Asymmetric Minwise Hashing for Indexing Binary Inner Products and Set Containment

Anshumali Shrivastava
Department of Computer Science
Computer and Information Science
Cornell University
Ithaca, NY 14853, USA
anshu@cs.cornell.edu

Ping Li
Department of Statistics and Biostatistics
Department of Computer Science
Rutgers University
Piscataway, NJ 08854, USA
pingli@stat.rutgers.edu

ABSTRACT

Minwise hashing (Minhash) is a widely popular indexing scheme in practice. Minhash is designed for estimating set resemblance and is known to be suboptimal in many applications where the desired measure is set overlap (i.e., inner product between binary vectors) or set containment. Minhash has inherent bias towards smaller sets, which adversely affects its performance in applications where such a penalization is not desirable. In this paper, we propose asymmetric minwise hashing (*MH-ALSH*), to provide a solution to this well-known problem. The new scheme utilizes asymmetric transformations to cancel the bias of traditional minhash towards smaller sets, making the final "collision probability" monotonic in the inner product. Our theoretical comparisons show that, for the task of retrieving with binary inner products, asymmetric minhash is provably better than traditional minhash and other recently proposed hashing algorithms for general inner products. Thus, we obtain an algorithmic improvement over existing approaches in the literature. Experimental evaluations on four publicly available high-dimensional datasets validate our claims. The proposed scheme outperforms, often significantly, other hashing algorithms on the task of near neighbor retrieval with set containment. Our proposal is simple and easy to implement in practice.

1. INTRODUCTION

Record matching (or linkage), data cleansing and plagiarism detection are among the most frequent operations in many large-scale data processing systems over the web. *Minwise hashing* (or minhash) [6, 7, 27] is a popular technique deployed by big data industries for these tasks. Minhash was originally developed for economically estimating the *resemblance* similarity between sets (which can be equivalently viewed as binary vectors). Later, because of its locality sensitive property [22], minhash became a widely used hash function for creating hash buckets leading to efficient algorithms for numerous applications including spam detection [6], collaborative filtering [4], news personalization [15], compressing social networks [13], graph sampling [14], record linkage [25], duplicate detection [21], all pair similarity [5], etc.

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741285 .

1.1 Sparse Binary Data, Set Resemblance, and Set Containment

Binary representations for web data are common, owing to the wide adoption of the "bag of words (n-gram)" representations for documents and images. It is often the case that a significant number of words (or combinations of words) occur rarely in a document and most of the higher-order n-grams in the document occur only once. Thus in practice, often only the presence or absence information suffices [9, 20, 24]. Leading search firms routinely use sparse binary representations in their large data systems, e.g., [8].

The underlying similarity measure of interest with minhash is the resemblance (also known as the Jaccard similarity). The resemblance similarity between two sets $x, y \subseteq \Omega = \{1, 2, ..., D\}$ is

$$\mathcal{R} = \frac{|x \cap y|}{|x \cup y|} = \frac{a}{f_x + f_y - a}, \quad \text{where } f_x = |x|,\ f_y = |y|,\ a = |x \cap y|. \tag{1}$$

Sets can be equivalently viewed as binary vectors with each component indicating the presence or absence of an attribute. The cardinality (e.g., f_x, f_y) is the number of nonzeros in the binary vector.

While the resemblance similarity is convenient and useful in numerous applications, there are also many scenarios where the resemblance is not the desired similarity measure [1, 11]. For instance, consider text descriptions of two restaurants:

1. "Five Guys Burgers and Fries Downtown Brooklyn New York"
2. "Five Kitchen Berkley"

Shingle (n-gram) based representations for strings are common in practice. Typical (first-order) shingle based representations of these names will be (i) {five, guys, burgers, and, fries, downtown, brooklyn, new, york } and (ii) {five, kitchen, berkley}. Now suppose the query is "Five Guys" which in shingle representation is {Five, Guys}. Suppose we hope to match and search the records, for this query "Five Guys", based on resemblance. Observe that the resemblance between query and record (i) is $\frac{2}{9}$ = 0.22, while that with record (ii) is $\frac{1}{4}$ = 0.25. Thus, simply based on resemblance, record (ii) is a better match for query "Five Guys" than record (i), which however should not be correct in this content.

Clearly the issue here is that resemblance penalizes the sizes of the sets involved. Shorter sets are unnecessarily favored over longer ones, which hurts the performance in (e.g.,) record matching [1]. There are other scenarios where such penalization is undesirable. For instance, in plagiarism detection, it is typically immaterial whether the text is plagiarized from a long or a short document.

To counter the often unnecessary penalization of the sizes of the sets with resemblance, a modified measure, the *set containment* (or Jaccard containment) was adopted [6, 1, 11]. Containment of set x and y with respect to x is defined as

$$\mathcal{J}_{\mathcal{C}} = \frac{|x \cap y|}{|x|} = \frac{a}{f_x}. \quad (2)$$

In the above example with query "Five Guys", the set containment with respect to query for record (i) will be $\frac{2}{2} = 1$ and with respect to record (ii) it will be $\frac{1}{2}$, leading to the desired ordering. It should be noted that for any fixed query x, the ordering under set containment with respect to the query, is the same as the ordering with respect to the intersection a (or binary inner product). Thus, near neighbor search problem with respect to $\mathcal{J}_{\mathcal{C}}$ is equivalent to the near neighbor search problem with respect to a.

1.2 Maximum Inner Product Search (MIPS) & Maximum Containment Search (MCS)

Formally, we state our problem of interest. We are given a collection $\mathcal{C}$ containing n sets (or binary vectors) over universe Ω with $|\Omega| = D$ (or binary vectors in $\{0,1\}^D$). Given a query $q \subset \Omega$, we are interested in the problem of finding $x \in \mathcal{C}$ such that

$$x = \arg\max_{x \in \mathcal{C}} |x \cap q| = \arg\max_{x \in \mathcal{C}} q^T x; \quad (3)$$

where $|\ |$ is the cardinality of the set. This is the so-called *maximum inner product search (MIPS)* problem.

For binary data, the MIPS problem is equivalent to searching with set containment with respect to the query, because the cardinality of the query does not affect the ordering and hence

$$x = \arg\max_{x \in \mathcal{C}} |x \cap q| = \arg\max_{x \in \mathcal{C}} \frac{|x \cap q|}{|q|}; \quad (4)$$

which we also refer to as the *maximum containment search (MCS)* problem.

1.3 Shortcomings of Inverted Index Based Approaches for MIPS (and MCS)

Owing to its practical significance, there have been many existing heuristics for solving the MIPS (or MCS) problem [31, 34, 12]. A notable recent work among them made use of the inverted index based approach [1]. Inverted indexes might be suitable for problems when the sizes of documents are small and each record only contains few words. This situation, however, is not always observed in practice. The documents over the web are large with huge vocabulary. Moreover, the vocabulary blows up very quickly once we start using higher-order shingles. In addition, there is an increasing interest in enriching the text with extra synonyms to make the search more effective and robust to semantic meanings [1], at the cost of a significant increase of the sizes of the documents. Furthermore, if the query contains many words then the inverted index is not very useful. To mitigate this issue several additional heuristics were proposed, for instance, the heuristic based on minimal infrequent sets [1]. Computing minimal infrequent sets is similar to the set cover problem which is hard in general and thus [1] resorted to greedy heuristics. The number of minimal infrequent sets could be huge in general and so these heuristics can be very costly. Also, such heuristics require the knowledge of the entire dataset before hand which is usually not practical in a dynamic environment like the web. In addition, inverted index based approaches do not have theoretical guarantees on the query time and their performance is very much dataset dependent. Not surprisingly, it was shown in [17] that simply using a sign of the projected document vector representation referred to as TOPSIG, which is also similar in nature to sign random projections (SRP) [18, 10], outperforms inverted index based approaches for querying.

1.4 Probabilistic Hashing

Locality Sensitive Hashing (LSH) [22] based randomized techniques are common and successful in industrial practice for efficiently solving NNS (*near neighbor search*). They are some of the few known techniques that do not suffer from the curse of dimensionality. Hashing based indexing schemes provide provably sub-linear algorithms for search which is a boon in this era of big data where even linear search algorithms are impractical due to latency. Furthermore, hashing based indexing schemes are massively parallelizable, which makes them ideal for modern distributed systems. The prime focus of this paper will be on efficient hashing based algorithms for binary inner products.

