ becomes feasible (cf. lines 9–11 of Algorithm 1).

Proof of Theorem 12. In the following, we always assume that $\mathrm{OPT} < 1/(1+\varepsilon)$, as otherwise any solution is a $(1+\varepsilon)$ approximation.

To round the set of linear Equations 1, we first recall and apply the following probabilistic bounds.

▶ **Theorem 13** (Multiplicative Chernoff-Bounds [31]). *Let $B_1, \dots B_d$ 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 14.** *Let S be a random binary vector obtained by rounding a fractional feasible solution of the set of Equations 1 and let $\varepsilon > 0$ be a constant. Assume that $OPT\cdot|X| \geq \frac{27\ln(4n)}{\varepsilon^2}$ for all $X \in N$. Then with probability at least $1/2$, the rounding procedure produces a binary solution S with $\max_{X\in N} D(X,S) \leq (1+\varepsilon)\cdot OPT$.*

Proof. Observe that $\mathbb{E}[|X\cup S|] \geq |X|$. We first derive concentration bounds on $|X \triangle S|$ and $|X \cup S|$. For any $X \in N$, Theorem 13 yields

$$\mathbb{P}\left[|X\cup S| < (1-\varepsilon/3)\cdot\mathbb{E}[|X\cup S|]\right] \leq \exp\left(-\frac{\varepsilon^2\cdot\mathbb{E}[|X\cup S|]}{18}\right) \leq \exp\left(-\frac{\varepsilon^2\cdot|X|}{18}\right) \leq \frac{1}{4n}$$

and

$$\begin{aligned}
& \mathbb{P}\left[|X\triangle S| > \mathbb{E}[|X\triangle S|] + \varepsilon/3\cdot\mathrm{OPT}\cdot\mathbb{E}[|X\cup S|]\right] \\
= \;& \mathbb{P}\left[|X\triangle S| > \left(1+\frac{\varepsilon\cdot\mathrm{OPT}\cdot\mathbb{E}[|X\cup S|]}{3\cdot\mathbb{E}[|X\triangle S|]}\right)\cdot\mathbb{E}[|X\triangle S|]\right] \\
\leq \;& \exp\left(-\frac{\varepsilon^2\cdot\mathrm{OPT}^2\cdot\mathbb{E}[|X\cup S|]^2}{27\cdot\mathbb{E}[|X\triangle S|]^2}\cdot\mathbb{E}[|X\triangle S|]\right) \\
\leq \;& \exp\left(-\varepsilon^2\cdot\mathrm{OPT}\cdot\mathbb{E}[|X\cup S|]/27\right) \leq \exp\left(-\varepsilon^2\cdot\mathrm{OPT}\cdot|X|/27\right) \leq \frac{1}{4n}.
\end{aligned}$$

Combining these two bounds, we have

$$\frac{|X\triangle S|}{|X\cup S|} \leq \frac{\mathbb{E}[|X\triangle S|] + \varepsilon/3\cdot\mathrm{OPT}\cdot\mathbb{E}[|X\cup S|]}{(1-\varepsilon/3)\cdot\mathbb{E}[|X\cup S|]} \leq \frac{\mathrm{OPT}+\varepsilon/3\cdot\mathrm{OPT}}{1-\varepsilon/3} \leq (1+\varepsilon)\cdot\mathrm{OPT}$$

with probability at least $1 - 1/2n$. Applying the union bound, we then obtain

$$\begin{aligned}
& \mathbb{P}\left[\max_{X\in N} D(X,S) \leq (1+\varepsilon)\cdot\mathrm{OPT}\right] \\
& = 1 - \mathbb{P}\left[\exists X\in N:\ \frac{|X\triangle S|}{|X\cup S|} > (1+\varepsilon)\cdot\mathrm{OPT}\right] \geq 1 - \frac{n}{2n} = 1/2.
\end{aligned}$$

◀

If $\mathrm{OPT}\cdot|X| > \frac{27\ln(4n)}{\varepsilon^2}$ for all $X \in N$, we can use the LP-based rounding scheme analyzed in Lemma 14 (cf. lines 4–7 of Algorithm 1). For the other cases, we will utilize Lemma 10 as follows. There exists at least one set Y with $\mathrm{OPT}\cdot|Y| \leq \frac{27\ln(4n)}{\varepsilon^2}$. With Fact 3, we have

