

Range Retrieval with Graph-Based Indices

Magdalen Dobson Manohar
 Carnegie Mellon University and
 Microsoft Azure
 USA
 mmanohar@microsoft.com

Taekseung Kim
 Carnegie Mellon University
 USA
 taekseuk@andrew.cmu.edu

Guy E. Blelloch
 Carnegie Mellon University and
 Google
 USA
 guyb@cs.cmu.edu

Abstract

Retrieving points based on proximity in a high-dimensional vector space is a crucial step in information retrieval applications. The approximate nearest neighbor search (ANNS) problem, which identifies the k nearest neighbors for a query, has been extensively studied in recent years. However, comparatively little attention has been paid to the related problem of finding all points within a given distance of a query, the *range retrieval* problem, despite its applications in areas such as duplicate detection, plagiarism checking, and facial recognition. In this paper, we present new techniques for range retrieval on graph-based vector indices, which are known to achieve excellent performance on ANNS queries. Since a range query may have anywhere from no matching results to thousands of matching results in the database, we introduce a set of range retrieval algorithms based on modifications of the standard graph search that adapt to terminate quickly on queries in the former group, and to put more resources into finding results for the latter group. Due to the lack of existing benchmarks for range retrieval, we also undertake a comprehensive study of range characteristics of existing embedding datasets, and select a suitable range retrieval radius for eight existing datasets with up to 1 billion points in addition to one existing benchmark. We test our algorithms on these datasets, and find up to 100x improvement in query throughput over a standard graph search and the FAISS-IVF range search algorithm. We also find up to 10x improvement over a previously suggested modification of the standard beam search, and strong performance up to 1 billion data points.

1 Introduction

Similarity search within databases of high-dimensional vectors has become increasingly important over the last decade due to the rise of semantic embeddings generated by neural networks and large language models (LLMs). Similarity search is a foundational building block of applications such as search, recommendations, advertising, and retrieval augmented generation (RAG). Correspondingly, there has been an explosion of work on efficient algorithms for approximate similarity search [16, 19, 27], as exact similarity search is prohibitively expensive due to the curse of dimensionality.

The object of approximate similarity search, or approximate nearest neighbor search (ANNS) is finding the top- k most similar embeddings. These top- k results can then be post-processed for the desired application. However, there are some applications for which retrieving top- k embeddings for a fixed k is a poor fit. Applications such as duplicate detection, plagiarism checking, and facial recognition require instead to retrieve all results within a certain *radius* of a query, with some post-processing after to verify whether a match exists out of the retrieved items [10, 22, 24]. Furthermore, in real-world search applications some queries may have tens of

thousands of matches while others have none or very few. In these cases, search with a small radius, or range retrieval, may be used to both differentiate these query types [28]. Range search can also be a useful subroutine in applications using nearest neighbor graphs, such as clustering or graph learning [14, 18].

Range search differs from top- k search by the diversity in size of the ground truth solutions for a set of queries. In practice, ground truth for a set of queries follows a skewed, Pareto-like distribution where the majority of queries tend to have no results within the chosen radius, while a smaller fraction have a small number of results, and a few outliers have thousands to tens of thousands of results. Queries in the middle category are well served by existing similarity search algorithms, which are already highly optimized and efficient. Thus, a good range search algorithm will match the efficiency of top- k search on the middle category of queries, while quickly terminating on queries with no results, and efficiently finding results for those few queries with thousands of results.

Despite the many applications of range search, there is very little work on designing algorithms specifically for the task of high-dimensional range search. From a practical perspective, it would be ideal if indices for top- k search could be reused or lightly adapted for range search, enabling an existing index to serve both types of queries. Indices for top- k search typically fall into one of two categories—partition-based indices, which partition the data set into cells and exhaustively search a small number of cells at query time [15, 16], and graph-based indices, which construct a proximity graph over the data points and use a variant of greedy search to answer queries [12, 19, 27, 30]. Graph-based indices are widely acknowledged as achieving equal or better similarity search performance than partition-based indices [31]. However, there is almost no work has studied the question of adapting graph-based indices for range retrieval. Furthermore, top- k search with IVF indices naturally explores thousands of candidates per query by exhaustively checking the distances between the query point and all the points in each cell, while the number of nodes a graph-based search typically explores is only a small multiple of k . It is thus a much less trivial algorithmic task to effectively adapt graph-based search to serve such queries (as we will show in Section 4, naive adaptations of existing search methods are not adequate for the task). The graph structure also suggests potential to terminate queries with no results after examining just a handful of vertices. It is natural then to ask whether graph-based indices can be adapted to efficiently serve range queries.

Our Contributions. In this paper, we present a set of algorithms designed for approximate high-dimensional range search on graph-based ANNS indices. Our techniques are modifications of the standard graph search algorithm, meaning that we enable one data

structure to efficiently serve both top- k queries and range queries. We devise an early-stopping heuristic that is capable of predicting whether a query has no results after only a few hops in the graph search. For queries with many results, we design an algorithm that extends the search path to return more results while minimizing wasted work, and compare and contrast situations where each algorithm dominates the other.

For our experimental evaluation, since there is only one publicly released dataset specifically for range search, we evaluate eight state-of-the-art publicly available metric embedding datasets with up to 1 billion points and select a suitable radius for each one. We also perform additional analysis of the characteristics of different range search datasets.

We evaluate our algorithms on the set of nine datasets (eight contributed by us, and one public benchmark) against both a simple adaptation of top- k search, an in-memory variant of the range search algorithm used in the Starling system [30] which we refer to as *doubling search*, and the range search algorithm from FAISS-IVF. We find that our range retrieval algorithms are capable of up to 100x speedup over the top- k and FAISS-IVF baselines, and up to 10x speedup over the doubling search. We additionally find that the speedup and scalability of our algorithms extends up to datasets with 1 billion points. Figure 1 shows a preview of our experimental results on three embedding datasets of sizes one million to 1 billion points.

Outline. Sections 1.1 and 2 cover related work and preliminaries, respectively. In Section 3 we explore the characteristics of range search datasets and provide heuristics for choosing an acceptable radius, which we use to adapt eight public embedding datasets for range searching. In Section 4 we present our algorithms for range search on graph-based indices, and in Section 5 we present experimental results.

1.1 Related Work

Work on Range Searching. Some earlier work addresses range retrieval in high dimensions from a theoretical standpoint [6], or with only minimal experiments that do not extend to the large embeddings used today [29]. Theory results on nearest neighbor search using locality sensitive hashing (LSH) also apply to range search, since their approximation guarantees are of the form of finding neighbors within a $(1 + \epsilon)$ radius of the distance to the true top- k result [15].

Meta AI released a range retrieval dataset aimed towards detecting misinformation [22], which was used as one of the competition datasets in the NeurIPS 2021 Big ANN Benchmarks Challenge [26], but competitors solved this task using naive adaptations of top- k search rather than novel range searching algorithms. A recent work by Szilvassy, Mazare, and Douze [28] addresses the combined task of retrieving range search results with one metric and then re-ranking them using a more sophisticated model. They design a range search metric specifically for *bulk* search—that is, range search over a group of queries with a fixed search budget—and benchmark its results on an image search application. They use an inverted file (IVF) index, a common data structure for similarity search [9], to

serve their database queries. The well-known FAISS-IVF library includes an option for range search on an inverted file index [9].

The Starling system [30] is an SSD-resident graph-based ANNS index optimized for I/O efficiency. Their work includes benchmarks on range search metrics, but the focus of their paper is not on comprehensive benchmarking of range search, and since their implementation is a disk-resident index and ours is fully in-memory, our experiments are not directly comparable to theirs. In Section 4 we describe the range search algorithm used in the Starling system, and how we adapted it for the in-memory case and made some small experimental improvements before using it as a baseline.

Other Related Work. Some variants on the high-level ideas behind our range search algorithms—that is, early termination of queries with few results, and extensions of beam search for queries with many results—have been used before in the context of nearest neighbor search. Two works [1, 17] use machine learning models to predict when a query can terminate early and still maintain high accuracy. Another work [32] uses graph search with two phases (essentially, an initial cheaper phase and a later, more compute intensive phase to gather additional candidates).

Not to be confused with range search, the *range filtering* problem addresses vectors with numerical labels such as timestamps, and asks to retrieve all vectors within a certain range on the labels (e.g., finding all vectors added to the database within a certain time range). This problem, which deals with external metadata on the vectors rather than finding all vectors within a ball of radius r , has been the subject of some recent work [11, 33, 35].

2 Preliminaries

In this work, we study a set $\mathcal{P} \subseteq \mathbb{R}^d$ of n points (vectors) in d dimensions. We denote the *distance* between two points $p, q \in \mathbb{R}^d$ as $d(p, q)$. Smaller distance indicates greater similarity.

Commonly-used distance functions include Euclidean distance (L_2 norm), and maximum inner product search, which uses negative inner product as a distance function [20].

Definition 2.1 (Range Search). Given a set of points \mathcal{P} in d -dimensions, a query point q and radius r , the range search problem finds a set $\mathcal{K}_q \subseteq \mathcal{P}$ such that $\max_{p \in \mathcal{K}_q} d(p, q) \leq r$.

Next we define average precision, a commonly used metric that captures the accuracy of a search algorithm. Note that the measure weights a point’s contribution to average precision proportionately to its size.

Definition 2.2 (Average Precision). Let \mathcal{P} be a set of points in d -dimensions and Q a query point set. For $q \in Q$, let \mathcal{K}_q be the true range neighbors of q in \mathcal{P} . Let $\mathcal{K}'_q \subset \mathcal{P}$ be an output of a range search algorithm, which is only allowed to consist of results that are truly within radius r of the query point. Then **Average Precision** is defined as
$$\frac{\sum_{q \in Q} |\mathcal{K}_q \cap \mathcal{K}'_q|}{\sum_{q \in Q} |\mathcal{K}_q|}$$

Throughout this work, we will use the notation $N_{out \in G}(p)$ to refer to the set of directed edges $(p, v) | v \in G$, where G is a directed graph.

