

BagMinHash – Minwise Hashing Algorithm for Weighted Sets

Otmar Ertl
Dynatrace
Linz, Austria
otmar.ertl@dynatrace.com

ABSTRACT

Minwise hashing has become a standard tool to calculate signatures which allow direct estimation of Jaccard similarities. While very efficient algorithms already exist for the unweighted case, the calculation of signatures for weighted sets is still a time consuming task. BagMinHash is a new algorithm that can be orders of magnitude faster than current state of the art without any particular restrictions or assumptions on weights or data dimensionality. Applied to the special case of unweighted sets, it represents the first efficient algorithm producing independent signature components. A series of tests finally verifies the new algorithm and also reveals limitations of other approaches published in the recent past.

CCS CONCEPTS

• **Theory of computation** → **Sketching and sampling**; *Near-est neighbor algorithms*; • **Mathematics of computing** → **Probabilistic algorithms**; *Dimensionality reduction*; • **Information systems** → **Similarity measures**; *Near-duplicate and plagiarism detection*; *Deduplication*;

KEYWORDS

Jaccard similarity; Weighted minwise hashing; Consistent weighted sampling; Locality-sensitive hashing; Sketching algorithms

ACM Reference Format:

Otmar Ertl. 2018. BagMinHash – Minwise Hashing Algorithm for Weighted Sets. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3219819.3220089>

1 INTRODUCTION

Increasing amounts of data require efficient algorithms to keep processing times reasonably small. Sketching algorithms can be used to trade accuracy for less computational costs. In case of tasks such as near-duplicate detection, classification, or clustering, which are based on proximity computations between objects, locality-sensitive hashing enabled large-scale applications by reducing the dimensionality of high-dimensional data [16].

Various distance metrics can be used to model similarity. If objects can be described as sets of features, a very popular metric is

the Jaccard distance $1 - J$ where the Jaccard similarity J of sets A and B is defined as

$$J = \frac{|A \cap B|}{|A \cup B|}. \quad (1)$$

Minwise hashing was the first algorithm to calculate hash signatures (also known as fingerprints) of sets which can be directly used for Jaccard similarity estimation [2]. It maps a set S to an m -dimensional vector $h_S = (h_{S1}, h_{S2}, \dots, h_{Sm})$ with statistically independent components. Ignoring potential hash collisions, the probability that same components of two different signatures h_A and h_B for sets A and B have identical value is equal to the Jaccard similarity

$$\mathbb{P}(h_{Ai} = h_{Bi}) = J, \quad \forall i \in \{1, 2, \dots, m\}. \quad (2)$$

This property allows unbiased estimation of J given the fraction of matching components

$$\hat{J}(h_A, h_B) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}(h_{Ai} = h_{Bi}) \quad (3)$$

where $\mathbb{1}$ denotes the indicator function. The variance of this estimator is

$$\text{Var}(\hat{J}(h_A, h_B)) = J(1 - J)/m, \quad (4)$$

because the number of equal components is binomially distributed with success probability J .

The calculation of signatures of size m using minwise hashing is expensive, because the computation time scales like $O(mn)$ with n denoting the cardinality of the input set. Typical values for m lie in the range from 100 to 10 000 and n can be in the millions [29].

1.1 Advanced Minwise Hashing Techniques

One permutation hashing is able to reduce the computational costs significantly and has optimal time complexity $O(m + n)$ [21]. However, this approach leads to biased Jaccard estimates for small n compared to m . This problem can be remedied by introducing a postprocessing step called densification [32]. Unfortunately, this approach increases the variance of the estimator compared to (4), because it breaks the statistical independence of individual signature components leading to correlated occurrences of matching components.

Recent developments such as fast similarity sketching [7] with time complexity $O(m \log m + n)$ or the SuperMinHash algorithm [10] with time complexity $O(m \log^2 m + n)$ also lead to signature component matches that are statistically dependent. But their correlation is negative, which even reduces the variance and thus improves the estimation error for small n . However, all these algorithms cannot replace the original approach, if independence of components is important.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '18, August 19–23, 2018, London, United Kingdom

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5552-0/18/08...\$15.00

<https://doi.org/10.1145/3219819.3220089>

An orthogonal technique that can be used with any minwise hashing approach is b -bit hashing [19, 20]. It maps signature components uniformly to b -bit integer values. The higher probability of matching signature components by chance introduces an estimation bias which is compensated by the estimator

$$\hat{j}^b(h_A^b, h_B^b) = \frac{\hat{J}(h_A^b, h_B^b) - 2^{-b}}{1 - 2^{-b}} = \frac{\left(\frac{1}{m} \sum_{i=1}^m \mathbf{1}(h_{Ai}^b = h_{Bi}^b)\right) - 2^{-b}}{1 - 2^{-b}}.$$

Here $h_A^b, h_B^b \in \{0, 1, \dots, 2^b - 1\}^m$ denote the signatures after reducing all components to b bits. The corresponding loss of information increases the variance

$$\text{Var}(\hat{j}^b(h_A^b, h_B^b)) = \text{Var}(\hat{J}(h_A, h_B)) + \frac{1 - J}{(2^b - 1)m}$$

where $\text{Var}(\hat{J}(h_A, h_B))$ denotes the variance of the signature components before reduction to b -bit values, which is given by (4) in case of original minwise hashing. Hence, to get the same estimation error, the signature size m needs to be increased. Nevertheless, if the Jaccard similarity J is not too small, the memory footprint of the whole signature can be significantly reduced. However, this comes with higher computational costs for calculating the additional signature components. Therefore, a fast hashing scheme is even more important for b -bit hashing.

1.2 Weighted Minwise Hashing

Sometimes it is more convenient to represent objects as bags also known as multisets, where each element is associated with a non-negative weight. For example, words or shingles in text documents are often weighted according to a measure called term frequency-inverse document frequency (TF-IDF) [16]. A similarity measure for bags is the weighted Jaccard similarity which generalizes (1). If two bags are described by weight functions w_A and w_B , respectively, which return a nonnegative weight for each element of a universe D , the weighted Jaccard similarity is given by

$$J = \frac{\sum_{d \in D} \min(w_A(d), w_B(d))}{\sum_{d \in D} \max(w_A(d), w_B(d))}. \quad (5)$$

If all weights are binary with values either zero or one, and we define sets A and B as $A := \{d \in D : w_A(d) > 0\}$ and $B := \{d \in D : w_B(d) > 0\}$, respectively, (5) will simplify to (1).

Finding signatures for the weighted case satisfying (2) is more challenging. For integral weights it is possible to add a corresponding number of replications and to apply conventional minwise hashing [13]. This is not efficient, because elements with large weights blow up the effective data size and hence also the computation time. Some optimizations to reduce the number of replications and to apply unweighted minwise hashing to real weights are described in [11] and [12].

The first algorithm inherently designed for real weights is based on rejection sampling [3, 15]. A faster approach is consistent weighted hashing with amortized time complexity $O(mn)$ [25]. Here $n = |\{d \in D : w(d) > 0\}|$ denotes the number of elements having nonzero weight. A further development of this concept called improved consistent weighted sampling (ICWS) achieves the same time complexity as worst case and has become the state-of-the-art algorithm for weighted minwise hashing [14].

Recently, a weighted hashing scheme was presented, which can lead to significant speedups [31]. The algorithm requires the universe D to be fixed and that weight upper bounds $w_{\max}(d)$ exist for all $d \in D$. The time complexity of this algorithm depends on the sparsity s of the weighted set and is $O(ms^{-1} + n)$. The sparsity is defined as $s := (\sum_{d \in D} w(d)) / (\sum_{d \in D} w_{\max}(d))$. If s^{-1} is small compared to the number of elements with nonzero weights ($s^{-1} \ll n$), the computation time can be significantly smaller compared to the ICWS algorithm. However, general applicability is limited by the required a priori knowledge of sharp upper bounds for $w_{\max}(d)$. Moreover, space requirements scale at least linearly with dimensionality $|D|$.

1.3 Applications

Weighted minwise hashing is used in many applications such as near duplicate image detection [4], duplicate news story detection [1], source code deduplication [26], time series indexing [24], hierarchical topic extraction [11], or malware classification [29] and detection [9]. Unfortunately, it was reported that weighted minwise hashing using the ICWS algorithm can be very expensive due to the $O(mn)$ time complexity, which sometimes limits its application [9, 29] or requires massive computing power using GPUs [26].

Recent works suggest to use weighted minwise hashing for machine learning too. As example, weighted minwise hashing was successfully applied to train generalized min-max kernel support vector machines [18, 22]. Furthermore, asymmetric locality-sensitive hashing [33], which can be realized using weighted minwise hashing, was used for efficient deep learning [34]. Finally, it could also be applied to random forests that are constructed using the weighted Jaccard index as similarity measure [30].

2 PRELIMINARIES

The aforementioned improvements on unweighted minwise hashing [7, 10, 32] achieve much better time complexities than $O(mn)$ of the original approach [2]. This was our main motivation to develop an algorithm for the weighted case, which reduces the $O(mn)$ complexity of the ICWS algorithm in a similar fashion.

