SIAM J. Comput.
Vol. 48, No. 2, pp. 350–372

ON THE DISTORTION OF LOCALITY SENSITIVE HASHING*

FLAVIO CHIERICHETTI[†], RAVI KUMAR[‡], ALESSANDRO PANCONESI[†], AND ERISA TEROLLI[§]

Abstract. Given a notion of pairwise similarity between objects, locality sensitive hashing (LSH) aims to construct a hash function family over the universe of objects such that the probability two objects hash to the same value is their similarity. LSH is a powerful algorithmic tool for large scale applications and much work has been done to understand LSHable similarities, i.e., similarities that admit an LSH. In this paper we focus on similarities that are provably non-LSHable and propose a notion of distortion to capture the approximation of such a similarity by an LSHable similarity. We consider several well-known non-LSHable similarities and show tight upper and lower bounds on their distortion.

Key words. locality sensitive hashing, distortion, similarity

AMS subject classifications. 68Q99, 68R99

DOI. 10.1137/17M1127752

1. Introduction.

The notion of similarity finds use in a large variety of fields beyond computer science. Often, the notion is tailored to the actual domain and application for which it is intended. Locality sensitive hashing (henceforth LSH) is a powerful algorithmic paradigm for computing similarities between data objects in an efficient way. Informally, an LSH scheme for a similarity is a probability distribution over a family of hash functions such that the probability the hash values of two objects agree is precisely the similarity between them. In many applications, computing similar objects (i.e., finding nearest neighbors) can be computationally very demanding and LSH offers an elegant and cost-effective alternative.

Intuitively, large objects can be represented compactly and yet accurately from the point of view of similarity, thanks to LSH. Thus, the similarity between two objects can be quickly estimated by picking a few random hash functions from the family and estimating the fraction of times the hash functions agree on the two objects. This paradigm has been very successful in a variety of applications dealing with large volumes of data, from near-duplicate estimation in text corpora to a nearest-neighbor search in a multitude of domains.

Given its success and importance,[1] researchers have looked for LSH schemes for more and more similarities. Thus a natural question arises: which similarities admit an LSH scheme? In [13] Charikar introduced two necessary criteria (the former weaker than the latter) for a similarity S to admit an LSH:

*Received by the editors April 27, 2017; accepted for publication (in revised form) January 23, 2019; published electronically April 2, 2019. A preliminary version of this paper appears in the Proceedings of the 8th Innovations in Theoretical Computer Science (ITCS 2017) conference.

http://www.siam.org/journals/sicomp/48-2/M112775.html

Funding: This work was partially supported by a Google Focused Research Award, by ERC Starting grant DMAP 680153, by SIR grant RBSI14Q743 of the Italian Ministry of Education, University and Research (MIUR), and by research grant C26M15ALKP of Sapienza University of Rome.

[†]Dipartimento di Informatica, Sapienza University of Rome, Rome 00198, Italy (flavio@di.uniroma1.it, ale@di.uniroma1.it).

[‡]Google, Mountain View, CA 94043 (ravi.k53@gmail.com).

[§]Max Planck Institute for Informatics, Saarbrücken 66123, Germany (eterolli@mpi-inf.mpg.de).

[1] The 2012 Paris Kanellakis Theory and Practice Award was given to Broder, Charikar, and Indyk for their work on LSH.

(T1) $1 - S$ must be a metric;

(T2) $1 - S$ must be isometrically embeddable in ℓ_1.

These two tests can be used to rule out the existence of LSH schemes for various similarities, for instance, the Sørensen–Dice and Sokal–Sneath similarities (see Table 1 or [16] for definitions).

TABLE 1

A list of similarities and of their lower and upper distortion bounds. The value n refers to the cardinality of the ground set or to the number of dimensions.

Name	$S(X,Y)$ $X \neq Y$	Distortion LB	Distortion UB
Jaccard	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\lvert X\triangle Y\rvert}$	1	1 (Shingles [9])
Hamming	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+\lvert X\triangle Y\rvert}$	1	1 (folklore)
Anderberg	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+2\lvert X\triangle Y\rvert}$	1	1 (RSS [14])
Rogers–Tanimoto	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+2\lvert X\triangle Y\rvert}$	1	1 (RSS [14])
Cosine	$\frac{X\cdot Y}{\ell_2(X)\cdot\ell_2(Y)}$	$\sqrt{n}$ (Theorem 4.3)	$6\sqrt{n}$ (Theorem 4.4)
Simpson	$\frac{\lvert X\cap Y\rvert}{\min\{\lvert X\rvert,\lvert Y\rvert\}}$	n (Theorem 4.2)	n (Shingles [9])
Braun–Blanquet	$\frac{\lvert X\cap Y\rvert}{\max\{\lvert X\rvert,\lvert Y\rvert\}}$	2 (Theorem 5.8)	2 (Shingles [9])
Sørensen–Dice	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+1/2\lvert X\triangle Y\rvert}$	2 (Theorem 4.2)	2 (Shingles [9])
Sokal–Sneath 1	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+1/2\lvert X\triangle Y\rvert}$	4/3 (Theorem 4.6)	2 (RSS [14])
Forbes	$\frac{n\,\lvert X\cap Y\rvert}{\lvert X\rvert\,\lvert Y\rvert}$	n (Theorem 7.1)	n (Theorem 7.1)
SORENSEN$_\gamma$	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\gamma\lvert X\triangle Y\rvert}$	$\max(1,1/\gamma)$ (Theorem 4.2)	$\max(1,1/\gamma)$ (Shingles [9], RSS [14])
SOKAL-SNEATH$_\gamma$	$\frac{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert}{\lvert X\cap Y\rvert+\lvert\overline{X\cup Y}\rvert+\gamma\lvert X\triangle Y\rvert}$	$\max(1,2/(1+\gamma))$ (Theorem 4.6)	$\max(1,1/\gamma)$ (RSS [14])

This brings us to a very natural question, the one we address in this paper: *if a similarity S does not admit an LSH scheme, then how well can it be approximated by another similarity S' that admits an LSH?*

Locality sensitive distortion. The two criteria (T1) and (T2) are one of the many points of contact between LSH schemes and the theory of embeddability in metric spaces, where the natural notion of "closeness" is distortion. We say that a similarity S has a *distortion* not larger than δ if there is a similarity S' defined by the same universe that admits an LSH and such that

$$\frac{S}{\delta} \leq S' \leq S.$$

The distortion is 1 if and only if S admits an LSH.

In this paper we begin a systematic investigation of the notion of distortion for LSH schemes and prove optimal distortion bounds for several well-known and widely used similarities such as cosine, Simpson, Braun–Blanquet (also known as "all-confidence"), Sørensen–Dice, and several others (see Table 1). We obtain our lower bounds by introducing two new combinatorial tools dubbed the *center method* and the *k-sets method*. In nearly all cases, we also exhibit matching distortion upper bounds by explicitly constructing an LSH. As concrete examples, we show that the distortion of the cosine similarity grows as the square root of the number of dimensions of the vectors, and that the distortion of the Braun–Blanquet, and Sørensen–Dice, similarities is 2 (the full picture is given in Table 1).

Each of the two methods leverages on the following basic idea. It is usually the case that, given a similarity S defined on pairs of objects coming from a universe $\mathcal{U}$, there exists a set $\mathcal{Z} \subseteq \binom{\mathcal{U}}{2}$ of pairs of elements of $\mathcal{U}$ such that S evaluates to zero on each of the pairs of $\mathcal{Z}$. For instance, the Jaccard similarity evaluates to zero on pairs of disjoint sets, and the cosine similarity evaluates to zero on pairs of orthogonal vectors. Suppose now that, for such a similarity, we can find a set of pairs $\mathcal{A} \subseteq \binom{\mathcal{U}}{2}$ such that

- the minimum value of S over the pairs in $\mathcal{A}$ is at least τ, i.e.,

$$\min_{\{a,b\}\in\mathcal{A}} S(a,b) \geq \tau,$$

- and, for each LSHable similarity S' such that $S'(a,b) = 0$ for each $\{a,b\} \in \mathcal{Z}$, the average of S' over pairs in $\mathcal{A}$ is at most τ/δ, for some $\delta > 1$, i.e.,

$$\operatorname*{avg}_{\{a,b\}\in\mathcal{A}} S'(a,b) \leq \frac{\tau}{\delta}.$$

Then, it must be the case that the distortion of S is at least δ, since all the LSHable S' that distort S by a finite amount have to evaluate to exactly zero on pairs in $\mathcal{Z}$. Although this is not apparent from this high level description, for many similarities, a judicious choice of $\mathcal{A}$ and $\mathcal{Z}$ allows us to pick large δ's and hence show large enough distortions. The center and the k-sets methods implement this plan in two different ways (that is, with two different pairs of $\mathcal{A}$ and $\mathcal{Z}$). The methods appear to be quite versatile for they give precise distortion bounds for many known similarities of interest. (We note in passing that one could obtain the same lower bound of δ on the distortion of S by weakening the first assumption to $\operatorname{avg}_{\{a,b\}\in\mathcal{A}} S(a,b) \geq \tau$. As it turns out, however, all our applications of the methods go through using the simpler uniform lower bound mentioned above.)

Our framework also expands the outreach of the tests (T1) and (T2) along two different dimensions. First, not only does it allow one to determine whether a given similarity is not LSHable, but it provides a quantitative framework to determine how far it is from being so. Second, it allows one to establish that similarities do not admit LSH schemes even when both tests (T1) and (T2) are passed. Indeed, we show that the Braun–Blanquet similarity has a distortion of exactly two, and therefore that it does not admit an LSH scheme. This similarity is particularly noteworthy because it passes both test (T1) and test (T2). To show this we prove that it is embeddable isometrically in ℓ_1, a result that may be of independent interest. Besides the two general methods discussed, which apply to many notable cases of interest, we also provide ad hoc distortion bounds for the Forbes similarity.

Of the two methods introduced in our work, the center method is easier to establish than the k-sets method. The former is applicable to many instances of similarity but the latter is necessary in the following sense. The Braun–Blanquet similarity not only passes (T1) and (T2) as noted earlier, but also the test provided by the center method. Thanks to the more powerful k-sets method, however, one can show a tight lower bound of two on its distortion, and hence its non LSHability. Other similarities to which the k-sets method applies are Sørensen–Dice and the family $\textsc{sorensen}_\gamma$.

