Hardness of Bichromatic Closest Pair with Jaccard Similarity

Rasmus Pagh
BARC, Copenhagen, Denmark
IT University of Copenhagen, Denmark
pagh@itu.dk

Nina Mesing Stausholm
BARC, Copenhagen, Denmark
IT University of Copenhagen, Denmark
nimn@itu.dk

Mikkel Thorup
BARC, Copenhagen, Denmark
University of Copenhagen, Denmark
mikkel2thorup@gmail.com

Abstract

Consider collections $\mathcal{A}$ and $\mathcal{B}$ of red and blue sets, respectively. Bichromatic Closest Pair is the problem of finding a pair from $\mathcal{A} \times \mathcal{B}$ that has similarity higher than a given threshold according to some similarity measure. Our focus here is the classic Jaccard similarity $|\mathbf{a} \cap \mathbf{b}|/|\mathbf{a} \cup \mathbf{b}|$ for $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$.

We consider the approximate version of the problem where we are given thresholds $j_1 > j_2$ and wish to return a pair from $\mathcal{A} \times \mathcal{B}$ that has Jaccard similarity higher than j_2 if there exists a pair in $\mathcal{A} \times \mathcal{B}$ with Jaccard similarity at least j_1. The classic locality sensitive hashing (LSH) algorithm of Indyk and Motwani (STOC '98), instantiated with the MinHash LSH function of Broder et al., solves this problem in $\tilde{O}(n^{2-\delta})$ time if $j_1 \geq j_2^{1-\delta}$. In particular, for $\delta = \Omega(1)$, the approximation ratio $j_1/j_2 = 1/j_2^{\delta}$ increases polynomially in $1/j_2$.

In this paper we give a corresponding hardness result. Assuming the Orthogonal Vectors Conjecture (OVC), we show that there cannot be a general solution that solves the Bichromatic Closest Pair problem in $O(n^{2-\Omega(1)})$ time for $j_1/j_2 = 1/j_2^{o(1)}$. Specifically, assuming OVC, we prove that for any $\delta > 0$ there exists an $\varepsilon > 0$ such that Bichromatic Closest Pair with Jaccard similarity requires time $\Omega(n^{2-\delta})$ for any choice of thresholds $j_2 < j_1 < 1 - \delta$, that satisfy $j_1 \leq j_2^{1-\varepsilon}$.

2012 ACM Subject Classification Theory of computation → Problems, reductions and completeness

Keywords and phrases fine-grained complexity, set similarity search, bichromatic closest pair, jaccard similarity

Digital Object Identifier 10.4230/LIPIcs.ESA.2019.74

Related Version A full version of the paper is available at `https://arxiv.org/abs/1907.02251`.

Funding This work was supported by Investigator Grant 16582, Basic Algorithms Research Copenhagen (BARC), from the VILLUM Foundation.
Rasmus Pagh: This research has received funding from the European Research Council under the European Union's 7th Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614331.

Acknowledgements We want to thank A. Rubinstein for helping us understand the background of his results in [9].

1 Introduction

Twitter is a well-known social network, in which a user can connect to other users by *following* them [5]. Users can read and write messages called *tweets* of up to 280 characters. An important service that Twitter provides is helping users discover other users that they might

27th Annual European Symposium on Algorithms (ESA 2019).
Editors: Michael A. Bender, Ola Svensson, and Grzegorz Herman; Article No. 74; pp. 74:1–74:13
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

like to follow, by making suggestions. This service is called the *You might also want to follow*-service and is better known as the WTF (Who To Follow) recommender system [6]. In order to suggest connections that the user might like, they should be similar to the user's existing connections. As an example, if a user is already connected to Cristiano Ronaldo, Twitter might suggest Lionel Messi as a new connection, since the connection to Ronaldo hints that the user likes famous soccer players. Hence, we need a way to decide if a connection is similar to an existing connection. We might for instance suggest a new connection if the tweets are similar to the tweets of an existing connection or if the connection has a lot of the same followers as an existing connection.

The main challenge is to find similar connections when the number of user accounts increases drastically and the task is particularly difficult when the similarity does not need to be significant, i.e., when we look for connections that have only little in common with existing ones, while they may still be of interest to the particular user [5]. This leads us to the notion of *similarity search*, which concerns the general problem of searching for similar objects in a collections of objects. Often we consider these objects as sets representing some concept or entity. An object could for example be a document that is represented by a set of words. Hence, we talk about *set similarity search.*

There are several versions of the problem addressing different situations. In this paper we consider a batched version of set similarity search, namely the Bichromatic Closest Pair which can be informally described as follows:

Suppose we are given collections $\mathcal{A}$ and $\mathcal{B}$, each of n sets from a universe of size $O(\log n)$. We refer to the sets in $\mathcal{A}$ as *red* and the sets in $\mathcal{B}$ as *blue.* Bichromatic Closest Pair is the problem of finding the pair consisting of a red and a blue set that is closest with respect to some distance or similarity measure. We will concern ourselves with Jaccard similarity, which is defined for a pair of sets $(\mathbf{a},\mathbf{b}) \in \mathcal{A} \times \mathcal{B}$ as

$$J(\mathbf{a},\mathbf{b}) = \frac{|\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|} = \frac{|\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a}| + |\mathbf{b}| - |\mathbf{a} \cap \mathbf{b}|}. \tag{1}$$

In particular, we consider the following *decision version* of Bichromatic Closest Pair with Jaccard similarity: decide whether there exists a pair $(\mathbf{a},\mathbf{b}) \in \mathcal{A} \times \mathcal{B}$ such that $J(\mathbf{a},\mathbf{b}) \geq j_1$ or if all pairs $(\mathbf{a},\mathbf{b}) \in \mathcal{A} \times \mathcal{B}$, has $J(\mathbf{a},\mathbf{b}) < j_2$ for given thresholds j_1 and j_2.

It is well-known that we can solve Bichromatic Closest Pair with Jaccard similarity for thresholds satisfying $j_1 \geq j_2^{1-\delta}$ in time $O(n^{2-\delta})$ (see Section 1.1). In particular, for $\delta = \Omega(1)$, the approximation ratio $j_1/j_2 = 1/j_2^{\delta}$ increases polynomially in $1/j_2$. In this paper, we will present a corresponding hardness result. The hardness is conditioned on one of the most well-known and widely believed hypotheses, namely the Orthogonal Vectors Conjecture [11].