Both, original minwise hashing as well as the ICWS algorithm calculate for every input element m different independent hash values, one for each signature component. Hence, n different hash values are computed for every signature component. The smallest of them finally defines the corresponding signature value. If $n \gg m$, an element is expected to generate at most one hash value small enough to contribute to the signature. Therefore, in many cases, it would be sufficient to determine for each input element only the smallest hash value including its assigned signature component. If this could be achieved in a more direct way, without the need to calculate all m different hash values first, a more efficient signature computation would be possible.

2.1 Discretization Error

The ICWS algorithm is exact for any nonnegative real weights. This is a nice property from a theoretical point of view. However, computers always approximate real numbers by values from a finite set. Therefore, we analyze the impact of weight discretization on the Jaccard index calculation. Let $V = \{v_0, v_1, \dots, v_L\}$ be the set

of discrete weight values satisfying $0 = v_0 < v_1 < \dots < v_L$. The weight function $w(d)$ can be approximated by

$$\tilde{w}(d) := v_{k(d)} \quad \text{with} \quad k(d) := \max(\{l : v_l \leq w(d)\}). \quad (6)$$

We assume that the discretization V is chosen in such a way that for all occurring weights $w(d)$

$$\tilde{w}(d) \leq w(d) \leq \tilde{w}(d)(1 + \varepsilon) \quad (7)$$

is fulfilled, where $\varepsilon \ll 1$ is a small positive constant. If an approximation \tilde{J} of the generalized Jaccard index (5) is calculated using approximations \tilde{w}_A and \tilde{w}_B for weight functions w_A and w_B , respectively, following error bounds can be established

$$J(1 - \varepsilon) \leq J(1 + \varepsilon)^{-1} \leq \tilde{J} \leq J(1 + \varepsilon). \quad (8)$$

They show that the relative error of \tilde{J} is limited by ε . Since $J \in [0, 1]$, ε is also a bound for the absolute error.

Assumption (7) is nothing special as it is usually satisfied by floating-point numbers as long as $w(d)$ is neither subnormal nor too large. As example, assume V represents the set of all non-negative single-precision floating-point numbers for which $L = 0x7FFFFFFF = 2\,139\,095\,039$. If all occurring nonzero weights are within the normal range of single-precision floating-point numbers which is roughly from $1.18E-38$ to $3.40E38$, (7) will be satisfied and ε will correspond to the machine epsilon which is approximately $1.19E-7$.

It makes sense to compare the discretization error $J\varepsilon$ to the error introduced by estimating the weighted Jaccard index from hash signatures using (3). According to (4) the standard error is $\sqrt{J(1-J)/m}$. The discretization error $J\varepsilon$ can be ignored as long as $\varepsilon \ll \sqrt{(1-J)/(Jm)}$. For instance, consider an extreme case where J is close to 1 and m is huge. As example, $J = 1 - 10^{-6}$ and $m = 10^6$, for which the estimation error would be in the order of 10^{-6} . The error is still one magnitude larger than the discretization error made by using single-precision weights and can therefore be neglected. This example shows that single precision will be sufficient for most applications.

The assumption of discrete weights is important for our new approach. As weight values are already discrete on computers, this is not really a restriction or a disadvantage compared to existing approaches, because one could simply choose V based on the given discretization. However, as demonstrated, a coarser choice will often be sufficient as long as the additional discretization error on the Jaccard index remains small compared to the expected estimation error.

2.2 Random Numbers

We assume an ideal hash function whose output can be treated as uniformly distributed and truly random number. Hashing the same input multiple times, but each time enriched with a different value from some predefined integer sequence, allows generating a sequence of independent hash values. Therefore, an ideal hash function can be used as ideal pseudorandom number generator (PRNG) which produces random bit sequences of arbitrary length as function of some given seed value. Hereinafter, we may write random number instead of pseudorandom number and always mean some value produced by a PRNG.

3 BAGMINHASH ALGORITHM

In the following we assume some fixed weight discretization V as discussed in Section 2.1. As starting point towards our new algorithm, we first consider the hash signature defined by

$$h_i := \min_{(d,l) \in Q} x_{dli} \quad (9)$$

with $Q := \bigcup_{d \in D} \bigcup_{l=1}^{k(d)} \{(d, l)\} = \bigcup_{d \in D} \{d\} \times \{1, 2, \dots, k(d)\}$ and $k(d)$ denoting the function introduced in (6). x_{dli} are independent exponentially distributed pseudorandom numbers $x_{dli} \sim \text{Exponential}(\alpha(v_l - v_{l-1}))$ with rate parameter $\alpha(v_l - v_{l-1})$ where α is some positive constant. The choice of α is arbitrary, because scaling all x_{dli} by the same factor preserves equalities between signature components h_i .

The probability, that the i -th signature component is equal for two different weighted sets A and B , is given by

$$\begin{aligned} \mathbb{P}(h_{Ai} = h_{Bi}) &= \mathbb{P}\left(\min_{(d,l) \in Q_A} x_{dli} = \min_{(d,l) \in Q_B} x_{dli}\right) = \\ &= \mathbb{P}\left(\min_{(d,l) \in Q_A \cap Q_B} x_{dli} < \min_{(d,l) \in Q_A \Delta Q_B} x_{dli}\right) = \mathbb{P}(y_{1i} < y_{2i}). \end{aligned} \quad (10)$$

Here we used that all random variables x_{dli} must be different as they are drawn from a continuous distribution. y_{1i} and y_{2i} are given by

$$y_{1i} := \min_{(d,l) \in Q_A \cap Q_B} x_{dli}, \quad y_{2i} := \min_{(d,l) \in Q_A \Delta Q_B} x_{dli}.$$

They are both the minimum of independent exponentially distributed random variables. Consequently, y_{1i} and y_{2i} are also independent and exponentially distributed. The corresponding rate parameters λ_1 and λ_2 can be calculated as the sum of all contributing rates [28]

$$\begin{aligned} \lambda_1 &= \sum_{(d,l) \in Q_A \cap Q_B} \alpha(v_l - v_{l-1}) = \alpha \sum_{d \in D} v_{\min(k_A(d), k_B(d))} \\ &= \alpha \sum_{d \in D} \min(v_{k_A(d)}, v_{k_B(d)}) = \alpha \sum_{d \in D} \min(\tilde{w}_A(d), \tilde{w}_B(d)) \end{aligned}$$

and similarly

$$\lambda_2 = \alpha \sum_{d \in D} \max(\tilde{w}_A(d), \tilde{w}_B(d)) - \min(\tilde{w}_A(d), \tilde{w}_B(d)).$$

Among independent exponentially distributed random numbers, the probability, that one is the smallest, is proportional to the rate parameter of its exponential distribution. Hence $\mathbb{P}(y_{1i} < y_{2i}) = \lambda_1 / (\lambda_1 + \lambda_2)$ [28], which finally gives in combination with (10)

$$\mathbb{P}(h_{Ai} = h_{Bi}) = \frac{\sum_{d \in D} \min(\tilde{w}_A(d), \tilde{w}_B(d))}{\sum_{d \in D} \max(\tilde{w}_A(d), \tilde{w}_B(d))} = \tilde{J}.$$

Therefore, signatures defined by (9) allow unbiased estimation of \tilde{J} using (3). They can also be used to estimate J , if the error introduced by the weight discretization can be neglected as discussed in Section 2.1. The number of equal components of two signatures follows a binomial distribution with success probability \tilde{J} , because individual components of the same signature are, like for original minwise hashing, statistically independent by definition.

Algorithm 1 is a straightforward implementation of signature definition (9). First, the signature components are initialized with infinity. For each pair (d, l) a PRNG R is created with the pair as

Algorithm 1: Straightforward implementation of signature definition (9).

Input: w
Output: h_1, h_2, \dots, h_m
 $(h_1, h_2, \dots, h_m) \leftarrow (\infty, \infty, \dots, \infty)$
forall $d \in D$ such that $w(d) \geq v_1$ **do**
 forall $l \in \{1, 2, \dots, L\}$ such that $v_l \leq w(d)$ **do**
 $R \leftarrow$ **new** PRNG with seed (d, l)
 for $i \leftarrow 1$ **to** m **do**
 $x \leftarrow R[\text{Exponential}(\alpha(v_l - v_{l-1}))]$
 if $x < h_i$ **then** $h_i \leftarrow x$

Algorithm 2: Transformation of a real-valued signature into a b -bit hash signature.