Upper bounds: Worst-case versus practice. The main motivation behind our work is to extend the range of applicability of LSH as far as possible, and our concept of distortion should be understood in these terms. For instance, even if a similarity is shown not to admit an LSH scheme it might be possible to approximate it efficiently by means of LSH schemes of other similarities that are close to it. Our results show that some cases, such as cosine, are a forlorn hope (since the distortion is not a constant), but in other instances, such as Sørensen–Dice and Braun–Blanquet, our bounds give reason to be optimistic. As a first "proof of concept" of the notion of distortion we performed a series of experiments with real-world text corpora. The results are encouraging, for they show that the distortion of real data sets is smaller than the worst case. In our tests the average distortion turned out to be approximately 1.4 as opposed to the worst-case bound of two.

In the same vein we also investigate experimentally for the first time the effectiveness of two recent LSH schemes for Anderberg and Rogers–Tanimoto similarities. Until the work in [14] it was not known whether these similarities admitted LSH schemes. That paper shows that they do, in a somewhat peculiar way; strictly speaking they might need exponentially many bits (albeit with low probability)! In this paper we report on experiments with real text corpora that show that in practice these schemes are quite efficient.

2. Related work.

LSH was formally developed over a series of papers [9, 10, 27, 28]. Broder et al. [9, 10] showed that min-wise independent permutations form an LSH for the Jaccard similarity. Indyk and Motwani [27] introduced sampling hash as an LSH scheme for the Hamming similarity. Pursuing the work of characterizing similarities that admit an LSH, Charikar [13] introduced (T1) and (T2) as necessary criteria. Chierichetti and Kumar [14] proposed the concept of LSH-preserving functions—that is, functions that preserve the LSH property of a similarity—showing that they all are the only the (possibly scaled-down) probability generating functions. From the point of view of applications, LSH has been widely used for solving the approximate or exact near-neighbor search [2] and similarity search [24, 32, 42] in high dimensional spaces. For a detailed bibliography on LSH, including pointers to implementations, see Alex Andoni's LSH page (www.mit.edu/~andoni/LSH/) and the surveys of Andoni and Indyk [3] and Wang et al. [47]. Our paper deals with upper and lower bounds on the minimum distortion that one has to apply to a similarity in order to obtain an LSH for it. This goal is somewhat orthogonal to a number of well-known results on LSH (e.g., [4, 35, 36]) that deal with lower bounds of an entirely different nature such as the minimum query time, and the minimum space, required by nearest-neighbor data structures based on Indyk–Motwani LSH schemes, and on more general approaches such as sketching algorithms.

Similarities are extensively used in various areas of computer science. The Hamming similarity, for instance, is widely used in information theory [6, 7, 20]. Areas like data mining and data management have seen the usage of Anderberg similarity [1], cosine similarity [11, 41], and Sokal–Sneath [45] similarity. Cosine similarity is also used in information retrieval [23, 34, 40, 50] and bioinformatics [12] and Sokal–Sneath is used in image processing [5]. We should note here that similarity algorithms/functions are also used outside computer science. For instance, Sørensen–Dice is commonly used in ecology [18, 30, 31], phytosociology [29, 46], plant taxonomy [48], biology [43], and even in lexicography [39]. Biology has also seen the usage of Sokal–

Sneath [44, 49], mentioned above. Other interesting examples are Simpson similarity used in microscopy [33] and biology [19], Braun–Blanquet in phytosociology [8] and ecology [37], and Rogers–Tanimoto in taxonomy [38].

The notion of distortion is studied in various areas of computer science and mathematics, especially in metric embedding problems. Here, we are given a source metric space (X, d) and a target metric space (X', d'), and we wish to find a map $f : X \to X'$ from points in X to points in X' that minimizes the distortion,

$$\max_{\{a,b\}\in\binom{X}{2}} \max\left(\frac{d(a,b)}{d'(f(a),f(b))}, \frac{d'(f(a),f(b))}{d(a,b)}\right).$$

Problems of this form have been studied for many source and target metric spaces (cf. [26]). Examples include embeddings into the Euclidean (ℓ_2) metric, into the ℓ_1 metric, or into tree metrics, from either shortest-path metrics on graphs or from normed spaces of large dimensionality. Even though the LSH distortion problem seems to resemble distorted metric embedding problems, an important difference is that we want to guarantee a multiplicative approximation to the "similarity" (as opposed to the "dissimilarity" or distance).

3. Preliminaries.

We use the notation 2^A to represent the set of all subsets of a set A. Also, for any set A, $\binom{A}{2}$ is the set of all pairs $\{a, b\}$ such that $a \neq b$ and $a, b \in A$. For a positive integer n, let $[n] = \{1, 2, \ldots, n\}$.

Let $\mathcal{U}$ be a (finite) universe of objects. A symmetric function $S : \mathcal{U} \times \mathcal{U} \to [0, 1]$ such that $S(X, X) = 1$ for all $X \in \mathcal{U}$ is called a *similarity.* See [16] for a rather complete illustration of the different types of similarities that are used in a practical context.

We first define what it means for a similarity to admit an LSH.

Definition 3.1 (LSH [13]). *An* LSH *for a similarity function $S : \mathcal{U}\times\mathcal{U} \to [0, 1]$ is a probability distribution over a set $\mathcal{H}$ of (hash) functions defined on $\mathcal{U}$ such that, for each $X, Y \in \mathcal{U}$, we have*

$$\Pr_{h\in\mathcal{H}}[h(X) = h(Y)] = S(X, Y).$$

(See [27] for a somewhat different definition of LSH in the same spirit.) A similarity is *LSHable* if there exists an LSH for it. The central notion we introduce in this paper is defined next.

Definition 3.2 (LSH distortion). *The* LSH distortion*, or* distortion*, of a similarity $S : \mathcal{U} \times \mathcal{U} \to [0, 1]$ is the minimum*[2] *$\delta \geq 1$ such that there exists an LSHable similarity $S' : \mathcal{U} \times \mathcal{U} \to [0, 1]$ for which*

$$\frac{1}{\delta} \cdot S(X, Y) \leq S'(X, Y) \leq S(X, Y) \quad \forall X, Y \in \mathcal{U}.$$

We denote $\mathsf{distortion}(S) = \delta$*.*

At first blush a more general definition seems possible. One could define the distortion of S as the minimum δ such that there exist an LSHable similarity S' and $\alpha, \beta \geq 1$, with $\alpha\beta = \delta$, such that, for all $X, Y \in \mathcal{U}$,

$$\frac{1}{\alpha} \cdot S(X, Y) \leq S'(X, Y) \leq \beta \cdot S(X, Y).$$

[2] A minimum δ exists because it is equal to the solution of a linear program (see, e.g., [15]) of size exponential in $|\mathcal{U}|$.

The next lemma, however, implies that Definition 3.2 can be adopted without loss of generality.

Lemma 3.3. *Let $S : \mathcal{U} \times \mathcal{U} \to [0,1]$ be an LSHable similarity. Then, for each $\gamma \in [0,1]$, the similarity*

$$S'(X,Y) = \begin{cases} \gamma \cdot S(X,Y), & X \neq Y, \\ 1, & X = Y \end{cases}$$

is also LSHable.

Proof. Let $\mathcal{H}$ be the hash function family for S given by Definition 3.1. We will build a family $\mathcal{H}'$ for S' by bijectively obtaining an h' for each $h \in \mathcal{H}$. To define h', consider the following procedure: with probability γ, let $h'(X) = h(X)$ for each $X \in \mathcal{U}$, while with probability $1-\gamma$, let $h'(X) = X$ for each $X \in \mathcal{U}$. Then, for each $X \neq Y$, $\Pr[h'(X) = h'(Y)] = \gamma \cdot S(X,Y)$. □

Now, suppose that for a given similarity S, we have an LSHable similarity S' satisfying $\frac{1}{\alpha} \cdot S(X,Y) \leq S'(X,Y) \leq \beta \cdot S(X,Y)$ with $\alpha\beta = \delta$. By applying Lemma 3.3 to S' we obtain an LSH for the similarity $S''(X,Y) = \frac{1}{\beta} \cdot S'(X,Y)$ (when $X \neq Y$) that satisfies

$$\frac{1}{\alpha\beta} \cdot S(X,Y) \leq \frac{1}{\beta} \cdot S'(X,Y) = S''(X,Y) \leq S(X,Y).$$

Hence Definition 3.2 is robust.

Known LSH for set similarities. Set similarities are a kind of similarity whose universe $\mathcal{U}$ satisfies $\mathcal{U} = 2^U$ for some finite ground set U. To give upper bounds on the distortions of various similarities we employ a number of LSH schemes for set similarities proposed in the literature. First and foremost, we employ *shingles* (also known as *MinHash*) [9, 10], which is an LSH scheme for the Jaccard similarity over sets ($\text{JACCARD}(X,Y) = |X \cap Y|/|X \cup Y|$) with universe $\mathcal{U} = 2^U$. To sample a hash function $h \in \mathcal{H}$ from this scheme, one picks a permutation π of the ground set U uniformly at random. Then, $h(X)$, for a set $X \neq \varnothing$, is equal to the element in X with smallest rank in π; here, $h(\varnothing)$ is identically equal to $\perp$. A simple calculation shows that $\Pr_{h \in \mathcal{H}}[h(X) = h(Y)] = \frac{|X \cap Y|}{|X \cup Y|}$ if $X \cup Y \neq \varnothing$, and $\Pr_{h \in \mathcal{H}}[h(\varnothing) = h(\varnothing)] = 1$.

We also use a generalization of shingles given in [13] for the weighted Jaccard similarity. Finally, we use some of the LSH schemes given in [14] for the various rational set similarities. We will use these results as black-boxes and hence we will not describe them.

4. The center method.

In this section we introduce our first lower bound tool for LSH distortion. It will be used to get tight bounds for the distortion of Simpson, and two infinite families of similarities, namely, SORENSEN_γ and the ℓ_p-norm dot product, that contain well-known similarities such as Sørensen–Dice and cosine as special cases. The main workhorse is given by the next theorem. Roughly, it says that if we can find a set of points in our universe that are mutually far apart, then its "center" is far apart from some point in the set. Later in this section, we will also present matching distortion upper bounds for these similarities.

Theorem 4.1. *Suppose that $S : \mathcal{U} \times \mathcal{U} \to [0,1]$ is a similarity admitting an LSH such that there exists $\varnothing \neq \mathcal{X} \subseteq \mathcal{U}$, with $S(X,X') = 0$ for each $\{X,X'\} \in \binom{\mathcal{X}}{2}$. Then, for each $Y \in \mathcal{U}$,*

$$\text{avg}_{X \in \mathcal{X}} \, S(X,Y) \leq \frac{1}{|\mathcal{X}|};$$

thus, there exists at least one $X^\star \in \mathcal{X}$ such that $S(X^\star, Y) \le 1/|\mathcal{X}|$.

