LSH-APG: Building Approximate Proximity Graphs Through Locality-Sensitive Hashing

ABSTRACT

The approximate nearest neighbor (ANN) search in high-dimensional spaces is a fundamental but computationally very expensive problem. Existing solutions can be mainly categorized into LSH-based methods and graph-based methods. The LSH-based methods can be costly to reach high query quality due to the hash-boundary issues, while the graph-based methods can achieve better query performance by greedy expansion in an approximate proximity graph (APG). However, the construction cost of APG can be one or two orders of magnitude higher than that for building hash-based indexes. In this paper, we propose a novel approach named LSH-APG to build APGs and facilitate fast ANN search using a lightweight LSH framework. LSH-APG builds an APG via consecutively inserting points based on their nearest neighbor relationship with an efficient and accurate LSH-based search strategy. A high-quality entry point selection technique and an LSH-based pruning condition are developed to accelerate index construction and query processing by reducing the number of points to be accessed during the search. Our approach is proven to be less affected by dataset cardinality. LSH-APG can be maintained incrementally as datasets evolve. Furthermore, this approach can be adapted to other graph-based methods. Extensive experiments on real-world and synthetic datasets demonstrate that LSH-APG incurs significantly less construction cost but achieves better query performance than existing graph-based methods.

KEYWORDS

Graph based method, nearest neighbor search, locality-sensitive hashing

ACM Reference Format:
. 2018. LSH-APG: Building Approximate Proximity Graphs Through Locality-Sensitive Hashing. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Given a dataset and a distance function, the nearest neighbor (NN) search problem aims to find the point in the dataset with the minimum distance to a given query point. It has applications in a wide range of areas such as machine learning [4], pattern recognition [34] and data mining [36]. It is well known that finding the exact NN in large high-dimensional datasets can be very time-consuming[20]. A more efficient and potentially more practical alternative is to perform an approximate nearest neighbor (ANN) search. For example, the c-approximate nearest neighbor (c-ANN) search problem finds a point whose distance to the query point q is at most cr, where r is the distance from q to its exact NN in the dataset and c is a user-given approximation ratio [8, 16].

A large number of solutions have been proposed for efficient ANN search in high-dimensional spaces. These methods can be broadly divided into two categories: locality-sensitive hashing (LSH)-based methods [2, 15, 16, 28, 35, 37] and graph-based methods [14, 17, 17, 31]. The LSH-based methods are known for their robust theoretical guarantee on query result accuracy and simple and efficient implementation. By mapping points in a high-dimensional space to low-dimensional spaces via a set of LSH functions, it is possible to find a c-ANN result with a guaranteed high probability by only checking the points around the query point in low-dimensional spaces[2, 16]. On the other hand, the graph-based methods build an approximate proximity graph (APG) where each data point is represented as a vertex in the graph, and there is an edge between two vertices if the corresponding points are sufficiently close to each other in the original space. Given a query q, the search begins from an arbitrary point o_1 in the APG and expands to its neighbor o_2 such that o_2 is the closest point to q. The process repeats until it reaches a point v, which is closer to q than any of its neighbors in the APG.

The graph-based methods have been extensively studied due to their higher query processing efficiency than the LSH-based methods in practice. Compared with LSH-based methods that may suffer from the problem of hash-boundary issues (*i.e.,* some points that are close in the original space may be mapped into different hash buckets, thus not being identified by LSH-based approaches as neighbors) [15, 16, 37], the graph-based methods can achieve much higher accuracy [14, 38]. However, the cost of APG construction can be one or two orders of magnitude higher than that of LSH-based methods for very large datasets [3, 25], as such an approach will require finding proper neighbors for each and every point in the dataset. In addition, the existing graph-based methods largely only consider static datasets. It remains to be a challenging problem to efficiently maintain APG (*i.e.,* without reconstructing the whole or a large part of the graph) as the dataset evolves.

Motivated by the above observations, in this paper, we propose LSH-APG, a novel approach to build APGs from lightweight LSH indexes to facilitate efficient ANN query processing. By exploiting the properties of LSH functions and analyzing the problem using different structures of proximity graphs, LSH-APG can overcome the drawbacks of both LSH (*i.e.,* the hash-boundary issues) and graph-based methods (*i.e.,* high graph construction time and poor maintainability for dynamic datasets). It adopts LSH indexes to quickly retrieve some query results as the entry point for a search in a proximity graph, followed by using graph-based techniques to further improve the accuracy of the query result. To boost query processing efficiency, an accurate and scalable pruning strategy is proposed to filter out those neighbors which are far away from the query point. This strategy can significantly reduce the number of points that need to be accessed during the search on the graph. A consecutive insertion strategy is used to build the graph index. It handles the high construction cost issue with the help of the LSH framework. All points are consecutively inserted into the index structure where each point is regarded as a query point and inserted into the graph index based on its nearest neighbors. This strategy not only reduces the construction cost due to improved search efficiency using the LSH framework but also enables a formal correctness and complexity analysis for our approach.

In summary, our main contributions in this paper include:

- Based on a comprehensive analysis of state-of-the-art LSH-based methods and graph-based methods, a new solution named LSH-APG is developed by adopting a well-designed LSH framework to accelerate indexing and query processing using proximity graphs with higher-quality entry point selection and LSH-based pruning conditions. LSH-APG has a much lower construction cost without scarifying query efficiency or query quality.
- The LSH framework is extended to other graph-based methods to demonstrate its effectiveness. In addition, an efficient update strategy is designed to maintain index structures as the database evolves.
- Assume the expected number of accessed points during the search is C_Q, LSH-APG can be built in an expected computational cost of $O(ndC_Q)$ and answer an ANN query with a constant success probability in expected $O(dC_Q)$ time, where n, d are the cardinality and dimensionality of the dataset. C_Q is demonstrated to be less affected by n. The effectiveness of the LSH framework is also theoretically proven.
- Extensive experiments show that LSH-APG can greatly reduce indexing cost and achieve the best trade-off between query processing efficiency and query accuracy compared with the existing methods for ANN queries.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the problem definition, basic concepts and analyzes the limitations in the existing methods. The structure of LSH-APG and query algorithm are presented in Section 4 and Section 5. The update and extension strategies are discussed in Section 6. Section 7 and Section 8 report theoretical analysis and experimental results. We conclude this paper in Section 9.

2 RELATED WORK

We discuss the mainstream ANNS methods in high-dimensional spaces, including LSH-based methods and graph-based methods.

2.1 LSH-based Methods

LSH is originally proposed in [8, 16, 20]. In these methods, LSH maps data points in the high-dimensional space into several low-dimensional hash buckets and the ANN query is answered by checking the buckets where the query falls. However, to guarantee a high query accuracy and sub-linear query cost, these methods require preparing multiple suit of LSH indexes with different bucket width, which causes undesirably large index sizes and limits their applications. [35] and [15] addressed this issue via the *virtual rehashing* technique, but they are hard to achieve a very high query quality due to the hash boundary issue, a shortcoming shared by all static LSH-based methods. To relieve the hash boundary issue, recent studies focus on designing dynamic LSH methods that dynamically construct query-centric hash buckets for every query point via novel LSH frameworks, such as collision counting based strategy [19, 28, 29] and metric based strategy [27, 41]. However, their query cost is no longer sub-linear due to the overhead of dynamic bucketing. In this year, Tian et al. combined the dynamic query strategy with the static LSH framework and proposed DB-LSH, which achieved the best query complexity theoretically among the existing LSH-based methods [37].

2.2 Graph-based Methods

Graph-based methods use the proximity graph to facilitate ANN search and have shown better query performance compared to LSH-based methods, in terms of both accuracy and efficiency recently [31, 32, 38]. However, the construction cost of exact proximity graph, *e.g.,* Delaunay graph (DG) [23], relative neighborhood graph (RNG) [21], minimum spanning tree (MST) [32] and kNN graph (KNNG) [9], is at least $O(n^2)$, which is unaffordable for large-scale datasets.

To lower the construction cost, several indexing strategies tried to build an approximate proximity graph (APG) at the cost of a bit of query quality. Dong et al.[9] proposed NN-Descent that approximates KNNG. In this method, an APG is built from a random graph and the edges for each point are updated by conducting local search iteratively among close neighbors of the query. The construction complexity of NN-Descent is lowered to $\tilde{O}(n^{1.14})$. Due to its efficiency compared to the brute-force manner, NN-Descent algorithm is used in many graph based methods, such as EFANNA [12], NSG [14] and others [26, 40] and some derivatives are developed [5]. However, it needs iterating about 10 times to find the high-quality neighbors, which is still time-consuming. NSW [30] builds the approximate KNNG and DG via consecutive insertion strategy, where the point to be inserted is regarded as a query in the current graph, and connected to its neighbors to finish the insertion. This construction manner is very efficient, but it can cause hubness issue, *i.e.,* high out-degrees for some vertexes, which makes the query inefficient. HNSW [31] adopts the same strategy as NSW but limits the maximum degree for each point to alleviate the hubness issue. HCNNG [32] and VRLSH [10] cluster or partition the data into many subgroups, and then build an APG in each subgroup. It could reduce the construction cost of graph index to $O(n)$ but requires building the graph multiple times to achieve a good performance.

To further boost query efficiency, many studies proposed the neighbor selection strategies to diversify the distribution of neighbors [14, 17, 25, 31]. In this manner, the similar edges will be cut off to reduce the average degree of the graph based on some occlusion rules. NSG [14] and HNSW consider the distribution of neighbors by their distance. In these two methods, for any two neighbors u and v of o, the value of $dist(u, v)$ must be larger than $dist(o, u)$ and $dist(o, v)$. Otherwise, the longer edge between (o, u) and (o, v)

Table 1: List of Key Notations.

Notation	Description
$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathcal{D}$	The dataset
n	The cardinality of dataset
d_i	The local intrinsic dimensionality of dataset
o, v, u	A data point
q	A query point
$\|o_1, o_2\|$	The distance between o_1 and o_2
$e = (o_1, o_2)$	The directed edge from o_1 to o_2
$h^*(o), h(o)$	Hash function
$\chi^2(m)$	The χ^2 distribution with freedom m
C_Q	The expected number of points accessed per query

would be discarded. DPG [25] and NSSG [13] consider the distribution of neighbors by the angle between two edges (o, u) and (o, v). DPG maximizes the average angle between any two edges of o and NSSG limits the angle between two edges no less than a given threshold. These strategies demonstrate better query performance, but all of them bring much larger construction costs due to the extra computation cost for occlusion rules.

3 PRELIMINARIES

In this section, we first introduce the problem definition, the concepts of LSH and graph-based methods. Then, a comprehensive analysis of limitations in the existing ANN methods is presented, which inspires us to design LSH-APG. Frequently used notations are summarized in the Table 1.

3.1 Problem Definition

In this paper, we study the c-ANN and (c, k)-ANN queries in the Euclidean space. Let $\mathcal{D}$ be a set of points in d-dimensional Euclidean space $\mathbb{R}^d$ with cardinality $|\mathcal{D}| = n$. Let $\|o_1, o_2\|$ denote the Euclidean distance between points $o_1, o_2 \in \mathcal{D}$.

