Down with the Hierarchy:
The ‘H’ in HNSW Stands for “Hubs”

Blaise Munyampirwa
Independent Researcher
Mountain View, CA

Vihan Lakshman
MIT CSAIL
Cambridge, MA

Benjamin Coleman
Google DeepMind
Mountain View, CA

Presented by Bosen Yang

2025-05-20

Outline

❑**Brief introduction of HNSW**

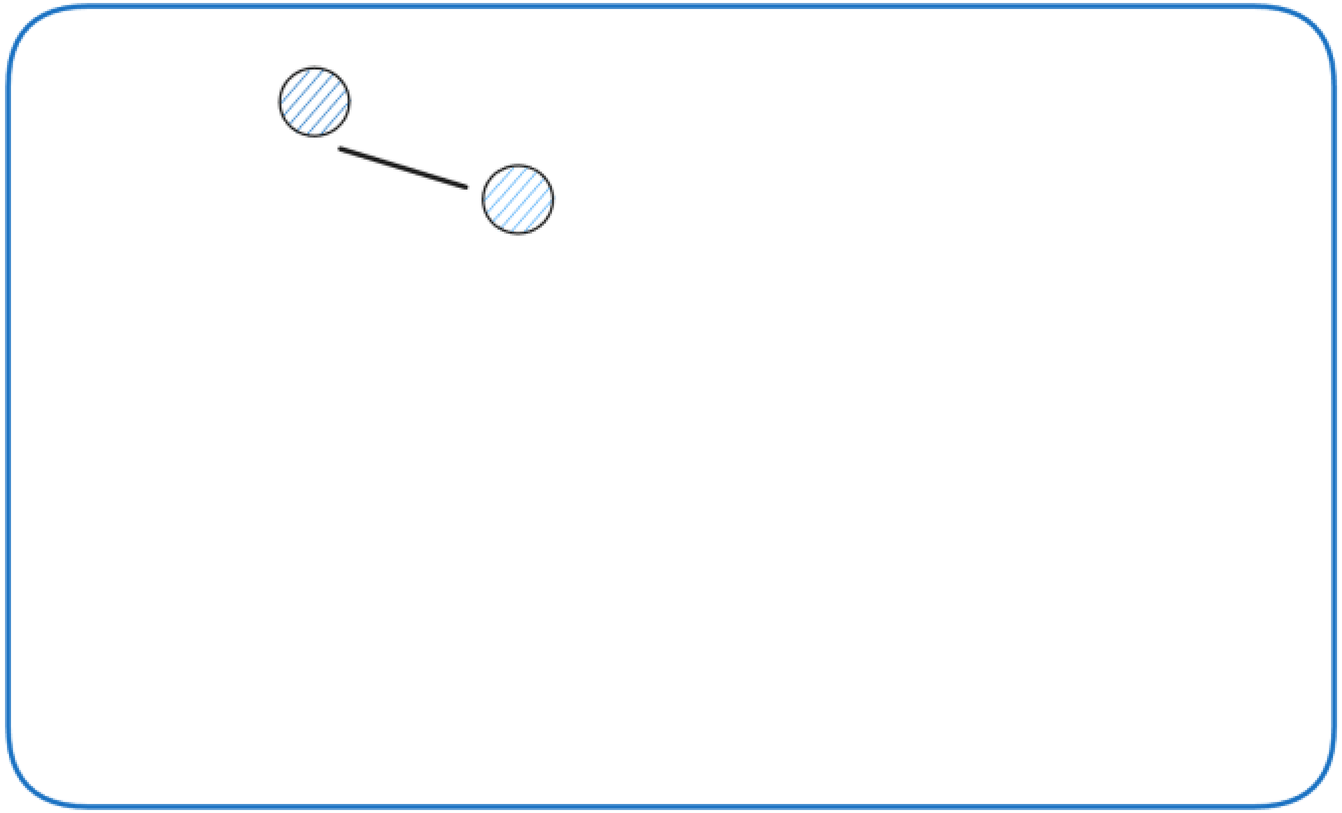
❑**Analysis of HNSW**

❑**Hubness highway hypothesis in high dimensional space**

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖Building parameter: M
  - ➢Newly inserted node will be connected to M nearest nodes in graph

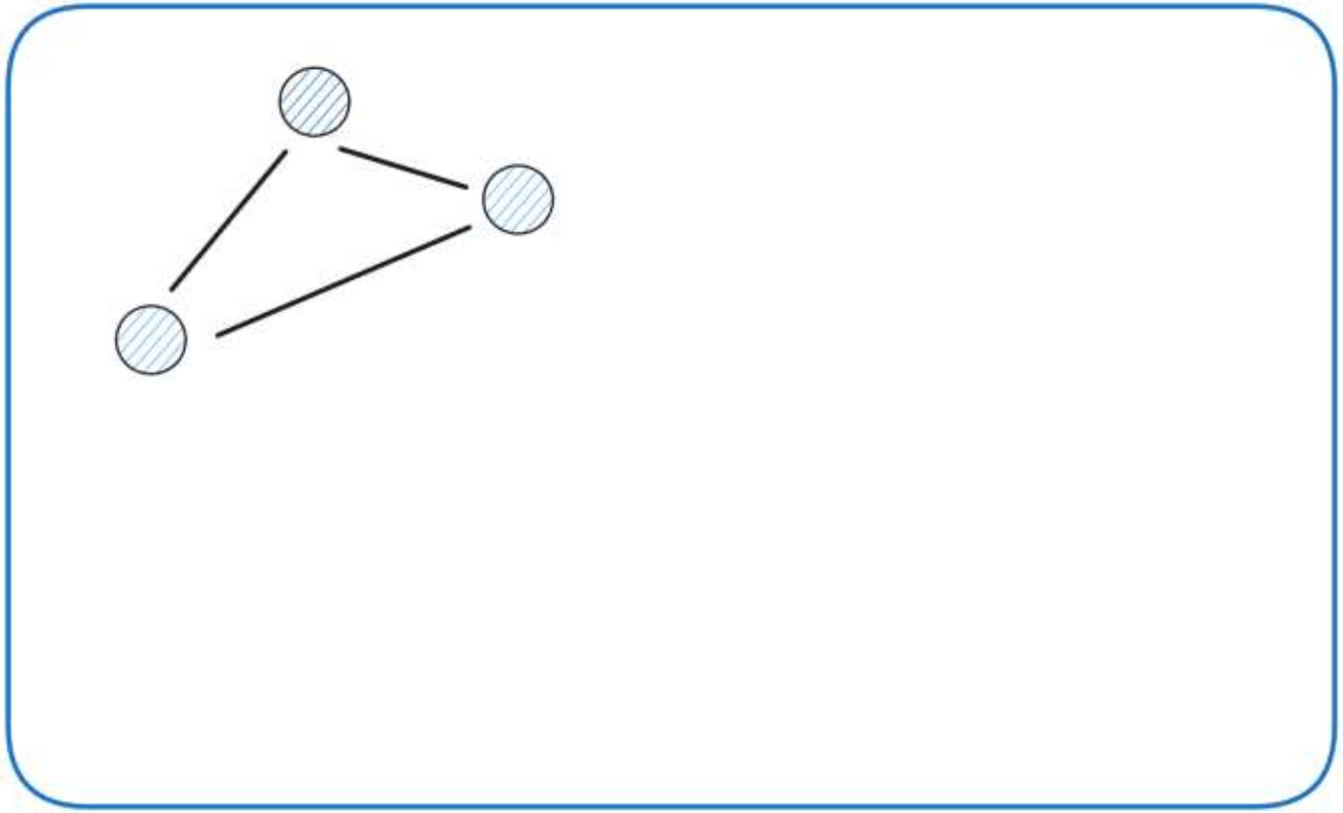


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- Building parameter: M
  - Newly inserted node will be connected to M nearest nodes in graph

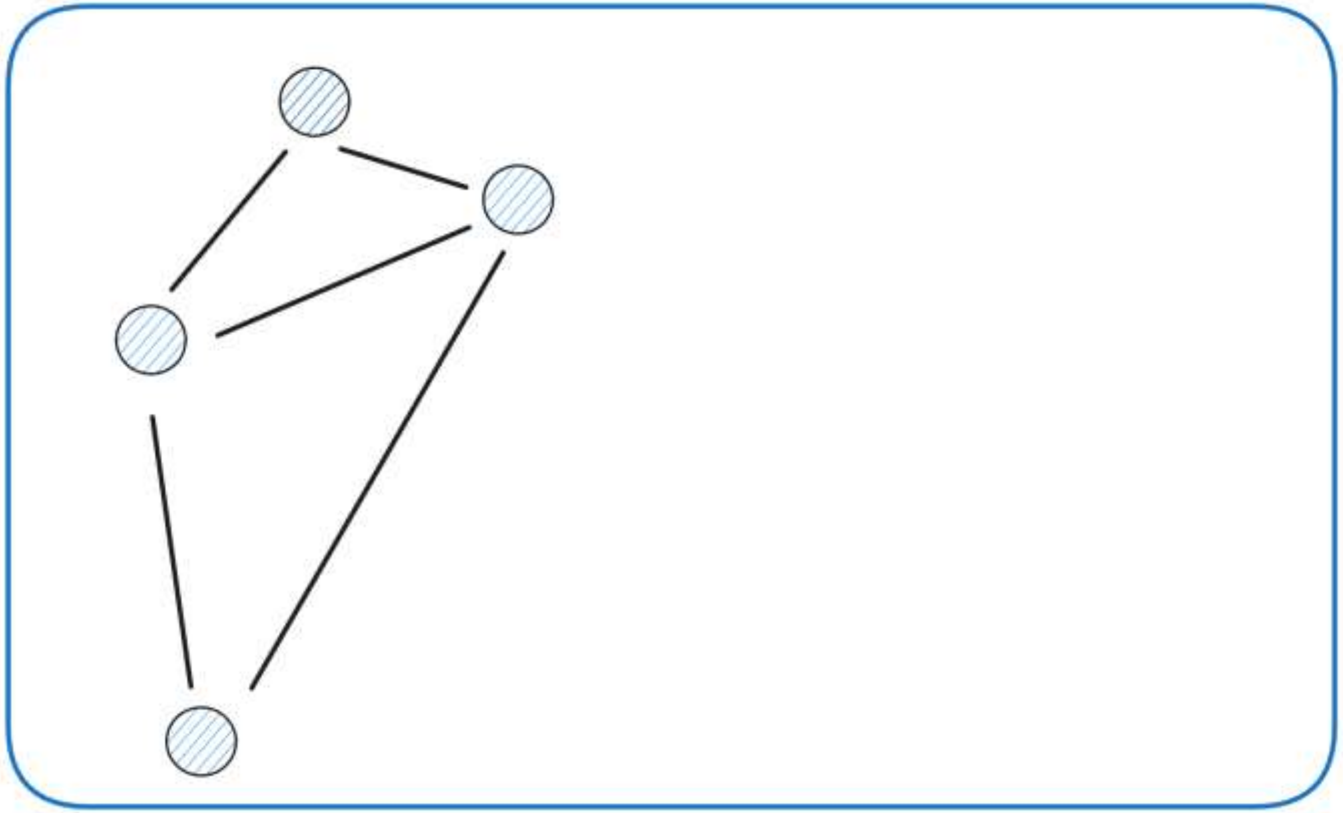


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖Building parameter: M
  - ➢Newly inserted node will be connected to M nearest nodes in graph