Input: h_1, h_2, \dots, h_m
Output: $h_1^b, h_2^b, \dots, h_m^b \in \{0, 1, \dots, 2^b - 1\}$
for $i \leftarrow 1$ **to** m **do**
 $R \leftarrow$ **new** PRNG with seed h_i
 $h_i^b \leftarrow R[\text{Uniform}(\{0, 1, \dots, 2^b - 1\})]$

seed to ensure independence. R is then used to generate m exponentially distributed random values which are used to update the signature components. The parameter of R describes the distribution from which a random value is drawn using R as random bit source. Obviously, Algorithm 1 is not very efficient, because its time complexity is $O(Lmn)$. This is much worse than $O(mn)$ of the ICWS algorithm, because L is expected to be huge. For example, L is in the billions, if weights are discretized using single-precision floating-point numbers. In the oncoming sections we will describe methods to make the calculation of the new signature much more efficient.

The signature components defined by (9) are all nonnegative real numbers. However, as mentioned in Section 1.1, integer values with a predefined number of bits are often more preferable. Similar to the proposal in [14], Algorithm 2 extracts b -bit integer values from the real-valued signature. A PRNG is initialized for each signature component with h_i as seed. It is used to generate a uniform random b -bit integer number which defines the value of the corresponding component of the transformed signature.

3.1 Interpretation as Poisson Process

Since the random variables x_{dli} in (9) are exponentially distributed with rate parameter $\alpha(v_l - v_{l-1})$, they can also be interpreted as first points of independent Poisson processes P_{dli} with rates $\alpha(v_l - v_{l-1})$, respectively. Poisson processes have some nice properties. Points can be generated in ascending order, because distances between successive points are independent and exponentially distributed. Moreover, Poisson processes can be combined and split [28].

These properties can be exploited to generate the first points of all P_{dli} in a different but statistically equivalent way. We consider the combined process $P_{dl} := \bigcup_{i=1}^m P_{dli}$ which has rate $\alpha m(v_l - v_{l-1})$. While generating points for P_{dl} in ascending order, P_{dl} is split again into the m constituent subprocesses P_{dli} . This can be done by distributing each point of P_{dl} randomly and uniformly to one of the m subprocesses P_{dli} . As soon as we have assigned at least one point to each of those subprocesses, which means we have found all their first points, we are done. Since the constant α is a free parameter, we set $\alpha := m^{-1}$ for simplicity such that P_{dl} has rate $v_l - v_{l-1}$ in the following.

Algorithm 3: Improved version of Algorithm 1 by generating random values in ascending order for fixed d and l .

Input: w
Output: h_1, h_2, \dots, h_m
 $(h_1, h_2, \dots, h_m) \leftarrow (\infty, \infty, \dots, \infty)$
forall $d \in D$ such that $w(d) \geq v_1$ **do**
 forall $l \in \{1, 2, \dots, L\}$ such that $v_l \leq w(d)$ **do**
 $R \leftarrow$ **new** PRNG with seed (d, l)
 $P \leftarrow$ **new** PoissonProcess($0, R, l - 1, l, w(d)$)
 $P.\text{Next}()$
 while $P.x \leq h_{\max}$ **do**
 if $P.x < h_{p,i}$ **then** $h_{p,i} \leftarrow P.x$
 $P.\text{Next}()$

Algorithm 3 demonstrates this idea. For each pair (d, l) a Poisson process P_{dl} with rate $v_l - v_{l-1}$ is simulated. A helper class given by Algorithm 4 is used to represent the corresponding state. A new Poisson process is initialized with starting point $x = 0$ and with indices $p = l - 1$ and $q = l$ which define the rate as $v_q - v_p$. Furthermore, since the Poisson process object requires an independent random source for point generation, a new PRNG is created. By choosing d and l as seed we ensure that all processes P_{dl} are independent. The helper class has a member function $\text{Next}()$ for generating the next point of the Poisson process by incrementing x by a value drawn from an exponential distribution with rate parameter $v_q - v_p$. In addition, a uniform random number i drawn from $\{1, 2, \dots, m\}$ defines the index of the subprocess P_{dli} the new point belongs to. The helper class also has some more functions as well as an additional field w for the weight $w(d)$ associated with d , which are not yet used in the context of Algorithm 3.

Instead of executing the while-loop until at least one point was generated for all m subprocesses P_{dli} , Algorithm 3 keeps track of $h_{\max} := \max(h_1, h_2, \dots, h_m)$. By definition, points greater than h_{\max} are not able to change any of the values h_1, h_2, \dots, h_m . Therefore, as soon as some point is greater than h_{\max} the Poisson process can be stopped. If h_{\max} is simultaneously updated each time some signature value h_i is replaced by a lower value, h_{\max} will decrease over time and the termination condition is satisfied earlier in subsequent iterations of d and l .

The efficient maintenance of h_{\max} is demonstrated by Algorithm 5. A binary tree is constructed over m leaf nodes with values h_1, h_2, \dots, h_m , all starting from positive infinity. The value of a parent node is defined to be the maximum of both children. Therefore, the root node always represents h_{\max} . If a Poisson process generates some point $x < h_{\max}$, an update of h_i and also of its ancestors might be necessary. Algorithm 5 makes a corresponding bottom-up traversal until no further change is necessary.

Conditioned on $x < h_{\max}$ the probability that some node is replaced by a smaller value is less than the reciprocal number of leaves in the corresponding subtree. This means, given $x < h_{\max}$, the probability, that h_i is updated, is (of course) at most 1. The probability, that the parent of h_i is updated, is at most $\frac{1}{2}$, because it is equally likely that the parent value is given by the sibling of h_i , and therefore an update of h_i will not change the parent. Continuing in this way shows that the expected number of node updates is bounded by the geometric series $1 + 1/2 + 1/4 + \dots$ and therefore takes amortized constant time and does not contribute to the overall time complexity of Algorithm 3.

Algorithm 4: Helper class for the representation of the combined Poisson process
 $P_{d,p;q} = \bigcup_{l=p+1}^q P_{dl} = \bigcup_{l=p+1}^q \bigcup_{i=1}^m P_{dli}$.

```

class PoissonProcess
  fields
    x                                current point
    R                                pseudorandom number generator
    p, q                            define the rate as  $v_q - v_p$ 
    w                                weight of  $d$ 
    i                                index of  $P_{dli}$  point  $x$  is belonging to
    PoissonProcess( $x', R', p', q', w'$ )    constructor, initializes a new object
    x ←  $x'$ 
    R ←  $R'$ 
    p ←  $p'$ , q ←  $q'$ 
    w ←  $w'$ 

  Next()                            generates the next point
    x ←  $x + R[\text{Exponential}(v_q - v_p)]$ 
    i ←  $R[\text{Uniform}(\{1, 2, \dots, m\})]$ 

  Split()                            splits process into two parts
    r ←  $\lfloor (p + q)/2 \rfloor$ 
     $R' \leftarrow$  new PRNG with seed ( $x, r$ )
    if  $R[\text{Bernoulli}((v_r - v_p)/(v_q - v_p))] = 1$  then
       $P' \leftarrow$  new PoissonProcess( $x, R', r, q, w$ )
      q ← r
    else
       $P' \leftarrow$  new PoissonProcess( $x, R', p, r, w$ )
      p ← r
    return  $P'$ 

  Splittable()
    return  $p + 1 < q$ 

  PartiallyRelevant()
    return  $v_{p+1} \leq w$ 

  FullyRelevant()
    return  $v_q \leq w$ 

```

Hence, the complexity is primarily dominated by the number of expected Next() calls. At the beginning, when all h_i and therefore also h_{\max} are still infinite, the while-loop is executed until all h_i have been updated at least once. Since points of P_{dl} are calculated in ascending order, h_{\max} will become finite exactly when the last infinite signature value is overwritten. Thus, h_{\max} will also have the same value as the current point of P and the termination condition of the while-loop will be immediately satisfied after advancing to the next point using $P.\text{Next}()$. In analogy of the coupon collector's problem [6], the average number of required steps until at least one point is assigned to each signature component will be mH_m where $H_m := 1 + \frac{1}{2} + \dots + \frac{1}{m} = O(\log m)$ denotes the m -th harmonic number.

Dependent on the weight $w(d)$ of the input element d , up to L inner for-loop iterations are required. Therefore, the expected number of Next() calls needed for the first input element is at most $L(1 + mH_m)$. The number of required cycles will be lower for further elements, because h_{\max} decreases with the number of inserted elements and the termination condition will be satisfied earlier. The decrease of h_{\max} will be faster, if elements with larger weights are processed first, because the corresponding random points are likely to be smaller than for elements with smaller weights. Therefore, it is preferable to process input elements in descending and not in ascending order of weight. If we assume random ordering, it can be shown that the expected number of Next() calls needed for the j -th input element is bounded by $L(1 + mH_m/j)$. Therefore, the

Algorithm 5: To efficiently keep track of the global maximum $h_{\max} := \max(h_1, h_2, \dots, h_m)$ while values h_i are successively decreased over time, the array is extended by h_{m+1}, \dots, h_{2m-1} which correspond to the parent nodes in a binary tree spanned over h_1, h_2, \dots, h_m . $h_{m+\lceil i/2 \rceil}$ is the parent of h_i . A parent node is defined as the maximum of its children $h_i = \max(h_{2(i-m)-1}, h_{2(i-m)})$ for $i > m$. If a leaf node h_i with $i \leq m$ is replaced by a smaller value $x < h_i$, following procedure will be necessary to update the root node h_{2m-1} which also represents h_{\max} .

```

Input: x, i
h ← x
while h <  $h_i$  do
   $h_i \leftarrow h$ 
  i ← m +  $\lceil i/2 \rceil$                 calculate parent index
  if i ≥ 2m then break
  h ←  $\max(h_{2(i-m)-1}, h_{2(i-m)})$ 

```

time complexity of Algorithm 3 is limited by $\sum_{j=1}^n L(1 + mH_m/j) = O(Lm \log^2 m + Ln)$. This is already a significant improvement over $O(Lmn)$ of Algorithm 1 for $n \gg m$ as we have eliminated the factor m that accompanies n .

