The Distortion of Locality Sensitive Hashing

Flavio Chierichetti[*1], Ravi Kumar[2], Alessandro Panconesi[*3], and Erisa Terolli[*4]

1 **Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy**
`flavio@di.uniroma1.it`
2 **Google, Mountain View, CA**
`ravi.k53@gmail.com`
3 **Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy**
`ale@di.uniroma1.it`
4 **Dipartimento di Informatica, Sapienza University of Rome, Rome, Italy**
`terolli@di.uniroma1.it`

Abstract

Given a pairwise similarity notion 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 a similarity that is LSHable. We consider several well-known non-LSHable similarities and show tight upper and lower bounds on their distortion.

1998 ACM Subject Classification F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY

Keywords and phrases Locality sensitive hashing, Distortion, Similarity

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

The notion of similarity finds its use in a large variety of fields above and beyond computer science. Often, the notion is tailored to the actual domain and the 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

* These authors were partially supported by a Google Focused Research Award, by the ERC Starting Grant DMAP 680153, and by the SIR Grant RBSI14Q743.

successful in a variety of applications dealing with large volumes of data, from near-duplicate estimation in text corpora to 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 [12] Charikar introduced two necessary criteria (the former weaker than the latter) for a similarity S to admit an 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 [15] for definitions).

This brings us to a very natural question, and 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 on 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 cosine similarity is $\Theta(\sqrt{n})$ and that of Braun–Blanquet and Sørensen–Dice similarities is two (the full picture is given in Table 1).

Our framework also greatly expands the outreach of the tests (T1) and (T2). We demonstrate its applicability by means of a few notable examples, in particular the Braun–Blanquet similarity whose distortion is proven to be exactly two. This similarity is particularly noteworthy because not only it passes test (T1) but also (T2). To show this we prove that this similarity is embeddable isometrically in ℓ_1, a result that may be of independent interest. Besides the two general methods discussed, we also provide ad hoc distortion bounds for Sokal–Sneath 1 and Forbes similarities.

Of the two methods introduced in our work, the center method is easier to apply than the k-sets method. The former is applicable to many instances of similarity but the latter is unavoidable in the following sense. Braun–Blanquet similarity not only, as remarked, passes (T1) and (T2), but also the test provided by the center method. However, the more powerful k-sets method can instead be used to show a distortion bound of two. Other similarities to which the k-sets method applies are Sørensen–Dice and the family $\textsc{sorensen}_\gamma$.

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

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 [8])
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 [13])
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 [13])
Cosine	$\frac{X\cdot Y}{\ell_2(X)\cdot\ell_2(Y)}$	$\sqrt{n}$ (Theorem 6)	$3\sqrt{n}$ (Theorem 7)
Simpson	$\frac{\lvert X\cap Y\rvert}{\min\{\lvert X\rvert,\lvert Y\rvert\}}$	n (Theorem 5)	n (Shingles [8])
Braun–Blanquet	$\frac{\lvert X\cap Y\rvert}{\max\{\lvert X\rvert,\lvert Y\rvert\}}$	2 (Theorem 16)	2 (Shingles [8])
Sørensen–Dice	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+1/2\lvert X\triangle Y\rvert}$	2 (Theorem 5)	2 (Shingles [8])
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 8)	2 (RSS [13])
Forbes	$\frac{n\lvert X\cap Y\rvert}{\lvert X\rvert\lvert Y\rvert}$	n (Theorem 20)	n (Theorem 20)
SORENSEN$_\gamma$	$\frac{\lvert X\cap Y\rvert}{\lvert X\cap Y\rvert+\gamma\lvert X\triangle Y\rvert}$	$\max(1,1/\gamma)$ (Theorem 5)	$\max(1,1/\gamma)$ (Shingles [8], RSS [13])
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 8)	$\max(1,1/\gamma)$ (RSS [13])

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.

Upper bounds: worst-case vs. 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 reasons 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 [13] 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 [8, 9, 25, 26]. Broder et al. [8, 9] showed that min-wise independent permutations form an LSH for the Jaccard similarity. Indyk and Motwani [25] introduced sampling hash as an LSH scheme for the Hamming similarity. Pursuing the work of characterizing similarities that admit an LSH, Charikar [12] introduced (T1) and (T2) as necessary criteria. Chierichetti and Kumar [13] proposed the concept of LSH-preserving functions, which are probability generating functions that preserve the LSH property of a similarity. From applications point of view, LSH has been widely used for solving the approximate or exact near-neighbor search [2] and similarity search [22, 30, 38] 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. [43].