Proof. Observe that if h is sampled from the LSH for S, then $h(X) \neq h(X')$ for every $\{X, X'\} \in \binom{\mathcal{X}}{2}$. Therefore, given any $Y \in \mathcal{U}$, and given any h having nonzero probability in the LSH for S, there can be at most one $X \in \mathcal{X}$ such that $h(X) = h(Y)$. Therefore,

$$\sum_{X \in \mathcal{X}} S(X, Y) = \sum_{X \in \mathcal{X}} \Pr[h(X) = h(Y)] \le 1.$$

By dividing the left- and the right-hand sides by $|\mathcal{X}|$ we get the first claim. The second follows trivially. □

We will use this characterization in the following way. For a given similarity, we will find a set $\mathcal{X} \subseteq \mathcal{U}$ of objects that are entirely dissimilar from one another (i.e., all their pairwise similarities are zero) and an additional object $Y \in \mathcal{U} \setminus \mathcal{X}$ (i.e., the *center*) that is more similar than $1/|\mathcal{X}|$ to each of the elements in $\mathcal{X}$. If we can prove a lower bound of $\alpha/|\mathcal{X}|$, $\alpha > 1$, on the similarities $S'(Y, X)$ for each $X \in \mathcal{X}$, then we can conclude that the similarity S' has to be distorted by at least α to admit an LSH. In the remainder of this section we apply Theorem 4.1 to a few notable examples.

4.1. Simpson and generalized Sørensen–Dice.

Let us begin by recalling the definition of the similarities to be discussed in this section. The Simpson similarity, operating on the subsets of the ground set $[n]$, is defined as

$$\textsc{simpson}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

if $|X|, |Y| \ge 1$, as $\textsc{simpson}(X, \varnothing) = 0$ if $|X| \ge 1$, and as $\textsc{simpson}(\varnothing, \varnothing) = 1$. The infinite family $\textsc{sorensen}_\gamma$, for $\gamma > 0$, operating on the subsets of $[n]$, is defined as

$$\textsc{sorensen}_\gamma(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + \gamma |X \triangle Y|}$$

if $|X| + |Y| \ge 1$, and $\textsc{sorensen}_\gamma(\varnothing, \varnothing) = 1$. The $\textsc{sorensen}_\gamma$ family subsumes as special cases several well-known similarities, for instance, Sørensen–Dice ($\gamma = \frac{1}{2}$), Jaccard ($\gamma = 1$), and Anderberg ($\gamma = 2$).

THEOREM 4.2. *For a ground set of n elements,*

$$\mathsf{distortion}(\textsc{simpson}) = n$$

and

$$\mathsf{distortion}(\textsc{sorensen}_\gamma) = \max(1/\gamma, 1) - O(1/n),$$

for each constant $\gamma > 0$.

Proof. First, we show the lower bound by exhibiting an instance on a ground set of n elements. Let $U = [n]$, $Y = U$, and $\mathcal{X} = \{X_1, \ldots, X_n\}$, where $X_i = \{i\}$ for $i \in [n]$. Observe that, for each $\{X_i, X_j\} \in \binom{\mathcal{X}}{2}$, we have that $\textsc{simpson}(X_i, X_j) = \textsc{sorensen}_\gamma(X_i, X_j) = 0$, while, for each $X_i \in \mathcal{X}$, we have $\textsc{simpson}(X_i, Y) = 1$ and $\textsc{sorensen}_\gamma(X_i, Y) = \frac{1}{\gamma n + (1-\gamma)}$.

By Theorem 4.1 we know that for every similarity S with an LSH that finitely distorts $\textsc{simpson}$ or $\textsc{sorensen}_\gamma$, there must exist at least one X_i such that $S(X_i, Y) \le \frac{1}{|\mathcal{X}|} = \frac{1}{n}$. The lower bounds follow.

Next, we show matching upper bounds for the distortion. Recall the definition of the Jaccard similarity:

$$\textsc{jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}.$$

Broder's shingles [9] and min-wise independent permutations [10] are a well-known LSH scheme for Jaccard similarity (see section 2). We use this to prove matching upper bounds for Theorem 4.2.

Min-wise independent permutations form an LSH scheme with distortion n for Simpson similarity since

$$\min(|X|, |Y|) \le |X \cup Y| \le n \cdot \min(|X|, |Y|),$$

as long as $|X|, |Y| \ge 1$. They also provide a distortion of $1/\gamma$ for $\textsc{sorensen}_\gamma$, for every $\gamma \in (0, 1]$ since

$$\gamma|X \cup Y| \le |X \cap Y| + \gamma|X \triangle Y| \le |X \cup Y|.$$

Finally, recall that a result in [14] proves that $\textsc{sorensen}_\gamma$ admits an LSH scheme as long as $\gamma \ge 1$. □

Figure 1 plots the minimum distortion of $\textsc{sorensen}_\gamma$ as γ varies.

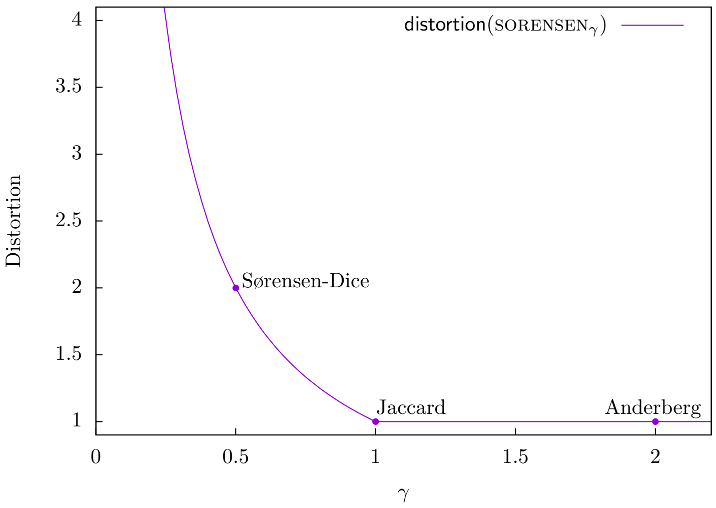


FIG. 1. *The minimum distortion of* $\textsc{sorensen}_\gamma$.

4.2. Cosine and unit ℓ_p-norm dot product.

Recall that given any $p \ge 1$, the ℓ_p norm of a vector $x \in \mathbf{R}^n$ is $\ell_p(x) = \left(\sum_{i=1}^n |x(i)|^p\right)^{1/p}$, and that the cosine similarity of two nonnegative vectors $x, y \in \mathbf{R}^n_+$ having unit ℓ_2 norm is $\sum_{i=1}^n x(i) \cdot y(i)$.

Furthermore, given $p \ge 1$, let

$$B_{p,n} := \left\{ x \in \mathbf{R}^n_+ \mid \sum_{i=1}^n x(i)^p \le 1 \right\} \quad \text{and} \quad S_{p,n} := \left\{ x \in \mathbf{R}^n_+ \mid \sum_{i=1}^n x(i)^p = 1 \right\}$$

be, respectively, the set of points contained in the p-ball of p-radius 1 with nonnegative coordinates, and the set of points lying on the p-sphere of p-radius 1 with nonnegative coordinates.

The universe of the dot product similarity (that we define next) is $B_{p,n}$, which is uncountably infinite. To avoid technical issues in giving a minimally distorted LSH for this similarity, we restrict the universe $B_{p,n}$ to any finite subset $F_{p,n}$ of $B_{p,n}$. Given any such subset, the similarity $\textsc{dot}_{p,n} : F_{p,n} \times F_{p,n} \to [0, \infty)$ is

$$\textsc{dot}_{p,n}(x, y) = \sum_{i=1}^{n} x(i) \cdot y(i).$$

Notice that $\textsc{dot}_{2,n}$ is the well-known cosine similarity (defined on the points of $S_{2,n}$). (Note that we have relaxed the notion of similarity to possibly have range outside $[0, 1]$; the distortion bounds will take care of this issue. For $p = 2$—that is, for the cosine similarity—the range is exactly $[0, 1]$.) We first show an upper bound on distortion and follow that with a matching lower bound.

THEOREM 4.3. *For* $p \geq 2$, $\mathsf{distortion}(\textsc{dot}_{p,n}) \leq 6n^{1-\frac{1}{p}}$.

Proof. We first define two hash schemes and combine them to obtain an LSH for the ℓ_p-norm dot product. Informally speaking, given two generic vectors x and y, the first hash scheme will take care of the coordinates where at least one of x and y have a "small" value, while the second one will take care of the coordinates where both x and y have "large" values.

The first scheme is as follows. First, pick an index $i \in [n]$ uniformly at random. Then, independently for each $x \in F_{p,n}$, select $h'(x)$ as follows: (i) $h'(x) = i$ with probability $\min\left(1, x(i) \cdot n^{1/p}\right)$, and (ii) $h'(x) = x$ with the remaining probability.

For notational convenience, let $\alpha_x^i := \min\left(1, x(i) \cdot n^{1/p}\right)$. Observe that if $x \neq y$, then

$$\Pr[h'(x) = h'(y)] = \frac{\sum_{i=1}^{n} \alpha_x^i \cdot \alpha_y^i}{n}.$$

Then,

$$\Pr[h'(x) = h'(y)] \leq n^{-1} \sum_{i=1}^{n} x(i) n^{\frac{1}{p}} \cdot y(i) n^{\frac{1}{p}} = n^{\frac{2}{p}-1} \sum_{i=1}^{n} x(i) \cdot y(i) \leq \sum_{i=1}^{n} x(i) \cdot y(i).$$

Now, let

$$C = C_{x,y} = \left\{ i \;\mid\; x(i) \leq n^{-\frac{1}{p}} \text{ or } y(i) \leq n^{-\frac{1}{p}} \right\}.$$

Then,

$$\Pr[h'(x) = h'(y)] \geq n^{-1} \sum_{i \in C} \left(x(i) \cdot y(i) \cdot n^{\frac{1}{p}} \right) = n^{\frac{1}{p}-1} \sum_{i \in C} x(i) \cdot y(i).$$

Let us now define the second type of hash function, denoted by h''. Given $x \in F_{p,n}$, we define the vector f_x as follows:

(i) For each coordinate $i \in [n]$, if the value of the ith coordinate of x is larger than $n^{-1/p}$, we let the value of the ith coordinate of f_x be equal to the value of the ith coordinate of x; otherwise, we set the value of the ith coordinate of f_x to 0.