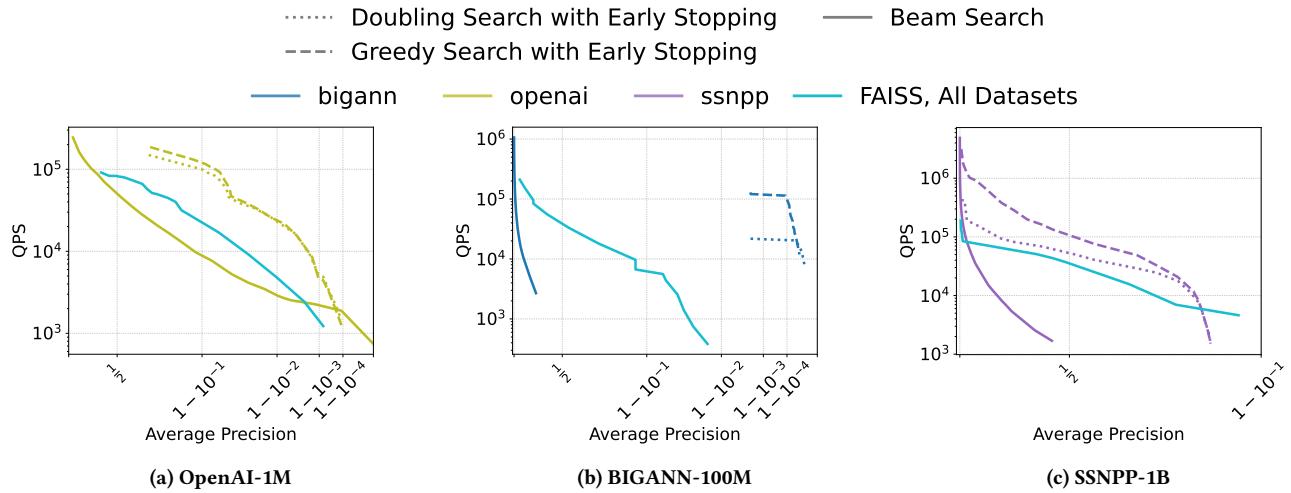


Figure 1: A preview of our experimental results on three datasets. The solid line shows the beam search baseline, while the dotted and dashed lines show our two new algorithms. Datasets and algorithms are described in detail in Sections 3 and 4, respectively.

3 Characteristics of Range Search Datasets

In this section, due to the fact that there is only one existing high-dimensional range search benchmark, we take eight existing embedding datasets from top- k retrieval benchmarks and show how to use them for range search by finding a suitable radius. We evaluate the characteristics of each dataset; alongside, we also evaluate the one existing range search benchmark.

For a dataset to be a meaningful range search benchmark, the same distance ϵ must consistently indicate a match for each query point. This is not always true for embedding datasets and is difficult to verify without full access to the source material. Thus, we assess existing benchmarks to determine suitability and select a radius accordingly. Real-world query sets typically follow a power-law distribution, where most points have no matches, but a few have thousands [22, 28]. Another key factor is whether a dataset has a “robust” radius—small perturbations should not drastically alter match distribution. Our experiments show that more robust datasets tend to outperform those sensitive to radius changes. See Table 1 for a description of each dataset used in our benchmarks.

3.1 Density of Matches by Dataset

We begin our investigation by computing a measure of density of each dataset, in order to both find a suitable radius and determine whether the choice of radius is “robust” in the sense that small changes to the radius do not result in drastically more points in each ball around a query point. To do this, we take each dataset and vary the radius, recording what we refer to as the “percent captured” for each radius—that is, for each query point q and radius r , what fraction of the dataset is captured in the ball of radius r around q ?

Figure 2 presents data for each dataset, separately plotting Euclidean and inner product results from the largest radius with 0% capture to the smallest achieving 100% capture. As shown, dataset responses to radius perturbations vary widely. BIGANN, GIST, DEEP,

Wikipedia, and MSMARCOWebSearch are the most robust in the relevant range (around 10^{-6} to 10^{-5} for million-size datasets, where most query points yield no results). In these datasets, fewer points lie near the boundary of radius r , reducing the effort needed to distinguish between those just inside or outside. Experimental results in Section 5 illustrate this effect. Using insights from this analysis, we select an appropriate radius for each dataset, with real (non-normalized) values shown in Figure 1

3.2 Frequency Distribution of Matches

Given a choice of radius for each dataset, we now evaluate the frequency distribution of the number of range results for each query point. The full distributions are found in Table 2. We observe a general pattern of most queries having no results and a few large outliers, which is more pronounced in BIGANN, DEEP, MSTuring, and Text2Image, and less pronounced in SSNPP, OpenAI, Wikipedia, and MSMARCOWebSearch. GIST is unusual among the datasets in having many extremely large outliers. For BIGANN, SSNPP, and Wikipedia, we also show the frequency distribution for a larger version of the same base dataset. Since these are effectively larger samples from the same distribution, we observe an increase in density of range results.

4 Range Search on Graph-Based ANNS Indices

This section introduces our range search algorithms, starting with the data structure and baseline beam search. Since beam search performs well for top- k queries, some range queries with an intermediate number of results can be efficiently handled using a small k . Thus, we focus on improving queries at the extremes—those with many results and those with none. Intuitively, both require dynamic adjustments: terminating early for queries with no results and dynamically extending the beam for those with many.

Dataset	Metric	Dimension	Query Size	Selected Radius	Source	Embedding Model
BIGANN	L2	128 (uint8)	10000	10000	Images [16]	SIFT descriptors
DEEP	L2	96 (float)	10000	.02	Images [5]	The last fully-connected layer of the GoogLeNet model
MSTuring	L2	100 (float)	100000	.3	Bing queries [34]	Turing AGI v5
GIST	L2	968 (float)	10000	.5	Images [16]	GIST descriptors
SSNPP	L2	200 (uint8)	100000	96237	Images [22]	SimSearchNet++
OpenAI	L2	1536 (float)	10000	.2	ArXiv articles [13]	OpenAI text-embedding-ada-002 model
Text2Image	IP	200 (float)	100000	-.6	Images [5]	Se-ResNext-101
Wikipedia	IP	768 (float)	5000	-10.5	Wikipedia articles [23]	cohere.ai multilingual 22-12
MSMARCO-WebSearch	IP	768 (float)	9374	-62	Clicked document-query pairs from the ClueWeb22 website corpus [7]	SimANS

Table 1: Descriptions of each dataset used in our experiments, along with the radius used for range search. Every dataset is publicly available [3, 25]. The SSNPP dataset is a range search benchmark, while the other datasets were originally published as top- k benchmarks. Unless otherwise indicated, the reference in the “Source” column covers all attributes of the dataset.

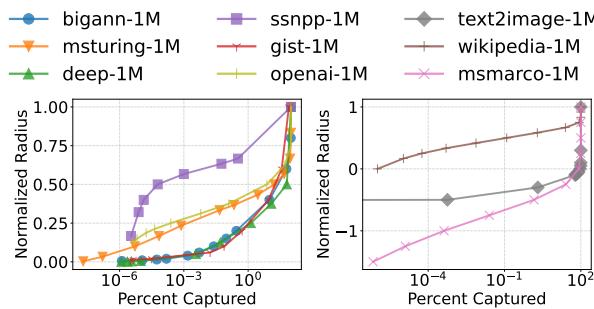


Figure 2: Plots of radius versus percent captured for each dataset: Euclidean datasets on the left and inner product datasets on the right. Radius is normalized to enable displaying multiple datasets on one plot. Note that inner product values can be less than zero, so negative ranges of radius are present on the right-hand figure.

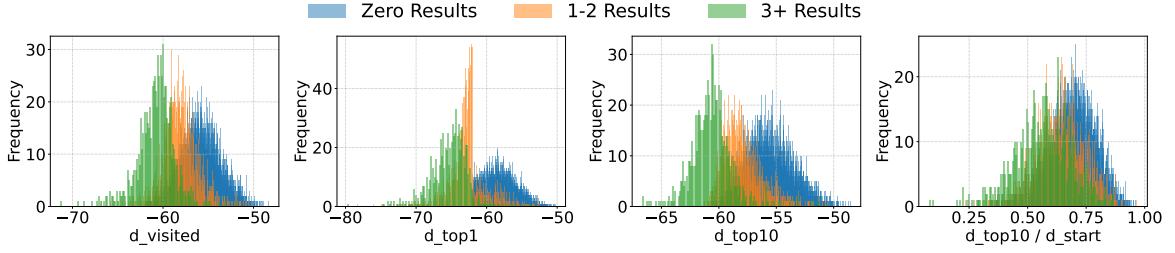
Dataset	0	$\leq 10^1$	$\leq 10^2$	$\leq 10^3$	$\leq 10^4$	$\leq 10^5$
BIGANN-1M	9728	143	84	45	0	0
BIGANN-10M	9590	173	99	93	45	0
BIGANN-100M	9413	212	136	101	93	45
DEEP-1M	9923	56	19	2	0	0
MSTuring-1M	95716	2443	20	21	0	0
GIST-1M	8487	830	160	143	134	246
SSNPP-1M	97422	2424	254	0	0	0
SSNPP-10M	91575	7310	954	161	0	0
SSNPP-100M	80795	13719	4357	971	158	0
OpenAI-1M	7030	2564	372	34	0	0
Text2Image-1M	99327	669	4	0	0	0
Wikipedia-1M	4445	482	73	0	0	0
Wikipedia-10M	3385	1328	277	10	0	0
MSMARCO-WebSearch-1M	7022	2199	152	3	0	0

Table 2: Table showing the distribution of result sizes for each dataset. Note that not all datasets have the same number of query points; see Figure 1 for the number of queries and corresponding radius for each dataset. No data point had more than 10^5 results.

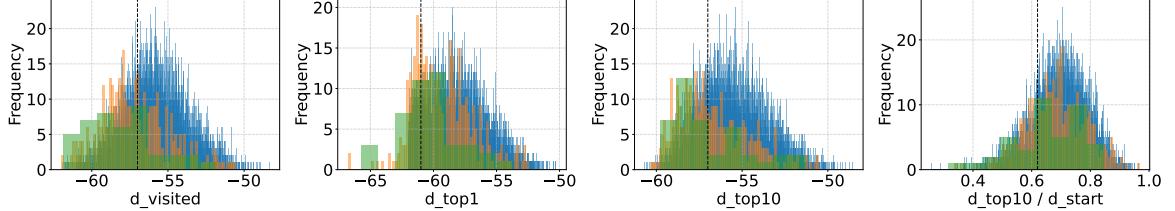
suggested in the Starling system [30], re-implementing it within the ParlayANN library for the in-memory setting for the fairest possible comparison.