Despite the interest in set containment and binary inner products, there were no hashing algorithms for these measures for a long time and minwise hashing is still a popular heuristic [1]. Recently, it was shown that general inner products for real vectors can be efficiently solved by using asymmetric locality sensitive hashing schemes [35, 37]. The asymmetry is necessary for the general inner products and an impossibility of having a symmetric hash function can be easily shown using elementary arguments. Thus, binary inner product (or set intersection) being a special case of general inner products also admits provable efficient search algorithms with these asymmetric hash functions which are based on random projections. However, it is known that random projections are suboptimal for retrieval in the sparse binary domain [39]. Hence, it is expected that the existing asymmetric locality sensitive hashing schemes for general inner products are likely to be suboptimal for retrieving with sparse high dimensional binary-like datasets, which are common over the web.

1.5 Our Contributions

We investigate hashing based indexing schemes for the problem of near neighbor search with binary inner products and set containment. The impossibility of existence of LSH for general inner products shown in [35] also hold for the binary case.

Recent results on hashing algorithms for maximum inner product search [35] have shown the usefulness of asymmetric transformations in constructing provable hash functions for new similarity measures, which were otherwise impossible. Going further along this line, we provide a novel (and still very simple) asymmetric transformation for binary data, that corrects minhash and removes the undesirable bias of minhash towards the sizes of the sets involved. Such an asymmetric correction eventually leads to a provable hashing scheme for binary inner products, which we call *asymmetric minwise hashing (MH-ALSH)*. Our theoretical comparisons show that for binary data, which are common over the web, the new hashing scheme is provably more efficient that the recently proposed asymmetric hash functions for general inner products [35, 37]. Thus, we obtain a provable algorithmic improvement over the state-of-the-art hashing technique for binary inner products. The construction of our asymmetric transformation for minhash could be of independent interest in itself.

The proposed asymmetric minhash significantly outperforms existing hashing schemes, in the tasks of ranking and near neighbor search with set containment as the similarity measure, on four real-world high-dimensional datasets. Our final proposed algorithm is simple and only requires minimal modifications of the traditional minhash and hence it can be easily adopted in practice.

2. BACKGROUND

2.1 c-Approximate Near Neighbor Search and the Classical LSH

Past attempts of finding efficient algorithms, for exact near neighbor search based on space partitioning, often turned out to be a disappointment with the massive dimensionality of modern datasets [40]. Due to the curse of dimensionality, theoretically it is hopeless to obtain an efficient algorithm for exact near neighbor search. Approximate versions of near neighbor search problem were proposed [22] to overcome the linear query time bottleneck. One commonly adopted such formulation is the c-approximate Near Neighbor (c-NN).

DEFINITION 1. *(c-Approximate Near Neighbor or c-NN). [22] 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 exists an S_0-near neighbor of q in P, it reports some cS_0-near neighbor.*

The usual notion of S_0-near neighbor is in terms of distance. Since we are dealing with similarities, we 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.

The popular technique, with near optimal guarantees for c-NN in many interesting cases, uses the underlying theory of *Locality Sensitive Hashing* (LSH) [22]. LSH relies on 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. More specifically, consider $\mathcal{H}$ a family of hash functions mapping $\mathbb{R}^D$ to some set $\mathcal{S}$.

DEFINITION 2. *(Locality Sensitive Hashing) A family $\mathcal{H}$ is called (S_0, cS_0, p_1, p_2) sensitive if for any two point $x, y \in \mathbb{R}^D$ 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. Note, $c < 1$ as we are defining neighbors in terms of similarity. To obtain distance analogy we can resort to $D(x, y) = 1 - Sim(x, y)$

FACT 1. *[22] 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 and space $O(n^{1+\rho})$, $\rho = \frac{\log 1/p_1}{\log 1/p_2} < 1$*

LSH trades off query time with extra preprocessing time and space that can be accomplished off-line. It requires constructing a one time data structure which costs $O(n^{1+\rho})$ space and further any c-approximate near neighbor queries can be answered in $O(n^\rho \log_{1/p_2} n)$ time in the worst case.

A particularly interesting sufficient condition for existence of LSH is the monotonicity of the collision probability in $Sim(x, y)$. Thus, if a hash function family $\mathcal{H}$ satisfies,

$$Pr_{h \in \mathcal{H}}(h(x) = h(y)) = g(Sim(x, y)), \tag{5}$$

where g is a monotonically increasing function, then the conditions of Definition 2 are automatically satisfied for all $c < 1$.

The quantity $\rho < 1$ is a property of the LSH family, and it is of particular interest because it determines the worst case query complexity of the c-approximate near neighbor search. It should be further noted, that the complexity depends on S_0 which is the operating threshold and c, the approximation ratio we are ready to tolerate. In case when we have two or more LSH families for a given similarity measure, then the LSH family with smaller value of ρ, for given S_0 and c, is preferred.

2.2 Minwise Hashing (Minhash)

Minwise hashing [6] is the LSH for the *resemblance*, also known as the *Jaccard similarity*, between sets. In this paper, we focus on binary data vectors which can be equivalent viewed as sets.

Given a set $x \in \Omega = \{1, 2, ..., D\}$, the minwise hashing family applies a random permutation $\pi : \Omega \rightarrow \Omega$ on x and stores only the minimum value after the permutation mapping. Formally minwise hashing (or minhash) is defined as:

$$h_\pi(x) = \min(\pi(x)). \tag{6}$$

Given sets x and y, it can be shown that the probability of collision is the resemblance $\mathcal{R} = \frac{|x \cap y|}{|x \cup y|}$:

$$Pr_\pi(h_\pi(x) = h_\pi(y)) = \frac{|x \cap y|}{|x \cup y|} = \frac{a}{f_x + f_y - a} = \mathcal{R}. \tag{7}$$

where $f_x = |x|$, $f_y = |y|$, and $a = |x \cap y|$. It follows from Eq. (7) that minwise hashing is (S_0, cS_0, S_0, cS_0)-sensitive family of hash function when the similarity function of interest is resemblance.

Even though minhash was really meant for retrieval with resemblance similarity, it is nevertheless a popular hashing scheme used for retrieving set containment or intersection for binary data [1]. In practice, the ordering of inner product a and the ordering or resemblance $\mathcal{R}$ can be different because of the variation in the values of f_x and f_y, and as argued in Section 1, which may be undesirable and lead to suboptimal results. We show later that by exploiting asymmetric transformations we can get away with the undesirable dependency on the number of nonzeros leading to a better hashing scheme for indexing set intersection (or binary inner products).

2.3 LSH for L2 Distance (L2LSH)

[16] presented a novel LSH family for all L_p ($p \in (0, 2]$) distances. In particular, when $p = 2$, this scheme provides an LSH family for L_2 distance. Formally, given a fixed number r, we choose a random vector w with each component generated from i.i.d. normal, i.e., $w_i \sim N(0, 1)$, and a scalar b generated uniformly at random from $[0, r]$. The hash function is defined as:

$$h^{L2}_{w,b}(x) = \left\lfloor \frac{w^T x + b}{r} \right\rfloor, \tag{8}$$

where $\lfloor \rfloor$ is the floor operation. The collision probability under this scheme can be shown to be

$$Pr(h^{L2}_{w,b}(x) = h^{L2}_{w,b}(y)) = F_r(d), \tag{9}$$

$$F_r(d) = 1 - 2\Phi(-r/d) - \frac{2}{\sqrt{2\pi} r/d}\left(1 - e^{-r^2/(2d^2)}\right) \tag{10}$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ is the cumulative density function (cdf) of standard normal distribution and $d = ||x - y||_2$ is the Euclidean distance between the vectors x and y. This collision probability $F(d)$ is a monotonically decreasing function of the distance d and hence $h^{L2}_{w,b}$ is an LSH for $L2$ distances. This scheme is also the part of LSH package [2]. Here r is a parameter.

