

---

# In Defense of MinHash Over SimHash

---

**Anshumali Shrivastava**

Department of Computer Science  
Computing and Information Science  
Cornell University, Ithaca, NY, USA

**Ping Li**

Department of Statistics and Biostatistics  
Department of Computer Science  
Rutgers University, Piscataway, NJ, USA

## Abstract

MinHash and SimHash are the two widely adopted Locality Sensitive Hashing (LSH) algorithms for large-scale data processing applications. Deciding which LSH to use for a particular problem at hand is an important question, which has no clear answer in the existing literature. In this study, we provide a theoretical answer (validated by experiments) that MinHash virtually always outperforms SimHash when the data are binary, as common in practice such as search.

The collision probability of MinHash is a function of *resemblance* similarity ( $\mathcal{R}$ ), while the collision probability of SimHash is a function of *cosine* similarity ( $\mathcal{S}$ ). To provide a common basis for comparison, we evaluate retrieval results in terms of  $\mathcal{S}$  for both MinHash and SimHash. This evaluation is valid as we can prove that MinHash is a valid LSH with respect to  $\mathcal{S}$ , by using a general inequality  $\mathcal{S}^2 \leq \mathcal{R} \leq \frac{\mathcal{S}}{2-\mathcal{S}}$ . Our **worst case** analysis can show that MinHash significantly outperforms SimHash in **high similarity** region.

Interestingly, our intensive experiments reveal that MinHash is also substantially better than SimHash even in datasets where most of the data points are not too similar to each other. This is partly because, in practical data, often  $\mathcal{R} \geq \frac{\mathcal{S}}{z-\mathcal{S}}$  holds where  $z$  is only slightly larger than 2 (e.g.,  $z \leq 2.1$ ). Our **restricted worst case** analysis by assuming  $\frac{\mathcal{S}}{z-\mathcal{S}} \leq \mathcal{R} \leq \frac{\mathcal{S}}{2-\mathcal{S}}$  shows that MinHash indeed significantly outperforms SimHash even in **low similarity** region.

We believe the results in this paper will provide valuable guidelines for search in practice, especially when the data are sparse.

## 1 Introduction

The advent of the Internet has led to generation of massive and inherently high dimensional data. In many industrial applications, the size of the datasets has long exceeded the memory capacity of a single machine. In web domains, it is not difficult to find datasets with the number of instances and the number of dimensions going into billions [1, 6, 28].

The reality that web data are typically sparse and high dimensional is due to the wide adoption of the “Bag of Words” (BoW) representations for documents and images. In BoW representations, it is known that the word frequency within a document follows power law. Most of the words occur rarely in a document and most of the higher order shingles in the document occur only once. It is often the case that just the presence or absence information suffices in practice [7, 14, 17, 23]. Leading search companies routinely use sparse binary representations in their large data systems [6].

**Locality sensitive hashing (LSH)** [16] is a general framework of indexing technique, devised for efficiently solving the approximate near neighbor search problem [11]. The performance of LSH largely depends on the underlying particular hashing methods. Two popular hashing algorithms are **MinHash** [3] and **SimHash** (sign normal random projections) [8].

MinHash is an LSH for **resemblance similarity** which is defined over binary vectors, while SimHash is an LSH for **cosine similarity** which works for general real-valued data. With the abundance of binary data over the web, it has become a practically important question: *which LSH should be preferred in binary data?* This question has not been adequately answered in existing literature. There were prior attempts to address this problem from various aspects. For example, the paper on *Conditional Random Sampling (CRS)* [19] showed that random projections can be very inaccurate especially in binary data, for the task of inner product estimation (which is not the same as near neighbor search). A more recent paper [26] empirically demonstrated that  $b$ -bit minwise hashing [22] outperformed SimHash and spectral hashing [30].

---

Appearing in Proceedings of the 17<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

**Our contribution:** Our paper provides an essentially conclusive answer that MinHash should be used for near neighbor search in binary data, both theoretically and empirically. To favor SimHash, our theoretical analysis and experiments evaluate the retrieval results of MinHash in terms of cosine similarity (instead of resemblance). This is possible because we are able to show that MinHash can be proved to be an LSH for cosine similarity by establishing an inequality which bounds resemblance by purely functions of cosine.

Because we evaluate MinHash (which was designed for resemblance) in terms of cosine, we will first illustrate the close connection between these two similarities.

## 2 Cosine Versus Resemblance

We focus on binary data, which can be viewed as sets (locations of nonzeros). Consider two sets  $W_1, W_2 \subseteq \Omega = \{1, 2, \dots, D\}$ . The cosine similarity ( $\mathcal{S}$ ) is

$$\mathcal{S} = \frac{a}{\sqrt{f_1 f_2}}, \quad \text{where} \quad (1)$$

$$f_1 = |W_1|, \quad f_2 = |W_2|, \quad a = |W_1 \cap W_2| \quad (2)$$

The resemblance similarity, denoted by  $\mathcal{R}$ , is

$$\mathcal{R} = \mathcal{R}(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|} = \frac{a}{f_1 + f_2 - a} \quad (3)$$

Clearly these two similarities are closely related. To better illustrate the connection, we re-write  $\mathcal{R}$  as

$$\mathcal{R} = \frac{a/\sqrt{f_1 f_2}}{\sqrt{f_1/f_2} + \sqrt{f_2/f_1} - a/\sqrt{f_1 f_2}} = \frac{\mathcal{S}}{z - \mathcal{S}} \quad (4)$$

$$z = z(r) = \sqrt{r} + \frac{1}{\sqrt{r}} \geq 2 \quad (5)$$

$$r = \frac{f_2}{f_1} = \frac{f_1 f_2}{f_1^2} \leq \frac{f_1 f_2}{a^2} = \frac{1}{\mathcal{S}^2} \quad (6)$$

There are two degrees of freedom:  $f_2/f_1$ ,  $a/f_2$ , which make it inconvenient for analysis. Fortunately, in Theorem 1, we can bound  $\mathcal{R}$  by purely functions of  $\mathcal{S}$ .

### Theorem 1

$$\mathcal{S}^2 \leq \mathcal{R} \leq \frac{\mathcal{S}}{2 - \mathcal{S}} \quad (7)$$

**Tightness** Without making assumptions on the data, neither the lower bound  $\mathcal{S}^2$  or the upper bound  $\frac{\mathcal{S}}{2 - \mathcal{S}}$  can be improved in the domain of continuous functions.