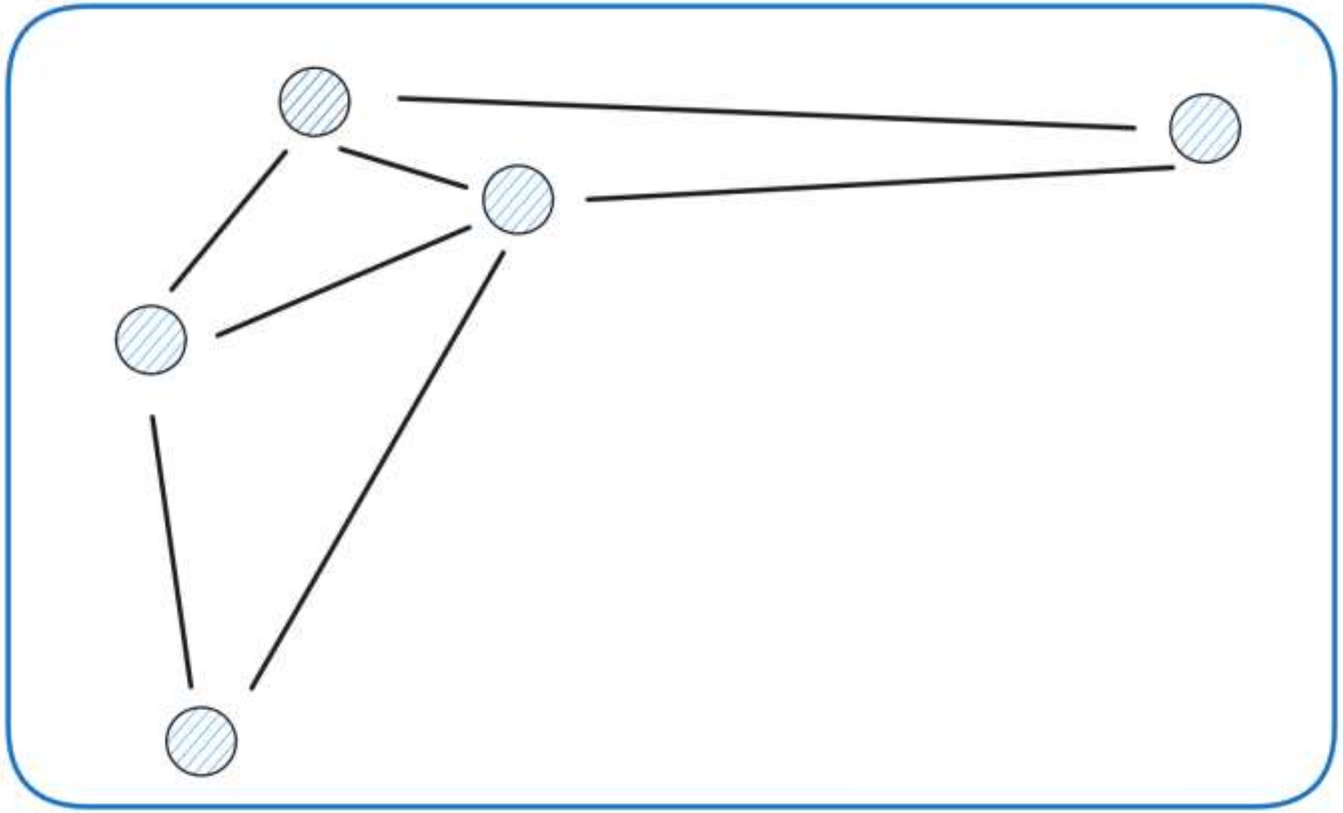


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖Building parameter: M
  - ➢Newly inserted node will be connected to M nearest nodes in graph

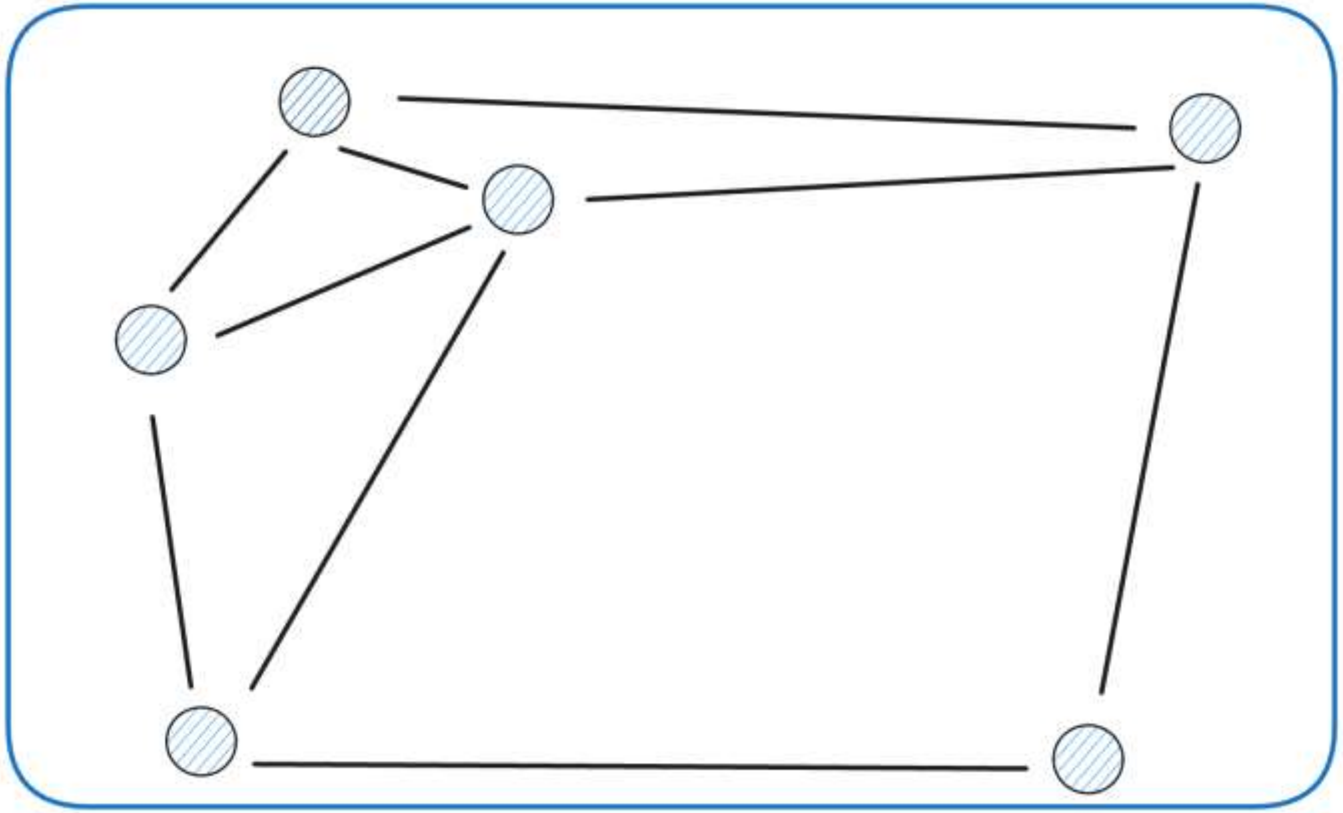


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖Building parameter: M
  - ➢Newly inserted node will be connected to M nearest nodes in graph

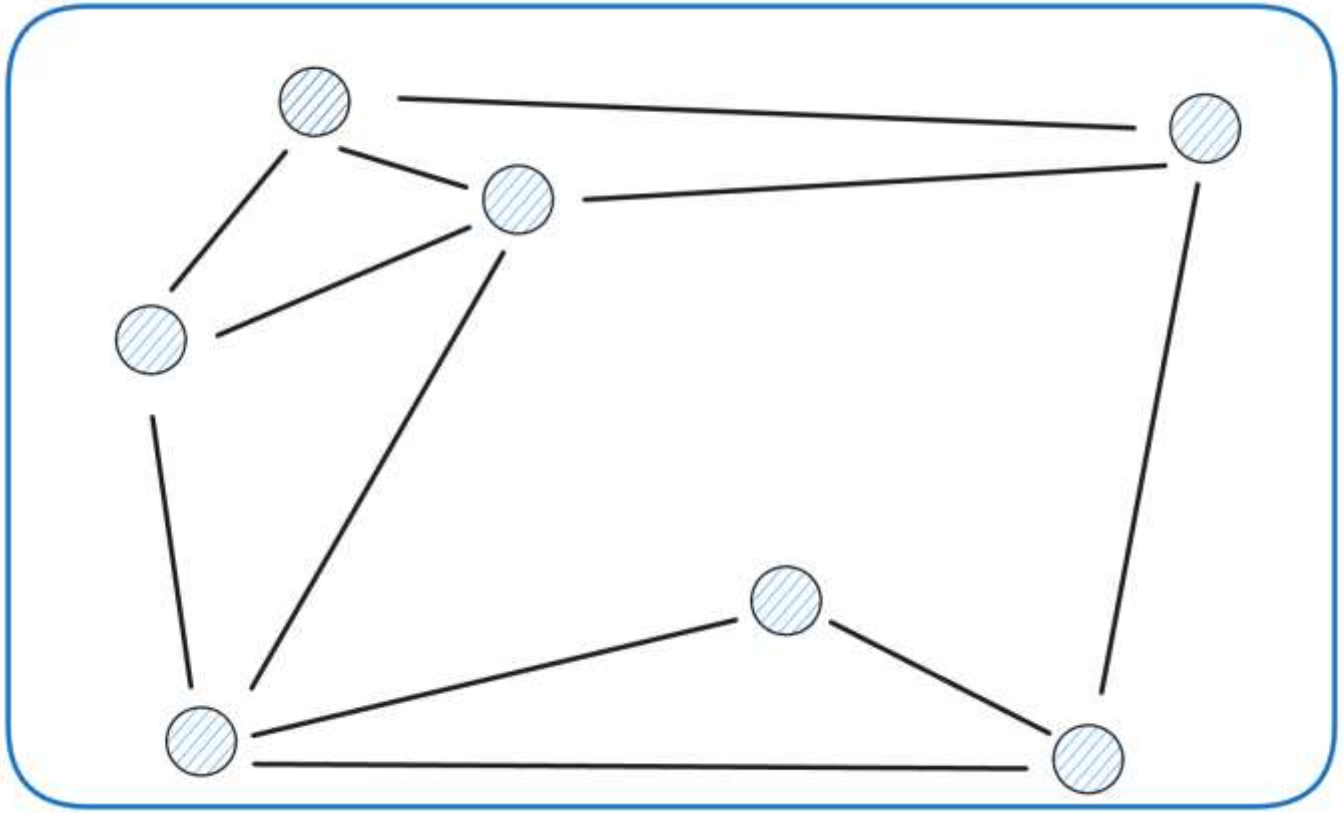


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖Building parameter: M
  - ➢Newly inserted node will be connected to M nearest nodes in graph

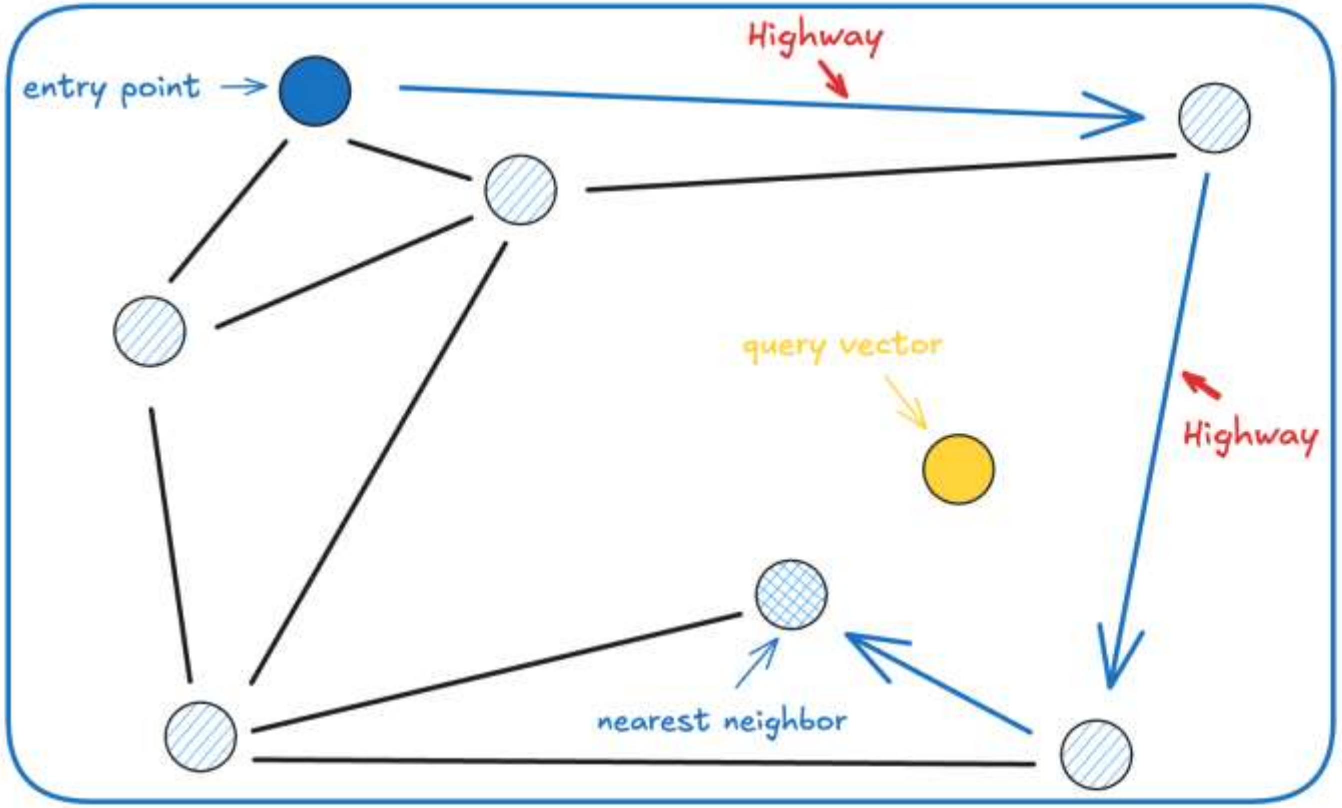


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

- ❖There exists “highway” in NSW
  - ➢Highway: The edges that can reach the nearest neighbor fast