4.2 Improving Range Search for Queries with Many Results

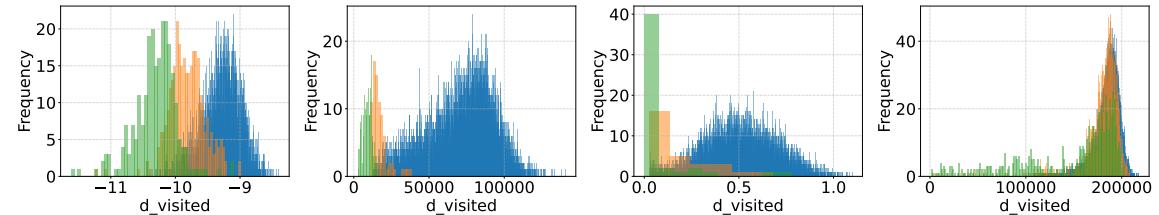
We first investigate how to approach graph-based range search for queries that have many results. A standard beam search can only capture as many points as are contained within the beam of size B , and thus any queries with significantly more than B results cannot



(a) Histograms plotting selected early stopping metrics for all query points using dataset MSMARCO-1M.



(b) Histograms plotting selected early stopping metrics on dataset MSMARCO-1M, excluding query points which have already found a candidate within the radius.

(c) Histograms showing the value of d_{visited} on more selected datasets. From left to right: Wikipedia-1M, BIGANN-1M, DEEP-1M, and SSNPP-1M.**Figure 3: Frequency distributions of selected metrics broken down by the number of matched points. All metric values were taken at step 20 of a beam search with beam 100. Queries are separated by color based on the number of range results.**

achieve high accuracy. The natural question is whether it is possible to identify such queries with more than B results and expand the set of returned candidates for only those queries. In this section we discuss two algorithmic approaches to this idea.

Doubling Beam Search. The first, and perhaps most natural, algorithm extending beam search to queries with variable number results is a dynamic beam search: that is, run the beam search with some starting beam B , then either terminate or double the beam and continue searching for more candidates. This algorithm was first proposed in the paper introducing the Starling system [30]. The algorithm terminates if fewer than some fraction λ of returned candidates are valid range candidates, and continues otherwise. In practice, we found that the value of λ did not change the results very much, so we set it to 1 as a rule (this finding seems to differ from that of the Starling authors, which may be attributed either to implementation differences or the in-memory setting). That algorithm is shown using beam search as a subroutine in Algorithm 3. In addition to rewriting the doubling search within the in-memory ParlayANN library, we made some minor optimizations such as using a faster approximate hash table for visited set lookups. For the fairest possible comparison, we also augment it with the early

stopping heuristic introduced in Section 4.3 during our experimental evaluation, but we note that the version in their paper contains no such heuristic.

Greedy Search. Another approach to dynamically resizing the queue length is to run a beam search with *unbounded* queue length, but much stricter criteria on which nodes are explored. Candidates are first generated by an initial beam search with beam B . If the number of candidates within the radius falls above a certain threshold (again, some fraction λ of the candidates returned), we next call our so-called greedy search routine, which has an unbounded queue but only adds nodes to the queue if they are within the ball of radius r around the query point for the specified radius. This approach takes advantage of the fact that nodes in the same ball of radius r are likely to be in connected clusters, and avoids the overhead of the doubling beam search by performing very few distance comparisons to points which are not valid results. See Algorithm 2 in the supplementary material for a complete description, and Section 5 for experimental analysis.

4.3 Improving Range Search for Queries with No Results

This section introduces an algorithm to improve range search for queries with no results. The search begins with a fixed beam B , continuing with Algorithm 2 or Algorithm 3 if only candidates within the radius are returned. Ideally, the initial beam search should terminate early when no results are found, using terms that are inexpensive to compute. A 2020 paper by Li et al. [17] explores similar termination criteria for top- k search, identifying several predictive terms: query point coordinates, distance to the current top- k neighbor for fixed k , and the ratio of that distance to the query point's distance from the start node.

To apply these ideas to range search, the key question is whether any of these metrics are effective at distinguishing queries with zero results from queries with one or more result. To investigate this, we look at each dataset and separate each query set into three groups: those with no results, those with 1-2 results, and those with more than three results. Then, we histogram the values of these metrics at different steps in the beam search, distinguishing between these three groups. We evaluate the following metrics, the last three of which are proposed in [17].

- d_{visited} : distance from query point q to the point being visited at step i of a beam search.
- $d_{\text{top}1}$: distance from query point q to the closest known neighbor of q at step i of a beam search.
- $d_{\text{top}10}$: distance from query point q to the tenth-closest known neighbor at step i of a beam search.
- $d_{\text{top}10}/d_{\text{start}}$: ratio of $d_{\text{top}10}$ to the distance from query point q to the starting point of the beam search.

A natural and general approach to an early stopping condition for range search, using a variety of metrics, is to apply the early stopping metric after a certain number of steps in the beam search. The search does not terminate early if it has already found at least one candidate in the specified range. Figure 3c shows one choice of metric— d_{visited} on four different datasets. Algorithm 4 shows formally how this condition can be applied.

Based on the histogram data of early stopping conditions, a larger set of which is shown in the supplemental material, we come to three conclusions: first, some datasets show potential for significant improvement using early stopping, while others do not. Second, if a dataset shows potential for improvement using one metric, it also tends to show improvement for all other metrics. Third, using d_{visited} seems to be the best early stopping metric. We hypothesize that d_{visited} is the best metric as it is the most frequently updated and contains the most granular information, but the positive results from other metrics suggests that approaches using machine learning such as those in the Li et al. paper [17] might also be useful. We present experimental analysis of this algorithm in Section 5. See Algorithm 5 for an example of how to apply the d_{visited} metric during a search, and Algorithm 6 for an example of the end-to-end algorithm using greedy search.

5 Experimental Results

We evaluate our algorithms on the nine datasets covered in Section 3.

Experimental Setup and Code. The algorithms introduced in Section 4 extend the ParlayANN [21] library, implemented in C++

Algorithm 1: BeamSearch($q, G, \mathcal{P}, S, r, b, \mathcal{M}$).

Input: Query point q , graph G , point set \mathcal{P} , starting points S , radius r , beam size b , early stopping metric \mathcal{M}

Output: Set \mathcal{B} of closest points, set \mathcal{V} of visited points

```

1  $\mathcal{B} \leftarrow S, \mathcal{V} \leftarrow \emptyset$ 
2 while  $\mathcal{B} \setminus \mathcal{V} \neq \emptyset$  do
3    $p^* \leftarrow \arg \min_{(p \in \mathcal{B} \setminus \mathcal{V})} \|p, q\|$ 
    //Early stopping conditions
4   if  $\mathcal{M}(q, \mathcal{B}, \mathcal{V}, r)$  then Break
5    $\mathcal{V} \leftarrow \mathcal{V} \cup \{p^*\}$   $\mathcal{B} \leftarrow \mathcal{B} \cup (N_{\text{out} \in G}(p^*) \setminus \mathcal{V})$ 
6   if  $|\mathcal{B}| > b$  then trim  $\mathcal{B}$  to size  $b$ , keeping the closest points to  $q$ 
7 return ( $\mathcal{B}, \mathcal{V}$ )

```

Algorithm 2: GreedySearch(q, G, \mathcal{P}, S, r).

Input: Query point q , graph G , point set \mathcal{P} , starting points S , radius r

Output: Set \mathcal{V} of visited points

```

1  $\mathcal{F} \leftarrow S, \mathcal{V} \leftarrow \emptyset$ 
2 while  $\mathcal{F} \setminus \mathcal{V} \neq \emptyset$  do
3    $p^* \leftarrow \arg \min_{(p \in \mathcal{F} \setminus \mathcal{V})} \|p, q\|$ 
4    $\mathcal{V} \leftarrow \mathcal{V} \cup \{p^*\}$ 
5   foreach  $f \in (N_{\text{out} \in G}(p^*) \setminus \mathcal{V})$  do
6     if  $\|f, q\| \leq r$  then  $\mathcal{F} = \mathcal{F} \cup \{f\}$ 
7 return  $\mathcal{V}$ 

```

Algorithm 3: DoublingSearch($q, G, \mathcal{P}, S, r, b, \mathcal{M}$).

Input: Query point q , graph G , point set \mathcal{P} , starting points S , radius r , beam size b , early stopping metric \mathcal{M}

Output: Set of neighbors \mathcal{N} within range

```

1 while true do
2    $\mathcal{N} \leftarrow \emptyset$ 
    // Early stopping is done inside the beam search
3    $(\mathcal{B}, \mathcal{V}) \leftarrow \text{BeamSearch}(q, G, \mathcal{P}, S, r, \mathcal{M}, b)$ 
4    $S \leftarrow S \cup \mathcal{V}$ 
5   for  $n \in \mathcal{B}$  do
6     if  $\|n, q\| \leq r$  then  $\mathcal{N} \leftarrow \mathcal{N} \cup \{n\}$ 
7   if  $|\mathcal{N}| < b$  then Break
8    $b \leftarrow 2 \times b$ 
9 return  $\mathcal{N}$ 

```

Algorithm 4: EarlyStopping($q, \mathcal{B}, \mathcal{V}, \mathcal{M}, C$).

Input: Query point q , current beam \mathcal{B} , visited set \mathcal{V} , early stopping metric \mathcal{M} , cutoff parameter set C

Output: Bool value for early stopping. True if conditions are met.

```

1 if  $\mathcal{M}(q, \mathcal{B}, \mathcal{V}, C)$  then return True

```

with ParlayLib [2] for fork-join parallelism and standard parallel primitives such as sorting and filtering. ParlayANN features a top-ranked DiskANN implementation, currently among the top five on ANN Benchmarks [4]. The code is publicly available at <https://github.com/cmuparlay/ParlayANN/tree/rangetest2>. Most experiments ran on an Azure Standard_L32s_v3 VM with a 3rd Gen Intel® Xeon® Platinum 8370C processor, 32 vCPUs, and 256 GB

Algorithm 5: EarlyStoppingExample($q, \mathcal{B}, \mathcal{V}, v, C = \{r, vl, esr\}$).