▶ **Conjecture 1** (Orthogonal Vectors Conjecture (OVC)). *For every $\delta > 0$ there exists $c = c(\delta)$ such that given two collections $\mathcal{A},\mathcal{B} \subset \{0,1\}^m$ of cardinality n, where $m = c\log n$, deciding if there is a pair $(\mathbf{a},\mathbf{b}) \in \mathcal{A} \times \mathcal{B}$ such that $\mathbf{a} \cdot \mathbf{b} = 0$ requires time $\Omega(n^{2-\delta})$.*

Assuming OVC, we show that there cannot be a general solution that solves the Bichromatic Closest Pair problem with Jaccard similarity in $O(n^{2-\Omega(1)})$ time for $j_1/j_2 = 1/j_2^{o(1)}$. More specifically, we show

▶ **Theorem 2.** *Assuming the Orthogonal Vectors Conjecture (OVC), the following holds: for any $\delta > 0$, there exists an $\varepsilon > 0$ such that for any given $j_2 < j_1 < 1 - \delta$ satisfying $j_1 \leq j_2^{1-\varepsilon}$, solving Bichromatic Closest Pair with Jaccard similarity for n red and n blue sets for sets from a universe of size $\ln(n)/j_2^{O(\log 1/j_1)}$ for thresholds j_1 and j_2 requires time $\Omega(n^{2-\delta})$.*

The dependence of ε on δ is unspecified because the function $c(\delta)$ in OVC is not specified, see discussion in Appendix B in the full version on ArXiv [8, App. B].

1.1 Techniques and Related Work

Similarity search can be performed in several ways – a popular technique is Locality Sensitive Hashing (LSH) [7] which attempts to collect similar items in buckets in order to reduce the number of sets needed to check similarity against. We can for example use Broder's MinHash [1] with locality sensitive hashing to solve Bichromatic Closest Pair with Jaccard similarity in time $\tilde{O}(n^{2-\varepsilon})$ when $j_1 \geq j_2^{1-\varepsilon}$ for any ε. This is done by ensuring that the collision probability for pairs with similarity j_2 is $1/n$ and the collision probability for pairs with similarity j_1 is $1/n^{1-\varepsilon}$. Hashing $n^{1-\varepsilon}$ times means that we find a pair with similarity j_1 if one exists. The ChosenPath method presented in [4] also uses the LSH framework to solve Bichromatic Closest Pair with Braun-Blanquet similarity in time $\tilde{O}(n^{2-\varepsilon})$ for thresholds $j_1 \geq j_2^{1-\varepsilon}$.

The proof of Theorem 2 will be based on a result by Rubinstein [9]: Assuming the Orthogonal Vectors Conjecture, a $(1+\varepsilon)$-approximation to Bichromatic Closest Pair with Hamming, Edit or Euclidean distance requires time $\Omega(n^{2-\delta})$. The required approximation factor $1+\varepsilon$ depends on δ, and tends to 1 as δ tends to zero. We translate this into an equivalent conditional lower bound for Jaccard similarity for certain constants j_1 and j_2.

In order to handle smaller subconstant values of j_1 and j_2 we use a technique that we call squaring, which allows us to increase the gap in similarities between pairs with high Jaccard similarity and pairs with low Jaccard similarity by computing the cartesian product of a binary vector with itself. A similar technique is used in [10] by Valiant. His technique is called *tensoring* and is used to amplify the gap between small and large inner products of vectors. We also see a similar technique in the LSH framework with MinHash, where we use concatenation of hash values (which are sampled set elements) to amplify the difference in collision probability, and hence in the Jaccard similarity.

Combining two simple reductions with the above squaring we show that for any δ, we can always find ε such that Bichromatic Closest Pair with Jaccard similarity cannot be solved in time $O(n^{2-\delta})$ for any pair $j_1, j_2 < 1-\delta$ when $j_1 \leq j_2^{1-\varepsilon}$. Contrast this with the above LSH upper bound of $\tilde{O}\left(n^{2-\delta}\right)$ for $j_1 \geq j_2^{1-\delta}$. We also know that there are parts of the parameter space where $j_1 = j_2^{1-\delta}$ that can be solved in $\tilde{O}\left(n^{2-\delta-\Omega(1)}\right)$ time, see the discussion in [4]. While LSH with MinHash is not the fastest possible algorithm in terms of the exponent achieved, it has been unclear how far from optimal it might be.

Other related work

Very recently, Chen and Williams [3] showed that assuming the OVC we cannot additively approximate our Bichromatic Closest Pair problem with Jaccard similarity. It might be possible to use Chen and Williams as a base for showing our main theorem, but this would require reductions quite different from the ones presented in this paper.

An earlier of result of Chen [2] shows that it is not possible (under OVC) to compute a $(d/\log n)^{o(1)}$-approximation to Maximum Inner Product (Max-IP) with two sets of n vectors from $\{0,1\}^d$ in time $O(n^{2-\Omega(1)})$.

2 Preliminaries

2.1 Notation

We will occasionally consider a set, $\mathbf{x}$, from a finite universe $U = \{u_1, ..., u_{|U|}\}$ as a vector $\mathbf{v}$ of dimension $|U|$ such that $v_i = [u_i \in x]$, in Iverson notation. We call this vector the *characteristic vector for* $\mathbf{x}$. Hence, we refer to the set of indexes and the universe interchangeably. We denote the Hamming weight of a binary vector $\mathbf{v}$ by $|\mathbf{v}|$. In the following, we will not only index vectors with integers, but also with vectors of integers. Hence, we will consider vectors of dimension d^2 with entries v_{ij}, for $i = (i_1, ..., i_d)$ and $j = (j_1, ..., j_d)$.

2.2 Bichromatic Closest Pair

Recall Jaccard similarity as is defined in (1). We define Bichromatic Closest Pair with Jaccard similarity for thresholds t_1 and t_2 as follows: Let U be a universe of size $O(\log n)$. Given collections $\mathcal{A}$ and $\mathcal{B}$, each of n sets from U, and thresholds $t_2 < t_1 < 1$, we will consider the problem of finding a pair of sets $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$ with $J(\mathbf{a}, \mathbf{b}) \geq t_2$ if there exists a pair $(\mathbf{a}^*, \mathbf{b}^*) \in \mathcal{A} \times \mathcal{B}$ with $J(\mathbf{a}^*, \mathbf{b}^*) \geq t_1$. If all pairs have $J(\mathbf{a}, \mathbf{b}) < t_2$, we must not return any pair of sets.

2.3 Useful instances of Bichromatic Closest Pair

The following lemma corresponds to Theorem 4.1 in [9] and will form the basis of our results. It includes the important properties of the instances constructed in the proof the theorem, which we will use actively to prove our own Theorem 2.