2.4 LSH for Cosine Similarity (SRP)

Sign Random Projections (SRP) or *simhash* is another popular LSH for the cosine similarity measure, which originates from the

concept of **Sign Random Projections (SRP)** [18, 10]. Given a vector x, SRP utilizes a random w vector with each component generated from i.i.d. normal, i.e., $w_i \sim N(0,1)$, and only stores the sign of the projection. Formally simhash is given by

$$h^{sign}(x) = sign(w^T x). \tag{11}$$

It was shown in the seminal work [18] that collision under SRP satisfies the following equation:

$$Pr_w(h^{sign}(x) = h^{sign}(y)) = 1 - \frac{\theta}{\pi}, \tag{12}$$

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 popular **cosine similarity**. For sets (or equivalently binary vectors), the cosine similarity reduces to

$$\mathcal{S} = \frac{a}{\sqrt{f_x f_y}} \tag{13}$$

The recent work on *coding for random projections* [28] has shown the advantage of SRP (and 2-bit random projections) over L2LSH for both similarity estimation and near neighbor search. Interestingly, another recent work [39] has shown that for binary data (actually even sparse non-binary data), minhash can significantly outperform SRP for near neighbor search even as we evaluate both SRP and minhash in terms of the cosine similarity (although minhash is designed for resemblance). This motivates us to design asymmetric minhash for achieving better performance in retrieving set containments. But first, we provide an overview of asymmetric LSH for general inner products (not restricted to binary data).

2.5 Asymmetric LSH (ALSH)

The term "ALSH" stands for *asymmetric LSH*, as used in a recent work [35]. Through an elementary argument, [35] showed that it is not possible to have a Locality Sensitive Hashing (LSH) family for general unnormalized inner products.

For inner products between vectors x and y, it is possible to have $x^T y \gg x^T x$. Thus for any hashing scheme h to be a valid LSH, we must have $Pr(h(x) = h(y)) > Pr(h(x) = h(x)) = 1$, which is an impossibility. It turns out that there is a simple fix, if we allow asymmetry in the hashing scheme. Allowing asymmetry leads to an extended framework of asymmetric locality sensitive hashing (ALSH). The idea to is have a different hashing scheme for assigning buckets to the data point in the collection $\mathcal{C}$, and an altogether different hashing scheme while querying.

DEFINITION 3. *(Asymmetric Locality Sensitive Hashing (ALSH)) A family $\mathcal{H}$, along with the two vector functions $Q : \mathbb{R}^D \mapsto \mathbb{R}^{D'}$ (**Query Transformation**) and $P : \mathbb{R}^D \mapsto \mathbb{R}^{D'}$ (**Preprocessing Transformation**), is called (S_0, cS_0, p_1, p_2)-sensitive if for a given c-NN instance with query q, and the hash function h chosen uniformly from $\mathcal{H}$ satisfies the following:*

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

Here x is any point in the collection $\mathcal{C}$. Asymmetric LSH borrows all theoretical guarantees of the LSH.

FACT 2. *Given a family of hash function $\mathcal{H}$ and the associated query and preprocessing transformations Q and P respectively, which is (S_0, cS_0, p_1, p_2) -sensitive, one can construct a data structure for c-NN with $O(n^\rho \log n)$ query time and space $O(n^{1+\rho})$, where $\rho = \frac{\log p_1}{\log p_2}$.*

[35] showed that using asymmetric transformations, the problem of **maximum inner product search (MIPS)** can be reduced to the problem of approximate near neighbor search in L_2. The algorithm first starts by scaling all $x \in \mathcal{C}$ by a constant large enough, such that $||x||_2 \leq U < 1$. The proposed ALSH family (**L2-ALSH**) is the LSH family for L_2 distance with the Preprocessing transformation $P : \mathbb{R}^D \mapsto \mathbb{R}^{D+m}$ and the Query transformation $Q : \mathbb{R}^D \mapsto \mathbb{R}^{D+2m}$ defined as follows:

$$P^{L2}(x) = [x; ||x||_2^2; ....; ||x||_2^{2^m}; 1/2; ...; 1/2] \tag{14}$$

$$Q^{L2}(x) = [x; 1/2; ...; 1/2; ||x||_2^2; ....; ||x||_2^{2^m}], \tag{15}$$

where [;] is the concatenation. $P^{L2}(x)$ appends m scalers of the form $||x||_2^{2^i}$ followed by m "1/2s" at the end of the vector x, while $Q^{L2}(x)$ first appends m "1/2s" to the end of the vector x and then m scalers of the form $||x||_2^{2^i}$. It was shown that this leads to provably efficient algorithm for MIPS.

FACT 3. *[35] For the problem of c-approximate MIPS in a bounded space, one can construct a data structure having $O(n^{\rho^*_{L2-ALSH}} \log n)$ query time and space $O(n^{1+\rho^*_{L2-ALSH}})$, where $\rho^*_{L2-ALSH} < 1$ is the solution to constrained optimization (16).*

$$\rho^*_{L2-ALSH} \tag{16}$$
$$= \min_{U<1, m\in N, r} \frac{\log F_r\left(\sqrt{m/2 - 2S_0\left(\frac{U^2}{V^2}\right) + 2U^{2^{m+1}}}\right)}{\log F_r\left(\sqrt{m/2 - 2cS_0\left(\frac{U^2}{V^2}\right)}\right)}$$
$$s.t. \quad \frac{U^{(2^{m+1}-2)}V^2}{S_0} < 1 - c,$$

Here the guarantees depends on the maximum norm of the space $V = \max_{x\in\mathcal{C}} ||x||_2$.

Quickly after, it was realized that a very similar idea can convert the MIPS problem in the problem of maximum cosine similarity search which can be efficiently solve by SRP leading to a new and better ALSH for MIPS **Sign-ALSH** [37] which works as follows: The algorithm again first starts by scaling all $x \in \mathcal{C}$ by a constant large enough, such that $||x||_2 \leq U < 1$. The proposed ALSH family (**Sign-ALSH**) is the SRP family for cosine similarity with the Preprocessing transformation $P^{sign} : \mathbb{R}^D \mapsto \mathbb{R}^{D+m}$ and the Query transformation $Q^{sign} : \mathbb{R}^D \mapsto \mathbb{R}^{D+2m}$ defined as follows:

$$P^{sign}(x) = [x; 1/2 - ||x||_2^2; ...; 1/2 - ||x||_2^{2^m}; 0; ...; 0] \tag{17}$$

$$Q^{sign}(x) = [x; 0; ...; 0; 1/2 - ||x||_2^2; ...; 1/2 - ||x||_2^{2^m}], \tag{18}$$

where [;] is the concatenation. $P^{sign}(x)$ appends m scalers of the form $1/2 - ||x||_2^{2^i}$ followed by m "0s" at the end of the vector x, while $Q^{sign}(x)$ appends m "0" followed by m scalers of the form $1/2 - ||x||_2^{2^i}$ to the end of the vector x. It was shown that this leads to provably efficient algorithm for MIPS.

As demonstrated by the recent work [28] on *coding for random projections*, there is a significant advantage of SRP over L2LSH for near neighbor search. Thus, it is not surprising that Sign-ALSH outperforms L2-ALSH for the MIPS problem. Similar to L2LSH, the runtime guarantees for Sign-ALSH can be shown as:

FACT 4. *For the problem of c-approximate MIPS, one can construct a data structure having $O(n^{\rho^*_{Sign-ALSH}} \log n)$ query time and space $O(n^{1+\rho^*_{Sign-ALSH}})$, where $\rho^*_{Sign-ALSH} < 1$ is the*

solution to constraint optimization problem

$$\rho^*_{Sign-ALSH} = \min_{U,m,} \frac{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\frac{S_0 \times \left(\frac{U^2}{V^2}\right)}{\frac{m}{4} + U^{2m+1}}\right)\right)}{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\min\{\frac{cS_0U^2}{V^2}, z^*\}\right)\right)} \tag{19}$$

$$z^* = \left[\frac{(m - m2^{m-1}) + \sqrt{(m - m2^{m-1})^2 + m^2(2^m - 1)}}{4(2^m - 1)}\right]^{2^{-m}}$$

There is a similar asymmetric transformation [3, 32] which followed by sign random projection leads to another ALSH having very similar performance to Sign-ALSH. The ρ values, which were also very similar to the $\rho_{Sign-ALSH}$ can be shown as

$$\rho_{Sign} = \frac{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\frac{S_0}{V^2}\right)\right)}{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\frac{cS_0}{V^2}\right)\right)} \tag{20}$$

Both L2-ALSH and Sign-ALSH work for any general inner products over $\mathbb{R}^D$. For sparse and high-dimensional binary dataset which are common over the web, it is known that minhash is typically the preferred choice of hashing over random projection based hash functions [39]. We show later that the ALSH derived from minhash, which we call asymmetric minwise hashing (*MH-ALSH*), is more suitable for indexing set intersection for sparse binary vectors than the existing ALSHs for general inner products.