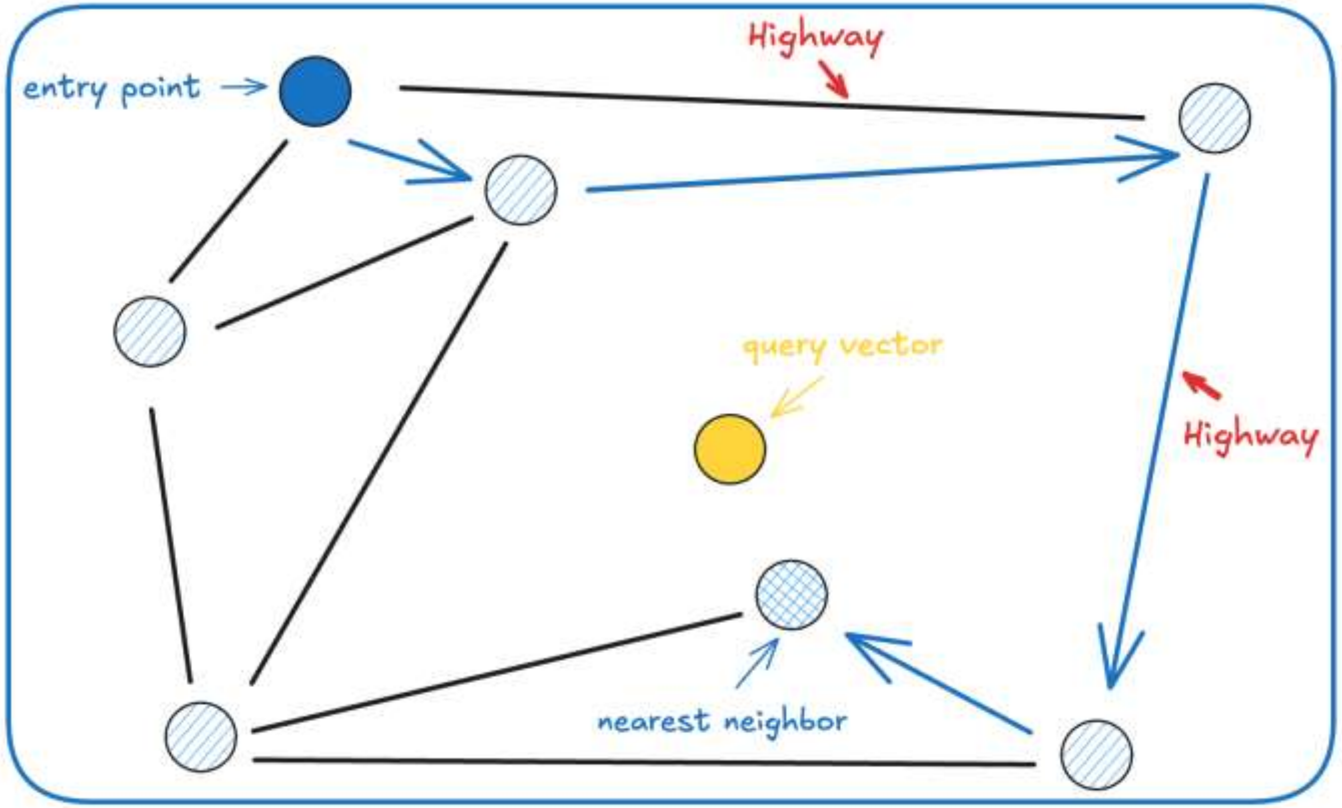


Example: M = 2

NSW

❑Brief introduction of NSW (Navigable Small World)

❖Although there exists “highway” in NSW, it’s unable to identify it

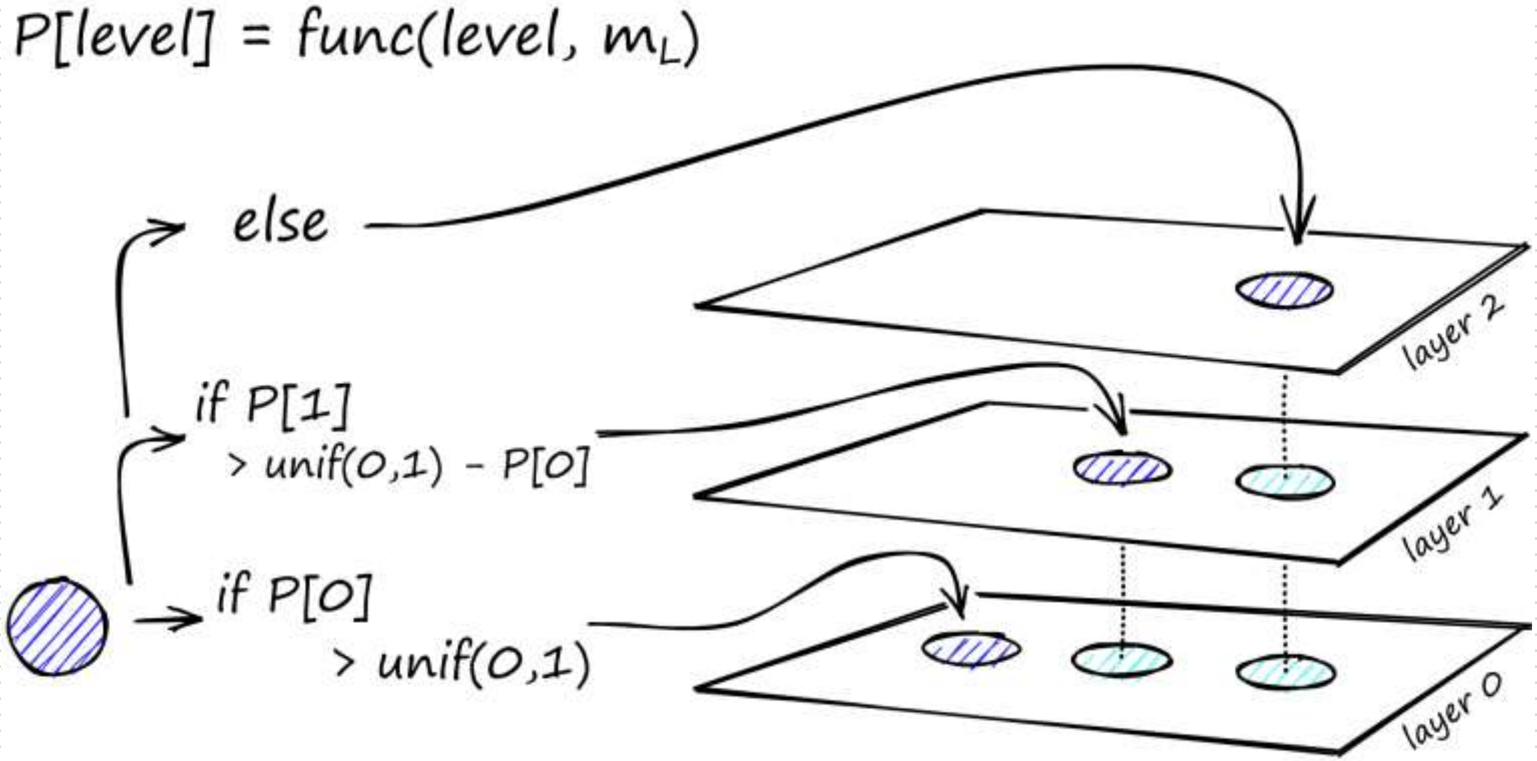


Example: M = 2

HNSW

❑Brief introduction of HNSW (Hierarchical Navigable Small World)

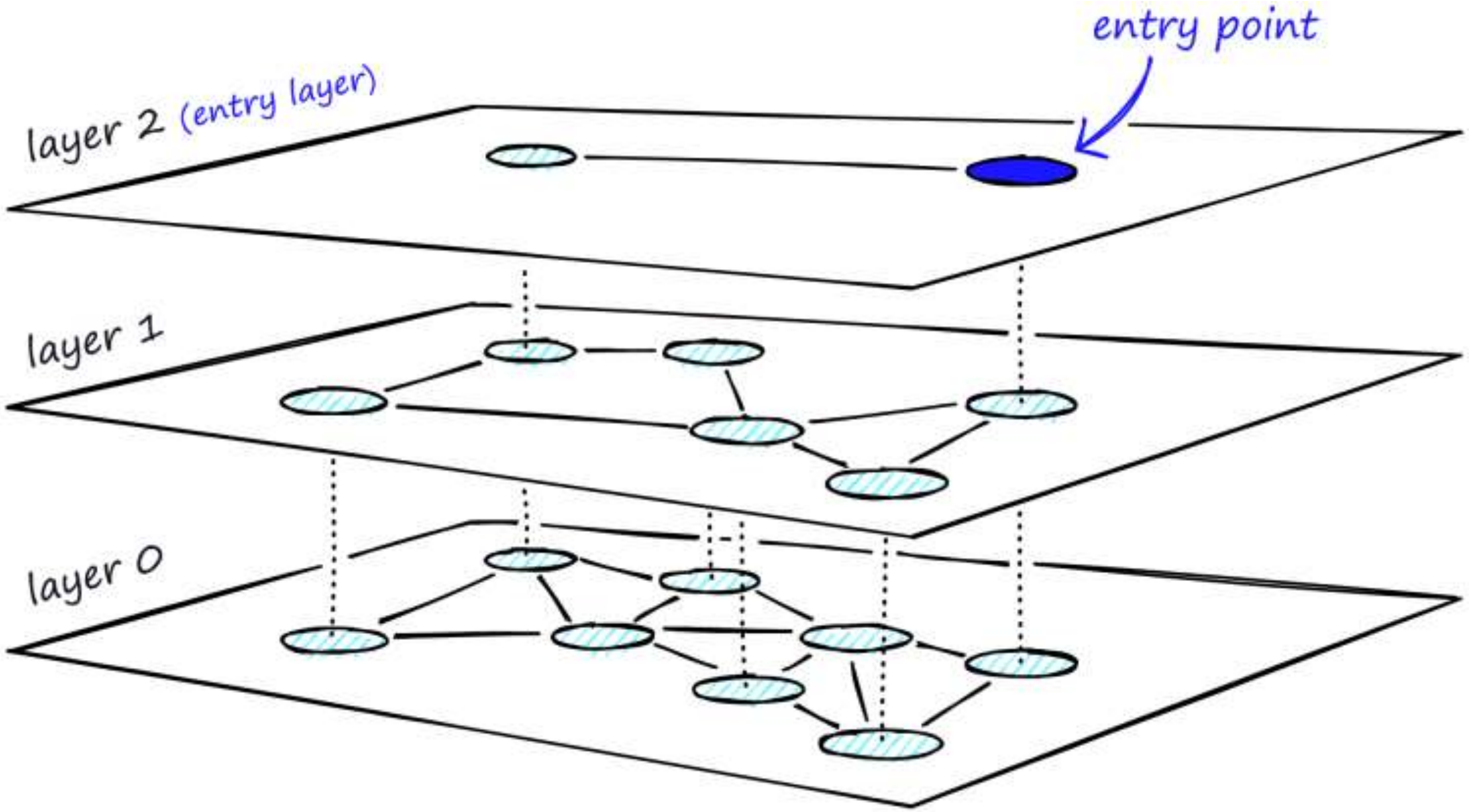
❖Each layer uses probability function to decide if a new node can be inserted



HNSW

❑Brief introduction of HNSW (Hierarchical Navigable Small World)