▶ **Lemma 3.** *Assume OVC. Given $\delta > 0$, there exist $\varepsilon > 0$ and values h_1, h_2 where $h_2 = (1+\varepsilon)h_1$ such that Bichromatic Closest Pair with Hamming distance for thresholds h_1 and h_2 requires time $\Omega(n^{2-\delta})$ for instances with n red and n blue sets from a universe of size $O(\log n)$. There are instances that require this time with the following properties, where we let $T = O\left(\frac{1}{\varepsilon}\right)$ and $m = O(\log n)$:*

- *All red sets have size Tm and all blue sets have size m.*
- *The thresholds h_1 and h_2 are $m(T-1)$ and mT, respectively.*
- *All sets in the instance come from a universe of size $2Tm$.*

In particular, the lemma states that we cannot compute a $(1+\varepsilon)$-approximation to Bichromatic Closest Pair with Hamming distance in truly subquadratic time. We will extend this result in a few steps, using the properties of the hard instances, to achieve Theorem 2.

2.4 Hardness of Bichromatic Closest Pair with Jaccard similarity

In order to prove Theorem 2, we need the following lemma, which extends Lemma 3 in the natural way to Jaccard similarity.

▶ **Lemma 4.** *Assuming OVC, we have the following: For any $\delta > 0$ there exist j_1, j_2 with $j_1 = 2 \cdot j_2$ such that Bichromatic Closest Pair with Jaccard similarity with thresholds j_1 and j_2 requires time $\Omega(n^{2-\delta})$.*

Proof. We use instances as described in Lemma 3. First, note that

$$J(\mathbf{a}, \mathbf{b}) = \frac{|\mathbf{a} \cap \mathbf{b}|}{|\mathbf{a} \cup \mathbf{b}|} = \frac{\frac{|\mathbf{a}|+|\mathbf{b}|-d_H(\mathbf{a},\mathbf{b})}{2}}{|\mathbf{a}| + |\mathbf{b}| - \frac{|\mathbf{a}|+|\mathbf{b}|-d_H(\mathbf{a},\mathbf{b})}{2}} = \frac{|\mathbf{a}| + |\mathbf{b}| - d_H(\mathbf{a}, \mathbf{b})}{|\mathbf{a}| + |\mathbf{b}| + d_H(\mathbf{a}, \mathbf{b})}$$

which implies that letting

$$j_1 = \frac{Tm + m - m(T-1)}{Tm + m + m(T-1)} = \frac{1}{T} \qquad \text{and} \qquad j_2 = \frac{Tm + m - Tm}{Tm + m + Tm} = \frac{1}{2T+1},$$

we cannot solve Bichromatic Closest Pair with Jaccard similarity in time $O(n^{2-\delta})$. Since $T = O\left(\frac{1}{\varepsilon}\right)$, as mentioned in Lemma 3, we get a lower bound for the approximation factor:

$$\frac{\frac{1}{T}}{\frac{1}{2T+1}} = \frac{2T+1}{T} = 2 + \frac{1}{T} = 2 + \Omega(\varepsilon).$$

In particular, we achieve hardness of a 2-approximation. ◀

3 Overview of reductions used

We prove Theorem 2 by combining several reductions into one. So let $(\mathcal{A}, \mathcal{B})$ be any instance of Bichromatic Closest Pair with Jaccard similarity as described in Lemma 3. We give a brief introduction to each of these reductions – note that all reductions are self-reductions. We give the details of the proof and the use of each reduction in Section 5. Further details can be found in Appendix B in the full version on ArXiv [8, App. B].

- **Adding common elements to sets:** Adding common elements to all sets in collections $\mathcal{A}$ and $\mathcal{B}$ increases the Jaccard similarity between any pair of red and blue sets.
- **Adding different elements to sets:** Adding elements to all sets in $\mathcal{A}$ decreases the Jaccard similarity between any pair of red and blue sets.
- **Squaring:** Consider all sets by their characteristic vector. We define squaring as follows: given vector $\mathbf{a} = (a_1, ..., a_d)$ the squared vector has entries

  $$a'_{ij} = a_i \cdot a_j \qquad \text{for } i, j \in \{1, ..., d\}.$$

  The resulting vector $\mathbf{a}'$, which is the characteristic vector for $\mathbf{a} \times \mathbf{a}$, has dimension d^2 as described in Section 2.1. Vector $\mathbf{a}'$ can equivalently be considered as a set from a universe of size d^2. We will use this reduction iteratively to reduce the Jaccard similarity between any pair of vectors in the instance of Bichromatic Closest Pair.
- **Sampling:** We will use sampling to reduce the size of the universe after each step of squaring. Hence, we consider squaring and sampling as a single reduction which first squares the vectors and then samples from the resulting vectors. We will use the squaring-and-sampling reduction iteratively.

4 The squaring-and-sampling reduction – details

In the proof of Theorem 2 we will take any instance of Bichromatic Closest Pair with Jaccard similarity with the properties described in Lemma 3 and use the squaring reduction described in Section 3 to decrease the Jaccard similarity of every pair of sets in the instance. We will argue that a solution for the new instance also provides a solution for the original instance. When squaring all sets, the Jaccard similarity between any pair of sets will decrease, so we need to capture this change in the thresholds, such that a solution for the new instance implies a solution for the initial instance. When squaring the sets in $\mathcal{A}$ and $\mathcal{B}$, the size of the sets will be squared and it is easy to see that so will the size of the intersection. Hence, the Jaccard similarity of a pair $(\mathbf{a}, \mathbf{b})$ after squaring i times, $(\mathbf{a}_i, \mathbf{b}_i)$ is

$$J(\mathbf{a}_i, \mathbf{b}_i) = \frac{|\mathbf{a} \cap \mathbf{b}|^{2^i}}{|\mathbf{a}|^{2^i} + |\mathbf{b}|^{2^i} - |\mathbf{a} \cap \mathbf{b}|^{2^i}}. \tag{2}$$

In order to keep down the size of the universe, we need to sample after each step of squaring. This might incur a small error in the Jaccard similarity. The next few sections will bound this error. From this point, we will denote the squaring-and-sampling reduction by f. Hence, applying the reduction f to a set, $\mathbf{v}$, i times will yield a set $\mathbf{f}(\mathbf{v}, \mathbf{i})$.

4.1 Subsampling

We bound the error incurred in each of $|\mathbf{a} \cap \mathbf{b}|$, $|\mathbf{a}|$ and $|\mathbf{b}|$ and combine these with a union bound to get a bound on the error in the Jaccard similarity. We shall see that when sampling sufficiently many elements from the universe the sets are taken from, we get that with high probability a solution for the constructed instance will provide a valid solution for the original instance.

The following lemmas will help us show that sampling after squaring will not distort the similarity of the resulting vectors too much.