3. SAMPLING BASED ALSH FOR INDEXING BINARY INNER PRODUCTS

In [35], it was shown that there cannot exist any LSH for general unnormalized inner product. Using a slightly different argument it can be shown that even for binary data we cannot have any LSH scheme. Note, for binary inner product $x^T y \leq x^T x$ and therefore we cannot use exactly the same argument as before. But we can have x,y and z such that $x^T y \gg z^T z$. Now, $Pr(h(x) = h(y)) > Pr(h(z) = h(z)) = 1$ is again impossible. However, with asymmetry it is not difficult to construct a provable hashing scheme for binary inner product.

The construction is based on sampling. Simply sampling a random component leads to the popular LSH for hamming distance [33]. The ordering of inner product is different from that of hamming distance. The hamming distance between x and query q is given by $f_x + f_q - 2a$, while we want the collision probability to be monotonic in the inner product a. f_x makes it non-monotonic in a. Note that f_q has no effect on ordering of $x \in \mathcal{C}$ because it is constant for every query. To construct an LSH monotonic in binary inner product, we need an extra trick.

Given a binary data vector x, we sample a random co-ordinate (or attribute). If the value of this co-ordinate is 1 (in other words if this attribute is present in the set), our hash value is a fixed number 0. If this randomly sampled co-ordinate has value 0 (or the attribute is absent) then ensure that the hash value of the query never matches the hash value of the data. Formally,

$$\mathcal{H}_S(f(x)) = \begin{cases} 0 & \text{if } x_i = 1,\ i \text{ drawn uniformly} \\ 1 & \text{if f = Q (for query)} \\ 2 & \text{if f = P (while preprocessing)} \end{cases} \tag{21}$$

Note the asymmetry, i.e., the hash functions are different for query and the dataset. We can also write it down more formally using $P(.)$ and $Q(.)$ but we avoid it for the sake of simplicity.

THEOREM 1. *Given two binary vectors x and y, we have*

$$Pr(\mathcal{H}_S(P(x)) = \mathcal{H}_S(Q(y))) = \frac{a}{D} \tag{22}$$

PROOF. The probability that both $H_S(P(x))$ and $H_S(Q(y))$ have value 0 is $\frac{a}{D}$. They cannot be equal otherwise □

COROLLARY 1. *$\mathcal{H}_S$ is $(S_0, cS_0, \frac{S_0}{D}, \frac{cS_0}{D})$-sensitive ALSH for binary inner product with $\rho_{\mathcal{H}_S} = \frac{\log\left(\frac{S_0}{D}\right)}{\log\left(\frac{cS_0}{D}\right)} < 1$*

3.1 Shortcomings

The above ALSH for binary inner product is likely to be very inefficient for sparse and high dimensional datasets. For those datasets, typically the value of D is very high and the sparsity ensures that a is very small. For modern web datasets, we can have D running into billions (or 2^{64}) while the sparsity is only in few hundreds or perhaps thousands [8]. Therefore, we have $\frac{a}{D} \simeq 0$ which essentially boils down to $\rho_{\mathcal{H}_S} \simeq 1$. In other words, the hashing scheme becomes worthless in sparse high dimensional domain. On the other hand, if we observe the collision probability of minhash Eq.(7), the denominator is $f_x + f_y - a$, which is usually of the order of a and much less than the dimensionality for sparse datasets.

Another way of realizing the problem with the above ALSH is to note that it is informative only if a randomly sampled co-ordinate has value equal to 1. For very sparse dataset with $a \ll D$, sampling a non zero coordinate has probability $\frac{a}{D} \simeq 0$. Thus, almost all of the hashes will be fixed numbers which are not informative.

3.2 Why Is Minhash Reasonable?

In this section, we argue why retrieving inner product based on plain minhash is a reasonable thing to do. Later, we will show a provable way to improve it using asymmetric transformations.

The number of nonzeros in the query, i.e., $|q| = f_q$ does not change the identity of $\arg\max$ in Eq.(4). Let us assume that we have data of bounded sparsity and define constant M as

$$M = \max_{x \in \mathcal{C}} |x| \tag{23}$$

where M is the maximum number of nonzeros (or maximum cardinality of sets) seen in the database. For sparse data seen in practice M is likely to be small compared to D. Outliers, if any, can be handled separately. By observing that $a \leq f_x \leq M$, we also have

$$\frac{a}{f_q + M - a} \leq \frac{a}{f_x + f_q - a} = \mathcal{R} \leq \frac{a}{f_q} \tag{24}$$

Thus, given the bounded sparsity, if we assume that the number of nonzeros in the query is given, then we can show that minhash is an LSH for inner products a because the collision probability can be upper and lower bounded by purely functions of a, M and f_q.

THEOREM 2. *Given bounded sparsity and query q with $|q| = f_q$, minhash is a $(S_0, cS_0, \frac{S_0}{f_q + M - S_0}, \frac{cS_0}{f_q})$ sensitive for inner products a with $\rho^q_{min} = \frac{\log \frac{S_0}{f_q + M - S_0}}{\log \frac{cS_0}{f_q}}$*

This explains why minhash might be a reasonable hashing approach for retrieving inner products or set intersection.

Here, if we remove the assumption that $|q| = f_q$ then in the worst case $\mathcal{R} \leq \frac{a}{f_q} \leq 1$ and we get $\log 1$ in the denominator. Note that the above is the worst case analysis and the assumption $|q| = f_q$ is needed to obtain any meaningful ρ with minhash. We show the power of ALSH in the next section, by providing a better hashing scheme and we do not even need the assumption of fixing $|q| = f_q$.

4. ASYMMETRIC MINWISE HASHING

In this section, we provide a very simple asymmetric fix to minhash, named *asymmetric minwise hashing (MH-ALSH)*, which makes the overall collision probability monotonic in the original inner product a. For sparse binary data, which is common in practice, we later show that the proposed hashing scheme is superior (both theoretically as well as empirically) compared to the existing ALSH schemes for inner product [35].

4.1 The New ALSH for Binary Data

We define the new preprocessing and query transformations $P' : [0,1]^D \rightarrow [0,1]^{D+M}$ and $Q' : [0,1]^D \rightarrow [0,1]^{D+M}$ as:

$$P'(x) = [x; 1; 1; 1; ...; 1; 0; 0; ...; 0] \quad (25)$$

$$Q'(x) = [x; 0; 0; 0; ...; 0], \quad (26)$$

For $P'(x)$ we append $M - f_x$ 1s and rest f_x zeros, while in $Q'(x)$ we simply append M zeros.

At this point we can already see the power of asymmetric transformations. The original inner product between $P'(x)$ and $Q'(x)$ is unchanged and its value is $a = x^T y$. Given the query q, the new resemblance R' between $P'(x)$ and $Q'(q)$ is

$$R' = \frac{|P'(x) \cap Q'(q)|}{|P'(x) \cup Q'(q)|} = \frac{a}{M + f_q - a}. \quad (27)$$

If we define our new similarity as $Sim(x, y) = \frac{a}{M+f_q-a}$, then the near neighbors in this new similarity are the same as near neighbors with respect to either set intersection a or set containment $\frac{a}{f_q}$. Thus, we can instead compute near neighbors in $\frac{a}{M+f_q-a}$ which is also the resemblance between $P'(x)$ and $Q'(q)$. We can therefore use minhash on $P'(x)$ and $Q'(q)$.

Observe that now we have $M + f_q - a$ in the denominator, where M is the maximum nonzeros seen in the dataset (the cardinality of largest set), which for very sparse data is likely to be much smaller than D. Thus, asymmetric minhash is a better scheme than $\mathcal{H}_S$ with collision probability $\frac{a}{D}$ for very sparse datasets where we usually have $M \ll D$.

From theoretical perspective, to obtain an upper bound on the query and space complexity of c-approximate near neighbor with binary inner products, we want the collision probability to be independent of the quantity f_q. This is not difficult to achieve. The asymmetric transformation used to get rid of f_x in the denominator can be reapplied to get rid of f_q.