**Data dependent bound** If the data satisfy  $z \leq z^*$ , where  $z$  is defined in (5), then

$$\frac{\mathcal{S}}{z^* - \mathcal{S}} \leq \mathcal{R} \leq \frac{\mathcal{S}}{2 - \mathcal{S}} \quad (8)$$

**Proof:** See Appendix A.  $\square$

Figure 1 illustrates that in high similarity region, the upper and lower bounds essentially overlap. Note that, in order to obtain  $\mathcal{S} \approx 1$ , we need  $f_1 \approx f_2$  (i.e.,  $z \approx 2$ ).

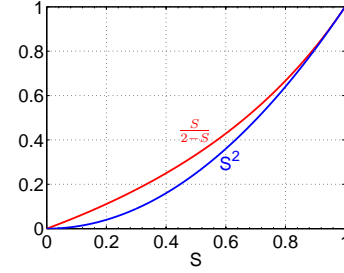


Figure 1: Upper (in red) and lower (in blue) bounds in Theorem 1, which overlap in high similarity region.

While the high similarity region is often of interest, we must also handle data in the low similarity region, because in a realistic dataset, the majority of the pairs are usually not similar. Interestingly, we observe that for the six datasets in Table 1, we often have  $\mathcal{R} = \frac{\mathcal{S}}{z - \mathcal{S}}$  with  $z$  only being slightly larger than 2; see Figure 2.

Table 1: Datasets

Dataset	# Query	# Train	# Dim
MNIST	10,000	60,000	784
NEWS20	2,000	18,000	1,355,191
NYTIMES	5,000	100,000	102,660
RCV1	5,000	100,000	47,236
URL	5,000	90,000	3,231,958
WEBSPPAM	5,000	100,000	16,609,143

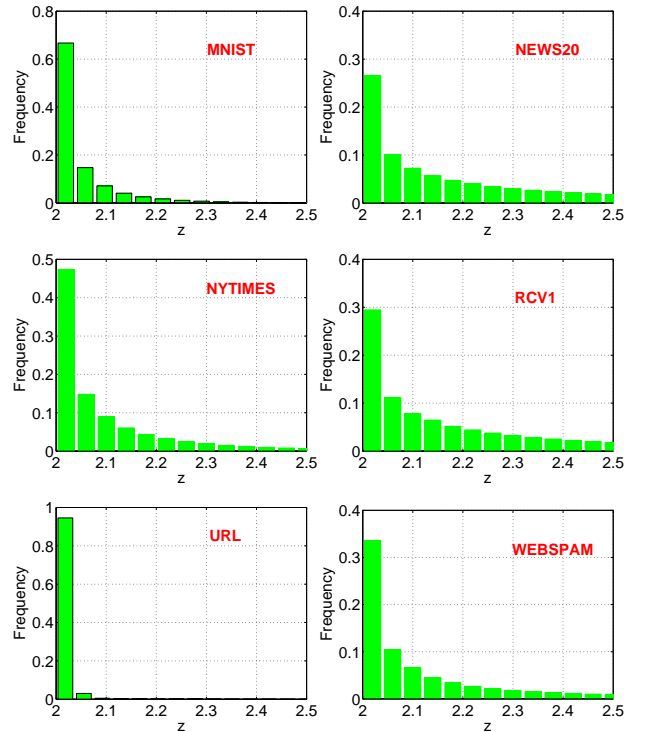


Figure 2: Frequencies of the  $z$  values for all six datasets in Table 1, where  $z$  is defined in (5). We compute  $z$  for every query-train pair of data points.

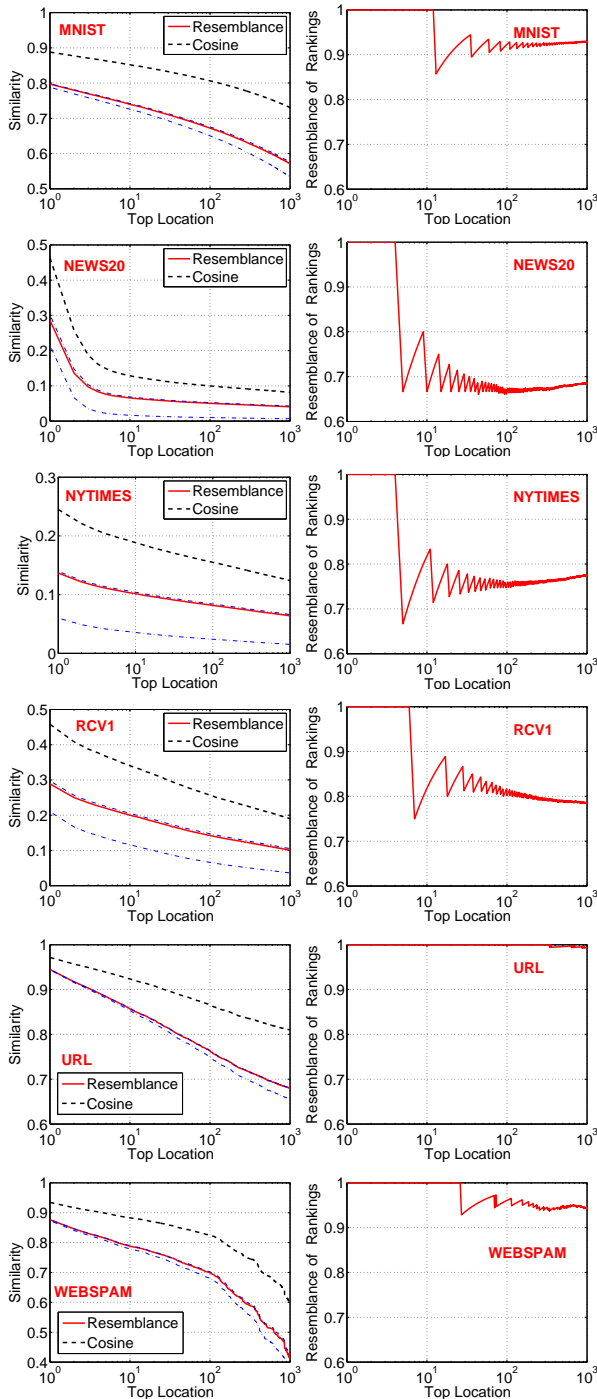


Figure 3: **Left panels:** For each query point, we rank its similarities to all training points in descending order. For every top location, we plot the median (among all query points) of the similarities, separately for cosine (dashed) and resemblance (solid), together with the lower and upper bounds of  $\mathcal{R}$  (dot-dashed). **Right Panels:** For every query point, we rank the training points in descending order of similarities, separately for cosine and resemblance. We plot the resemblance of two ranked lists at top- $T$  ( $T = 1$  to 1000).