▶ **Lemma 5.** *Let* $0 < m' < m < 1$ *and let* $\mathbf{p}$ *be a set from a universe of size* s^2 *for an integer* s*. Assume that* $(m' \cdot s)^2 \le |\mathbf{p}| \le (m \cdot s)^2$*. Sample* s' *elements from the universe uniformly at random,* $\mathbf{z}$*, thus generating sample set* $\mathbf{p} \cap \mathbf{z}$*. We have*

$$(1-\gamma) \cdot m'^2 \cdot s' \le |\mathbf{p} \cap \mathbf{z}| \le (1+\gamma) \cdot m^2 \cdot s'$$

with probability at least $1 - 2n^{-10}$ *when sampling* $s' \ge \frac{30\ln(n)}{\gamma^2 m'^2}$ *elements.*

Proof. The result is an immediate consequence of the Chernoff bound: when we sample $s' \ge \frac{20\ln(n)}{\gamma^2 m'^2}$ elements, we have with probability at least $1 - n^{-10}$ that

$$(1-\gamma)\,(m' \cdot s)^2 \ \cdot\ \frac{s'}{s^2} \le\ |\mathbf{p}\ \cap\ \mathbf{z}|.$$

A similar result gives the upper bound on $|\mathbf{p} \cap \mathbf{z}|$ for $s' \ge \frac{30\ln(n)}{\gamma^2 m^2}$. As $m' \le m$, we maximize s' by $\frac{30\ln(n)}{\gamma^2 m'^2}$ and thus ensure both bounds with probability at least $1 - 2n^{-10}$ using a union bound. ◀

We are generally going to use γ as the same fixed parameter (to be determined later) every time we invoke the sampling of Lemma 5.

Lemma 5 will be used to show that sampling after squaring will not distort the Jaccard similarity of a pair of vectors too much, and hence we get the benefits of squaring without the exploding vector dimensions. We start by bounding the resulting sizes for each of $|\mathbf{a}|, |\mathbf{b}|$ and $|\mathbf{a} \cap \mathbf{b}|$ for any choice of $\mathbf{a}, \mathbf{b} \in \mathcal{A} \times \mathcal{B}$ from squaring and sampling i times.

▶ **Lemma 6.** *Let* $\mathbf{v}$ *be a set from a universe of size* d *or the intersection of such two sets. Let* $\mathbf{f}(\mathbf{v}, \mathbf{i})$ *denote the resulting set after running* i *iterations of the squaring-and-sampling reduction on set* $\mathbf{v}$ *for* $i \ge 1$*. We have*

$$(1-\gamma)^{2^i} \frac{|\mathbf{v}|^{2^i}}{d^{2^i}} s_i \le |\mathbf{f}(\mathbf{v}, \mathbf{i})| \le (1+\gamma)^{2^i} \frac{|\mathbf{v}|^{2^i}}{d^{2^i}} s_i$$

with probability at least $1 - 2in^{-10}$ *where* $s_i \ge \frac{30\ln(n) d^{2^i}}{\gamma^2 (1-\gamma)^{2^i-2} |\mathbf{v}|^{2^i}}$*.*

Proof. Let $\mathbf{v}$ be as described. We show the lemma by induction on i. Clearly, when squaring the vector $\mathbf{v}$ once, i.e., for $i = 1$, the resulting vector has Hamming weight $|\mathbf{v}|^2$ and dimension d^2. Hence, by Lemma 5 we have

$$(1-\gamma)\frac{|\mathbf{v}|^2}{d^2} \cdot s_1 \le |\mathbf{f}(\mathbf{v}, \mathbf{1})| \le (1+\gamma)\frac{|\mathbf{v}|^2}{d^2} \cdot s_1$$

with probability at least $1 - 2n^{-10}$ for our choice of s_1. Assume now that after $i-1$ iterations the following bounds hold:

$$(1-\gamma)^{2^{i-1}-1} \frac{|\mathbf{v}|^{2^{i-1}}}{d^{2^{i-1}}} s_{i-1} \le |\mathbf{f}(\mathbf{v}, \mathbf{i-1})| \le (1+\gamma)^{2^{i-1}-1} \frac{|\mathbf{v}|^{2^{i-1}}}{d^{2^{i-1}}} s_{i-1}. \tag{3}$$

Then Lemma 5 gives that after i iterations of the squaring-and-sampling reduction, we have

$$(1-\gamma)^{2^i-1} \frac{|\mathbf{v}|^{2^i} s_{i-1}^2}{d^{2^i}} \cdot \frac{s_i}{s_{i-1}^2} \le |\mathbf{f}(\mathbf{v}, \mathbf{i})| \le (1+\gamma)^{2^i-1} \frac{|\mathbf{v}|^{2^i} s_{i-1}^2}{d^{2^i}} \cdot \frac{s_i}{s_{i-1}^2}$$

with probability at least $1-2n^{-10}$ for $s_i \geq \frac{30\ln(n)d^{2^i}}{\gamma^2(1-\gamma)^{2^i-2}|\mathbf{v}|^{2^i}}$. This particularly means that

$$(1-\gamma)^{2^i}\frac{|\mathbf{v}|^{2^i}}{d^{2^i}}\cdot s_i \leq |\mathbf{f}(\mathbf{v},\mathbf{i})| \leq (1+\gamma)^{2^i}\frac{|\mathbf{v}|^{2^i}}{d^{2^i}}\cdot s_i.$$

Now, to ensure these bounds, we assumed that $|\mathbf{f}(\mathbf{v},\mathbf{i-1})|$ satisfies certain bounds (see (3)). So in order to ensure that $\mathbf{f}(\mathbf{v},\mathbf{i})$ satisfies the given bounds, we need $\mathbf{f}(\mathbf{v},\mathbf{j})$ to satisfy similar bounds for every $1 \leq j \leq i$. By a union bound, we see that $|\mathbf{f}(\mathbf{v},\mathbf{j})|$ satisfies both upper and lower bounds for all $1 \leq j \leq i$ (simultaneously) with probability at least $1-2in^{-10}$ when sampling $s_j \geq \frac{30\ln(n)d^{2^j}}{\gamma^2(1-\gamma)^{2^j-2}|\mathbf{v}|^{2^j}}$ at step j. Hence, $|\mathbf{f}(\mathbf{v},\mathbf{i})|$ satisfies the given bound with probability at least $1-2in^{-10}$. ◀

The next section will use Lemma 6 to bound the Jaccard similarity after i iterations of the squaring/sampling reduction.