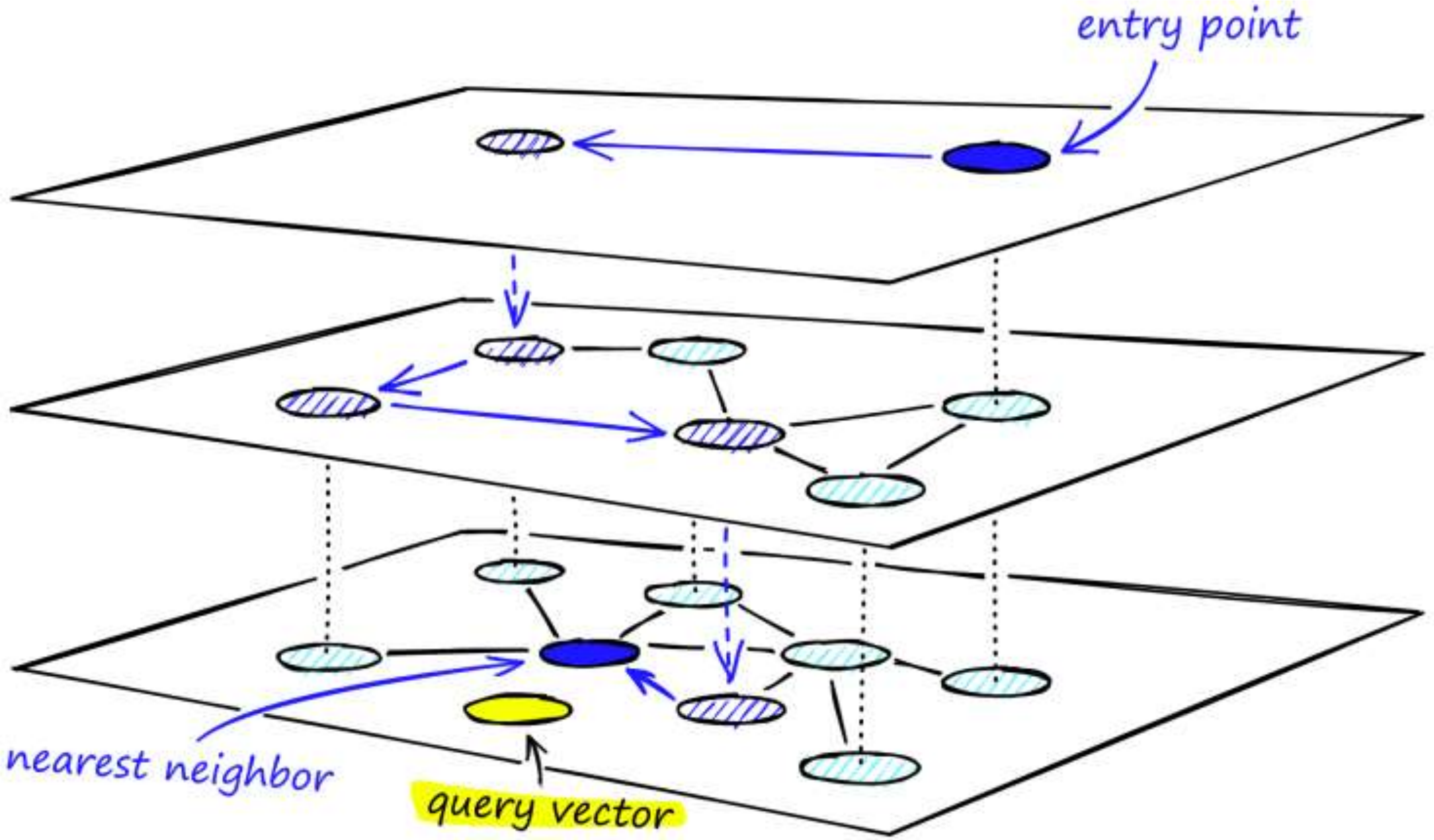
❖Build “highway” in hierarchical layers



HNSW

❑Brief introduction of HNSW (Hierarchical Navigable Small World)

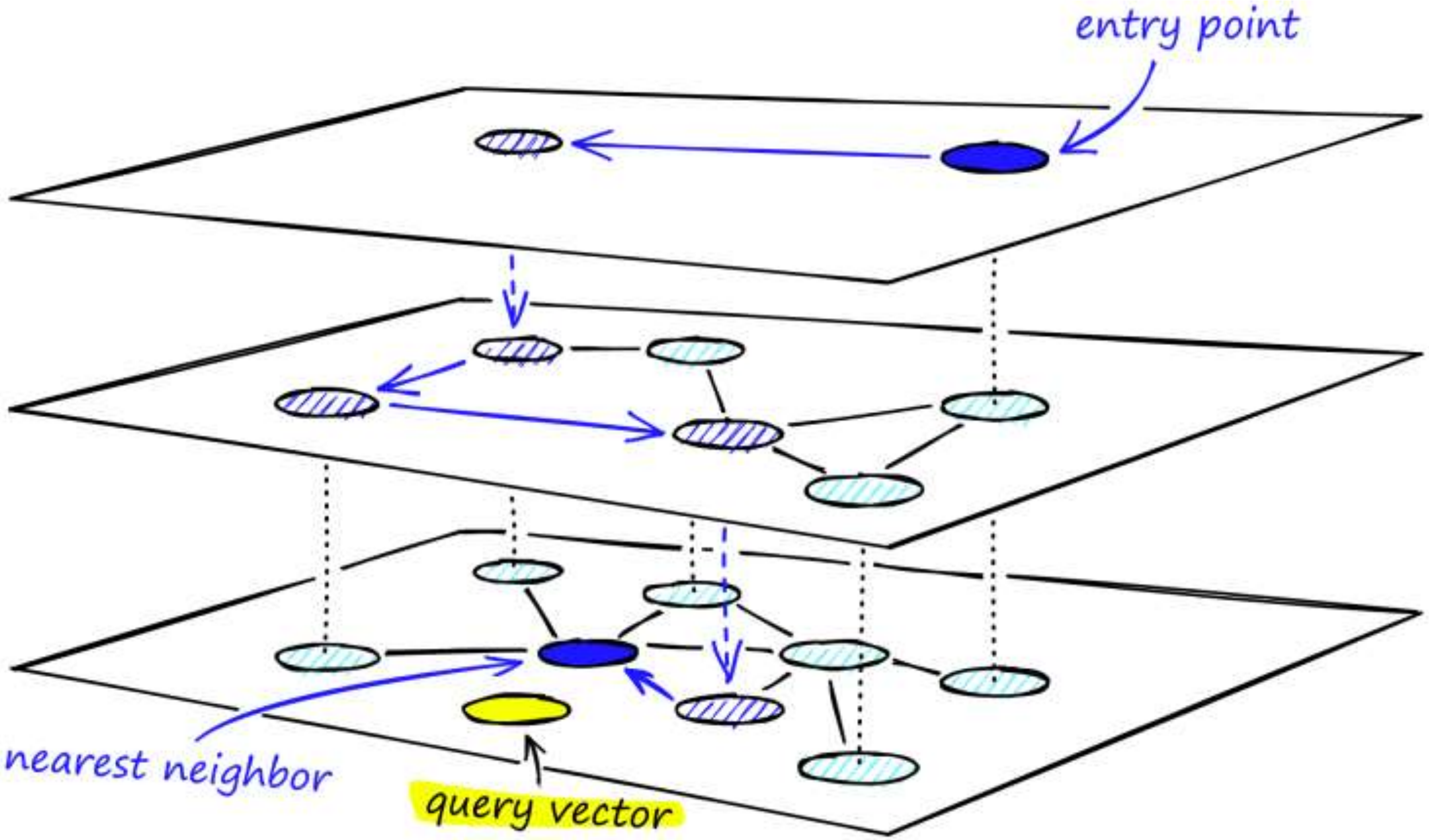
❖Reach the “highway” in the hierarchical layer in search



HNSW

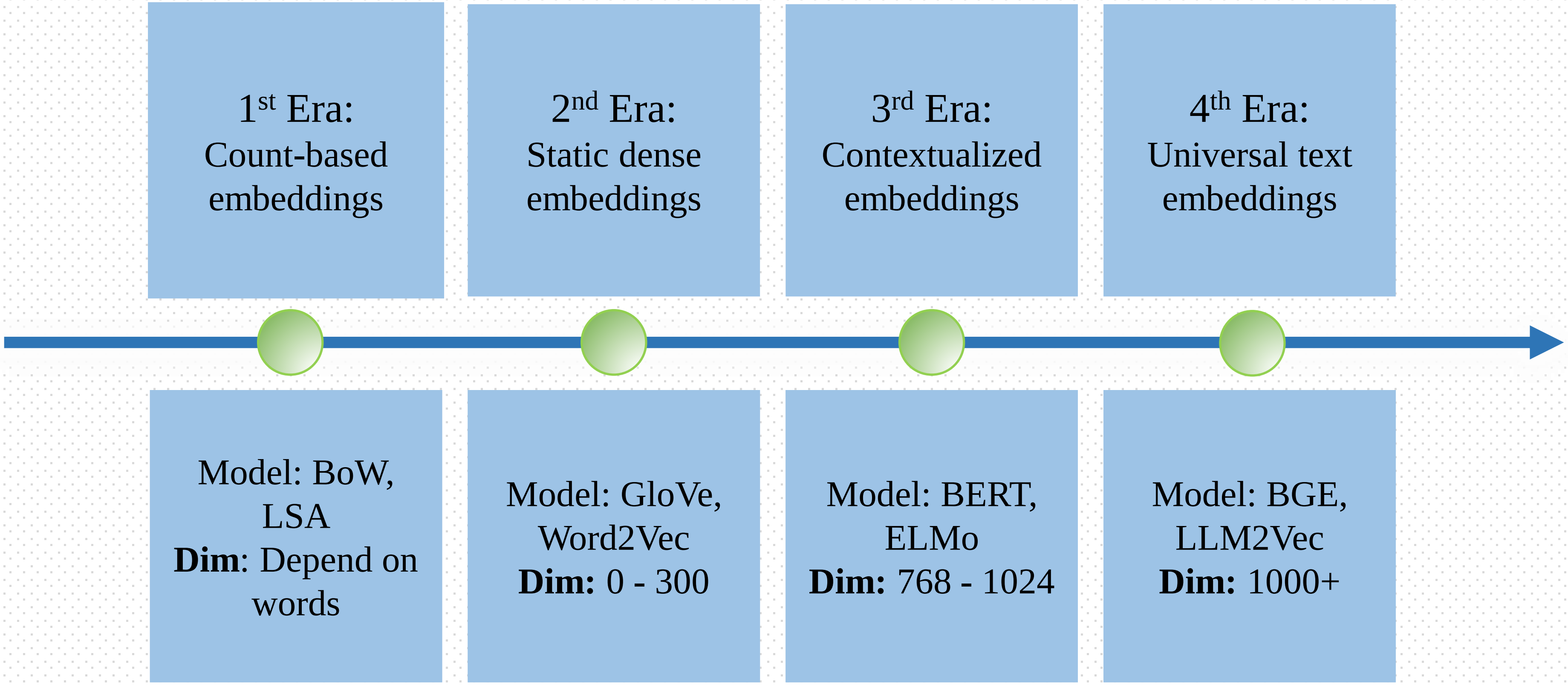
❑Brief introduction of HNSW (Hierarchical Navigable Small World)

❖With the hierarchical layer, HNSW performs well and is widely used in ANN search



Background

❑Dimensions of vectors become increasingly higher



1st Era:
Count-based embeddings

Model: BoW, LSA
Dim: Depend on words

2nd Era:
Static dense embeddings

Model: GloVe, Word2Vec
Dim: 0 - 300

3rd Era:
Contextualized embeddings

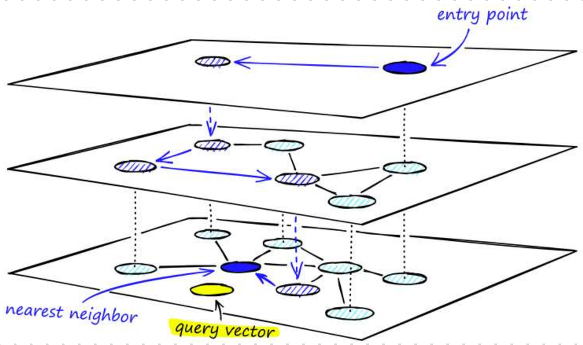
Model: BERT, ELMo
Dim: 768 - 1024

4th Era:
Universal text embeddings

Model: BGE, LLM2Vec
Dim: 1000+

Background

❑Dimensions of vectors become increasingly higher



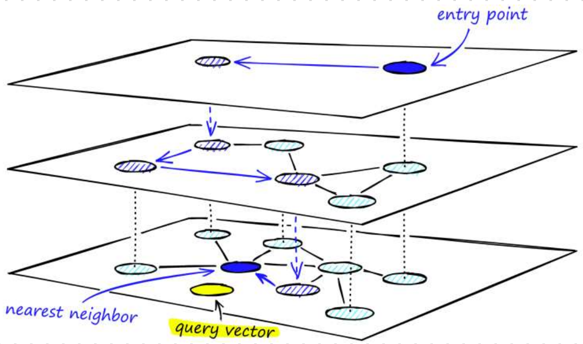
HNSW

"HNSW is the only vector index supported by PASE, Milvus, and Elasticsearch in common." [VBASE OSDI'23]

Model: BoW, LSA **Dim**: Depend on words	Model: GloVe, Word2Vec **Dim:** 0 - 300	Model: BERT, ELMo **Dim:** 768 - 1024	Model: BGE, LLM2Vec **Dim:** 1000+

Background

- Dimensions of vectors become increasingly higher



HNSW

Is HNSW still effective?