Similarities are extensively used in various areas of computer science. Hamming similarity, for instance, is widely used in information theory [5, 6, 18]. Areas like data mining and data management have seen the usage of Anderberg similarity [1], Cosine similarity [10, 37], and Sokal–Sneath [40] similarity. Cosine similarity is also ubiquitous in information retrieval [21, 32, 36, 46] and bioinformatics [11] whereas Sokal–Sneath is used in image processing [4]. We should note here that the success of similarity algorithms/functions is not limited only within computer science. For instance, Sørensen–Dice is commonly used in ecology [16, 28, 29], phytosociology [27, 42], plant taxonomy [44], biology [39] and even in lexicography [35]. Biology has also seen the usage of Sokal–Sneath [41, 45], mentioned above. Other interesting examples are Simpson similarity used in microscopy [31] and biology [17], Braun–Blanquet in phytosociology [7] and ecology [33], and Rogers–Tanimoto used in taxonomy [34].

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. [24]). Examples include embeddings into the Euclidean (ℓ_2) metric, into the ℓ_1 metric, or into tree metrics from 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 "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 [15] 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 a locality sensitive hash (LSH).

▶ **Definition 1** (LSH [12]). 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 [25] for a somewhat different definition of LSH in the same spirit.) A similarity is *LSHable* if there exists an LSH for it. The basic notion we introduce in this paper is defined next.

▶ **Definition 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 $X, Y \in \mathcal{U}$,

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

The next lemma however implies that Definition 2 can be adopted without loss of generality.

▶ **Lemma 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 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 $h'(X) = (0, h(X))$ for each $X \in \mathcal{U}$, while with probability $1 - \gamma$, let $h'(X) = (1, X)$, 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 to S' we obtain an LSH for the similarity $S''(X,Y) = \frac{1}{\beta} \cdot S'(X,Y)$ (when $X \neq Y$) which 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 2 is robust.

Known LSH for set similarities. Set similarities are those similarities whose universe $\mathcal{U}$ satisfies $\mathcal{U} = 2^U$, for some finite ground set U. To give upper bounds on the distortions

[2] A minimum δ exists because it is equal to the solution of a linear program (see, e.g., [14]) of size polynomial in $|\mathcal{U}|$.

of various similarities we employ a number of LSH schemes for set similarities proposed in the literature. First and foremost, we employ *shingles* [8,9], which is an LSH scheme for the Jaccard similarity over sets ($\textsc{jaccard}(X,Y) = |X \cap Y|/|X \cup Y|$), over the 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. (And, $h(\varnothing)$ 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 [12] for the weighted Jaccard similarity. Finally, we use some of the LSH schemes given in [13] 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, S_γ and 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.** *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 $Y \in \mathcal{U}$, there exists at least one $X^\star \in \mathcal{X}$ such that $S(X^\star, Y) \leq 1/|\mathcal{X}|$.*

Proof. Let $\mathcal{H}$ be the hash function family for S. Observe that no LSH with a finite distortion can assign a non-zero probability to any of the pairs in $\binom{\mathcal{X}}{2}$, since their pairwise similarities are zero. Therefore each $Y \in \mathcal{U}$ can be hashed to the same value of at most one element of $\mathcal{X}$ for each hash function. In other words, each hash function $h \in \mathcal{H}$ must satisfy $|\{h(X) \mid X \in \mathcal{X}\}| = |\mathcal{X}|$. Therefore,

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

By averaging, it follows that there must exists at least one $X^\star \in \mathcal{X}$ such that $S(X^\star, Y) \leq 1/|\mathcal{X}|$. ◀

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 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 $|X|, |Y| \geq 1$, as $\textsc{simpson}(X, \varnothing) = 0$ if $|X| \geq 1$ 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 $|X| + |Y| \geq 1$, 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 5.** *For a ground set of n elements,*

$$\begin{aligned} \mathsf{distortion}(\textsc{simpson}) &= n, \text{ and} \\ \mathsf{distortion}(\textsc{sorensen}_\gamma) &= \max(1/\gamma, 1) - O(1/n). \end{aligned}$$

Proof. First, we show the lower bound by exhibiting an instance on a universe 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 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 $S(X_i, Y) \leq \frac{1}{|\mathcal{X}|} = \frac{1}{n}$. 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 [8] and minwise independent permutations [9] are a well-known LSH scheme for Jaccard similarity (see § 2). We use this to prove matching upper bounds for Theorem 5.

Minwise independent permutations form an LSH scheme with distortion n for Simpson similarity since

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

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

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

Finally, recall that a result in [13] proves that the similarity H_γ admits an LSH scheme as long as $\gamma \geq 1$. ◀

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

4.2 Cosine and unit ℓ_p-norm dot product

Recall that given any $p \geq 1$, 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 non-negative vectors $x, y \in \mathbf{R}^n_+$ having unit ℓ_2 norm is $\sum_{i=1}^n x(i) \cdot y(i)$.