Formally, we can define $P'' : [0,1]^D \rightarrow [0,1]^{D+2M}$ and $Q'' : [0,1]^D \rightarrow [0,1]^{D+2M}$ as :

$$P''(x) = Q'(P'(x)); \quad Q''(x) = P'(Q'(x)); \quad (28)$$

where in $P''(x)$ we append $M - f_x$ 1s and rest $M + |f_x|$ zeros, while in $Q''(x)$ we append M zeros, then $M - f_q$ 1s and rest zeros

Again the inner product a is unaltered, and the new resemblance then becomes

$$R'' = \frac{|P''(x) \cap Q''(q)|}{|P''(x) \cup Q''(q)|} = \frac{a}{2M - a}. \quad (29)$$

which is independent of f_q and is monotonic in a. This allows us to achieve a formal upper bound on the complexity of c-approximate maximum inner product search with the new asymmetric minhash.

From the collision probability expression, i.e., Eq. (29), we have

THEOREM 3. *Minwise hashing along with Query transformation Q'' and Preprocessing transformation P'' defined by Equation 28 is a $(S_0, cS_0, \frac{S_0}{2M-S_0}, \frac{cS_0}{2M-cS_0})$ sensitive asymmetric hashing family for set intersection.*

This leads to an important corollary.

COROLLARY 2. *There exists an algorithm for c-approximate set intersection, with bounded sparsity M, that requires space $O(n^{1+\rho_{MH-ALSH}})$ and query time $O(n^{\rho_{MH-ALSH}} \log n)$, where*

$$\rho_{MH-ALSH} = \frac{\log \frac{S_0}{2M-S_0}}{\log \frac{cS_0}{2M-cS_0}} < 1 \quad (30)$$

Given query q and any point $x \in \mathcal{C}$, the collision probability under traditional minhash is $R = \frac{a}{f_x+f_q-a}$. This penalizes sets with high f_x, which in many scenarios is not desirable. To balance this negative effect, asymmetric transformation penalizes sets with smaller f_x. Note, that $M - f_x$ ones added in the transformations $P'(x)$ gives additional chance in proportion to $M - f_x$ for minhash of $P'(x)$ not to match with the minhash of $Q'(x)$. This asymmetric probabilistic correction balances the penalization inherent in minhash. This is a simple way of correcting the probability of collision which could be of independent interest in itself. We will show in our evaluation section, that despite this simplicity such correction leads to significant improvement over plain minhash.

4.2 Efficient Sampling

Our transformations P'' and Q'' always create sets with $2M$ nonzeros. In case when M is big, hashing might take a lot of time. We can use (improved) consistent weighted sampling [30, 23] for efficient generation of hashes. We can instead use transformations P''' and Q''' that makes the data non-binary as follows

$$P'''(x) = [x; M - f_x; 0] \quad (31)$$

$$Q'''(x) = [x; 0; M - f_x]$$

It is not difficult to see that the weighted resemblance (or weighted Jaccard similarity) between $P'''(x)$ and $Q'''(q)$ for given query q and any $x \in \mathcal{C}$ is

$$\mathcal{R}_W = \frac{\sum_i \min(P'''(x)_i, Q'''(q)_i)}{\sum_i \max(P'''(x)_i, Q'''(q)_i)} = \frac{a}{2M - a}. \quad (32)$$

Therefore, we can use fast consistent weighted sampling for weighted resemblance on $P'''(x)$ and $Q'''(x)$ to compute the hash values in time constant per nonzero weights, rather than maximum sparsity M. In practice we will need many hashes for which we can utilize the recent line of work that make minhash and weighted minhash significantly much faster [29, 36, 38, 19].

5. THEORETICAL COMPARISONS

For solving the MIPS problem in general data types, we already know two asymmetric hashing schemes, *L2-ALSH* and *Sign-ALSH*, as described in Section 2.5. In this section, we provide theoretical comparisons of the two existing ALSH methods with the proposed asymmetric minwise hashing (*MH-ALSH*). As argued, the LSH scheme described in Section 3 is unlikely to be useful in practice because of its dependence on D; and hence we can safely ignore it for simplicity of the discussion.

Before we formally compare various asymmetric LSH schemes for maximum inner product search, we argue why asymmetric minhash should be advantageous over traditional minhash for retrieving inner products. Let q be the binary query vector, and f_q denotes the number of nonzeros in the query. The $\rho_{MH-ALSH}$ for asymmetric minhash in terms of f_q and M is straightforward from the collision probability Eq.(27):

$$\rho^q_{MH-ALSH} = \frac{\log \frac{S_0}{f_q+M-S_0}}{\log \frac{cS_0}{f_q+M-cS_0}} \quad (33)$$

For minhash, we have from theorem 2 $\rho^q_{min} = \frac{\log \frac{S_0}{f_q+M-S_0}}{\log \frac{cS_0}{f_q}}$. Since M is the upper bound on the sparsity and cS_0 is some value of inner product, we have $M - cS_0 \geq 0$. Using this fact, the following theorem immediately follows

THEOREM 4. *For any query q, we have* $\rho^q_{MH-ALSH} \leq \rho^q_{min}$.

This result theoretically explains why asymmetric minhash is better for retrieval with binary inner products, compared to plain minhash.

For comparing asymmetric minhash with ALSH for general inner products, we compare $\rho_{MH-ALSH}$ with the ALSH for inner products based on sign random projections. Note that it was shown that Sign-ALSH has better theoretical ρ values compared to L2-ALSH [37]. Therefore, it suffices to show that asymmetric minhash outperforms sign random projection based ALSH. Both $\rho_{MH-ALSH}$ and ρ_{sign} can be rewritten in terms of ratio $\frac{S_0}{M}$ as follows. Note that for binary data we have $M = \max_{x \in \mathcal{C}} ||x||^2 = V^2$

$$\rho_{MH-ALSH} = \frac{\log \frac{S_0/M}{2-S_0/M}}{\log \frac{cS_0/M}{2-cS_0/M}}; \quad \rho_{Sign} = \frac{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\frac{S_0}{M}\right)\right)}{\log\left(1 - \frac{1}{\pi}\cos^{-1}\left(\frac{cS_0}{M}\right)\right)} \tag{34}$$

Observe that M is also the upper bound on any inner product. Therefore, we have $0 \leq \frac{S_0}{M} \leq 1$. We plot the values of $\rho_{MH-ALSH}$ and ρ_{sign} for $\frac{S_0}{M} = \{0.1,\ 0.2, ..., 0.8,\ 0.9,\ 0.95\}$ with c. The comparison is summarized in Figure 1. Note that here we use ρ_{Sign} based on the slightly more convenient transformation from [3, 32] instead of $\rho_{Sign-ALSH}$ for convenience although the two schemes perform essentially the same.

Clearly, irrespective of the choice of threshold $\frac{S_0}{M}$ or the approximation ratio c, asymmetric minhash outperforms sign random projection based ALSH in terms of the theoretical ρ values. This is not surprising, because it is known that minhash based methods are often significantly powerful for binary data compared to SRP (or simhash) [39]. Therefore ALSH based on minhash outperforms ALSH based on SRP as shown by our theoretical comparisons. Our proposal thus leads to an algorithmic improvement over state-of-the-art hashing techniques for retrieving binary inner products.

6. EVALUATIONS

In this section, we compare the different hashing schemes on the actual task of retrieving top-ranked elements based on set Jaccard containment. The experiments are divided into two parts. In the first part, we show how the ranking based on various hash functions correlate with the ordering of set containment. In the second part, we perform the actual LSH based bucketing experiment for retrieving top-ranked elements and compare the computational saving obtained by various hashing algorithms.

6.1 Datasets

We used four publicly available high dimensional sparse datasets: *EP2006*, *MNIST*, *NEWS20*, and *NYTIMES*. (Note that "EP2006" is a short name for "E2006LOG1P" from LIBSVM web site.) Except for MNIST, the other three datasets are binary "BoW" representation of the corresponding text corpus. MNIST is an image dataset consisting of 784 pixel image of handwritten digits. Binarized versions of MNIST are commonly used in literature. The pixel values in MNIST were binarized to 0 or 1 values. For each of the four datasets, we generate two partitions. The bigger partition was used to create hash tables and is referred as the *training partition*. The small partition which we call the *query partition* is used for querying. The statistics of these datasets are summarized in Table 1. The datasets cover a wide spectrum of sparsity and dimensionality.