$\mathrm{OPT} \cdot |C| \leq \mathrm{OPT} \cdot |Y|/(1-\mathrm{OPT}) \leq \frac{27\cdot(1+\varepsilon)\cdot\ln(4n)}{\varepsilon^3}$. For any two sets $X_1, X_2 \in N$, we then have

$$\begin{aligned} |X_1 \triangle X_2| &\leq 2 \cdot \mathrm{OPT} \cdot |X_1 \cup X_2| \leq 2 \cdot \mathrm{OPT} \cdot (|X_1| + |X_2|) \\ &\leq 4 \cdot \mathrm{OPT} \frac{|C|}{1-\mathrm{OPT}} \leq \frac{108 \cdot (1+\varepsilon)^2 \cdot \ln(4n)}{\varepsilon^4}. \end{aligned}$$

Let M now be a collection of sets satisfying the guarantee of Lemma 10 with $A_M = \bigcap_{X\in M} X$ and $O_M = \bigcup_{X,Y\in M} X \triangle Y$. Such a collection can be determined in time $n^{O(\varepsilon^{-1})}$ by iterating through all subsets of N of cardinality $O(\varepsilon^{-1})$. Since $|O_M| \leq \sum_{X_i\in M}\sum_{X_j\in M} |X_i \triangle X_j| \leq |M|^2 \cdot \max_{X_i,X_j\in M} |X_i \triangle X_j| \in O(\log n \cdot \varepsilon^{-6})$, we can compute an optimal solution of $\max_{X\in N} \min_{S\subseteq O_M} D(X, A_M \cup S)$ in time $2^{|O_M|} = 2^{O(\log n\cdot\varepsilon^{-6})}$.

The total running time amounts to d^2 calls to the LP given via Equations 1 or d^2 applications of Lemma 10 with a running time of $2^{O(\log n\cdot\varepsilon^{-6})} = n^{O(\varepsilon^{-6})}$. ◀

6 An FPT Algorithm for Binary Jaccard Center

Our second application of core-covers is an FPT algorithm in the parameter $k = \max_{X\in N} |X \triangle C|$ where C is an arbitrary optimal solution. The main technical difficulty is to efficiently construct a core-cover without enumerating all possible core-covers. We first bound the size of an anchored 1-core-cover given by Lemma 10 in terms of k.

▶ **Lemma 15.** *For any collection N of subsets and an optimal center C with cost $OPT < 1$, let $k = \max_{X\in N} |X \triangle C|$. Then*

$$\min\left\{\frac{\log(OPT\cdot|C|)}{\log(2-OPT)} + 1, |C|\right\} \leq 2k \text{ and } \log\frac{OPT\cdot|C|}{1-OPT} \leq 3\log k.$$

Proof. There exists an $X \in N$ such that $k \geq |X \triangle C| = \mathrm{OPT} \cdot |X \cup C| \geq \mathrm{OPT} \cdot |C|$. We first note that both terms are increasing with OPT, hence we assume $\mathrm{OPT} > 1/2$. Then $|X \triangle C|/|X \cup C| = \mathrm{OPT}$ for some $X \in N$ implying

$$(1-\mathrm{OPT}) = \mathrm{OPT} \cdot |X \cap C|/|X \triangle C| \geq \frac{1}{2|X \triangle C|} \geq \frac{1}{2k}.$$

Therefore, we have $1/(1-\mathrm{OPT})) \leq 2k$,

$$\log(2-\mathrm{OPT}) = \log(1+1-\mathrm{OPT}) = \frac{\ln(1+1-\mathrm{OPT})}{\ln 2} \geq \frac{1-\mathrm{OPT}}{2\ln 2} \geq \frac{1}{4k\ln 2},$$

and

$$\min\left\{\frac{\log(\mathrm{OPT}\cdot|C|)}{\log(2-\mathrm{OPT})} + 1, |C|\right\} \leq |C| \leq 2k \text{ and } \log\frac{\mathrm{OPT}\cdot|C|}{1-\mathrm{OPT}} \leq 1 + 2\log k.$$ ◀

For a given estimate of OPT, the algorithm initially chooses two arbitrary sets to be included in the anchored core-cover M. If the optimal solution $A_M \cup S$ with $S \subseteq O_M$ satisfies $\max_{X\in N} D(X, A_M \cup S) < \mathrm{OPT}$ then we can reduce our estimate of OPT. Otherwise, we add any set X at distance greater than OPT to M. The set X improves the core-cover, either by increasing $|C \cap (A_M \cup O_M)|$ or by decreasing $|A_M \setminus C|$ for some optimal center C. Lemma 15 allows us to bound the number of times this happens before M satisfies the anchored core-cover guarantee, upon which we can recover the optimum solution.