Furthermore, given $p \geq 1$, let

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

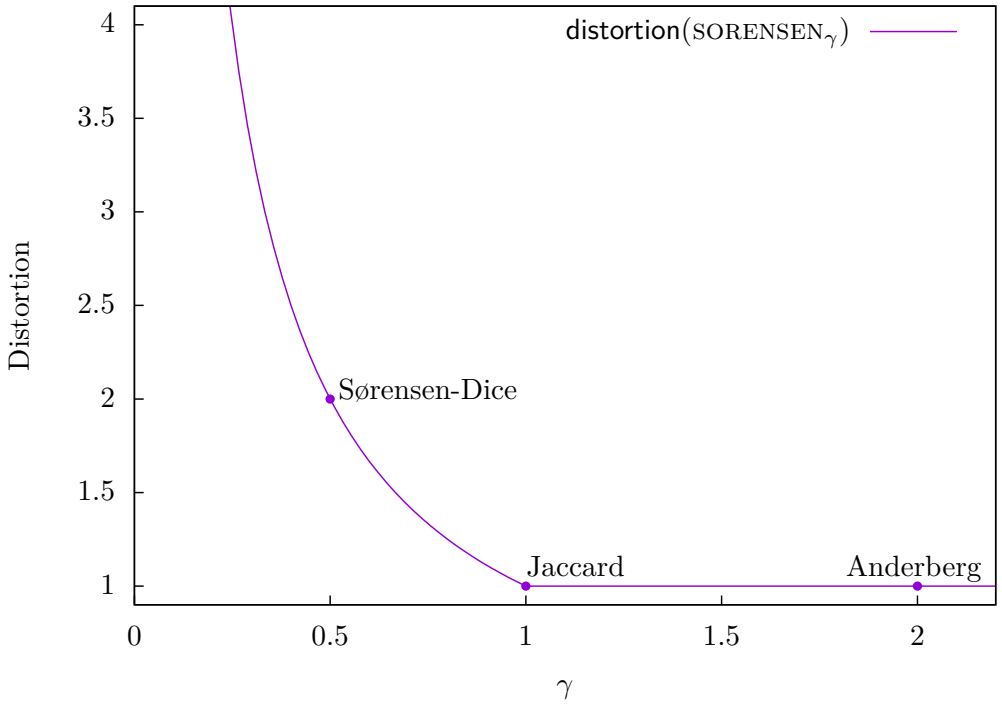


Figure 1 The minimum distortion of SORENSEN$_\gamma$.

be, respectively, the set of points contained in the p-ball of p-radius 1 with non-negative coordinates and the set of points lying on the p-sphere of p-radius 1 with non-negative 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 $\text{DOT}_{p,n} : F_{p,n} \times F_{p,n} \to [0, \infty)$ is

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

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

▶ **Theorem 6.** *For $p \geq 2$,* $\text{distortion}(\text{DOT}_{p,n}) \leq 3n^{1-\frac{1}{p}}$.

Proof. We first define two hash schemes and combine them to obtain an LSH for the ℓ_p-norm dot product.

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 $x \neq y$,

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

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; moreover,

(ii) we will furthermore add to the vector $f(x)$ one coordinate for each element of $F_{p,n}$; the value of $f(x)$ in the 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, by definition, $\ell_1(f(x)) = n^{1-\frac{1}{p}}$. We now apply the LSH scheme of [12] for weighted Jaccard to $f(F_{p,n})$. For $x \neq y$, let $\overline{C} = [n] \setminus C$ be the set of coordinates where both x and y have value greater than $n^{-\frac{1}{p}}$. Observe that, in any coordinate which is not in $\overline{C}$, at least one of $f(x)$ and $f(y)$ has a value of 0. Then, we have:

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

Observe that

$$\sum_{i \in [n]} \min(x(i), y(i)) \leq \ell_1(x) \leq \ell_p(x) \cdot n^{1-\frac{1}{p}} \leq n^{1-\frac{1}{p}},$$

and thus,

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

where the last inequality follows from the fact that, for each $i \in \overline{C}$, the largest of $x(i)$ and $y(i)$ is at least $n^{-\frac{1}{p}}$. Therefore, when $p \geq 2$, we have that

$$\max(x(i), y(i)) \geq n^{-\frac{1}{p}} \geq n^{\frac{1}{p}-1}.$$

Moreover,

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

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

$$\frac{1}{3n^{1-\frac{1}{p}}} \cdot \sum_{i\in[n]} x(i)\cdot y(i) \le \Pr[h(x)=h(y)] \le \sum_{i\in[n]} x(i)\cdot y(i).$$

Thus, there exists an LSH for a similarity that is within distortion $3n^{1-\frac{1}{p}}$ of the dot product similarity on non-negative vectors having ℓ_p norm at most 1. ◀

Now we show that the distortion of Theorem 6 is close to optimal using, once again, the center method.

▶ **Theorem 7.** *For $p \ge 1$, $\mathsf{distortion}(\text{DOT}_{p,n}) \ge n^{1-\frac{1}{p}}$.*

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 $X = \{u_1, u_2, \ldots, u_n\}$. Observe that for each $x \in X$, we have $\ell_p(x) = 1$ and $\ell_p(u_\star) = 1$.