(ii) Moreover, we add to the vector f_x one coordinate for each element of $F_{p,n}$; the value of f_x in its new coordinate associated to x will be equal to

$$n^{1-\frac{1}{p}} - \sum_{\substack{i \\ x(i) > n^{-1/p}}} x(i).$$

The value of f_x in any other new coordinate will be set to 0. Observe that the value of the new coordinate of f_x associated to x will be nonnegative. Indeed, $\sum_{i:x(i)>n^{-1/p}} x(i) \le \ell_1(x)$, and by the Cauchy–Schwarz inequality we get

$$\ell_1(x) \le n^{1-\frac{1}{p}} \ell_p(x) \le n^{1-\frac{1}{p}}.$$

Thus, by definition, $\ell_1(f_x) = n^{1-\frac{1}{p}}$. We now apply the LSH scheme of [13] for weighted Jaccard to the set $\{f_x \mid x \in F_{p,n}\}$. For $x \neq y$, let $\overline{C} = [n]\backslash C$ be the set of coordinates where both x and y have value greater than $n^{-\frac{1}{p}}$. Observe that in any coordinate in C at least one of f_x and f_y has a value of 0. Then, we have

$$\begin{aligned}\Pr[h''(f_x) = h''(f_y)] &= \frac{\sum_i \min(f_x(i), f_y(i))}{\sum_i \max(f_x(i), f_y(i))}\\ &= \frac{\sum_{i\in\overline{C}} \min(x(i), y(i))}{\sum_i \max(f_x(i), f_y(i))}\\ &\le \frac{\sum_{i\in\overline{C}} \min(x(i), y(i))}{\ell_1(f_x)}.\end{aligned}$$

Now, recall that for each $i \in \overline{C}$, each of $x(i)$ and $y(i)$ is larger than $n^{-\frac{1}{p}}$. Therefore, when $p \ge 2$, we have that

$$\max(x(i), y(i)) \ge n^{-\frac{1}{p}} \ge n^{\frac{1}{p}-1} = \frac{1}{\ell_1(f_x)}.$$

Thus,

$$\begin{aligned}\Pr[h''(f_x) = h''(f_y)] &\le \frac{\sum_{i\in\overline{C}} \min(x(i), y(i))}{\ell_1(f_x)} \le \sum_{i\in\overline{C}} \left(\min(x(i), y(i)) \cdot \max(x(i), y(i))\right)\\ &= \sum_{i\in\overline{C}} \left(x(i)\cdot y(i)\right) \le \sum_{i=1}^{n} \left(x(i)\cdot y(i)\right).\end{aligned}$$

Moreover,

$$\Pr[h''(f_x) = h''(f_y)] \ge \frac{\sum_{i\in\overline{C}} \min(x(i), y(i))}{\ell_1(f_x) + \ell_1(f_y)} \ge \frac{\sum_{i\in\overline{C}} (x(i)y(i))}{2n^{1-\frac{1}{p}}}.$$

Therefore, if a hash function h is chosen from the mixture $\frac{1}{3}h' + \frac{2}{3}h''$, we obtain

$$\frac{1}{6n^{1-\frac{1}{p}}} \cdot \sum_{i=1}^{n} x(i)\cdot y(i) \le \Pr[h(x) = h(y)] \le \sum_{i=1}^{n} x(i)\cdot y(i).$$

Thus, there exists an LSH for a similarity that is within distortion $6n^{1-\frac{1}{p}}$ of the dot product similarity on nonnegative vectors having ℓ_p norm at most 1. ∎

Now we show that the distortion of Theorem 4.3 is close to optimal by using, once again, the center method.

THEOREM 4.4. *For $p \ge 1$, even for some finite $F_{p,n} \subseteq S_{p,n}$, it holds that*

$$\mathsf{distortion}(\text{DOT}_{p,n}) \ge n^{1-\frac{1}{p}}.$$

Proof. Consider the n vectors u_i defined as $u_i(i) = 1$, and $u_i(j) = 0$ for each $i \in [n]$ and for each $j \in [n] \setminus \{i\}$. Also, let $u_\star$ be the vector such that $u_\star(i) = n^{-\frac{1}{p}}$, for each $i \in [n]$, and let $X = \{u_1, \ldots, u_n\}$. Observe that for each $x \in X$, we have $\ell_p(x) = 1$ and $\ell_p(u_\star) = 1$—that is, $u_\star \in S_{p,n}$ and $X \subseteq S_{p,n}$.

Suppose that S is an LSHable similarity that distorts $\textsc{dot}_{p,n}$ by the minimum possible amount. Since $S(u_i, u_j) = 0$ for every $i \neq j$, by Theorem 4.1 we know that there exists $u_i \in X$ such that $S(u_i, u_\star) \leq \frac{1}{n}$. Since $\textsc{dot}_{p,n}(u_i, u_\star) = n^{-\frac{1}{p}}$, the distortion is at least $n^{1-\frac{1}{p}}$. □

As a simple corollary, we observe that the distortion for the cosine similarity is $\Theta(\sqrt{n})$ and that the distortion bound is tight for $p \geq 2$.[3] We conjecture that it is generally tight for all $p \geq 1$, i.e., that Theorem 4.3 could be strengthened to all $p \geq 1$.

Conjecture 4.5. *For each* $p \geq 1$, $\mathsf{distortion}(\textsc{dot}_{p,n}) = \Theta(n^{1-\frac{1}{p}})$.

4.3. Sokal–Sneath similarities.

Finally, we look at the Sokal–Sneath similarities. For $\gamma > 0$, let

$$\textsc{sokal-sneath}_\gamma(X, Y) = \frac{|X \cap Y| + \left|\overline{X \cup Y}\right|}{|X \cap Y| + \left|\overline{X \cup Y}\right| + \gamma\, |X \triangle Y|}.$$

Observe that $\textsc{sokal-sneath}_1$ is the Hamming similarity, $\textsc{sokal-sneath}_{1/2}$ is the Sokal–Sneath 1 similarity, and $\textsc{sokal-sneath}_2$ is the Rogers–Tanimoto similarity.

In [14] it is proved that $\textsc{sokal-sneath}_\gamma$ has an LSH iff $\gamma \geq 1$. Thus, the Hamming similarity and the Rogers–Tanimoto similarity admit an LSH, while the Sokal–Sneath 1 similarity does not admit an LSH.

We use the center method to prove a lower bound on the LSH-distortion of $\textsc{sokal-sneath}_\gamma$.

Theorem 4.6. *For any* $0 < \gamma < 1$,

$$\frac{2}{1+\gamma} \leq \mathsf{distortion}(\textsc{sokal-sneath}_\gamma) \leq \frac{1}{\gamma}.$$

Proof. We begin with the lower bound. Given any ground set $[n]$ of even cardinality, consider the three sets $X = [n/2]$, $X' = [n] \setminus [n/2]$, and $Y = [n]$. We have $\textsc{sokal-sneath}_\gamma(X, X') = 0$, $\textsc{sokal-sneath}_\gamma(X, Y) = \textsc{sokal-sneath}_\gamma(X', Y)$, and

$$\textsc{sokal-sneath}_\gamma(X, Y) = \frac{1/2}{1/2 + \gamma/2} = \frac{1}{1+\gamma}.$$

Consider any set similarity S on the ground set $[n]$ that admits an LSH, and that guarantees that $S(X, X') = 0$. By Theorem 4.1, there must exist $X^\star \in \{X, X'\}$ such that $S(X^\star, Y) \leq 1/2$. It follows that the distortion is at least $\frac{\frac{1}{1+\gamma}}{\frac{1}{2}} = \frac{2}{1+\gamma}$.

As for the upper bound, observe that for $0 < \gamma < 1$, we can approximate $\textsc{sokal-sneath}_\gamma$ with $\textsc{sokal-sneath}_1$ by introducing a distortion of $1/\gamma$. Since $\textsc{sokal-sneath}_1$ admits an LSH [14], it follows that $\mathsf{distortion}(\textsc{sokal-sneath}_\gamma) \leq 1/\gamma$. □

[3] While, as we prove in this paper, the cosine similarity does not admit a bounded-distortion LSH, the so-called SimHash scheme [13] provides an LSH scheme for a related similarity, namely $1 - \theta(u,v)/\pi$, where $\theta(u, v)$ is the angle between the two nonzero vectors u and v.

5. The k-sets method.

In this section we introduce our second tool for lower bounding the distortion of LSH. This method is geared towards set similarities. The main ingredient is the following theorem. Let $\mathcal{U}_{n,k}$ denote $\binom{[n]}{k}$.

Theorem 5.1. *Let $k = o(\sqrt{n})$, and let $S : \mathcal{U}_{n,k} \times \mathcal{U}_{n,k} \to [0, 1]$ be a similarity such that $S(X, Y) = 0$ if $X \cap Y = \varnothing$. If S admits an LSH, then*

$$f(S) := \operatorname*{avg}_{\substack{\{X,Y\}\in\binom{\mathcal{U}_{n,k}}{2}\\ |X\cap Y|=1}} S(X,Y) \,\le\, \alpha_k + O\left(\frac{k}{n}\right), \quad \text{where} \quad \alpha_k := \frac{1}{2k-1}.$$

This will be used in the following way. Suppose that we have a similarity S' defined on sets such that $S'(X, Y) = 0$ whenever X and Y are disjoint (not all, but many set similarities satisfy this property), and suppose also that $S'(X, Y) \geq d \cdot \alpha_k$ whenever X and Y are such that $|X| = |Y| = k$ and $|X \cap Y| = 1$. If S is LSHable, how small can its distortion be with respect to S'? By Theorem 5.1, there must exist a pair of sets such that $S(X, Y) \leq \alpha_k + O(k/n)$, which implies that the distortion of any LSHable S with respect to S' is at least $d - O(k^2/n)$.

In what follows, we begin with some technical Lemmas (section 5.1) to prove Theorem 5.1 (section 5.2) and then apply it (section 5.3) to Braun–Blanquet similarity, establishing optimal distortion bounds for it. We conclude with a discussion on the error term in Theorem 5.1 (section 5.4). We remark that this "k-sets method" applies to other similarities such as Sørensen–Dice and $\textsc{sorensen}_\gamma$, for which the simpler center method has already been shown to give optimal results (section 4). By contrast, we show (section 6) that neither the center method nor (T1) nor (T2) (see section 1) can be used to lower bound the distortion of Braun–Blanquet.

5.1. Extremal partitions.

A partition of a set is a collection of pairwise disjoint subsets of that set whose union equals that set. Observe that a hash function h on $\mathcal{U}$ naturally induces a partition of $\mathcal{U}$ in the following sense: two objects $X, Y \in \mathcal{U}$ belong to the same side of the partition iff $h(X) = h(Y)$. This view is particularly useful for our purposes, and from now on we will identify a hash function with the partition that it induces.

