

# Compressed bitmap indexes: beyond unions and intersections

Owen Kaser<sup>1\*</sup> and Daniel Lemire<sup>2</sup>

<sup>1</sup>*Dept. of CSAS, University of New Brunswick, Saint John, NB, Canada*

<sup>2</sup>*LICEF, TELUQ, Université du Québec, Montreal, QC, Canada*

## SUMMARY

Compressed bitmap indexes are used to speed up simple aggregate queries in databases. Indeed, set operations like intersections, unions and complements can be represented as logical operations (AND, OR, NOT) that are ideally suited for bitmaps. However, it is less obvious how to apply bitmaps to more advanced queries. For example, we might seek products in a store that meet some, but maybe not all, criteria. Such threshold queries generalize intersections and unions; they are often used in information-retrieval and data-mining applications. We introduce new algorithms that are sometimes three orders of magnitude faster than a naïve approach. Our work shows that bitmap indexes are more broadly applicable than is commonly believed.

**KEY WORDS:** T-overlap queries; compressed bitmaps; threshold functions; symmetric functions; opt-threshold queries

## 1. INTRODUCTION

There are many applications for bitmap indexes, from conventional databases (e.g., Oracle [1]) all the way to information retrieval [2] and column stores [3]. They are used in data-warehouse platforms such as Apache Hive, LucidDB [4], Druid [?] and Sybase IQ [5].

We are primarily motivated by the application of bitmap indexes to common databases (i.e., row stores). In this case, it has long been established that bitmap indexes can speed up several queries, e.g., joins [6], as well as intersections and unions (e.g., `SELECT * WHERE A=1 AND B=2`).

Databases are commonly used for data mining and machine learning. An algorithm could seek to identify all movies that are “similar” to a target movie, or all customers that “almost” fit a given profile. Such queries need neither an intersection nor a union, but something in-between: a threshold function where only some of the criteria need to be satisfied. We aim to show that such queries (specifically *Many-Criteria* queries and *Similarity* queries, see § 4) can be answered efficiently using bitmap indexes. Because the result of the query is itself a bitmap, we can then further process it using the standard operations permitted on bitmaps (OR, AND, XOR, NOT) to answer more complicated queries efficiently.

Of course, the set of basic operations supported by bitmap indexes may be sufficient to synthesize any required function. However, the efficiency of such approaches is unknown. To our knowledge, the efficient computation of threshold functions over bitmaps has never been investigated in depth: the exception is Rinfret et al. [7] where two algorithms are compared on a related problem (top- $K$  queries).

---

\*Correspondence to: Owen Kaser, Dept. of CSAS, University of New Brunswick, 100 Tucker Park Rd, Saint John, NB E2L 4L5 Canada. email: owen@unbsj.ca

Table I. Algorithms considered in this paper.

Algorithm	Source	Section	Main idea
SCANCOUNT	[9]	§ 5.1	Allocate array of counters, scan values while incrementing counters and, finally, scan array of counters for matching counts.
MGOPT	[10, 11]	§ 5.2	Set aside the largest $T - 1$ inputs, merge the remaining $N - T + 1$ inputs using a heap, then look up matching values in the largest inputs.
DSK	[9]	§ 5.2	Similar to MGOPT, but during the merger of the small inputs, some values are skipped; requires a tuning parameter.
BSTM	modified from [7, 12]	§ 5.3.1	Transforms the query into a Boolean circuit to be evaluated on the bitmaps.
W2CTI	novel	§ 6.1	Merge inputs two-by-two starting with lowest-cardinality inputs while maintaining counters, prune results as early as possible.
LOOPED	novel	§ 6.2	Allocates $T$ temporary bitmaps corresponding to the count values $1, 2, \dots, T$ , the first bitmap updates the first temporary bitmap, the second bitmap updates the first two temporary bitmaps, and so on.
RBMRG	novel, inspired by [13]	§ 6.3	Using a heap, merge RLE-compressed words.

This paper considers several algorithms for threshold functions over compressed bitmap indexes (see Table I). Some of these algorithms are novel (LOOPED, W2CTI), whereas other algorithms are adaptations of known algorithms that had operated over sorted integer lists (SCANCOUNT, MGOPT, DSK) or over bitmaps (BSTM). (A companion report [8] considers additional algorithms that do not perform as well, and also considers the use of uncompressed bitmaps.) The theoretical analyses of these alternatives, summarized in Table III<sup>†</sup>, suggests that there would be no single best algorithm for all cases, as the algorithms’ running times depend on different factors. Experiments described in § 7 confirm this conclusion. Thus one of our contributions is a set of rules for automatically choosing algorithms.

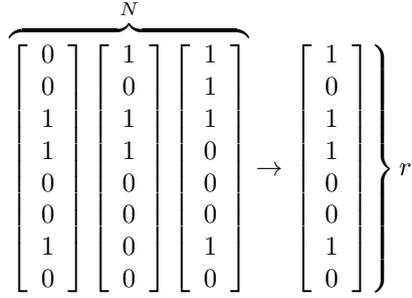
Our work is organized as follows. In § 2, we formalize the problem. In § 3, we present some background material and related work. In § 4, we present the queries over database tables that we use for benchmarking. In § 4.1, we begin our experimental report by showing that using a bitmap index, albeit naïvely, is better than a full table scan: the indexed version is anywhere from 1.1 to 6 times faster. In § 5 and § 6, we present our various algorithms. Finally, in § 7 we assess them experimentally and show that one can do significantly better than a naïve approach: up to  $1100\times$  better, in one case. Over a large workload that we constructed, we could more than triple performance.

## 2. FORMULATION

We take  $N$  sorted sets over a universe having  $r$  distinct values. For our purposes, we represent sets as bitmaps using  $r$  bits. For example, if  $N = 2$  and  $r = 8$ , we might have the sets  $\{1, 4, 5\}$  and  $\{4, 5, 7\}$  of integers in  $[0, 8)$  represented using the bitmaps 00110010 and 10110000, where the least-significant bit represents the smallest value in the universe (see § 3.1). The notation we use throughout is described in Table II.

The sum of the cardinalities of the  $N$  sets is  $B$ : in our example  $B = 3 + 3 = 6$ . The cardinality of a set is also given by the number of 1s in the corresponding bitmap. (By extension, the cardinality of a bitmap is the number of 1s it contains.) Therefore, the value  $B$  is also the total number of 1s in all bitmaps. We apply a threshold  $T$  ( $1 \leq T \leq N$ ), seeking those elements that occur in at least  $T$  sets

<sup>†</sup>The notation used throughout can be found in Table II.

Figure 1. Solution to a threshold query with  $T = 2$  over  $N = 3$  bitmaps.

Name	City	Montreal	Toronto	Paris
John	Montreal	1	0	0
Peter	Montreal	1	0	0
Jack	Toronto	0	1	0
Jack	Toronto	0	1	0
Jill	Toronto	0	1	0
Lucy	Paris	0	0	1
Mary	Toronto	0	1	0

Figure 2. Bitmap index of the attribute City.

(see Fig. 1). Because the cases  $T = 1$  and  $T = N$  correspond to intersections and unions, which are well understood, we assume that  $2 \leq T \leq N - 1$ . These queries are often called  $T$ -overlap [14, 15],  $T$ -occurrence [9, 16] or  $T$ -threshold [11, 17] queries.

We can map a  $T$ -overlap query to a query over bitmaps using a Boolean threshold function: given  $N$  bits, the  $T$ -threshold function  $\vartheta(T, \{b_1, \dots, b_N\})$  returns true if at least  $T$  bits are true; it returns false otherwise. For example, given  $T = N$ , such a function would just be a logical conjunction (AND) and given  $T = 1$ , it would be a logical disjunction (OR). That is, we have  $\vartheta(N, \{b_1, \dots, b_N\}) = b_1 \wedge \dots \wedge b_N$  and  $\vartheta(1, \{b_1, \dots, b_N\}) = b_1 \vee \dots \vee b_N$ .

A (unary) bitmap index over a table has as many bitmaps as there are distinct attribute values (see Fig. 2). Each attribute value (say value  $v$  of attribute  $a$ ) has a bitmap that encodes the set of row IDs that satisfy the predicate  $a = v$ . A  $T$ -overlap query seeks row IDs that satisfy at least  $T$  of  $N$  predicates. Since each predicate is encoded as a bitmap, we need to compute a bitwise threshold function over the  $N$  chosen bitmaps.

Threshold functions are a subset of the *symmetric Boolean functions* (see § 3.2). They include the majority function: given  $N$  bits, the majority function returns true when  $1 + \lfloor N/2 \rfloor$  or more bits are true, and it returns false otherwise. We can compute the majority function as any other threshold function. Other potentially useful generalizations include setting up a maximum (no more than  $T$  bits are set) or setting a range (the number of set bits is in  $[T_1, T_2]$ ). They can be rewritten in terms of threshold functions: e.g., we can determine whether at most  $T$  bits are set by evaluating  $\vartheta(N - T, \{\neg b_1, \dots, \neg b_N\})$ . We do not consider such generalizations further.

We denote the processor's native word length as  $W$  (typically<sup>‡</sup>  $W = 64$ ). An uncompressed bitmap will have  $\lceil r/W \rceil$  words; given  $N$  bitmaps, there are  $N \lceil r/W \rceil$  words. To simplify, we assume  $\log N < W < r$  as well<sup>§</sup> as  $\log r \leq W$ , which would typically be the case in the applications we envision. Also, we assume that  $B \geq N$ , which would be true if there is no bitmap containing only 0s: such *empty* bitmaps could be virtually deleted without harm.

<sup>‡</sup>Common 64-bit PCs have SIMD instructions that work over 128-bit (SSE and AVX), 256-bit (AVX2) or 512-bit (AVX-512) vectors. These instructions might be used automatically by compilers and interpreters. Other general purpose processors in embedded or mobile devices sometimes have a 32-bit word size.

<sup>§</sup>In this paper,  $\log n$  means  $\log_2 n$ .

Table II. Notation used in analyses.

Symbol	Meaning
$A_i$ or $B_i$	$i^{\text{th}}$ bitmap
$ B_i $	number of 1s in $i^{\text{th}}$ bitmap
$B_i[j]$	value of the $j^{\text{th}}$ bit in the $i^{\text{th}}$ bitmap
$B$	$\sum_i  B_i $
$B'$	number of 1s <b>not</b> in the $T - 1$ largest bitmaps
EWAHSIZE	storage size in bytes of a collection of compressed bitmaps
$N$	Number of bitmaps in the query
$r$	length of bitmaps (largest index covered)
RUNCOUNT	number of runs of 0s and 1s in a collection of bitmaps
$T$	Minimum threshold
$\vartheta(T, \{b_1, \dots, b_N\})$	threshold function over bits $b_i$
$W$	machine word size
$\oplus$ , XOR	exclusive or
$\wedge$ , AND	logical and
$\vee$ , OR	logical or
$\neg$ , NOT	logical negation

Table III. Time and memory complexity of threshold algorithms over RLE-compressed bitmap indexes.

Algorithm	Time complexity	Memory	Comment
SCANCOUNT	$O(r + B)$	$O(r)$	Efficient access pattern
MGOPT	$O(B'(\log(N - T) + T) + B - B')$	$O(N)$	Pruning can reduce $B$
DSK	$O(B'(\log(N - T) + T) + B - B')$	$O(N)$	Pruning can reduce $B'$ & $B$
BSTM	$O(Nr/W \times \log N)$	$O(\log N \times r/W)$	Note <sup>1</sup>
w2CTI	$O(B(N - T))$	$O(B)$	
LOOPEd	$O(NTr/W)$	$O(Tr/W)$	Note <sup>2</sup>
RBMrg	$O(\text{RUNCOUNT} \times \log N)$	$O(N)$	

<sup>1</sup>  $O(N \log N)$  basic bitmap operations are used, producing temporary results. There are  $O(\log N)$  temporaries live at any time. An  $O(Nr/W \times \log N)$  time bound ignores any benefits of compression for storage or processing.

<sup>2</sup> Fewer than  $2NT$  basic bitmap operations are used, and  $T$  temporary bitmaps are used. The bounds shown ignore compression's benefits.

Our focus is on algorithms that run in main memory; we assume that the  $N$  bitmaps involved in the threshold query have already been read into main memory. Our memory bounds in Table III are based on the *additional* working memory required, not including the input and output.

**A lower bound (and beating it):** Towards a lower bound for the problem, note that if the output indicates that  $X$  entries meet the threshold, at least  $TX$  1s have been observed in the input. If each such observation triggers  $\Omega(1)$  work (as it does with SCANCOUNT (§ 5.1), when a counter is incremented), this implies an  $\Omega(TX)$  lower bound. Barbay and Kenyon [11] have established a data-dependent lower bound for the problem, assuming the data is presented in sorted arrays and using a model where comparisons are the only allowed operations on array elements. However, both bounds leave open the possibility of using parallelism. One such approach, parallelization of SCANCOUNT on GPUs, is described by Li et al. [14]. We can use bit-level parallelism (readily available in bitmap inputs) to process several events per machine operation. (See § 5.3 and § 6.2.) The bounds also leave open the possibility of using Run Length Encoding (RLE), whereby many consecutive events can be succinctly represented and processed together. Our compressed bitmap inputs are suitable for such an approach: see § 6.3.

### 3. BACKGROUND AND RELATED WORK

We review some key concepts on compressed bitmaps, Boolean circuits, and Boolean functions—especially symmetric and threshold functions.

#### 3.1. Bitmaps

We find bitmap indexes in several database systems, going as far back as the MODEL 204 database engine, commercialized in 1972 [18]. Most commonly, a bitmap index associates a bitmap (also called bitset or bit vector) with every attribute value  $v$  of every attribute  $a$ ; the bitmap represents the predicate  $a = v$ . In the example of Fig. 2, we see that we could identify all rows where the value of the attribute is either Montreal or Toronto by computing the bitwise OR between two bitmaps. Such bitwise operations can be computed quickly by most processors. In an experimental evaluation using the Oracle database system, Sharma found that a bitmap index is preferable to a B-tree when the data is infrequently updated [19].

We consider compressed and uncompressed bitmaps. The *density* of a bitmap is the fraction of its bits that are 1s. A bitmap with low density is *sparse*, and such bitmaps arise in many applications. A bitmap with density closer to 1 (perhaps 5% or more) is *dense*.

**Uncompressed Bitmaps** An uncompressed bitmap represents a sorted set  $S$  over  $\{0, 1, \dots, r-1\}$  using  $\lfloor (r+W-1)/W \rfloor$  consecutive words. The  $W$  bits in the first word record which values in  $[0, W-1]$  are present in  $S$ . The bits in the second word record the values in  $[W, 2W-1]$  that are in  $S$ , and so forth. Within a word, the least-significant bit represents the smallest value. For example, the set  $\{1, 2, 7, 9\}$  is represented as 10000110 00000010 with  $W = 8$ . The first word (10000110) represents the first 3 integers ( $\{1, 2, 7\}$ ) whereas the second word (00000010) is used to store the value 9. The density is  $\frac{4}{16}$ . The exact mapping between integers and the bits within a word is unimportant, as long as it is always consistent (e.g.,  $\{1, 2, 7\}$  could be written as 10000110 or 01100001). The number of 1s is always equal to the cardinality of the set.

Uncompressed bitmaps have the advantages of a fixed size (updates do not change the size) and an efficient membership test. However, if  $r$  is large, the bitmap occupies many words and uses much memory—even when representing a small set ( $B \ll r$ ).