4.2 Combining the bounds

For a given pair of vectors $\mathbf{a}$ and $\mathbf{b}$, Lemma 6 gives upper and lower bounds on the Jaccard similarity $J = J\Big(\mathbf{f}(\mathbf{a},\mathbf{i}),\mathbf{f}(\mathbf{b},\mathbf{i})\Big)$. We claim that with probability at least $1-6in^{-10}$:

$$\begin{aligned} J &\geq \frac{(1-\gamma)^{2^i-1}\frac{|\mathbf{a}\cap\mathbf{b}|^{2^i}}{d^{2^i}}s_i}{(1+\gamma)^{2^i-1}\frac{|\mathbf{a}|^{2^i}}{d^{2^i}}s_i+(1+\gamma)^{2^i-1}\frac{|\mathbf{b}|^{2^i}}{d^{2^i}}s_i-(1-\gamma)^{2^i-1}\frac{|\mathbf{a}\cap\mathbf{b}|^{2^i}}{d^{2^i}}s_i} \\ &\geq \frac{(1-\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1+\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}\right)-(1-\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}\end{aligned}$$

$$\begin{aligned} J &\leq \frac{(1+\gamma)^{2^i-1}\frac{|\mathbf{a}\cap\mathbf{b}|^{2^i}}{d^{2^i}}s_i}{(1-\gamma)^{2^i-1}\frac{|\mathbf{a}|^{2^i}}{d^{2^i}}s_i+(1-\gamma)^{2^i-1}\frac{|\mathbf{b}|^{2^i}}{d^{2^i}}s_i-(1+\gamma)^{2^i-1}\frac{|\mathbf{a}\cap\mathbf{b}|^{2^i}}{d^{2^i}}s_i} \\ &\leq \frac{(1+\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1-\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}\right)-(1+\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}.\end{aligned}$$

This is easily seen by taking a union bound over the probabilities that each of $|\mathbf{a}|$, $|\mathbf{b}|$ and $|\mathbf{a}\cap\mathbf{b}|$ violate either the upper or the lower bound. Next, we claim that these bounds imply:

$$J \geq \frac{(1-\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1+4\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}-|\mathbf{a}\cap\mathbf{b}|^{2^i}\right)} \geq \frac{(1-\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1+\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}\right)-(1-\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}$$

$$J \leq \frac{(1+\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1-\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}\right)-(1+\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}} \leq \frac{(1+\gamma)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{(1-4\gamma)^{2^i}\left(|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}-|\mathbf{a}\cap\mathbf{b}|^{2^i}\right)}.$$

The argument can be found in Appendix A in the full version on ArXiv [8, App. A]. In particular, we have argued for the following lemma. We ignore the sample size for now and discuss it in Section 4.3.

▶ **Lemma 7.** *Let $\mathcal{A}$ and $\mathcal{B}$ be an instance of Bichromatic Closest Pair with Jaccard similarity. After applying the Squaring and Sampling mapping, f, i times as previously described to each set in $\mathcal{A}$ and $\mathcal{B}$, we have for all n^2 pairs $(\mathbf{a},\mathbf{b}) \in \mathcal{A}\times\mathcal{B}$ in the instance that:*

$$\frac{\left(\frac{1-\gamma}{1+4\gamma}\right)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}-|\mathbf{a}\cap\mathbf{b}|^{2^i}} \le J\Big(\mathbf{f}(\mathbf{a},\mathbf{i}),\mathbf{f}(\mathbf{b},\mathbf{i})\Big) \le \frac{\left(\frac{1+\gamma}{1-4\gamma}\right)^{2^i}|\mathbf{a}\cap\mathbf{b}|^{2^i}}{|\mathbf{a}|^{2^i}+|\mathbf{b}|^{2^i}-|\mathbf{a}\cap\mathbf{b}|^{2^i}}$$

with probability at least $1-6in^{-8}$*.*

Hence, with high probability none of the Jaccard similarities diverge too much from (2) due to sampling. This was exactly what we wanted, as this allows us to reduce the dimension by sampling.

4.3 Summing up

Recall that in our setting we reduce from instances where the set sizes of all red and blue sets are fixed. We now describe thresholds such that solving the instances constructed by the reduction f cannot be done in truly subquadratic time.

▶ **Lemma 8.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two collections of n sets from a universe of dimension d, where all sets in $\mathcal{A}$ have size y and all sets in $\mathcal{B}$ have size z. Assume that $(\mathcal{A},\mathcal{B})$ is taken from a family of instances of Bichromatic Closest Pair with Jaccard similarity, which require time $\Omega(n^{2-\delta})$ for thresholds $t_1=\frac{x_1}{y+z-x_1}$ and $t_2=\frac{x_2}{y+z-x_2}$. The reduction which applies f i times to each set in $\mathbf{s}\in\mathcal{A}\cup\mathcal{B}$ for $i\ge 1$ constructs an instance of Bichromatic Closest Pair with Jaccard similarity, which requires time $\Omega(n^{2-\delta})$ time for thresholds*

$$t_1'=\left(\frac{1-\gamma}{1+4\gamma}\right)^{2^i}\frac{x_1^{2^i}}{y^{2^i}+z^{2^i}-x_1^{2^i}}, \qquad \text{and} \qquad t_2'=\left(\frac{1+\gamma}{1-4\gamma}\right)^{2^i}\frac{x_2^{2^i}}{y^{2^i}+z^{2^i}-x_2^{2^i}}.$$

whose solution provides a valid solution to the original instance with high probability when sampling $s_j>\frac{30\ln(n)d^{2^j}}{\gamma^2(1-\gamma)^{2^j-2}x_2^{2^j}}$ at each step $1\le j\le i$.

Proof. Lemma 7 ensures that with high probability a solution to the constructed instance provides a valid solution to the original instance, since no pair of sets is likely to have Jaccard similarities that deviate beyond the chosen thresholds.

In Lemma 7 we skipped the discussion of the sample size at each iteration – we will argue for it now. From Lemma 6, it is easily seen that we maximize the needed sample size for all of $|\mathbf{a}|$, $|\mathbf{b}|$ or $|\mathbf{a}\cap\mathbf{b}|$ for any choice of $\mathbf{a}$ and $\mathbf{b}$ in iteration i by

$$s_i>\frac{30\ln(n)d^{2^i}}{\gamma^2(1-\gamma)^{2^i-2}\min_{(\mathbf{a},\mathbf{b})\in\mathcal{A}\times\mathcal{B}}\{|\mathbf{a}\cap\mathbf{b}|\}^{2^i}}.$$

Hence, sampling s_i elements from the universe will ensure that each of the upper and lower bounds for either $|\mathbf{a}|$, $|\mathbf{b}|$ or $|\mathbf{a}\cap\mathbf{b}|$ will fail with probability at most n^{-10} in that iteration. As $\min_{(\mathbf{a},\mathbf{b})\in\mathcal{A}\times\mathcal{B}}\{|\mathbf{a}\cap\mathbf{b}|\}$ is unknown, we instead use x_2, which was the intersection size for a pair with Jaccard similarity j_2. Such a pair need not exist, but as the set sizes are fixed, x_2 can be easily computed.