Definition 5.2 (acceptable partition). *A partition $\mathcal{P}$ of $\mathcal{U}_{n,k}$* induces *a pair $\{X, Y\}$ (with $X \neq Y$) if X, Y belong to the same part of $\mathcal{P}$. A partition is* acceptable *if it induces no pair $\{X, Y\}$ such that X and Y are disjoint. The* value *of a partition is the number of pairs induced by it.*

Our first goal is to prove that no acceptable partition of $\mathcal{U}_{n,k}$ has value greater than

$$\left(1 + O\left({}^{k^2}\!/_{n}\right)\right) \cdot \frac{n^{2k-1}}{2(2k-1)((k-1)!)^2}.$$

Definition 5.3 (nice partition). *An acceptable partition $\mathcal{P} = \{P_1, \ldots, P_t\}$ of $\mathcal{U}_{n,k}$ is* nice *if there exists a partition $I_1, \ldots, I_t$ of $[n]$ such that for each $i \in [t]$,*

$$P_i = \left\{X \in \mathcal{U}_{n,k} \mid I_i \subseteq X \text{ and } X \cap \left(\cup_{j=1}^{i-1} I_j\right) = \varnothing\right\}.$$

We first show that nice partitions satisfy a slightly stronger version of the above bound; we will then reduce any partition to a nice one.

Lemma 5.4. *The value of a nice partition of $\mathcal{U}_{n,k}$ is at most*

$$\frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2}.$$

Proof. The value v of a nice partition of $\mathcal{U}_{n,k}$ is equal to the sum of the numbers of pairs of sets in each part of the partition. Let $I_1, \ldots, I_t$ be the partition of $[n]$ induced by the given nice partition. Let $a_i = |I_i| \geq 1$ and $b_i = \sum_{j=1}^{i-1} |I_j|$. Then, we have

$$v \leq \sum_{i=1}^{t} \binom{\binom{n-a_i-b_i}{k-a_i}}{2} \leq \sum_{i=1}^{t} \frac{\binom{n-a_i-b_i}{k-a_i}^2}{2} \leq \sum_{i=1}^{t} \frac{\binom{n-1-b_i}{k-1}^2}{2},$$

where the last step follows from $\binom{s}{t} \leq \binom{s+1}{t+1}$. Using this,

$$\begin{aligned} v &\leq \sum_{i=1}^{t} \frac{\binom{n-1-b_i}{k-1}^2}{2} \leq \sum_{i=1}^{n-1} \frac{\binom{n-i}{k-1}^2}{2} \leq \sum_{i=1}^{n-1} \frac{(n-i)^{2k-2}}{2\left((k-1)!\right)^2} \leq \frac{1}{2\left((k-1)!\right)^2} \sum_{i=0}^{n-1} i^{2k-2} \\ &\leq \frac{1}{2\left((k-1)!\right)^2} \int_{x=1}^{n} x^{2k-2} dx = \frac{1}{2\left((k-1)!\right)^2} \left[\frac{x^{2k-1}}{2k-1}\right]_1^n \leq \frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2}. \qquad \square \end{aligned}$$

We will make use of the following result of Hilton and Milner [25] (see [22] for a short proof), which bounds the maximum cardinality of an Erdös–Ko–Rado [21] family that is not a star.

THEOREM 5.5 (Hilton–Milner [25]). *Let $\mathcal{F} \subseteq \mathcal{U}_{n,k}$ be a family of sets with pairwise nonempty intersection with $n \geq 2k$. If $\bigcap_{F \in \mathcal{F}} F = \varnothing$, then $|\mathcal{F}| \leq \binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1$.*

We will also need this simple bound for the difference of two binomial coefficients.

FACT 5.6. $\binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1 \leq O\left(k \cdot \frac{n^{k-2}}{(k-2)!}\right)$.

Now, we can finally bound the value of an acceptable partition.

LEMMA 5.7. *The value of an acceptable partition of $\mathcal{U}_{n,k}$ is at most*

$$\left(1 + O\left(\frac{k^2}{n}\right)\right) \cdot \frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2}.$$

Proof. Let $\mathcal{P}$ be an acceptable partition, and let $P_1, \ldots, P_t$ be its parts. Let $p_i = |P_i|$, and let $m_i = \binom{p_i}{2}$ be the number of pairs that belong to P_i. Let $m = \sum_{i=1}^{t} m_i$ be the total number of pairs of $\mathcal{P}$, i.e., let m be the value of $\mathcal{P}$.

Let $\hat{\mathcal{P}} = \{P_i \mid P_i \in \mathcal{P} \wedge \bigcap_{X \in P_i} X = \varnothing\}$, i.e., let $\hat{\mathcal{P}}$ be the set of parts of $\mathcal{P}$ whose sets have an empty intersection. Moreover, let $\hat{p} = \sum_{P_i \in \hat{\mathcal{P}}} p_i$ and $\hat{m} = \sum_{P_i \in \hat{\mathcal{P}}} m_i$. If $P_i \in \hat{\mathcal{P}}$, then Theorem 5.5 entails that $p_i = O\big(k \frac{n^{k-2}}{(k-2)!}\big)$. Therefore, $m_i = O\left(p_i^2\right) = O\big(p_i k \frac{n^{k-2}}{(k-2)!}\big)$ and

$$\hat{m} = \sum_{P_i \in \hat{\mathcal{P}}} m_i = \sum_{P_i \in \hat{\mathcal{P}}} O\left(p_i k \frac{n^{k-2}}{(k-2)!}\right) = O\left(\hat{p} k \frac{n^{k-2}}{(k-2)!}\right).$$

Define M to be $M \triangleq \frac{n^{2k-2}}{(k-1)! \cdot (k-2)!}$. Then, by definition, $\hat{p} \leq \sum_{i=1}^{t} p_i = \binom{n}{k} = O\left(\frac{n^k}{k!}\right)$. Thus, $\hat{m} = O(M)$. Now, let us consider the partition $\mathcal{P}'$ obtained by splitting into singletons all the sets $P_i \in \hat{\mathcal{P}}$. If m' is the total number of pairs in $\mathcal{P}'$, we have that $m \leq m' + M$. Without loss of generality, let $\mathcal{P}' = \{P'_1, \ldots, P'_{t'}\}$ and $|P'_1| \geq \cdots \geq |P'_{t'}|$. Observe that for each P'_i we have $\bigcap_{X \in P'_i} X \neq \varnothing$.

```
ALGORITHM 1. **Require:** $(\mathcal{P}'_i)$ {A partition $\mathcal{P}'_i$, such that $\forall P \in \mathcal{P}'_i$ it holds that $\bigcap_{X\in P} X \neq \varnothing$}
A greedy selection rule.
Let $\mathcal{P}'_i = \{Q_1, \dots, Q_{t'}\}$ with $|Q_1| \geq \cdots \geq |Q_{t'}|$
for $i = 1, \dots, t' - 1$ do
  if there exists a set $T \in \bigcup_{j=i+1}^{t'} Q_j$ such that $T \cap \bigcap_{P\in Q_i} P \neq \varnothing$ then
    remove $T$ from its part and add it to $Q_i$
    let the resulting partition be $\mathcal{P}'_{i+1}$
    return $\mathcal{P}'_{i+1}$
  end if
end for
return $\mathcal{P}'_i$
```

Let $\mathcal{P}'_0 = \mathcal{P}'$ and $m'_0 = m'$. Algorithm 1 is a greedy selection rule that can be used to produce a sequence $\mathcal{P}'_1, \dots, \mathcal{P}'_\ell$ of acceptable partitions, where $\mathcal{P}'_0$ is defined as above and $\mathcal{P}'_{i+1} = \textbf{Greedy}(\mathcal{P}'_i)$. The sequence stops at the smallest ℓ such that $\mathcal{P}'_\ell = \textbf{Greedy}(\mathcal{P}'_\ell)$, and it satisfies the following properties: (i) $\mathcal{P}'_\ell$ is (by definition) a nice partition, and if we let m'_i be the value of partition $\mathcal{P}'_i$, it holds that (ii) $m'_0 \leq m'_1 \leq \cdots \leq m'_\ell$. Observe that in each iteration where the partition is modified, i.e., where the algorithm moves a set from Q_j to Q_i with $j > i$, the number of pairs in the partition (i.e., its value) gets reduced by $|Q_j| - 1$, but it gets increased by $|Q_i|$. By $j < i$ we have $|Q_i| \geq |Q_j|$, and therefore the total number of pairs increases by at least one unit; therefore $m'_{i+1} > m'_i$, and property (ii) has been proved.

Returning to our main goal, we have that $m \leq m' + O(M) \leq m'_\ell + O(M)$, where m'_ℓ is the value of a nice partition. By Lemma 5.4, we have $m'_\ell \leq \frac{n^{2k-1}}{2\cdot(2k-1)\cdot((k-1)!)^2}$. Thus,

$$m \leq \frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2} + O\left(\frac{n^{2k-2}}{(k-1)!\cdot(k-2)!}\right)$$
$$= \left(1 + O\left(\frac{k^2}{n}\right)\right) \cdot \frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2}.$$

5.2. Proof of Theorem 5.1. Let

$$\alpha = \operatorname*{avg}_{\substack{\{X,Y\}\in\binom{\mathcal{U}_{n,k}}{2}\\ |X\cap Y|=1}} S(X,Y)$$

be the average similarity between pairs of sets of cardinality k having an intersection of cardinality 1. Let σ be the total amount of similarity between unordered pairs of sets of cardinality k having intersection 1. To count the number of ordered pairs, observe that we can select the intersection in n possible ways, so we can then select the other elements of the first set in $\binom{n-1}{k-1}$ ways, and the other elements of the second set in $\binom{n-k}{k-1}$ ways. Moreover, each such unordered pair can be ordered in exactly two ways, so that the number of these unordered pairs is equal to $\frac{n}{2} \cdot \binom{n-1}{k-1} \cdot \binom{n-k}{k-1}$. Therefore,

$$\sigma = \frac{n\binom{n-1}{k-1}\binom{n-k}{k-1}}{2}\alpha.$$

Recall that, if $1 \leq c < \frac{n}{\ell^2}$, we have

$$\binom{n-(c-1)\ell}{\ell} \geq \frac{(n-c\cdot\ell)^\ell}{\ell!} = \frac{n^\ell\left(1-\frac{c\ell}{n}\right)^\ell}{\ell!} \geq \frac{n^\ell}{\ell!}\left(1-\frac{c\cdot\ell^2}{n}\right).$$