However, since L can be very large, further improvements are necessary to compete with $O(mn)$ of the ICWS algorithm. Algorithm 3 is only useful for small L . In particular, by choosing $L = 1$ and $V = \{0, 1\}$ binary weight values can be accurately described. Therefore, Algorithm 3 can be used for the unweighted case and represents an interesting alternative to the advanced minwise hashing algorithms discussed in Section 1.1. The reason is that the time complexity $O(m \log^2 m + n)$ is comparable to other advanced minwise hashing algorithms discussed in Section 1.1. But, in contrast to them, the signature components are statistically independent like for original minwise hashing.

3.2 In-Order Point Generation

The individual treatment of processes P_{dl} is responsible for the factor L in the time complexity of Algorithm 3. To eliminate or at least to reduce this factor, we need to process all P_{dl} for a given input element d in a collective fashion. We consider the combined process $P_d := \bigcup_{l=1}^L P_{dl}$ which has rate $\sum_{l=1}^L v_l - v_{l-1} = v_L$. In this context we will call the processes P_{dl} elementary subprocesses. The idea is now to generate points for P_d in ascending order. To comply with (9) we must filter points that belong to elementary subprocesses P_{dl} which satisfy $v_l \leq w(d)$. In the following, we will call these elementary subprocesses and also their points relevant.

The properties of Poisson processes allow iterative splitting of P_d until we arrive at some elementary subprocess. For this, we introduce the notation $P_{d,p;q} := \bigcup_{l=p+1}^q P_{dl}$, which means that $P_d = P_{d,0:L}$ and $P_{dl} = P_{d,l-1:l}$. Assume we have a process $P_{d,p;q}$ for which the first point x has already been generated. If it is not an elementary process, which is equivalent to $p + 1 < q$, it is called splittable and can be divided into two subprocesses $P_{d,p;r}$ and $P_{d,r;q}$ with $p < r < q$. The probability that x is a point of $P_{d,p;r}$ or $P_{d,r;q}$ is proportional to the corresponding rates, respectively. Hence, the result of a Bernoulli trial can be used to decide to which subprocess the point x actually belongs to. The first point of the other subprocess can be generated by adding to x a random number drawn from an exponential distribution with rate parameter equal to the rate of that subprocess.

Algorithm 4 shows how the splitting operation can be realized. The Split() method first calculates the split index as $r = \lfloor (p + q)/2 \rfloor$,

such that the weight index range is divided into two parts of almost equal size. A Bernoulli trial with success probability $(v_r - v_p)/(v_q - v_p)$ determines whether x is a point of $P_{d,p:r}$. If not, x must belong to $P_{d,r:q}$. The parent process object can be reused for the elected subprocess by simply adjusting either q or p , respectively. A new process object P' is needed to represent the second subprocess. It is initialized with starting point x and a new PRNG. Independence can be ensured by using x together with the splitting index r as seed. The splitting index makes the seed unique, because further splittings are not possible with same index. The object of the second child P' is finally returned by the Split() method. A call to $P'.Next()$ must follow later to get the first point of P' .

In order to calculate the relevant points of P_d in ascending order, we first generate the very first point of P_d . Next, we determine, to which elementary subprocess the point belongs to. The corresponding subprocess can be found through iterative splitting, always proceeding with the child that contains the point of the parent process. This procedure corresponds to searching for the smallest point in a binary tree, where each node is equal to the minimum of both children. The next smallest point of P_d can be found analogously. Among all subprocesses, which resulted from all splittings so far, we choose the subprocess with smallest current point and restart the same splitting procedure as before from there. In this way all points of P_d and its associated elementary subprocesses P_{dl} can be found in ascending order.

Since we are only interested in relevant points, we must omit irrelevant ones. However, v_L is often much larger than $w(d)$, which makes point-wise filtering not very efficient. Luckily, the hierarchical structure allows efficient skipping of irrelevant points. A subprocess $P_{d,p:q}$ only contains relevant points, if at least one of its elementary processes is relevant, which is the case if $v_{p+1} \leq w(d)$. In this case, we say $P_{d,p:q}$ is partially relevant. If $P_{d,p:q}$ is not partially relevant, it does not contain any relevant points and can immediately be discarded without further processing. We also define a subprocess $P_{d,p:q}$ to be fully relevant, if all its elementary subprocesses and corresponding points are relevant, which is the case if $v_q \leq w(d)$. Full relevance implies partial relevance. If a process is an elementary subprocess and therefore not further splittable, partial and full relevance will be equivalent.

Algorithm 6 called BagMinHash 1 implements the described procedure. The algorithm loops over all elements having nonzero discrete weight ($w(d) \geq v_1$). For each element d a new Poisson process object P is created that represents P_d . The weight $w(d)$ is passed to the Poisson process object, which is needed by the PartiallyRelevant() and FullyRelevant() member functions. After generating the first point and updating the corresponding signature value in case P is already fully relevant, a while-loop is started. Therein, the process is split as long as possible and as long as it is partially relevant. After each split operation, P and also its sibling P' are checked for full relevance. If this is the case, the corresponding signature value $h_{P,i}$ or $h_{P',i}$ and if necessary also h_{\max} are updated. These early updates allow leaving the outer while-loop as soon as possible. If the sibling P' is partially relevant, it needs to be processed later and must be saved in the meantime. However, since the current point of P' represents a lower bound for all its further points, P' can be discarded as a whole, if $P'.x$ is already larger than

Algorithm 6: BagMinHash 1.

```

Input:  $w$ 
Output:  $h_1, h_2, \dots, h_m$ 
 $(h_1, h_2, \dots, h_m) \leftarrow (\infty, \infty, \dots, \infty)$ 
forall  $d \in D$  such that  $w(d) \geq v_1$  do
     $R \leftarrow$  new PRNG with seed  $d$ 
     $P \leftarrow$  new PoissonProcess( $0, R, 0, L, w(d)$ )
     $P.Next()$ 
    if  $P.FullyRelevant()$  and  $P.x < h_{P,i}$  then  $h_{P,i} \leftarrow P.x$ 
    create new min-heap for Poisson processes  $P$  with  $P.x$  as key
    while  $P.x \leq h_{\max}$  do
        while  $P.Splittable()$  and  $P.PartiallyRelevant()$  do
             $P' \leftarrow P.Split()$ 
            if  $P.FullyRelevant()$  and  $P.x < h_{P,i}$  then  $h_{P,i} \leftarrow P.x$ 
            if  $P'.PartiallyRelevant()$  then
                 $P'.Next()$ 
                if  $P'.FullyRelevant()$  and  $P'.x < h_{P',i}$  then  $h_{P',i} \leftarrow P'.x$ 
                if  $P'.x \leq h_{\max}$  then push  $P'$  to heap
            if  $P.FullyRelevant()$  then
                 $P.Next()$ 
                if  $P.x < h_{P,i}$  then  $h_{P,i} \leftarrow P.x$ 
                if  $P.x \leq h_{\max}$  then push  $P$  to heap
            if heap is empty then break
         $P \leftarrow$  pop from heap

```

h_{\max} . Otherwise, we put P' into a min-heap with $P'.x$ as key. In this way the subprocess with smallest current point can be efficiently retrieved later.

The inner while-loop can be left due to two reasons: First, if P is no longer partially relevant, which may be the consequence of the last split operation, we can immediately discard P . Second, if P is no longer splittable and corresponds to an elementary subprocess, while P is still partially relevant, it must also be fully relevant. In this case, we generate the next point of P , update the corresponding signature value if necessary, and push it into the min-heap. In both cases we continue with the process having the smallest current point, which can be retrieved from the min-heap. Once the point of the current process is larger than h_{\max} , which also means that all remaining points of any other processes are larger, we are finished as further signature value updates can be ruled out.

The time complexity of this algorithm can be bounded by $O(\varphi(m, L) + (\log L)(\log \log L)n)$ where φ denotes a function independent of n . The reason is that for large data sizes $n \gg m$, point generation can be stopped most of the time after finding the smallest relevant point. The corresponding binary search over all L elementary subprocesses explains the $\log L$ factor. The min-heap insertions and removals contribute the $\log \log L$ factor. Hence, if $\log L$ is much smaller than m , the new approach is potentially much faster than ICWS.