**Compressed Bitmaps** In a bitmap, there are runs of consecutive 0s and runs of consecutive 1s. The number of such runs is called the `RUNCOUNT` of a bitmap, or of a collection of bitmaps [20]. For example, in the bitmap index illustrated by Fig. 2, there are  $2 + 4 + 3 = 9$  runs. In the unary bitmap index of an attribute containing  $N$  distinct attribute values, given that there are  $r$  rows, the number of runs must be between  $3N - 2$  and  $2r + N - 2$ . Correspondingly, for  $r \gg N$ , the average length of the runs is between  $\approx r/3$  and  $\approx N/2$ . In many situations where bitmaps are generated, we expect to find many long runs (e.g., with length greater than  $W$ ).

Though there are alternatives [21], the most popular compression techniques are based on the (word-aligned) RLE compression model inherited from Oracle (BBC [1]): WAH [22], Concise [23], PLWAH [24], EWAH [13], COMPAX [25], VAL-WAH [26], among others. The  $r$  bits of the bitmap are partitioned into sequences of  $W'$  consecutive bits, where  $W' \approx W$  depends on the technique used; for EWAH,  $W' = W$ ; for WAH,  $W' = W - 1$ . When such a sequence contains only 1s or only 0s, it is a *fill* word, otherwise it is a *dirty* word. For example, using  $W' = 8$ , the uncompressed bitmap 0000000001010000 contains two words, a fill word (00000000) and a dirty word (01010000). Techniques such as BBC, WAH or EWAH typically use special marker words to compress long sequences of identical fill words. When accessing these formats, it may be necessary to read every compressed word to determine whether it indicates a sequence of fill words, or a dirty word. The EWAH format [13] supports a limited form of skipping because it uses marker words not only to mark the length of the sequences of fill words, but it also uses these markers to indicate the lengths of the sequences of consecutive dirty words. Because of this feature, one can skip sequences of dirty words when using EWAH.

Though there are many good compressed formats to choose from, we have picked EWAH. In a benchmark between various formats where the authors used our implementation (the JavaEWAH library [27]), Guzun et al. [26] found that “Although EWAH does not compress well, (...) it offers the best query time for all distributions.” Moreover, EWAH is used in a major data database system (Apache Hive). We refer the reader to previous work for the exact format specification [13].

Compressed bitmaps are often appropriate for storing sets that cannot be efficiently handled by uncompressed bitmaps. For instance, consider the bitmap consisting of a million 0s followed by a million 1s. This data has two runs ( $\text{RUNCOUNT} = 2$ ) but a million 1s. It can be stored using EWAH in only a few words.

However, some RLE compressed bitmaps are not efficient for storing extremely sparse data that does not have dense clusters. For instance, consider EWAH: sparse data with very long runs of 0s between elements will result in a marker word and a dirty word for each 1 bit. Because EWAH uses 64-bit words by default, we would use 128 bits per element. This would be less efficient than explicitly listing the set elements (e.g., 32 bits) by a factor of 4. Observe, however, that using (compressed) bitmaps for such sets is likely inefficient in any case: bitmaps are efficient due to bit-level parallelism when there are many words containing a mix of 1s and 0s.

Software libraries for compressed bitmaps will typically include an assortment of basic Boolean operations that operate directly on the compressed bitmaps. One would expect to find operations for AND, OR, and often one finds XOR, ANDNOT, and NOT. EWAH, like most other RLE-based formats, allows the operations AND, OR, XOR and ANDNOT between two compressed bitmaps ( $B_1$  and  $B_2$ ) to execute in time  $O(\text{EWAHSIZE}(B_1) + \text{EWAHSIZE}(B_2))$ . Moreover, the output of such an aggregate has compressed size bounded by the size of the input ( $\text{EWAHSIZE}(B_1) + \text{EWAHSIZE}(B_2)$ ). (For AND, the output is bounded by  $\min(\text{EWAHSIZE}(B_1), \text{EWAHSIZE}(B_2))$ .)

Some libraries support only binary operations, whereas others support *wide* queries: for instance, a wide AND would allow us to intersect four bitmaps in a single operation, rather than having to AND bitmaps together pairwise. Explicit support for wide operations can allow for better performance [28]. Threshold functions are wide queries when  $N > 2$ .

Our complexity analysis (Table III) assumes that we can iterate over the 1s in a compressed bitmap in  $\Theta(1)$  time each. We can indeed iterate over the 1s in a compressed EWAH bitmap quickly. Runs of fill words are not problematic: e.g., 64-bit EWAH uses 32-bit counters for the length of such runs, so runs of up to  $2^{32} \times 2^W$  identical bits can be marked with a single marker word. Moreover, we can also extract 1s from dirty words quickly. In Java, we can use the `Long.numberOfTrailingZeros` function and a simple loop: this function is commonly compiled to efficient machine instructions by the JVM (e.g., `bsr` on Intel and AMD processors).

### 3.2. Boolean Functions and Circuits

A Boolean function is a function of the form  $f : \{0, 1\}^k \rightarrow \{0, 1\}$ . For relevant background on Boolean functions, see Knuth [29]. A Boolean circuit over some basis (e.g., AND, OR, NOT) is a directed acyclic graph where each vertex is either a basis function or an input, and where some of the vertices are outputs. Boolean functions can be computed by Boolean circuits. As discussed in § 2, some Boolean functions are *symmetric*. These functions are unchanged under any permutation of their inputs. I.e., a symmetric function is completely determined if one knows the number of 1s (the Hamming weight) in its inputs. An example symmetric function outputs 0  $\iff$  the Hamming weight of its inputs is a multiple of 2: this is the XOR function.

### 3.3. Threshold Functions

Threshold functions, in the guise of  $T$ -overlap queries, have been used for approximate searching. Specifically, Sarawagi and Kirpal [10] show how to avoid unnecessary and expensive pairwise distance computations (such as edit-distance computations) by using threshold functions to screen out items that cannot be approximate matches. Their observation was that strings  $s_1$  and  $s_2$  must have many ( $T$ )  $q$ -grams in common, if they have a chance of being approximate matches to one another. Given  $s_1$  and seeking suitable  $s_2$  values, we take the set of  $q$ -grams of  $s_1$ . Each  $q$ -gram

is associated with a set of the words (more specifically, with their row IDs) that contain that  $q$ -gram at least once. Taking these  $N$  sets, we use a threshold function to determine values  $s_2$  that can be compared more carefully against  $s_1$ . Using  $q$ -grams, Sarawagi and Kirpal showed that  $T = |s_1| + q - 1 - kq$  will not discard any string that might be within edit distance  $k$  of  $s_1$ . In applications where  $k$  and  $q$  are small but the strings are long, this will create queries where  $T \approx N$ . (Similar formulae are known for Jaccard, cosine and dice similarities [9, 10].)

Closely related to  $T$ -overlap queries, we have Opt-threshold queries [11, 30]. In these queries,  $T$  is unspecified: the algorithm is responsible for choosing the largest threshold value that leads to a non-empty result. We could further generalize such queries by asking for the largest value  $T$  such that the result of the  $T$ -overlap query contains at least  $K$  elements. “Top- $K$ ” versions of the problem [7] are closely related, but are not symmetric bitwise Boolean operations—if the Opt-threshold result yields two elements, a top-1 query will return only one of them, despite both meeting the same threshold.

#### 4. ADVANCED QUERIES

To obtain results that correspond to a practical applications of bitmap indexes, we focus on using threshold functions over bitmap indexes to answer two different types of queries, *Many-Criteria* queries and *Similarity* queries.

**Many-Criteria Queries:** The first type of query has a set of criteria, and we are seeking those records that meet some minimum number of the criteria, but perhaps not all. E.g., consider a query that might be typical of some human-resources system (in pseudo-SQL).

```
SELECT * FROM table WHERE Gender="F" AND
  (City="Montreal" OR City="Vancouver") AND
  experience >= 24 AND education >= college;
```

If it corresponds to an application where we filter job candidates, maybe applying all constraints at once could lead to a small (or empty) result set. Or maybe we want to include exceptional candidates who fail to satisfy a few conditions. So we are willing to relax the constraint somewhat, by maybe requiring that only three of the constraints hold, as in the following example.

```
SELECT * FROM table WHERE
  CASE WHEN Gender="F" THEN 1 ELSE 0 END
+ CASE WHEN City="Montreal" THEN 1 ELSE 0 END
+ CASE WHEN City="Vancouver" THEN 1 ELSE 0 END
+ CASE WHEN experience >= 2 THEN 1 ELSE 0 END
+ CASE WHEN education >= college THEN 1 ELSE 0 END
>= 3;
```

**Similarity Queries:** The second type of query presents a prototypical item. We determine the criteria that this item meets, and then seek all items that meet (at least)  $T$  of these criteria. For example, if a user liked a given movie, he might be interested in other similar movies (e.g., same director, or same studio, or same leading star, or same date of release). As part of a recommender system, we might be interested in identifying quickly all movies satisfying at least  $T$  of these criteria. This might be viewed as setting a threshold on the Hamming distance between tuples.

Once the criteria have been defined, SQL can handle the rest of the query, as in the previous example. Critchley [31] proposes an alternative SQL-only solution using joins and SQL aggregation. We consider the evaluation of such external-memory approaches outside our current scope. Similarity queries have been used with approximate string matching [9, 10]. In this case, items are small chunks of text, and the occurrence of a particular 3-gram (a sequence of 3 consecutive letters) is a criterion. In that previous work, an index maps each 3-gram to a sorted list of integers that specify the chunks of text containing it. More recent work by others [32, 33] solves similar problems using bitmaps, one for each 2-gram.

---

**Algorithm 1** Row-scanning approach over a row store.

---

**Require:** A table with  $D$  attributes. A set  $\kappa$  of  $N \leq D$  attributes, and for each such attribute a desired value. Some threshold  $T$ .

```

1: Create an initially empty set  $s$ 
2: for each row in the table do
3:   counter  $c \leftarrow 0$ 
4:   for for each attribute  $k$  in  $\kappa$  do
5:     if attribute  $k$  of the row has the desired value then
6:       increment  $c$ 
7:   if  $c \geq T$  then
8:     add the row (via a reference to it) to  $s$ 
9: return the set of matching rows,  $s$ 

```

---

A generalization of a Similarity query presents *several* prototypical items, then determines the criteria met by at least one of them. We then proceed as before, finding all items in the database that meet at least  $T$  of the criteria. If there are  $n$  prototypes, we have a “Similarity( $n$ )” query.

Assuming one has a bitmap index, can one answer Many-Criteria and Similarity queries better than using the row-scan that would be done by a typical database engine? One of our contributions is to show that it is indeed the case. In § 4.1, we show that a simple bitmap-based algorithm (SCANCOUNT, see § 5.1) is able to outperform a row scan (e.g., by a factor of 6). Then in § 7 we show that other bitmap-based algorithms can outperform this simple approach (SCANCOUNT), sometimes by hundreds of times.

#### 4.1. An index is better than no index

Could a simple T-occurrence query can be more effectively answered without using a bitmap index? Before continuing our investigation with various novel algorithms, we want to establish that bitmap indexes can accelerate some T-occurrence queries. Our purpose is merely motivational: detailed experiments, including a description of our queries and datasets is given in § 7.

As a reference, we use a full table scan (see Algorithm 1), where the table is stored in RAM. To test the basic usefulness of a bitmap index, we use a simple algorithm (SCANCOUNT, see § 5.1 for details): we create an array of  $r$  counters initialized to zero. Then the bits of each bitmap are scanned in sequence, one bitmap at a time. When a 1-bit is found, the corresponding counter is incremented. The algorithm concludes with a full scan of the all counters.

We made 30 trials, on each of the datasets CensusIncome, Weather and TWEED. These are described in § 7.2 and have 42, 19 and 53 attributes, respectively. We randomly chose one value per attribute and randomly chose a threshold between 1 and the number of attributes, exclusively. This query corresponds to a Many-Criteria query. Table IV shows that using an EWAH index for this query was 4–6 times faster than scanning the table. The advantage persisted, but was smaller, when we did a Similarity query against a randomly chosen row. It is reassuring that a bitmap index using SCANCOUNT answered our queries faster than they would be computed from the base table. It remains to determine whether we can surpass SCANCOUNT. Section 7 shows that two algorithms can run at least  $1000\times$  faster than SCANCOUNT on certain queries, although speedups of  $3\times$  to  $5\times$  seem more typical.

## 5. EXISTING APPROACHES FOR THRESHOLD FUNCTIONS

We next present several different approaches to computing threshold functions that have been proposed in the literature. Several generalize to handle all symmetric functions, and several can be modified to solve Opt-threshold queries.

Table IV. Total time (ms) required for queries in our workload.  
 Top: Many-Criteria query. Bottom: Similarity query.

	CensusIncome	Weather	TWEED
EWAH SCANCOUNT	109	201	6
Row Scan (no index)	487	1212	23
Row Scan/SCANCOUNT(%)	450	600	380
EWAH SCANCOUNT	327	508	20
Row Scan (no index)	557	1344	22
Row Scan/SCANCOUNT(%)	170	260	110

### 5.1. Counter-based approaches

In information retrieval, it is common practice to solve threshold queries using sets of counters [34]. The simple SCANCOUNT algorithm of Li et al. [9] (previewed in § 4.1) uses an array of counters, one counter per item. The input is scanned, one bitmap at a time. If an item (as a bit set to 1) is seen in the current bitmap, its counter is incremented. A final pass over the counters can determine which items have been seen at least  $T$  times. In our case, items correspond to positions in the bitmap. If the maximum bit position is known in advance, if this position is not too large, and if one can efficiently iterate over the bit positions in a bitmap, then SCANCOUNT is easily implemented. These conditions are frequently met when the bitmaps represent the sets of row IDs in a table that is not exceptionally large.

SCANCOUNT is part of a family of counter-based approaches that have the characteristic that they count the occurrences of each item. They can handle arbitrary symmetric functions, since one can provide a user-defined function mapping  $[0, N]$  to Booleans. However, some counter-based approaches can be optimized specifically to compute threshold functions (see § 6.1).

To analyze SCANCOUNT, note that it uses  $\Theta(r)$  counters. We assume  $N < 2^W$ , so each counter occupies a single machine word. Even if counter initialization can be avoided (see Li et al. for details) the algorithm compares each counter against  $T$ . Also, the total number of counter increments is  $B$ . Together, these imply a time complexity of  $\Theta(r + B)$  and a space complexity of  $\Theta(r)$ . Aside from the effect of  $N$  on  $B$  (on average, a linear effect), note that this algorithm does not depend on  $N$ . (Li et al. [9] also present an alternative SCANCOUNT algorithm that generates an unsorted list in  $O(B)$  time. Generating a RLE-compressed bitmap would require sorting this output, and this could be a major overhead for queries with large outputs. Thus we do not consider this variation.)

The SCANCOUNT approach fits modern hardware well: the counters are accessed in sequence, during the  $N$  passes over them when they are incremented. Experimentally, we found that using 8-bit byte counters when  $N < 128$  usually brought a small (perhaps 15%) speed gain compared with 32-bit int counters. Perhaps more importantly, this also quarters the memory consumption of the algorithm. One can also experiment with other memory-reduction techniques: e.g., if  $T < 128$ , one could use a saturating 8-bit counter. Experimentally, we found that the gains usually were less than the losses that come from the additional conditional check required to ensure saturation. Based on our experimental results, the SCANCOUNT implementation used in § 7 switches between byte, short and int counters based on  $N$ , but does not use the saturating-count approach.

SCANCOUNT fails when the bitmaps have extreme  $r$  values. If we restrict ourselves to bitmaps that arise within a bitmap index, this implies that we have indexed a table with an extreme number of rows. However, instead of using  $r$  counters, we could use a small number and effectively partition the problem: choose a fixed number of counters  $r'$  and execute SCANCOUNT  $\lceil r/r' \rceil$  times, always reusing the same counters. We exploit this idea in § 6.3 with the RBMRG scheme.