Suppose that S is an LSHable similarity that distorts $\text{DOT}_{p,n}$ by the minimum possible amount. Since $S(u_i, u_j) = 0$ for every $i \ne j$, by Theorem 4 we know that there exists $u_i \in X$ such that $S(u_i, u_\star) \le \frac{1}{n}$. Since $\text{DOT}_{p,n}(u_i, u_\star) = n^{-\frac{1}{p}}$, 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 $p \ge 2$. We conjecture that it is generally tight for all $p \ge 1$, i.e., that Theorem 6 could be strengthened to all $p \ge 1$.

4.3 Sokal–Sneath similarities

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

$$\text{SOKAL-SNEATH}_\gamma(X,Y) = \frac{|X\cap Y| + \left|\overline{X\cup Y}\right|}{|X\cap Y| + \left|\overline{X\cup Y}\right| + \gamma\,|X \triangle Y|}.$$

Observe that SOKAL-SNEATH_1 is the Hamming similarity, $\text{SOKAL-SNEATH}_{1/2}$ is the Sokal–Sneath 1 similarity, and SOKAL-SNEATH_2 is the Rogers–Tanimoto similarity.

Rational set similarities [13] prove that $\text{SOKAL-SNEATH}_\gamma$ has an LSH iff $\gamma \ge 1$. 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 $\text{SOKAL-SNEATH}_\gamma$.

▶ **Theorem 8.** *For any $0 < \gamma < 1$,*

$$\frac{2}{1+\gamma} \le \mathsf{distortion}(\text{SOKAL-SNEATH}_\gamma) \le \frac{1}{\gamma}.$$

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

$$\text{SOKAL-SNEATH}_\gamma(X,Y) = \frac{1/2}{1/2 + \gamma/2} = \frac{1}{1+\gamma}.$$

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, there must exists $X^\star \in \{X, X'\}$ such that $S(X^\star, Y) \le 1/2$. 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 $\text{SOKAL-SNEATH}_\gamma$ with SOKAL-SNEATH_1 by introducing a distortion of $1/\gamma$. Since SOKAL-SNEATH_1 admits an LSH [13], it follows that $\mathsf{distortion}(\text{SOKAL-SNEATH}_\gamma) \le 1/\gamma$. ◀

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 tool is the following theorem. Let $\mathcal{U}_{n,k}$ denote $\binom{[n]}{k}$.

▶ **Theorem 9.** *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*

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

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 9, 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 (§ 5.1) to prove Theorem 9 (§ 5.2) and then apply it, in § 5.3, to Braun–Blanquet similarity, establishing optimal distortion bounds for it. We conclude with a discussion on the error term in Theorem 9 (§ 5.4). We remark that this "k-sets method" applies to other similarities such as Sørensen–Dice and SORENSEN$_\gamma$, for which the simpler center method has already been shown to give optimal results (§ 4). By contrast, we show (§ 6) that neither the center method, nor (T1) nor (T2) (see § 1) can be used to lower bound the distortion of Braun–Blanquet.

5.1 Extremal partitions

A hash function h on $\mathcal{U}$ naturally induces a partition in the following sense: two objects $X, Y \in \mathcal{U}$ belong to the same side of the partition if $h(X) = h(Y)$. 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 10** (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 11** (Nice partition)**.** An acceptable partition $\mathcal{P}$ of $\mathcal{U}_{n,k}$ is *nice* if it contains n parts $P_1, \ldots, P_t$, and if there exists a partition $I_1, \ldots, I_t$ of $[n]$ (with $I_i \neq \varnothing$ for $i \in [t]$, $\cup_{i=1}^t I_i = [n]$ and $I_i \cap I_j = \varnothing$ for each $\{i,j\} \in \binom{[t]}{2}$) 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 the above bound; we will then reduce any partition to a nice one.

▶ **Lemma 12.** *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 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, also, $a_i=|I_i|\geq 1$ and $b_i=\sum_{j=1}^{i-1}|I_j|$. Then, we have

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

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

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

◀

We now make use of the following theorem of Hilton and Milner [23] (see [20] for a short proof), which bounds the maximum cardinality of an Erdös–Ko–Rado [19] family that is not a star.

▶ **Theorem 13** (Hilton–Milner [23])**.** *Let $\mathcal{F}\subseteq\mathcal{U}_{n,k}$ be a family of sets with pairwise non-empty intersection with $n\geq 2k$. If $\bigcap_{F\in\mathcal{F}}F=\varnothing$ then $|\mathcal{F}|\leq\binom{n-1}{k-1}-\binom{n-k-1}{k-1}+1$.*

▶ Fact 14. $\binom{n-1}{k-1}-\binom{n-k-1}{k-1}+1\leq O\left(k\cdot\frac{n^{k-2}}{(k-2)!}\right)$.