Input: Query point q , current beam \mathcal{B} , visited set \mathcal{V} , number of visited points v , radius r , limit of number of visits vl , early stopping radius esr

Output: Bool value for early stopping.

//When the closest point in the beam is not within the radius

- 1 **return** ($\arg \min_{p \in \mathcal{B}} \|p, q\| > r$, and
- //When the number of visited points is larger than the given cutoff
- 2 $v \geq vl$, and
- //When the current point visited is not within the early stopping radius
- 3 $\arg \min_{p^* \in \mathcal{V} \setminus \mathcal{B}} \|p^*, q\| > esr$)

Algorithm 6: GreedyRangeSearch($Q, G, \mathcal{P}, S, r, b, \mathcal{M}$).

Input: Query points Q , graph G , point set \mathcal{P} , starting points S , radius r , beam size b , early stopping metric \mathcal{M}

Output: Set of neighbors \mathcal{V} within range for each of the query points

- 1 **parallel for** $q \in Q$ **do**
- 2 $\mathcal{N} \leftarrow \emptyset$
- 3 $(\mathcal{B}, \mathcal{V}) \leftarrow \text{BeamSearch}(q, G, \mathcal{P}, S, r, b, \mathcal{M})$
- // Early stopping is done inside the beam search
- 4 **foreach** $n \in \mathcal{B}$ **do**
- 5 **if** $\|n, q\| \leq r$ **then** $\mathcal{N} \leftarrow \mathcal{N} \cup \{n\}$
- 6 **if** $|\mathcal{N}| < b$ **then** $\mathcal{V}[q] \leftarrow \mathcal{N}$
- 7 **else**
- // Add results from beam search as starting points for greedy search
- 8 $\mathcal{V}[q] \leftarrow \text{GreedySearch}(Q, G, \mathcal{P}, \mathcal{N}, r)$
- 9 **return** \mathcal{V}

memory, utilizing all vCPUs unless noted. Billion-scale tests were conducted on a 72-core Dell R930 with 4x Intel® Xeon® E7-8867 v4 processors, each with 18 cores, 2.4GHz, and 45MB L3 cache, and 1 TB main memory.

To produce a curve of QPS versus average precision, the starting beamwidth is varied on a number of searches and the Pareto frontier is reported. The build parameters used for the DiskANN graphs, as well as the early stopping threshold for each dataset, are reported in the supplemental material. We use scalar quantization to 8 bits for all floating-point datasets, with re-ranking to ensure that all reported results are within the radius.

Competitors. To evaluate our algorithms against another standard baseline, we also use the range search algorithms from the well-known FAISS-IVF library [9]. At a high level, FAISS-IVF uses an inverted file index produced via clustering, with the option to construct an HNSW index over the cluster centroids to help select clusters more efficiently at query time. A query probes a selected number of cells and returns the candidates within the radius. The parameters chosen for FAISS, along with information on how we configured them, are shown in the supplemental material. To produce a curve of QPS versus average precision for FAISS, we varied both the number of cells probed and the *ef_search* of the centroid graph and report the Pareto frontier.

QPS and Precision Tradeoffs. Our experimental evaluation plots queries per second (QPS) versus precision for nine datasets (Figure 4). Figure 5 examines size scaling effects for SSNPP and BIGANN, from 10 million to 1 billion. While our algorithms improve upon the naive baseline across all datasets, the degree of improvement and algorithm rankings vary. DEEP, BIGANN, and MSTuring achieve near-100x throughput gains over FAISS-IVF and the beam search baseline, with doubling beam search similar than greedy search at 1 million scale but up to 10x slower on larger datasets. OpenAI sees a 10x improvement over beam search and FAISS, with both doubling and greedy search performing similarly. Wikipedia and MSMARCO-WebSearch improve moderately over the FAISS and beam search baselines, with greedy search slightly ahead of doubling. GIST stands out due to its unique frequency distribution, as the only dataset with hundreds of points exceeding 10,000 range results, making greedy search more effective. Text2Image shows modest gains below 90% recall. SSNPP, analyzed across four scales, exhibits little improvement at 1 million but increasingly benefits from the specialized range search algorithms as dataset size grows.

Comparison with FAISS-IVF. Greedy search and doubling search are overall faster than FAISS-IVF, with occasional exceptions at the highest recall values. Interestingly, FAISS-IVF is faster than the beam search baseline about half the time. These results reflect the fact that the FAISS algorithm is more intrinsically suited to range search than the beam search baseline, since it considers every point in each cell as a potential range candidate. However, greedy search and doubling search with early stopping are overall more successful at precisely searching the correct number of points for each range query.

Scaling. Figure 5 examines the performance of our algorithms on two datasets at three different scales: 10 million, 100 million, and 1 billion. Similarly to existing range search benchmarks, we use the same radius for each data scale, so each order of magnitude effectively represents a more dense space, and the average number of results per query point increases. Unsurprisingly, the advantage of greedy search and doubling search increase significantly over the beam search baseline, which is only capable of returning up to about 1000 points per query before becoming hopelessly inefficient. Interestingly, we also see that the greedy search seems to have the most significant advantage over the doubling beam search as the density of the dataset increases (this is also true of GIST-1M, which has some extraordinarily dense outliers). This is likely because in denser graphs, the greedy search is particularly adept at efficiently finding all points within the radius, which are likely to be in a dense connected cluster in the graph structure. In the supplemental material we investigate this hypothesis directly by fixing one dataset and measuring performance of both algorithms while varying the chosen radius. Greedy search had better performance for larger choices of radius (indicating more matches per point on average).

Analyzing Time Spent on Algorithm Components. To aid in our analysis, we investigate the time spent on different phases of each algorithm in Figure 6. Each algorithm consists of an initial beam search phase—either with or without the early stopping criteria in Algorithm 4—and then a following phase of either greedy search or successive calls to doubling beam search. We show this visual

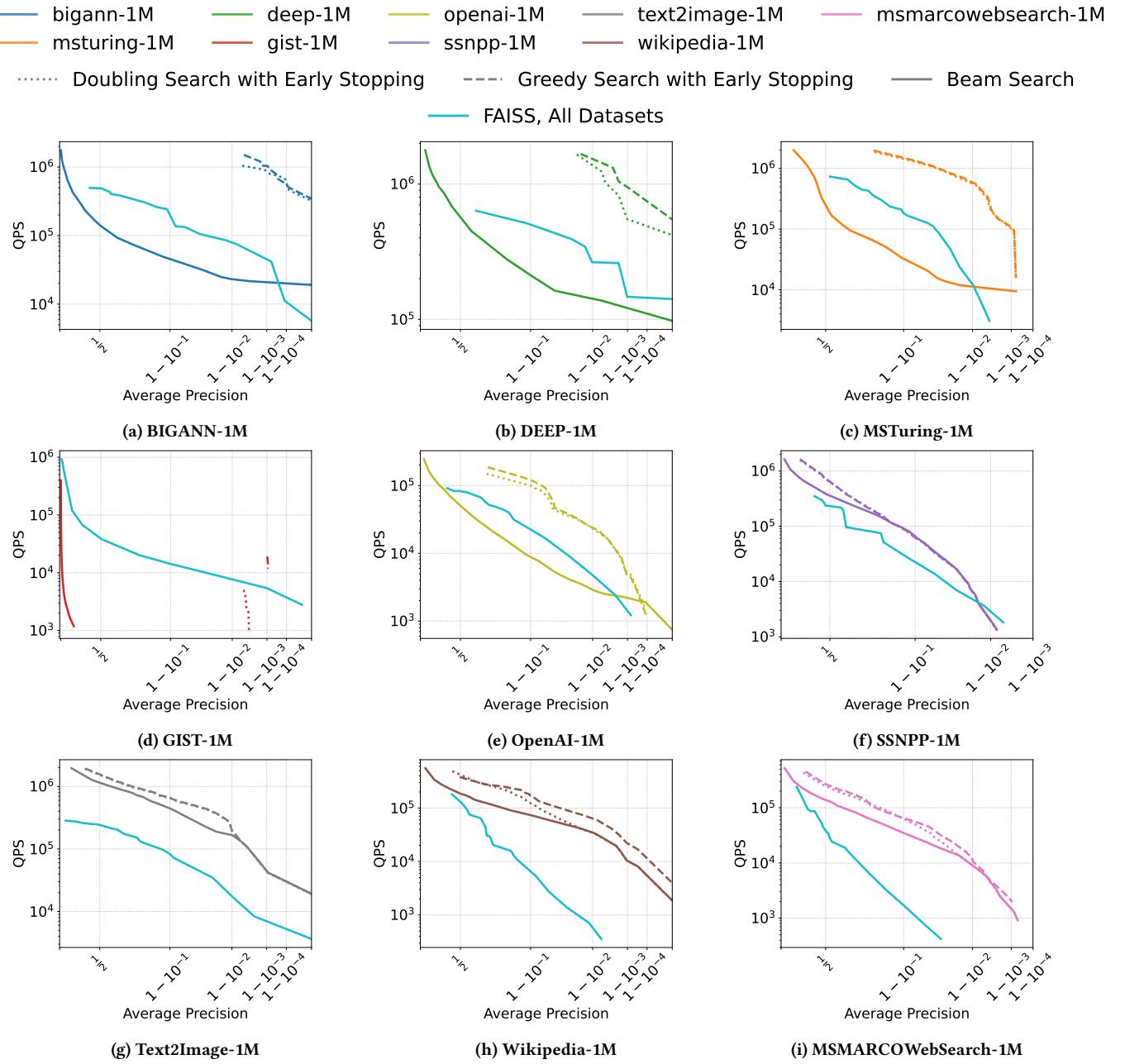


Figure 4: Average precision vs QPS for eight datasets and four range search algorithms. For GIST-1M, the lines for doubling search and greedy search are very short due to even the smallest of initial beam sizes producing recall in the .99-.999 range.

breakdown for three datasets at selected fixed average precision: BIGANN-1M, BIGANN-100M, and MSMarcoWebSearch-1M. The time breakdowns provide significant insight into the results in Figures 4 and 5. First, note that for all datasets, the initial beam search phase is a sizable portion of the overall cost of computation. This is because the initial beam search takes place on all points, but further searches only take place on points with one or more valid

range result (since the distances used are exact, a point with zero results will never pass on to the second round of computation).