We have left to argue that the pairs with intersection smaller than x_2 also satisfy the bounds in Lemma 7 with high probability. The main observation is that they only need to satisfy the upper bound, as the resulting Jaccard similarities need only to stay below the lower threshold, t_2' – the Jaccard similarities can become arbitrarily small without affecting the result.

By bounding the size of each term as we did in Lemma 6 using the chosen s_i, we see that the error probabilities are still at most n^{-10} for each of $|\mathbf{a}|$, $|\mathbf{b}|$ and $|\mathbf{a}\cap\mathbf{b}|$ for any choice of $(\mathbf{a},\mathbf{b})\in\mathcal{A}\times\mathcal{B}$. ◀

5 Main Result

We are now ready to prove Theorem 2. We first give some intuition behind the proof and state a few lemmas to ease the proof. For convenience we restate Theorem 2.

▶ **Theorem 9.** *Assuming the Orthogonal Vectors Conjecture (OVC), the following holds: for any $\delta > 0$, there exists an $\varepsilon > 0$ such that for any given $j_2 < j_1 < 1 - \delta$ satisfying $j_1 \leq j_2^{1-\varepsilon}$, solving Bichromatic Closest Pair with Jaccard similarity for n red and n blue sets for sets from a universe of size $\ln(n)/j_2^{O(\log(1/j_1))}$ for thresholds j_1 and j_2 requires time $\Omega(n^{2-\delta})$.*

5.1 Intuition

The proof of Theorem 2 reduces instances of Bichromatic Closest Pair as described in Section 2.3 by composing three reductions, that together construct instances of Bichromatic Closest Pair with Jaccard similarity, which requires time $\Omega(n^{2-\delta})$ for the given thresholds j_1 and j_2 and some ε. A short description of each of the reductions can be found in Section 3. Below, we give three lemmas showing that these reductions preserve hardness.

The first lemma states that adding common elements to all sets in the instance will preserve hardness. This reduction increases the Jaccard similarity of all pairs of red and blue sets, and by choice of the number of added elements, we ensure that pairs of sets that initially had Jaccard similarity higher than the *lower* threshold will get Jaccard similarity greater than $1 - \delta$. Hence, we get hardness for thresholds that are greater than $1 - \delta$. From this point we can decrease the thresholds using two other reductions to achieve the given thresholds, that by assumption are less than $1 - \delta$.

The second lemma states that the squaring-and-sampling reduction, discussed in detail in Section 4, preserves hardness. The squaring-and-sampling reduction allows us to decrease the thresholds, so they come close to j_1 and j_2. Finally, the third lemma states that the reduction, which adds elements to only red sets will still preserve hardness. This reduction ensures that we can decrease the Jaccard similarity further. We will use it in such a way, that we effectively multiply the upper bound by a well-chosen α that ensures that the upper threshold is j_1 after this reduction. The proof ends by picking an ε, such that j_2 is strictly greater than the current lower threshold, and thus preserves hardness for the thresholds j_1 and j_2.

5.2 Supporting Lemmas

In the following, assume that $\mathcal{A}$ and $\mathcal{B}$ are collections of n red and n blue sets from a universe U, respectively.

▶ **Lemma 9.** *Let $0 < \delta \leq 1$ be given and let $(\mathcal{A}, \mathcal{B})$ be any instance of Bichromatic Closest Pair with Jaccard similarity as described in Lemma 3. Define $\ell := \max_{\mathbf{q} \in \mathcal{A} \cup \mathcal{B}}\{|\mathbf{q}|\} \cdot (1/\delta - 1)$ and $\mathbf{x} := \{x_1, ..., x_\ell\}$ such that $\mathbf{x} \cap (\mathcal{A} \cup \mathcal{B}) = \emptyset$, and further define the mapping $g : \mathcal{A} \cup \mathcal{B} \to \mathcal{A}' \cup \mathcal{B}'$ by $g(\mathbf{v}) = \mathbf{v} \cup \mathbf{x}$ where $\mathcal{A}' = \mathcal{A} \cup \mathbf{x}$ and equivalently $\mathcal{B}' = \mathcal{B} \cup \mathbf{x}$. The reduction that applies g to every element of $\mathcal{A}$ and $\mathcal{B}$ generates an instance $(\mathcal{A}', \mathcal{B}')$ of Bichromatic Closest Pair with Jaccard similarity that requires time $\Omega(n^{2-\delta})$ for some thresholds $t_1', t_2' \geq 1 - \delta$.*

Proof. First, note that if $\mathbf{v} \in \mathcal{A}$, then $g(\mathbf{v}) \in \mathcal{A}'$ and similarly if $\mathbf{v} \in \mathcal{B}$ then $g(\mathbf{v}) \in \mathcal{B}'$. We recall that instances of Bichromatic Closest Pair as described in Lemma 3 are constructed such that all red sets have the same size and all blue sets have the same size. We also have $\max_{\mathbf{q} \in \mathcal{A} \cup \mathcal{B}}\{|\mathbf{q}|\} = |\mathbf{a}|$, for any $\mathbf{a} \in \mathcal{A}$, since the sets in $\mathcal{A}$ were larger than the sets in $\mathcal{B}$. It is easy to see that hardness is preserved under the reduction.

We finally argue that the resulting thresholds are larger than $1-\delta$: Let $(\mathbf{a},\mathbf{b})$ be any pair from $\mathcal{A}\times\mathcal{B}$ which has Jaccard similarity at least t_2 and let $\mathbf{a}'=g(\mathbf{a})$ and $\mathbf{b}'=g(\mathbf{b})$. We argue that any such pair satisfies $|\mathbf{a}\cap\mathbf{b}|\geq\frac{|\mathbf{b}|}{2}$: Note that with these particular instances of Bichromatic Closest Pair and from the proof of Lemma 4, we have

$$J(\mathbf{a},\mathbf{b})=\frac{|\mathbf{a}\cap\mathbf{b}|}{|\mathbf{a}\cup\mathbf{b}|}=\frac{|\mathbf{a}\cap\mathbf{b}|}{Tm+m-|\mathbf{a}\cap\mathbf{b}|}\geq t_2=\frac{t_1}{2}=\frac{1/T}{2}.$$

Since $|\mathbf{b}|=m\geq|\mathbf{a}\cap\mathbf{b}|$, this implies

$$|\mathbf{a}\cap\mathbf{b}|\geq\frac{m}{2}+\frac{m}{2T}-\frac{|\mathbf{a}\cap\mathbf{b}|}{2T}\quad\Rightarrow\quad|\mathbf{a}\cap\mathbf{b}|\geq m/2=|\mathbf{b}|/2.$$