Model: BoW, LSA
Dim: Depend on words

Model: GloVe, Word2Vec
Dim: 0 - 300

Model: BERT, ELMo
Dim: 768 - 1024

Model: BGE, LLM2Vec
Dim: 1000+

Benchmarking Experiments

❑Goal

- ❖Evaluate performance of HNSW in high dimensional space

❑Code

- ❖HNSW: hnswlib [IEEE TPAMI’16]
- ❖NSW: flatnav
  - ➢Built from hnswlib
  - ➢Separate the confounding impact of performance engineering

Benchmarking Experiments

❑Goal

- ❖Evaluate performance of HNSW in high dimensional space

❑Dataset

Dataset	Dimensionality	# Points	# Queries
Synthetic Uniform	4, 8, and 16		
Yandex DEEP	96	100M	10K
Microsoft SpaceV	100	100M	29.3K
BigANN	128	100M	10K
NYTimes	256	290K	10K
GIST	960	1M	1K

Benchmarking Experiments

❑Result

- ❖Memory (GB) consumption

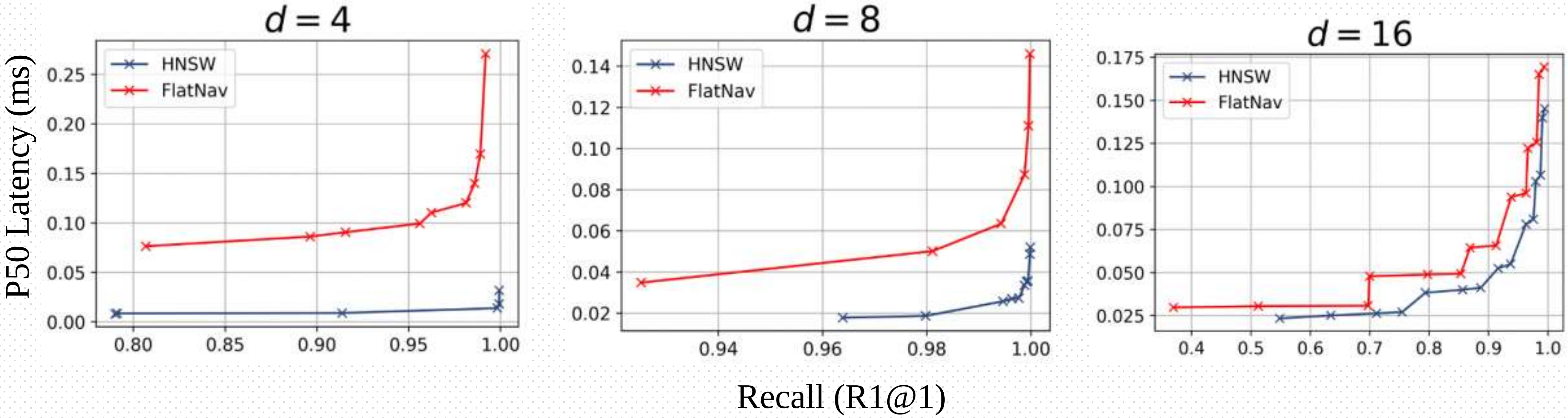
Dataset	# Data	Dimensionality	Hnswlib Memory	Flatnav Memory
BigANN	100M	128	183	113
Microsoft SpaceV	100M	100	104	85.5
Yandex DEEP	100M	96	100	60.7

NSW can reduce about 40% memory compared to HNSW

Benchmarking Experiments

❑Result

❖Synthetic Uniform

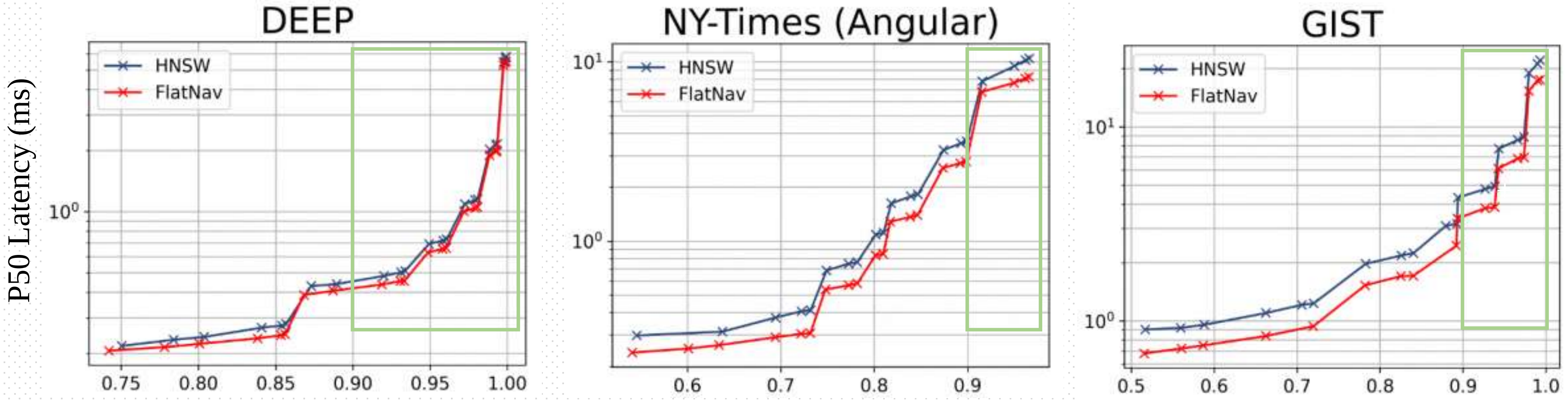


In low dimensional space, HNSW performs better than NSW

Benchmarking Experiments

❑Result

❖High dimensional dataset



Dim: 96

Dim: 256

Recall (R100@100)

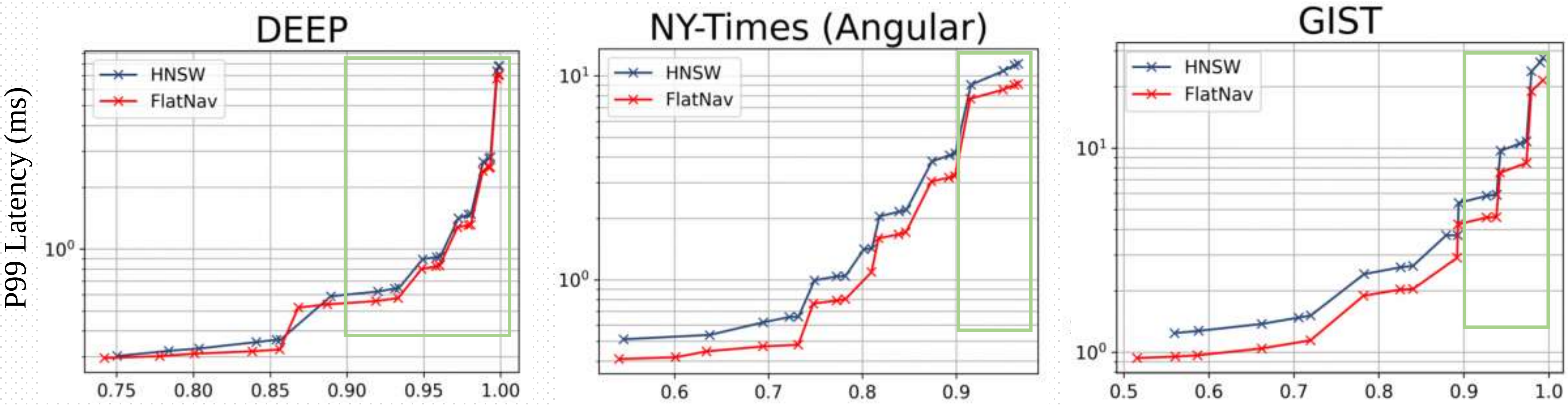
Dim: 960

In high dimensional space, HNSW provides no tangible benefit

Benchmarking Experiments

❑Result

❖High dimensional dataset



Dim: 96

Dim: 256

Dim: 960

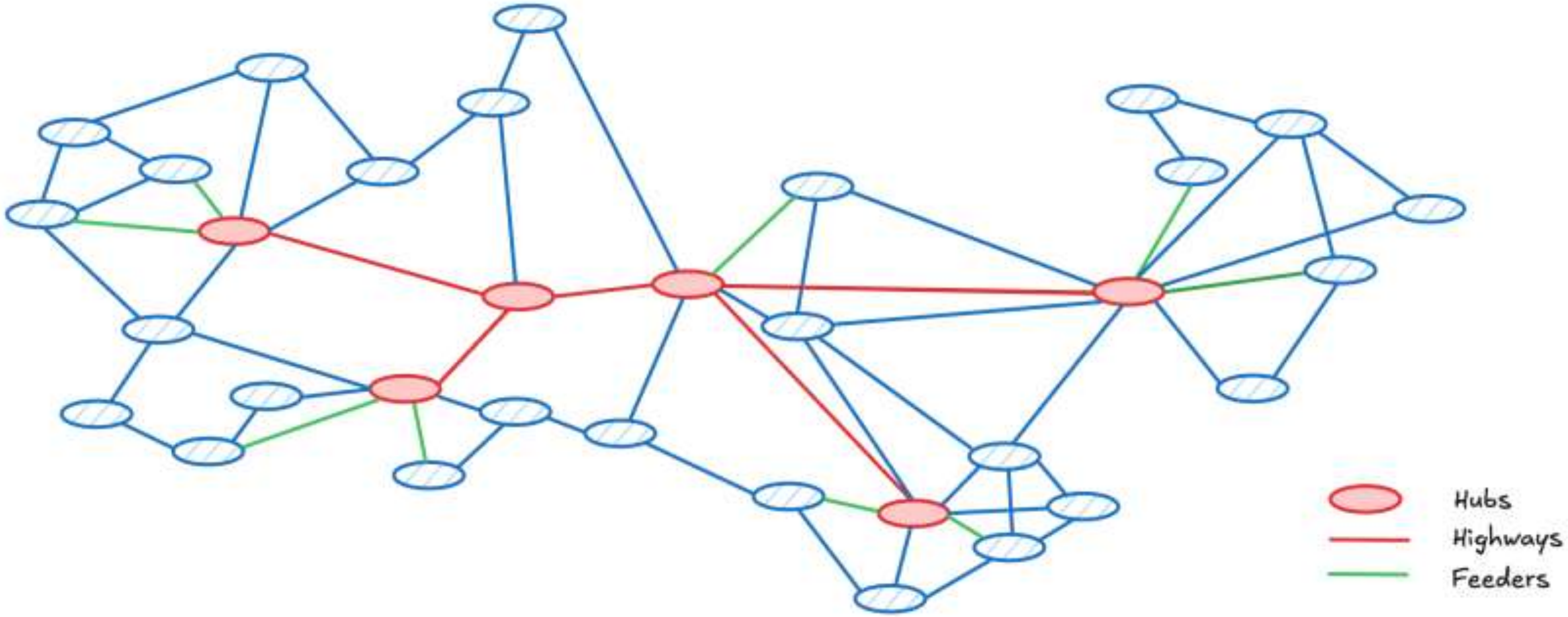
Recall (R100@100)