For each dataset, we compute both cosine and resemblance for every query-train pair (e.g.,  $10000 \times 60000$  pairs for MNIST dataset). For each query point, we rank its similarities to all training points in descending order. We examine the top-1000 locations as in Figure 3. In the left panels, for every top location, we plot the median (among all query points) of the similarities, separately for cosine (dashed) and resemblance (solid), together with the lower and upper bounds of  $\mathcal{R}$  (dot-dashed). We can see for NEWS20, NYTIMES, and RCV1, the data are not too similar. Interestingly, for all six datasets,  $\mathcal{R}$  matches fairly well with the upper bound  $\frac{\mathcal{S}}{2-\mathcal{S}}$ . In other words, the lower bound  $\mathcal{S}^2$  can be very conservative even in low similarity region.

The right panels of Figure 3 present the comparisons of the orderings of similarities in an interesting way. For every query point, we rank the training points in descending order of similarities, separately for cosine and resemblance. This way, for every query point we have two lists of numbers (of the data points). We truncate the lists at top- $T$  and compute the resemblance between the two lists. By varying  $T$  from 1 to 1000, we obtain a curve which roughly measures the “similarity” of cosine and resemblance. We present the averaged curve over all query points. Clearly Figure 3 shows there is a strong correlation between the two measures in all datasets, as one would expect.

### 3 Locality Sensitive Hashing (LSH)

A common formalism for approximate near neighbor problem is the  $c$ -approximate near neighbor or  $c$ -NN.

**Definition:** ( $c$ -Approximate Near Neighbor or  $c$ -NN). Given a set of points in a  $d$ -dimensional space  $\mathbb{R}^d$ , and parameters  $S_0 > 0$ ,  $\delta > 0$ , construct a data structure which, given any query point  $q$ , does the following with probability  $1 - \delta$ : if there exist an  $S_0$ -near neighbor of  $q$  in  $P$ , it reports some  $cS_0$ -near neighbor of  $q$  in  $P$ .

The usual notion of  $S_0$ -near neighbor is in terms of the distance function. Since we are dealing with similarities, we can equivalently define  $S_0$ -near neighbor of point  $q$  as a point  $p$  with  $Sim(q, p) \geq S_0$ , where  $Sim$  is the similarity function of interest.

A popular technique for  $c$ -NN, uses the underlying theory of *Locality Sensitive Hashing* (LSH) [16]. LSH is a family of functions, with the property that similar input objects in the domain of these functions have a higher probability of colliding in the range space than non-similar ones. In formal terms, consider  $\mathcal{H}$  a family of hash functions mapping  $\mathbb{R}^D$  to some set  $\mathcal{S}$ .

**Definition: Locality Sensitive Hashing** A family  $\mathcal{H}$  is called  $(S_0, cS_0, p_1, p_2)$ -sensitive if for any two points  $x, y \in \mathbb{R}^d$  and  $h$  chosen uniformly from  $\mathcal{H}$  satisfies the following:

- if  $Sim(x, y) \geq S_0$  then  $Pr_{\mathcal{H}}(h(x) = h(y)) \geq p_1$
- if  $Sim(x, y) \leq cS_0$  then  $Pr_{\mathcal{H}}(h(x) = h(y)) \leq p_2$

For approximate nearest neighbor search typically,  $p_1 > p_2$  and  $c < 1$  is needed. Since we are defining neighbors in terms of similarity we have  $c < 1$ . To get distance analogy we can use the transformation  $D(x, y) = 1 - Sim(x, y)$  with a requirement of  $c > 1$ .

The definition of LSH family  $\mathcal{H}$  is tightly linked with the similarity function of interest  $Sim$ . An LSH allows us to construct data structures that give provably efficient query time algorithms for  $c$ -NN problem.

**Fact:** Given a family of  $(S_0, cS_0, p_1, p_2)$ -sensitive hash functions, one can construct a data structure for  $c$ -NN with  $O(n^\rho \log_{1/p_2} n)$  query time, where  $\rho = \frac{\log p_1}{\log p_2}$ .

The quantity  $\rho < 1$  measures the efficiency of a given LSH, the smaller the better. In theory, in the worst case, the number of points scanned by a given LSH to find a  $c$ -approximate near neighbor is  $O(n^\rho)$  [16], which is dependent on  $\rho$ . Thus given two LSHs, for the same  $c$ -NN problem, the LSH with smaller value of  $\rho$  will achieve the same approximation guarantee and at the same time will have faster query time. LSH with lower value of  $\rho$  will report fewer points from the database as the potential near neighbors. These reported points need additional re-ranking to find the true  $c$ -approximate near neighbor, which is a costly step. It should be noted that the efficiency of an LSH scheme, the  $\rho$  value, is dependent on many things. It depends on the similarity threshold  $S_0$  and the value of  $c$  which is the approximation parameter.

### 3.1 Resemblance Similarity and MinHash

Minwise hashing [4] is the LSH for resemblance similarity. The minwise hashing family applies a random permutation  $\pi : \Omega \rightarrow \Omega$ , on the given set  $W$ , and stores only the minimum value after the permutation mapping. Formally MinHash is defined as:

$$h_\pi^{min}(W) = \min(\pi(W)). \quad (9)$$

Given sets  $W_1$  and  $W_2$ , it can be shown by elementary probability argument that

$$Pr(h_\pi^{min}(W_1) = h_\pi^{min}(W_2)) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|} = \mathcal{R}. \quad (10)$$

It follows from (10) that minwise hashing is  $(\mathcal{R}_0, c\mathcal{R}_0, \mathcal{R}_0, c\mathcal{R}_0)$  sensitive family of hash function when the similarity function of interest is resemblance i.e  $\mathcal{R}$ . It has efficiency  $\rho = \frac{\log \mathcal{R}_0}{\log c\mathcal{R}_0}$  for approximate resemblance based search.