Substituting $k-1$ for ℓ, we obtain

$$\sigma \geq \left(1-O\left(\frac{k^2}{n}\right)\right)\frac{n^{2k-1}}{2((k-1)!)^2}\cdot\alpha,$$

where the $O(\cdot)$ term tends to 0, since $k = o(\sqrt{n})$. Since $S(X,Y) = 0$ whenever $|X \cap Y| = 0$, we cannot give positive probability to a hash function placing two such sets X and Y in the same part; otherwise we would have infinite distortion. Hence, we can only use acceptable partitions. Suppose that S has an LSH and assume without loss of generality that this LSH gives positive probabilities $p_1, \ldots, p_h > 0$ to partitions $\mathcal{P}_1, \ldots, \mathcal{P}_h$, and that it gives probability 0 to other partitions. Let $v_1, \ldots, v_h$ be the values of partitions $\mathcal{P}_1, \ldots, \mathcal{P}_h$, and observe that $\sum_{i=1}^h p_i = 1$. Then, we have

$$\sigma = \sum_{\substack{\{X,Y\}\in\binom{\mathcal{U}_{n,k}}{2}\\ |X\cap Y|=1}} S(X,Y) \leq \sum_{i=1}^{h}(p_i v_i),$$

i.e., the total amount of similarity mass that an acceptable partition brings to our similarity's values is no larger than the probability that the LSH assigns to the partition times the number of the partition's pairs or, equivalently, to its own value. By Lemma 5.7, the value of an acceptable partition is at most

$$\tau = \left(1+O\left(\frac{k^2}{n}\right)\right)\frac{n^{2k-1}}{2(2k-1)\left((k-1)!\right)^2}.$$

Therefore, $\sigma \leq \sum_{i=1}^h (\tau p_h) = \tau$, that is, if S admits an LSH, then $\tau \geq \sigma$. Thus, we must have

$$1 \geq \frac{\sigma}{\tau} \geq \left(1-O\left(\frac{k^2}{n}\right)\right)\frac{\frac{n^{2k-1}}{2((k-1)!)^2}\cdot\alpha}{\frac{n^{2k-1}}{2(2k-1)((k-1)!)^2}} = \left(1-O\left(\frac{k^2}{n}\right)\right)\cdot\alpha\cdot(2k-1),$$

which implies

$$\alpha \leq \left(1+O\left(\frac{k^2}{n}\right)\right)\frac{1}{2k-1} = \frac{1}{2k-1}+O\left(\frac{k}{n}\right).$$

5.3. The distortion of Braun–Blanquet.

Recall the definition of Braun–Blanquet, which operates on the subsets of the ground set $[n]$:

$$\text{BRAUN-BLANQUET}(X,Y) = \frac{|X\cap Y|}{\max(|X|,|Y|)}$$

if $|X| + |Y| \geq 1$, and $\text{BRAUN-BLANQUET}(X,Y) = 1$ if $X = Y = \varnothing$.

Observe that for sets $X, Y \subseteq [n]$ such that $|X| = |Y| = k \geq 1$, both Braun–Blanquet and Sørensen–Dice evaluate to $1/k$ if $|X \cap Y| = 1$, and that they evaluate to 0 when $|X \cap Y| = 0$. Therefore, Theorem 5.1 implies that they have to be distorted by at least $(1-o_n(1))\cdot(2-1/k)$ when applied on such pairs of k-sets. By letting k grow to infinity, we obtain an asymptotically tight lower bound of 2 on their distortions. More precisely, let $k=\Theta\left(n^{1/3}\right)$, and let n grow to infinity. We will prove that the distortions of the two similarities can be lower bounded by $2-\Theta\left(n^{-1/3}\right)$. Indeed, if we denote with S any of the two similarities, with S' any LSHable similarity with the same domain, and with X,Y any two sets (with $|X|=|Y|=k$ and $|X\cap Y|=1$) that minimize $S'(X,Y)$, we obtain,

$$\frac{S(X,Y)}{S'(X,Y)} \geq \frac{\frac{1}{k}}{\frac{1}{2k-1}+O(\frac{k}{n})} = \frac{2-\frac{1}{k}}{1+O(\frac{k^2}{n})} = 2-O(n^{-1/3}).$$

We finally observe that min-wise independent permutations [9, 10] achieve a distortion of $2-\Theta\left(n^{-1}\right)$ for Braun–Blanquet. Thus, we have the following theorem.

THEOREM 5.8. $\mathsf{distortion}(\text{BRAUN-BLANQUET}) = 2-o(1)$.

5.4. Tightness of Theorem 5.1. We do not know whether the error term of Theorem 5.1 is tight. Here, we give a lower bound on that error term.

LEMMA 5.9. *Fix any $k\geq 2$ and let $n\geq 2k-1$. Then, there exists an LSHable similarity $S:\mathcal{U}_{n,k}\times\mathcal{U}_{n,k}\to[0,1]$ such that $S(X,Y)=0$ if $X\cap Y=\varnothing$ and*

$$\operatorname{avg}_{\substack{\{X,Y\}\in\binom{\mathcal{U}_{n,k}}{2}\\ |X\cap Y|=1}} S(X,Y) \geq \frac{1}{2k-1}+\Omega\left(\binom{n}{2k-1}^{-1}\right).$$

Proof. We use a variant of min-wise independent permutations. Pick a permutation $\pi:[n]\to[n]$ uniformly at random. For a set $X\in\mathcal{U}_{n,k}$, let $m(X)=m_\pi(X)$ be the minimum i such that $\pi(i)\in X$. Then, the hash function will map X to $m(X)$ if $m(X)\leq n-2k+1$, and to $\star$ otherwise.

Now, for any two sets $X,Y\in\mathcal{U}_{n,k}$, (i) if $|X\cap Y|=1$, then the probability that X and Y will be hashed together is at least $\frac{1}{2k-1}+\Omega\big(\binom{n}{2k-1}^{-1}\big)$, and (ii) if $|X\cap Y|=0$, the probability that X and Y will be hashed together is 0. The claim follows. □

6. Is the k-sets method necessary?

In this section we prove that Braun–Blanquet satisfies the following:

(i) $1-$ BRAUN-BLANQUET is a metric that can be embedded isometrically into ℓ_1, i.e., it passes the tests (T1) and (T2), and

(ii) the center method of section 4 is useless in determining the distortion of Braun–Blanquet.

On the other hand, we know from Theorem 5.8 that its distortion is $2-o(1)$. Thus, the k-sets method is the only method known to prove that Braun–Blanquet does not admit an LSH scheme (and, also, to give a tight bound on the distortion of Braun–Blanquet).

6.1. ℓ_1-embeddability.

LEMMA 6.1. $1-$ BRAUN-BLANQUET *can be isometrically embedded into ℓ_1.*

Proof. Recall that a distance $d:\mathcal{U}\times\mathcal{U}\to[0,\infty)$ can be embedded into ℓ_1 iff it can be expressed as a nonnegative linear combination of cut metrics [17], i.e., iff there exists a nonnegative weighting $w:2^{\mathcal{U}}\to[0,\infty)$ of the subsets of $\mathcal{U}$ such that, for all $\{x,x'\}\in\binom{\mathcal{U}}{2}$, it holds that

$$\sum_{\substack{\varnothing\subset Y\subset\mathcal{U}\\ |\{x,x'\}\cap Y|=1}} w(Y) = d(x,x').$$

We first exhibit such a weighting, and then prove that it satisfies the required equations. Recall that for Braun–Blanquet $\mathcal{U} = 2^{[n]}$. For $i \in [n]$ and $c \in [n]$, let $Y_{i,c} \subseteq \mathcal{U}$ be defined as

$$Y_{i,c} = \{X \in \mathcal{U} \mid X \ni i \text{ and } |X| \le c\}\,.$$

Define w as follows:

(i) $w(\{\varnothing\}) = \frac{1}{2}$;

(ii) $w(Y_{i,c}) = \frac{1}{2c^2+2c}$ for each $i \in [n]$ and $c \in [n-1]$;

(iii) $w(Y_{i,n}) = \frac{1}{2n}$ for each $i \in [n]$; and

(iv) every other set has weight equal to 0.

(To simplify notation, for $n = 1$ we have given positive weight both to a set and to its complement.)

We now prove that w satisfies the required equations. First, note that for integers $1 \le a \le b$, we have

$$\sum_{j=a}^{b-1} \frac{1}{2j^2+2j} = \frac{1}{2} \cdot \sum_{j=a}^{b-1} \left(\frac{1}{j} - \frac{1}{j+1}\right) = \frac{1}{2} \cdot \left(\frac{1}{a} - \frac{1}{b}\right).$$

Consider two distinct nonempty sets $X, X' \in \mathcal{U}$. We have that

$$\begin{aligned}
\ell_1(X,X') &= \sum_{\substack{\varnothing \subset Y \subset \mathcal{U} \\ |\{X,X'\} \cap Y| = 1}} w(Y) \\
&= \sum_{i \in X \setminus X'} \left(\sum_{c=|X|}^{n-1} \left(\frac{1}{2c^2+2c}\right) + \frac{1}{2n} \right) \\
&\quad + \sum_{i \in X' \setminus X} \left(\sum_{c=|X'|}^{n-1} \left(\frac{1}{2c^2+2c}\right) + \frac{1}{2n} \right) \\
&\quad + \sum_{i \in X \cap X'} \left(\sum_{c=\min(|X|,|X'|)}^{\max(|X|,|X'|)-1} \left(\frac{1}{2c^2+2c}\right) \right) \\
&= |X \setminus X'| \left(\frac{1}{2|X|} - \frac{1}{2n} + \frac{1}{2n}\right) \\
&\quad + |X' \setminus X| \left(\frac{1}{2|X'|} - \frac{1}{2n} + \frac{1}{2n}\right) \\
&\quad + |X \cap X'| \left(\frac{1}{2\min(|X|,|X'|)} - \frac{1}{2\max(|X|,|X'|)}\right).
\end{aligned}$$

Let us assume without loss of generality that $|X| \le |X'|$. Then,

$$\begin{aligned}
\ell_1(X,X') &= \sum_{\substack{\varnothing \subset Y \subset \mathcal{U} \\ |\{X,X'\} \cap Y| = 1}} w(Y) \\
&= \frac{|X \setminus X'|}{2|X|} + \frac{|X' \setminus X|}{2|X'|} + \frac{|X \cap X'|}{2|X|} - \frac{|X \cap X'|}{2|X'|} \\
&= \frac{|X|}{2|X|} + \frac{|X'| - 2|X \cap X'|}{2|X'|} \\
&= 1 - \frac{|X \cap X'|}{|X'|} = 1 - \text{BRAUN-BLANQUET}(X,X').
\end{aligned}$$

It remains only to consider the case where exactly one of the two sets is empty. Let $\varnothing \subset X \subseteq [n]$. Then

$$\begin{aligned}\ell_1(X,\varnothing) &= \sum_{\substack{\varnothing \subset Y \subset \mathcal{U}\\ |\{X,\varnothing\}\cap Y|=1}} w(Y)\\ &= w\left(\{\varnothing\}\right) + \sum_{i\in X}\left(\sum_{c=|X|}^{n-1}\left(\frac{1}{2c^2+2c}\right)+\frac{1}{2n}\right)\\ &= \frac{1}{2} + |X|\cdot\left(\frac{1}{2|X|}-\frac{1}{2n}+\frac{1}{2n}\right)\\ &= 1 = 1 - \textsc{braun-blanquet}(X,\varnothing).\end{aligned}$$

The proof is concluded. □

6.2. Inapplicability of the center method. We next show that Theorem 4.1 is inapplicable to the case of the Braun–Blanquet similarity.