▶ **Lemma 15.** *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 $\mathcal{P}$.

Let $\hat{\mathcal{P}}=\{P_i\mid\bigcap_{X\in P_i}X=\varnothing\}$, $\hat{p}=\sum_{P_i\in\hat{\mathcal{P}}}p_i$ and $\hat{m}=\sum_{P_i\in\hat{\mathcal{P}}}m_i$. If $P_i\in\hat{\mathcal{P}}$, Theorem 13 entails that $p_i\leq O\left(k\frac{n^{k-2}}{(k-2)!}\right)$. Thus, $m_i\leq O\left(p_i^2\right)\leq O\left(p_ik\frac{n^{k-2}}{(k-2)!}\right)$. Therefore,

$$\hat{m}=\sum_{P_i\in\hat{\mathcal{P}}}m_i\leq\sum_{P_i\in\hat{\mathcal{P}}}O\left(p_ik\frac{n^{k-2}}{(k-2)!}\right)\leq O\left(\hat{p}k\frac{n^{k-2}}{(k-2)!}\right)\leq O\left(\frac{n^{2k-2}}{(k-1)!\cdot(k-2)!}\right)=M,$$

where by definition $\hat{p}\leq\sum_{i=1}^{t}p_i=\binom{n}{k}\leq O\left(\frac{n^k}{k!}\right)$. 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 $m\leq m'+M$. Wlog, let $\mathcal{P}'=\{P'_1,\ldots,P'_{t'}\}$ and $|P'_1|\geq|P'_2|\geq\cdots\geq|P'_{t'}|$. Observe that, for each P'_i, we have $\bigcap_{X\in P'_i}X\neq\varnothing$.

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 $\mathcal{P}'_0,\mathcal{P}'_1,\ldots,\mathcal{P}'_\ell$ of acceptable partitions, (i) with $\mathcal{P}'_\ell$ being a nice partition, (ii) with $\mathcal{P}'_i$ having m'_i pairs, and with $m'_0\leq m'_1\leq\cdots\leq m'_\ell$. We stop executing the greedy selection rule as soon as we hit a nice partition, so property (i) is trivial.

Algorithm 1 A greedy selection rule.

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

Observe that in every successful iteration, moving a set from Q_j to Q_i, the number of pairs gets reduced by $|Q_j| - 1$, but it gets increased by $|Q_i|$. Since the algorithm selects $i < j$ it will hold $|Q_i| \geq |Q_j|$, and therefore the total number of pairs increases by at least one unit, hence $m'_{i+1} > m'_i$, and property (ii) has been proved.

Returning to our main goal, we have that $m \leq m' + M \leq m'_\ell + M$, where m'_ℓ is the value of a nice partition. By Lemma 12, we have $m'_\ell \leq \frac{n^{2k-1}}{2\cdot(2k-1)\cdot((k-1)!)^2}$. Moreover, $M = O\left(\frac{n^{2k-2}}{(k-1)!\cdot(k-2)!}\right)$. Thus,

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

5.2 Proof of Theorem 9

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. It is equal to:

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

Recall that, in general, we have that

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

Substituting k 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, for otherwise we would have infinite distortion. Hence, we can only use acceptable partitions. Suppose that the S has an LSH and assume wlog that this LSH gives positive probabilities $p_1, \ldots, p_h > 0$ to partitions $P_1, \ldots, P_h$, and that it gives probability 0 to other partitions. Let $v_1, \ldots, v_h$ be the values of partitions $P_1, \ldots, P_h$, and observe that $\sum_{i=1}^{h} p_i = 1$. Then, we have

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

i.e., the total amount of similarity mass that an acceptable partition brings to our similarity's values is equal to the probability that the LSH assigns to the partition times the number of the partition's pairs, equivalently, its own value. By Lemma 15, 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, $\sigma \leq \sum_{i=1}^{h} (\tau p_h) = \tau$. I.e., if S admits an LSH, then $\tau \geq \sigma$. Thus, we must have

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

which implies

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

5.3 The distortion of Braun–Blanquet

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

$$\textsc{braun-blanquet}(X, Y) = \frac{|X \cap Y|}{\max\left(|X|, |Y|\right)},$$

if $|X| + |Y| \geq 1$, and $\textsc{braun-blanquet}(X, Y) = 1$ 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 9 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, by selecting $k = \Theta\left(n^{1/3}\right)$, and by letting n grow to infinity, their distortion is at least $2 - \Theta\left(n^{-1/3}\right)$. 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 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 [8, 9] achieve a distortion of $2 - \Theta\left(n^{-1}\right)$ for Braun–Blanquet. Thus, we have the following theorem:

▶ **Theorem 16.** $\text{distortion}(\textsc{braun-blanquet}) = 2 - o(1)$.

5.4 Tightness of Theorem 9

We do not know whether the error term of Theorem 9 is tight. Here, we give a lower bound on that error term.