### 3.2 SimHash and Cosine Similarity

SimHash is another popular LSH for the cosine similarity measure, which originates from the concept of

*sign random projections (SRP)* [8]. Given a vector  $x$ , SRP utilizes a random vector  $w$  with each component generated from i.i.d. normal, i.e.,  $w_i \sim N(0, 1)$ , and only stores the sign of the projected data. Formally, SimHash is given by

$$h_w^{sim}(x) = \text{sign}(w^T x) \quad (11)$$

It was shown in [12] that the collision under SRP satisfies the following equation:

$$Pr(h_w^{sim}(x) = h_w^{sim}(y)) = 1 - \frac{\theta}{\pi}, \quad (12)$$

where  $\theta = \cos^{-1} \left( \frac{x^T y}{\|x\|_2 \|y\|_2} \right)$ . The term  $\frac{x^T y}{\|x\|_2 \|y\|_2}$ , is the cosine similarity for data vectors  $x$  and  $y$ , which becomes  $\mathcal{S} = \frac{a}{\sqrt{f_1 f_2}}$  when the data are binary.

Since  $1 - \frac{\theta}{\pi}$  is monotonic with respect to cosine similarity  $\mathcal{S}$ . Eq. (12) implies that SimHash is a  $\left( S_0, cS_0, \left( 1 - \frac{\cos^{-1}(S_0)}{\pi} \right), \left( 1 - \frac{\cos^{-1}(cS_0)}{\pi} \right) \right)$  sensitive hash function with efficiency  $\rho = \frac{\log \left( 1 - \frac{\cos^{-1}(S_0)}{\pi} \right)}{\log \left( 1 - \frac{\cos^{-1}(cS_0)}{\pi} \right)}$ .

## 4 Theoretical Comparisons

We would like to highlight here that the  $\rho$  values for MinHash and SimHash, shown in the previous section, are not directly comparable because they are in the context of different similarity measures. Consequently, it was not clear, before our work, if there is any theoretical way of finding conditions under which MinHash is preferable over SimHash and vice versa. It turns out that the two sided bounds in Theorem 1 allow us to prove MinHash is also an LSH for cosine similarity.

### 4.1 MinHash as an LSH for Cosine Similarity

We fix our gold standard similarity measure to be the cosine similarity  $Sim = \mathcal{S}$ . Theorem 1 leads to two simple corollaries:

**Corollary 1** *If  $\mathcal{S}(x, y) \geq S_0$ , then we have  $Pr(h_\pi^{min}(x) = h_\pi^{min}(y)) = \mathcal{R}(x, y) \geq S_0^2$*

**Corollary 2** *If  $\mathcal{S}(x, y) \leq cS_0$ , then we have  $Pr(h_\pi^{min}(x) = h_\pi^{min}(y)) = \mathcal{R}(x, y) \leq \frac{cS_0}{2-cS_0}$*

Immediate consequence of these two corollaries combined with the definition of LSH is the following:

**Theorem 2** *For binary data, MinHash is  $(S_0, cS_0, S_0^2, \frac{cS_0}{2-cS_0})$  sensitive family of hash function for cosine similarity with  $\rho = \frac{\log S_0^2}{\log \frac{cS_0}{2-cS_0}}$ .*

### 4.2 1-bit Minwise Hashing

SimHash generates a single bit output (only the signs) whereas MinHash generates an integer value. Recently

proposed  $b$ -bit minwise hashing [22] provides simple strategy to generate an informative single bit output from MinHash, by using the parity of MinHash values:

$$h_{\pi}^{\text{min},1\text{bit}}(W_1) = \begin{cases} 1 & \text{if } h_{\pi}^{\text{min}}(W_1) \text{ is odd} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

For 1-bit MinHash and very sparse data (i.e.,  $\frac{f_1}{D} \rightarrow 0$ ,  $\frac{f_2}{D} \rightarrow 0$ ), we have the following collision probability

$$Pr(h_{\pi}^{\text{min},1\text{bit}}(W_1) = h_{\pi}^{\text{min},1\text{bit}}(W_2)) = \frac{\mathcal{R} + 1}{2} \quad (14)$$

The analysis presented in previous sections allows us to theoretically analyze this new scheme. The inequality in Theorem 1 can be modified for  $\frac{\mathcal{R}+1}{2}$  and using similar arguments as for MinHash we obtain

**Theorem 3** For binary data, 1-bit MH (minwise hashing) is  $(S_0, cS_0, \frac{S_0^2+1}{2}, \frac{1}{2-cS_0})$  sensitive family of hash function for cosine similarity with  $\rho = \frac{\log \frac{2}{S_0^2+1}}{\log(2-cS_0)}$ .

### 4.3 Worst Case Gap Analysis

We will compare the gap ( $\rho$ ) values of the three hashing methods we have studied:

$$\text{SimHash: } \rho = \frac{\log\left(1 - \frac{\cos^{-1}(S_0)}{\pi}\right)}{\log\left(1 - \frac{\cos^{-1}(cS_0)}{\pi}\right)} \quad (15)$$

$$\text{MinHash: } \rho = \frac{\log S_0^2}{\log \frac{cS_0}{2-cS_0}} \quad (16)$$

$$\text{1-bit MH: } \rho = \frac{\log \frac{2}{S_0^2+1}}{\log(2-cS_0)} \quad (17)$$

This is a worst case analysis. We know the lower bound  $\mathcal{S}^2 \leq \mathcal{R}$  is usually very conservative in real data when the similarity level is low. Nevertheless, for high similarity region, the comparisons of the  $\rho$  values indicate that MinHash significantly outperforms SimHash as shown in Figure 4, at least for  $S_0 \geq 0.8$ .