In high dimensional space, HNSW provides no tangible benefit

Why hierarchy fails

❑Hub Highway Hypothesis in high dimensional space

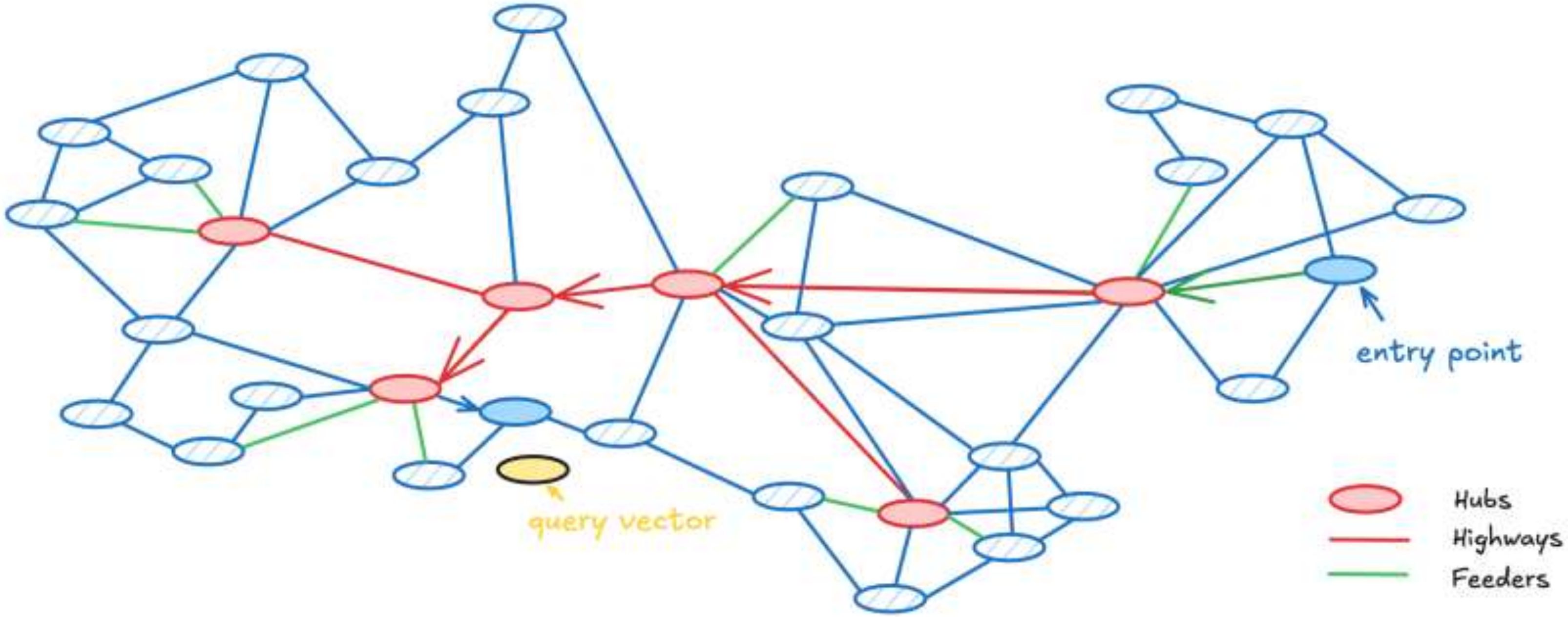
❖There exists well-connected and heavily traversed nodes



Why hierarchy fails

❑Hub Highway Hypothesis in high dimensional space

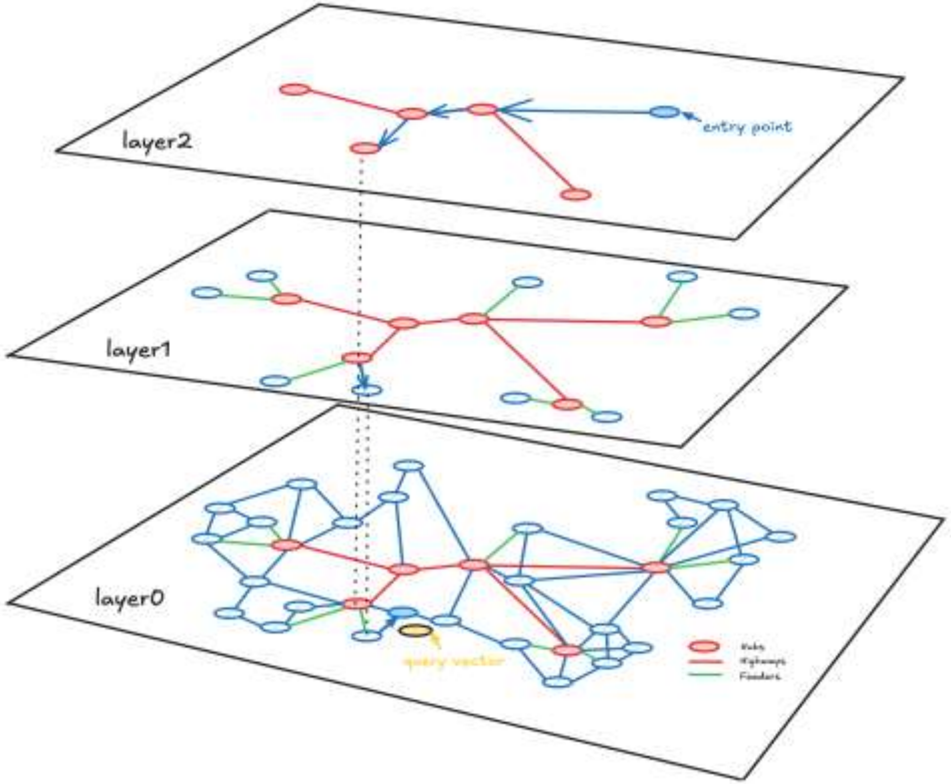
- ❖There exists well-connected and heavily traversed nodes



Why hierarchy fails

❑Hub Highway Hypothesis in high dimensional space

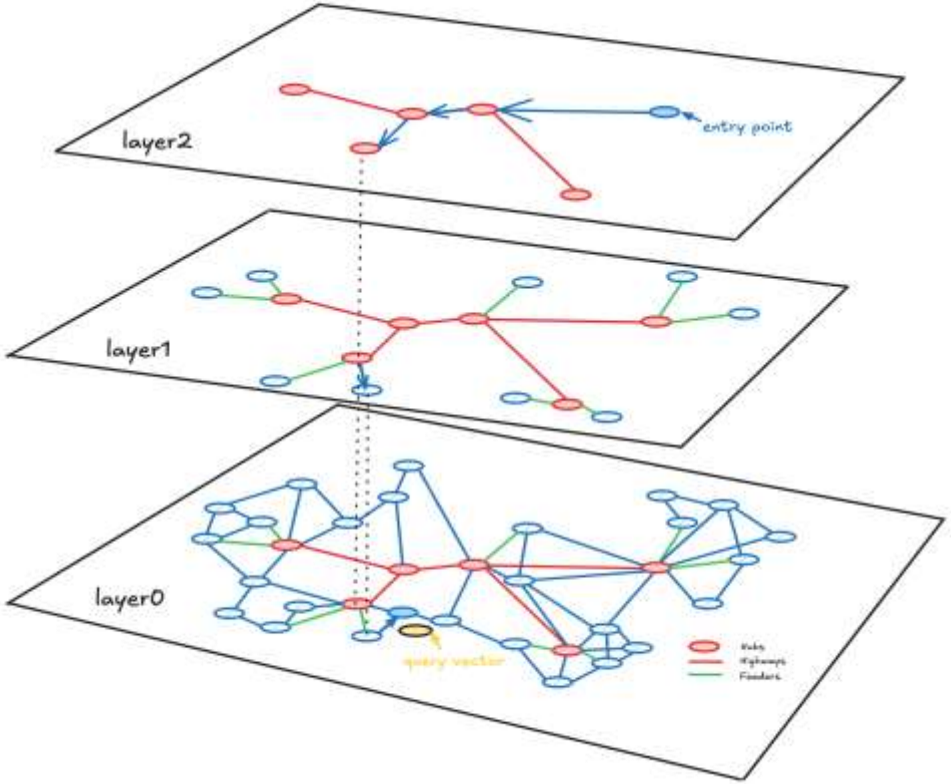
❖There exists well-connected and heavily traversed nodes



Why hierarchy fails

❑Hub Highway Hypothesis in high dimensional space

- ❖There exists well-connected and heavily traversed nodes



Hierarchical structure repeatedly identifies highways

Hub Highway Hypothesis

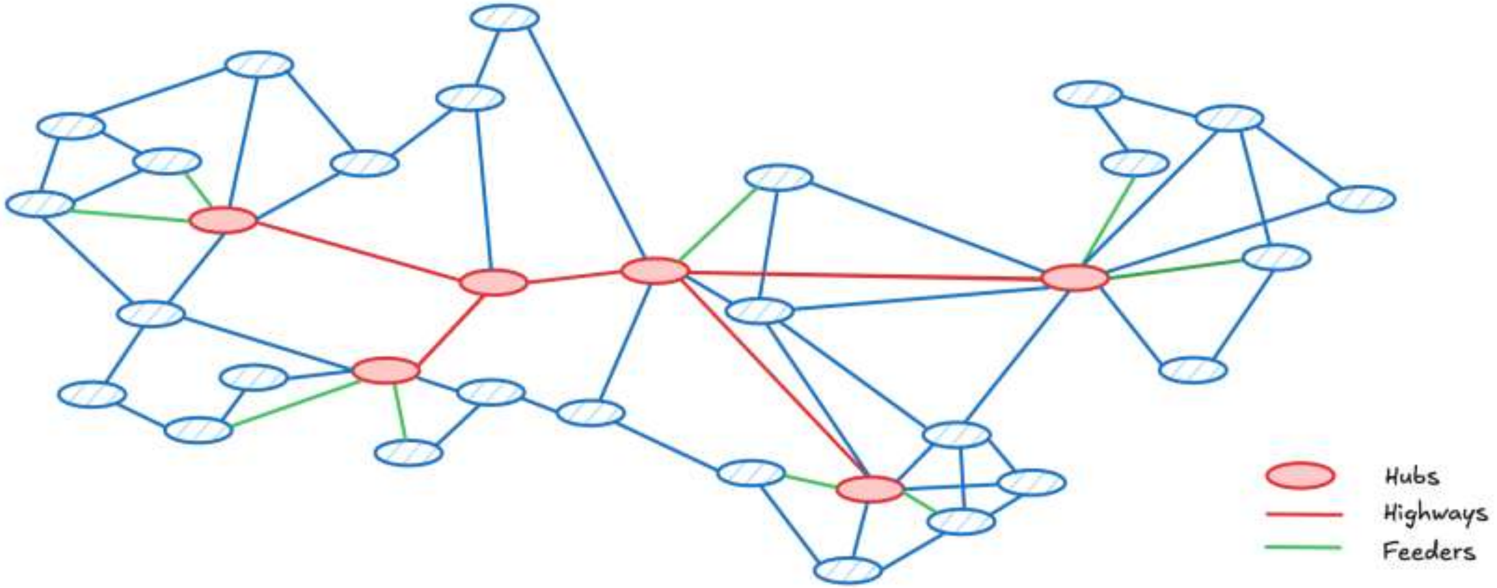
❑Methodology

Claim 1 ***Some nodes are visited by queries much more frequently than others***

Claim 2 ***The hub nodes tend to be connected to each other***

Claim 3 ***Queries visit many hub nodes early in the search process***

❖If these three claims can be satisfied, it indicates that the hypothesis is correct.