3.3 Optimization

The new algorithm can be further optimized. The processing time for inserting a new element d decreases with the number of already inserted elements, because h_{\max} is continuously decreasing. In particular, the processing time of the very first element takes a lot of time, because at least one point needs to be found for each signature position. However, if there are much more input elements to come, it is very likely that most of these signature values are overwritten by points from subsequent input elements. Therefore,

Algorithm 7: BagMinHash 2, an optimized variant of Algorithm 6.

Input: w
Output: h_1, h_2, \dots, h_m
 $(h_1, h_2, \dots, h_m) \leftarrow (\infty, \infty, \dots, \infty)$
create empty buffer
forall $d \in D$ such that $w(d) \geq v_1$ **do**
 create new min-heap for Poisson processes P with $P.x$ as key
 $R \leftarrow$ **new** PRNG with seed d
 $P \leftarrow$ **new** PoissonProcess($0, R, 0, L, w(d)$)
 $P.Next()$
 if $P.FullyRelevant()$ **and** $P.x < h_{p,i}$ **then** $h_{p,i} \leftarrow P.x$
 while $P.x \leq h_{\max}$ **do**
 while $P.Splittable()$ **and** $P.PartiallyRelevant()$ **and not** $P.FullyRelevant()$ **do**
 $P' \leftarrow P.Split()$
 if $P.FullyRelevant()$ **and** $P.x < h_{p,i}$ **then** $h_{p,i} \leftarrow P.x$
 if $P'.PartiallyRelevant()$ **then**
 $P'.Next()$
 if $P'.FullyRelevant()$ **and** $P'.x < h_{p',i}$ **then** $h_{p',i} \leftarrow P'.x$
 if $P'.x \leq h_{\max}$ **then** push P' to heap
 if $P.FullyRelevant()$ **then**
 push P to heap
 break
 if heap is empty **then break**
 $P \leftarrow$ pop from heap
 add all heap elements P with $P.x \leq h_{\max}$ to buffer
 create new min-heap from all buffer elements P with $P.x \leq h_{\max}$
 while not heap is empty **do**
 $P \leftarrow$ pop from heap
 if $P.x > h_{\max}$ **then break**
 while $P.Splittable()$ **and** $P.PartiallyRelevant()$ **do**
 $P' \leftarrow P.Split()$
 if $P.FullyRelevant()$ **and** $P.x < h_{p,i}$ **then** $h_{p,i} \leftarrow P.x$
 if $P'.PartiallyRelevant()$ **then**
 $P'.Next()$
 if $P'.FullyRelevant()$ **and** $P'.x < h_{p',i}$ **then** $h_{p',i} \leftarrow P'.x$
 if $P'.x \leq h_{\max}$ **then** push P' to heap
 if $P.FullyRelevant()$ **then**
 $P.Next()$
 if $P.x < h_{p,i}$ **then** $h_{p,i} \leftarrow P.x$
 if $P.x \leq h_{\max}$ **then** push P to heap

we can save some computation time, if we divide the processing of individual input elements into two phases. In a first phase, we calculate only the smallest relevant point of P_d for each input element d and store any subprocesses, which result from splittings and which have first points not greater than h_{\max} , in a buffer for later processing. If there are enough input elements $n \gg m$, h_{\max} will already be very small after the first step. In a second phase, we put all buffered processes into a new min-heap, which allows processing them in order of their current points. Subprocesses of further splittings are again pushed into this min-heap. As soon as h_{\max} is smaller than the first element of the min-heap, the signature computation is finished.

BagMinHash 2 shown as Algorithm 7 implements the described optimization and is logically equivalent to BagMinHash 1. The first for-loop represents the first phase of the improved approach. As soon as the smallest relevant point is found for an input element d , which is the case once the first fully relevant subprocess has been found, the iterative splitting process is stopped. All subprocesses, that have been forked from P_d so far, are stored in a buffer before continuing with the next input element. The second step starts with the construction of a min-heap filled with all the subprocesses

that have been collected in the buffer while processing all input elements in the first phase. Then, these subprocesses are processed as in Algorithm 6 until all remaining processes have points larger than h_{\max} .

In order to keep space requirements for the buffer small, it can be realized as max-heap. In this way, since h_{\max} is decreasing during the first step, subprocesses with points greater than h_{\max} can be immediately evicted in an efficient way.

4 EXPERIMENTAL RESULTS

For our tests we have implemented both variants of the BagMinHash algorithm using C++¹. Since BagMinHash assumes an ideal PRNG, its implementation requires some special care. We have chosen to use the xxHash64 hashing algorithm [5], which we applied to the given seed in combination with values from a predefined integer sequence to get as many different 64-bit random numbers as needed. To keep the number of hash function evaluations small, we tried to use as few random bits as possible to generate random numbers for different distributions as required by BagMinHash. Bernoulli trials are realized using an algorithm that takes only two random bits on average [8]. Uniform discrete values can be efficiently generated as described in [23]. Moreover, we used the ziggurat algorithm to produce exponentially distributed random numbers without the need of costly logarithm evaluations [27].

In our experiments we primarily used the set of nonnegative single-precision floating point numbers as described in Section 2.1 for weight discretization. However, double-precision is used for all calculations and the representation of points of Poisson processes. Moreover, we assumed that all input elements are 64-bit integers. This corresponds to a universe D with dimensionality $|D| = 2^{64}$, which is large enough for most real-world applications. Any bag from a different universe can be mapped to a bag of 64-bit integers by hashing its elements.

4.1 Verification

For the verification of the BagMinHash algorithm we used synthetic data. The reason is that it is difficult to extract many pairs of input vectors from real data sets that have some predefined Jaccard index. Furthermore, a couple of recently published algorithms with theoretical flaws have been wrongly justified by tests on real-world data. As we will see these algorithms did not pass our tests, which is an indication that our tests are quite selective.

Each of our test cases is characterized by a bag of weight pairs $\bigcup_u \{(w_{Au}, w_{Bu})\}$. Each pair stands for some unique input element d which has weight $w_A(d) = w_{Au}$ in the first bag A and weight $w_B(d) = w_{Bu}$ in the second bag B , respectively. Since we are free to choose the elements that are associated with the weight pairs, we can simply draw 64-bit random numbers. In this way we can generate as many pairs of bags as needed for our evaluations. All of them will have the same Jaccard similarity. To avoid that elements are processed in some particular order, the elements of a bag are always shuffled before hashing.

Table 1 lists all test cases together with the corresponding Jaccard similarity, which we used to test our new BagMinHash algorithm

¹Source code that has been used to produce the results presented in this paper is made available on <https://github.com/oertl/bagminhash>.

Table 1: Analysis of the estimation error of various weighted minwise hashing algorithms for different test cases and different hash sizes. The empirical mean squared errors MSE and corresponding z-scores have been calculated from 10 000 random examples, respectively. MSE values with magnitude greater than 3 indicate exceptional behavior and are colored red.