Table 1: Datasets

Dataset	# Query	# Train	# Dim	nonzeros (mean $\pm$ std)
EP2006	2,000	17,395	4,272,227	6072 ± 3208
MNIST	2,000	68,000	784	150 ± 41
NEWS20	2,000	18,000	1,355,191	454 ± 654
NYTIMES	2,000	100,000	102,660	232 ± 114

6.2 Competing Hash Functions

We consider the following hash functions for evaluations:

1. **Asymmetric minwise hashing (Proposed):** This is our proposal, the asymmetric minhash described in Section 4.1.
2. **Traditional minwise hashing (MinHash):** This is the usual minwise hashing, the popular heuristic described in Section 2.2. This is a symmetric hash function, we use h_π as defined in Eq.(6) for both query and the training set.
3. **L2 based Asymmetric LSH for Inner products (L2-ALSH):** This is the asymmetric LSH of [35] for general inner products based on LSH for L2 distance.
4. **SRP based Asymmetric LSH for Inner Products (Sign-ALSH):** This is the asymmetric hash function of [37] for general inner products based on SRP.

6.3 Ranking Experiment: Hash Quality Evaluations

We are interested in knowing, how the orderings under different competing hash functions correlate with the ordering of the underlying similarity measure which in this case is the set containment. For this task, given a query q vector, we compute the top-100 gold standard elements from the training set based on the set containment $\frac{a}{f_q}$. Note that this is the same as the top-100 elements based on binary inner products. Give a query q, we compute K different hash codes of the vector q and all the vectors in the training set. We then compute the number of times the hash values of a vector x in the training set matches the hash values of query q defined by

$$Matches_x = \sum_{t=1}^{K} \mathbf{1}(h_t(q) = h_t(x)), \tag{35}$$

where $\mathbf{1}$ is the indicator function. t subscript is used to distinguish independent draws of the underlying hash function. Based on $Matches_x$ we rank all elements in the training set. This procedure generates a sorted list for every query for every hash function. For asymmetric hash functions, in computing total collisions, on the query vector we use the corresponding Q function (query transformation) followed by underlying hash function, while for elements in the training set we use the P function (preprocessing transformation) followed by the corresponding hash function.

We compute the precision and the recall of the top-100 gold standard elements in the ranked list generated by different hash functions. To compute precision and recall, we start at the top of the ranked item list and walk down in order, suppose we are at the p^{th} ranked element, we check if this element belongs to the gold standard top-100 list. If it is one of the top 100 gold standard elements, then we increment the count of *relevant seen* by 1, else we move to $p + 1$. By p^{th} step, we have already seen p elements, so the *total*

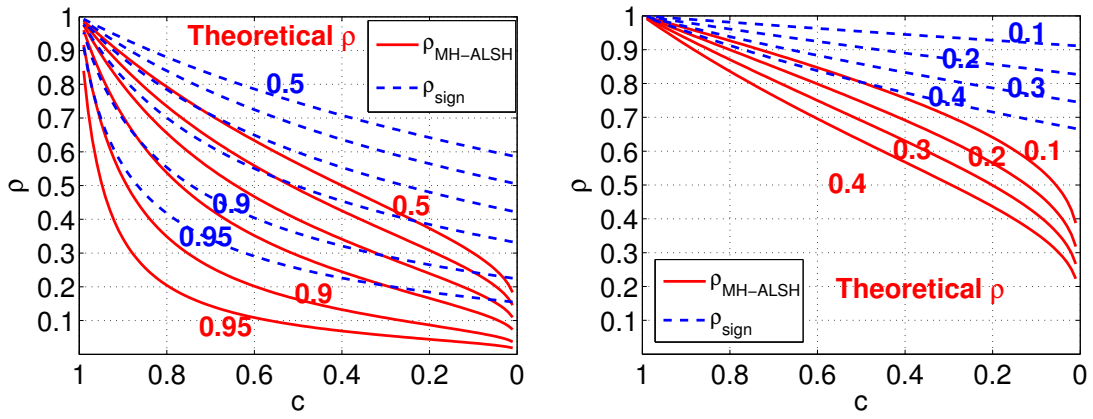

Figure 1: Values of $\rho_{MH-ALSH}$ and ρ_{sign} (lower is better) with respect to approximation ratio c for different $\frac{S_0}{M}$. The curves show that asymmetric minhash (solid curves) is noticeably better than ALSH based on sign random projection (dashed curves) in terms of their ρ values, irrespective of the choices of $\frac{S_0}{M}$ or c. For clarity, the results are shown in two panels.

elements seen is p. The precision and recall at that point is then computed as:

$$Precision = \frac{\text{relevant seen}}{p}, \qquad Recall = \frac{\text{relevant seen}}{100} \tag{36}$$

It is important to balance both. Methodology which obtains higher precision at a given recall is superior. Higher precision indicates higher ranking of the relevant items. We finally average these values of precision and recall over all elements in the query set. The results for $K \in \{32,\ 64,\ 128\}$ are summarized in Figure 2.

We can clearly see, that the proposed hashing scheme always achieves better, often significantly, precision at any given recall compared to other hash functions. The two ALSH schemes are usually always better than traditional minwise hashing. This confirms that fact that ranking based on collisions under minwise hashing can be different from the rankings under set containment or inner products. This is expected, because minhash in addition penalizes the number of nonzeros leading to a ranking very different from the ranking of inner products. Sign-ALSH usually performs better than L2-LSH, this is in line with the results obtained in [37].

It should be noted that ranking experiments only validate the monotonicity of the collision probability. Although, better ranking is certainly a good indicator of good hash function, it does not always mean that we will achieve faster sub-linear LSH algorithm. For bucketing the probability sensitivity around a particular threshold is the most important factor, see [33] for more details. What matters is the **gap** between the collision probability of good and the bad points. In the next subsection, we compare these schemes on the actual task of near neighbor retrieval with set containment.

6.4 LSH Bucketing Experiment: Computational Savings in Near Neighbor Retrieval

In this section, we evaluate the four hashing schemes on the standard (K, L)-parameterized bucketing algorithm [2] for sub-linear time retrieval of near neighbors based on set containment. In (K, L)-parameterized LSH algorithm, we generate L different meta-hash functions. Each of these meta-hash functions is formed by concatenating K different hash values as

$$B_j(x) = [h_{j1}(x); h_{j2}(x); ...; h_{jK}(x)], \tag{37}$$

where $h_{ij}, i \in \{1, 2, ..., K\}$ and $j \in \{1, 2, ..., L\}$, are KL different independent evaluations of the hash function under consideration. Different competing scheme uses its own underlying randomized hash function h.

In general, the (K, L)-parameterized LSH works in two phases:

i) **Preprocessing Phase:** We construct L hash tables from data by storing element x, in the training set, at location $B_j(P(x))$ in the hash-table j. Note that for vanilla minhash which is a symmetric hashing scheme $P(x) = x$. For other asymmetric schemes, we use their corresponding P functions. Preprocessing is a one time operation, once the hash tables are created they are fixed.

ii) **Query Phase:** Given a query q, we report the union of all the points in the buckets $B_j(Q(q))\ \forall j \in \{1, 2, ..., L\}$, where the union is over L hash tables. Again here Q is the corresponding Q function of the asymmetric hashing scheme, for minhash $Q(x) = x$.

Typically, the performance of a bucketing algorithm is sensitive to the choice of parameters K and L. Ideally, to find best K and L, we need to know the operating threshold S_0 and the approximation ratio c in advance. Unfortunately, the data and the queries are very diverse and therefore for retrieving top-ranked near neighbors there are no common fixed threshold S_0 and approximation ratio c that work for all the queries.

Our objective is to compare the four hashing schemes and minimize the effect of K and L, if any, on the evaluations. This is achieved by finding best K and L at every recall level. We run the bucketing experiment for all combinations of $K \in \{1, 2, 3, ...40\}$ and $L \in \{1, 2, 3, ..., 400\}$ for all the four hash functions independently. These choices include the recommended optimal combinations at various thresholds. We then compute, for every K and L, the mean recall of Top-T pairs and the mean number of points reported, per query, to achieve that recall. The best K and L at every recall level is chosen independently for different Ts. The plot of the mean fraction of points scanned with respect to the recall of top-T gold standard near neighbors, where $T \in \{5, 10, 20, 50\}$, is summarized in Figure 3.