LEMMA 6.2. *For each $Y \subseteq [n]$, and for each $\mathcal{X} \subseteq 2^{[n]}$ such that* BRAUN-BLANQUET $(X, X') = 0$ *for all* $\{X, X'\} \in \binom{\mathcal{X}}{2}$, *it holds that*

$$\text{avg}_{X\in\mathcal{X}}\ \textsc{braun-blanquet}(X,Y) \le \frac{1}{|\mathcal{X}|}.$$

Thus, there exists $X \in \mathcal{X}$ such that

$$\textsc{braun-blanquet}(X,Y) \le \frac{1}{|\mathcal{X}|}.$$

Proof. Observe that for $\mathcal{X}$ to satisfy the premise, one has to have that $\{X, X'\} \in \binom{\mathcal{X}}{2}$ implies $X \cap X' = \varnothing$, i.e., the sets in $\mathcal{X}$ have to be pairwise disjoint.

Now, take any $\varnothing \subsetneq Y \subseteq [n]$. We must have

$$\sum_{X\in\mathcal{X}} \textsc{braun-blanquet}(X,Y) = \sum_{X\in\mathcal{X}} \frac{|X\cap Y|}{\max(|X|,|Y|)} \le \sum_{X\in\mathcal{X}} \frac{|X\cap Y|}{|Y|} \le 1,$$

where the last step follows from the pairwise disjointness of the sets in $\mathcal{X}$. If instead $Y = \varnothing$, we have

$$\begin{aligned}&\sum_{X\in\mathcal{X}} \textsc{braun-blanquet}(X,\varnothing)\\ &\le \sum_{\substack{X\in\mathcal{X}\\ X\neq\varnothing}} \frac{0}{\max(|X|,|Y|)} + \textsc{braun-blanquet}(\varnothing,\varnothing)\\ &= 1.\end{aligned}$$

Thus, in general, $\sum_{X\in\mathcal{X}} \textsc{braun-blanquet}(X,Y) \le 1$. It follows that

$$\text{avg}_{X\in\mathcal{X}}\ \textsc{braun-blanquet}(X,Y) \le |\mathcal{X}|^{-1},$$

and the proof is complete. □

7. Ad hoc approaches.

In this section we discuss another similarity, whose distortion bound we prove through a simple ad hoc approach.

7.1. Forbes similarity. The Forbes similarity is defined as $\textsc{forbes}(X,Y) = n \cdot \frac{|X\cap Y|}{|X|\,|Y|}$ if $|X|,|Y| \geq 1$, $\textsc{forbes}(X,\varnothing) = 0$ if $|X| \geq 1$, and $\textsc{forbes}(\varnothing,\varnothing) = 1$. Since $F(\{1\},\{1\}) = n$, we have the following simple observation.

THEOREM 7.1. $\mathsf{distortion}(\textsc{forbes}) = n$.

Proof. The lower bound is trivial since $\textsc{forbes}(\{1\},\{1\}) = n$, and no LSH can assign a value larger than 1 to a pair of sets.

We give an LSH for the similarity $\textsc{forbes}/n$, thus proving an upper bound of n on its distortion. The hash function h will be chosen as follows: $h(\varnothing) = \perp$ and, for each $X \neq \varnothing$ independently, $h(X)$ will be chosen uniformly at random from the elements of X. Then, if $X \neq Y$, we have $\Pr[h(X) = h(Y)] = \frac{|X\cap Y|}{|X|\cdot|Y|}$. □

8. Experiments.

In this section we report on the outcome of two types of experiments. As we have seen in the previous sections the distortion of Braun–Blanquet and of Sørensen–Dice is $2 - o(1)$, and this bound can be matched by Jaccard, which is LSHable. Distortion being a worst-case notion, it is conceivable that the typical behavior of Jaccard with real-world datasets could be somewhat better. This is exactly what our experiments with three real world data sets show. We stress that our results are preliminary, but they give reason for hope and might justify a more comprehensive experimental assessment. The average distortion turns out to be as low as 1.3 for some of our data sets and always less than two. The second set of experiments is a feasibility study of the LSH scheme for Anderberg and Rogers–Tanimoto similarities that until recently were not known to be LSHable. As shown in [14] they are, but in a somewhat peculiar way, because the LSH schemes might need exponentially many bits (with low probability). The goal of our tests is to see whether such schemes are practical. Our study shows that they are, and that in fact they can be very effective with very few bits. We begin by describing our data sets.

8.1. Datasets.

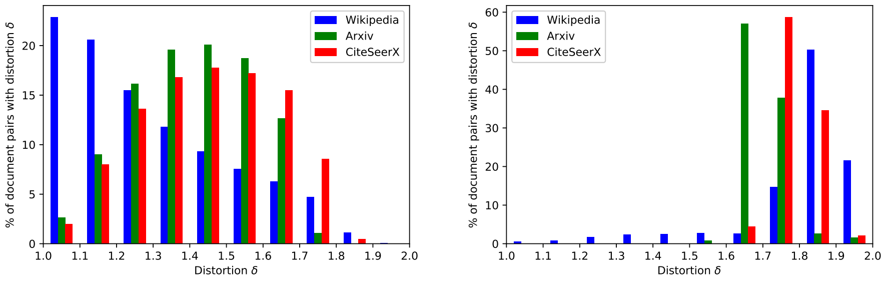
We use three publicly available datasets: (i) a collection of more than 110K scientific papers downloaded from CiteSeerX, (ii) 29K scientific articles downloaded from ArXiv, and (iii) 104K Wikipedia articles. The collection of XML metadata of CiteSeerX and ArXiv was accessed using the OAI protocol for metadata harvesting, which is supported by both digital libraries. The Wikipedia collection was obtained from en.wikipedia.org/wiki/Wikipedia:Database_download. The words in each paper were transformed into lowercase, and each document became a bag of words (no repetitions).

For the experiments of section 8.3 the documents underwent the following “cleaning” procedure: (i) all words not included in the top 1000 most frequent words of the whole dataset were removed, and (ii) every word was hashed to a unique integer. As a result, the papers are represented as vectors containing integers in the range $[1000] = \{1, 2, \ldots, 1000\}$.

8.2. Distortion on real data.

From each corpus, we selected 50 million random pairs of documents and computed the distortion, i.e., the ratio between the Jaccard value (computed exactly) and the two similarities Braun–Blanquet and Sørensen–Dice. Figure 2(a) shows the distortion w.r.t. Braun–Blanquet for our three datasets: ArXiv, CiteSeer, and Wikipedia. For each value of the distortion on the x-axis, the plot gives, on the y-axis, the fraction of pairs with that distortion. Similarly, Figure 2(b) shows the distortion w.r.t. Sørensen–Dice. Table 2 displays the average distortion and the standard deviation of these experiments.

Overall, these tests show that in real-world scenarios the average distortion of



(a) Braun–Blanquet similarity

(b) Sørensen–Dice similarity

FIG. 2. *The fractions of document pairs with a given distortion on three datasets. The distortion values have been bucketed in the intervals* $[1.0, 1.1], (1.1, 1.2], \dots, (1.9, 2.0]$.

TABLE 2
Experimental results.

	Braun–Blanquet		Sørensen–Dice	
	μ	σ	μ	σ
ArXiv	1.45	0.2	1.78	0.09
CiteSeerX	1.4	0.16	1.7	0.05
Wikipedia	1.29	0.21	1.81	1.14

Braun–Blanquet and Sørensen–Dice can be significantly smaller than the worst case bound.

8.3. LSH schemes for rational set similarities. Let us start by recalling the definitions of the similarities we deal with in this section. The Anderberg similarity is defined as follows. Given two nonempty sets X, Y of n elements,

$$\textsc{anderberg}(X, Y) = \frac{|X \cap Y|}{|X \cap Y| + 2|X \triangle Y|},$$

where $\triangle$ is the symmetric difference. (Note that S_2 is the Anderberg similarity.) The value is zero if exactly one of the two sets is empty, and it is 1 whenever $X = Y$. In [14] it is proven that the following is an LSH scheme for the Anderberg similarity. Pick a positive integer r at random with probability 2^{-r}. Let $h_1, \dots, h_r$ be r shingles picked independently. Then, $h(X) := (h_1(X), \dots, h_r(X))$ is an LSH scheme for Anderberg, i.e., $\textsc{anderberg}(X, Y) = \Pr[h(X) = h(Y)]$.

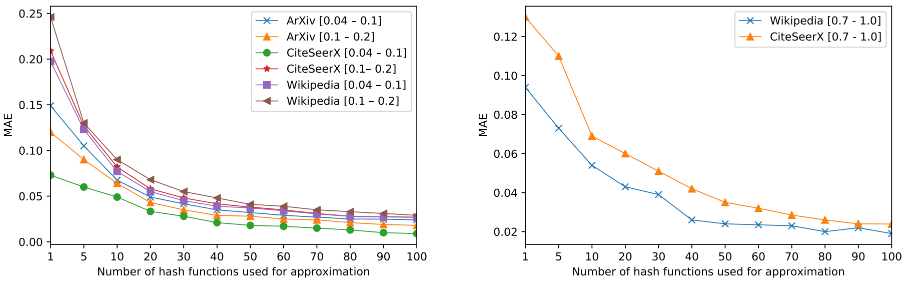
The Rogers–Tanimoto similarity is defined as

$$\textsc{rogers-tanimoto}(X, Y) = \frac{|X \cap Y| + \left|\overline{X \cup Y}\right|}{|X \cap Y| + \left|\overline{X \cup Y}\right| + 2|X \triangle Y|}.$$

(Note that $\textsc{sokal-sneath}_2$ is the Rogers–Tanimoto similarity.) The following is the LSH scheme for Rogers–Tanimoto proposed in [14]. Pick r as before, and then pick r i.i.d. elements $e_1, \dots, e_r$ uniformly at random. The random hash function h is defined as follows. For a set X, we let $h(X) := (e_1 \in X, \dots, e_r \in X)$, where $e_i \in X$ is a Boolean value. Given two sets X and Y, $h(X) = h(Y)$ iff the two vectors coincide on each coordinate (for each element $e = e_1, \dots, e_r$, either both sets have it or neither do).

Recall that in this experiment our corpora consists of bag of words in which only the one thousand most popular words are retained. So each document can be thought of as a binary vector of one thousand coordinates, where coordinate i is one iff the ith most popular word is in the document.