test case	m	BagMinHash (float)			BagMinHash (binary)		ICWS [14]		0-bit [17]		CCWS [35]		PCWS [36]		I ² CWS [37]	
		$\mathbb{E}(\widehat{\text{MSE}})$	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score	$\widehat{\text{MSE}}$	z-score
{(1, 10)}	4	2.25E-2	2.19E-2	-1.49	N/A	N/A	2.23E-2	-0.43	8.10E-1	1.93E3	1.00E-2	-3.07E1	2.25E-2	-0.04	2.23E-2	-0.44
	16	5.63E-3	5.75E-3	1.44	N/A	N/A	5.55E-3	-0.91	8.10E-1	9.39E3	1.00E-2	5.11E1	5.67E-3	0.55	5.70E-3	0.82
	64	1.41E-3	1.42E-3	0.75	N/A	N/A	1.38E-3	-1.30	8.10E-1	3.99E4	1.00E-2	4.24E2	1.41E-3	0.12	1.42E-3	0.76
	256	3.52E-4	3.48E-4	-0.66	N/A	N/A	3.50E-4	-0.36	8.10E-1	1.62E5	1.00E-2	1.93E3	3.59E-4	1.45	3.42E-4	-1.94
	1024	8.79E-5	8.85E-5	0.50	N/A	N/A	8.91E-5	1.00	8.10E-1	6.51E5	1.00E-2	7.96E3	8.95E-5	1.28	8.56E-5	-1.82
	4096	2.20E-5	2.26E-5	1.91	N/A	N/A	2.18E-5	-0.44	8.10E-1	2.61E6	1.00E-2	3.21E4	2.18E-5	-0.65	2.19E-5	-0.09
{(9, 10)}	4	2.25E-2	2.27E-2	0.40	N/A	N/A	2.24E-2	-0.33	1.00E-2	-3.07E1	8.10E-1	1.93E3	2.25E-2	0.02	2.28E-2	0.77
	16	5.62E-3	5.77E-3	1.66	N/A	N/A	5.66E-3	0.46	1.00E-2	5.11E1	8.10E-1	9.39E3	5.66E-3	0.42	5.60E-3	-0.35
	64	1.41E-3	1.39E-3	-1.00	N/A	N/A	1.36E-3	-2.15	1.00E-2	4.24E2	8.10E-1	3.99E4	1.38E-3	-1.21	1.41E-3	0.19
	256	3.52E-4	3.55E-4	0.63	N/A	N/A	3.50E-4	-0.40	1.00E-2	1.93E3	8.10E-1	1.62E5	3.54E-4	0.55	3.52E-4	0.07
	1024	8.79E-5	8.63E-5	-1.30	N/A	N/A	8.70E-5	-0.76	1.00E-2	7.96E3	8.10E-1	6.51E5	9.02E-5	1.87	8.79E-5	-0.02
	4096	2.20E-5	2.20E-5	0.11	N/A	N/A	2.16E-5	-1.13	1.00E-2	3.21E4	8.10E-1	2.61E6	2.20E-5	0.24	2.18E-5	-0.66
{(3, 20), (30, 7)}	4	4.00E-2	4.01E-2	0.09	N/A	N/A	4.07E-2	1.19	8.04E-2	7.04E1	4.00E-2	0.00	3.45E-2	-9.50	3.32E-2	-1.18E1
	16	1.00E-2	1.00E-2	0.00	N/A	N/A	9.88E-3	-0.83	3.73E-2	1.92E2	4.00E-2	2.11E2	1.09E-2	6.48	2.00E-2	7.05E1
	64	2.50E-3	2.51E-3	0.35	N/A	N/A	2.52E-3	0.53	2.63E-2	6.72E2	4.00E-2	1.06E3	4.87E-3	6.69E1	1.69E-2	4.07E2
	256	6.25E-4	6.21E-4	-0.51	N/A	N/A	6.38E-4	1.47	2.35E-2	2.59E3	4.00E-2	4.45E3	3.47E-3	3.22E2	1.62E-2	1.76E3
	1024	1.56E-4	1.55E-4	-0.73	N/A	N/A	1.58E-4	0.86	2.28E-2	1.02E4	4.00E-2	1.80E4	3.08E-3	1.32E3	1.59E-2	7.13E3
	4096	3.91E-5	3.87E-5	-0.60	N/A	N/A	3.89E-5	-0.35	2.26E-2	4.08E4	4.00E-2	7.23E4	2.99E-3	5.34E3	1.59E-2	2.87E4
{(0, 2), (3, 4), (6, 3), (2, 4)}	4	6.25E-2	6.20E-2	-0.66	N/A	N/A	6.41E-2	2.07	6.82E-2	7.47	2.50E-1	2.45E2	6.74E-2	6.34	8.17E-2	2.51E1
	16	1.56E-2	1.54E-2	-0.83	N/A	N/A	1.54E-2	-1.08	2.19E-2	2.96E1	2.50E-1	1.10E3	2.09E-2	2.47E1	3.86E-2	1.07E2
	64	3.91E-3	3.80E-3	-1.90	N/A	N/A	3.88E-3	-0.43	1.10E-2	1.29E2	2.50E-1	4.49E3	8.76E-3	8.86E1	2.78E-2	4.37E2
	256	9.77E-4	9.84E-4	0.53	N/A	N/A	9.64E-4	-0.89	8.12E-3	5.19E2	2.50E-1	1.81E4	5.91E-3	3.58E2	2.53E-2	1.76E3
	1024	2.44E-4	2.44E-4	-0.15	N/A	N/A	2.42E-4	-0.49	7.51E-3	2.11E3	2.50E-1	7.24E4	5.25E-3	1.45E3	2.46E-2	7.05E3
	4096	6.10E-5	6.18E-5	0.91	N/A	N/A	6.06E-5	-0.55	7.33E-3	8.42E3	2.50E-1	2.90E5	5.09E-3	5.83E3	2.44E-2	2.82E4
{(4, 2) ¹⁵ , (1, 4) ¹⁰ , (12, 0) ⁵ }	4	4.69E-2	4.73E-2	0.60	N/A	N/A	4.66E-2	-0.45	4.69E-2	0.08	6.25E-2	2.46E1	4.30E-2	-6.10	4.40E-2	-4.52
	16	1.17E-2	1.17E-2	-0.22	N/A	N/A	1.17E-2	-0.18	1.17E-2	0.16	6.25E-2	3.10E2	1.18E-2	0.77	2.54E-2	8.36E1
	64	2.93E-3	2.99E-3	1.46	N/A	N/A	2.93E-3	0.03	3.03E-3	2.44	6.25E-2	1.44E3	4.55E-3	3.92E1	2.04E-2	4.24E2
	256	7.32E-4	7.24E-4	-0.82	N/A	N/A	7.32E-4	0.01	8.08E-4	7.33	6.25E-2	5.97E3	2.63E-3	1.84E2	1.92E-2	1.78E3
	1024	1.83E-4	1.84E-4	0.43	N/A	N/A	1.82E-4	-0.62	2.53E-4	2.69E1	6.25E-2	2.41E4	2.13E-3	7.52E2	1.89E-2	7.21E3
	4096	4.58E-5	4.53E-5	-0.68	N/A	N/A	4.52E-5	-0.90	1.15E-4	1.08E2	6.25E-2	9.65E4	1.99E-3	3.01E3	1.88E-2	2.89E4
$\bigcup_{u=0}^{1000} \{(1.001^u, 1.002^u)\}$ $J = 0.538308$	4	6.21E-2	6.14E-2	-1.02	N/A	N/A	6.18E-2	-0.37	6.14E-2	-0.94	2.84E-1	2.90E2	6.75E-2	6.99	1.02E-1	5.28E1
	16	1.55E-2	1.57E-2	0.98	N/A	N/A	1.56E-2	0.53	1.55E-2	-0.33	2.82E-1	1.25E3	1.98E-2	2.03E1	5.99E-2	2.09E2
	64	3.88E-3	3.78E-3	-1.90	N/A	N/A	3.88E-3	-0.02	3.83E-3	-0.97	2.81E-1	5.09E3	7.93E-3	7.42E1	4.97E-2	8.41E2
	256	9.71E-4	9.82E-4	0.79	N/A	N/A	9.68E-4	-0.22	9.60E-4	-0.82	2.81E-1	2.04E4	4.96E-3	2.91E2	4.72E-2	3.37E3
	1024	2.43E-4	2.48E-4	1.40	N/A	N/A	2.36E-4	-1.81	2.36E-4	-2.07	2.81E-1	8.19E4	4.31E-3	1.19E3	4.67E-2	1.35E4
	4096	6.07E-5	6.09E-5	0.30	N/A	N/A	6.10E-5	0.43	6.10E-5	0.36	2.81E-1	3.28E5	4.12E-3	4.74E3	4.66E-2	5.42E4
{(0, 1), (1, 0), (1, 1)}	4	5.56E-2	5.53E-2	-0.41	5.59E-2	0.46	5.50E-2	-0.74	5.57E-2	0.24	5.63E-2	1.02	5.67E-2	1.57	5.60E-2	0.60
	16	1.39E-2	1.39E-2	0.06	1.39E-2	-0.11	1.41E-2	0.95	1.37E-2	-1.05	1.41E-2	1.29	1.39E-2	-0.20	1.38E-2	-0.44
	64	3.47E-3	3.44E-3	-0.58	3.48E-3	0.06	3.50E-3	0.66	3.42E-3	-1.05	3.43E-3	-0.81	3.45E-3	-0.52	3.61E-3	2.85
	256	8.68E-4	8.70E-4	0.17	8.55E-4	-1.09	8.84E-4	1.33	8.60E-4	-0.64	8.76E-4	0.65	8.77E-4	0.73	8.66E-4	-0.13
	1024	2.17E-4	2.13E-4	-1.24	2.13E-4	-1.45	2.25E-4	2.45	2.20E-4	0.86	2.18E-4	0.20	2.21E-4	1.24	2.15E-4	-0.54
	4096	5.43E-5	5.50E-5	0.98	5.36E-5	-0.80	5.44E-5	0.21	5.43E-5	0.08	5.52E-5	1.27	5.39E-5	-0.44	5.43E-5	0.02
{(0, 1) ³⁰ , (1, 0) ¹⁰ , (1, 1) ¹⁶⁰ }	4	4.00E-2	3.98E-2	-0.39	3.92E-2	-1.47	4.00E-2	-0.07	3.97E-2	-0.44	4.02E-2	0.37	4.02E-2	0.37	4.02E-2	0.29
	16	1.00E-2	9.95E-3	-0.33	1.00E-2	0.33	1.00E-2	0.28	9.81E-3	-1.33	9.81E-3	-1.34	9.95E-3	-0.34	1.01E-2	0.93
	64	2.50E-3	2.55E-3	1.28	2.42E-3	-2.32	2.56E-3	1.81	2.45E-3	-1.48	2.50E-3	0.12	2.48E-3	-0.63	2.52E-3	0.58
	256	6.25E-4	6.19E-4	-0.66	6.16E-4	-1.02	6.23E-4	-0.25	6.17E-4	-0.89	6.31E-4	0.64	6.07E-4	-2.00	6.31E-4	0.68
	1024	1.56E-4	1.54E-4	-1.10	1.55E-4	-0.40	1.57E-4	0.18	1.56E-4	0.05	1.55E-4	-0.54	1.52E-4	-2.07	1.57E-4	0.24
	4096	3.91E-5	3.94E-5	0.66	3.85E-5	-0.98	3.89E-5	-0.25	3.95E-5	0.86	3.93E-5	0.43	3.92E-5	0.27	3.82E-5	-1.53
{(0, 1) ³⁰⁰ , (1, 0) ⁵⁰⁰ , (1, 1) ¹²⁰⁰ }	4	6.00E-2	5.97E-2	-0.38	5.96E-2	-0.56	5.92E-2	-1.05	5.86E-2	-1.90	6.07E-2	0.88	6.13E-2	1.78	6.02E-2	0.21
	16	1.50E-2	1.52E-2	0.97	1.51E-2	0.29	1.49E-2	-0.43	1.49E-2	-0.45	1.47E-2	-1.36	1.51E-2	0.71	1.50E-2	-0.05
	64	3.75E-3	3.78E-3	0.52	3.83E-3	1.49	3.70E-3	-0.90	3.73E-3	-0.38	3.75E-3	-0.01	3.81E-3	1.18	3.73E-3	-0.33
	256	9.37E-4	9.30E-4	-0.54	9.53E-4	1.17	9.52E-4	1.08	9.36E-4	-0.12	9.36E-4	-0.10	9.36E-4	-0.14	9.32E-4	-0.41
	1024	2.34E-4	2.32E-4	-0.70	2.37E-4	0.79	2.34E-4	0.00	2.32E-4	-0.58	2.40E-4	1.55	2.33E-4	-0.56	2.33E-4	-0.35
	4096	5.86E-5	5.76E-5	-1.24	5.93E-5	0.81	5.89E-5	0.41	5.78E-5	-0.99	5.89E-5	0.32	5.76E-5	-1.25	5.95E-5	1.12