Greedy Search vs Doubling Search. Next, we examine the differences between doubling and greedy search. Doubling search spends more time in the second phase but achieves higher recall from a smaller starting beam. Dataset characteristics determine the best tradeoff between QPS and average precision. Greedy search requires a larger initial beam but has a cheaper second round since

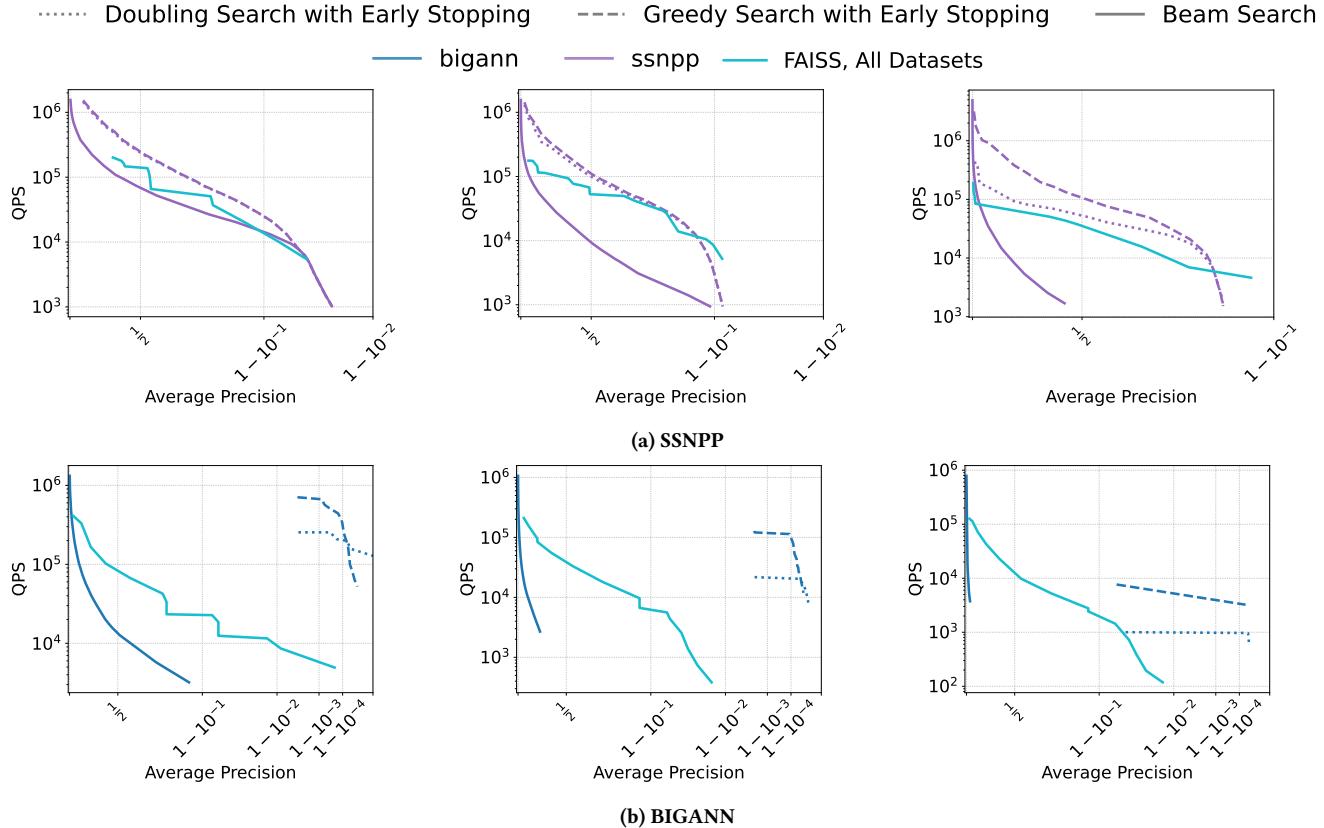


Figure 5: Average precision vs QPS for the SSNPP and BIGANN datasets. From left to right, slices of size 10 million, 100 million and 1 billion.

it only visits points within the radius. In contrast, doubling search starts with a smaller beam but spends more time in the second phase, performing additional distance comparisons and incurring overhead from repeated beam search calls. Experimental results suggest doubling search dominates greedy search when (a) a small initial beam is enough for doubling search to achieve high recall, and (b) few rounds of doubling are needed. These conditions hold for BIGANN-1M (and DEEP-1M), but GIST-1M requires excessive doubling rounds, and MSMARCOWebSearch-1M demands a large initial beam. Scaling studies in Figure 5 show greedy search overtakes doubling search as match density increases, likely due to the growing number of doubling rounds required.

Effects of Increasing Radius. To better isolate the reason for greedy search’s superior performance, in Figure 7, we measure the effect of varying the radius for a fixed dataset for three selected values of average precision, annotating each data point with the beam used for the initial beam search. To aid the reader’s understanding, instead of marking the x-axis with the raw value of the radius used, we instead measure the average number of matches per query induced by the chosen radius. We see that in most cases greedy search outperforms or performs equally to doubling search, with the effect becoming more pronounced as the number of matches per point increases. This confirms the observations in the size scaling study.

Effects of Early Stopping. Next, we examine the impact of early stopping. It significantly reduces search time for a fixed beam; for example, in Figure 6b, early stopping with beam 44 takes only about twice the time as a search with beam 2 and no early stopping. However, as Figure 6c shows, early stopping can reduce accuracy by prematurely terminating points with valid range results. Appendix A details its effects on selected datasets. Overall, early stopping is beneficial when there is clear separation between points with no results and those with one or more results.

Comparison with Top-k Search. We add a final note on comparison with top- k search. To investigate this, we ran top-10 searches on OpenAI-1M and MSTuring-1M, and measured the QPS at 90% recall. For OpenAI-1M, the QPS at 90% recall was around 5000, compared to around 10000 QPS for range search at 90% precision. For MSTuring-1M, the QPS at 90% recall for top-10 search was about 30000, compared to over 100000 for range search using our algorithms. This suggests range search may actually be an easier problem than top- k search.

6 Conclusion

We present a set of algorithms for range retrieval on high-dimensional datasets using graph-based indices. To improve the performance of queries with no range results, we introduce an early stopping

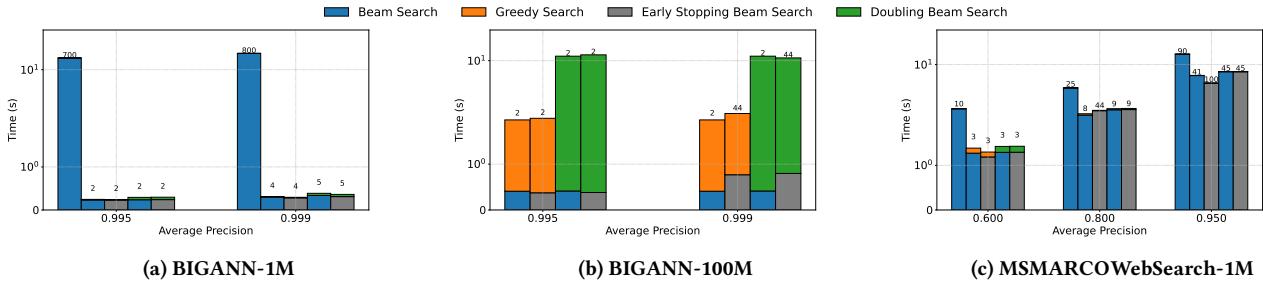


Figure 6: Figures breaking down the cost in seconds (single threaded time) of each type of search for three datasets for selected average precision. The label at the top of each column indicates the beam width of the initial search. Each collection from left to right shows: the beam search baseline (which is sometimes omitted when it cannot achieve the desired average precision), beam search followed by greedy search, beam search with early stopping followed by greedy search, beam search followed by doubling beam search, and beam search with early stopping followed by doubling beam search. Note that in some cases, the time spent on the second phase of search is so small that it is not visible.

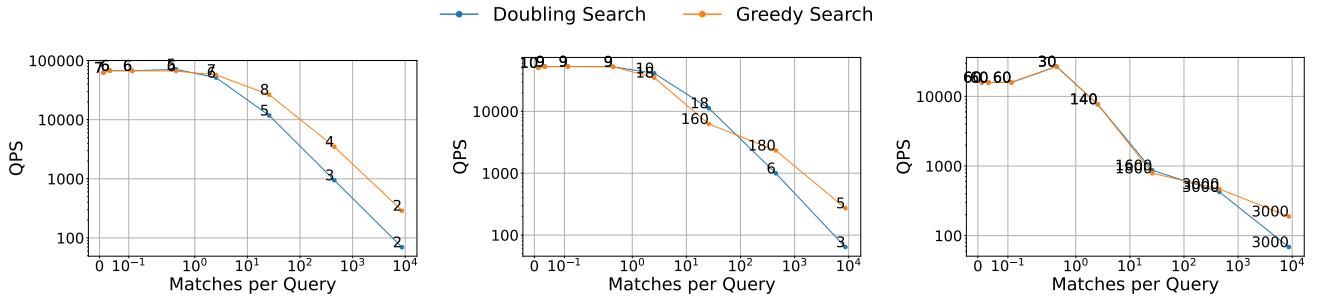


Figure 7: Figures showing the effect of varying the search radius (expressed here as average number of matches per query) on the performance of greedy search and doubling search. Each figure shows the Wikipedia-10M dataset for selected average precision (from left to right: .9, .95, .99). Each data point is annotated with the beam used for the initial beam search. Early stopping is not used for either search type.

mechanism capable of significantly decreasing the time taken to terminate searches on such queries. We also introduce two extensions of beam search, greedy search and doubling search, for efficient range queries with hundreds to tens of thousands of results. Combined, our techniques achieved up to 100x speedup over a naive baseline, with 5-10x speedup in almost all cases. We introduce a benchmark of nine datasets, including datasets with up to 100 million points, and demonstrate strong performance and scalability for our algorithms on these datasets.