It is easy to obtain an Opt-Threshold algorithm: SCANCOUNT begins as usual and obtains the  $r$  counters.  $T$  is the maximum value in the counters, and the algorithm then returns those elements whose counters equal  $T$ .

### 5.2. $T$ -occurrence algorithms for integer sets

Prior work [9, 10] has studied the case when the data is presented as sorted lists of integers rather than bitmaps. We consider the following  $T$ -occurrence algorithms: WHEAP [10], MGOPT [10, 11], and DSK [9]. For full details of these algorithms, see the papers that introduced them. All can be viewed as modifications to the basic WHEAP approach. This approach essentially uses an  $N$ -element min-heap that contains one element per input. Using the heap, it merges the sorted input sequences. As items are removed from the heap, we count duplicates and thereby know which elements had at least  $T$  duplicates. This approach can be generalized to compute any symmetric function, but it requires that we process the 1s in each list, inserting (and then removing) the position of each into an  $N$  element min-heap. The total time cost is thus  $O(B \log N)$  for sorted lists.

The WHEAP approach has been shown to have worse performance than MGOPT or DSK [8, 9, 10] and thus is not considered further.

The remaining algorithms are also based around heaps (MGOPT and DSK), but they are designed to exploit characteristics of real data, such as skew, that allow us to skip certain input elements. In contrast with other algorithms (e.g., WHEAP, RBMRG and SCANCOUNT), MGOPT and DSK do not generalize to arbitrary symmetric functions because such functions preclude skipping any input. This is illustrated by the (wide) XOR function, whose output always depends on all input bits—knowing all but one input bit is never enough to determine the output.

**Algorithm MGOPT:** Sarawagi and Kirpal’s MGOPT algorithm [10] sets aside the largest  $T - 1$  inputs. Any item contained only in these inputs cannot meet the threshold. Then it uses an approach similar to WHEAP with threshold 1 on the smallest  $N - T + 1$  inputs. For each item found in the smallest inputs, say with count  $t$ , the algorithm checks whether at least  $T - t$  instances of the item are found in the largest  $T - 1$  inputs. The items are checked in the largest inputs in ascending sequence. If one of the largest inputs is checked for occurrence of item  $x$ , and the next check is for the occurrence of item  $y$ , we know that  $y > x$ . Items between  $x$  and  $y$  in the big input will never be needed, and can be skipped over without inspection. Whereas we use bitmaps as inputs, Sarawagi and Kirpal use sorted lists of integers as inputs. Thus they can use a doubling/bootstrapping binary search to find the smallest value at least as big as  $y$ , without needing to scan all values between  $x$  and  $y$ . The portions skipped have been pruned.

As noted in § 3.1, providing random access is not a standard part of a RLE-based compressed bitmap library, although it is essentially free for uncompressed bitmaps. However, with certain compressed bitmap indexes one can “fast forward”, skipping portions of the index in a limited way: the JavaEWAH library [27] uses the fact that we can skip runs of dirty words (e.g., when computing intersections).

To bound the running time, we can distinguish the  $B - B'$  1s in the  $T - 1$  largest bitmaps from the  $B'$  1s in the remaining  $N - T + 1$  bitmaps. A heap of size  $O(N - T + 1)$  is made of the  $N - T + 1$  remaining bitmaps, and  $O(B')$  items will pass through the heap, at a cost of  $O(\log(N - T + 1))$  each. As each item is removed from the heap, it will be sought in  $O(T)$  bitmaps. Because the items sought are in ascending order, the  $T - 1$  bitmaps will each be processed in a single ascending scan that handles all the searches. Each of the  $B - B'$  1s in the remaining bitmaps should cost us  $O(1)$  effort. Thus we obtain a bound of  $O(B'(\log(N - T + 1) + T) + B - B') = O(B'(\log(N - T) + T) + B - B')$  for the time complexity of MGOPT.

A similar algorithm was earlier presented by Barbay and Kenyon [11]. Any input may appear in their heap, but at any time there will be  $T - 1$  inputs that are not in the heap. Setting aside the *largest* items (as with Sarawagi and Kirpal) seems like a useful enhancement. Indeed, consider our complexity bound of  $O(B'(\log(N - T) + T) + B - B')$ : each of the  $B'$  elements has a multiplicative cost factor of  $\log(N - T) + T$  whereas each of the other  $B - B'$  elements has a cost factor of 1. This reflects the fact that the  $B'$  elements are stored in a heap whereas the  $B - B'$  elements are merely accessed sequentially. Thus we prefer to minimize  $B'$ , which is done by setting aside the largest bitmaps.

Our analysis does not take fully into account the effect of pruning, because we might be able to skip many of the  $B - B'$  1s as we search forward through the largest  $T - 1$  bitmaps. Since these are

the *largest* bitmaps, if  $T$  is close to  $N$  or if the sizes (number of 1s) in the bitmaps vary widely, pruning could make a large difference. This depends on the data. Barbay and Kenyon present a detailed running-time analysis (with input as sorted integer lists) in terms of a “ $t$ -alternation” parameter for the problem instance. It matches their comparison-based lower bound for the problem in many cases, and in all cases it is within a factor of  $O(\log(N - T + 1))$  of the optimal complexity.

Barbay and Kenyon also describe how to obtain an Opt-threshold algorithm from any  $T$ -overlap algorithm by successively trying  $T = N, T = N - 1, \dots$  until a non-empty answer is obtained. Although naïve, the empty  $T$ -overlap queries have a predictable cost for MGOPT (no worse than the final query), whereas a binary search for  $T$  may make some more expensive queries.

**Algorithm DSK:** Algorithm DSK is essentially a hybrid of MGOPT and another pruning algorithm called MERGESKIP. MERGESKIP [9] is like WHEAP except that, when removing copies of an item from the heap, if there are not enough copies to meet the threshold, we remove some extra items. This is done in such a way that the extra items removed (and not subsequently re-inserted) could not possibly meet the threshold. (MERGESKIP is not described further here, because its performance is worse than DSK [8, 9]). Algorithm DSK processes the heap as in MERGESKIP, while it sets apart the largest bitmaps as in MGOPT. However, rather than following MGOPT and always setting apart the  $T - 1$  largest sets, it chooses the  $L$  largest sets where  $L$  is a tuning parameter. Li et al. determine another tuning parameter  $\mu$  experimentally, for a workload of queries against a given dataset. From  $\mu$  and the length of the longest input, Li et al. use a heuristic formula for  $L$  (see § 7.3). With a suitable  $L$ , we would not expect DSK to perform significantly worse than MGOPT.

Our running-time complexity bound for DSK is identical to that for MGOPT, and based on the same reasoning that ignores pruning. We cannot easily account for the pruning opportunities that DSK inherits from MERGESKIP and MGOPT. However, as with MGOPT, data-dependent pruning could reduce the  $B - B'$  term. As with MERGESKIP, the multiplicative  $B'$  factor can be reduced by data-dependent pruning [8].

Considering memory, note that MGOPT and DSK partition the inputs into two groups. Regardless of group, each compressed bitmap input will have an iterator constructed for it. The first group also go into a heap that accepts one element per input. Thus we end up with a memory bound of  $O(N)$ .

### 5.3. Boolean synthesis

A typical bitmap implementation provides a set of basic operations, typically AND, OR, NOT, XOR and sometimes ANDNOT<sup>¶</sup>. Since one can synthesize any Boolean function using AND, OR and NOT operations in combination, any desired bitwise bitmap function can be “compiled” into a sequence of primitive operations. For instance, the threshold functions over individual bits are, for  $N = 3$ ,

- $(T = 1) \vartheta(1, \{b_1, b_2, b_3\}) = b_1 \vee b_2 \vee b_3$ ,
- $(T = 2) \vartheta(2, \{b_1, b_2, b_3\}) = (b_1 \wedge b_2) \vee (b_2 \wedge b_3) \vee (b_1 \wedge b_3)$ ,
- $(T = 3) \vartheta(3, \{b_1, b_2, b_3\}) = b_1 \wedge b_2 \wedge b_3$ .

As a bitwise operation over EWAH bitmaps  $B_1, B_2$  and  $B_3$ , we have the corresponding Java expressions:

- $(T = 1)$  `EWAHCompressedBitmap.or(B1, B2, B3)`,
- $(T = 2)$  `EWAHCompressedBitmap.or(B1.and(B2), B2.and(B3), B1.and(B3))`,
- $(T = 3)$  `EWAHCompressedBitmap.and(B1, B2, B3)`.

Of course, these are examples: several Boolean expressions are equivalent to a given threshold function, and some are more efficient than others. For example, we can also write

<sup>¶</sup>The x86 extensions SSE2 and AVX2 support AND NOT, as do several bitmap libraries (EWAH included). Specifically, Intel has the `pandn` and `v pandn` instructions; however it does not appear the standard x86 instruction set has a corresponding instruction.

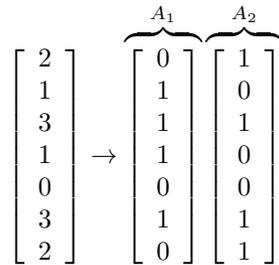


Figure 3. Example of a bit-sliced index [12].

$\vartheta(2, \{b_1, b_2, b_3\}) = (b_2 \wedge (b_1 \vee b_3)) \vee (b_1 \wedge b_3)$ —saving one Boolean operation over the alternative  $((b_1 \wedge b_2) \vee (b_2 \wedge b_3) \vee (b_1 \wedge b_3))$ .

In § 5.3.1 and 6.2 we introduce threshold algorithms BSTM and LOOPED that synthesize the desired bitmap function from standard bitmap operations (binary AND, OR, XOR, and ANDNOT operations). One major advantage is that this approach allows us to use a bitmap library as a black box, although it is crucial that the primitive operations have efficient algorithms and implementations. O’Neill and Quass [12] and Rinfret et al. [7] implicitly used this idea when doing arithmetic and comparison operations bitwise over a bit-sliced index. They note the opportunities for bit-level parallelism that arise. For example, the expression  $\vartheta(2, \{b_1, b_2, b_3\}) = (b_1 \wedge b_2) \vee (b_2 \wedge b_3) \vee (b_1 \wedge b_3)$  can actually compute 64 thresholds using only 5 bitwise operations on a 64-bit architecture. Without bit-level parallelism, we would need at least  $3 \times 64/2 = 96$  binary operations so that each input is used once: the benefits of bit-level parallelism are at least a factor of  $96/5 = 19.2$  in this case.

Unfortunately, it is computationally infeasible to determine the fewest required primitive operations that realize a desired Boolean function, except in the simplest cases [29]. In any case, for RLE compressed bitmaps, the relative costs of the primitive operations depend on the data.

*5.3.1. Adding: The BSTM algorithm.* Rinfret et al. [7] used the Boolean synthesis approach to solve a problem closely related to thresholding. In their information-retrieval problem, one seeks the top  $k$  documents that best match a set of keywords. The input is provided as a collection of bitmaps, one bitmap for each keyword. Set bits in a bitmap indicate the presence of the keyword in a document. The result of the query is a bitmap with  $k$  bits set.

While the “top- $k$ ” aspect means that the required computation is not a bitwise Boolean function, their method of solution can be adapted to solve our threshold problem, leading to the following algorithm, BSTM.

The algorithm begins with a Boolean bitwise function that views each of the  $N$  input bitmaps as representing a vector of single-bit numbers. Conceptually, these vectors of single-bit numbers are successively added (pointwise) to an accumulator vector whose entries may eventually grow to require  $\Theta(\log N)$  bits each. The multi-bit accumulator is represented as a “bit-sliced index” [12], a collection of bitmaps  $A_1, A_2, \dots, A_{\lfloor \log 2N \rfloor}$ , where bitmap  $A_1$  stores the least-significant bits of the totals,  $A_2$  stores the next-least-significant bits, and so forth (see Fig. 3). The totals can be considered to give the bitwise Hamming weight of the inputs; see Fig. 4. We express a Hamming weight using  $\lfloor \log 2N \rfloor$  bits, the minimal number of bits required to write  $N$  in binary form.

(Successive addition into an accumulator is not necessarily the best approach to adding  $N$  1-bit numbers to obtain  $\lfloor \log 2N \rfloor$ -bit Hamming weights. It is also possible [8] to use a balanced binary tree of adders, a “carry-save” adder approach [35], or (perhaps best) a “sideways-sum” circuit presented by Knuth [29, 7.1.2]. However, we choose to present the approach that most closely resembles the published BSTM algorithm.)

Once we have the Hamming counts, we need to check them to see which meet threshold  $T$ . For this, we can simplify the Range Predicate computation for bit-sliced indexes, Algorithm 4.2 of O’Neil and Quass [12]. Rather than check for  $\geq T$ , we do a greater-than comparison against  $T - 1$ . In Fig. 4, for  $T = 2$  we should compute the bitmap 1010..., since the Hamming counts of

Inputs			Hamming weight	Outputs	
$B_1$	$B_2$	$B_3$		$A_2$	$A_1$
0	1	1	$0 + 1 + 1 = 2$	1	0
0	0	1	$0 + 0 + 1 = 1$	0	1
1	1	1	$1 + 1 + 1 = 3$	1	1
1	0	0	$1 + 0 + 0 = 1$	0	1
	$\vdots$		$\vdots$	$\vdots$	$\vdots$

Figure 4. Computing the bitwise Hamming function.

the first and third rows exceed 2 – 1:  $10_2 > 1$ ,  $01_2 \not> 1$ ,  $11_2 > 1$  and  $01_2 \not> 1$ . (Again, it is possible to improve somewhat on the number of bitmap operations [8], but we choose to use the previously published algorithm, specialized to compute only greater-than.)

The BSTM algorithm is presented in Algorithm 2. The correctness of the computations of Hamming counts and greater-than have been previously established [7, 12].

We can illustrate the algorithm as follows:

1. Suppose we begin with three bitmaps:  $B_1 = 0011$ ,  $B_2 = 1010$ ,  $B_3 = 1110$ . Before the main loop of the algorithm, we have  $A_1 = B_1 = 0011$  and  $A_2 = 0000$ .
2. During the first pass through the main loop ( $i = 2$ ), we first compute  $C = 1010 \wedge 0011 = 0010$  and  $A_1 = 1010 \oplus 0011 = 1001$ . Because  $C$  is not empty, we further need to update  $A_2$  to  $0010$ . We now have  $j_{\max} = 2$ .
3. During the second pass through the main loop ( $i = 3$ ), we first set  $C = 1110 \wedge 1001 = 1000$  and  $A_1 = 1110 \oplus 1001 = 0111$ . Because  $C$  is not empty, we have to update  $A_2$  to  $1000 \oplus 0010 = 1010$ .
4. At the end of the main loop, we have  $A_1 = 0111$  and  $A_2 = 1010$  with  $j_{\max} = 2$ .
5. Suppose that the threshold is  $T = 2$ , then the last loop in the algorithm runs from 2 to 1. When  $j = 2$ , we set  $b_{\text{gt}} = A_2 = 1010$  and  $b_{\text{eq}} = \neg A_2 = 0101$ . When  $j = 1$ , we set  $b_{\text{eq}} = 0101 \wedge 0111 = 0101$ . The final answer is 1010.