and to compare it to other weighted minwise hashing algorithms. We also varied the signature size $m \in \{4, 16, 64, 256, 1024, 4096\}$. For each test case $c = 10\,000$ pairs of bags have been generated. For each pair an estimate \hat{J}_J of the Jaccard similarity was obtained using the corresponding hash signatures and (3). We calculated the empirical mean squared error (MSE) as $\widehat{\text{MSE}} = \frac{1}{c} \sum_{j=1}^c (\hat{J}_j - J)^2$. Since hash collisions can be ignored for 64-bit signature values, the number of matching signature positions follows a binomial distribution with success probability J . The expectation and the variance of $\widehat{\text{MSE}}$ can be derived as $\mathbb{E}(\widehat{\text{MSE}}) = \frac{J(1-J)}{m}$ and $\text{Var}(\widehat{\text{MSE}}) =$

$\frac{J^2(1-J)^2}{m^2c} \left(2 - \frac{6}{m}\right) + \frac{J(1-J)}{m^3c}$, respectively. These formulas can be used to normalize $\widehat{\text{MSE}}$, which yields the corresponding z-score

$$\text{z-score} = (\widehat{\text{MSE}} - \mathbb{E}(\widehat{\text{MSE}})) / \sqrt{\text{Var}(\widehat{\text{MSE}})}.$$

Table 1 shows the empirical MSEs and the corresponding z-scores of our simulations for various algorithms. In case of BagMinHash we got identical results for both variants, which is expected as both variants are logically equivalent. The results do not show any evidence for unexpected behavior, because all z-scores have magnitudes smaller than 3. Values with larger magnitudes would

have been shown in red color. We also applied BagMinHash with binary weight discretization ($V = \{0, 1\}$, $L = 1$) to test cases with only weights equal to 0 or 1. Again, no anomalies have been observed and the theoretically predicted MSEs have been confirmed. As expected, the ICWS also passes all our test cases.

Much more interesting are a couple of algorithms that have been recently presented as supposed improvements over ICWS. They share the same time complexity $O(mn)$ of the ICWS algorithm, but introduce some modifications to save a couple of operations. As example, 0-bit hashing [17] was proposed to simplify the generation of integer signature values and to avoid a post-processing step like Algorithm 2. It was reported that the resulting signatures approximate (2) reasonably well in practice. Unfortunately, a rigorous analysis of the validity range of this approximation was not made. Our results show that the algorithm has problems for $n < m$ except for test cases with binary weights.

There is also a series of recently published modifications of ICWS algorithm that did not pass our tests either: Canonical consistent weighted sampling (CCWS) [35], practical consistent weighted sampling (PCWS) [36], and improved ICWS (I²CWS) [37]. When looking for the reason, we discovered theoretical flaws described in the following using the notation of the corresponding papers. In [35] the samples of y_k are drawn from $[S_k - r_k, S_k]$ instead from $[0, S_k]$ which, however, is necessary for consistent sampling. In [36] the expression $\text{pdf}(y, a) = \frac{1}{S}(yu_1^{-1})e^{-(yu_1^{-1})a}$ is turned into $\text{pdf}(y, a) = \frac{1}{S}(Se^{-Sa})$ by replacing yu_1^{-1} by its expectation $\mathbb{E}(yu_1^{-1}) = S$ which is an invalid transformation. And finally, in [37] the proof of consistency assumes that $t_{k,2}^S = t_{k,2}^T$ follows from $y_{k_*} \leq T_{k_*} \leq S_{k_*}$, which is not true either. Incorrect argumentation even led to the claim that the ICWS algorithm is wrong, which we could neither confirm theoretically nor empirically.

4.2 Performance

We still need to check if BagMinHash is really fast in practice. Therefore, we measured the performance for various signature sizes $m \in \{256, 1024, 4096, 16384\}$ and bag sizes n ranging from 1 to 10^7 . For each case we randomly generated 100 bags of 64-bit random integer numbers with weights drawn from an exponential distribution with rate parameter 1. The average calculation times on an Intel Core i5-2500K CPU for both BagMinHash variants and ICWS are shown in Fig. 1. The calculation time of ICWS is extrapolated for $n > 10000$ by utilizing its perfect proportional scaling with n .

The BagMinHash algorithms are significantly slower than ICWS for small n . The reason can be seen when considering the case $n = 1$ for which, on average, the BagMinHash algorithms require the generation of $mH_m = O(m \log m)$ points until at least one point is found for every signature component as discussed in Section 3.1. Furthermore, the iterative splitting of the Poisson process to generate one point contributes at least an additional $\log L$ factor. In contrast, the calculation time for ICWS is only proportional to m in this case.

However, the BagMinHash algorithms are orders of magnitude faster for $n \gg m$, where their calculation times are approaching a linear function of n . The gap to the calculation time of the ICWS algorithm increases with m , which indicates that the dominating

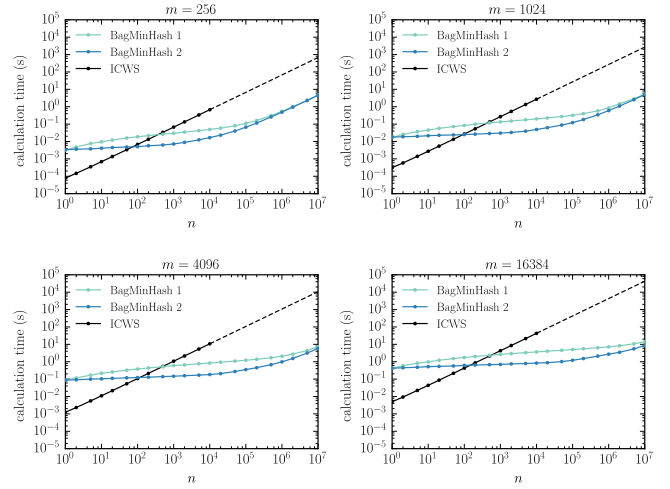


Figure 1: Average calculation time for random bags of size n with weights drawn from an exponential distribution with rate 1.

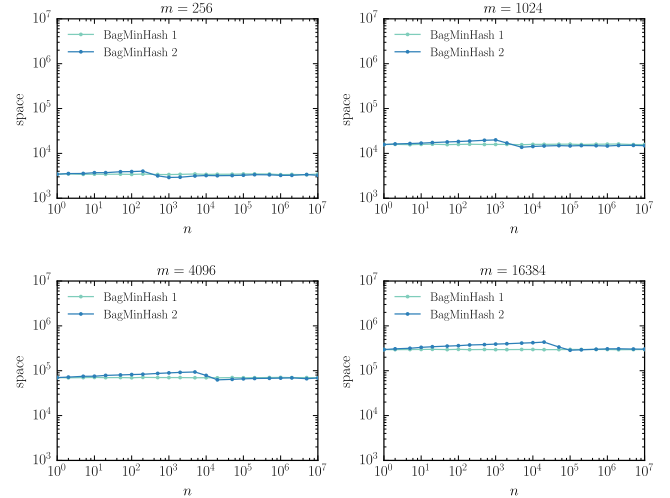


Figure 2: The average space requirements in terms of stored Poisson process objects as defined in Algorithm 4.

term in the time complexity of BagMinHash does not depend on m . The optimization described in Section 3.3 and implemented by BagMinHash 2 leads to a significant improvement of the calculation time for intermediate input sizes n . The break-even point of BagMinHash 2 compared to ICWS is around $n = 100$.