Algorithm 2: FPT-algorithm for the Jaccard center problem

Input : Collection N of subsets, Parameter $k = \max_{X \in N} |X \triangle C|$
Output : Optimal Jaccard center C

1 Let $D = \{\frac{i}{j} \mid 1 \leq j \leq d \text{ and } 0 \leq i < j\}$.
2 Initialize list $C = \emptyset$.
3 **foreach** $\widehat{OPT} \in D$ **do**
4 Initialize $M = \{X, Y\}$ with arbitrary $X, Y \in N$ and $X \neq Y$.
5 **for** $i = 1$ **to** $5k$ **do**
6 Compute optimal solution $K_{\widehat{\text{OPT}}} = A_M \cup S$ with $S \subseteq O_M$ (cf. Lemma 10).
7 **if** $\exists X \in N : D(X, K_{\widehat{OPT}}) > \widehat{OPT}$ **then**
8 $M = M \cup \{X\}$
9 **else**
10 Add $K_{\widehat{\text{OPT}}}$ to C
11 **break**
12 **return** $\operatorname{argmin}_{\widehat{\text{OPT}} \in D} \{K_{\widehat{\text{OPT}}} \in C\}$

▶ **Theorem 16.** *Algorithm 2 computes an optimal Jaccard center C satisfying $\max_{X \in N} |X \triangle C| = k$ in time $2^{O(k^3)} \cdot n \cdot d^3$.*

Proof. Let $\widehat{\text{OPT}} \in D$ be a guess for our optimal value OPT. If $\widehat{\text{OPT}} < \text{OPT}$ then the loop terminates without finding a center. Let $\widehat{\text{OPT}} \geq \text{OPT}$. Using Observation 1, we know that

$$D(X, K_{\widehat{\text{OPT}}}) \leq D(X, K_{\widehat{\text{OPT}}} \cap C) + \frac{|K_{\widehat{\text{OPT}}} \setminus C| - 2 \cdot |(X \cap K_{\widehat{\text{OPT}}}) \setminus C)|}{|X \cup K_{\widehat{\text{OPT}}}|}.$$

If $D(X, K_{\widehat{\text{OPT}}}) > \widehat{\text{OPT}}$ then we distinguish between two cases:

Case $|K_{\widehat{\text{OPT}}} \setminus C| - 2 \cdot |(X \cap K_{\widehat{\text{OPT}}}) \setminus C)| \leq 0$:
Then $D(X, K_{\widehat{\text{OPT}}} \cap C) > \widehat{\text{OPT}}$ and we can apply the analysis of Lemma 9. Thus, we add at most $2k$ sets to M until this case can no longer occur (Lemma 15).

Case $|K_{\widehat{\text{OPT}}} \setminus C| - 2 \cdot |(X \cap K_{\widehat{\text{OPT}}}) \setminus C)| > 0$:
Then $|(X \cap K_{\widehat{\text{OPT}}}) \setminus C)| \leq |K_{\widehat{\text{OPT}}} \setminus C|/2$ and we can apply the analysis of Lemma 10. Thus, we add at most $3 \log k$ points to M until this case can no longer occur (Lemma 15).

The computation of $K_{\widehat{\text{OPT}}}$ can be done in time $2^{O(k^3)}$ by an exhaustive search over all possible subsets of O_M since

$$|O_M| \leq |M^2| \cdot \max_{X,Y \in N} |X \triangle Y| \leq O(k^2) \cdot \max_{X,Y \in N} (|X \triangle C| + |Y \triangle C|) = O(k^3).$$

We perform the exhaustive search $O(k)$ times and for each solution we evaluate the objective value for each set. Since $|D| = O(d^2)$ and we examine every set in line 7 of Algorithm 2, the algorithm terminates in time $2^{O(k^3)} \cdot n \cdot d^3$. ◀