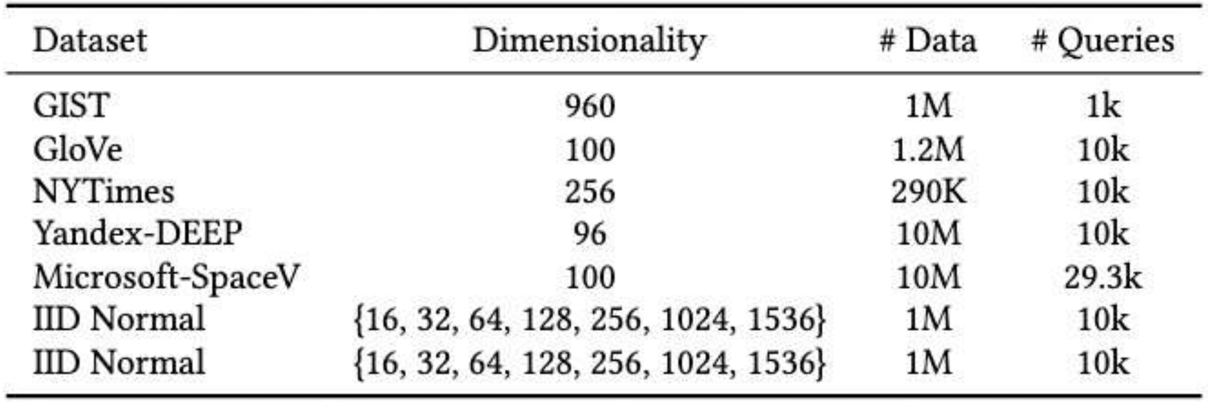
Empirical Evidence

❑Experiment1: Prove claim1

Claim 1 *Some nodes are visited by queries much more frequently than others*

❑Setup

❖Dataset

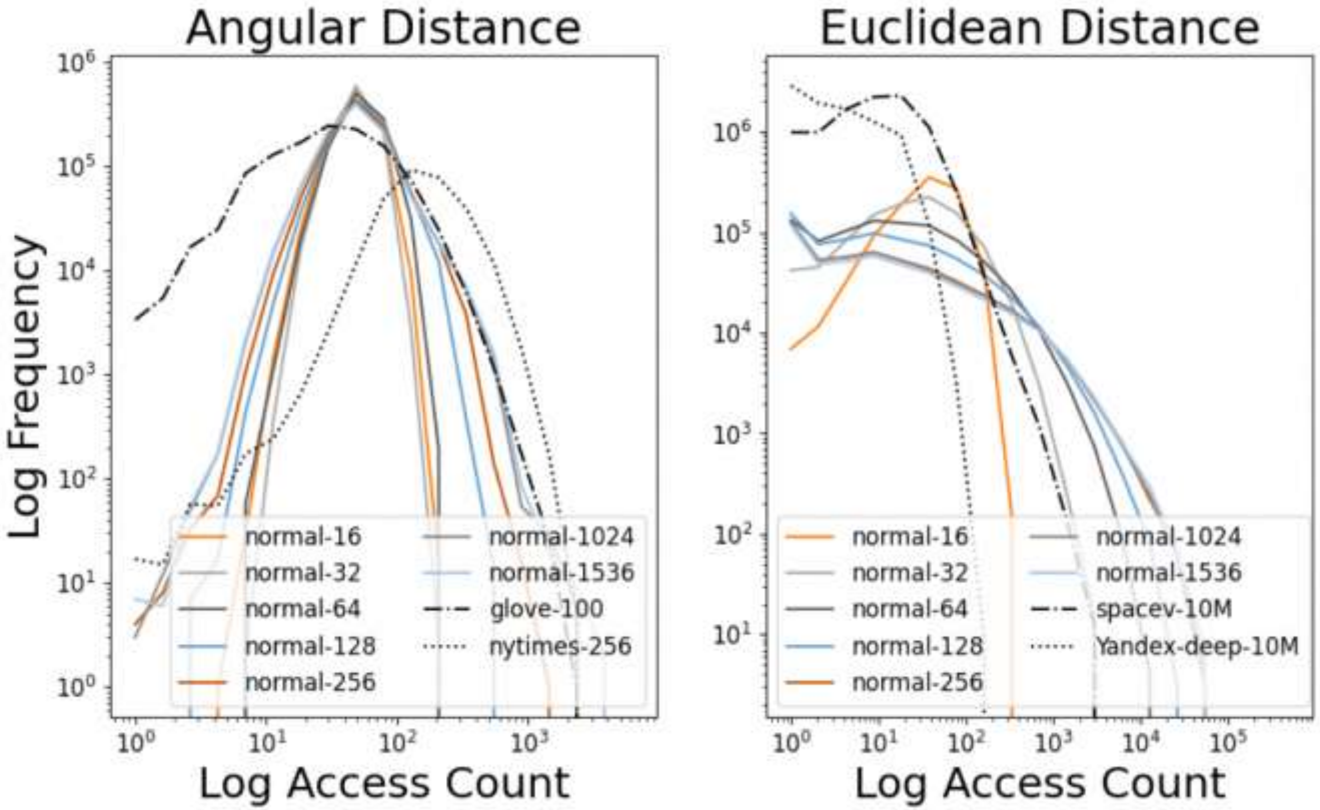


Dataset	Dimensionality	# Data	# Queries
GIST	960	1M	1k
GloVe	100	1.2M	10k
NYTimes	256	290K	10k
Yandex-DEEP	96	10M	10k
Microsoft-SpaceV	100	10M	29.3k
IID Normal	{16, 32, 64, 128, 256, 1024, 1536}	1M	10k
IID Normal	{16, 32, 64, 128, 256, 1024, 1536}	1M	10k

❖Check if the distribution of node access count is skewed

Empirical Evidence

❑Skewness of the Node Access Distribution

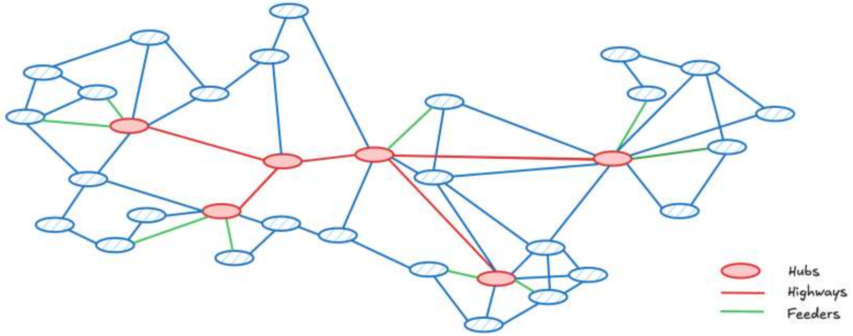


The distribution is indeed skewed to the right

Empirical Evidence

❑Experiment2: Prove claim2

Claim 2 ***The hub nodes tend to be connected to each other***



❑Experimental Design Approach

❖How to identify hub nodes

➢Use P95/P99 threshold of the node access distribution based on Experiment1

❖How to prove the claim2

➢Estimate the likelihood (L1) of hub nodes among the neighbors of hub nodes

➢Estimate the likelihood (L2) of hub nodes among the neighbors of non-hub nodes

➢Propose null hypothesis : there is no difference between L1 and L2

➢Use Mann-Whitney U-test and two-sample t-test to reject null hypothesis

Empirical Evidence

❑Connectivity between hub nodes

Dataset	Dim	P95 Can non-hypothesis be rejected?	P99 Can non-hypothesis be rejected?
Yandex-DEEP	96	No	Yes
Microsoft-SpaceV	100	No	Yes
GloVe	100	Yes	Yes
NYTimes	256	Yes	Yes
GIST	960	Yes	Yes

In most cases, non-hypothesis can be rejected

Empirical Evidence

❑Experiment3: Prove claim3

Claim 3 ***Queries visit many hub nodes early in the search process***

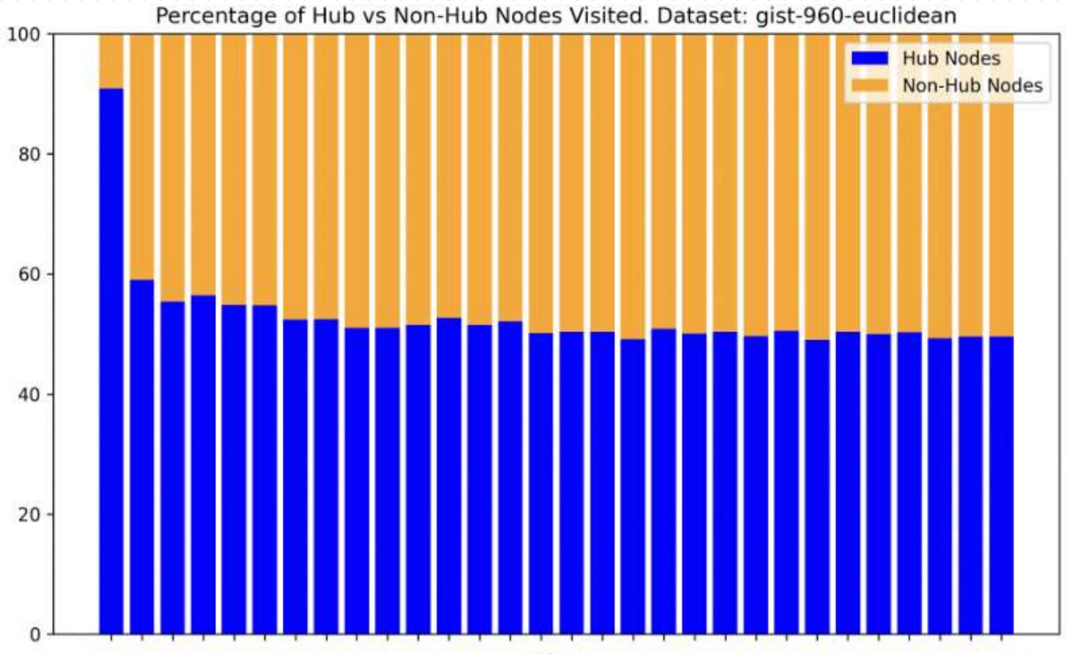
❑Experimental Design Approach

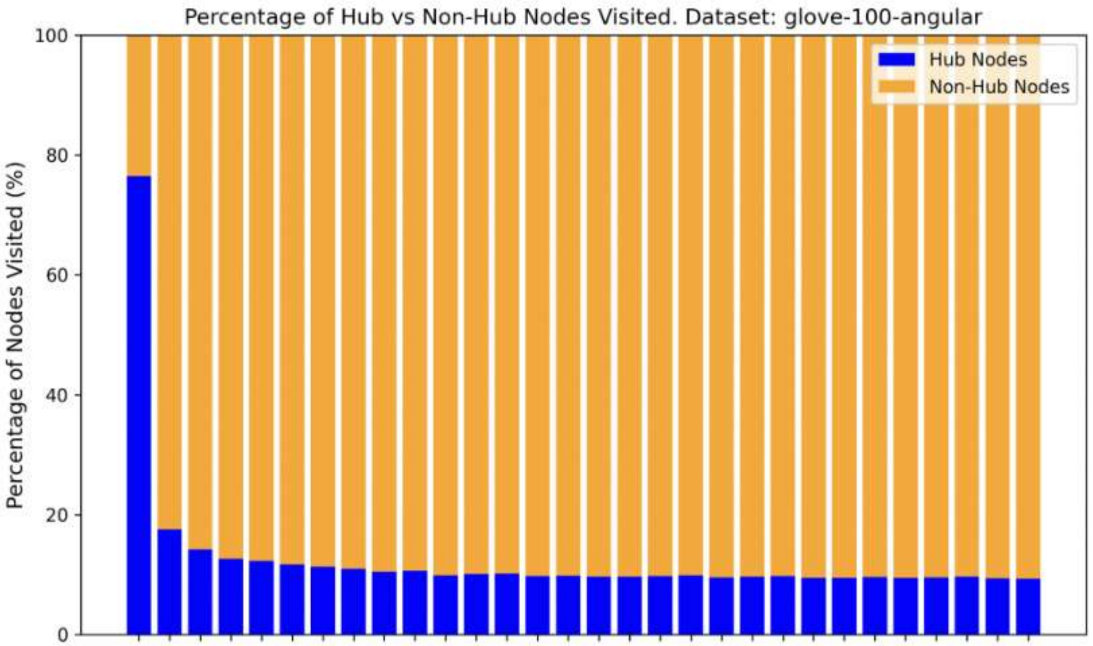
- ❖How to identify hub nodes
  - ➢Use P95/P99 threshold of the node access distribution based on Experiment1
- ❖Examine the fraction of time spent on hub nodes in different phases of search

Empirical Evidence

❑Hub-Highway Nodes Enable Fast Traversal

Percentage of nodes visted (%)