### 4.4 Restricted Worst Case Gap Analysis

The worst case analysis does not make any assumption on the data. It is obviously too conservative when the data are not too similar. Figure 2 has demonstrated that in real data, we can fairly safely replace the lower bound  $\mathcal{S}^2$  with  $\frac{\mathcal{S}}{z-\mathcal{S}}$  for some  $z$  which, defined in (5), is very close to 2 (for example, 2.1). If we are willing to make this assumption, then we can go through the same analysis for MinHash as an LSH for cosine and compute the corresponding  $\rho$  values:

$$\text{MinHash: } \rho = \frac{\log \frac{S_0}{z-S_0}}{\log \frac{cS_0}{2-cS_0}} \quad (18)$$

$$\text{1-bit MH: } \rho = \frac{\log \frac{2(z-S_0)}{z}}{\log(2-cS_0)} \quad (19)$$

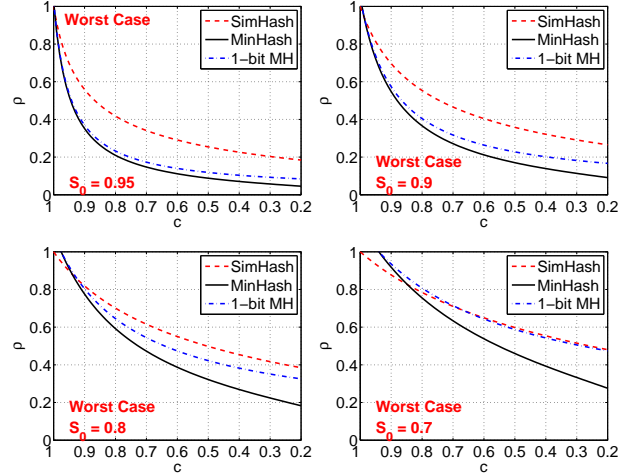


Figure 4: Worst case gap ( $\rho$ ) analysis, i.e., (15) (16) (17), for high similarity region; lower is better.

Note that this is still a worst case analysis (and hence can still be very conservative). Figure 5 presents the  $\rho$  values for this restricted worst case gap analysis, for two values of  $z$  (2.1 and 2.3) and  $S_0$  as small as 0.2. The results confirm that MinHash still significantly outperforms SimHash even in low similarity region.

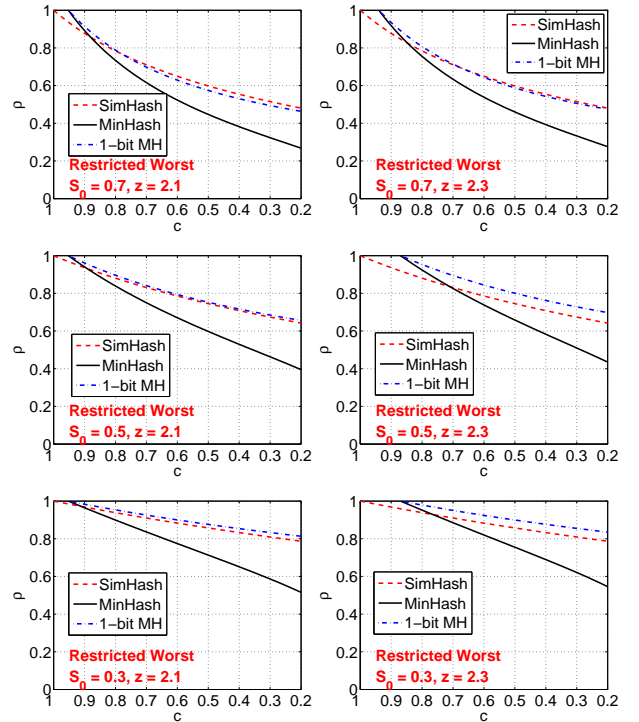


Figure 5: Restricted worst case gap ( $\rho$ ) analysis by assuming the data satisfy  $\frac{\mathcal{S}}{z-\mathcal{S}} \leq \mathcal{R} \leq \frac{\mathcal{S}}{2-\mathcal{S}}$ , where  $z$  is defined in (5). The  $\rho$  values for MinHash and 1-bit MinHash are expressed in (18) and (19), respectively.

Both Figure 4 and Figure 5 show that 1-bit MinHash

can be less competitive when the similarity is not high. This is expected as analyzed in the original paper of  $b$ -bit minwise hashing [20]. The remedy is to use more bits. As shown in Figure 6, once we use  $b = 8$  (or even  $b = 4$ ) bits, the performance of  $b$ -bit minwise hashing is not much different from MinHash.

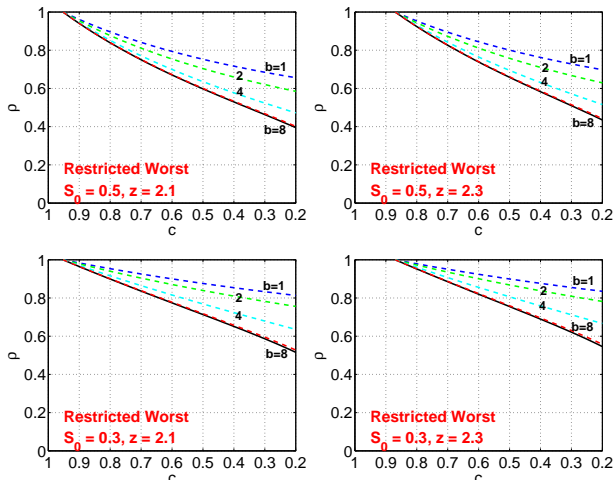


Figure 6: Restricted worst case gap ( $\rho$ ) analysis for  $b$ -bit minwise hashing for  $b = 1, 2, 4, 8$ .

### 4.5 Idealized Case Gap Analysis

The restricted worst case analysis can still be very conservative and may not fully explain the stunning performance of MinHash in our experiments on datasets of low similarities. Here, we also provide an analysis based on fixed  $z$  value. That is, we only analyze the gap  $\rho$  by assuming  $\mathcal{R} = \frac{S}{z-S}$  for a fixed  $z$ . We call this idealized gap analysis. Not surprisingly, Figure 7 confirms that, with this assumption, MinHash significantly outperform SimHash even for extremely low similarity. We should keep in mind that this idealized gap analysis can be somewhat optimistic and should only be used as some side information.