To analyze the number of bitmap operations, we consider the following worst-case situation. The first item occurs in every bitmap and hence has a Hamming count of  $N$ . The second item occurs in every bitmap except the first, and in general the  $i^{\text{th}}$  item, for  $1 \leq i \leq N$ , occurs in all bitmaps except for the first  $i - 1$ . With this worst case, the first for loop iterates  $N - 1$  times, doing 2 bitmap operations before beginning the while loop. On beginning the  $i^{\text{th}}$  iteration, the items have Hamming counts ranging from 0 to  $i - 1$ ; in particular, some have Hamming counts with  $\Theta(\log i)$  trailing 1s. Thus there will be  $\Theta(\log i)$  other slices where carry propagation (involving two operations) is done. Together, we have  $2(N - 1) + \sum_{i=1}^{N-1} \Theta(\log i)$  operations to compute the Hamming weights. This quantity is  $\Theta(N \log N)$ , so the number of operations grows more than linearly in  $N$ , in the worst case. There are a few operations required to compare the Hamming weights against  $T$ . In the worst case,  $j_{\max} = \lfloor \log 2N \rfloor$ , and this many iterations are done. Each iteration does 3 bitmap operations (counting ANDNOT as a single operation), except when a bit of  $T - 1$  is 1; in that case, only 1 bitmap operation is done. If  $\#(T - 1)$  denotes the Hamming weight of  $T - 1$ , we need  $3\lfloor \log 2N \rfloor - 2\#(T - 1)$  bitmap operations. When  $N$  is large, the number of operations for comparison is inconsequential, due to the  $\Theta(N \log N)$  worst-case cost to compute Hamming weights. Nevertheless, this algorithm can do very few operations in some cases (approximately  $2N$  if the maximum Hamming weight is 1 and  $T > 1$ ).

An Opt-threshold algorithm can be obtained from a bit-sliced index with  $O(\log N)$  bitmap operations using ideas from Rinfret et al. [7].

**Symmetric functions beyond threshold:** We could apply a bit-sliced index to compute general symmetric functions. One could use the previous approach to compute the  $\lfloor \log 2N \rfloor$ -bit Hamming weights of the inputs followed by a computation of basic bitmap operations for the corresponding

---

**Algorithm 2** BSTM algorithm. Each input bitmap  $B_i$  is treated as a bit-slice index encoding 1-bit numbers.

---

**Require:**  $N$  bitmaps  $B_1, B_2, \dots, B_N$ , a threshold parameter  $T \in \{2, \dots, N - 1\}$

```

1: //  $A$  is the bit-slice-index accumulator for the Hamming weights
2: create  $\lfloor \log 2N \rfloor$  empty bitmaps  $A_1, A_2, \dots, A_{\lfloor \log 2N \rfloor}$ 
3:  $A_1 \leftarrow B_1$ 
4:  $j_{\max} \leftarrow 1$ 
5: // keep track of the  $A_j$ 's being modified
6: // Add remaining bitmaps (1-bit numbers) to the accumulator
7: for  $i \leftarrow 2$  to  $N$  do
8:    $C \leftarrow B_i \wedge A_1$ ;  $A_1 \leftarrow B_i \oplus A_1$ 
9:   // Propagate carries (C) to other slices
10:   $j \leftarrow 2$ 
11:  while  $C$  is not empty do
12:     $C, A_j \leftarrow C \wedge A_j, C \oplus A_j$ 
13:     $j \leftarrow j + 1$ 
14:     $j_{\max} \leftarrow \max(j, j_{\max})$ 
15: // Compare Hamming weights against  $T - 1$ 
16:  $b_{\text{eq}} \leftarrow 1111 \dots$ 
17:  $b_{\text{gt}} \leftarrow 0000 \dots$ 
18: if  $j_{\max} < \lfloor \log(2(T - 1)) \rfloor$  then
19:   return  $0000 \dots$ 
20: for  $j \leftarrow j_{\max}$  down to  $1$  do
21:   if bit  $j$  is set in  $T - 1$  then
22:     $b_{\text{eq}} \leftarrow b_{\text{eq}} \wedge A_j$ 
23:   else
24:     $b_{\text{gt}} \leftarrow b_{\text{gt}} \vee b_{\text{eq}} \wedge A_j$ 
25:     $b_{\text{eq}} \leftarrow b_{\text{eq}} \wedge \neg A_j$ 
26: return  $b_{\text{gt}}$ 

```

---

test (e.g., is the result between  $T_1$  and  $T_2$ ?) in lieu of the  $>$  computation making up the second half of Algorithm 2.

In cases where  $N$  is small, we are guaranteed to use few operations. Indeed, Knuth [29, 7.1.2] observes that since he has calculated the minimum number of operations (12) to realize any 5-input Boolean function, we can realize any symmetric Boolean function of  $N \leq 31$  inputs using no more than  $12 + s(N)$  operations, where  $s(N) = 5N - 2\#(N) - 3\lfloor \log N \rfloor - 3$  is the number of operations that a sideways-sum circuit uses to compute the Hamming weight [29, Prob. 7.1.2.30]. (For instance, if  $N = 31$  we use  $5 \times 31 - 2 \times 5 - 3 \times 4 - 3 = 130$  operations to compute the Hamming weight; with at most another 12 we can realize *any* symmetric function.)

## 6. NEW APPROACHES FOR THRESHOLD FUNCTIONS

In addition to existing approaches for computing threshold functions, we also propose a few novel techniques. They can be modified to solve Opt-threshold queries.

### 6.1. Mergeable-count structures.

A common approach to computing intersections and unions of several sets is to do it two sets at a time. To generalize the idea to symmetric queries, we represent each set as an array of values coupled with an array of counters. For example, the set  $\{1, 14, 24\}$  becomes  $\{1, 14, 24\}, \{1, 1, 1\}$ , where the second array indicates the frequency of each element (respectively). If we are given a second set ( $\{14, 24, 25, 32\}$ ), we supplement it with its own array of counters  $\{1, 1, 1, 1\}$  and

can then merge the two: the result is the union of two sets along with an array of counters ( $\{1, 14, 24, 25, 32\}, \{1, 2, 2, 1, 1\}$ ). From this final answer, we can deduce both the intersection and the union, as well as other symmetric operations.

Algorithm w2CTI takes this approach. Given  $N$  input bitmaps, it orders them by increasing cardinality and then merges each input, starting with the shortest, into an accumulating total. (The merge step is akin to the merge operation in the merge-sort algorithm.) A worst-case input has bitmaps of equal cardinality, each containing  $B/N$  items that are disjoint from any other input. At the  $i^{\text{th}}$  step the accumulating array of counters will have  $Bi/N$  entries and this will dominate the merge cost for the step. The total time complexity for this worst-case input is  $\Theta(\sum_{i=1}^{N-1} Bi/N) = \Theta(BN)$ . For memory use, the same input ends up growing an accumulating array of counters of size  $B$ .

Algorithm w2CTI refines this basic approach: although it ends up reading its entire input, during the merging stages it can discard elements that cannot achieve the required threshold. For instance, we can check the accumulating counters during each merge step. If there are  $i$  inputs left to merge, then any element that has not achieved a count of at least  $T - i$  can be removed from consideration (“pruned”).

In large-threshold cases, this pruning is beneficial. For instance, suppose  $T = N - \tau$  for some  $\tau \geq 1$ . Any item that has not occurred in one of the first  $\tau + 1$  bitmaps will be pruned. As these are the smallest bitmaps, they can contain no more than  $(\tau + 1)B/N$  items, and this bounds the size of the accumulator in any of the  $N$  merge operations. The total cost of the merge operations is thus in  $O(B + N(\tau + 1)B/N) = O(\tau B) = O((N - T)B)$ . However, pruning is mostly unhelpful with the worst-case input, if  $T = 2$ . We cannot discard any item until the final merge is done, because the last input set could push the count (currently 1) of any accumulated item to 2, meeting the threshold. Thus, with  $T = 2$  we find a worst-case time bound of  $\Omega((N - T)B)$ .

## 6.2. LOOPED algorithm

Given  $N$  bitmaps  $B_1, B_2, \dots, B_N$ , the LOOPED algorithm (see Algorithm 3) seeks to compute the threshold problem for all thresholds  $1, 2, \dots, T$  using corresponding temporary bitmaps  $C_1, C_2, \dots, C_T$ . Let us consider a concrete example:  $B_1 = 0011$ ,  $B_2 = 1110$  and  $B_3 = 1000$  with  $T = 2$ . At first, we process bitmap  $B_1$  and get  $C_1 = 0011$ ,  $C_2 = 0000$ . We then process bitmap  $B_2$  and get  $C_1 = 1111$ ,  $C_2 = 0010$ . We then process the last bitmap to get  $C_1 = 1111$ ,  $C_2 = 1010$ . As with BSTM, the LOOPED approach also combines basic bitmap operations to synthesize the threshold operation.

Our algorithm uses dynamic programming and is based on the following recurrence formula:  $\vartheta(T, \{b_1, b_2, \dots, b_N\}) = \vartheta(T, \{b_1, \dots, b_{N-1}\}) \vee \vartheta(T - 1, \{b_1, \dots, b_{N-1}\}) \wedge b_N$ . I.e., we can achieve a given threshold  $T$  over  $N$  bits, either by achieving it over  $N - 1$  bits, or by having a 1-bit for  $b_N$  and achieving threshold  $T - 1$  over the remaining  $N - 1$  bits. We can use bit-level parallelism to express this as a computation over bit vectors; loops can compute the result specified by the recurrence. Although  $\Theta(NT)$  bit-vector operations are used, we need only  $\Theta(T)$  working bitmaps during the computation, in addition to our  $N$  inputs.

The number of binary bitmap operations is  $2NT - N - T^2 + T - 1$  and depends linearly on  $T$ , which is unusual compared with our other algorithms. However, the number of bitmap operations is not necessarily a good predictor of performance when using compressed bitmaps. It depends on the dataset.

An Opt-threshold algorithm is easily obtained from LOOPED: first do the calculation with the maximum permitted value of  $T$ — i.e.,  $N$  or  $N - 1$ . Then find the maximum value  $i$  such that  $C_i$  is not empty. This algorithm does  $\Theta(N^2)$  bitmap operations, requiring  $\Theta(N^2r/W)$  time if we assume bitmap compression is ineffective.

**Algorithm 3** LOOPED algorithm.**Require:**  $N$  bitmaps  $B_1, B_2, \dots, B_N$ , a threshold parameter  $T \in \{2, \dots, N - 1\}$ 


---

```

1: create  $T$  bitmaps  $C_1, C_2, \dots, C_T$  initialized with false bits
2:  $C_1 \leftarrow B_1$ 
3: for  $i \leftarrow 2$  to  $N$  do
4:   for  $j \leftarrow \min(T, i)$  down to 2 do
5:      $C_j \leftarrow C_j \vee (C_{j-1} \wedge B_i)$ 
6:    $C_1 \leftarrow C_1 \vee B_i$ 
7: return  $C_T$ 

```

---

	$B_1$	$B_2$	Count
→	0	0	0
→	0	1	1
→	1	1	2
	1	1	2
	1	1	2
→	0	0	0
→	0	1	1
→	0	1	1

Figure 5. Runs, showing positions where new runs begin (and where the current Hamming-weight count needs to be adjusted).

### 6.3. Exploiting run-length coding: RBMRG

Algorithm RUNNINGBITMAPMERGE (henceforth RBMRG) is a refinement of an algorithm presented in Lemire et al. [13]. The simplest form of the algorithm is for bitmaps that have been run-length encoded; handling word alignment adds additional complexity that is discussed in § 6.4.

See Algorithm 4 and Fig. 5. The approach considers runs as integer intervals, and each bitmap provides a sorted sequence of intervals. For example, the bitmap  $B_1 = 00111000$  might be viewed as the sequence (bit: 0, range [0, 1]; bit 1, range [2, 4]; bit 0, range [5, 7]).

Heap  $H$  enables us to quickly find, in sorted order, those points where intervals begin (and the bitmaps involved). At such points, we calculate the function on its revised inputs; in the case of symmetric functions such as threshold, this can be quick. As we sweep through the data, we update the current count. Whenever a new interval of 1s begins, the count increases; whenever a new interval of 0s begins, the count decreases. Assuming  $\log N \leq W$ , the new value of a threshold function can be determined in  $\Theta(1)$  time whenever an interval changes. (The approach can be used with Boolean functions in general, but the complexity analysis might differ.)

Every run passes through a  $N$ -element heap, giving a running time of  $O(\text{RUNCOUNT} \log N)$ . One can implement the  $N$  required iterators in  $O(1)$  space each, leaving a memory bound of  $O(N)$ .

As an extreme example where this approach would excel, consider a case where each bitmap is either entirely 1s or entirely 0s. Then  $\text{RUNCOUNT} = N$ , and in  $O(N \log N)$  time we can compute the output, regardless of  $r$  or  $B$ .

### 6.4. Implementing RBMRG with EWAH

The EWAH implementation of RBMRG processes runs of clean words as described, but word alignment means that we must consider dirty words also. If the interval from  $a'$  to  $a$  corresponds to  $N_{\text{clean}}$  bitmaps with clean runs, of which  $k$  are clean runs of 1s, the implementation distinguishes three cases:

1.  $T - k \leq 0$ : the output is 1, and there is no need to examine the  $N - N_{\text{clean}}$  bitmaps that contain dirty words. This pruning will help cases when  $T$  is small.

2.  $T - k > N - N_{\text{clean}}$ : the output is 0, and there is no need to examine the dirty words. This pruning will help cases when  $T$  is large.
3.  $1 \leq T - k \leq N - N_{\text{clean}}$ : the output will depend on the dirty words. We can do a  $(T - k)$ -threshold over the  $N - N_{\text{clean}}$  bitmaps containing dirty words.

We process the  $N - N_{\text{clean}}$  dirty words as follows.

- (a) If  $T - k = 1$  (resp.  $T - k = N - N_{\text{clean}}$ ), we compute the bitwise OR (resp. AND) between the dirty words.
- (b) If  $T - k \geq 128$ , we always use SCANCOUNT using 64 counters (see § 5.1).
- (c) Otherwise, we compute  $\beta$ , the number of 1s in the dirty words. This can be done efficiently in Java since the `Long.bitCount` function on desktop processors is typically compiled to fast machine code. If  $2\beta \geq (N - N_{\text{clean}})(T - k)$ , we use the LOOPED algorithm (§ 6.2), otherwise we use SCANCOUNT again.

We arrived at this particular approach by trial and error: we find that it gives reasonable performance.

Like MGOPT and DSK, RBMRG has minimal memory usage ( $O(N)$ , see Table III). Indeed, the memory usage of RBMRG does not depend on the length of the bitmaps ( $r$ ) in contrast to competitive schemes like SCANCOUNT, BSTM and LOOPED. This might make RBMRG especially suitable for multicore processing where all cores share the same limited cache memory.

When the bitmaps are poorly compressible, we can view RBMRG as a memory-conscious version of SCANCOUNT. Indeed, whereas SCANCOUNT uses  $r$  counters, RBMRG uses only 64 counters—constantly recycling them.

The algorithm would be a suitable addition to compressed bitmap index libraries that are RLE-based; as a result of this work, we have added it to JavaEWAH [27]—the complete implementation is freely available online.

To illustrate the algorithm, consider the following problem involving 4 bitmaps and a threshold query with  $T = 3$ .

1. Without compression, but in terms of 64-bit words, our 4 bitmaps are  
 $B_1 = \{0x0, 0x0F, \underline{0x00}, \underline{0x00}, \underline{0x00}, 0x0F, 0x01\}$ ,  
 $B_2 = \{0x0, 0xF0F, \underline{0xF\cdots F}, \underline{0xF\cdots F}, 0x0F, 0x0F, 0x01\}$  and  
 $B_3 = B_4 = \{\underline{0xF\cdots F}, \underline{0xF\cdots F}, \underline{0xF\cdots F}, \underline{0xF\cdots F}, 0x0F, 0x0F, 0x01\}$ .