Definition 1 ((c, k)-ANN Query [37]). *Given a query point q, an approximation ratio $c > 1$ and a positive integer k, a (c, k)-approximate nearest neighbor query returns k points $o_1, \ldots, o_k$ that are sorted in ascending order w.r.t. their distances to q. If o_i^* is the i-th nearest neighbor of q in $\mathcal{D}$, it satisfies that $\|q, o_i\| \leq c \cdot \|q, o_i^*\|$.*

Remark 1. *The c-ANN query is the (c, k)-ANN query with $k = 1$.*

Remark 2. *In the graph-based methods, we usually do not explicitly use c to control the query quality. Without confusion, we abbreviate (c, k)-ANN as kANN for simplicity.*

3.2 Locality Sensitive Hashing

Definition 2 (Locality Sensitive Hashing (LSH) [37, 41]). *Given a distance r and an approximation ratio $c > 1$, a family of hash functions $\mathcal{H} = \{h : \mathbb{R}^d \to \mathbb{R}\}$ is called (r, cr, p_1, p_2)-locality-sensitive, if for $\forall o_1, o_2 \in \mathbb{R}^d$, it satisfies both conditions below:*

(1) If $\|o_1, o_2\| \leq r, \Pr[h(o_1) = h(o_2)] \geq p_1$;

(2) If $\|o_1, o_2\| > cr, \Pr[h(o_1) = h(o_2)] \leq p_2$,

where $h \in \mathcal{H}$ is chosen at random, p_1, p_2 are collision probabilities and $p_1 > p_2$.

A typical LSH family in the Euclidean space is defined as follows [19]:

$$h^*(o) = \vec{a} \cdot \vec{o}, \tag{1}$$

where $\vec{o}$ is the vector representation of a point $o \in \mathbb{R}^d$ and $\vec{a}$ is a d-dimensional vector where each entry is chosen independently from the standard normal distribution.

Lemma 1. *Let $P(o) = (h_1^*(o), \ldots, h_m^*(o))$ be an m-dimensional vector where h_i^* is chosen from the LSH family in Eq. 1. Then, for $\forall$ $o_1, o_2 \in \mathcal{D}$, we have $\dfrac{\|P(o_1), P(o_2)\|^2}{\|o_1, o_2\|^2} \sim \chi^2(m)$.*

Proof. The proof of this lemma can be found in [41]. □

Another commonly used LSH family in the Euclidean space is defined as follows [8]:

$$h(o) = \left\lfloor \frac{h^*(o) + b}{w} \right\rfloor, \tag{2}$$

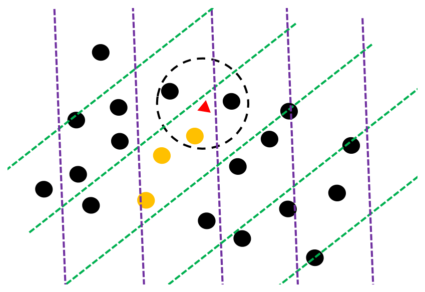
where w is a pre-defined integer and b is a real number chosen uniformly from $[0, w)$. To distinguish these two LSH families, we call h^* as the projected function and h as the hash function.

3.3 Graph based Methods

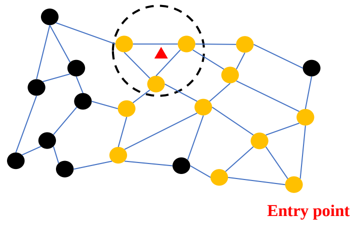
The foundational structure of graph-based methods is a proximity graph, denoted as $G = (V, E)$, where the vertex set V represents all data points in the dataset $\mathcal{D}$, and the edge set E is the collection of all edges between vertexes if the corespondent points are sufficiently close to each other in the original space. There are mainly three strategies to build the proximity graph [38].

- **Cluster & Merge [32].** The dataset $\mathcal{D}$ is clustered into several small groups and the exact proximity graph is built in each group. Then, we obtain the approximate proximity graph by merging the all subgraphs.
- **Iteration [9, 14].** The APG is built from an initial random graph. For each vertex, we iteratively update its neighbors according to the local information. The final APG is obtained when the iteration converges.
- **Consecutive Insertion [30, 31].** In this manner, the graph is built by inserting the point one by one, like that in R-Tree.

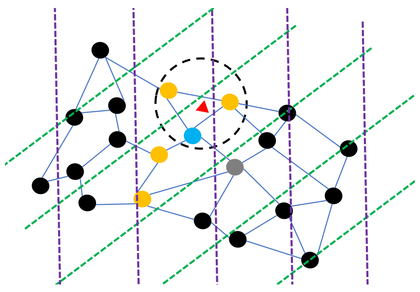
The number of edges directly affects the query performance[38]. Intuitively, the more the out-edges of a vertex, the more candidates and computations need to be performed, and the slower the query processing will be. There are two commonly used neighbor selection strategies to control the distribution of edges, *i.e.,* the simple selection strategy and the heuristic selection strategy. The simple selecting strategy is to choose the closest M neighbors, where M is a pre-defined threshold. In the heuristic selecting strategy, neighbors are chosen based on the distribution so as to preserve their diversity. Take HNSW and NSG for example, if the point o has two edges (o, u) and (o, v) that satisfies $\|o, u\| < \|u, v\| < \|o, v\|$, the edge (o, v) is said to be conflicted to (o, u) and the longer edge (o, v) will be discarded. The behind idea is that edges (o, u) and (o, v) are too similar, and thus there is no need to store both of them.



(a) Search in the LSH-Based Method



(b) Search in the Graph-Based Method



(c) Search in LSH-APG

Figure 1: Difference in the LSH-based method, graph-based method and LSH-APG. The red triangle is the query point q. The three points in the black dashed circle are the 3NN of q. The orange points denote those that are accessed during the search.

3.4 Limitations in the existing ANN methods

In this subsection, we analyze the drawbacks of the LSH-based and graph-based methods and how LSH-APG improves them.

LSH-based methods. Figure 1(a) shows the framework of the LSH-based methods, where the data points are mapped into several hash buckets (parallelograms formed by purple and green lines). Then, the points in the 2-dimensional projected space are indexed by some simple structures, such as hash table, B+-Tree and R-Tree. To answer the ANN query, we only check the points in the bucket where the query point falls, *i.e.,* the 3 orange points in Figure 1(a). Both the LSH family in Equation 1 and 2 satisfy that the collision probability decreases monotonically with the distance between points. Thus, LSH-based methods provide a guarantee of query quality. However, it is hard for these methods to achieve a very high recall due to their simple structures. As shown in the Figure 1(a), 2 of 3-NN results are not found by this LSH index. To find them, it requires to build more LSH indexes, and thus incurs higher query cost.

Graph-based methods. Figure 1(b) shows the framework of the graph-based methods, where each vertex in the graph has 2-4 neighbors. The query procedure begins from the bottom right point (a random entry point) and approaches to the correct results via the greedy search in the APG. The orange points are the points accessed during the search for the 3ANN of q. Although these methods hardly offer a theoretical guarantee, they always perform better than LSH-based methods. Experimental results show that to reach the recall of 0.95, the cost of DB-LSH is about 500 times higher than that of graph-based methods. The bottleneck of graph-based methods, however, comes from their huge construction cost. In the cluster and merge strategy, the subgraphs built for each cluster are disconnected and their graph quality is unsatisfactory since many close point pairs are divided into different clusters. Thus, we need to repeat the cluster and merge operations several times to improve the graph quality, which is time-consuming. In the iteration strategy, the construction cost is high because it requires updating the neighbors for every vertex in each iteration and the number of iterations to convergence increases with the data cardinality. Compared to them, the consecutive insertion strategy is more efficient, but it depends on the specific query strategy. Besides, the neighbor selection strategies have a significant impact on the construction cost. The heuristic selection strategy greatly increases the computational cost since we need to compare every two neighbors. To lower its cost, some methods, such as HNSW [31], only adopt the heuristic selection strategy to cut off the edges of a vertex when its degree reaches a given maximum capacity. The simple selecting strategy is more efficient, but some edges could be similar especially when the data is in a dense region, which incurs the unnecessary computational cost in the query processing.

Our solution. From the above analysis, we can see that LSH indexes are fast to build but difficult to achieve the high recall. On the contrary, graph-based methods perform well on the query processing but suffer from the high construction cost. To address this dilemma, we propose LSH-APG with following the consecutive insertion strategy and simple selection strategy. An LSH framework is used to accelerate the construction of the APG. First, we adopt a suit of lightweight LSH indexes to quickly find a better entry point for the query in the graph, which reduces the number of hops required to terminate the algorithm. As shown in Figure 1(c), we select the closest point to the query point q in the bucket where the q falls as the entry point *i.e.,* the red point, and thus the number of hops is reduced to 2, half of that in Figure 1(b). Then, we design an LSH-based pruning condition to avoid accessing all the neighbors during the search. In the figure, the grey point is one of neighbors of the red point, but we do not need to check it since its distance to q is much larger than that of the red point. With the help of the LSH framework, LSH-APG can be built much faster than the competitors with maintaining the same query performance. We demonstrate the effectiveness of the LSH framework via the theoretical analysis and experimental results.

4 LSH-APG STRUCTURE

LSH-APG consists of two parts, the hash indexes $\mathcal{I}_H$ and the graph index $\mathcal{I}_G$. In this section, we give a detailed description of them.

4.1 Building the LSH indexes

In LSH-APG, a fast LSH framework is required for quickly finding a high-quality entry point. Following the idea in [35], we build our hash indexes $\mathcal{I}_H$ as follows: first, we randomly choose K LSH functions $h_1, \dots, h_K$ as described by Eq. 2. For a point o to be inserted, we compute its K hash values $h_1(o), \dots, h_K(o)$ and consider them as a K-dimensional point $H(o) = (h_1(o), \dots, h_K(o))$. Then,

Algorithm 1: Construction of LSH-APG($\mathcal{D}, T, T'$)

```
Input: Dataset $\mathcal{D}$ and parameter $T, T'$
Output: LSH-APG index $\mathcal{I}_G$ and $\mathcal{I}_H$
1  $\mathcal{I}_H \leftarrow \emptyset, \mathcal{I}_G \leftarrow \emptyset$;
2  for each point $o \in \mathcal{D}$ do
3      candidates $\leftarrow$ call
        kANN-Query($o, \mathcal{I}_G, \mathcal{I}_H, p_\tau = 0.95, T$);
4      for each $e \in$ candidates do
5          $\mathcal{I}_G \leftarrow \mathcal{I}_G \cup \{(o,e),(e,o)\}$;
6          if $e.degree > T'$ then
7              $o_f \leftarrow e$'s furthest neighbor in $\mathcal{I}_G$;
8              $\mathcal{I}_G \leftarrow \mathcal{I}_G - \{(e, o_f)\}$;
9      Insert $o$ into the corresponding LSB-Tree in the $\mathcal{I}_H$;
10 return $\mathcal{I}_H$ and $\mathcal{I}_G$;
```

we transform $H(o)$ into a one-dimensional value $z(H(o))$ via the Z-order curve [7] and store the $z(H(o))$ in a B+-Tree. Repeating this process L times, we build L B+-Trees, which forms $\mathcal{I}_H$. For the convenience, we denote the K LSH functions used the for i-th projected space as $H_i = \{h_{i,1}, \dots, h_{i,K}\}, i = 1, \dots, L$, *i.e.,* there are total $L \times K$ LSH functions used.