## 5 Experiments

We evaluate both MinHash and SimHash in the actual task of retrieving top- $k$  near neighbors. We implemented the standard  $(K, L)$  parameterized LSH [16] algorithms with both MinHash and SimHash. That is, we concatenate  $K$  hash functions to form a new hash function for each table, and we generate  $L$  such tables (see [2] for more details about the implementation). We used all the six binarized datasets with the query and training partitions as shown in Table 1. For each dataset, elements from training partition were used for constructing hash tables, while the elements of the query partition were used as query for top- $k$  neighbor search. For every query, we compute the gold standard top- $k$  near neighbors using the cosine similarity as the underlying similarity measure.

In standard  $(K, L)$  parameterized bucketing scheme

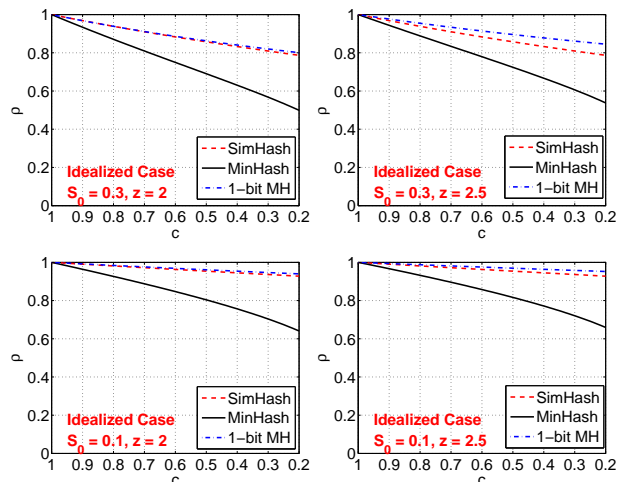


Figure 7: Idealized case gap ( $\rho$ ) analysis by assuming  $\mathcal{R} = \frac{S}{z-S}$  for a fixed  $z$  ( $z = 2$  and  $z = 2.5$  in the plots).

the choice of  $K$  and  $L$  is dependent on the similarity thresholds and the hash function under consideration. In the task of top- $k$  near neighbor retrieval, the similarity thresholds vary with the datasets. Hence, the actual choice of ideal  $K$  and  $L$  is difficult to determine. To ensure that this choice does not affect our evaluations, we implemented all the combinations of  $K \in \{1, 2, \dots, 30\}$  and  $L \in \{1, 2, \dots, 200\}$ . These combinations include the reasonable choices for both the hash function and different threshold levels.

For each combination of  $(K, L)$  and for both of the hash functions, we computed the mean recall of the top- $k$  gold standard neighbors along with the average number of points reported per query. We then compute the least number of points needed, by each of the two hash functions, to achieve a given percentage of recall of the gold standard top- $k$ , where the least was computed over the choices of  $K$  and  $L$ . We are therefore ensuring the best over all the choices of  $K$  and  $L$  for each hash function independently. This eliminates the effect of  $K$  and  $L$ , if any, in the evaluations. The plots of the fraction of points retrieved at different recall levels, for  $k = 1, 10, 20, 100$ , are in Figure 8.

A good hash function, at a given recall should retrieve less number of points. MinHash needs to evaluate significantly less fraction of the total data points to achieve a given recall compared to SimHash. MinHash is consistently better than SimHash, in most cases very significantly, irrespective of the choices of dataset and  $k$ . It should be noted that our gold standard measure for computing top- $k$  neighbors is cosine similarity. This should favor SimHash because it was the only known LSH for cosine similarity. Despite this “disadvantage”, MinHash still outperforms SimHash in top near neighbor search with cosine similarity. This nicely confirms our theoretical gap analysis.