When using EWAH compression, we have that  $B_1$  contains two runs of fill words (containing 0s and shown underlined) and two runs of dirty words. We have that  $B_2$  contains two runs of fill words, and two runs of dirty words,  $B_3$  contains one run of fill words and one run of dirty words. Finally,  $B_4$  is identical to  $B_3$ .

2. The algorithm considers four runs (one for each bitmap). Initially, it considers a run of 0s from  $B_1$  (of length 1 word), a run of 0s from  $B_2$  (of length 1 word), and two other runs of 1s (of length 4 words) from  $B_3$  and  $B_4$ . Using a heap, it determines that the shortest run has length 1 word. The Hamming weight of the fill words is 2 and there is no dirty word, so immediately it outputs a single fill word of 0s by case 2.
3. We have a run of one dirty word from  $B_1$  (0x0F), a run of one dirty word from  $B_2$  (0xF0F) and the same run of fill words from  $B_3$  and  $B_4$  (with a remaining length of 3 words). Because  $T = 2$  and we have one fill word made of 1s, the algorithm outputs the bitwise OR of the two dirty words (0xF0F) by case 3a.
4. The algorithm then looks at the beginning of a run of 0s (of length 3 words) in bitmap  $B_1$ , and at runs of 1s (of length 2 words) in  $B_2$ ,  $B_3$  and  $B_4$ . The algorithm immediately outputs two fill words of 1s by case 1.
5. We have a run of 0s of length 1 word in  $B_1$ , and runs of dirty words from  $B_2$ ,  $B_3$  and  $B_4$ . The algorithm thus outputs the bitwise AND between the first dirty words from  $B_2$ ,  $B_3$  and  $B_4$  (0x0F) by case 3a.
6. The algorithm looks at 4 runs of dirty words of length 2 words from  $B_1$ ,  $B_2$ ,  $B_3$  and  $B_4$ . In this instance, case 3c applies. It collects the first 4 dirty words from the 4 bitmaps (0x0F, 0x0F, 0x0F, 0x0F). The algorithm computes the number of 1s ( $\beta = 16$ ) and it uses the LOOPED

**Algorithm 4** Algorithm RBMRG.

---

**Require:**  $N$  bitmaps  $B_1, \dots, B_N$  over  $r$  bits, some Boolean function  $\gamma$  such as  $\vartheta(T, \{\cdot\})$   
 $I_i \leftarrow$  iterator over the runs of identical bits of  $B_i$   
 $\Gamma \leftarrow$  a new buffer to store the aggregate of  $B_1, \dots, B_N$  (initially empty)  
 $\gamma \leftarrow$  the bit value determined by  $\gamma(I_i, \dots, I_N)$   
 $H \leftarrow$  a new  $N$ -element min-heap storing ending values of the runs along with their iterators  
 $a' \leftarrow 0$   
**while true do**  
  let  $a$  be the minimum of all ending values for the runs of  $I_1, \dots, I_N$ , determined from  $H$   
  append run  $[a', a]$  to  $\Gamma$  with value  $\gamma$   
   $a' \leftarrow a + 1$   
  **for** iterator  $I_i$  with a run ending at  $a$  (selected from  $H$  as root element) **do**  
    increment  $I_i$ ; if  $I_i$  has reached the end, terminate the algorithm  
    Update  $\gamma$  with the new value of  $I_i$   
    Update the heap  $H$  with the new value of  $I_i$

---

algorithm, outputting 0x0F. On the next four dirty words (0x01, 0x01, 0x01, 0x01), it finds that  $\beta = 4$  and uses the SCANCOUNT algorithm on the last four dirty words; it outputs 0x01.

7. The algorithm concludes with the solution  $\{0x0, 0xF0F, 0xF\cdots F, 0xF\cdots F, 0x0F, 0x0F, 0x01\}$ .

## 7. DETAILED EXPERIMENTS

We conducted extensive experiments on the various threshold algorithms, using EWAH compressed bitmaps generated from real datasets. The various bitmaps in our study, even within a particular dataset, vary drastically in characteristics such as density. We discuss this in more detail before giving the experimental results.

### 7.1. Platform

Experimental results were gathered on a desktop with an Intel Core i7 2600 (3.4 GHz, 8 MB of L3 CPU cache) processor with 16 GB of memory (DDR3-1333 RAM with dual channel). Because all algorithms are benchmarked after the data has been loaded in memory, disk performance is irrelevant.

The system ran Ubuntu 12.04LTS with Linux kernel 3.2. During experiments, we disabled dynamic overclocking (Turbo Boost) and dynamic frequency scaling (SpeedStep). Software was written in Java (version 1.7), compiled and run using OpenJDK (IcedTea 2.4.7) and the OpenJDK 64-bit server JVM.

We used the JavaEWAH software library [27], version 0.8.1, for our EWAH compressed bitmaps. It includes an implementation of the RBMRG algorithm. Our measured times were in wall-clock milliseconds. All our software is single-threaded.

### 7.2. Data

Real data tests were done with datasets IMDB-3gr, PGDVD, PGDVD-2gr, CensusIncome, TWEED and Weather<sup>||</sup>. Our first three datasets (IMDB-3gr, PGDVD and PGDVD-2gr) are similar to datasets used in related work [9]. They are *not* indexed as if they were database tables. The last three datasets (Weather, TWEED and CensusIncome) are more representative of content from relational databases and they are indexed as such (see Fig. 2).

<sup>||</sup>See <http://lemire.me/data/symmetric2014.html>.

Table V. Characteristics of real datasets. Overall bitmap density is the number of 1s, divided by the product of the number of rows and the number of bitmaps ( $B/(Nr)$ ).

Dataset	$r$	Attributes	Bitmaps	Average Density		
				Overall	In M-C workload	In Sim workload
IMDB-3gr	1783816	—	50663	$4.1 \times 10^{-4}$	—	$1.9 \times 10^{-2}$
PGDVD	2439448	—	11118	$2.9 \times 10^{-4}$	—	$3.7 \times 10^{-3}$
PGDVD-2gr	3513575	—	755	$2.8 \times 10^{-1}$	—	$6.1 \times 10^{-1}$
CensusIncome	199523	42	103419	$4.1 \times 10^{-4}$	$1.5 \times 10^{-1}$	$3.4 \times 10^{-1}$
TWEED	11245	53	1167	$4.5 \times 10^{-2}$	$2.0 \times 10^{-1}$	$5.5 \times 10^{-1}$
Weather	1015367	19	18647	$1.0 \times 10^{-3}$	$7.6 \times 10^{-2}$	$1.2 \times 10^{-1}$

IMDB-3gr is based on descriptions of a dataset used in the work of Li et al. [9], in an application looking for actor names that are at a small edit distance from a (possibly misspelt) name. Each bitmap corresponds to a 3-gram found in some actor’s name. The  $k^{\text{th}}$  bit in the bitmap indicates whether the  $k^{\text{th}}$  actor’s name contains this 3-gram.

The PGDVD dataset has a bitmap for each of 11 118 files on the Project Gutenberg DVD [36]. Each bitmap represents the vocabulary set found in that file (the total vocabulary had over 2.4 million words).

PGDVD-2gr is similar to IMDB-3gr except that, instead of actor names, we formed 2-grams from chunks of text from the Project Gutenberg DVD. Each chunk was obtained by concatenating paragraphs until we accumulated at least 1000 characters. We rejected any paragraph with over 20 000 characters—this protected us from some non-text content (e.g., the digits of  $\pi$ ) on the Project Gutenberg DVD.

We also chose three more conventional datasets in the context of relational databases. Two have many attributes, CensusIncome [13, 37] and TWEED [38]. The former is a census extract; the latter is a small dataset containing historical information on terrorist attacks in Europe, for which we used all attributes, rather than the projection used by Webb et al. [39]. We also used the entries for September 1985 of a larger dataset (Weather) [40, 41]. This particular month has been used previously [41], although the previous use had projected only 9 attributes, whereas we used all of them. We selected just one month of data because the full dataset (123 million rows) caused several of the tested algorithms to run out of memory. It would have been difficult to report meaningful aggregate results with such failures.

A bitmap index was built for each conventional dataset, and it had a bitmap for every attribute value. Row reordering can improve RLE-compressed bitmap indexes [13], but it is not always possible. Our indexes used the given (unsorted) row order. In CensusIncome, one attribute is responsible for 99 800 of the bitmaps; the remaining 3619 bitmaps are much denser than these 99 800. Together, our real datasets cover a range of application areas, lengths, widths and densities.

By design, our work does not consider external-memory indexes on very large datasets. Thus our datasets are chosen so the bitmaps for each query fit in RAM. However, even if our machines had more than 16 GB of RAM, we might still want to partition the problems so that bitmaps do not span much more than a few million bits, to alleviate caching issues.

### 7.3. Queries Used

To assess our algorithms, we generated two random workloads, one with 5000 Many-Criteria queries and the other with 5000 Similarity queries (see § 4).

- To generate a Many-Criteria query, we randomly chose a dataset. Many-Criteria queries do not make much sense for IMDB-3gr, PGDVD or PGDVD-2gr. For instance, almost all 3-grams have extremely sparse bitmaps and empty results can be expected, even with  $N$  large and  $T$  small. Therefore, we chose the dataset with equal probability from {CensusIncome, TWEED, Weather}. Having determined the dataset, we chose  $N$  next. We could pick random values of  $N$  uniformly at random in a range ( $[3, 1000]$ ), but most values of

$N$  would then be relative large ( $\gg 10$ ). Instead, we used a discretized log-uniform distribution with  $\log N \sim U[\log 3, \log 1000]$ , which resulted in a workload where small values of  $N$  were more common, but large values of  $N$  sometimes occurred. We then chose (uniformly at random, with replacement)  $N$  attributes on which criteria were established, by choosing one of their bitmaps uniformly. We finally randomly chose an integer threshold  $T$  uniformly from  $[2, N' - 1]$ , where  $N'$  is the number of attributes on which criteria had been established.

- We considered Similarity queries with  $n$  prototypes (henceforth Similarity( $n$ )). For such a query, we selected (with equal probability) one of our datasets. Then we chose  $n$  distinct prototypes  $\{r_i \mid i \leq n\}$ , each represented by a row identifier chosen uniformly from  $[0, r)$ . We then found the set of bitmaps matching at least one of them,  $\bigcup\{B_i \mid \exists j \text{ such that } B_i[r_j] = 1\}$ ; we have that  $N$  was the number of matching bitmaps.

The probability of Similarity(1), Similarity(5), Similarity(10), Similarity(15) and Similarity(20) queries were 20 % each.

In the last columns of Table V, we present the average density of the bitmaps involved in our query workloads.

We agree with Jia et al. [16] that it does not make sense to time queries whose answers are empty. Regardless whether we had a Many-Criteria or a Similarity query, if the answer to the threshold query was empty and  $T > 2$ , we chose (uniformly at random) a new value of  $T$  between 2 and the existing value of  $T$ . If the threshold query had an empty answer when  $T = 2$ , we discarded the query and generated a new one.

Considering the 10 000 queries in the two workloads, there were 54 queries with  $N > 1000$ : the maximum value of  $N$  was 11 115 whereas the average  $N$  was 165. The maximum value of  $T$  was 10 863, but the average was 42. The largest set of input data, in terms of storage, was 185 MB; in terms of cardinality, it was 740 million items.

The bitmaps involved in our queries are denser than the average bitmap. Indeed, the last three columns in Table V differ: the first shows the average density of bitmaps from the dataset, whereas the second and third respectively show the average density of the bitmaps actually selected in our two workloads. We see that the latter are denser (anywhere from twice as dense to 1000 times denser). This is a consequence of how we pick the queries.

- Many-Criteria queries tend to choose dense bitmaps because the sparsest bitmaps frequently come from the same (high cardinality) attribute, and all attributes are given an equal probability.
- For Similarity queries, we note that denser bitmaps are more likely to appear in  $\bigcup\{B_i \mid \exists j \text{ such that } B_i[r_j] = 1\}$ .

**Choosing  $\mu$  for DSK:** The DSK algorithm requires a tuning parameter  $\mu$ , which depends on the dataset. Li et al. [9] sketch a process for choosing  $\mu$ :

- For each dataset, select a representative workload of queries.
- For each query, execute DSK with various choices of  $\mu$ , recording the  $\mu$  that produced the fastest answer for that query.
- Average the recorded  $\mu$  values for a dataset.

We followed their approach. For the workload, we generated 500 queries using the random query generation process already described for Many-Criteria queries. We tried up to 20 values of  $\mu$  for each query, using the relationship  $L = T/(\mu \log M + 1)$  given by Li et al. ( $M$  is the cardinality of the largest bitmap) to choose  $\mu$  values. When  $T \leq 20$ , we tried  $L = 1, L = 2, \dots, L = T - 1$ . Otherwise, we tried all values in  $[T - 5, T) \cup \{\lceil \frac{T-6}{15}i \rceil \mid i \in [1, 15]\}$ . For CensusIncome, TWEED and Weather, the respective  $\mu$  values were 0.0388, 0.0452 and 0.0444. We then repeated the process with 500 Similarity queries, obtaining  $\mu$  values of 0.180 (IMDB-3gr), 0.0752 (PGDVD), 0.004 81 (PGDVD-2gr) 0.0560 (CensusIncome), 0.0112 (TWEED) and 0.0351 (Weather).

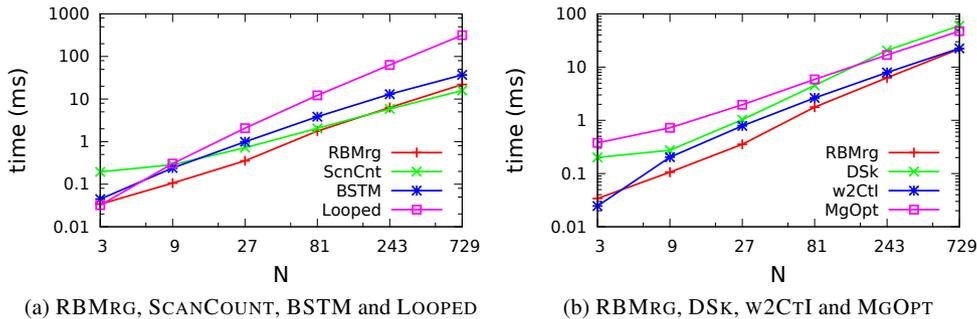


Figure 6. Effect of  $N$  on the running times of the algorithms, for Many-Criteria majority queries on CensusIncome. For clarity, we use two plots to represent the 7 algorithms: the same RBMRG timings appear in the two plots.

**Competitions:** We assess the effectiveness of the various algorithms by measuring their wall-clock times on the queries in our workload. Each query can be viewed as a competition between algorithms.

Unfortunately, for some of the larger queries, w2CTI (§ 6.1) was not able to complete without running out of memory. It is unfair to give the algorithm a nearly infinite running time when trying to compute its aggregate performance over the workload. However, it is also unfair to omit the running time from an average, as it is excusing a result where even a good algorithm would take a long time. Our solution is to assign the running time of the slowest algorithm that *did* complete the competition.

#### 7.4. Experimental Effects of $N$ and $T$

As previewed in Table III, our theoretical bounds suggest that the various algorithms' running times are all affected\*\* by  $N$ . Some algorithms are affected by  $T$  and others are highly sensitive to the characteristics of the datasets being processed. A few anecdotal examples given here illustrate these effects and help confirm/augment our theoretical bounds, as well give some idea of the constants that are abstracted away during our asymptotic analyses. (See [8] for more extensive experiments.)