▶ **Lemma 17.** *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 $\frac{1}{2k-1} + \Omega\left(\binom{n}{2k-1}^{-1}\right)$, 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 § 4 is useless in determining the distortion of BRAUN-BLANQUET.

On the other hand, we know from Theorem 16 that its distortion is $2 - o(1)$. Thus, the k-sets method is the only method known to gauge its distortion.

6.1 ℓ_1-embeddability

▶ **Lemma 18.** $1 -$ 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 if there exists a non-negative 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 $Y_{i,c} \subseteq U$ be defined as

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

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 $1 \leq a \leq b$, 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 non-empty 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 wlog. that $|X| \leq |X'|$. Then,

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

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

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

The proof is concluded. ◀

6.2 Inapplicability of the center method

We prove the following Lemma, which shows the inapplicability of Theorem 4 to the case of the Braun–Blanquet similarity.

▶ **Lemma 19.** *For each $Y \subseteq [n]$, and for each $\mathcal{X} \subseteq 2^{[n]}$ such that* $\text{BRAUN-BLANQUET}(X, X') = 0$ *for all* $\{X, X'\} \in \binom{\mathcal{X}}{2}$*, it holds*

$$\text{avg}_{X \in \mathcal{X}} \text{BRAUN-BLANQUET}(X, Y) \leq \frac{1}{|\mathcal{X}|}.$$

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

$$\text{BRAUN-BLANQUET}(X, Y) \leq \frac{1}{|\mathcal{X}|}.$$

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:

$$\begin{aligned}\sum_{X \in \mathcal{X}} \text{BRAUN-BLANQUET}(X, Y) &= \sum_{X \in \mathcal{X}} \frac{|X \cap Y|}{\max(|X|, |Y|)} \\ &\leq \sum_{X \in \mathcal{X}} \frac{|X \cap Y|}{|Y|} \leq 1,\end{aligned}$$

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

$$\begin{aligned}&\sum_{X \in \mathcal{X}} \text{BRAUN-BLANQUET}(X, \varnothing) \\ &\quad\leq \sum_{\substack{X \in \mathcal{X} \\ X \neq \varnothing}} \frac{0}{\max(|X|, |Y|)} + \text{BRAUN-BLANQUET}(\varnothing, \varnothing) \\ &\quad= 1.\end{aligned}$$

Thus, in general, $\sum_{X \in \mathcal{X}} \text{BRAUN-BLANQUET}(X, Y) \leq 1$. It follows that,

$$\text{avg}_{X \in \mathcal{X}} \text{BRAUN-BLANQUET}(X, Y) \leq |\mathcal{X}|^{-1},$$

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 $\text{FORBES}(X, Y) = n \cdot \frac{|X \cap Y|}{|X|\,|Y|}$ if $|X|, |Y| \geq 1$, $\text{FORBES}(X, \varnothing) = 0$ if $|X| \geq 1$, and if $\text{FORBES}(\varnothing, \varnothing) = 1$. Since $F(\{1\}, \{1\}) = n$, we have the following simple observation.

▶ **Theorem 20.** $\text{distortion}(\text{FORBES}) = n$.

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

We give an LSH for the similarity FORBES/n, thus proving an upper bound of n on its distortion. The hash function h will be chosen as follows: $h(\varnothing) = \varnothing$ 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 reasons 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 [13] they are, but in a somewhat peculiar way, for 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 where 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 § 8.3 the documents underwent the following "cleaning" procedure: (i) all words not included in 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 2a 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 2b shows the distortion w.r.t. Sørensen–Dice. Table 2 displays the average distortion and the variance of these experiments.

Overall, these tests show that in real-world scenarios the average distortion of Braun–Blanquet and Sørensen–Dice can be significantly smaller than the worst case bound.

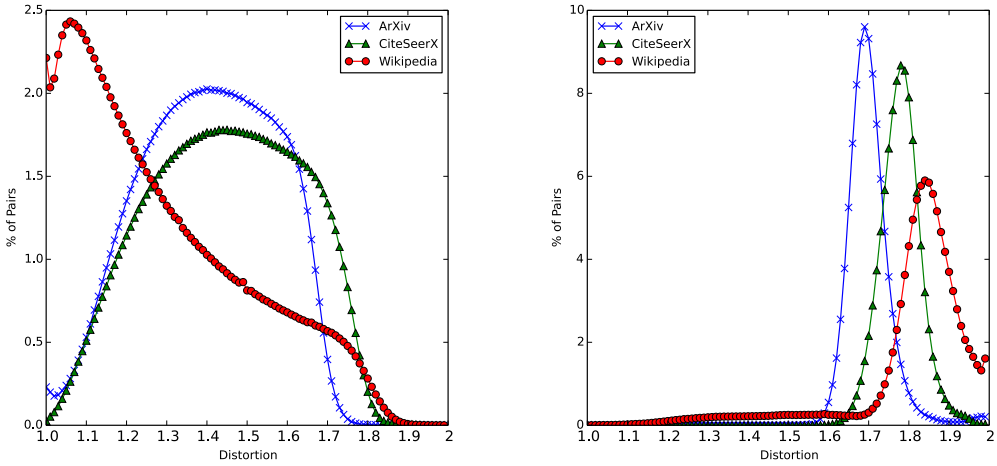
8.3 LSH schemes for rational set similarities

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,

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

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



(a) Braun–Blanquet similarity. **(b)** Sørensen–Dice similarity.