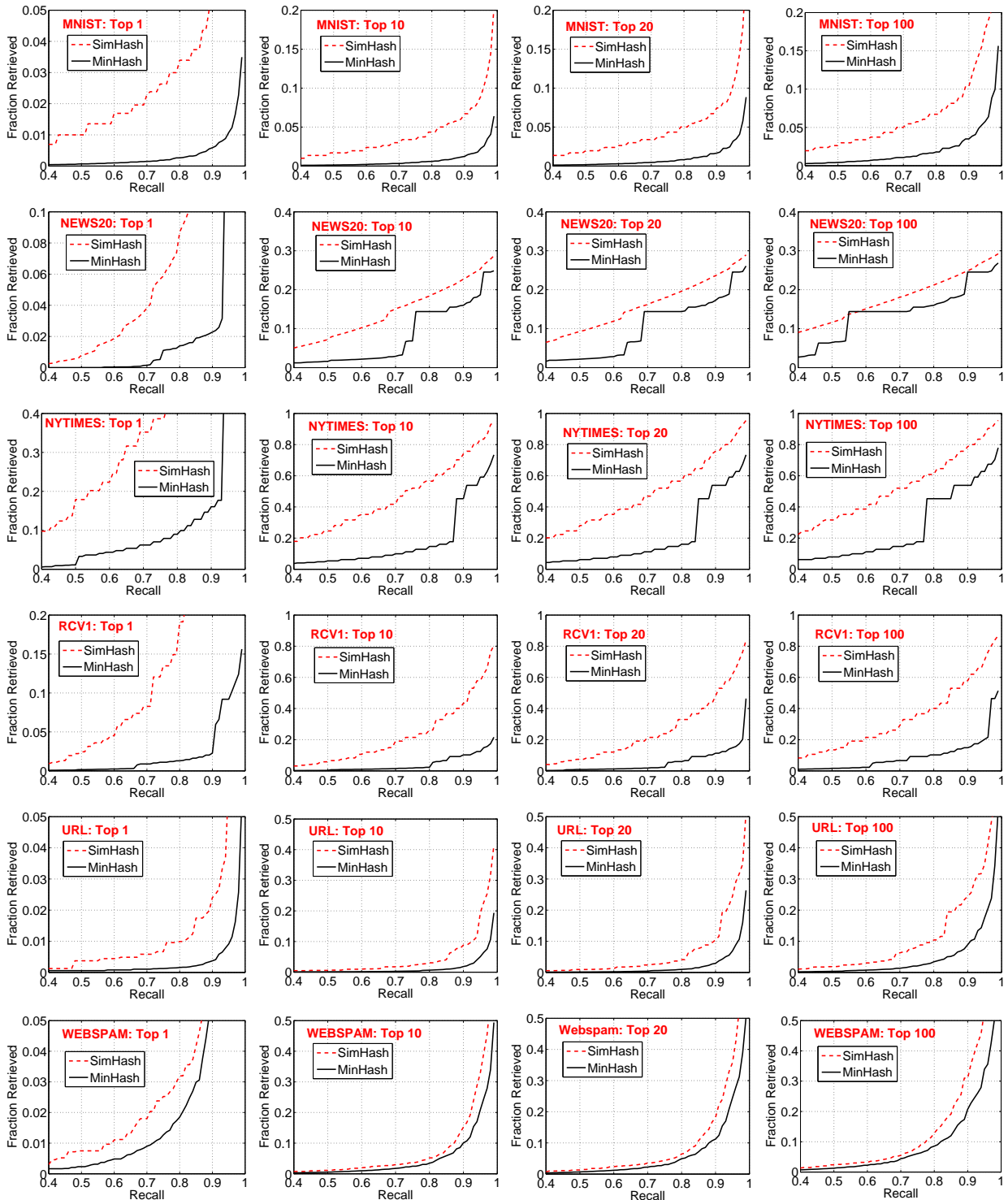


Figure 8: Fraction of data points retrieved (y-axis) in order to achieve a specified recall (x-axis), for comparing SimHash with MinHash. Lower is better. We use top- $k$  (cosine similarities) as the gold standard for  $k = 1, 10, 20, 100$ . For all 6 binarized datasets, MinHash significantly outperforms SimHash. For example, to achieve a 90% recall for top-1 on MNIST, MinHash needs to scan, on average, 0.6% of the data points while SimHash has to scan 5%. For fair comparisons, we present the optimum outcomes (i.e., smallest fraction of data points) separately for MinHash and SimHash, by searching a wide range of parameters ( $K, L$ ), where  $K \in \{1, 2, \dots, 30\}$  is the number of hash functions per table and  $L \in \{1, 2, \dots, 200\}$  is the number of tables.

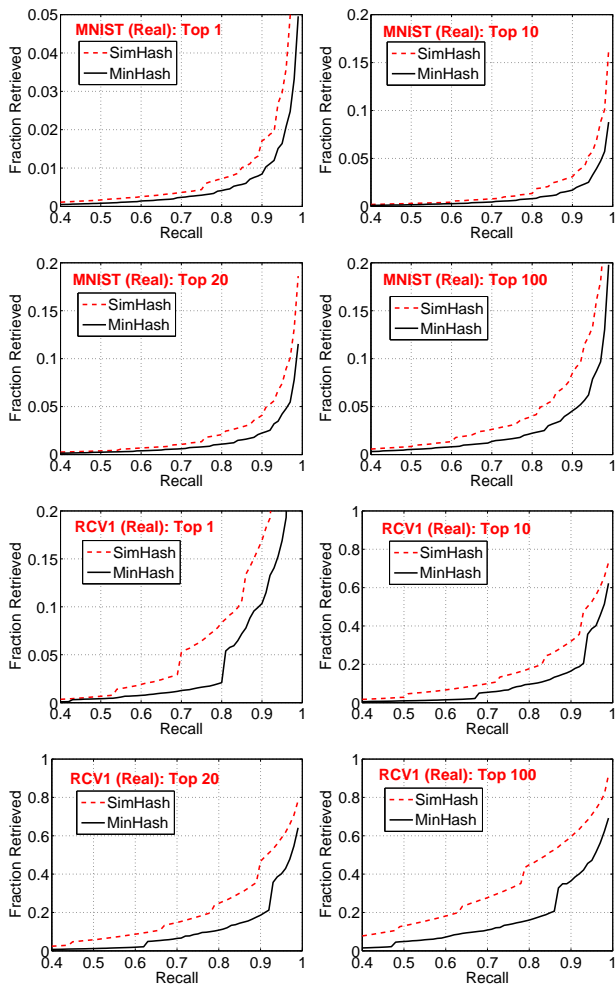


Figure 9: Retrieval experiments on the original real-valued data. We apply SimHash on the original data and MinHash on the binarized data, and we evaluate the retrieval results based on the cosine similarity of the original data. MinHash still outperforms SimHash.

To conclude this section, we also add a set of experiments using the original (real-valued) data, for MNIST and RCV1. We apply SimHash on the original data and MinHash on the binarized data. We also evaluate the retrieval results based on the cosine similarities of the original data. This set-up places MinHash in a very disadvantageous place compared to SimHash. Nevertheless, we can see from Figure 9 that MinHash still noticeably outperforms SimHash, although the improvements are not as significant, compared to the experiments on binarized data (Figure 8).

## 6 Conclusion

Minwise hashing (MinHash), originally designed for detecting duplicate web pages [3, 10, 15], has been widely adopted in the search industry, with numerous applications, for example, large-sale machine learning systems [23, 21], Web spam [29, 18], content match-

ing for online advertising [25], compressing social networks [9], advertising diversification [13], graph sampling [24], Web graph compression [5], etc. Furthermore, the recent development of *one permutation hashing* [21, 27] has substantially reduced the preprocessing costs of MinHash, making the method more practical.

In machine learning research literature, however, it appears that SimHash is more popular for approximate near neighbor search. We believe part of the reason is that researchers tend to use the cosine similarity, for which SimHash can be directly applied.

It is usually taken for granted that MinHash and SimHash are theoretically incomparable and the choice between them is decided based on whether the desired notion of similarity is cosine similarity or resemblance. This paper has shown that MinHash is provably a better LSH than SimHash even for cosine similarity. Our analysis provides a first provable way of comparing two LSHs devised for different similarity measures. Theoretical and experimental evidence indicates significant computational advantage of using MinHash in place of SimHash. Since LSH is a concept studied by a wide variety of researchers and practitioners, we believe that the results shown in this paper will be useful from both theoretical as well as practical point of view.

**Acknowledgements:** Anshumali Shrivastava is a Ph.D. student supported by NSF (DMS0808864, SES1131848, III1249316) and ONR (N00014-13-1-0764). Ping Li is partially supported by AFOSR (FA9550-13-1-0137), ONR (N00014-13-1-0764), and NSF (III1360971, BIGDATA1419210).