4.2 Building the graph index

The graph index $\mathcal{I}_G = (V, E)$ is a directed graph, where $V = \mathcal{D}$ and E is constructed as follows: first, we conduct a kANN query for the incoming point o with $k = T$ in the current graph index. T is a pre-defined integer. Suppose that the found T ANNs are $o_1, \dots, o_T$. Next, we construct two edges for each point $o_i, i = 1, \dots, T$, one the edge $e = (o, o_i)$ from o to o_i, the other an edge $e = (o_i, o)$ from o_i to o. Let $N(o) = \{v | (o, v) \in E\}$ be the set of all neighbors of o and $|N(o)|$ be the degree (or out-degree) of o. Then, for each neighbor o_i of o, we check whether $|N(o_i)|$ reaches the maximum capacity T'. If so, we select the closest T' points to o_i in $N(o_i)$, following the simple neighbor selection strategy. To control the quality of $\mathcal{I}_G$, we set $T' = 2T$ by default,*i.e.,* the degree of each vertex is in the range $[T, 2T]$. LSH-APG with this setting usually achieves a better performance than LSH-APG with $T' = T$. An intuition explanation is that: the real-world dataset is usually distributed unevenly. The data points in dense regions are more likely to have more neighbors, and thus have a higher degree than data points in sparse regions. Therefore, it is hard to choose a unified degree for all data points. The degrees in a range $[T, 2T]$ better fit the data distribution. Algorithm 1 describes the construction of $\mathcal{I}_H$ and $\mathcal{I}_G$ in details.

5 QUERY PROCESSING

We proceed to introduce the ANN query processing algorithm using LSH-APG. The query algorithm not only decides the efficiency and accuracy of ANN search, but also affects the index quality and indexing efficiency.

Algorithm 2: kANN Query($q, \mathcal{I}_G, \mathcal{I}_H, p_\tau, k$)

```
Input: A query point $q$, LSH-APG index $\mathcal{I}_G$ and $\mathcal{I}_H$,
       parameter $p_\tau$, $k$
Output: $k$ nearest points to $q$
1  Compute $q$'s projected values $h_1^*(q), \dots, h_{L\times K}^*(q)$;
2  $EPs \leftarrow$ the set of $k$ approximate nearest points to $q$ in $\mathcal{I}_H$;
3  $V \leftarrow$ the set of visited points during the above search;
4  $R \leftarrow EPs$; //the result set of $k$ best results found so far
5  $m \leftarrow K, P(q) \leftarrow (h_1^*(q), \dots, h_m^*(q))$;
6  $t \leftarrow \sqrt{\chi^2_{p_\tau}(m)}$;
7  while $|EPs| > 0$ do
8      $e_p \leftarrow$ pop the nearest element in $EPs$ to $q$;
9      $R_k \leftarrow$ the furthest points in $R$ to $q$;
10     if $\|e_p, q\| > \|q, R_k\|$ then
11         break;
12     for each $o \in N(e_p)$ do
13         if $o \notin V$ then
14             $V \leftarrow V \cup \{o\}$;
15             if $\|P(q), P(o)\| < t \cdot \|q, R_k\|$ then
16                 Compute $\|q, o\|$;
17                 update $R$ and $EPs$;
18 return $R$;
```

5.1 ANN Query in LSH-APG

Algorithm 2 outlines the ANN query processing, which consists of two parts: find entry points using $\mathcal{I}_H$ (Line 1 - 4) and improve query quality using $\mathcal{I}_G$ (Line 5 - 18). The algorithm starts by computing the hash values of q (Line 1). Then, we can conduct a kANN search for q using $\mathcal{I}_H$ (Line 2). Since the search procedure in $\mathcal{I}_H$ is the same as that in the LSB-Tree [35], we do not give detailed description here due to the space limitation. According to the property of LSH, we can obtain high-quality results in $O(L)$ costs [35]. Next, we take points in EPs as the entry points to conduct a kANN query in $\mathcal{I}_G$. Specifically, in each iteration (hop), the nearest point e_p in EPs to q is popped from EPs (Line 8) and we denote the current found k-th NN as R_k (Line 9). If the distance between q and e_p is already greater than that between q and R_k, the algorithm terminates immediately (Line 10). Otherwise, we continue the search from e_p by checking its neighbors. To further boost query efficiency, we design an LSH-based pruning condition (Line 15) to reduce the number of neighbors checked. That is, for each neighbor o of e_p, if $\|P(q), P(o)\| > t \cdot \|q, R_k\|$, we filter out o without computing the distance $\|q, o\|$. According to Lemma 2, to be introduced in Section 5.2, we can guarantee that $\|q, o\| > \|q, R_k\|$ with the high probability p_τ, and thus it is reasonable to prune o directly. By updating EPs and the result set R iteratively, we get kANN of q finally. Generating entry points via the LSH framework greatly reduces the initial search radius. According to Lemma 8, to be introduced in Section 7, the reduction of the initial search radius can lower the hop numbers required to terminate the algorithm, and thus improve the query efficiency. Moreover, a closer entry point decreases the probability that the query terminates at a local minimal where the final search radius is still very high. The above query strategy satisfies the following conclusion.

Theorem 1. *Assume that the expected edge length in LSH-APG is r_o and for each vertex its neighbors are uniformly distributed around it, the query cost of LSH-APG is expected to depend only on the local intrinsic dimensionality of the dataset. Moreover, the final search radius s is expected to be bounded by $(1+\gamma)r_o$ where $\gamma < 1$.*

Proof. This theorem is derived from Lemma 7 and Theorem 2 in Section 7. It guarantees the query quality and efficiency. □

5.2 LSH-Based Pruning Condition

The query cost of a graph-based method comes from the number of distance computations between q and the neighbors of e_p when accessing e_p. The previous query strategy checks all e_p's neighbors, which is unnecessary and time-consuming since some neighbors are obviously far away from q and unlikely to be the required result. Hence, we design an LSH-based pruning condition as follow to filter out some neighbors that might be far:

$$\|P(q), P(o)\| < \sqrt{\chi^2_{p_\tau}(m)} \cdot d_k, \tag{3}$$

where $P(o)$ is the m-dimensional projected vector defined as in Lemma 1, $\chi^2_p(m)$ is the quantile of $\chi^2(m)$ distribution at p_τ and d_k is the current found k-th best NN result. The intuition of the pruning condition is that when $\|o, q\|$ is greater than d_k, $\|P(q), P(o)\|$ will also be greater than $\sqrt{\chi^2_{p_\tau}(m)} \cdot d_k$ with high probability p_τ. Then, it is reasonable to discard o without computing the distance $\|q, o\|$.

Lemma 2. *With the LSH-based pruning condition, the probability that a point o is filtered increases with $\|q, o\|$. Assume $p_\tau = \frac{1}{2}$ and the current search radius is r, for any $c > 1$, the point whose distance to q is less than r will not be filtered with at least the probability of $\frac{1}{2}$ and we access at most $O(n^\alpha)$ points whose distance to q is greater than cr, where $\alpha = 1 - \frac{9\kappa(c^{-2/3}-1)^2}{4}$ and $\kappa = \frac{m}{\log n}$.*

Proof. The proof is after Lemma 10 in Section 7. □

This lemma indicates the point whose distance to q is less than the search radius r will be checked with at least the probability of $\frac{1}{2}$. On the contrary, the farther a point to q, the more likely it is filtered out. The number of checked points whose distances to q are greater than cr is bounded by $O(n^\alpha)$, which depends on the values of c, n and m. The total number of points filtered by the pruning condition depends on how many "far" points we meet. This pruning condition helps reduce the query cost and still guarantees the query quality. Usually, a large p_τ is used to better balance the query quality and query efficiency. We only prove this lemma with $p_\tau = \frac{1}{2}$ as an example. For other values of p_τ, α is also bounded.

6 EXTENSION

In this section, we provide an efficient update strategy and extend the LSH framework to other graph-based methods.

6.1 Updates

The updates of LSH-APG includes the insertion and deletion. It is simple and natural to insert a new point into LSH-APG since LSH-APG is built via the consecutive insertion strategy. Therefore, we focus on designing an efficient deletion strategy. To delete a point o, we need to discard all the out-edges and in-edges of o in $\mathcal{I}_G$ and remove o from $\mathcal{I}_H$. It is trivial to remove a point from the $\mathcal{I}_H$, but it is challenging to delete in-edges in $\mathcal{I}_G$ since they are not recorded in the graph.

Let $RN(o) = \{v | (v, o) \in E\}$ be the set of reverse neighbors of o in $\mathcal{I}_G$ and $d_m = \max_{v \in RN(o)} \|o, v\|$ be the maximum length. We store the in-degree of o, $|RN(o)|$, and d_m in LSH-APG to facilitate the deletion, which is much space-saving than storing all the in-edges. To delete o, we first mark o and all the out-edges of o as the Deleting status. Then, we conduct a range search in LSH-APG with the search radius d_m from o's closest neighbor in $\mathcal{I}_G$. Once a point u is found, we check whether u is in the $RN(o)$. If so, (u, o) is discarded and the in-degree of o is decreased by one. Moreover, we increase the number of u's neighbors to T' by finding points in neighbors of u's neighbors once the degree of u is less than T. By this manner, we maintain the degree of each vertex in the range $[T, T']$ and reduce the influence of the deletion on the graph quality.

However, d_m could be very large in some cases and it is costly to check all the points whose distance to q is within d_m. Moreover, $\mathcal{I}_G$ is a directed graph and some points in $RN(o)$ is unreachable from o. Hence, we first set a maximum search cost C_{Dm} to control the cost of the range search during the deletion. The larger C_{Dm}, the more likely it is to find an in-edge, but it incurs a higher cost. Then, for an in-edge of o not found in the range search, we leave it to the deletion procedure in the following queries. If it is found during an ANN query later, we discard it and decrease the in-degree of o by one. Once the in-degree of o becomes 0, we discard all the out-edges of o and o itself. Finally, to avoid some in-edges not being found for a long time and thus wasting the space, we also traverse the graph and discard all the edges to be deleted when their amount reaches 10% of the total number of edges in $\mathcal{I}_G$.