Our results yield significant improvement over the naive baseline, and established a wide-ranging benchmark for performance. As we tackle a relatively unexplored problem, many areas for further improvement are still available. One natural extension would be to explore more sophisticated metrics for our early stopping algorithm, perhaps using machine learning methods, as better early termination of queries with few results has potential for significant impact on overall search times. Another would be to explore extensions of the range retrieval benchmark itself; for example, Szilvassy et al. [28] argue that the range search benchmark should differentiate between results which are very close to the query point and points which are close to the radius boundary, giving the

former more weight. In preliminary experiments, we attempted to use a simple weighting mechanism based on a point’s distance to the query point when calculating precision, but we were unable to find a weighting that changed the precision values significantly. However, future work identifying such mechanisms would be very practical and promising.

Finally, there is a notable lack of benchmarks designed for range searching. Contributions of new datasets with natural range search benchmarks would significantly further this area of exploration.

Acknowledgments

This work was supported in part by the National Science Foundation grants CCF-2119352 and CCF-1919223.

References

- [1] [n. d.]. ([n. d.]).
- [2] [n. d.]. ParlayLib: A Toolkit for Programming Parallel Algorithms on Shared-Memory Multicore Machines. <https://cmuparlay.github.io/parlaylib/> ([n. d.]).
- [3] Laurent Amsaleg and Hervé Jegou. 2010. Datasets for approximate nearest neighbor search. <http://corpus-texmex.irisa.fr/>.
- [4] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2021. ANN Benchmarks. <http://ann-benchmarks.com/>.
- [5] Dmitry Baranchuk and Artem Babenko. 2021. Deep1B Dataset. <https://research.yandex.com/blog/benchmarks-for-billion-scale-similarity-search>.
- [6] Bernard Chazelle, Ding Liu, and Avner Magen. 2008. Approximate range searching in higher dimension. *Comput. Geom.* 39, 1 (2008), 24–29. <https://doi.org/10.1016/j.comgeo.2007.05.008>
- [7] Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang. 2024. MS MARCO Web Search: A Large-scale Information-rich Web Dataset with Millions of Real Click Labels. In *Companion Proceedings of the ACM on Web Conference 2024 (WWW '24)*.
- [8] Matthijs Douze. 2022. Faiss Wiki. Webpage. Retrieved April 29, 2025 from <https://github.com/facebookresearch/faiss/wiki>
- [9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *CoRR* (2024). <https://doi.org/10.48550/arXiv.2401.08281>
- [10] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomás Jenícek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, Andrej Chum, and Cristian Canton-Ferrer. 2021. The 2021 Image Similarity Dataset and Challenge. *CoRR* (2021). <https://arxiv.org/abs/2106.09672>
- [11] Joshua Engels, Benjamin Landrum, Shangdi Yu, Laxman Dhulipala, and Julian Shun. 2024. Approximate Nearest Neighbor Search with Window Filters. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=8t8zBaGFar>
- [12] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai. 2019. Fast Approximate Nearest Neighbor Search With The Navigating Spreading-out Graph. *Proc. VLDB Endow.* 12, 5 (2019), 461–474. <http://www.vldb.org/pvldb/vol12/p461-fu.pdf>
- [13] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. <https://openai.com/index/new-and-improved-embedding-model/>.
- [14] Jonathan Halcrow, Alexandru Mosoi, Sam Ruth, and Bryan Perozzi. 2020. Grale: Designing Networks for Graph Learning. In *SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2523–2532. <https://doi.org/10.1145/3394486.3403302>
- [15] Piotr Indyk and Rajeev Motwani. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*. ACM, 604–613. <https://doi.org/10.1145/276698.276876>
- [16] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33, 1 (2010).
- [17] Conglong Li, Minjia Zhang, David G. Andersen, and Yuxiong He. 2020. Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*. ACM, 2539–2554. <https://doi.org/10.1145/3318464.3380600>
- [18] Hao Li, Xiaojie Liu, Tao Li, and Rundong Gan. 2020. A novel density-based clustering algorithm using nearest neighbor graph. *Pattern Recognit.* 102 (2020), 107206. <https://doi.org/10.1016/j.patcog.2020.107206>
- [19] Yury Malkov and Dmitry Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 4 (2018).
- [20] Magdalen Dobson Manohar, Zheqi Shen, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Harsha Vardhan Simhadri, and Yihan Sun. 2024. ParlayANN: Scalable and Deterministic Parallel Graph-Based Approximate Nearest Neighbor Search Algorithms. In *Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming, PPoPP 2024, Edinburgh, United Kingdom, March 2-6, 2024*. ACM, 270–285. <https://doi.org/10.1145/3627535.3638475>
- [21] Magdalen Dobson Manohar, Zheqi Shen, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Harsha Vardhan Simhadri, and Yihan Sun. 2025. ParlayANN. Webpage. Retrieved February 7, 2025 from <https://github.com/cmuparlay/ParlayANN>
- [22] MetaAI. 2020. Using AI to detect COVID-19 misinformation and exploitative content. Webpage. Retrieved January 26, 2022 from <https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/>
- [23] Nils Reimers. 2022. Datasets: Cohere/wikipedia-22-12-en-embeddings. [https://huggingface.co/datasets/Coherewikipedia-22-12-en-embeddings](https://huggingface.co/datasets/Cohere/wikipedia-22-12-en-embeddings).
- [24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 815–823. <https://doi.org/10.1109/CVPR.2015.7298682>
- [25] Harsha Simhadri, George Williams, Martin Aumüller, Artem Babenko, Dmitry Baranchuk, Qi Chen, Matthijs Douze, Lucas Hosseini, Ravishankar Krishnaswamy, Gopal Srinivasa, Suhas Jayaram Subramanya, and Jingdong Wang. 2021. BigANN Benchmarks: Billion-Scale Approximate Nearest Neighbor Search Challenge. <https://big-ann-benchmarks.com/>
- [26] Harsha Vardhan Simhadri, George Williams, Martin Aumüller, Matthijs Douze, Artem Babenko, Dmitry Baranchuk, Qi Chen, Lucas Hosseini, Ravishankar Krishnaswamy, Gopal Srinivasa, Suhas Jayaram Subramanya, and Jingdong Wang. 2021. Results of the NeurIPS'21 Challenge on Billion-Scale Approximate Nearest Neighbor Search. In *NeurIPS 2021 Competitions and Demonstrations Track, 6-14 December 2021, Online (Proceedings of Machine Learning Research, Vol. 176)*. PMLR, 177–189. <https://proceedings.mlr.press/v176/simhadri22a.html>
- [27] Suhas Jayaram Subramanya, Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnaswamy, and Rohan Kadekodi. 2019. Rand-NSG: Fast Accurate Billion-point Nearest Neighbor Search on Single Node. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 13748–13758. <https://proceedings.neurips.cc/paper/2019/hash/09853c7fb1d3f8ee67a61b6bf4a7f8e6-Abstract.html>
- [28] Gergely Szilvassy, Pierre-Emmanuel Mazaré, and Matthijs Douze. 2024. Vector search with small radiiuses. *CoRR* abs/2403.10746 (2024). <https://doi.org/10.48550/arXiv.2403.10746>
- [29] Jie Wang, Jian Lu, Zheng Fang, Tingjian Ge, and Cindy X. Chen. 2013. PL-Tree: An Efficient Indexing Method for High-Dimensional Data. In *Advances in Spatial and Temporal Databases - 13th International Symposium, SSTD 2013, Munich, Germany, August 21-23, 2013. Proceedings (Lecture Notes in Computer Science, Vol. 8098)*. Springer, 183–200. https://doi.org/10.1007/978-3-642-40235-7_11
- [30] Mengzhao Wang, Weizhi Xu, Xiaomeng Yi, Songlin Wu, Zhangyang Peng, Xiangyu Ke, Yunjun Gao, Xiaoliang Xu, Rentong Guo, and Charles Xie. 2024. Starling: An I/O-Efficient Disk-Resident Graph Index Framework for High-Dimensional Vector Similarity Search on Data Segment. *Proc. ACM Manag. Data* 2, 1 (2024), V2mod014:1–V2mod014:27. <https://doi.org/10.1145/3639269>
- [31] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search. *Proc. VLDB Endow.* 14, 11 (2021), 1964–1978. <http://www.vldb.org/pvldb/vol14/p1964-wang.pdf>
- [32] Xiaoliang Xu, Mengzhao Wang, Yuxiang Wang, and Dingcheng Ma. 2021. Two-stage routing with optimized guided search and greedy algorithm on proximity graph. *Knowl. Based Syst.* 229 (2021), 107305. <https://doi.org/10.1016/j.knosys.2021.107305>
- [33] Yuexuan Xu, Jianyang Gao, Yutong Gou, Cheng Long, and Christian S. Jensen. 2024. iRangeGraph: Improvising Range-dedicated Graphs for Range-filtering Nearest Neighbor Search. *Proc. ACM Manag. Data* 2, 6 (2024), 239:1–239:26. <https://doi.org/10.1145/3698814>
- [34] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic Intent Representation in Web Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 65–74. <https://doi.org/10.1145/3331184.3331198>
- [35] Chaoji Zuo, Miao Qiao, Wencho Zhou, Feifei Li, and Dong Deng. 2024. SeRF: Segment Graph for Range-Filtering Approximate Nearest Neighbor Search. *Proc. ACM Manag. Data* 2, 1 (2024), 69:1–69:26. <https://doi.org/10.1145/3639324>

A Supplementary Material

A.1 Additional Data on Early Stopping

We provide data on early stopping metrics for four datasets in Figures 8, 9, 10, and 11. In each figure, we histogram the values of each metric, first with all results included, and then with only points where the search has not already found a point inside the radius. The figure also shows vertical lines indicating a potential cutoff for Algorithm 4; points with more than one range result to the right of the line would be “incorrectly” cut off from further search, while points with zero points to the right of the line would be “incorrectly” allowed to proceed instead of terminating early. The goal is to identify a cutoff that yields a more favorable QPS/recall curve.