We will consider the Jaccard similarity of $\mathbf{a}'$ and $\mathbf{b}'$:

$$\begin{aligned}J(\mathbf{a}',\mathbf{b}')&=\frac{|\mathbf{a}\cap\mathbf{b}|+|\mathbf{a}|(1/\delta-1)}{(|\mathbf{a}|+|\mathbf{a}|(1/\delta-1))+(|\mathbf{b}|+|\mathbf{a}|(1/\delta-1))-(|\mathbf{a}\cap\mathbf{b}|+|\mathbf{a}|(1/\delta-1))}\\&=\frac{|\mathbf{a}\cap\mathbf{b}|+|\mathbf{a}|(1/\delta-1)}{|\mathbf{a}|/\delta+|\mathbf{b}|-|\mathbf{a}\cap\mathbf{b}|}.\end{aligned}$$

By assumption $|\mathbf{a}\cap\mathbf{b}|\geq\frac{|\mathbf{b}|}{2}$, so:

$$\begin{aligned}\frac{|\mathbf{a}\cap\mathbf{b}|+|\mathbf{a}|(1/\delta-1)}{|\mathbf{a}|/\delta+|\mathbf{b}|-|\mathbf{a}\cap\mathbf{b}|}&\geq\frac{|\mathbf{b}|/2+|\mathbf{a}|(1/\delta-1)}{|\mathbf{a}|/\delta+|\mathbf{b}|/2}\geq1-\delta\\\Leftrightarrow\quad&\frac{|\mathbf{b}|}{2}+|\mathbf{a}|(1/\delta-1)\geq|\mathbf{a}|(1/\delta-1)+\frac{|\mathbf{b}|}{2}-\frac{|\mathbf{b}|\delta}{2}\end{aligned}$$

which is always satisfied. Hence, $J(\mathbf{a}',\mathbf{b}')\geq1-\delta$ for any choice of $\delta>0$, and so, we construct an instance where every pair with Jaccard similarity higher than t_2 will have Jaccard similarity higher than $1-\delta$. Thus, there are thresholds that are greater than $1-\delta$, that make the constructed instance hard. ◀

▶ **Lemma 10.** *Let $0<\delta\leq1$ be given and consider any instance of Bichromatic Closest Pair with Jaccard similarity, $(\mathcal{A},\mathcal{B})$, from a family of instances which require time $\Omega(n^{2-\delta})$ for thresholds t_1 and t_2. Using the reduction f defined in Section 4 on each $\mathbf{v}\in\mathcal{A}\cup\mathcal{B}$ for i iterations where $i\geq1$, we construct a valid instance of Bichromatic Closest Pair with Jaccard similarity with high probability, which requires time $\Omega(n^{2-\delta})$ for thresholds that are decreasing functions of i.*

Proof. The lemma follows immediately from Lemma 8. ◀

▶ **Lemma 11.** *Let $0<\delta\leq1$ be given and consider any instance of Bichromatic Closest Pair with Jaccard similarity, $(\mathcal{A},\mathcal{B})$, from a family of instances which require time $\Omega(n^{2-\delta})$ for thresholds t_1 and t_2. Define $\ell:=\max_{\mathbf{q}\in\mathcal{A}\cup\mathcal{B}}\{|\mathbf{q}|\}\cdot(1/\alpha-1)$ and $\mathbf{y}:=\{y_1,...,y_\ell\}$ such that $\mathbf{y}\cap(\mathcal{A}\cup\mathcal{B})=\emptyset$. Define mapping $h:\mathcal{A}\to\mathcal{A}'$ where $\mathcal{A}'=\mathcal{A}\cup Y$ by $h(\mathbf{a})=\mathbf{a}\cup\mathbf{y}$. The reduction that applies h to every element of $\mathcal{A}$ generates an instance $(\mathcal{A}',\mathcal{B})$ of Bichromatic Closest Pair with Jaccard similarity that requires time $\Omega(n^{2-\delta})$ for some thresholds t_1',t_2'.*

Proof. Clearly, hardness is preserved under the reduction that simply adds new elements to all red sets. In particular this reduction decreases the thresholds by decreasing the similarity between red and blue pairs. ◀

5.3 Proof outline for Theorem 2

Proof. For simplicity and readability we leave out most of the calculations – details can be found in Appendix B in the full version on ArXiv [8, App. B].

Let $\delta > 0$ be given and let j_1, j_2 be given such that $j_2 < j_1 < 1 - \delta$. Take any instance of Bichromatic Closest Pair with Jaccard similarity satisfying the properties described in Lemma 3. Recall from this lemma that $T = O\left(\frac{1}{\varepsilon}\right)$.

Apply the reductions from first Lemma 9 to achieve an instance, which requires time $\Omega(n^{2-\delta})$ for thresholds greater than $1 - \delta$. We wish to reduce to an instance that is hard for smaller thresholds j_1 and j_2. The reduction from Lemma 10 is used to decrease the thresholds, where we pick the largest i, such that the resulting upper threshold t_1 is no smaller than j_1, i.e., $t_1 \geq j_1$. This reduction decreases the thresholds until the upper threshold is only slightly greater than j_1. Now, let $\alpha = \frac{j_1}{t_1}$ and apply the reduction from Lemma 11 to ensure that the resulting upper threshold is now equal to j_1. This eventually gives an instance of Bichromatic Closest Pair with Jaccard similarity, which cannot be solved in time $O(n^{2-\delta})$ for thresholds

$$t_1' = \alpha \left(\frac{1-\gamma}{1+4\gamma}\right)^{2^i} \left(\frac{\delta}{T} + 1 - \delta\right)^{2^i}$$

$$t_2' = \left(\frac{1+\gamma}{1-4\gamma}\right)^{2^i} \frac{\left(\frac{\delta}{2T} + 1 - \delta\right)^{2^i}}{\frac{1}{\alpha} + \left(\frac{\delta}{T} + 1 - \delta\right)^{2^i} - \left(\frac{\delta}{2T} + 1 - \delta\right)^{2^i}}$$

where we observe that by construction $t_1' = \alpha \cdot t_1 = j_1$. We refer to Appendix B in the full version on ArXiv for the calculations [8, App. B]. So we have constructed an instance which is hard for thresholds j_1 and t_2'.

Set $t_2^* = \left(\frac{1+\gamma}{1-4\gamma}\right)^{2^i} \left(\frac{\delta}{2T} + 1 - \delta\right)^{2^i}$. Then $t_2' < \alpha t_2^*$ and so the hardness for $t_1' = j_1$ and t_2' implies hardness for $t_1' = j_1$ and αt_2^*. We show that there is an ε that only depends on δ such that $\alpha t_2^* < j_2$. Then the hardness for $t_1' = j_1$ and αt_2^* implies hardness for the given j_1 and j_2.