The deletion cost of LSH-APG includes the cost of finding o's in-edges, adding new neighbors for some points and C_{Dm}. Due to the property of the NN-graph, *i.e.,* neighbors are more likely to be neighbors of each other [9], $C_{Dm} = C_Q$ can ensure that the most of in-edges are found. To add new neighbors for a point u, it is sufficient to find points in neighbors of e's neighbors rather than conduct a query for u. C_{Dm} is bounded by C_Q and thus we have the following conclusion as follows:

Lemma 3. *The expected deletion cost of LSH-APG is only depends on the local intrinsic dimensionality of the dataset.*

6.2 The Extension of LSH Framework

The LSH framework provides better entry points and allows us not to check all the neighbors, and thus reduces the query cost of LSH-APG. Since the most graph-based methods adopt the similar query strategy as LSH-APG, we can extend our LSH framework on top of them to accelerate the query processing. In addition, the LSH framework can improve the indexing efficiency of these methods under different construction strategies.

For the method with the iteration strategy, we can build an initial graph via LSH indexes, which reduces the number of iteration to convergence. For the consecutive insertion strategy, the benefit from LSH framework can be obtained by improving the query strategy via LSH. For the methods with cluster & merge strategy, such as HCNNG and VRLSH, the LSH indexes can be adopted to partition the data. Then, when building the subgraph in a group, we can only compute the distances of points in the same hash buckets instead of all point pairs in the group since the points in the same buckets are more likely to be close. However, the clustering-based methods are hard to acquire this kind of information. Experimental results show that with the LSH framework, the construction cost of HCNNG can be reduced by 40-90% on SIFT100M (See Section 8.5).

7 QUALITY ANALYSIS

In this section, we provide a theoretical analysis of LSH-APG. First, we give a model to analyze the query cost and the query quality of LSH-APG via C_Q and the final search radius, *i.e.,* the search radius when the query terminates. We find that C_Q and the final search radius depend little on n, which is surprising but in line with our experimental results. With the model, we compute the benefit from the LSH framework. Then, we analyze the indexing, space and query complexities of LSH-APG.

7.1 Analysis of LSH-APG

We aim to find a model to analyze the correctness and cost of LSH-APG. For the correctness, we prove the final search radius is bounded. For the cost, we prove C_Q is independent of n and can be reduced by the LSH framework. To compute C_Q and the final search radius, we consider the the hop numbers l and the final search radius s are the functions of the initial search radius, and denote them as the $l(r)$ and $s(r)$, respectively.

The models of $l(r)$ and $s(r)$. We assume that the length of the edges in $\mathcal{I}_G$ is r_o, and for each vertex, its neighbors are uniformly distributed around it. The hop number l depends on the initial search radius r. Given the entry point e_p with $\|q, e_p\| = r$, we denote as $l(r)$ the hop numbers of query processing starting with e_p. To solve $l(r)$, we need to know how fast the search radius decreases and where the query terminates. We define $p(r)$ as the probability that the query terminates when the search radius is r and $\delta(r) = r - r'$ as the hop length, where r' is the search radius of the next hop.

Lemma 4. *Assume that $l(r) \gg \delta(r)$ and $l(r)$ is differentiable, $l(r)$, $p(r)$ and $\delta(r)$ satisfy the following equation:*

$$[1 - p(r)]\delta(r)l'(r) + p(r)l(r) = 1, \tag{4}$$

where $l'(r)$ is the derivative of $l(r)$ with respect to r.

We can derive a similar conclusion for $s(r)$ as follows.

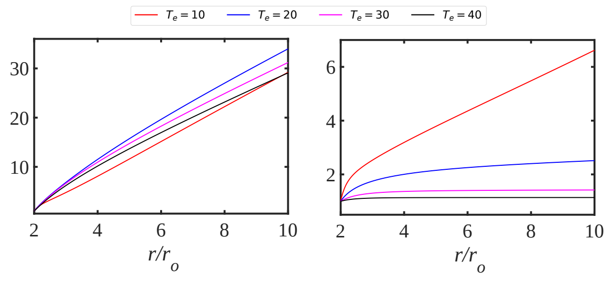
Lemma 5. *Assume that $s(r) \gg \delta(r)$ and $s(r)$ is differentiable, $s(r)$, $p(r)$ and $\delta(r)$ satisfy the following equation:*

$$[1 - p(r)]\delta(r)s'(r) + p(r)s(r) = rp(r), \tag{5}$$

where $s'(r)$ is the derivative of $s(r)$ with respect to r.

Proof. Due to the space limitation, we put the proofs of these two lemmas in our technical report [33]. □

The correctness and the query cost of LSH-APG. It is impossible to obtain the closed-form expressions of $l(r)$ and $s(r)$ from Eq. 4 and 5 since the expressions of $\delta(r)$ and $p(r)$ are extremely complex. Due to the space limitation, the details are in the technical report[33]. Here we give some useful conclusions about them.



(a) $l(r)$ (b) $s(r)/r_o$

Figure 2: The curves of $l(r)$ and $s(r)$

Lemma 6. *Assume that for each vertex its neighbors are uniformly distributed around it, $l(r)$ and $s(r)$ are independent on n and only affected by the local intrinsic dimensionality of the dataset.*

Proof. Due to the space limitation, we put the proof in our technical report [33]. □

The above lemma ensures that $l(r)$ only depends on d_i, and thus we denote l as $O(\varphi(d_i))$. $C_Q = O(Tl)$, so it is also $O(\varphi(d_i))$.

Lemma 7. *The expected query cost in LSH-APG only relies on d_i.*

As for $s(r)$, we aim to further prove that it is close to r_o. $s(r_i) = \int_{r_o}^{r_i} s'(r)dr + s(r_o)$ where r_i is the initial search radius and $s(r_o) < r_o$. Hence, we prove that $\int_{r_o}^{r_i} s'(r)dr$ is bounded and gives the following conclusion:

Theorem 2. *The expected final search radius of LSH-APG, $s(r_i)$, is at most $(1 + \gamma)r_o$, where γ only depends on the d_i and r_o is the expected edge length in LSH-APG. Usually, $\gamma < 1$.*

Proof. Due to the space limitation, we put the proof in our technical report [33]. □

Note that Lemma 7 and Theorem 2 are the general properties for approximate kNN graphs. Based on these two conclusions, we prove Theorem 1. Figure 2 shows the curves of $l(r)$ and $s(r)$ with varying r where T_e is the average degree. From Figure 2(a) we can see that $l(r)$ is nearly linear with r when $r > 2$. We can estimate the value of $\varphi(d_i)$ according to the curve. From Figure 2(b) we can see that $s(r)$ converges to the maximum value s_{max} when r increases and a larger T_e incurs a smaller s_{max}. For a given LID, we can reduce the value of s_{max} to at most $2r_o$ by increasing T_e, which is in line with $\gamma < 1$.

The benefit from $\mathcal{I}_H$. Since $l(s)$ depends on the initial search radius r_i. By using $\mathcal{I}_H$, we can obtain the closer entry point o_h, which essentially reduces the initial search radius. Assume that $\|q, o_h\| = r_1$ and a random entry point has the distance r_2 to q $(r_0 < r_1 < r_2)$. Then, the benefit from $\mathcal{I}_H$ is $B_{\mathcal{I}_H} = l(r_2) - l(r_1)$. To demonstrate the $\mathcal{I}_H$ is effective, we first prove the results returned from $\mathcal{I}_H$ is good enough.

Lemma 8 (Theorem 1 of [35]). *By setting $K = O(\log n)$ and $L = n^\rho$ where $\rho = O(1/c')$, $\mathcal{I}_H$ can answer a c'^2-ANN $(c' \geq 2)$ in $O(dL)$ query cost with at least constant probability of $1/2 - 1/e$.*

The lemma indicates that we can find a c'^2-ANN result with a small query cost $O(dL)$. As stated in [35], even single LSB-Tree is also expected to return results with high quality. In our method,

$c' = 2$ and we set $K = O(1)$ and $L = O(1)$ following the mainstream LSH methods [29, 37].

Lemma 9. *Let $\Delta r = r_2 - r_1$ be the reduction of the initial search radius compared to the random entry point, the expected $B_{\mathcal{I}_H}$ is $\frac{\Delta r}{\lambda r_o}$ where $\lambda < 1$ and depends on LID and the average degree T_e.*

Proof. Due to the space limitation, we put the proof in our technical report [33]. □

This lemma indicates that the benefit of the entry point selection is $O(\Delta r)$, which is in line with Figure 2(a). When $r > r_1$, $\delta(r)$ is nearly a constant and the search radius decreases slowly, which incurs a large query cost in the graph-based methods. On the contrary, although the LSH-based methods are hard to achieve as a high query performance as graph-based methods, they can quickly find a point o with $\|q, o\| = r_1 \le c'r$, and thus improve the query performance of LSH-APG.

7.2 Analysis of the Pruning Condition

Before proving Lemma 2, we introduce the following conclusion:

Lemma 10 (Lemma 4 in [41]). *Given a query q, an approximation ratio c and parameter t, we define the following two events:*

- ***E1:** For a point o that $\|q, o\| \le r$, its projected distance to q, $\|P(q), P(o)\|$, is smaller than tr.*
- ***E2:** There are fewer than βn ($\beta > \alpha_2$) points whose distances to q exceed cr but projected distances to q are smaller than tr.*

Then, we have that the probability that E1 occurs is at least α_1, and the probability that E2 occurs is at least $1 - \frac{\alpha_2}{\beta}$, where $\alpha_1 = F(t^2; m)$ and $\alpha_2 = F(t^2/c^2; m)$.

Now, we prove the Lemma 2.

Proof. According to the Lemma 10, those points whose distance to q exceed cr are very likely to have projected distances to q larger than tr. The total number of this kind of points is less than $2\alpha_2 n$ and the total number of points to be verified is at most $2\alpha_2 n + k$. We try to bound α_1 and $2\alpha_2 n + k$ where $p_\tau = \alpha_1$. By using the Wilson–Hilferty transformation [22, 39], for a random variable $X \sim \chi^2(m)$, $\sqrt[3]{X/k}$ is approximately distributed with $N(1 - \frac{2}{9m}, \frac{2}{9m})$, and thus $Y = \frac{\sqrt[3]{X/k} - (1 - \frac{2}{9m})}{\sqrt{\frac{2}{9m}}} \sim N(0, 1)$. Consider $t^2 = m(1 - \frac{2}{9m})^3$, $\alpha_1 = \Pr[X \le m(1 - \frac{2}{9m})^3] = \Pr[Y \le 0] = 0.5 = p_\tau$, which is coincident with our setting. Likewise, $\alpha_2 = \Phi(c_1(\sqrt{m} - \frac{2}{9}\sqrt{1/m}))$, where $c_1 = \sqrt{\frac{9}{2}}(c^{-2/3} - 1) < 0$ and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Next, we prove that α_2 can be bounded by $\exp(-\frac{c_1^2 m}{2})$. Denote $g(u) = \Phi(c_1(u - \frac{2}{9u})) - \exp(-\frac{c_1^2 u^2}{2})$, it is easy to demonstrate that $g(u)$ first decreases and then increases with u in $(0, +\infty)$, so $g(\sqrt{m}) < \max\{g(0), g(+\infty)\}$. $g(0) = g(+\infty) = 0$ and thus $g(\sqrt{m}) < 0$, which implies $\alpha_2 < \exp(-\frac{c_1^2 m}{2})$. Considering that $m = \kappa \log n$, we have $\exp(-\frac{c_1^2 m}{2}) = n^{-\frac{\kappa c_1^2}{2}}$ and the total number of points to be verified is at most $2n^{1-\frac{\kappa c_1^2}{2}} + k$. □