As shown in Fig. 2, we also investigated the space requirements in terms of Poisson process objects, that need to be simultaneously kept in memory. The maximum space required by BagMinHash 1 is dominated by the heap size needed for processing the very first element. Since the first element is faced with the largest h_{\max} values, more points need to be generated and more Poisson process splittings are necessary than for later elements. This explains why

the space requirements are essentially independent of n . Our BagMinHash 2 implementation uses a max-heap as buffer, which allows efficient discarding of process objects while h_{\max} is decreasing. This is the reason, why the space requirements are not very different to those of BagMinHash 1. Since the observed values are small enough and also the dependence on m seems to be at most quasilinear, the space requirements are not an issue for today's computers. We did not investigate the space requirements of ICWS. It only needs some vectors of size m , which is negligible in practice.

The presented performance results refer to the single-precision weight discretization discussed in Section 2.1. Since the discretization V can be freely chosen, it can be optimized to achieve a certain precision, if all non-zero weights are from some known value range. For example, if an estimation error of ε is acceptable for the Jaccard index, one could choose $v_l = v_1(1 + \varepsilon)^{l-1}$ with $1 \leq l \leq L$ together with appropriate $v_1 > 0$ and L to cover the desired value range. In this way (7) will be satisfied, while minimizing L and thus also computational costs.

5 CONCLUSION

We have presented BagMinHash, a new weighted minwise hashing algorithm for the calculation of hash signatures that can be used to estimate the generalized Jaccard similarity of weighted sets. Without being less accurate, BagMinHash is orders of magnitude faster for large input than the current state of the art. Therefore, we expect that our new approach will improve the performance of existing applications and also enable new use cases for which computational costs had been a limiting factor so far.

REFERENCES

- [1] O. Alonso, D. Fetterly, and M. Manasse. 2013. Duplicate News Story Detection Revisited. In *Proc. 9th Asia Information Retrieval Societies Conf. (AIRS)*. 203–214.
- [2] A. Z. Broder. 1997. On the Resemblance and Containment of Documents. In *Proc. Compression and Complexity of Sequences*. 21–29.
- [3] M. S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proc. 34th Symp. on Theory of Computing (STOC)*. 380–388.
- [4] O. Chum, J. Philbin, and A. Zisserman. 2008. Near Duplicate Image Detection: Min-Hash and TF-IDF Weighting. In *Proc. British Machine Vision Conf. (BMVC)*. 812–815.
- [5] Y. Collet. 2016. xxHash – Extremely Fast Hash Algorithm. <https://github.com/Cyan4973/xxHash>.
- [6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. 2009. *Introduction to Algorithms*. MIT Press.
- [7] S. Dahlgaard, M. B. T. Knudsen, and M. Thorup. 2017. Fast Similarity Sketching. In *Proc. 58th Symp. on Foundations of Computer Science (FOCS)*. 663–671.
- [8] L. Devroye. 1986. *Non-Uniform Random Variate Generation*. Springer, New York.
- [9] J. Drew, M. Hahsler, and T. Moore. 2017. Polymorphic Malware Detection Using Sequence Classification Methods and Ensembles. *EURASIP J. on Information Security* (2017), 2:1–2:12.
- [10] O. Ertl. 2017. SuperMinHash – A New Minwise Hashing Algorithm for Jaccard Similarity Estimation. (2017). [arXiv:1706.05698](https://arxiv.org/abs/1706.05698)
- [11] S. Gollapudi and R. Panigrahy. 2006. Exploiting Asymmetry in Hierarchical Topic Extraction. In *Proc. 15th Int. Conf. on Information and Knowledge Management (CIKM)*. 475–482.
- [12] B. Haeupler, M. S. Manasse, and K. Talwar. 2014. Consistent Weighted Sampling Made Fast, Small, and Easy. [arXiv:1410.4266](https://arxiv.org/abs/1410.4266)
- [13] T. Haveliwala, A. Gionis, and P. Indyk. 2000. Scalable Techniques for Clustering the Web. In *Proc. 3rd Int. Workshop on the Web and Databases (WebDB)*. 129–134.
- [14] S. Ioffe. 2010. Improved Consistent Sampling, Weighted Minhash and L1 Sketching. In *Proc. 10th Int. Conf. on Data Mining (ICDM)*. 246–255.
- [15] J. Kleinberg and E. Tardos. 2002. Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields. *J. of the ACM* 49, 5 (2002), 616–639.
- [16] J. Leskovec, A. Rajaraman, and J. D. Ullman. 2014. *Mining of Massive Datasets*. Cambridge University Press.
- [17] P. Li. 2015. 0-Bit Consistent Weighted Sampling. In *Proc. 21th Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 665–674.
- [18] P. Li. 2017. Linearized GMM Kernels and Normalized Random Fourier Features. In *Proc. 23rd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 315–324.
- [19] P. Li and A. C. König. 2010. b-Bit Minwise Hashing. In *Proc. 19th Int. Conf. on World Wide Web (WWW)*. 671–680.
- [20] P. Li and A. C. König. 2011. Theory and Applications of b-bit Minwise Hashing. *Communications of the ACM* 54, 8 (2011), 101–109.
- [21] P. Li, A. Owen, and C. Zhang. 2012. One Permutation Hashing. In *Proc. 26th Conf. on Advances in Neural Information Processing Systems (NIPS)*. 3113–3121.
- [22] P. Li and C.-H. Zhang. 2017. Theory of the GMM Kernel. In *Proc. 26th Int. Conf. on World Wide Web (WWW)*. 1053–1062.
- [23] J. Lumbroso. 2013. Optimal Discrete Uniform Generation from Coin Flips, and Applications. (2013). [arXiv:1304.1916](https://arxiv.org/abs/1304.1916)
- [24] C. Luo and A. Shrivastava. 2016. SSH (Sketch, Shingle, & Hash) for Indexing Massive-Scale Time Series. In *Proc. of Machine Learning Research*, Vol. 55. 38–58.
- [25] M. Manasse, F. McSherry, and K. Talwar. 2010. *Consistent Weighted Sampling*. Technical Report. <https://www.microsoft.com/en-us/research/publication/consistent-weighted-sampling/>
- [26] V. Markovtsev and E. Kant. 2017. Topic Modeling of Public Repositories at Scale Using Names in Source code. (2017). [arXiv:1704.00135](https://arxiv.org/abs/1704.00135)
- [27] G. Marsaglia and W. W. Tsang. 2000. The Ziggurat Method for Generating Random Variables. *J. of Statistical Software* 5, 8 (2000), 1–7.
- [28] M. Mitzenmacher and E. Upfal. 2005. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press.
- [29] E. Raff and C. Nicholas. 2017. Malware Classification and Class Imbalance via Stochastic Hashed LZJD. In *Proc. 10th ACM Workshop on Artificial Intelligence and Security (AISec)*. 111–120.
- [30] S. Sathe and C. C. Aggarwal. 2017. Similarity Forests. In *Proc. 23rd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 395–403.
- [31] A. Shrivastava. 2016. Simple and Efficient Weighted Minwise Hashing. In *Advances in Neural Information Processing Systems 29 (NIPS)*. 1498–1506.
- [32] A. Shrivastava. 2017. Optimal Densification for Fast and Accurate Minwise Hashing. In *Proc. 34th Int. Conf. on Machine Learning (ICML)*. 3154–3163.
- [33] A. Shrivastava and P. Li. 2015. Asymmetric Minwise Hashing for Indexing Binary Inner Products and Set Containment. In *Proc. 24th Int. Conf. on World Wide Web (WWW)*. 981–991.
- [34] R. Spring and A. Shrivastava. 2017. Scalable and Sustainable Deep Learning via Randomized Hashing. In *Proc. 23rd Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 445–454.
- [35] W. Wu, B. Li, L. Chen, and C. Zhang. 2016. Canonical Consistent Weighted Sampling for Real-Value Weighted Min-Hash. In *Proc. 16th Int. Conf. on Data Mining (ICDM)*. 1287–1292.
- [36] W. Wu, B. Li, L. Chen, and C. Zhang. 2017. Consistent Weighted Sampling Made More Practical. In *Proc. 26th Int. Conf. on World Wide Web (WWW)*. 1035–1043.
- [37] W. Wu, B. Li, L. Chen, C. Zhang, and P. S. Yu. 2017. Improved Consistent Weighted Sampling Revisited. (2017). [arXiv:1706.01172](https://arxiv.org/abs/1706.01172)