The performance of a hashing based method varies with the variations in the similarity levels in the datasets. It can be seen that the proposed asymmetric minhash always retrieves much less number of points, and hence requires significantly less computations, compared to other hashing schemes at any recall level on all the four

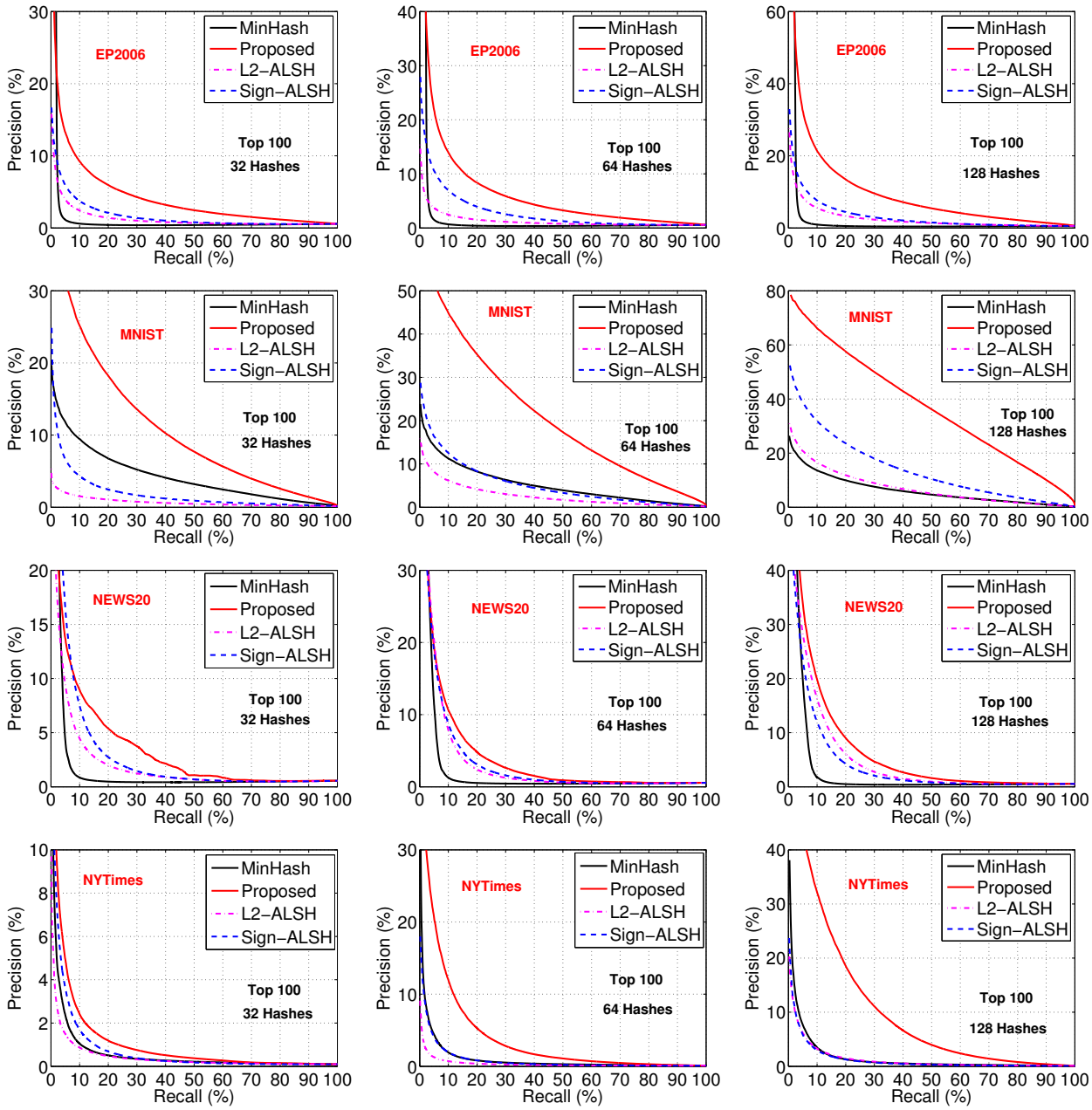

Figure 2: Ranking Experiments. Precision Vs Recall curves for retrieving top-100 items, for different hashing schemes on 4 chosen datasets. The precision and the recall were computed based on the rankings obtained by different hash functions using 32, 64 and 128 independent hash evaluations. Higher precision at a given recall is better.

datasets. Asymmetric minhash consistently outperforms other hash functions irrespective of the operating point. The plots clearly establish the superiority of the proposed scheme for indexing set containment (or inner products).

L2-ALSH and Sign-ALSH perform better than traditional minhash on EP2006 and NEWS20 datasets while they are worse than plain minhash on NYTIMES and MNIST datasets. If we look at the statistics of the dataset from Table 1, NYTIMES and MNIST are precisely the datasets with less variations in the number of nonzeros and hence minhash performs better. In fact, for MNIST dataset with very small variations in the number of nonzeros, the performance of plain minhash is very close to the performance of asymmetric minhash. This is of course expected because there is negligible effect of penalization on the ordering. EP2006 and NEWS20 datasets have huge variations in their number of nonzeros and hence minhash performs very poorly on these datasets. What is exciting is that despite these variations in the nonzeros, asymmetric minhash always outperforms other ALSH for general inner products.

The difference in the performance of plain minhash and asymmetric minhash clearly establishes the utility of our proposal which is simple and does not require any major modification over traditional minhash implementation. Given the fact that minhash is widely popular, we hope that our proposal will be adopted.