Table 2: Summary of Datasets

Datasets	Cardinality	Dim.	LID.	Size (GB)
MNIST	60,000	784	12.7	0.184
Deep1M	1,000,000	256	26.0	1.00
Gauss10M	10,000,000	32	26.3	1.19
Rand10M	10,000,000	32	23.9	1.19
Gist1M	1,000,000	960	36.2	3.58
SIFT10M	10,000,000	128	22.0	4.77
SIFT100M	100,000,000	128	23.7	47.7
Tiny80M	79,302,017	384	44.6	113

7.3 Complexity Analysis

Following the mainstream graph-based methods [14, 31, 32], we consider T as a constant. According to Theorem 1 and Lemma 3, we have the following conclusion:

Theorem 3. *LSH-APG has the space complexity of $O(n)$ and can be built in $O(nd\varphi(d_i))$ time, where d_i is the LID of $\mathcal{D}$ and $\varphi(d_i)$ is the expectation of C_Q. The costs of query, insertion and deletion operations using LSH-APG are all $O(d\varphi(d_i))$.*

8 EXPERIMENTAL STUDY

In this section, we conduct extensive experiments on real-world and synthetic datasets to provide a comprehension analysis on LSH-APG. We implement LSH-APG[1] and the competitors in C++ compiled with g++ using -Ofast optimization and openMP for parallelism. All experiments are run on a Ubuntu server with 2 Intel(R) Xeon(R) Gold 5218 CPUs @ 2.30GHz (64 threads) and 254 GB RAM.

8.1 Experimental Settings

Datasets. We employ 8 datasets varying in cardinality and dimensionality, whose information is summarized in Table 2 in the ascending order of their sizes. **LID.**[2] in the table is used to measure the difficulty of answering NN queries in the dataset [1]. A larger LID implies that it is harder to find NN on the dataset. Among the datasets, Six are real-world datasets widely used in NN search methods[14, 31, 37, 38], including our competitors. The rest 2 synthetic datasets, **Rand10M** and **Gauss10M** are generated from uniform distribution $U(-1, 1)$ and Gaussian distribution $N(0, 1)$ on each dimensionality independently. For NN queries, we randomly select 100 points as query points and remove them from the datasets.

Competitors. To demonstrate the indexing and query performance of LSH-APG, we compare it with the best existing NN search methods, including LSH-based and graph-based methods. Among LSH-based methods, DB-LSH [37] is proven to have the lowest query complexity. Among graph-based methods, a comprehensive experimental comparison of graph-based methods [38] has shown that HNSW [31], HCNNG [32] and NSG [14] are always the best three ones in terms of the query performance. Therefore, we choose DB-LSH, HNSW, HCNNG and NSG as our competitor algorithms.

Parameter Settings. We consider the (c, k)-ANN queries with $k = 50$ for all algorithms in the default settings. Following the settings in the paper or source code of our competitors, we adopt the fixed parameters for each algorithm as shown in Table 3. For HCNNG,

[1]https://github.com/LSH-APG/LSH-APG-Code

[2]There is no unified method to compute the LID and we compute it according to the Definition 1 in [1] with x being the average distance between the query points and their 50-th NN.

Table 3: Parameter Settings of Algorithms

Algorithm	Parameters
LSH-APG	$K = 16, L = 2, T = 24, T' = 2T, p_\tau = 0.95$
DB-LSH	$c = 1.5, K = 12, L = 5$
HNSW	$M = 48, ef = 80$
HCNNG	$MC = 500, NC = 10$
NSG	$L = 40, R = 50, C = 500$

MC is the maximum size of the cluster and NC is the number executions of hierarchical clustering procedures.

Evaluation Metric. We compare the algorithms from four aspects: indexing quality, indexing efficiency, query quality and query efficiency. The former two are the parts of the indexing performance and the latter two are the parts of the query performance. We report the average degrees, the range of degrees, the standard deviation of degrees of each graph index and evaluate the quality of them by using the normalized maximum common subgraph (NMCS). Maximum common subgraph (MCS) is used to measure the similarity of two graphs [11]. Here, we adopt **NMCS**, a derived definition from MCS, to compute the similarity between a graph index G for the ANN query and the exact NN graph, which can reflect the quality of the graph. NMCS is defined as follow,

Definition 3 (NMCS (Normalized maximum common subgraph)). *Let $G = (V, E)$ be an APG and $G_E = (V, E')$ be the exact NN graph of V that satisfies:*

- *For any a point $v \in V$, $|G(v)| = |G_E(v)|$ where $G(v)$ is the neighbors of v in G;*
- *Let $k' = |G_E(v)|$. $G_E(v)$ is the k'-NN result of v in $V - v$.*

Then

$$NMCS = \frac{\sum_{v \in V} |G(e) \cap G_E(e)|}{\sum_{v \in V} |G(e)|} \tag{6}$$

Since the exact NN graph is nearly impossible to compute, we randomly choose 200 vertexes in V to estimate the NMCS.

We evaluate the query quality via recall. Given a query point q, for a (c, k)-ANN query, assume that the algorithm return the set $R = \{o_1, \dots, o_k\}$ and the exact kNN of q is $R^* = \{o_1^*, \dots, o_k^*\}$, then recall are defined as follows [37]

$$Recall = \frac{|R \cap R^*|}{k}. \tag{7}$$

Since all algorithms except DB-LSH adopt openMP parallelizing, their running times are unstable and liable to be affected by the current system status. Therefore, we use the number of distance computations per insertion (**CPI**) and the number of distance computations per query (**CPQ**) instead of the running time to evaluate the indexing and query efficiency, respectively. For fairness, we consider the cost of computing hash values as a part of distance computations in LSH-APG.

8.2 Self Evaluation

In this subsection, we demonstrate the effectiveness of the LSH framework on LSH-APG and analysis the effect of p_τ.

Evaluation of the LSH framework. To demonstrate the functionality of the LSH framework, we denote by Naive-G our solution without using the LSH framework and compare it with LSH-APG. Here we report the results on DEEP1M and SIFT100M for brevity,

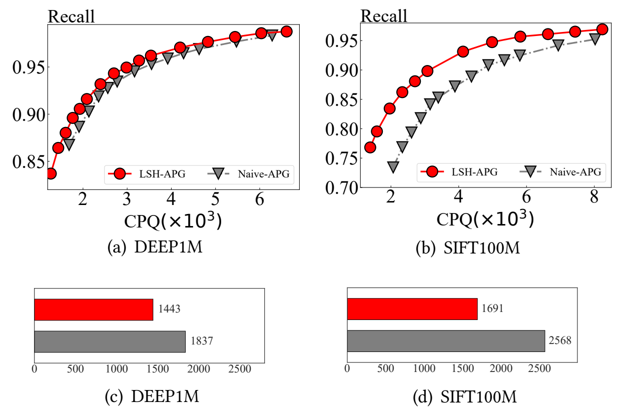


Figure 3: Comparison on LSH-APG and Naive-G

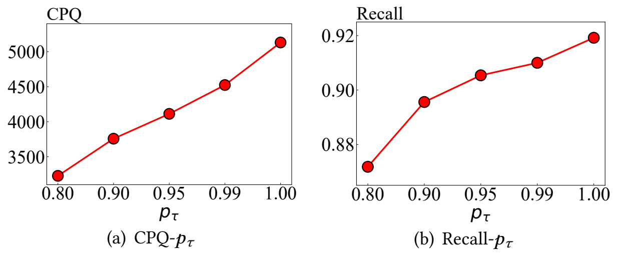


Figure 4: Performance of LSH-APG when varying p_τ

as shown in Figure 3. To compare their query performance, we plot their Recall-CPQ curves. The results show that to reach a given recall, LSH-APG reduce about 20% CPQ on DEEP1M and 50% CPQ on SIFT100M, which demonstrates the improvement of the LSH framework on query processing. To compare their indexing performance, we notice that LSH-APG and Naive-G have nearly the same graph indexes and LSH framework only affects the indexing cost. Thus, we report the CPIs of LSH-APG and Naive-G. The results show that LSH-APG (the red bar) reduce about 20% CPI on DEEP1M and 60% CPI on SIFT100M, respectively. It indicates that LSH framework plays also an important role in the indexing phase.

Parameter study on p_τ. We discuss the effect of p_τ on query performance by varying p_τ in range $\{0.8, 0.9, 0.95, 0.95, 1.0\}$. Here, we only show results on DEEP1M for brevity. p_τ controls the pruning threshold in the query phase and a smaller p_τ makes it more likely to filter out a neighbor. As shown in Figure 4, both the recall and CPQ increases with p_τ, which indicates that the pruning condition in LSH-APG can reduce the query cost but slightly damages the query quality to some extent because it inevitably filter out some exact results. Taking both query efficiency and quality into consideration, we set $p_\tau = 0.9$.

8.3 Evaluation of Indexing Performance

We study the index quality and indexing efficiency of all algorithms with the default settings. The results are shown in Table 4. In this table, μ and σ denote the average of degrees and the standard deviation of degrees. **Range** is the interval where degrees distribute. NSG fails to build the index on Tiny80M due to the out of memory issue, so we do not report its results (including the query results)

Table 4: Overview of index information.