We first fixed the dataset (CensusIncome) and kind of query (Many-Criteria majority) to examine the effect of  $N$  in that particular scenario. For chosen values of  $N$ , we took 100 queries on CensusIncome and, for each algorithm, averaged their running times. These are majority queries: threshold queries with  $T = \lceil N/2 \rceil$  and  $N$  odd. (Unlike our normal workload, we had a mixture of queries returning empty and non-empty results.) In this scenario, we have that  $r$  and  $W$  are fixed while  $B$ ,  $B'$ ,  $T$  and  $N - T$  grow with  $N$ . From Table III, we might expect (using an admittedly naïve analysis) the running time of SCANCOUNT to grow linearly ( $N$ ), the running time of MGOPT, BSTM, DSK and RBMRG to grow as  $N \log N$  and the running time of w2CTI and LOOPED to grow quadratically ( $N^2$ ). (Experimentally, it is often difficult to distinguish linear growth from  $N \log N$ , but quadratic growth will stand out as having a larger slope on a log-log plot.) Figure 6 shows how our algorithms behaved in this particular test as  $N$  was changed.

Focusing on  $N \geq 9$ , we see that RBMRG had the best absolute performance except when  $N$  was very large. In such cases, SCANCOUNT was faster. In terms of growth rate (corresponding to slope in a log-log plot), LOOPED stands out, with a growth rate that corresponds to approximately  $N^{\log_3 5} \approx N^{1.5}$  — better than our  $O(N^2)$  worst-case bound suggests, but still worse than the other algorithms. The slopes of RBMRG and DSK are higher than those of the other algorithms. In other tests we rarely see w2CTI performing well, due to its large memory requirement. However, on these queries against our small CensusIncome dataset, it seemed to display a slightly sub-linear

\*\*For table entries (such as that for SCANCOUNT) where  $B$  is given but  $N$  is not explicit, note that  $B$  grows as  $N$  grows: given a set of  $N$  bitmaps with  $B$  1s, if a new non-empty bitmap is added, the total number of 1s increases.

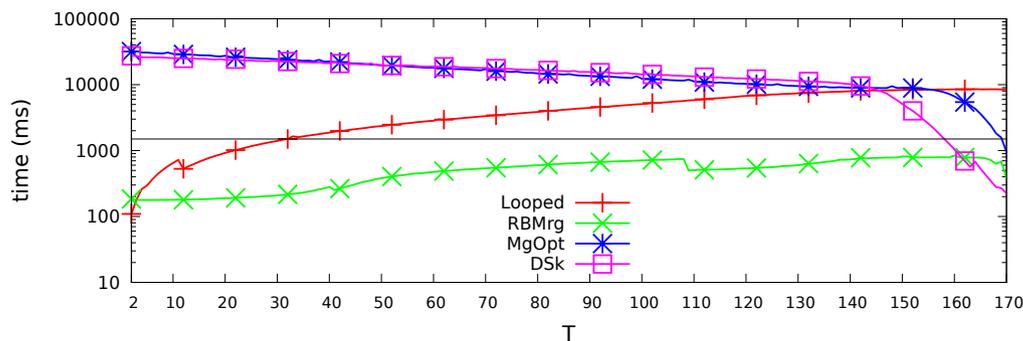


Figure 7. Effect of  $T$  on the running times of several algorithms.  $N = 171$ , and the dataset is PGDVD-2gr. Other algorithms were less affected by  $T$ . The SCANCOUNT and BSTM algorithms took about 1500 ms for all values of  $T$ . The w2CTI algorithm dropped steadily from about 13 000 ms for small  $T$  to about 10 000 ms for the largest  $T$  values.

running time growth in  $N$ , far better than our quadratic bound indicated. SCANCOUNT was similar in apparently having sub-linear growth. The query-generation approach means that  $B$ , the number of set bits, grows proportionally with  $N$ . However, the total number of possible items,  $r$ , is constant and, for this dataset, we have  $r$  approximately 8 times larger than the number of set bits in an average input bitmap. With  $N = 3$ , we have  $B$  significantly less than  $r$ . They are comparable at  $N = 9$ . This can explain apparently sublinear growth as  $N$  grew from 3 to moderately large values of  $N$ . The explanation for w2CTI appears simpler: on this dataset, our majority queries had empty answers for  $N > 27$ . The ever-larger thresholds presented more opportunities for pruning that w2CTI exploited. We might have expected similar improvements from both MGOPT and DSK, but only MGOPT seems to have had them. The reason may be that DSK prunes especially well when  $N - T$  is small. Since we have  $T \approx \frac{N}{2}$ , we see DSK outperforming MGOPT for small  $N$ , but then becoming closer to MGOPT as  $N$  increases.

Discussion during our theoretical analyses has indicated that large  $T$  values improve pruning possibilities—and hence should lead to improved running times—for several algorithms (w2CTI, MGOPT, DSK and RBMRG). However, these pruning effects are data dependent and hence, except for w2CTI, were not reflected in our asymptotic bounds. (In fact, our bounds for MGOPT and DSK actually suggest running time might *increase* somewhat with  $T$ .) Our LOOPED algorithm is expected to grow linearly with  $T$ , due to a  $NT$  term. Moreover, small values of  $T$  can lead to pruning in RBMRG. Experiments can help us see, at least in specific cases, the effects of pruning that are not captured by our asymptotic running-time bounds.

We then chose an arbitrary query (a Similarity(1) query against PGDVD-2gr): Fig. 7 shows the effect of varying  $T$ , on one particular set of 171 bitmaps. Absolute times are shown, but on a logarithmic scale. Increasing  $T$  (and thereby decreasing the size of the answer) affected algorithms differently, and DSK is particularly notable, improving two orders of magnitude and going from one of the worst algorithms for small  $T$ , to the best for large  $T$ . It is difficult to see, but RBMRG had a 43% speedup when  $T$  increased from 169 to 170. Overall, it tended to perform best when  $T$  was small, however. The different pruning opportunities can affect which algorithm is fastest for a given  $T$ . For this collection of bitmaps, we got best results from LOOPED at  $T = 2$ , then RBMRG until  $T \approx 160$ , after which DSK was fastest. The potentially enhanced pruning of DSK over MGOPT was not manifest until  $T = 145$  on this dataset, whereas for Fig. 6 even majority queries usually showed an advantage for DSK.

### 7.5. Comparing Algorithms on Our Workloads

Because the state of our system varies slightly over time, we make an error when measuring the time required by the implementation of an algorithm. We think of the *true* performance of the implementation of the algorithm as its best possible speed on a given query. We can measure this

best possible speed with little error by repeating the execution hundreds of times. However, given our 10 000 queries and 7 algorithms, these repeated tests would require more than a year to complete. Thus we tested each algorithm on each query only a few times.

Moreover, when comparing algorithms, we did not merely want to decide whether an algorithm is superior to another; for this purpose a standard statistical test would have sufficed. Instead we wanted to compare the results numerically, so we first estimated our measurement error by generating 5 random queries. Because timings errors are always additive, given a query, we ran each algorithm on each query 200 times. The minimum timing is assumed to be error-free: the fastest test out of 200 tests is a good approximation of the fastest possible result. Any larger timing is in error. We have 6 datasets and 7 algorithms, so we collected  $200 \times 6 \times 7$  measurement errors per query. We found that the 99<sup>th</sup>-percentile error was less than 10 % for the 5 reference queries.

We then considered our set of 10 000 queries. For each query, dataset and algorithm, we fixed the number of repetitions so that the total running time is at least 1 s. Supposing the measured running times on some query are  $t_1$  and  $t_2$  for two given algorithms, we say that the first algorithm is faster only if  $t_1 < 0.8 \times t_2$ . We anticipate mis-identifying a superior algorithm less than 1 % of the time.

Tables VI and VII compare each pair of algorithms using our two workloads. The cell associated with the row for algorithm  $\mathcal{A}_1$  and the column for algorithm  $\mathcal{A}_2$  gives the number of times that  $\mathcal{A}_1$  had performance superior to that of  $\mathcal{A}_2$ .

For all cases when  $t_1 \leq t_2$ , we record the percentage improvement measured. (A percentage improvement of  $x$  means that  $\mathcal{A}_1$  is  $1/(1-x)$  times faster than  $\mathcal{A}_2$ . That is, improvements of 99 %, 90 %, 80 %, 50 % indicate that we have  $100\times$ ,  $10\times$ ,  $5\times$  and  $2\times$  the speed.) To assess these performance improvements (ignoring the possibility of measurement error), we show the time reductions that could be obtained for the query, by using  $\mathcal{A}_1$  instead of  $\mathcal{A}_2$ . We show the 50<sup>th</sup>- and 75<sup>th</sup>-percentile and maximum time reductions in percentage.

We round percentage reductions down, thus percentage reductions of 99 mean speedups of *at least* 2 orders of magnitude are possible by switching algorithms. Note that even the weakest algorithm outperforms each of the others (excepting RBMRG), even if rarely. The final column in the table shows the number of workload queries where the row’s algorithm was the best (ignoring possible measurement error). We see that results are similar on the two workloads, and the superior algorithms are RBMRG (80 % of the queries), SCANCOUNT (15 %).

The final row represents the case where an oracle picks the fastest algorithm for each query. As expected, because RBMRG is best about 80 % of the time, the median of the percentage improvements is zero for this algorithm. The final row shows that *every* algorithm performs badly on at least one instance (e.g., RBMRG is beaten by 80 % once, which means that another algorithm is  $5\times$  faster). We see that SCANCOUNT, DSK, LOOPED, MGOPT and W2CTI are sometimes at least two orders of magnitude slower than necessary. At the 75<sup>th</sup> percentile level RBMRG is the clear winner (which is expected, given that it is best 80 % of the time). SCANCOUNT and BSTM are similar : each is typically about five times slower than the best algorithm. Although SCANCOUNT is the fastest algorithm at least 15 % of the time versus 0 % for BSTM, the comparison may not seem so lopsided when we consider that BSTM was clearly superior to SCANCOUNT more than 20 % of the time.

**Beating SCANCOUNT:** In § 4.1 we suggested that SCANCOUNT could be beaten; indeed, we can see this by inspecting the SCANCOUNT column in Tables VI and VII. To be more precise, our workloads contained a query that, compared to SCANCOUNT, was answered  $1100\times$  faster using RBMRG, another query that was also  $1100\times$  faster with LOOPED, one that was  $300\times$  faster with BSTM, one that was  $70\times$  faster using DSK, one that was  $34\times$  faster with W2CTI, and one where MGOPT was  $81\times$  faster. These extreme cases involve 3 datasets with long bitmaps ( $r$  is large) and queries involving a few especially sparse input bitmaps ( $N \leq 4$  and  $B$  is small)—conditions especially difficult for SCANCOUNT. At least in such cases, SCANCOUNT can be beaten by orders of magnitude.

Table VI. Percentage of competitions (Similarity workload) where the row’s algorithm was at least 20 % faster than the column’s algorithm, and beneath it, the percentage improvements from the row’s algorithm. We show the median, 75<sup>th</sup>-percentile, and maximum percentage improvement. An improvement of 99 % means at least 100× speed.

The final column shows the percentage of cases where the row’s algorithm was measured to be fastest.

vs	RBMRG	SCNCNT	LOOPED	DSK	w2CTI	BSTM	MGOPT	fastest
RBMRG		76 % 73 86 99	96 % 86 91 99	94 % 91 97 99	99 % 93 97 99	100 % 75 81 98	98 % 91 96 99	80 %
SCNCNT	12 % 56 69 80		73 % 72 88 99	77 % 86 90 96	90 % 83 87 97	58 % 52 68 96	82 % 85 90 98	15 %
LOOPED	2 % 30 46 66	19 % 62 82 99		54 % 74 90 99	62 % 65 83 99	17 % 38 68 96	51 % 69 89 99	3 %
DSK	1 % 17 39 64	17 % 57 74 92	38 % 72 86 99		31 % 32 66 99	24 % 50 71 94	22 % 21 40 96	3 %
w2CTI	0 % 13 17 30	8 % 60 75 92	27 % 54 72 99	36 % 25 36 69		14 % 35 51 89	29 % 22 36 90	0 %
BSTM	0 %	21 % 28 38 99	70 % 53 67 99	64 % 83 89 96	76 % 80 87 98		71 % 78 87 95	0 %
MGOPT	0 % 7 12 29	13 % 58 74 93	37 % 58 76 99	21 % 15 23 49	27 % 25 52 98	15 % 38 64 88		0 %
fastest	0 0 80	65 84 99	87 92 99	90 97 99	93 97 99	77 83 98	91 96 99	

Table VII. Results on the Many-Criteria workload, in the same format as Table VI.

vs	RBMRG	SCNCNT	LOOPED	DSK	w2CTI	BSTM	MGOPT	fastest
RBMRG		75 % 66 80 99	91 % 75 86 98	99 % 88 94 99	98 % 89 95 99	99 % 66 79 99	99 % 86 93 99	77 %
SCNCNT	9 % 20 29 48		56 % 72 84 96	82 % 85 89 96	95 % 80 83 87	63 % 51 59 75	82 % 82 86 95	18 %
LOOPED	3 % 26 51 76	34 % 58 76 98		65 % 70 89 99	70 % 71 87 99	33 % 33 58 93	61 % 64 86 99	5 %
DSK	0 % 5 13 16	12 % 57 77 98	19 % 35 53 89		24 % 49 76 99	9 % 37 54 87	7 % 15 26 65	0 %
w2CTI	0 % 16 30 57	4 % 47 68 97	16 % 33 50 83	50 % 30 42 77		2 % 36 55 86	23 % 16 30 70	1 %
BSTM	0 % 2 6 11	23 % 42 62 98	45 % 49 66 91	83 % 72 83 99	95 % 67 76 98		84 % 65 76 98	0 %
MGOPT	0 % 7 14 28	13 % 54 78 98	21 % 33 51 84	41 % 19 27 63	30 % 32 65 99	8 % 36 55 85		0 %
fastest	0 0 76	59 78 99	74 86 98	89 95 99	89 95 99	67 80 99	87 93 99	

### 7.6. Performance Across Workload Subsets

Table VIII shows the total time taken by each algorithm across both workloads, or across a portion of the workload(s) meeting certain criteria shown in the first column. (Since Tables VI and VII showed such similar results, we combine the two workload into an overall composite workload.) Table VIII shows the effect of large  $N$ , small  $T$  or  $N$ , the kind of query, or the dataset.