Note that we have chosen $\alpha \geq t_1$, since otherwise i could not be maximal. So we have:

$$\frac{\log(j_1)}{\log(\alpha t_2^*)} = \frac{\log(\alpha t_1)}{\log(\alpha t_2^*)} \leq \frac{\log\left(t_1^2\right)}{\log\left(t_1 \cdot \left(\frac{1+\gamma}{1-4\gamma}\right)^{2^i} \cdot \left(\frac{\delta}{2T} + 1 - \delta\right)^{2^i}\right)}$$

$$= \frac{2^i \cdot \log\left(\left(\frac{1-\gamma}{1+4\gamma}\right)^2 \cdot (\delta/T + 1 - \delta)^2\right)}{2^i \cdot \log\left(\left(\frac{1-\gamma}{1+4\gamma}\right) \cdot (\delta/T + 1 - \delta)\left(\frac{1+\gamma}{1-4\gamma}\right) \cdot \left(\frac{\delta}{2T} + 1 - \delta\right)\right)}.$$

We need to show that this expression is bounded by $1 - \varepsilon$ for some ε that depends on δ, but not on j_1 and j_2. Observe that the factors 2^i cancel out and we may pick γ small enough that it can essentially be ignored. We show in Appendix B in the full version on ArXiv [8, App. B] that we can use any $\gamma < \min\left\{\frac{1}{2^{i+1}}, \frac{\delta}{20T}\right\}$. Then for given δ, there exists an ε such that the expression is bounded by $1 - \varepsilon$, since T can be considered a constant for a fixed δ. Recall that T was defined in Lemma 3. By the assumption $j_1 \leq j_2^{1-\varepsilon}$ we then have $\alpha t_2^* < j_2$. Then the hardness of t_1' and αt_2^* where $t_1' = j_1$ and $\alpha t_2^* < j_2$, implies the desired hardness for the given j_1 and j_2.

We finally argue about the size of the universe of the instance constructed by the compositions of reductions described. In the following, d is the size of the universe of the initial instance of Bichromatic Closest Pair with Jaccard instance. In the proof of Lemma 8, we argued that we could use x_2, which was the size of the intersection for a pair with Jaccard similarity j_2, in the sample size s_i, which means that

$$\begin{aligned} s_i &\geq \frac{30\ln(n)d^{2^i}}{\gamma^2(1-\gamma)^{2^i}x_2^{2^i}} = \frac{30\ln(n)d^{2^i}}{\gamma^2(1-\gamma)^{2^i}(j_2(|a|+|b|-x_2))^{2^i}} \\ &= \frac{30\ln(n)}{\gamma^2(1-\gamma)^{2^i}j_2^{2^i}} \cdot \left(\frac{\delta+1}{1+\frac{\delta}{2T}}\right)^{2^i} . \end{aligned}$$

Again, the calculations can be found in Appendix B in the full version on ArXiv [8, App. B]. Hence, the sets constructed by the composition of reductions come from a universe whose size is bounded by

$$|U| \leq s_i + s_i(1/\alpha - 1) = \frac{s_i}{\alpha} \leq \frac{30\ln(n)}{\gamma^2 j_2^{2^i}} \left(\frac{\delta+1}{\left(\frac{\delta}{T}+1-\delta\right)\left(\frac{\delta}{2T}+1\right)}\right)^{2^i} \left(\frac{1+4\gamma}{(1-\gamma)^2}\right)^{2^i}$$

By Assumption $t_1'^2 < j_1 \leq t_1'$, which implies that $2^i = O\left(\frac{\log j_1}{\log c}\right) = O\left(\log \frac{1}{j_1}\right)$ for constant $c < 1$. Hence, we conclude that the size of the universe is $\ln(n)/j_2^{O(\log 1/j_1)}$. This finishes the proof of Theorem 2. ◀

6 Final Comments

On a final note, we remark that one can obtain a result similar to Theorem 2 for Braun-Blanquet similarity. Recall that we define Braun-Blanquet similarity for a pair of sets $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$ as

$$BB(\mathbf{a}, \mathbf{b}) = \frac{|\mathbf{a} \cap \mathbf{b}|}{\max\left\{|\mathbf{a}|, |\mathbf{b}|\right\}} \in [0, 1]$$

In fact, the proof is slightly simpler than the one given in Section 5.3 and the calculations are somewhat nicer. The proof ideas, i.e., the choice and order of reductions, are exactly the same and should be easy to carry out by following the structure of the proof of Theorem 2.

The main open problem we leave is whether existing upper bounds are near-optimal when ε is an arbitrary constant between 0 and 1. Our techniques only work when ε is sufficiently small.

References

1 Andrei Z Broder. On the resemblance and containment of documents. In *Compression and complexity of sequences 1997. proceedings*, pages 21–29. IEEE, 1997.

2 Lijie Chen. On The Hardness of Approximate and Exact (Bichromatic) Maximum Inner Product. In *33rd Computational Complexity Conference, CCC 2018, June 22-24, 2018, San Diego, CA, USA*, pages 14:1–14:45, 2018. doi:10.4230/LIPIcs.CCC.2018.14.

3 Lijie Chen and Ryan Williams. An Equivalence Class for Orthogonal Vectors. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 21–40, 2019. doi:10.1137/1.9781611975482.2.

4 Tobias Christiani and Rasmus Pagh. Set similarity search beyond MinHash. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 1094–1107, 2017. doi:10.1145/3055399.3055443.

5 Ashish Goel, Aneesh Sharma, Dong Wang, and Zhijun Yin. Discovering similar users on twitter. In *11th Workshop on Mining and Learning with Graphs*, 2013.

6 Pankaj Gupta, Ashish Goel, Jimmy J. Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. WTF: the who to follow service at twitter. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 505–514, 2013. doi:10.1145/2488388.2488433.

7 Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*, pages 604–613, 1998. doi:10.1145/276698.276876.

8 Rasmus Pagh, Nina Stausholm, and Mikkel Thorup. Hardness of Bichromatic Closest Pair with Jaccard Similarity, 2019. arXiv:1907.02251.

9 Aviad Rubinstein. Hardness of approximate nearest neighbor search. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 1260–1268, 2018. doi:10.1145/3188745.3188916.

10 Gregory Valiant. Finding Correlations in Subquadratic Time, with Applications to Learning Parities and the Closest Pair Problem. *J. ACM*, 62(2):13:1–13:45, 2015. doi:10.1145/2728167.

11 Virginia Vassilevska Williams. Some Open Problems in Fine-Grained Complexity. *SIGACT News*, 49(4):29–35, 2018. doi:10.1145/3300150.3300158.