		(μ, σ)	Range	NMCS	CPI			(μ, σ)	Range	NMCS	CPI
MNIST	**LSH-APG**	(37.51,9.19)	[24,48]	0.7655	583.59	**GIST1M**	**LSH-APG**	(32.43,9.33)	[24,48]	0.4159	1864.0
	HNSW	(17.21,8.83)	[1,48]	0.4006	928.50		**HNSW**	(10.07,9.86)	[1,48]	0.1159	2144.3
	NSG	(18.19,5.72)	[1,50]	0.4499	3801.7		**NSG**	(16.50,12.53)	[1,77]	0.2929	7027.1
	HCNNG	(11.39,3.03)	[1,26]	0.5353	4801.9		**HCNNG**	(17.50,5.63)	[1,30]	0.1375	5091.8
	DB-LSH	-	-	-	60		**DB-LSH**	-	-	-	60
DEEP1M	**LSH-APG**	(35.84,9.42)	[24,48]	0.5915	1442.9	**SIFT10M**	**LSH-APG**	(36.61,9.31)	[24,48]	0.5194	1476.8
	HNSW	(23.55,10.84)	[1,48]	0.3036	2138.8		**HNSW**	(25.84,10.80)	[1,48]	0.3018	2578.5
	NSG	(20.96,9.17)	[1,50]	0.4510	6291.6		**NSG**	(21.11,9.29)	[1,51]	0.4143	4782.5
	HCNNG	(17.49,4.35)	[2,30]	0.2786	5091.8		**HCNNG**	(17.83,4.09)	[1,30]	0.2163	3341.7
	DB-LSH	-	-	-	50		**DB-LSH**	-	-	-	60
Gauss10M	**LSH-APG**	(32.47,9.21)	[24,48]	0.3645	2266.4	**SIFT100M**	**LSH-APG**	(36.31,9.36)	[24,48]	0.5110	1691.3
	HNSW	(25.43,13.00)	[1,48]	0.2463	8776.5		**HNSW**	(26.68,11.05)	[24,48]	0.2577	3262.2
	NSG	(24.31,13.95)	[1,51]	0.3365	7279.9		**NSG**	(26.84,12.07)	[1,52]	0.3734	7024.7
	HCNNG	(19.68,6.09)	[6,30]	0.0793	3341.8		**HCNNG**	(18.18,4.17)	[1,30]	0.1719	4164.7
	DB-LSH	-	-	-	50		**DB-LSH**	-	-	-	60
Rand10M	**LSH-APG**	(35.76,9.21)	[24,48]	0.5508	2159.2	**Tiny80M**	**LSH-APG**	(32.18,9.15)	[24,48]	0.3159	2494.6
	HNSW	(33.92,9.41)	[1,48]	0.4544	9547.2		**HNSW**	(13.34,10.97)	[1,48]	0.0698	3567.6
	NSG	(36.02,11.78)	[2,50]	0.4126	6573.0		**NSG**	/	/	/	/
	HCNNG	(19.58,4.92)	[7,30]	0.1005	3341.8		**HCNNG**	(17.75,5.99)	[1,30]	0.1436	6380.3
	DB-LSH	-	-	-	60		**DB-LSH**	-	-	-	60

on Tiny80M. From the table we have the following observations when comparing each evaluation metric:

(1) LSH-APG have the the large average degrees. It is because LSH-APG adopt the simple neighbor selection strategy and does not discard the similar edges. HNSW has the extremely low degree on Gist1M and Tiny80M, although these two datasets have high LIDs. This abnormal phenomenon is caused by the heuristic neighbor selection and will damage the query performance to some extent. Comparing the degree among different datasets, we find the distribution of degrees on LSH-APG changes little but the distributions on other graph indexes do not. It is hard to say which is better.

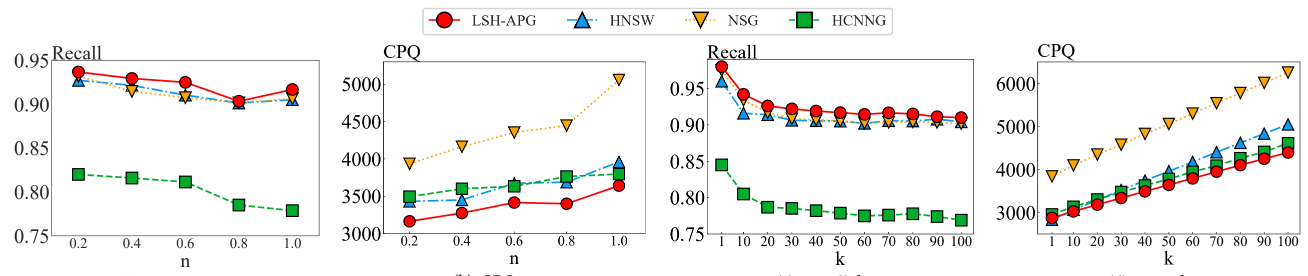
(2) LSH-APG always has the highest NMCS among the graph based algorithms, which indicates it is more similar as an exact NN graph and has the highest-quality edges, which benefits the query quality since it is more likely that we find the exact NN in LSH-APG via greedy expansion.

(3) DB-LSH achieves the smallest CPI among the graph-based methods on all datasets. The CPI of LSH-based methods comes from the cost of computing hash values and the number of hash functions will not be very high. So, LSH indexes can be built much faster than APGs. Among the graph-based algorithms, LSH-APG has the smallest CPI and only HNSW's CPI is comparable with it. NSG and HCNNG both have the larger CPIs. The reason of this phenomenon is that: First, LSH-APG avoids many unnecessary distance computation by the better entry point selection and LSH-based pruning condition when inserting a point. Second, compared to HNSW, LSH-APG does not adopt the time-consuming heuristic selection strategy like that in HNSW or NSG. In NSG, half the CPI comes from building an approximation kNN graph with a large degree K via NN-Descent strategy. Moreover, the CPI of building the NSG in the base of an NNG is also high due to the heuristic selection strategy. In HCNNG, it requires computing the distance between any two points in a group and executes hierarchical clustering procedures multiple times, which incurs the large CPI.

(4) Comparing the NMCS and CPI among different datasets, we can find all the algorithms follows the similar tendency. For example, LSH-APG has the smaller CPI and higher NMCS on DEEP1M than that on GIST1M. Then, HNSW, NSG and HCNNG all have the smaller CPI and higher NMCS on DEEP1M than that on GIST1M, respectively. Moreover, the CPI and NMCS of each algorithm vary with the datasets. For LSH-APG, the CPI increase gradually with the data cardinality when comparing that on SIFT10M and SIFT100M. The CPI and NMCS of LSH-APG are greatly influenced by LID. A larger LID incurs a higher CPI and a smaller NMCS. For other algorithms, their NMCS and CPI are influenced by data cardinality and LID. Astonishingly, HNSW has the highest CPIs on Gauss10M and Rand10M (8776.5 and 9547.2), which is even much higher than that on SIFT100M and Tiny80M (3262.2 and 3567.6). It indicates HNSW is hard to process the random dataset.

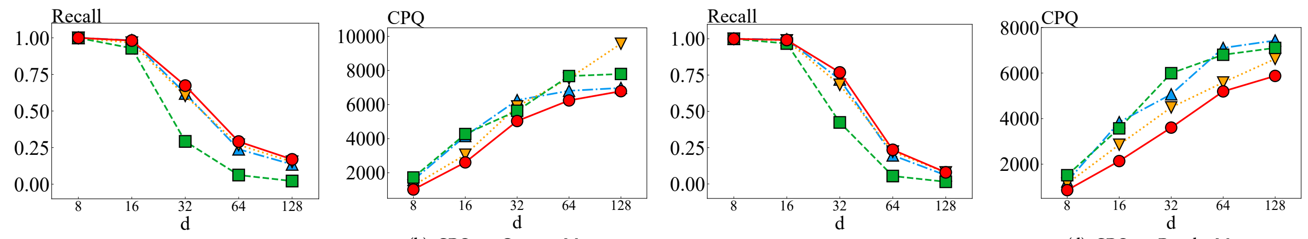
8.4 Evaluation of Query Performance

In these sets of experiments, we study the query performance of all algorithms. We first study how the characteristics of dataset, *i.e.,* data cardinality n and dimensionality d, affect the query performance. Then, we study the effect of k. Finally, we analysis the trade-off between the query quality and query efficiency by increasing the number of checked points. Since DB-LSH always needs the much larger query cost to reach the similar recall than the graph-based methods. For example, the CPQ of DB-LSH to reach a recall of 0.95 is about $8M$, which is 500 times higher than the worst graph-based methods. We do not report its results in the rest experiments.



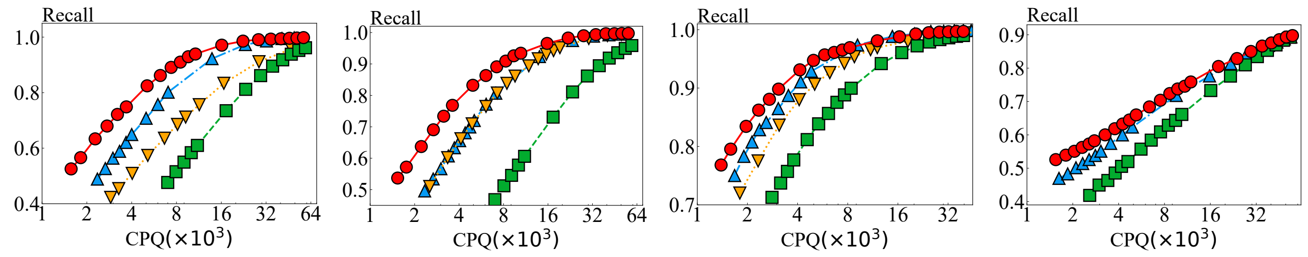
(a) Recall-n (b) CPQ-n (c) Recall-k (d) CPQ-k

Figure 5: Performance on SIFT100M when Varying n and k



(a) Recall on Gauss10M (b) CPQ on Gauss10M (c) Recall on Rand10M (d) CPQ on Rand10M

Figure 6: Performance on Rand10M and Gauss10M when Varying d



(a) Gauss10M (b) Rand10M (c) SIFT100M (d) Tiny80M

Figure 7: Recall-CPQ Curves on all Datasets

8.4.1 ***Effect of*** n. By randomly selecting some points from the original dataset, we compare the query performance of all algorithms in the default settings. Limited by the space, we only report the result on SIFT100M, as shown in Figure 5(a)-(b). The range of the cardinality n is $\{0.2N_0, 0.4N_0, \ldots, N_0\}$, where N_0 is the cardinality of the original datasets. As shown in the figures, each algorithm has the larger CPQ and nearly the smaller recall when n increases. But the increase on CPQ of LSH-APG is rather slight, which is in line with our conclusion that the query cost LSH-APG is less affected by the data cardinality.

8.4.2 ***Effect of*** k. We compare the query performance of all algorithms when varying k of (c, k)-ANN query in $\{1, 10, 20, \ldots, 100\}$. Limited by the space, we only report the result on SIFT100M, as shown in Figure 5(c)-(d). From the figures we can see that CPQ of each algorithm increases nearly linearly with k but LSH-APG has the smallest slope. Moreover, LSH-APG always achieves the smallest CPQ and the highest recall, which indicates it performs best on query processing among the competitor algorithms. NSG and HNSW achieves the surprisingly close recall and their CPQ-k curves also have about the same slope, which further demonstrate the similarity between them.