Figure 2 Percentage of document pairs with distortion δ with respect to shingles as δ increases.

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 [13] it is proven that the following is an LSH scheme for it. Pick a positive integer r at random with probability 2^{-r}. Let $h_1, \ldots, h_r$ be r shingles picked independently. Then, $h(X) := (h_1(X), \ldots, h_r(X))$ is an LSH scheme for A, i.e., $\textsc{anderberg}(X, Y) = \Pr[h(X) = h(Y)]$.

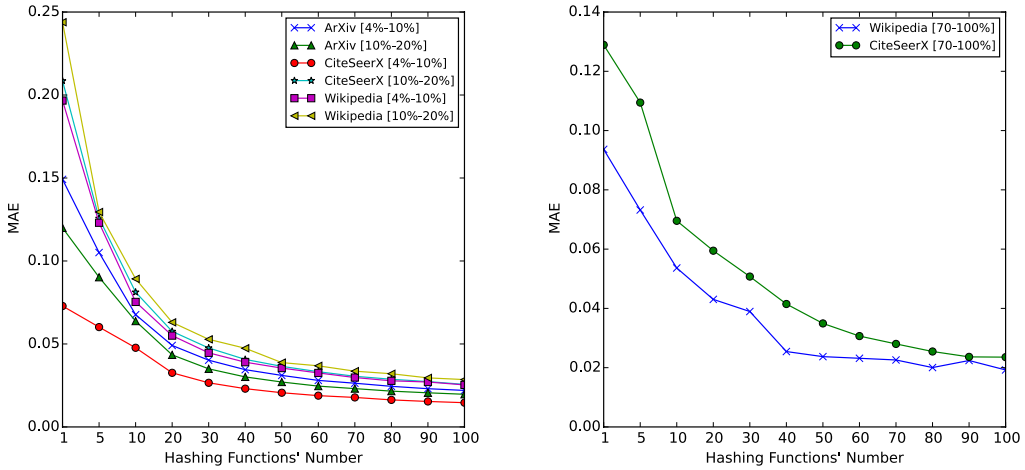
The Rogers–Tanimoto similarity is defined as

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

(Note that H_2 is the Rogers–Tanimoto similarity.) The following is the LSH scheme for Rogers–Tanimoto proposed in [13]. Pick r as before, and then pick r elements $e_1, e_2, \ldots, e_r$ from the ground set independently at random. The random hash function h is defined as follows. For a set X, we let $h(X) := (e_1 \in X, \ldots, e_r \in X)$ (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 $e = e_1, e_2, \ldots, e_r$, either both sets have it or they both do not).

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 binary vector of one thousand coordinates (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



(a) Anderberg similarity. **(b)** Rogers-Tanimoto similarity.

Figure 3 Mean Average Error as number of hash functions applied varies.

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 $i \in [k]$.

Figure 3a 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 time-wise and space-wise. Similar conclusions apply to Rogers–Tanimoto, as Figure 3b 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 locality sensitive hashing schemes for a number of widely-used similarities that do not admit exact such 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 dictated by the worst-case bounds.

It will be interesting to consider other non-LSHable similarities and study their distortion. The encyclopedia [15] is a rich source for such similarities.

Acknowledgments We thank the anonymous reviewers for several useful comments and suggestions.

References

1 Michael R Anderberg. *Cluster Analysis for Applications*. Academic Press, Inc., New York, 1973.

2 Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, pages 459–468, 2006.

3 Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *C. ACM*, 51(1):117–122, 2008.

4 Ismail Avcibaş, Mehdi Kharrazi, Nasir Memon, and Bülent Sankur. Image steganalysis with binary similarity measures. *EURASIP Journal on Applied Signal Processing*, 2005:2749–2757, 2005.

5 LR Bahl, J Cocke, F Jelinek, and J Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE TOIT*, 20(2):284–287, 1974.

6 Charles H. Bennett and Peter W. Shor. Quantum information theory. *IEEE TOIT*, 44(6):2724–2742, 1998.

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

8 A. Broder. On the resemblance and containment of documents. In *Proc. SEQUENCES*, pages 21–29, 1997.

9 Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *JCSS*, 60(3):630–659, 2000.