## A Proof of Theorem 1

The only less obvious step is the **Proof of tightness:** Let a continuous function  $f(\mathcal{S})$  be a sharper upper bound i.e.,  $\mathcal{R} \leq f(\mathcal{S}) \leq \frac{\mathcal{S}}{2-\mathcal{S}}$ . For any rational  $\mathcal{S} = \frac{p}{q}$ , with  $p, q \in \mathbb{N}$  and  $p \leq q$ , choose  $f_1 = f_2 = q$  and  $a = p$ . Note that  $f_1, f_2$  and  $a$  are positive integers. This choice leads to  $\frac{\mathcal{S}}{2-\mathcal{S}} = \mathcal{R} = \frac{p}{2q-p}$ . Thus, the upper bound is achievable for all rational  $\mathcal{S}$ . Hence, it must be the case that  $f(\mathcal{S}) = \frac{\mathcal{S}}{2-\mathcal{S}} = \mathcal{R}$  for all rational values of  $\mathcal{S}$ . For any real number  $c \in [0, 1]$ , there exists a Cauchy sequence of rational numbers  $\{r_1, r_2, \dots, r_n, \dots\}$  such that  $r_n \in \mathbb{Q}$  and  $\lim_{n \rightarrow \infty} r_n = c$ . Since all  $r_n$ 's are rational,  $f(r_n) = \frac{r_n}{2-r_n}$ . From the continuity of both  $f$  and  $\frac{\mathcal{S}}{2-\mathcal{S}}$ , we have  $f(\lim_{n \rightarrow \infty} r_n) = \lim_{n \rightarrow \infty} \frac{r_n}{2-r_n}$  which implies  $f(c) = \frac{c}{2-c}$  implying  $\forall c \in [0, 1]$ .

For tightness of  $\mathcal{S}^2$ , let  $\mathcal{S} = \sqrt{\frac{p}{q}}$ , choosing  $f_2 = a = p$  and  $f_1 = q$  gives an infinite set of points having  $\mathcal{R} = \mathcal{S}^2$ . We now use similar arguments in the proof tightness of upper bound. All we need is the existence of a Cauchy sequence of square root of rational numbers converging to any real  $c$ .  $\square$



## References

- [1] Alekh Agarwal, Olivier Chapelle, Miroslav Dudik, and John Langford. A reliable effective terascale linear learning system. Technical report, arXiv:1110.4198, 2011.
- [2] Alexandr Andoni and Piotr Indyk. E2lsh: Exact euclidean locality sensitive hashing. Technical report, 2004.
- [3] Andrei Z. Broder. On the resemblance and containment of documents. In *the Compression and Complexity of Sequences*, pages 21–29, Positano, Italy, 1997.
- [4] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. In *STOC*, pages 327–336, Dallas, TX, 1998.
- [5] Gregory Buehrer and Kumar Chellapilla. A scalable pattern mining approach to web graph compression with communities. In *WSDM*, pages 95–106, Stanford, CA, 2008.
- [6] Tushar Chandra, Eugene Ie, Kenneth Goldman, Tomas Lloret Llinares, Jim McFadden, Fernando Pereira, Joshua Redstone, Tal Shaked, and Yoram Singer. Sibyl: a system for large scale machine learning. Technical report, 2010.
- [7] Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik. Support vector machines for histogram-based image classification. 10(5):1055–1064, 1999.
- [8] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, Montreal, Quebec, Canada, 2002.
- [9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. On compressing social networks. In *KDD*, pages 219–228, Paris, France, 2009.
- [10] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet L. Wiener. A large-scale study of the evolution of web pages. In *WWW*, pages 669–678, Budapest, Hungary, 2003.
- [11] Jerome H. Friedman, F. Baskett, and L. Shustek. An algorithm for finding nearest neighbors. *IEEE Transactions on Computers*, 24:1000–1006, 1975.
- [12] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, 42(6):1115–1145, 1995.
- [13] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *WWW*, pages 381–390, Madrid, Spain, 2009.
- [14] Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pages 136–143, Barbados, 2005.
- [15] Monika Rauch Henzinger. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR*, pages 284–291, 2006.
- [16] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, pages 604–613, Dallas, TX, 1998.
- [17] Yugang Jiang, Chongwah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, pages 494–501, Amsterdam, Netherlands, 2007.
- [18] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *WSDM*, pages 219–230, Palo Alto, California, USA, 2008.
- [19] Ping Li, Kenneth W. Church, and Trevor J. Hastie. Conditional random sampling: A sketch-based sampling technique for sparse data. In *NIPS*, pages 873–880, Vancouver, BC, Canada, 2006.
- [20] Ping Li and Arnd Christian König. b-bit minwise hashing. In *Proceedings of the 19th International Conference on World Wide Web*, pages 671–680, Raleigh, NC, 2010.
- [21] Ping Li, Art B Owen, and Cun-Hui Zhang. One permutation hashing. In *NIPS*, Lake Tahoe, NV, 2012.
- [22] Ping Li, Anshumali Shrivastava, and Arnd Christian König. b-bit minwise hashing in practice. In *Internetware*, Changsha, China, 2013.
- [23] Ping Li, Anshumali Shrivastava, Joshua Moore, and Arnd Christian König. Hashing algorithms for large-scale learning. In *NIPS*, Granada, Spain, 2011.
- [24] Marc Najork, Sreenivas Gollapudi, and Rina Panigrahy. Less is more: sampling the neighborhood graph makes salsa better and faster. In *WSDM*, pages 242–251, Barcelona, Spain, 2009.
- [25] Sandeep Pandey, Andrei Broder, Flavio Chierichetti, Vanja Josifovski, Ravi Kumar, and Sergei Vassilvitskii. Nearest-neighbor caching for content-match applications. In *WWW*, pages 441–450, Madrid, Spain, 2009.
- [26] Anshumali Shrivastava and Ping Li. Fast near neighbor search in high-dimensional binary data. In *ECML*, Bristol, UK, 2012.
- [27] Anshumali Shrivastava and Ping Li. Densifying one permutation hashing via rotation for fast near neighbor search. In *ICML*, 2014.
- [28] Simon Tong. Lessons learned developing a practical large scale machine learning system. <http://googleresearch.blogspot.com/2010/04/lessons-learned-developing-practical.html>, 2008.
- [29] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking web spam with html style similarities. *ACM Trans. Web*, 2(1):1–28, 2008.
- [30] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *NIPS*, 2008.