8.4.3 ***Effect of*** d. In the default settings, we compare the query performance of all algorithms when varying the dimensionality d of datasets. Due to the unbalanced distribution in real-world datasets, it is not meaningful to pick up parts of dimensionality on them. So, we only compare the algorithms on two synthetic datasets, Rand10M and Gauss10M with the range of d $\{8, 16, 32, 64, 128\}$. The results are shown in Figure 6. As shown in the figures, the CPQ of each algorithm increases with d. The increasing tendencies of LSH-APG, HNSW and HCNNG are sublinear but that of NSG are nearly linear, which indicates NSG is more likely to be affected by the dimensionality. As for the recall, we surprisingly find that the recall of each algorithm decreases very rapidly with d. When d is 8 or 16, all algorithms can reach the recall of nearly 1.0; when d comes to 32, recalls all declines to about 0.6 on Gauss10M and 0.75 on Rand10M; when d reaches 64 or even 128, recalls have gone down to less than 0.3. This results show that the effect of the dimensionality on the recall is much greater than that of the cardinality. This result can be explained by the "curse of dimensionality",*i.e.,* when the dimensionality is large enough, the distance between any two points is almost identical and it becomes very difficult to differentiate the NNs and other points for any query. The "curse of dimensionality"

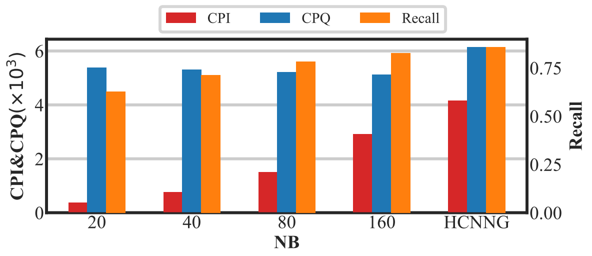


Figure 8: Performance of HCNNG and its LSH Extension

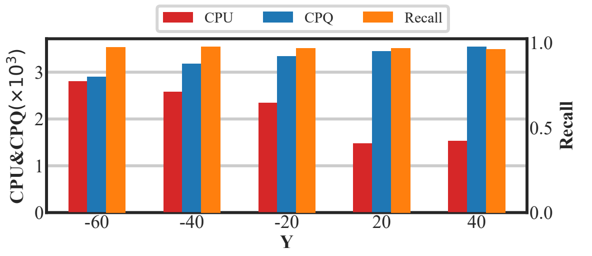


Figure 9: Performance on DEEP1M when Updating

phenomenon becomes very obvious on the random datasets than that on the real-world datasets, even when d is only 64 or 128. This is because with the same dimensionality, the random datasets have the much higher LID than the real-world datasets.

8.4.4 ***Recall-CPQ Curve***. For any ANN methods, a more accurate result can be obtained by increasing the number of checked points, which consequently damages the query efficiency. Therefore, there is a trade-off between the query quality and query efficiency. In this set of experiments, we analyze the trade-off between them via the Recall-CPQ Curve. To reach the same target recall, the algorithm with the smaller CPQ has a higher query performance. Figure 7 present the results on four datasets.

From the figures we have the following observations: (1) As the CPQ increases, all algorithms achieve the higher recall, which is in line with the philosophy of ANN methods, *i.e.,* trading accuracy for efficiency. Moreover, the required CPQ to reach a given recall *Rec* increases nearly exponentially with *Rec*, which indirectly demonstrates that finding the exact NNs is extremely hard and time-consuming. (2) Among all algorithms, LSH-APG requires the smallest CPQ to reach the same recall, which indicates LSH-APG achieves the best trade-off between the query quality and query efficiency. As analyzed in the previous part, the advantages of LSH-APG are due to the better entry point selection found in the LSH indexes and the pruning condition that enables us to filter out some neighbors. (3) NSG and HNSW always exhibit the similar tendency and even nearly the same curve on Rand10M, which demonstrates that they may have the similar structure since they adopt the totally identical heuristic selection strategy (This is proven in [38]). However, in the most cases, HNSW is the better algorithm. (4) HCNNG always achieves the worst query performance. On Gauss10M and Rand10M, to reach the same recall, HCNNG requires nearly 4 times as large CPQ as LSH-APG. This can be explianed that: HCNNG adopts the cluster-based approach to construct subgraphs. However, the data in the high-dimensional space are not "well" clustered and we could not find proper clusters.

8.5 Evaluation of the LSH Extension

In this experiment, we take HCNNG as example and evaluate its LSH extension, LSH-HCNNG, on SIFT100M. As said in Section 6.1, we adopt the LSH indexes to partition the data and only compute the distance of the points in the same hash buckets for LSH-HCNNG. We set $MC = 500$ and $NC = 20$ for LSH-HCNNG and vary the number of points in a bucket, NB, in $\{25, 40, 80, 160\}$. The results are reported in Figure 8. We can see LSH-HCNNG has a much smaller CPI than HCNNG. When $NB = 20$, its CPI is only 10% of HCNNG's. Even though with $NB = 160$, LSH-HCNNG has about a 40% smaller CPI than HCNNG. Moreover, the CPQ of LSH-HCNNG is also smaller due to the better entry point and the pruning condition. However, the recall of LSH-HCNNG is a bit lower than HCNNG, especially when $NB = 20, 40$. This is because LSH-HCNNG only check a small part of point pairs in a group and build an approximate PG; while HCNNG build the exact PG (the minimal spanning tree) in a group and has a higher graph quality. Compared to the huge improvement on CPQ and CPI, the loss on the recall is acceptable and the LSH framework is proven to be effective for HCNNG.

8.6 Evaluation of Updating

We follow [18] and evaluate the update performance of LSH-APG via a batch insertion or deletion. Given the graph index of LSH-APG, $\mathcal{I}_G = (V_0, E_0)$, we denote a batch insertion (resp. deletion) of the APG as $Y\%$ update where $Y = \frac{|V|-|V_0|}{|V_0|}$ and $|V|$ is the number of points in the graph after insertion or deletion. In this set of experiments, we study the the effect of Y on DEEP1M with $V_0 = 600K$ when varying Y in $\{-60, -40, -20, 20, 40\}$ ($Y < 0$ denotes the deletion operation) and report the query performance and the cost per update (CPU) on Figure 9. As for the updating cost, the insertion cost is obvious smaller than the deletion cost and 40% insertion has the higher cost than 20% insertion. The deletion cost increases slightly with $|Y|$, which can be explained that the total number of points in the graph decreases and The probability that a vertex connecting to the points to be deleted increases, and thus it takes more time to add the new edges. For the query performance, we find the recall is steady and CPQ gradually increases with Y since the number of the points increases. It indicates that the update operations do not damage the query performance.

9 CONCLUSION

In this paper, we have proposed a novel approach, called LSH-APG, to efficiently build an APG and facilitate ANN query processing in high-dimensional spaces with quality guarantees. LSH-APG handles the high construction cost issue in the graph-based methods by designing an efficient and accurate LSH-based query strategy to consecutively insert data points based on their nearest neighbors in the APG. A high-quality entry point selection technique and LSH-based pruning condition have been developed to reduce the number of points to be checked in the search. The expected query cost has been proven to be less affected by dataset cardinality, allowing us to simultaneously reduce the query processing time and the indexing time. Therefore, LSH-APG is well positioned to deal with large-scale datasets. In addition, It is proven that LSH-APG can be maintained incrementally in a low cost as the dataset evolves and well adapted to the other graph-based methods. A thorough range of experiments showed that LSH-APG can reduce the construction cost by an average of 40% compared to the best competitors, HNSW, while still maintaining the best query performance.

REFERENCES

[1] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. 2015. Estimating Local Intrinsic Dimensionality. In *KDD*. 29–38.

[2] Alexandr Andoni and Piotr Indyk. 2016. LSH Algorithm and Implementation (E2LSH). https://www.mit.edu/~andoni/LSH

[3] Martin Aumüller, Erik Bernhardsson, and Alexander John Faithfull. 2020. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Inf. Syst.* 87 (2020).

[4] Mahendra Awale and Jean-Louis Reymond. 2019. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* 59, 1 (2019), 10–17.

[5] Brankica Bratic, Michael E. Houle, Vladimir Kurbalija, Vincent Oria, and Milos Radovanovic. 2018. NN-Descent on High-Dimensional Data. In *WIMS*. 20:1–20:8.

[6] T. Tony Cai, Jianqing Fan, and Tiefeng Jiang. 2013. Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.* 14, 1 (2013), 1837–1864.

[7] H. K. Dai and Hung-Chi Su. 2003. On the Locality Properties of Space-Filling Curves. In *ISAAC (Lecture Notes in Computer Science)*, Vol. 2906. 385–394.

[8] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on Computational Geometry*. 253–262.

[9] Wei Dong, Moses Charikar, and Kai Li. 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In *WWW*. 577–586.

[10] Carlos Eiras-Franco, David Martínez-Rego, Leslie Kanthan, César Piñeiro, Antonio Bahamonde, Bertha Guijarro-Berdiñas, and Amparo Alonso-Betanzos. 2020. Fast Distributed *k*NN Graph Construction Using Auto-tuned Locality-sensitive Hashing. *ACM Trans. Intell. Syst. Technol.* 11, 6 (2020), 71:1–71:18.

[11] Mirtha-Lina Fernández and Gabriel Valiente. 2001. A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognit. Lett.* 22, 6/7 (2001), 753–758.

[12] Cong Fu and Deng Cai. 2016. EFANNA : An Extremely Fast Approximate Nearest Neighbor Search Algorithm Based on kNN Graph. *CoRR* abs/1609.07228 (2016).

[13] Cong Fu, Changxu Wang, and Deng Cai. 2021. High Dimensional Similarity Search with Satellite System Graph: Efficiency, Scalability, and Unindexed Query Compatibility. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3067706

[14] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai. 2019. Fast Approximate Nearest Neighbor Search With The Navigating Spreading-out Graph. *PVLDB* 12, 5 (2019), 461–474.

[15] Junhao Gan, Jianlin Feng, Qiong Fang, and Wilfred Ng. 2012. Locality-sensitive hashing scheme based on dynamic collision counting. In *SIGMOD*. 541–552.

[16] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity Search in High Dimensions via Hashing. In *VLDB*. 518–529.

[17] Ben Harwood and Tom Drummond. 2016. FANNG: Fast Approximate Nearest Neighbour Graphs. In *CVPR*. IEEE Computer Society, 5713–5722.

[18] Kai Huang, Huey-Eng Chua, Sourav S. Bhowmick, Byron Choi, and Shuigeng Zhou. 2021. MIDAS: Towards Efficient and Effective Maintenance of Canned Patterns in Visual Graph Query Interfaces. In *SIGMOD Conference*. 764–776.

[19] Qiang Huang, Jianlin Feng, Yikai Zhang, Qiong Fang, and Wilfred Ng. 2015. Query-Aware Locality-Sensitive Hashing for Approximate Nearest Neighbor Search. *PVLDB* 9, 1 (2015), 1–12.

[20] Piotr Indyk and Rajeev Motwani. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *STOC*. 604–613.

[21] Jerzy W. Jaromczyk and Miroslaw Kowaluk. 1991. Constructing the relative neighborhood graph in 3-dimensional Euclidean space. *Discret. Appl. Math.* 31, 2 (1991), 181–191.

[22] NL Johnson, S Kotz, and N Balakrishnan. 1994. Chi-squared distributions including chi and Rayleigh. *Continuous univariate distributions* 1 (1994), 415–493.

[23] Rolf Klein. 2016. Voronoi Diagrams and Delaunay Triangulations. In *Encyclopedia of Algorithms*. 2340–2344.

[24] Cornelius Lanczos. 1964. A precision approximation of the gamma function. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 1, 1 (1964), 86–96.

[25] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. 2020. Approximate Nearest Neighbor Search on High Dimensional Data - Experiments, Analyses, and Improvement. *IEEE Trans. Knowl. Data Eng.* 32, 8 (2020), 1475–1488.