The table shows that LOOPED, DSK, MGOPT and w2CTI can have some extremely expensive queries, although fewer than 25 % of the queries are extremely expensive. The apparent preference for RBMRG toward the top of the table partly breaks down when we examine individual datasets at the bottom of the table. The large size of PGDVD-2gr and the excellent performance of RBMRG

Table VIII. Total time to process queries of various groups. The top line of each group is the total time. Then four lines give the 25<sup>th</sup>-, 50<sup>th</sup>-, 75<sup>th</sup>-percentile, and maximum query times. Values for RBMRG are absolute (seconds for total time; ms for percentile values). Values for all other algorithms are relative—the measured time has been normalized by dividing it by the corresponding time for RBMRG.

data	RBMRg	ScnCnt	Looped	DSk	w2CtI	BSTM	MgOpt
$N \leq 15$	$3.9 \times 10^0$	3.18	2.99	4.55	5.56	2.65	4.18
	0	4.3	1.5	3.8	5.4	2.1	3.3
	1	6.0	1.9	3.3	4.9	2.2	3.2
	3	4.0	2.4	4.0	5.3	2.5	4.0
	15	2.7	5.8	12.9	9.5	3.2	9.7
$N \geq 16$	$1.3 \times 10^3$	1.40	32.69	15.94	16.33	2.78	17.33
	2	3.3	5.7	13.3	11.7	5.2	12.5
	13	2.6	5.7	13.9	12.9	3.4	12.0
	102	0.9	5.1	6.3	5.6	2.7	6.1
	2484	1.0	827.7	25.7	31.1	5.4	24.6
$T < 5$	$3.2 \times 10^1$	2.11	1.32	27.54	17.74	3.62	27.87
	0	5.5	1.6	7.2	8.4	3.4	6.0
	1	6.2	1.9	6.8	7.3	2.5	6.0
	7	2.2	1.7	9.4	7.5	3.1	8.0
	619	3.1	1.0	103.3	103.3	6.9	98.8
Many Criteria	$2.7 \times 10^2$	<b>0.87</b>	6.33	7.63	5.17	2.16	6.36
	0	4.1	3.0	10.9	13.7	3.8	9.1
	4	2.5	4.0	13.8	11.0	3.3	11.1
	34	1.4	4.3	10.7	7.8	3.2	8.5
	745	0.7	19.6	10.6	5.3	2.1	8.9
IMDB-3gr	$1.0 \times 10^2$	<b>0.39</b>	5.48	2.01	2.69	1.95	2.69
	11	1.2	4.6	1.7	2.6	3.2	2.3
	21	1.2	5.9	2.8	3.6	3.1	3.1
	228	0.3	3.5	2.0	2.6	1.9	2.6
	615	0.2	9.0	1.8	2.4	1.3	2.3
PGDVD	$6.6 \times 10^1$	<b>0.43</b>	346.58	2.44	2.55	5.12	7.84
	1	7.8	2.4	3.3	3.1	2.6	3.2
	4	2.7	5.5	3.0	2.6	3.7	3.3
	22	0.9	11.6	2.8	2.4	4.9	3.8
	2484	0.3	827.7	3.5	6.2	5.4	16.4
PGDVD-2gr	$8.0 \times 10^2$	1.75	19.93	21.55	23.02	2.85	23.77
	550	2.5	8.0	20.7	28.6	3.4	25.9
	872	2.0	12.0	23.6	20.3	2.5	26.2
	1514	1.3	17.8	19.0	15.6	2.7	19.0
	2071	1.2	37.8	30.9	37.3	2.3	29.5
CensusIncome	$7.5 \times 10^1$	1.30	6.20	11.27	7.29	2.70	9.26
	3	2.7	2.1	9.0	10.9	3.4	7.9
	7	4.7	7.1	32.9	25.5	4.5	24.5
	29	1.7	5.2	17.8	11.3	3.9	13.4
	290	0.8	17.5	9.7	3.8	2.2	8.8
TWEED	$1.7 \times 10^0$	3.70	17.14	42.68	25.07	5.75	32.76
	0	7.9	3.8	21.7	37.9	7.4	21.6
	0	8.4	8.3	55.8	54.6	6.2	43.3
	1	3.8	9.0	41.7	28.1	5.8	36.2
	6	1.9	44.8	42.3	14.8	4.5	36.0
Weather	$2.3 \times 10^2$	<b>0.96</b>	6.28	7.58	5.74	2.23	6.40
	7	3.9	4.3	5.4	14.5	4.7	5.5
	21	2.5	3.7	12.8	12.1	2.8	11.6
	88	1.2	4.9	8.1	7.1	3.0	7.3
	745	0.7	19.6	10.6	5.3	2.1	8.9
All	$1.3 \times 10^3$	1.41	32.60	15.91	16.30	2.78	17.29
	1	4.0	4.4	13.0	17.4	3.9	12.4
	7	2.9	6.3	10.0	11.1	4.4	8.9
	60	1.1	4.8	8.9	6.5	3.1	8.6
	2484	1.0	827.7	25.7	31.1	5.4	24.6

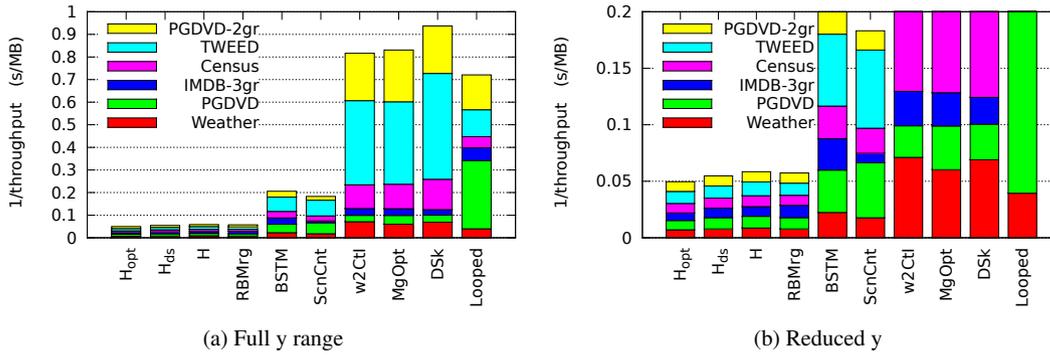


Figure 8. Aggregate throughput on each dataset. Bar height represents the number of seconds for a workload containing 1 MB of bitmap data from each dataset.  $H_{opt}$ ,  $H_{ds}$  and  $H$  are discussed in § 8.2.

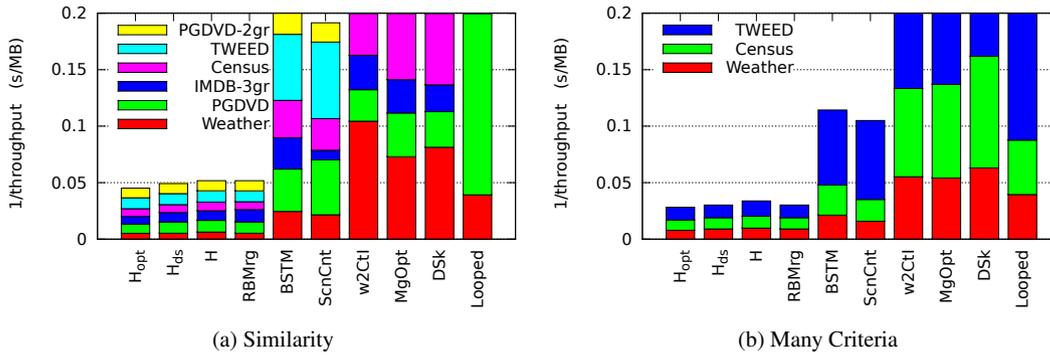


Figure 9. RBMRG excels in both workloads, but for the Similarity workload, SCANCOUNT does a better job on IMDB-3gr.

on this large dataset act together to dominate the overall results. Also, results are dominated by larger values of  $N$ , despite our generating workloads so that small- $N$  queries were more frequent than large- $N$  ones. For instance, in all the cases where SCANCOUNT did best overall (Many-Criteria queries, IMDB-3gr, PGDVD and Weather), note that RBMRG significantly outperformed SCANCOUNT at the median level. As well, costs were dominated by Similarity queries; while equal in number to Many-Criteria queries, they included the queries with the largest values of  $N$ .

To visualize or aggregate this data, we should consider that the workload involves datasets of widely different size: there are three orders of magnitude difference between the total volume of data for our TWEED queries and our PGDVD-2gr queries. Instead of merely timing the queries, we measure their throughput: amount of input data divided by the time necessary to complete the query, expressing the result in MB/s. Given an algorithm and a dataset, we use the harmonic mean to obtain an aggregate throughput value. However, for display purposes it is convenient to show the reciprocal throughput. For instance, the stacked bar charts in Fig. 8 can be viewed as representing times (in seconds) on some hypothetical workload in which 1 MB of bitmap data had been processed by the queries for each dataset. For our workload, RBMRG, SCANCOUNT and BSTM are strongly preferred to the others. Figure 9 shows that, for our relational datasets—the only ones that were used with both Many-Criteria and Similarity queries—RBMRG is the clear winner.

In several applications we can expect that  $N$  will not be particularly large, or that typical queries will usually have  $T \approx N$ . Figure 10 shows the results for these cases. On such queries, while RBMRG is best, BSTM is better than SCANCOUNT: for a typical query with  $N$  small, the cost of initializing and scanning the  $r$  counters is being amortized over a small volume of bitmap data. We also see that LOOPED is a viable algorithm for  $N \leq 16$ .

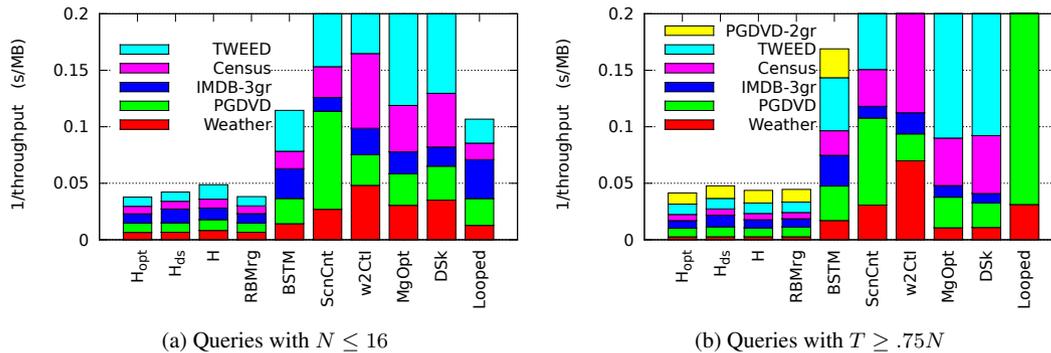


Figure 10. Normalized workload times for queries with small  $N$  and for queries with  $T \geq .75N$ .

Figure 10 also shows the situation for the workload queries where  $T \geq 0.75N$ . This situation is one where pruning-based algorithms such as w2CTI, MGOPT and DSK can excel. Indeed, we see them doing well on IMDB-3gr, Weather and (for w2CTI and DSK) PGDVD. However, there are other datasets where they still perform badly. Altogether, on workloads similar to ours, the benefits from pruning do not seem to be worth the risks of using these algorithms. Nevertheless, there are some applications where the requirement for a large  $T$  can be met. For instance, using the formula of Sarawagi and Kirpal [10] with strings of length  $N = 64$ , if we are interested in finding the strings of edit distance at most two from some target using trigrams, the appropriate threshold is  $T = 64 + 3 - 1 - 2 \times 3 = 60$ .

*7.6.1. The advantage of RBMRG.* Our results show that RBMRG was usually the fastest algorithm, especially over datasets coming from relational tables and for Many-Criteria queries. This speed advantage is due to fewer executed instructions, rather than cache effects: experiments showed that the processor executed about two instructions per cycle (IPC) for all implementations. (We saw 1.9 IPC for SCANCOUNT; 2.0 for w2CTI; 2.1 for BSTM, LOOPED and RBMRG; and 2.2 for MGOPT and DSK.)

One reason for the advantage of RBMRG is that it ends up solving a threshold problem over the dirty words, and our implementation adaptively switches between algorithms LOOPED and SCANCOUNT. In essence, it gets a benefit from RLE encoding, and then combines the strengths of two other efficient algorithms. An initial implementation had done a naïve computation (iterating over all bit positions) and this implementation of RBMRG was usually not competitive with SCANCOUNT or BSTM. Solving the threshold subproblem effectively on the dirty words was crucial, and our hybrid of SCANCOUNT and LOOPED made the revised implementation fast.

## 8. HYBRID ALGORITHMS

Our success in handling dirty words adaptively suggests that an adaptive, hybrid approach might also be a better way to solve threshold problems on compressed bitmaps.

### 8.1. An Execution-Time Model

To guide an adaptive algorithm, we need to estimate the running times of the more promising algorithms, in terms of the limited data that a DBMS might be expected to maintain.

Table IX shows estimates of the running-time functions over our workload. They were derived by least-squares fitting our running-time bounds in § 5–6 and Table III to the measured times for the competitions in the workload. To account for bitmap compression, we substitute EWAHSIZE where  $Nr/W$  occurred in Table III. Given a bound of  $O(f(x_1, x_2, \dots, x_k))$ , we modeled the running time as  $cf(x_1, x_2, \dots, x_k)$  and fitted  $c$  according to the measured running times. Algorithm SCANCOUNT

Table IX. Running time estimates for good algorithms (this excludes w2CTI, MGOPT and DSK).

Algorithm	Time complexity estimate	Fitted coefficients
SCANCOUNT	$c_{sc,1} \times r + c_{sc,2} \times B$	$c_{sc,1} = 2.072 \times 10^{-5} \pm 7.6 \times 10^{-7}$ $c_{sc,2} = 2.683 \times 10^{-6} \pm 6.1 \times 10^{-9}$
LOOPED	$c_{LOOPEd} \times T \times \text{EWAHSIZE}$	$c_{LOOPEd} = 1.306 \times 10^{-6} \pm 2.9 \times 10^{-9}$
BSTM	$c_{BSTM} \times \text{EWAHSIZE} \times \ln N$	$c_{BSTM} = 3.133 \times 10^{-5} \pm 1.6 \times 10^{-7}$
RBMrg	$c_{RBMRg} \times \text{EWAHSIZE} \times \ln N$	$c_{RBMRg} = 1.592 \times 10^{-6} \pm 5.3 \times 10^{-9}$

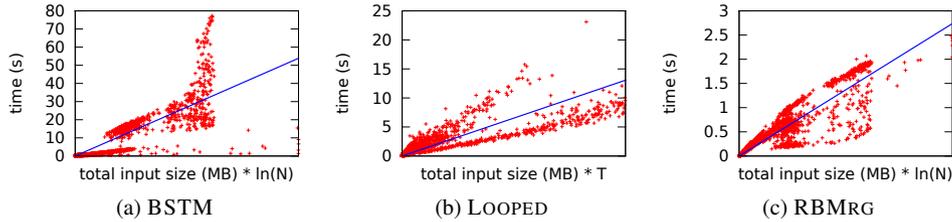


Figure 11. The running times for BSTM and RBMRg depend on the total compressed size of the input bitmaps and a  $\log N$  factor. The running time for LOOPEd depends on the total compressed size and  $T$ . We show least-squares lines (passing through the origin) to fit these models.

had two independent terms and we used a separate constant for each term. Also, for RBMRg, we felt it would be unreasonable to expect the `RUNCOUNT` of the bitmaps to be cataloged. Instead, we used `EWAHSIZE` as a proxy.

Our time-complexity estimates for BSTM, LOOPEd and RBMRg are shown against the actual data in Fig. 11. We see the fits are not particularly good, but we seldom underestimate running times by more than a factor of 2. (Our overestimates are frequently off by larger factors.) This may be good enough to avoid selecting an algorithm that is badly suited for a query.

## 8.2. Algorithms

We experimented with hybrid algorithms  $H$  and  $H_{ds}$ , described below. For comparison purposes,  $H_{opt}$  is the hybrid algorithm that always chooses the fastest algorithm for any query.

**Hybrid algorithm,  $H$ .** Since we have multiple alternative algorithms for the same problem, there are sophisticated approaches for choosing the best algorithm for a given application [42]. However, we can get reasonably good performance with two simple approaches for choosing the appropriate algorithm.

Our first approach, hybrid algorithm  $H$ , evaluates the running-time estimates given in Table IX. It then selects the algorithm predicted to run fastest. Mathematically, the estimate for BSTM is about 20 times larger than the estimate for RBMRg, so we should never choose BSTM. Algebraic manipulation of the time estimates for RBMRg and LOOPEd shows that when  $T < \frac{c_{RBMRg}}{c_{LOOPEd}} \ln N$  we should choose LOOPEd in preference to RBMRg. Conveniently, one does not need to know `EWAHSIZE`.