In Figure 12 and Figure 13, we examine the impact of early stopping on each dataset at the 1 million scale, measuring improvement over both greedy search and doubling search.

A.2 Parameter Values

In Figure 14, we show our choice of FAISS parameters for each dataset. For each dataset, we chose a starting set of parameters

based on both instructions from the FAISS Wiki [8] and the parameter settings from the NeurIPS’21 Big ANN Benchmarks Competition [25]. In each case, we use two-level clustering to significantly speed up build times. We then experimented with increasing and decreasing the choice of both the size of the PQ codes and the number of centroids, and kept the parameters that gave the optimal QPS/Average Precision curve. For SSNPP-100M, we also made use of the Hamming threshold parameter as suggested from the configurations in the Big ANN Benchmarks repository. The curves were drawn by varying the HNSW *ef_search* parameter as well as the number of centroids probed. FAISS has a build-in range search method [8], which we used to conduct searches. For Wikipedia and MSMARCOWebSearch, we used no quantization during range search as the quantization seemed to distort distances for inner product datasets and produced nonsensical results.

In Figure 15 we show the choices of build and search parameters for the DiskANN graphs used in our experiments. For the build parameters, we used $R = 64, L = 128$ wherever possible, although we found it was not sufficient for SSNPP. The α parameter was configured experimentally for each dataset, as was the choice of early stopping threshold.

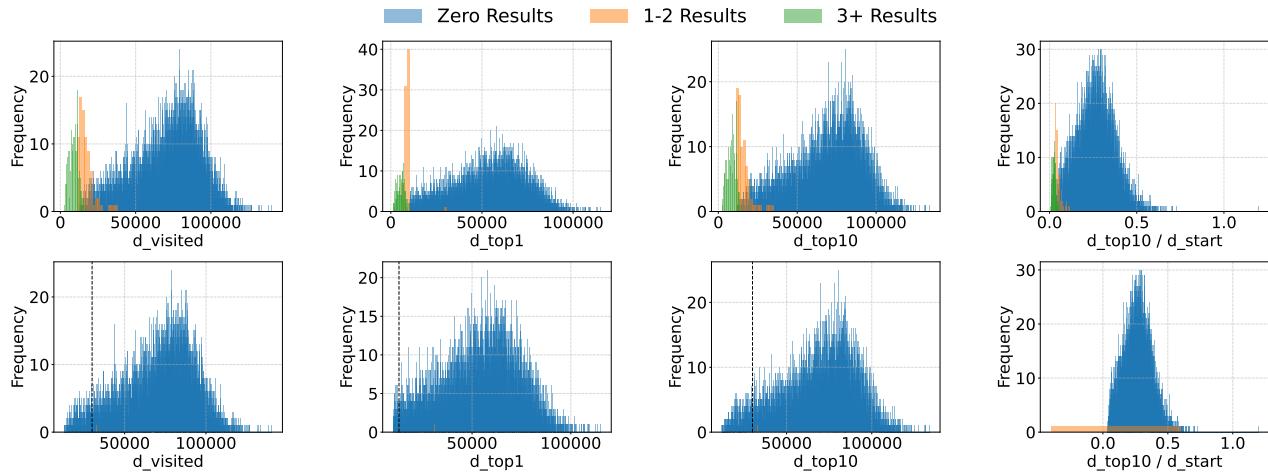


Figure 8: Early stopping metrics for BIGANN-1M, taken at step 20 of a beam search with beam 100. The top row shows all results, while the bottom row shows only results for beam searches that have not yet found a candidate within the radius.

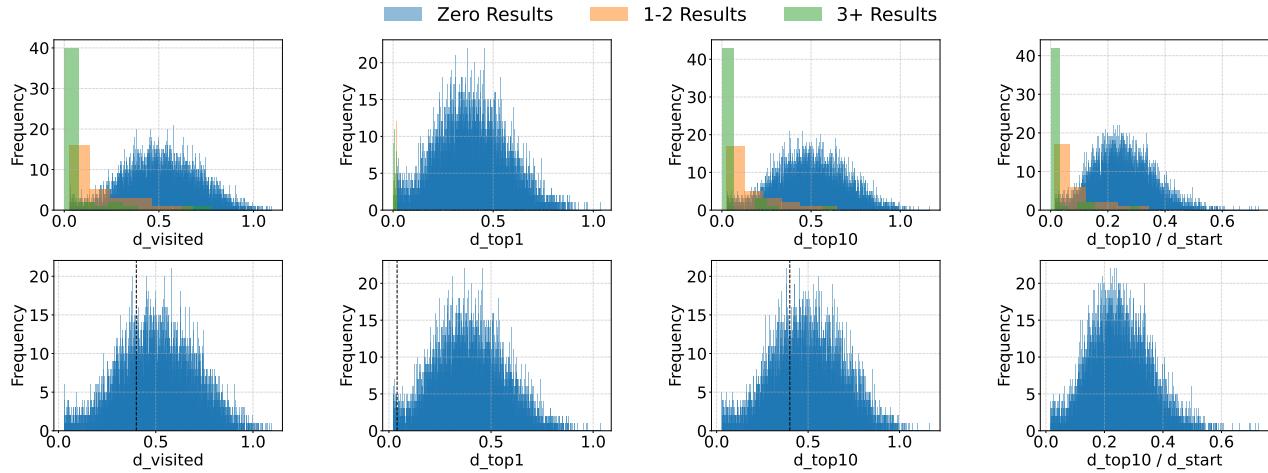


Figure 9: Early stopping metrics for DEEP-1M, taken at step 20 of a beam search with beam 100. The top row shows all results, while the bottom row shows only results for beam searches that have not yet found a candidate within the radius.

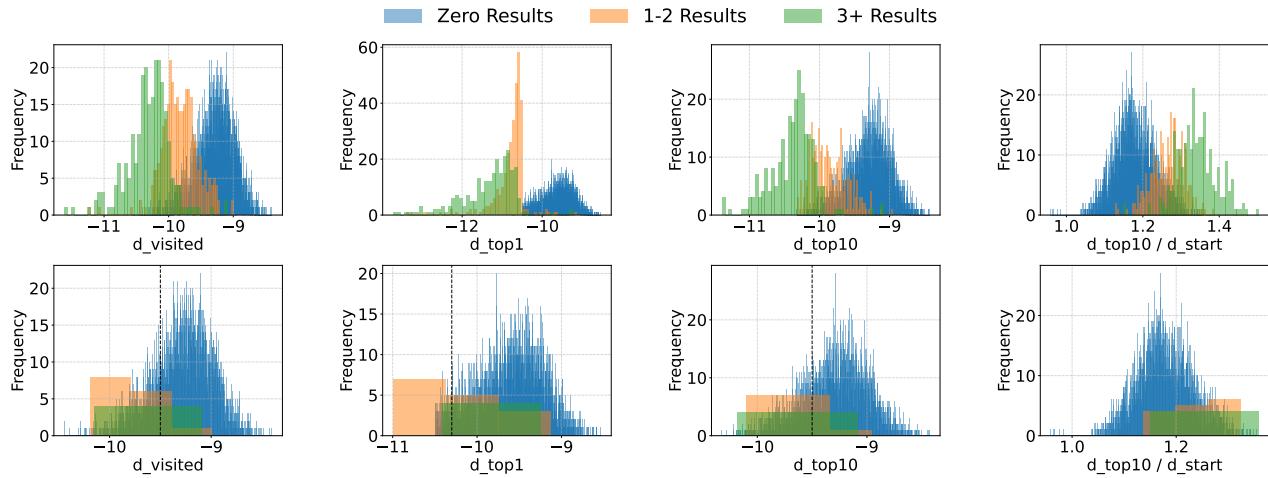


Figure 10: Early stopping metrics for Wikipedia-1M, taken at step 20 of a beam search with beam 100. The top row shows all results, while the bottom row shows only results for beam searches that have not yet found a candidate within the radius.

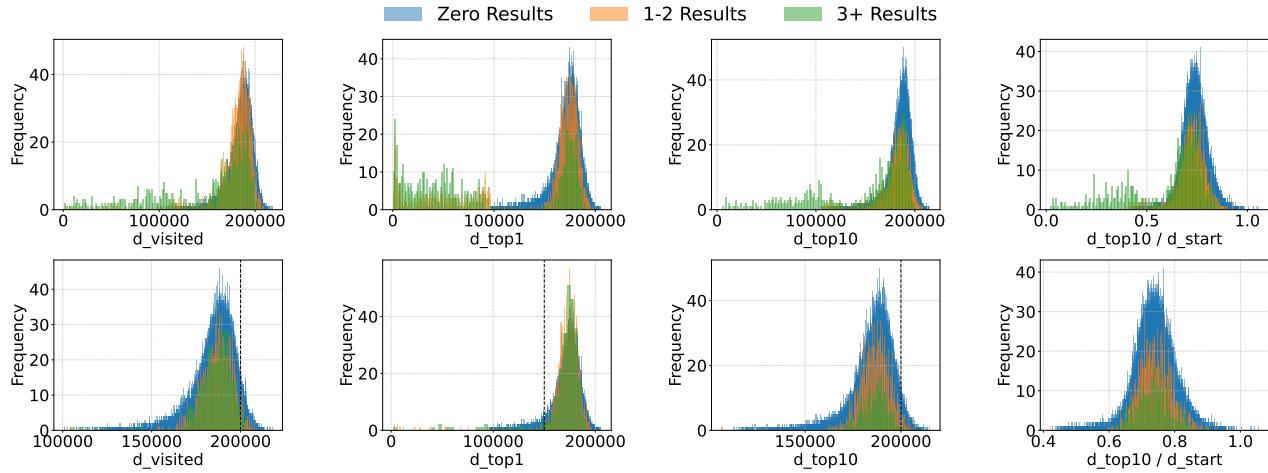


Figure 11: Early stopping metrics for SSNPP-1M, taken at step 20 of a beam search with beam 100. The top row shows all results, while the bottom row shows only results for beam searches that have not yet found a candidate within the radius.