7 A Note on Continuous Jaccard Center

We conclude by briefly describing how to find the continuous Jaccard center. 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 $X^j \in N$ 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. We further use the variables $a_{i,j}$ and $b_{i,j}$ for all $i \in \{1, \ldots, d\}$ and $j \in \{1, \ldots n\}$ to denote the maximum and minimum of X_i^j and c_i. We then use the constraints

$$\sum_{i=1}^{d} b_{i,j} \geq (1 - dist) \cdot \sum_{i=1}^{d} a_{i,j} \qquad \text{for all } j \in \{1, \ldots n\}$$

$$b_{i,j} \leq c_i, X_i^j \leq a_{i,j} \qquad \text{for all } j \in \{1, \ldots n\}, i \in \{1, \ldots d\}$$

$$a_{i,j}, b_{i,j}, c_i \geq 0 \qquad \text{for all } j \in \{1, \ldots n\}, i \in \{1, \ldots d\}.$$

Note that the top most equation $\sum_{i=1}^{d} \min(c_i, X_i^j) \geq (1 - dist) \cdot \sum_{i=1}^{d} \max(c_i, X_i^j)$ is equal to $1 - \frac{\sum_{i=1}^{d} \min(X_i, Y_i)}{\sum_{i=1}^{d} \max(X_i, Y_i)} \leq dist$.

References

1. P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.
2. C. Bachmaier, F. J. Brandenburg, A. Gleißner, and A. Hofmeier. On the hardness of maximum rank aggregation problems. *J. Discrete Algorithms*, 31:2–13, 2015.
3. M. Badoiu and K. L. Clarkson. Optimal core-sets for balls. *Comput. Geom.*, 40(1):14–22, 2008.
4. M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 250–257, 2002.
5. T. C. Biedl, F.-J. Brandenburg, and X. Deng. On the complexity of crossings in permutations. *Discrete Mathematics*, 309(7):1813–1823, 2009.
6. A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, SEQUENCES'97, pages 21–, Washington, DC, USA, 1997. IEEE Computer Society.
7. 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.
8. L. Cai and D. W. Juedes. On the existence of subexponential parameterized algorithms. *J. Comput. Syst. Sci.*, 67(4):789–807, 2003.
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. K. L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms*, 6(4), 2010.
12. 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.
13. 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.

14 C. de la Higuera and F. Casacuberta. Topology of strings: Median string is NP-complete. *Theor. Comput. Sci.*, 230(1-2):39–48, 2000.

15 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.

16 M. Frances and A. Litman. On covering problems of codes. *Theory Comput. Syst.*, 30(2):113–119, 1997.

17 M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

18 L. Gasieniec, J. Jansson, and A. Lingas. Efficient approximation algorithms for the Hamming center problem. In *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms, 17-19 January 1999, Baltimore, Maryland.*, pages 905–906, 1999.

19 T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.

20 J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48, 1986.

21 J. Gramm, R. Niedermeier, and P. Rossmanith. Fixed-parameter algorithms for CLOSEST STRING and related problems. *Algorithmica*, 37(1):25–42, 2003.

22 S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Inf. Syst.*, 25(5):345–366, 2000.

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

24 R. Impagliazzo, R. Paturi, and F. Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001.

25 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.

26 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.

27 J. K. Lanctôt, M. Li, B. Ma, S. Wang, and L. Zhang. Distinguishing string selection problems. *Inf. Comput.*, 185(1):41–55, 2003.

28 M. Li, B. Ma, and L. Wang. On the closest string and substring problems. *J. ACM*, 49(2):157–171, 2002.

29 D. Marx. Closest substring problems with small distances. *SIAM J. Comput.*, 38(4):1382–1410, 2008.

30 N. Megiddo. Linear programming in linear time when the dimension is fixed. *J. ACM*, 31(1):114–127, 1984.

31 M. Mitzenmacher and E. Upfal. *Probability and computing – randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

32 F. Nicolas and E. Rivals. Hardness results for the center and median string problems under the weighted and unweighted edit distances. *J. Discrete Algorithms*, 3(2-4):390–415, 2005.

33 V. Y. Popov. Multiple genome rearrangement by swaps and by element duplications. *Theor. Comput. Sci.*, 385(1-3):115–126, 2007.

34 R. Real and J. M. Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385, 1996.

35 H. Späth. The minisum location problem for the Jaccard metric. *Operations-Research-Spektrum*, 3(2):91–94, 1981.

36 J. J. Sylvester. A Question in the Geometry of Situation. *Quarterly Journal of Pure and Applied Mathematics*, 1, 1857.

37 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.

38 E. Welzl. *Smallest enclosing disks (balls and ellipsoids)*, pages 359–370. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.

39 P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

40 E. A. Yildirim. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.