A weakness of this approach is that it is based explicitly on the performance of our particular test computers. While slightly inaccurate estimates may not lead to bad decisions, those using this approach on systems that differ significantly should conduct their own benchmarks and adjust the coefficients.

**Adjust-by-dataset algorithm,  $H_{ds}$ .** Faced with a collection of queries over disparate datasets, an obvious approach is to select the algorithm entirely on the basis of the dataset. Perhaps some initial profile runs would be used to select the algorithm to be used consistently on a dataset. This  $H_{ds}$

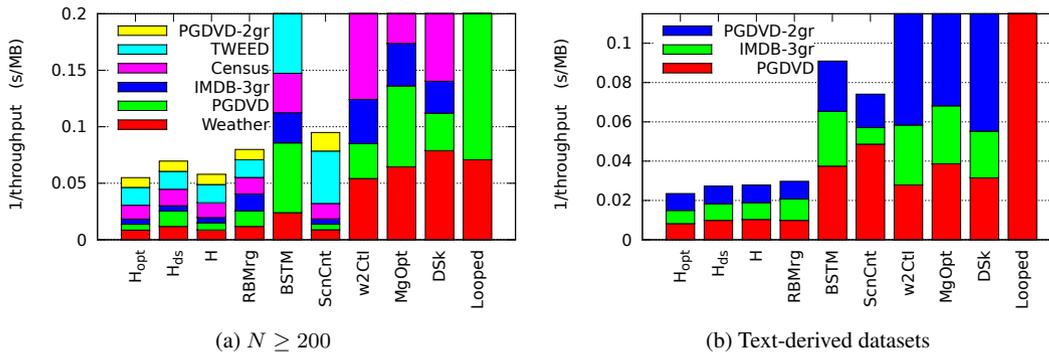


Figure 12. Some cases when it is best to mix RBMRG and SCANCOUNT.

approach was tested; on our workloads we used SCANCOUNT for all queries against IMDB-3gr, and RBMRG for all other queries. (We chose this combination by inspecting Fig. 8.)

**Optimal hybrid algorithm,  $H_{opt}$ .** For comparison purposes, we can determine the effect of the optimal hybrid algorithm,  $H_{opt}$ , which always selects the best algorithm for any competition as an oracle would. Since we have already run every algorithm during the competition, this is easy for us to do. Of course, in practice one would not have this information— $H_{opt}$  exists only to make comparisons.

### 8.3. Evaluation of Hybrid Algorithms

As Figs. 8–10 show, sometimes there is little reason to rely on anything other than RBMRG, and choosing otherwise only hurts performance. Yet Fig. 12 shows two other cases where hybrid approaches helped. On queries with large  $N$ , H had a 28% improvement over RBMRG. ( $H_{opt}$  and  $H_{ds}$  had respective improvements over RBMRG of 31% and 13%.) For our text-derived datasets the improvements were smaller: 4%, 9% and 23% respectively for H,  $H_{ds}$  and  $H_{opt}$ . Looking at Figs. 9b and 10a we see cases where H and  $H_{ds}$  incorrectly chose to use algorithms other than RBMRG, leading to slightly worse results.

Algorithm RBMRG requires a detailed knowledge of the internal workings of a RLE-compressed bitmap representation and is best added by the maintainers of a compressed bitmap package. Thus, it is reasonable to look at the tradeoffs that come from hybrid algorithms that omit RBMRG. Our comparison corresponds to bar heights in Fig. 12, and the time for H increased by 66% for the  $N \geq 200$  case and 154% for the text-derived datasets. Excluding RBMRG, the best non-hybrid algorithm was SCANCOUNT. When H could not choose RBMRG, its result was 2% worse than SCANCOUNT on the text-derived datasets but 6% better for the  $N \geq 200$  cases. Note that hybrid algorithms can do much better: if  $H_{opt}$  chooses between BSTM, SCANCOUNT and LOOPED, on the text-derived datasets we can get a result 56% better than SCANCOUNT (and only 13% worse than RBMRG).

## 9. CONCLUSION AND FUTURE WORK

We reviewed several novel and several known algorithms for computing thresholds. We found that a novel algorithm (RBMRG) was generally superior to alternatives, sometimes being orders of magnitude faster.

Although RBMRG could be considered the overall winner, each algorithm examined was weak in some circumstances. However, we combine them in a hybrid algorithm that improves on any

individual algorithm. In future work, we might create better hybrid algorithms, perhaps by applying machine-learning processes to choose the fastest threshold algorithm [43, 44].

Our work has considered  $N$  values up to a few thousand (at most 11 115). Yet datasets whose indexes have millions of bitmaps are not out of the question. Would there be applications where a threshold computation with  $N = 1\,000\,000$  would be useful? If so, which algorithms should be used? Can new algorithms be developed for this case?

When possible, data should be indexed in sorted order [13]: this can improve compression and processing speed. Some algorithms might benefit more than others from sorting, and this warrants further investigation.

Finally, algorithms can be parallelized, and while most of our threshold computations take only a few milliseconds, some take tens of seconds. If we try extremely large  $N$  values, this may increase. At some point, it may become important to have one threshold computation run faster than is possible using a single core. For multicore processing, a particular challenge with current architectures is that all cores compete for access to L3 and RAM. E.g., this means that it is best if intermediate results fit in L2 cache. It might be advisable to partition the problems.

#### REFERENCES

1. Antoshenkov G. Byte-aligned bitmap compression. *Data Compression Conference, DCC'95*, IEEE Computer Society: Washington, DC, USA, 1995; 476, doi:10.1109/DCC.1995.515586.
2. Culpepper JS, Moffat A. Efficient set intersection for inverted indexing. *ACM Transactions on Information Systems* Dec 2010; **29**(1):1:1–1:25, doi:10.1145/1877766.1877767.
3. Stonebraker M, Abadi DJ, Batkin A, Chen X, Cherniack M, Ferreira M, Lau E, Lin A, Madden S, O’Neil E, et al.. C-Store: a column-oriented DBMS. *Proceedings of the 31st International Conference on Very Large Data Bases, VLDB'05*, ACM: New York, NY, USA, 2005; 553–564.
4. Thomsen C, Pedersen TB. A survey of open source tools for business intelligence. *Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches*, Taniar D, Chen L (eds.). chap. 10, IGI Global: Hershey, PA, USA, 2011; 237–257, doi:10.4018/978-1-60960-537-7.ch010.
5. Yang F, Tschetter E, Léauté X, Ray N, Merlino G, Ganguli D. Druid: a real-time analytical data store. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, ACM: New York, NY, USA, 2014; 157–168, doi:http://doi.acm.org/10.1145/2588555.2595631.
6. MacNicol R, French B. Sybase IQ Multiplex — designed for analytics. *Proceedings of the 30th International Conference on Very Large Data Bases, VLDB'04*, VLDB Endowment, 2004; 1227–1230, doi:10.1016/B978-012088469-8.50111-X.
7. O’Neil P, Graefe G. Multi-table joins through bitmapped join indices. *SIGMOD Record* Sep 1995; **24**(3):8–11, doi:10.1145/211990.212001.
8. Rinfret D, O’Neil P, O’Neil E. Bit-sliced index arithmetic. *Proceedings of the 2001 ACM SIGMOD international conference on Management of Data*, ACM: New York, NY, USA, 2001; 47–57, doi:http://doi.acm.org/10.1145/375663.375669.
9. Kaser O, Lemire D. Threshold and symmetric functions over bitmaps. *Technical Report TR-14-001*, Dept. of CSAS, University of New Brunswick 2014. ArXiv:1402.4073 [cs.DB].
10. Li C, Lu J, Lu Y. Efficient merging and filtering algorithms for approximate string searches. *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE'08*, IEEE Computer Society: Washington, DC, USA, 2008; 257–266, doi:10.1109/ICDE.2008.4497434.
11. Sarawagi S, Kirpal A. Efficient set joins on similarity predicates. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, SIGMOD'04*, ACM: New York, NY, USA, 2004; 743–754, doi:10.1145/1007568.1007652.
12. Barbay J, Kenyon C. Deterministic algorithm for the t-threshold set problem. *Algorithms and Computation, Lecture Notes in Computer Science*, vol. 2906, Ibaraki T, Katoh N, Ono H (eds.). Springer Berlin Heidelberg, 2003; 575–584, doi:10.1007/978-3-540-24587-2\_59.
13. O’Neil P, Quass D. Improved query performance with variant indexes. *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, 1997; 38–49.
14. Lemire D, Kaser O, Aouiche K. Sorting improves word-aligned bitmap indexes. *Data & Knowledge Engineering* 2010; **69**(1):3–28, doi:10.1016/j.datak.2009.08.006.
15. Li M, Jia L, You J, Xi J, Qin H, Zeng R. Fast T-overlap query algorithms using graphics processor units and its applications in web data query. *World Wide Web* 2013; :1–17,doi:10.1007/s11280-013-0232-6.
16. Behm A, Ji S, Li C, Lu J. Space-constrained gram-based indexing for efficient approximate string search. *Proceedings IEEE 25th International Conference on Data Engineering, ICDE'09*, IEEE, 2009; 604–615, doi:10.1109/ICDE.2009.32.
17. Jia L, Xi J, Li M, Liu Y, Miao D. ETI: an efficient index for set similarity queries. *Frontiers of Computer Science* 2012; **6**(6):700–712, doi:10.1007/s11704-012-1237-5.
18. Barbay J, Kenyon C. Adaptive intersection and t-threshold problems. *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'02*, Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2002; 390–399.

19. O’Neil PE. Model 204 architecture and performance. *2nd International Workshop on High Performance Transaction Systems*, Springer-Verlag: London, UK, 1989; 40–59.
20. Sharma V. Bitmap index vs. B-tree index: Which and when? online: <http://www.oracle.com/technetwork/articles/sharma-indexes-093638.html> [April 25th, 2014] March 2005.
21. Lemire D, Kaser O. Reordering columns for smaller indexes. *Information Sciences* June 2011; **181**(12):2550–2570, doi:10.1016/j.ins.2011.02.002.
22. Navarro G, Provedel E. Fast, small, simple rank/select on bitmaps. *Experimental Algorithms, Lecture Notes in Computer Science*, vol. 7276, Klasing R (ed.). Springer Berlin Heidelberg, 2012; 295–306, doi:10.1007/978-3-642-30850-5\_26.
23. Wu K, Stockinger K, Shoshani A. Breaking the curse of cardinality on bitmap indexes. *Scientific and Statistical Database Management, Lecture Notes in Computer Science*, vol. 5069, Ludäscher B, Mamouli N (eds.). Springer Berlin Heidelberg, 2008; 348–365, doi:10.1007/978-3-540-69497-7\_23.
24. Colantonio A, Di Pietro R. Concise: Compressed ‘n’ composable integer set. *Information Processing Letters* Jul 2010; **110**(16):644–650, doi:10.1016/j.ipl.2010.05.018.
25. Deliège F, Pedersen TB. Position list word aligned hybrid: optimizing space and performance for compressed bitmaps. *Proceedings of the 13th International Conference on Extending Database Technology (EDBT’10)*, ACM: New York, NY, USA, 2010; 228–239, doi:10.1145/1739041.1739071. URL <http://doi.acm.org/10.1145/1739041.1739071>.
26. Fusco F, Stoecklin MP, Vlachos M. NET-FLI: On-the-fly compression, archiving and indexing of streaming network traffic. *Proceedings of the VLDB Endowment* 2010; **3**:1382–1393.
27. Guzun G, Canahuate G, Chiu D, Sawin J. A tunable compression framework for bitmap indices. *Proceedings IEEE 30th International Conference on Data Engineering, ICDE’14*, IEEE, 2014.
28. Lemire D, Moon C, McIntosh D, Becho R, Ranger C, Zenz V, Kaser O. JavaEWAH - GitHub page. online: <https://github.com/lemire/javaewah> [25 April 2014].
29. Wu K, Otoo E, Shoshani A. On the performance of bitmap indices for high cardinality attributes. *Proceedings of the 30th International Conference on Very Large Data Bases, VLDB’04*, Morgan Kaufmann, 2004; 24–35, doi: 10.1016/B978-012088469-8.50006-1.
30. Knuth DE. *Combinatorial Algorithms, Part 1, The Art of Computer Programming*, vol. 4A. Addison-Wesley: Boston, Massachusetts, 2011.
31. Tellez ES, Chávez E, Navarro G. Succinct nearest neighbor search. *Information Systems* 2013; **38**(7):1019–1030, doi:10.1016/j.is.2012.06.005.
32. Critchley A. Finding similar rows in SQL. *Code Project (Webzine)*. CodeProject.com, 2013. Online: <http://www.codeproject.com/Articles/610103/FindingplusSimilarplusRowsplusInplusSQL> [25 April 2014].
33. Ferro A, Giugno R, Puglisi PL, Pulvirenti A. An efficient duplicate record detection using q-grams array inverted index. *Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science*, vol. 6263, Bach Pedersen T, Mohania M, Tjoa A (eds.). Springer Berlin Heidelberg, 2010; 309–323, doi:10.1007/978-3-642-15105-7\_25.
34. Montanari D, Puglisi PL. Near duplicate document detection for large information flows. *Multidisciplinary Research and Practice for Information Systems, Lecture Notes in Computer Science*, vol. 7465, Quirchmayr G, Basl J, You I, Xu L, Weippl E (eds.). Springer Berlin Heidelberg, 2012; 203–217, doi:10.1007/978-3-642-32498-7\_16.
35. Perry SA, Willett P. A review of the use of inverted files for best match searching in information retrieval systems. *Journal of Information Science* 1983; **6**(2-3):59–66, doi:10.1177/016555158300600204.
36. Ellingsen E. Bit tricks, part III: Fast vertical counter. online:<http://www.steike.com/code/bits/vertical-counter/> 2009. Last checked 2014-02-17.
37. Project Gutenberg Literary Archive Foundation. July 2006 Gutenberg DVD. online: [http://www.gutenberg.org/wiki/Gutenberg:The\\_CD\\_and\\_DVD\\_Project](http://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project) [25 April 2014].
38. Frank A, Asuncion A. UCI machine learning repository. online: <http://archive.ics.uci.edu/ml> [25 April 2014].
39. Engine JO. Five decades of terrorism in Europe: The TWEED dataset. *Journal of Peace Research* 2007; **44**(1):109–121, doi:10.1177/0022343307071497.
40. Webb H, Lemire D, Kaser O. Diamond dicing. *Data & Knowledge Engineering* 2013; **86**:1–18, doi:10.1016/j.datak.2013.01.001.
41. Hahn CJ, Warren SG, London J. Edited synoptic cloud reports from ships and land stations over the globe, 1982–1991. online: <ftp://cdiac.ornl.gov/pub2/ndp026b/> [25 April 2014].
42. Beyer K, Ramakrishnan R. Bottom-up computation of sparse and iceberg CUBEs. *SIGMOD Record* 1999; **28**(2):359–370, doi:10.1145/304181.304214.
43. Hoos HH. Programming by optimization. *Communications of the ACM* Feb 2012; **55**(2):70–80, doi:10.1145/2076450.2076469.
44. Lagoudakis MG, Littman ML. Algorithm selection using reinforcement learning. *Proceedings of the 17th International Conference on Machine Learning, ICML’00*, Morgan Kaufmann: San Francisco, CA, USA, 2000; 511–518.
45. Horvitz EJ. Reasoning under varying and uncertain resource constraints. *Proceedings of the 7th National Conference on Artificial Intelligence, AAAI’88*, The MIT Press: Cambridge, MA, USA, 1988; 111–116.