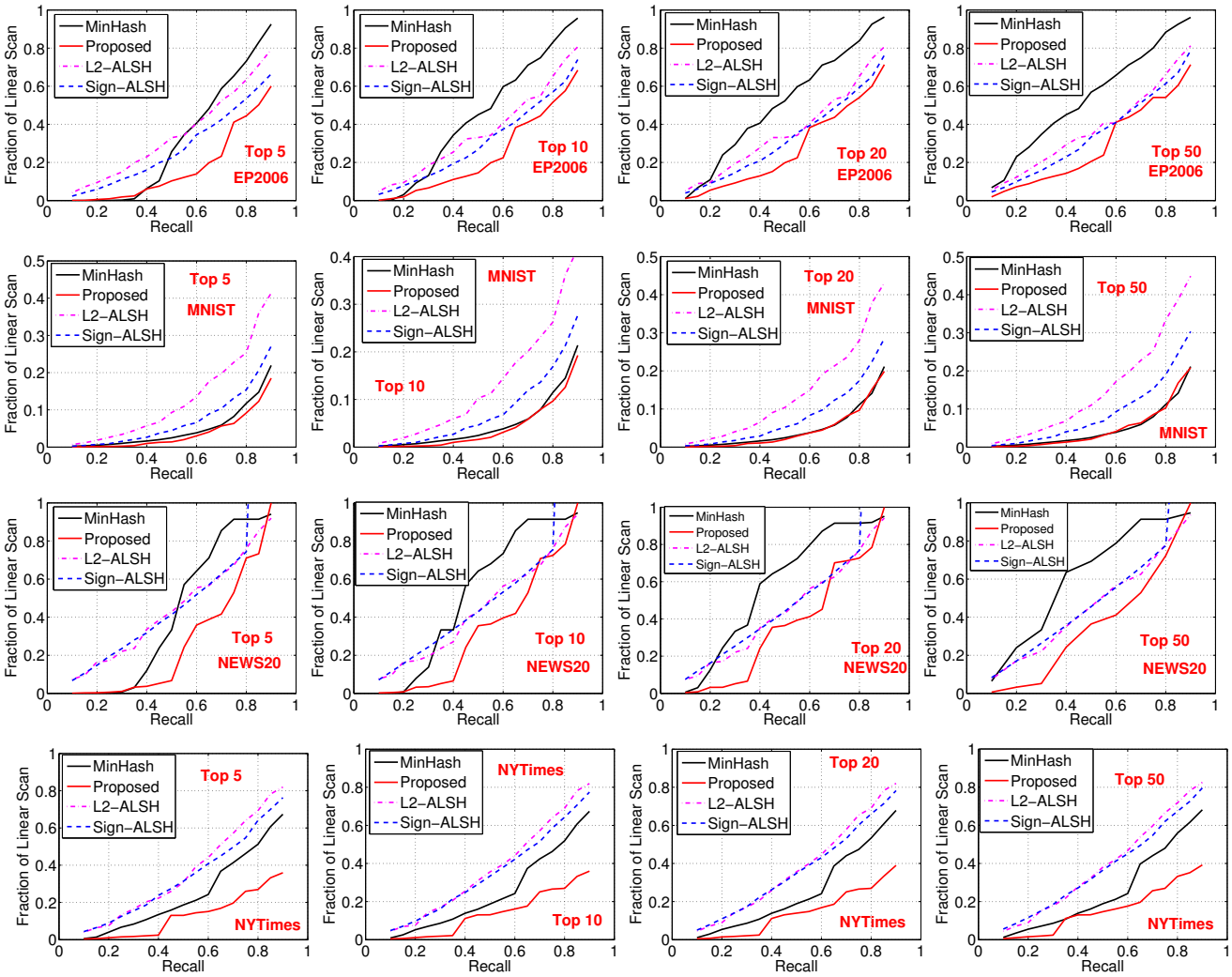

Figure 3: LSH Bucketing Experiments. Average number of points retrieved per query (lower is better), relative to linear scan, evaluated by different hashing schemes at different recall levels, for top-5, top-10, top-20, top-50 nearest neighbors based on set containment (or equivalently inner products), on four datasets. We show that results at the best K and L values chosen at every recall value, independently for each of the four hashing schemes.

7. CONCLUSION AND FUTURE WORK

Minwise hashing (minhash) is a widely popular indexing scheme in practice for similarity search. Minhash was originally designed for estimating set resemblance (i.e., normalized size of set intersections). In many applications the performance of minhash is severely affected because minhash has a bias towards smaller sets. In this study, we propose asymmetric corrections (asymmetric minwise hashing, or MH-ALSH) to minwise hashing that remove this often undesirable bias. Our corrections lead to a provably superior algorithm for retrieving binary inner products in the literature. Rigorous experimental evaluations on the task of retrieving maximum inner products clearly establish that the proposed approach can be significantly advantageous over the existing state-of-the-art hashing schemes in practice, when the desired similarity is the inner product (or containment) instead of the resemblance. Our proposed method requires only minimal modification of the original minwise hashing algorithm and should be straightforward to implement in practice.

Future work: One immediate direction for future work would be *asymmetric consistent weighted sampling* for hashing weighted intersection: $\sum_{i=1}^{D} \min\{x_i, y_i\}$, where x and y are general real-valued vectors. One proposal of the new asymmetric transformation is the following:

$$P(x) = [x; M - \sum_{i=1}^{D} x_i; 0], \qquad Q(x) = [x; 0; M - \sum_{i=1}^{D} x_i],$$

where $M = \max_{x \in \mathcal{C}} \sum_i x_i$. It is not difficult to show that the weighted Jaccard similarity between $P(x)$ and $Q(y)$ is monotonic in $\sum_{i=1}^{D} \min\{x_i, y_i\}$ as desired. At this point, we can use existing methods for consistent weighted sampling [30, 23, 19, 26]. on the new data after asymmetric transformations

Acknowledgement

The work is partially supported by NSF-DMS-1444124, NSF-III-1360971, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137.

8. REFERENCES

[1] P. Agrawal, A. Arasu, and R. Kaushik. On indexing error-tolerant set containment. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 927–938. ACM, 2010.

[2] A. Andoni and P. Indyk. E2lsh: Exact euclidean locality sensitive hashing. Technical report, 2004.

[3] Y. Bachrach, Y. Finkelstein, R. Gilad-Bachrach, L. Katzir, N. Koenigstein, N. Nice, and U. Paquet. Speeding up the xbox recommender system using a euclidean transformation for inner-product spaces. In *RecSys*, 2014.

[4] Y. Bachrach, E. Porat, and J. S. Rosenschein. Sketching techniques for collaborative filtering. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, 2009.

[5] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In *WWW*, pages 131–140, 2007.

[6] A. Z. Broder. On the resemblance and containment of documents. In *the Compression and Complexity of Sequences*, pages 21–29, Positano, Italy, 1997.

[7] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. In *STOC*, pages 327–336, Dallas, TX, 1998.

[8] T. Chandra, E. Ie, K. Goldman, T. L. Llinares, J. McFadden, F. Pereira, J. Redstone, T. Shaked, and Y. Singer. Sibyl: a system for large scale machine learning.

[9] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[10] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, Montreal, Quebec, Canada, 2002.

[11] S. Chaudhuri, V. Ganti, and R. Kaushik. A primitive operatior for similarity joins in data cleaning. In *ICDE*, 2006.

[12] S. Chaudhuri, V. Ganti, and D. Xin. Mining document collections to facilitate accurate approximate entity matching. *Proceedings of the VLDB Endowment*, 2(1):395–406, 2009.

[13] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, and P. Raghavan. On compressing social networks. In *KDD*, pages 219–228, Paris, France, 2009.

[14] G. Cormode and S. Muthukrishnan. Space efficient mining of multigraph streams. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 271–282. ACM, 2005.

[15] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM, 2007.

[16] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokn. Locality-sensitive hashing scheme based on p-stable distributions. In *SCG*, pages 253 – 262, Brooklyn, NY, 2004.

[17] S. Geva and C. M. De Vries. Topsig: Topology preserving document signatures. In *CIKM*, pages 333–338, 2011.

[18] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, 42(6):1115–1145, 1995.

[19] B. Haeupler, M. Manasse, and K. Talwar. Consistent weighted sampling made fast, small, and easy. Technical report, arXiv:1410.4266, 2014.

[20] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pages 136–143, Barbados, 2005.

[21] M. R. Henzinger. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR*, pages 284–291, 2006.

[22] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, pages 604–613, Dallas, TX, 1998.

[23] S. Ioffe. Improved consistent sampling, weighted minhash and L1 sketching. In *ICDM*, pages 246–255, Sydney, AU, 2010.

[24] Y. Jiang, C. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, pages 494–501, Amsterdam, Netherlands, 2007.

[25] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 802–803. ACM, 2006.

[26] P. Li. Min-max kernels. Technical report, arXiv:1503.0173, 2015.

[27] P. Li and A. C. König. Theory and applications b-bit minwise hashing. *Commun. ACM*, 2011.

[28] P. Li, M. Mitzenmacher, and A. Shrivastava. Coding for random projections and approximate near neighbor search. Technical report, arXiv:1403.8144, 2014.

[29] P. Li, A. B. Owen, and C.-H. Zhang. One permutation hashing. In *NIPS*, Lake Tahoe, NV, 2012.

[30] M. Manasse, F. McSherry, and K. Talwar. Consistent weighted sampling. Technical Report MSR-TR-2010-73, Microsoft Research, 2010.

[31] S. Melnik and H. Garcia-Molina. Adaptive algorithms for set containment joins. *ACM Transactions on Database Systems (TODS)*, 28(1):56–99, 2003.

[32] B. Neyshabur and N. Srebro. A simpler and better lsh for maximum inner product search (mips). Technical report, arXiv:1410.5518, 2014.

[33] A. Rajaraman and J. Ullman. *Mining of Massive Datasets*. http://i.stanford.edu/ ullman/mmds.html.

[34] K. Ramasamy, J. F. Naughton, and R. Kaushik. Set containment joins: The good, the bad and the ugly.

[35] A. Shrivastava and P. Li. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (mips). In *NIPS*, Montreal, CA, 2014.

[36] A. Shrivastava and P. Li. Densifying one permutation hashing via rotation for fast near neighbor search. In *ICML*, Beijing, China, 2014.

[37] A. Shrivastava and P. Li. Improved asymmetric locality sensitive hashing (ALSH) for maximum inner product search (MIPS). *arXiv:1410.5410 (submitted to AISTATS)*, 2014.

[38] A. Shrivastava and P. Li. Improved densification of one permutation hashing. In *UAI*, Quebec City, CA, 2014.

[39] A. Shrivastava and P. Li. In defense of minhash over simhash. In *AISTATS*, 2014.

[40] R. Weber, H.-J. Schek, and S. Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *VLDB*, pages 194–205, 1998.