10 Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

11 Julie Chabalier, Jean Mosser, and Anita Burgun. A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8(1):235, 2007.

12 Moses Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. STOC*, pages 380–388, 2002.

13 Flavio Chierichetti and Ravi Kumar. LSH-preserving functions and their applications. In *SODA*, pages 1078–1094, 2012. URL: `http://dl.acm.org/citation.cfm?id=2095116.2095201`.

14 Flavio Chierichetti, Ravi Kumar, and Mohammad Mahdian. The complexity of LSH feasibility. *Theoretical Computer Science*, 530:89 – 101, 2014. URL: `http://www.sciencedirect.com/science/article/pii/S0304397514001467`, `doi:http://dx.doi.org/10.1016/j.tcs.2014.02.030`.

15 MichelMarie Deza and Elena Deza. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009. URL: `http://dx.doi.org/10.1007/978-3-642-00234-2_1`, `doi:10.1007/978-3-642-00234-2_1`.

16 Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

17 Sarah J Dixon, Nina Heinrich, Maria Holmboe, Michele L Schaefer, Randall R Reed, Jose Trevejo, and Richard 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. *Journal of Chemometrics*, 23(1):19–31, 2009.

18 William HR Equitz and Thomas M Cover. Successive refinement of information. *IEEE TOIT*, 37(2):269–275, 1991.

19 Paul Erdös, Chao Ko, and Richard Rado. Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford*, 12(2):313–320, 1961.

20 Péter Frankl and Zoltan Füredi. Non-trivial intersecting families. *JCT, Series A*, 41(1):150–153, 1986. URL: http://www.sciencedirect.com/science/article/pii/0097316586901214, doi:http://dx.doi.org/10.1016/0097-3165(86)90121-4.

21 George W Furnas, Scott Deerwester, Susan T Dumais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR*, pages 465–480, 1988.

22 Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *VLDB*, pages 518–529, 1999.

23 A.J.W. Hilton and E.C. Milner. Some intersection theorems for systems of finite sets. *Quart. J. Math. Oxford*, 18:369–384, 1967.

24 Piotr Indyk and Jiri Matousek. Low-distortion embeddings of finite metric spaces. *Handbook of Discrete and Computational Geometry*, pages 177–196, 2004.

25 Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, pages 604–613, 1998. URL: http://doi.acm.org/10.1145/276698.276876, doi:10.1145/276698.276876.

26 Piotr Indyk, Rajeev Motwani, Prabhakar Raghavan, and Santosh Vempala. Locality-preserving hashing in multidimensional spaces. In *STOC*, pages 618–625, 1997. URL: http://doi.acm.org/10.1145/258533.258656, doi:10.1145/258533.258656.

27 Dennis H Knight. A phytosociological analysis of species-rich tropical forest on Barro Colorado Island, Panama. *Ecological Monographs*, pages 259–284, 1975.

28 Patricia Koleff, Kevin J Gaston, and Jack J Lennon. Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, 72(3):367–382, 2003.

29 J Looman and JB Campbell. Adaptation of sorensen's k (1948) for estimating unit affinities in prairie vegetation. *Ecology*, pages 409–416, 1960.

30 Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *VLDB*, pages 950–961, 2007.

31 EMM Manders, FJ Verbeek, and JA Aten. Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy*, 169(3):375–382, 1993.

32 Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 775–780, 2006.

33 MED Poore. The use of phytosociological methods in ecological investigations: I. The Braun–Blanquet system. *The Journal of Ecology*, pages 226–244, 1955.

34 David J Rogers and Taffee T Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

35 Pavel Rychlỳ. A lexicographer-friendly association score. In *RASLAN*, pages 6–9, 2008.

36 Gerard Salton. Developments in automatic text retrieval. *Science*, 253(5023):974, 1991.

37 Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

38 Venu Satuluri and Srinivasan Parthasarathy. Bayesian locality sensitive hashing for fast similarity search. *VLDB*, 5(5):430–441, 2012.

39 AVI Shmida and Mark V Wilson. Biological determinants of species diversity. *Journal of Biogeography*, pages 1–20, 1985.

40 Peter H. A. Sneath and Robert R Sokal. *Numerical Taxonomy: The principles and practice of numerical classification.* W. H. Freeman, 1973.

41 PHA Sneath and R Johnson. The influence on numerical taxonomic similarities of errors in microbiological tests. *Journal of General Microbiology*, 72(2):377–392, 1972.

42 Thorvald 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.*, 5:1–34, 1948.

43 Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. Technical report, CoRR abs/1408.2927, 2014.

44 Robert H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972.

45 ST Williams, M Goodfellow, G Alderson, EMH Wellington, PHA Sneath, and MJ Sackin. Numerical classification of Streptomyces and related genera. *Journal of General Microbiology*, 129(6):1743–1813, 1983.

46 SK Michael Wong, Wojciech Ziarko, and Patrick CN Wong. Generalized vector spaces model in information retrieval. In *SIGIR*, pages 18–25, 1985.