[26] Peng-Cheng Lin and Wan-Lei Zhao. 2019. On the Merge of k-NN Graph. *CoRR* abs/1908.00814 (2019).

[27] Yingfan Liu, Jiangtao Cui, Zi Huang, Hui Li, and Heng Tao Shen. 2014. SK-LSH: An Efficient Index Structure for Approximate Nearest Neighbor Search. *PVLDB* 7, 9 (2014), 745–756.

[28] Kejing Lu and Mineichi Kudo. 2020. R2LSH: A Nearest Neighbor Search Scheme Based on Two-dimensional Projected Spaces. In *ICDE*. IEEE, 1045–1056.

[29] Kejing Lu, Hongya Wang, Wei Wang, and Mineichi Kudo. 2020. VHP: Approximate Nearest Neighbor Search via Virtual Hypersphere Partitioning. *Proc. VLDB Endow.* 13, 9 (2020), 1443–1455.

[30] Yury Malkov, Alexander Ponomarenko, Andrey Logvinov, and Vladimir Krylov. 2014. Approximate nearest neighbor algorithm based on navigable small world graphs. *Inf. Syst.* 45 (2014), 61–68.

[31] Yury A. Malkov and Dmitry A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (2020), 824–836.

[32] Javier Alvaro Vargas Muñoz, Marcos André Gonçalves, Zanoni Dias, and Ricardo da Silva Torres. 2019. Hierarchical Clustering-Based Graphs for Large Scale Approximate Nearest Neighbor Search. *Pattern Recognit.* 96 (2019).

[33] Technical Report. 2022. Available at:. https://github.com/LSH-APG/Technical-Report

[34] Bo Tang and Haibo He. 2015. ENN: Extended Nearest Neighbor Method for Pattern Recognition [Research Frontier]. *IEEE Comput. Intell. Mag.* 10, 3 (2015), 52–60.

[35] Yufei Tao, Ke Yi, Cheng Sheng, and Panos Kalnis. 2009. Quality and efficiency in high dimensional nearest neighbor search. In *SIGMOD*. 563–576.

[36] Yuan Tian, David Lo, and Chengnian Sun. 2012. Information Retrieval Based Nearest Neighbor Classification for Fine-Grained Bug Severity Prediction. In *WCRE*. IEEE Computer Society, 215–224.

[37] Yao Tian, Xi Zhao, and Xiaofang Zhou. 2022. DB-LSH: Locality-Sensitive Hashing with Query-based Dynamic Bucketing. *CoRR* (2022).

[38] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. A Comprehensive Survey and Experimental Comparison of Graph-Based Approximate Nearest Neighbor Search. *PVLDB* 14, 11 (2021), 1964–1978.

[39] Edwin B Wilson and Margaret M Hilferty. 1931. The distribution of chi-square. *proceedings of the National Academy of Sciences of the United States of America* 17, 12 (1931), 684.

[40] Wan-Lei Zhao. 2018. k-NN Graph Construction: a Generic Online Approach. *CoRR* abs/1804.03032 (2018).

[41] Bolong Zheng, Xi Zhao, Lianggui Weng, Nguyen Quoc Viet Hung, Hang Liu, and Christian S. Jensen. 2020. PM-LSH: A Fast and Accurate LSH Framework for High-Dimensional Approximate NN Search. *Proc. VLDB Endow.* 13, 5 (2020), 643–655.

APPENDIX: PROOFS

A. Proof of Lemma 4

PROOF. When the query starts with search radius r, the query could terminate with the probability $p(r)$ and go to the next hop with the probability $1 - p(r)$. If the query terminates, $l(r) = 1$; otherwise, $l(r) = l(r - \delta(r)) + 1$. Hence,

$$l(r) = p(r) \cdot 1 + [1 - p(r)][l(r - \delta(r)) + 1] \tag{8}$$

Considering that $l(r) \gg \delta(r)$, we have

$$\frac{l(r) - l(r - \delta(r))}{\delta(r)} \approx l'(r)$$

Thus, $l(r - \delta(r)) = l(r) - \delta(r)l'(r)$ and

$$l(r) = p(r) \cdot 1 + [1 - p(r)][l(r) - \delta(r)l'(r) + 1] \tag{9}$$

Simplifying the above equation, we prove the lemma. □

B. Proof of Lemma 5

PROOF. As that in Proof 9, we have

$$s(r) = p(r) \cdot r + [1 - p(r)]s(r - \delta(r)) \tag{10}$$

Considering that

$$\frac{s(r) - s(r - \delta(r))}{\delta(r)} \approx s'(r)$$

Thus,

$$[1 - p(r)]\delta(r)s'(r) + p(r)s(r) = rp(r) \tag{11}$$

□

C. Solving $\delta(r)$ and $p(r)$.

We now begin to solve $\delta(r)$. Let o_n be the closest points to q among ep's neighbors and θ_m be the vectorial angle between $\vec{e_pq}$ and $\vec{e_po_n}$. We have the following observation:

Observation 1. *$\delta(r)$ satisfies $\delta(r) = r - \sqrt{r^2 + r_o^2 - 2rr_0 \cos\theta_m}$. When $r \gg r_0$, $\delta(r) = r_0 \cos\theta_m$.*

Proof. Since $\|e_p, q\| = r$, $\|e_p, o_n\| = r$ and the vectorial angle between $\vec{e_pq}$ and $\vec{e_po_n}$ is θ_m. According to the law of cosines, $\|o_n, q\| = \sqrt{r^2 + r_o^2 - 2rr_0 \cos\theta_m}$. Thus, $\delta(r) = \|e_p, q\| - \|o_n, q\| = r - \sqrt{r^2 + r_o^2 - 2rr_0 \cos\theta_m}$. □

To analyze the distribution of θ_m, we first consider the distribution of the vectorial angle θ_o between $\vec{e_pq}$ and $\vec{e_po}$, where o is one of neighbors of e_p. Since e_p's neighbors are uniformly distributed around it, θ_o follows the distribution of the vectorial angle between any two random vectors in $\mathbb{R}^{d_i}$, where d_i is the local intrinsic dimensionality (LID) of $\mathcal{D}$ [1].

Lemma 11 (Theorem 1 in [6]). *The probability density function of θ_o is*

$$f_{\theta_o}(\theta) = \frac{\Gamma(\frac{d_i}{2}) sin^{d_i-2}\theta}{\sqrt{\pi}\Gamma(\frac{d_i-1}{2})}, \tag{12}$$

where $\Gamma(x)$ is the Gamma function [24].

Then, the cumulative distribution function of θ_o can be written as $F_{\theta_o}(\theta) = \int_0^\theta f_{\theta_o}(t)dt$. Assume e_p has T_e neighbors. Based on the above assumptions, $\vec{e_po_n}$ is the edge that has the minimal angle to $\vec{e_pq}$. Then, θ_m is the minimal one among T_e vectorial angles, and thus we have

Corollary 1. *The cumulative distribution function of θ_m is*

$$F_{\theta_m}(\theta) = 1 - [1 - F_{\theta_o}(\theta)]^{T_e}, \tag{13}$$

Correspondingly, the probability density function of θ_m can be written as $f_{\theta_m}(\theta) = F'_{\theta_m}(\theta)$.

We now compute $p(r)$ and $p(r)$ is the probability that the query terminates, which indicates that $\delta(r) < 0$. Hence, we have

Lemma 12. *$p(r)$ satisfies*

$$p(r) = 1 - F_{\theta_m}(\arccos\frac{r_0}{2r}). \tag{14}$$

Proof. $p(r) = \Pr[\delta(r) < 0] = \Pr[\cos\theta_m < \frac{r_0}{2r}] = \Pr[\theta_m > \frac{r_0}{2r}] = 1 - F_{\theta_m}(\arccos\frac{r_0}{2r})$. □

D. Proof of Lemma 6

Proof. Since $p(r)$ and $\delta(r)$ are independent on n, $l(r)$ and $s(r)$ will also be independent on n and only affected by the local intrinsic dimensionality of the dataset. □

E. Proof of Lemma 9

Proof. When $r > r_1$, the query is very unlikely to terminate and when can consider $p(r) = 0$. Thus, Eq. 4 can be simplified as $\delta(r)l'(r) = 1$ and we have

$$l(r) = \int_{r_1}^{r} \frac{1}{\delta(r)}dr + l(r_1).$$

Then, $B_{\mathcal{I}_H} = l(r_2) - l(r_1) = \int_{r_1}^{r_2} \frac{1}{\delta(r)}dr$. Further, consider $r \gg r_o$, we approximate $\delta(r)$ as

$$\begin{aligned}\delta(r) =& r - r\sqrt{1 + (\frac{r_o}{r})^2 - 2\frac{r_o}{r}\cos\theta_m} \\ \approx& r - r[1 + \frac{r_o^2}{2r^2} - \frac{r_o}{r}\cos\theta_m] \approx r_o\cos\theta_m\end{aligned}$$

Hence, we have $B_{\mathcal{I}_H} = \frac{r_2 - r_1}{r_o \cos\theta_m}$ and $r_2 - r_1 = \Delta r$. Therefore, the expected $B_{\mathcal{I}_H}$ is $\frac{\Delta r}{r_o \mathbb{E}[\cos\theta_m]}$. Denote $\lambda = \mathbb{E}[\cos\theta_m]$, which satisfies $\lambda < 1$, we prove the lemma. □

F. Proof of Theorem 2

Proof. Consider that $r > r_o$ and we express $s(r)$ as $s(r) = \int_{r_o}^{r_i} s'(r)dr + s(r_o)$. Since $s(r_o) \leq r_o$, denote $\int_{r_o}^{r_i} s'(r)dr = \beta r_o$ and we prove that β is bounded.

According to Lemma 5, we have

$$[1 - p(r)]\delta(r)s'(r) < rp(r). \tag{15}$$

$p(r)$ decreases monotonically with r and thus $1 - p(r) > 1 - p(r_o)$. From the proof of Lemma 9 we know that $\delta(r) \approx \lambda r_o$. Hence,

$$s'(r) < \frac{rp(r)}{[1 - p(r_o)]r_o\cos\theta_m}.$$

Then,

$$\beta r_o = \int_{r_o}^{r_i} s'(r)dr < \frac{\int_{r_o}^{r_i} rp(r)dr}{[1 - p(r_o)]r_o\cos\theta_m}$$

$p(r) = 1 - F_{\theta_m}(\arccos\frac{r_0}{2r}) = [1 - F_{\theta_o}(\arccos\frac{r_0}{2r})]^{T_e}$ Let $u = r/r_o$ and $u_i = r_i/r_o$, we have

$$\beta < \frac{\int_1^{u_i} u[1 - F_{\theta_o}(\arccos\frac{1}{2u})]^{T_e}du}{\lambda[1 - p(r_o)]}.$$

Although it is hard to compute this integral, we can find it can be reduced by increasing T_e. Since $\lim_{T_e \to = \infty} \beta = 0$, we can guarantee that $\beta < 1$ when given a large enough T_e. □