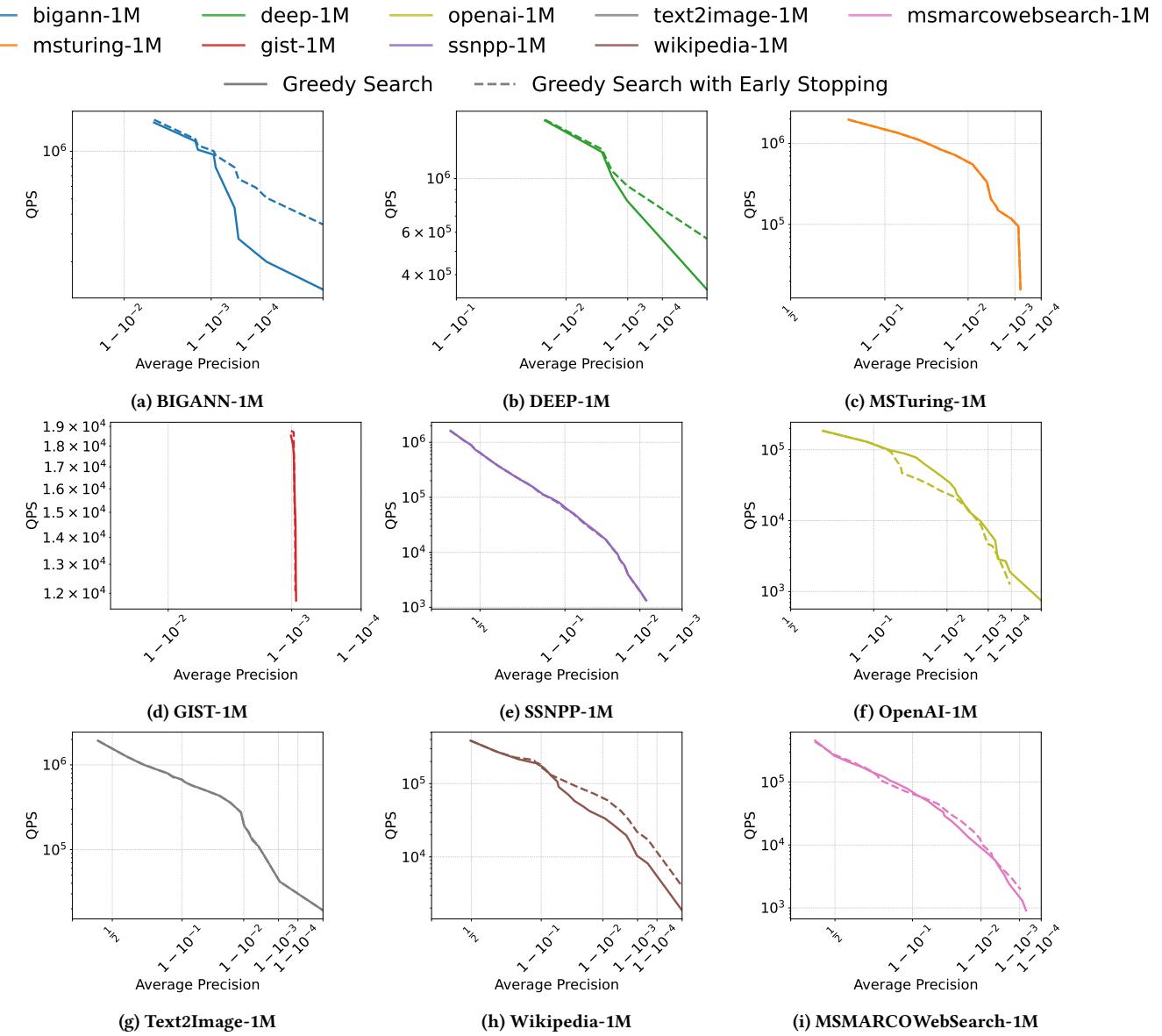


Figure 12: Average precision vs QPS for all nine datasets using greedy search, comparing use of early stopping versus without early stopping.

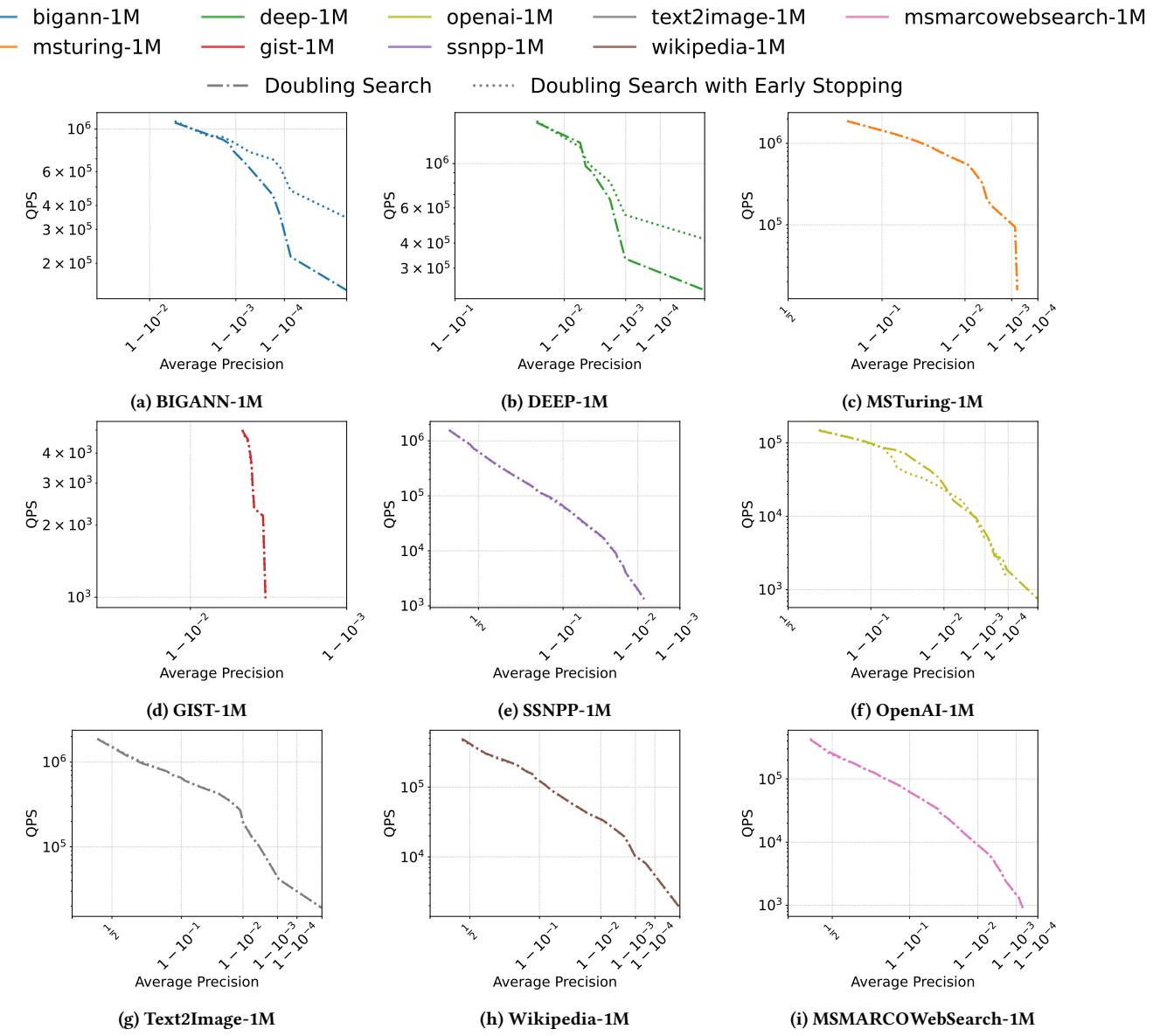


Figure 13: Average precision vs QPS for all nine datasets using doubling search, comparing use of early stopping versus without early stopping.

Dataset	Parameters
BIGANN-1M	"OPQ32_64,IVF65536_HNSW32,PQ64x4fsr"
BIGANN-10M	"OPQ32_64,IVF1048576_HNSW32,PQ64x4fsr"
BIGANN-100M	"OPQ64_128,IVF1048576_HNSW32,PQ128x4fsr"
DEEP-1M	"OPQ32_64,IVF65536_HNSW32,PQ64x4fsr"
MSTuring-1M	"OPQ32_64,IVF16384_HNSW32,PQ64x4fsr"
GIST-1M	"OPQ32_64,IVF16384_HNSW32,PQ64x4fsr"
SSNPP-1M	"OPQ32_64,IVF65536_HNSW32,PQ64x4fsr"
SSNPP-10M	"OPQ64_128,IVF65536_HNSW32,PQ128x4fsr"
SSNPP-100M	"OPQ32_128,IVF1048576_HNSW32,PQ32"
OpenAI-1M	"OPQ32_64,IVF16384_HNSW32,PQ64x4fsr"
Text2Image-1M	"OPQ64_128,IVF16384_HNSW32,PQ128x4fsr"
Wikipedia-1M	"IVF16384_HNSW32,Flat"
MSMARCOWebSearch-1M	"IVF16384_HNSW32,Flat"

Figure 14: Choices of FAISS parameters for each dataset.

Dataset	Build Parameters	Early Stopping Threshold
BIGANN (all sizes)	$R = 64, L = 128, \alpha = 1.15$	10000
DEEP-1M	$R = 64, L = 128, \alpha = 1.15$.4
MSTuring-1M	$R = 64, L = 128, \alpha = 1.15$	1.6
GIST-1M	$R = 64, L = 128, \alpha = 1.15$	2.5
SSNPP (all sizes)	$R = 80, L = 200, \alpha = 1.1$	200000
OpenAI-1M	$R = 64, L = 128, \alpha = 1.15$.28
Text2Image-1M	$R = 64, L = 128, \alpha = 1.0$	-1.0
Wikipedia-1M	$R = 64, L = 128, \alpha = 1.0$	-9.5
MSMARCOWebSearch-1M	$R = 64, L = 128, \alpha = 1.0$	-57

Figure 15: Choices of DiskANN parameters for each dataset.