The experiment is as follows. Let h denote a generic hash function of the LSH scheme that we are testing. From each corpus, we picked one hundred thousand random pairs of documents. Then, for every $k \in [100]$, we selected k hash functions $h_1, \ldots, h_k$ and estimated the similarity of the random pair in the usual fashion, i.e., as the fraction of times that $h_i(X) = h_i(Y)$, for $i \in [k]$.



(a) Anderberg similarity

(b) Rogers–Tanimoto similarity

FIG. 3. *The Mean Average Error (MAE) produced by the LSH schemes for Anderberg and Rogers–Tanimoto similarities. The similarity is computed in the natural way as the fraction of collisions over the number of hash functions used. The latter is reported on the x-axis, while on the y-axis the corresponding MAE is shown. The pairs from the three datasets were partitioned into buckets, according to their actual similarity value. For instance, the bottom curve in Figure* 3(a) *shows the MAE obtained for all pairs of documents from CiteSeerX whose Anderberg similarity lies in the range* $[0.04, 0.1]$. *Not all buckets appear in the figure, but the data shown exemplify the general trend.*

Figure 3(a) shows, for each value of k on the x-axis, the mean absolute error (MAE) w.r.t. the real value of Anderberg. Note that already for $k = 20$ the MAE is below 0.05. Since the expected number of shingles used in each h is two (with very small variance) this shows the LSH scheme is inexpensive both timewise and spacewise. Similar conclusions apply to Rogers–Tanimoto, as Figure 3(b) shows.

The experimental results show that the MAE decreases as the number of hashing functions applied increase for each of the databases and similarities tested, reinforcing the theoretical aspects of LSH applied to specific group of similarities that admit such an LSH.

9. Conclusions.

In this paper we studied the notion of distorted LSH schemes for a number of widely-used similarities that do not admit exact LSH schemes. For most of them, we have obtained tight bounds on the minimum distortion required for obtaining an LSH. In doing so, we developed two lower bounding tools that could be useful for bounding the distortion of other similarities that are not LSHable.

To complement our theoretical bounds, we also studied the behavior of our proposed distorted LSH schemes on real datasets. Our main observation is that in practice, the average distortion is milder than what is produced by the worst-case bounds.

It will be interesting to consider other non-LSHable similarities and study their distortion. The encyclopedia [16] is a rich source of such similarities.

REFERENCES

[1] M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[2] A. Andoni and P. Indyk, *Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions*, in Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, IEEE, Washington, DC, 2006, pp. 459–468.

[3] A. Andoni and P. Indyk, *Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions*, Comm. ACM, 51 (2008), pp. 117–122.

[4] A. Andoni, T. Laarhoven, I. Razenshteyn, and E. Waingarten, *Optimal Hashing-Based Time-Space Trade-Offs for Approximate Near Neighbors*, in Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '17, SIAM, Philadelphia, 2017, pp. 47–66, http://dl.acm.org/citation.cfm?id=3039686.3039690.

[5] I. Avcibaş, M. Kharrazi, N. Memon, and B. Sankur, *Image steganalysis with binary similarity measures*, EURASIP J. Adv. Signal Process., 2005 (2005), pp. 2749–2757.

[6] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, *Optimal decoding of linear codes for minimizing symbol error rate*, IEEE Trans. Inform. Theory, 20 (1974), pp. 284–287.

[7] C. H. Bennett and P. W. Shor, *Quantum information theory*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2724–2742.

[8] J. Braun, *Die Vegetationsverhältnisse der Schneestufe in den Rätisch-Lepontischen Alpen: Ein Bild des Pflanzenlebens an seinen äussersten Grenzen*, Schweizerische Naturforschende Gesellschaft, 1913.

[9] A. Z. Broder, *On the resemblance and containment of documents*, in Proceedings of the Compression and Complexity of Sequences, IEEE, Washington, DC, 1997, pp. 21–29.

[10] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, *Min-wise independent permutations*, J. Comput. System Sci., 60 (2000), pp. 630–659.

[11] R. Burke, *Hybrid recommender systems: Survey and experiments*, User Model. User-Adapt. Interact., 12 (2002), pp. 331–370.

[12] J. Chabalier, J. Mosser, and A. Burgun, *A transversal approach to predict gene product networks from ontology-based similarity*, BMC Bioinformatics, 8 (2007), 235.

[13] M. Charikar, *Similarity estimation techniques from rounding algorithms*, in Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, ACM, New York, 2002, pp. 380–388.

[14] F. Chierichetti and R. Kumar, *LSH-preserving functions and their applications*, J. ACM, 62 (2015), 33.

[15] F. Chierichetti, R. Kumar, and M. Mahdian, *The complexity of LSH feasibility*, Theoret. Comput. Sci., 530 (2014), pp. 89–101.

[16] M. M. Deza and E. Deza, *Encyclopedia of Distances*, Springer-Verlag, Berlin, 2009.

[17] M. M. Deza and M. Laurent, *Geometry of Cuts and Metrics*, 1st ed., Springer, Heidelberg, 2009.

[18] L. R. Dice, *Measures of the amount of ecologic association between species*, Ecology, 26 (1945), pp. 297–302.

[19] S. J. Dixon, N. Heinrich, M. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevejo, and R. G. Brereton, *Use of cluster separation indices and the influence of outliers: Application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles*, J. Chemom., 23 (2009), pp. 19–31.

[20] W. H. Equitz and T. M. Cover, *Successive refinement of information*, IEEE Trans. Inform. Theory, 37 (1991), pp. 269–275.

[21] P. Erdös, C. Ko, and R. Rado, *Intersection theorems for systems of finite sets*, Quart. J. Math. Oxford (2), 12 (1961), pp. 313–320.

[22] P. Frankl and Z. Füredi, *Non-trivial intersecting families*, J. Combin. Theorey Ser. A, 41 (1986), pp. 150–153.

[23] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, *Information retrieval using a singular value decomposition model of latent semantic structure*, in Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, 1988, pp. 465–480.

[24] A. Gionis, P. Indyk, R. Motwani, et al., *Similarity search in high dimensions via hashing*, in Proceedings of the 25th International Conference on Very Large Data Bases, Morgan Kaufmann, San Francisco, 1999, pp. 518–529.

[25] A. Hilton and E. Milner, *Some intersection theorems for systems of finite sets*, Quart. J. Math. Oxford Ser. (2), 18 (1967), pp. 369–384.

[26] P. Indyk and J. Matousek, *Low-Distortion Embeddings of Finite Metric Spaces*, Handbook of Discrete and Computational Geometry, CRC Press, Boca Raton, FL, 2004, pp. 177–196.

[27] P. Indyk and R. Motwani, *Approximate nearest neighbors: Towards removing the curse of dimensionality*, in Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM, New York, 1998, pp. 604–613.

[28] P. Indyk, R. Motwani, P. Raghavan, and S. Vempala, *Locality-preserving hashing in multidimensional spaces*, in Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, ACM, New York, 1997, pp. 618–625.

[29] D. H. Knight, *A Phytosociological Analysis of Species-Rich Tropical Forest on Barro Colorado Island, Panama*, Ecol. Monogr., 45 (1975), pp. 259–284.

[30] P. Koleff, K. J. Gaston, and J. J. Lennon, *Measuring beta diversity for presence–absence data*, Jo. Animal Ecol., 72 (2003), pp. 367–382.

[31] J. Looman and J. Campbell, *Adaptation of Sorensen's k* (1948) *for estimating unit affinities in prairie vegetation*, Ecology, 41 (1960), pp. 409–416.

[32] Q. Lv, W. Josephson, Z. Wang, M. Charikar, and K. Li, *Multi-probe LSH: Efficient indexing for high-dimensional similarity search*, in Proceedings of the 33rd International Conference on Very Large Data Bases, 2007, pp. 950–961.

[33] E. Manders, F. Verbeek, and J. Aten, *Measurement of co-localization of objects in dual-colour confocal images*, J. Microsc., 169 (1993), pp. 375–382.

[34] R. Mihalcea, C. Corley, and C. Strapparava, *Corpus-based and knowledge-based measures of text semantic similarity*, in Proceedings of the 21st National Conference on Artificial Intelligence, 2006, pp. 775–780.

[35] R. Motwani, A. Naor, and R. Panigrahy, *Lower bounds on locality sensitive hashing*, SIAM J. Discret. Math., 21 (2007), pp. 930–935, https://doi.org/10.1137/050646858.

[36] R. O'Donnell, Y. Wu, and Y. Zhou, *Optimal lower bounds for locality-sensitive hashing (except when q is tiny)*, ACM Trans. Comput. Theory, 6 (2014), 5, https://doi.org/10.1145/2578221.

[37] M. Poore, *The use of phytosociological methods in ecological investigations:* I. *The Braun–Blanquet system*, J. Ecol., 43 (1955), pp. 226–244.

[38] D. J. Rogers and T. T. Tanimoto, *A computer program for classifying plants*, Science, 132 (1960), pp. 1115–1118.

[39] P. Rychlỳ, *A lexicographer-friendly association score*, in Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN, 2008, pp. 6–9.

[40] G. Salton, *Developments in automatic text retrieval*, Science, 253 (1991), pp. 974–980.

[41] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, *Item-based collaborative filtering recommendation algorithms*, in Proceedings of the 10th International Conference on World Wide Web, 2001, ACM, New York, pp. 285–295.

[42] V. Satuluri and S. Parthasarathy, *Bayesian locality sensitive hashing for fast similarity search*, Proceedings of the VLDB Endowment, 5 (2012), pp. 430–441.

[43] A. Shmida and M. V. Wilson, *Biological determinants of species diversity*, J. Biogeogr., 12 (1985), pp. 1–20.

[44] P. Sneath and R. Johnson, *The influence on numerical taxonomic similarities of errors in microbiological tests*, J. Gen. Microbiol., 72 (1972), pp. 377–392.

[45] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W. H. Freeman, San Francisco, 1973.

[46] T. Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons*, Biol. Skr., 5 (1948), pp. 1–34.

[47] J. Wang, H. T. Shen, J. Song, and J. Ji, *Hashing for Similarity Search: A Survey*, preprint, https://arxiv.org/abs/1408.2927, 2014.

[48] R. H. Whittaker, *Evolution and measurement of species diversity*, Taxon, (1972), pp. 213–251.

[49] S. Williams, M. Goodfellow, G. Alderson, E. Wellington, P. Sneath, and M. Sackin, *Numerical classification of Streptomyces and related genera*, Journal of General Microbiology, 129 (1983), pp. 1743–1813.

[50] S. M. Wong, W. Ziarko, and P. C. Wong, *Generalized vector spaces model in information retrieval*, in Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, 1985, pp. 18–25.