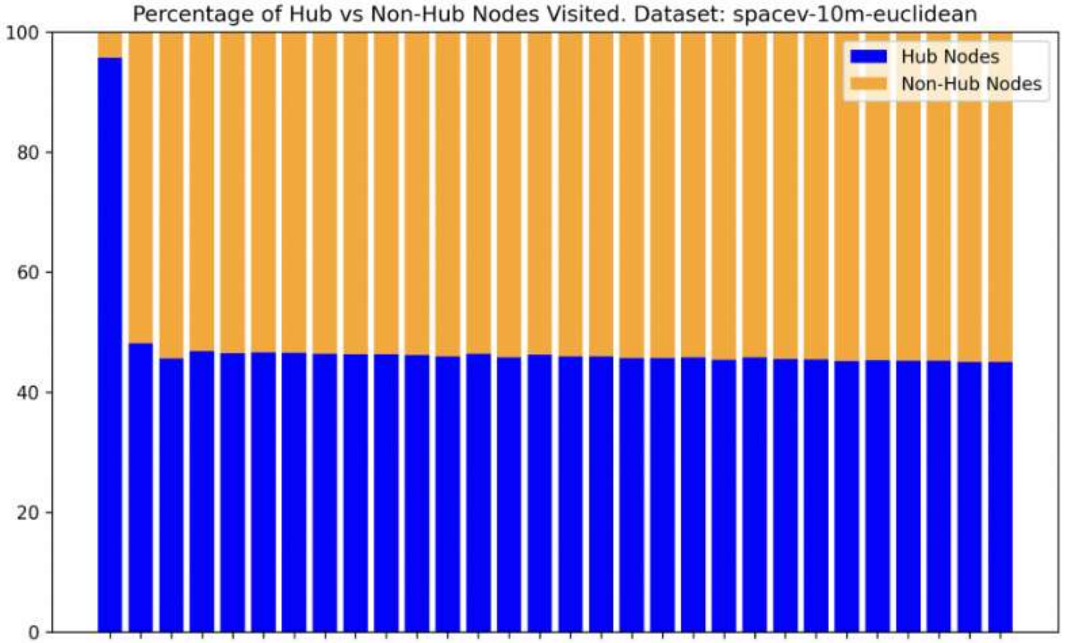


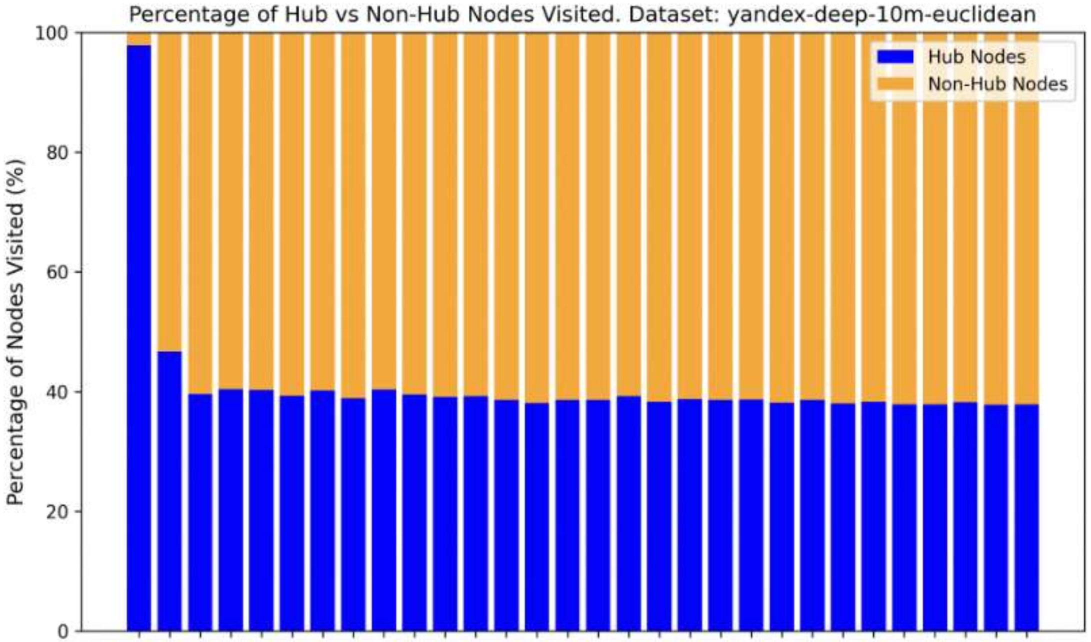
Interval index

Empirical Evidence

❑Hub-Highway Nodes Enable Fast Traversal

Percentage of nodes visted (%)





Interval index

Queries tend to concentrate in the highway structures early in search

Summary

❑Contribution

- ❖Make benchmark experiments to check the performance of HNSW
- ❖Propose Hub Highway Hypothesis and prove it

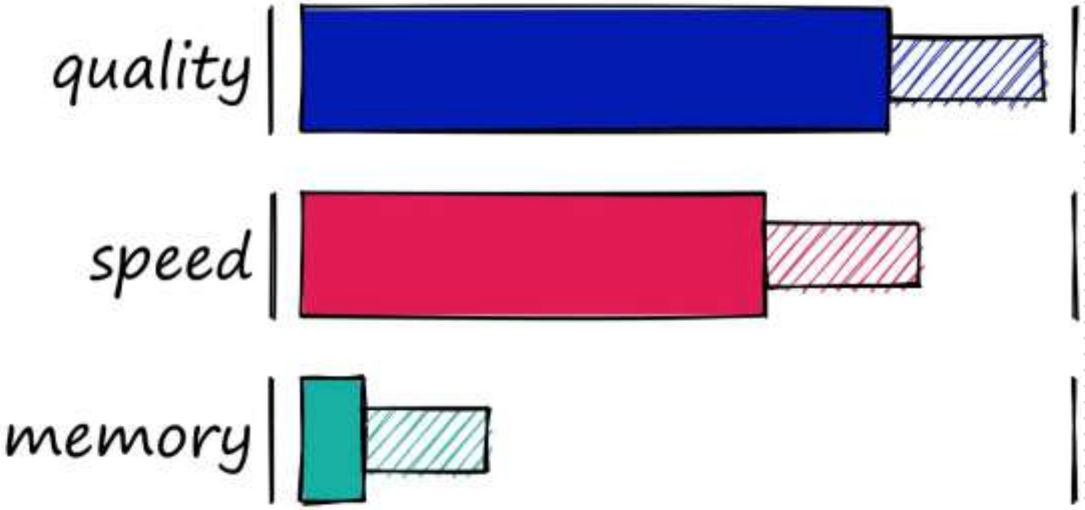
❑Drawback

- ❖Lack further innovation point
- ❖Some experimental results do not exhibit a clear trend of change with increasing dimensionality

Background

❑Brief introduction of HNSW (Hierarchical Navigable Small World)

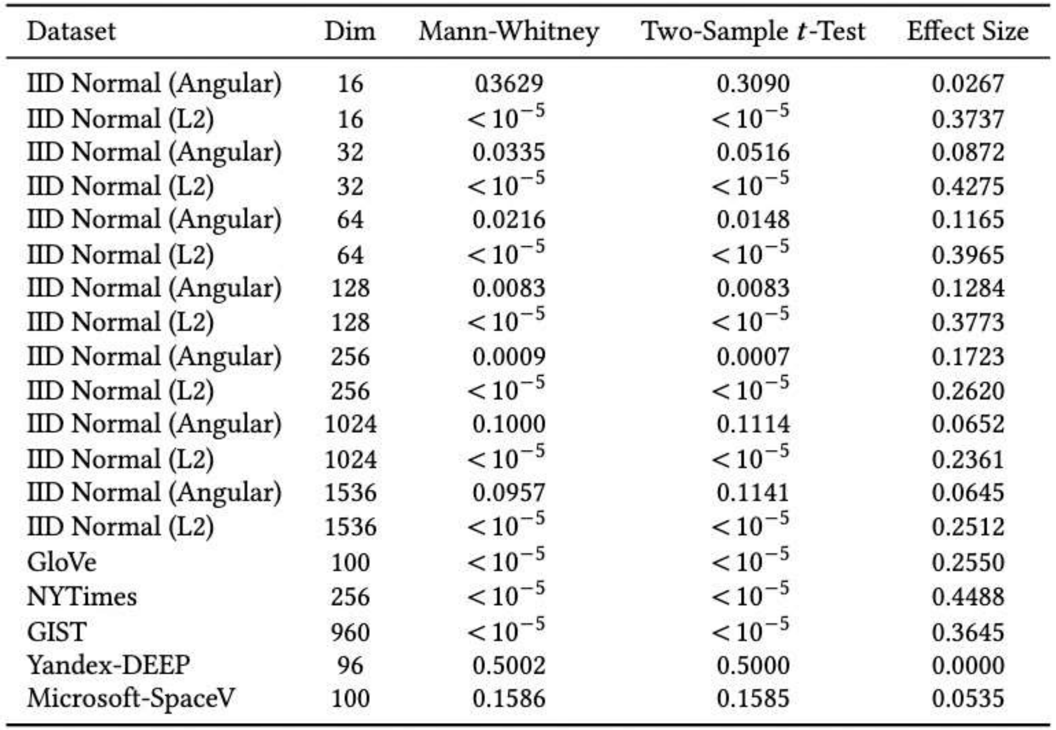
- ❖Reach the “highway” in the hierarchical layer in search
  - ➢With the hierarchical layer, HNSW performs well and is widely used in ANN search



HNSW performs well in ANN search

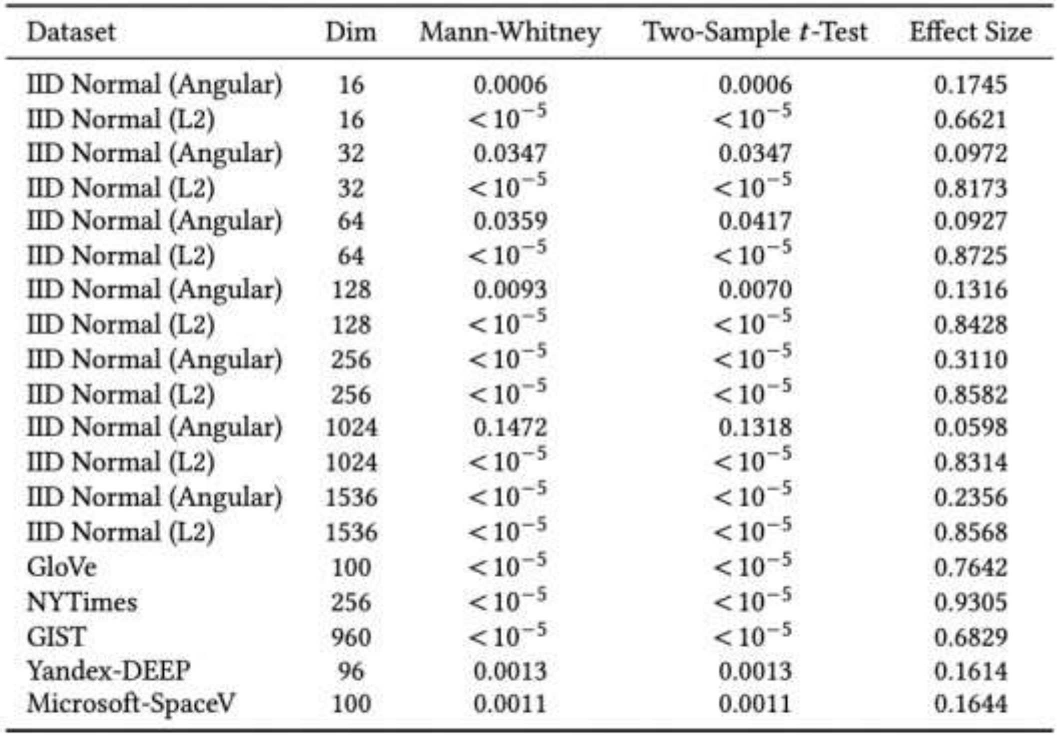
Empirical Evidence

❑Connectivity between hub nodes



Dataset	Dim	Mann-Whitney	Two-Sample t-Test	Effect Size
IID Normal (Angular)	16	0.3629	0.3090	0.0267
IID Normal (L2)	16	$< 10^{-5}$	$< 10^{-5}$	0.3737
IID Normal (Angular)	32	0.0335	0.0516	0.0872
IID Normal (L2)	32	$< 10^{-5}$	$< 10^{-5}$	0.4275
IID Normal (Angular)	64	0.0216	0.0148	0.1165
IID Normal (L2)	64	$< 10^{-5}$	$< 10^{-5}$	0.3965
IID Normal (Angular)	128	0.0083	0.0083	0.1284
IID Normal (L2)	128	$< 10^{-5}$	$< 10^{-5}$	0.3773
IID Normal (Angular)	256	0.0009	0.0007	0.1723
IID Normal (L2)	256	$< 10^{-5}$	$< 10^{-5}$	0.2620
IID Normal (Angular)	1024	0.1000	0.1114	0.0652
IID Normal (L2)	1024	$< 10^{-5}$	$< 10^{-5}$	0.2361
IID Normal (Angular)	1536	0.0957	0.1141	0.0645
IID Normal (L2)	1536	$< 10^{-5}$	$< 10^{-5}$	0.2512
GloVe	100	$< 10^{-5}$	$< 10^{-5}$	0.2550
NYTimes	256	$< 10^{-5}$	$< 10^{-5}$	0.4488
GIST	960	$< 10^{-5}$	$< 10^{-5}$	0.3645
Yandex-DEEP	96	0.5002	0.5000	0.0000
Microsoft-SpaceV	100	0.1586	0.1585	0.0535

P95 threshold



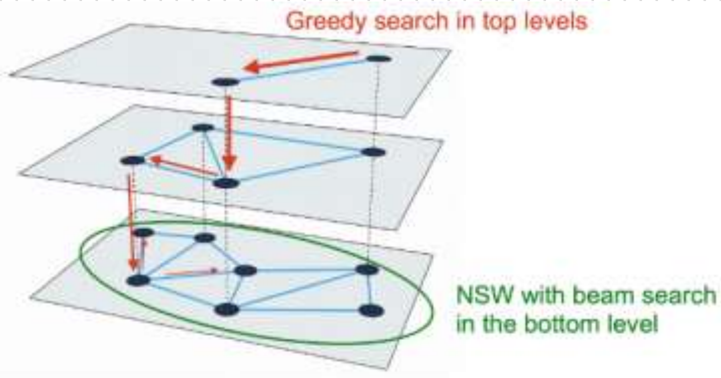
Dataset	Dim	Mann-Whitney	Two-Sample t-Test	Effect Size
IID Normal (Angular)	16	0.0006	0.0006	0.1745
IID Normal (L2)	16	$< 10^{-5}$	$< 10^{-5}$	0.6621
IID Normal (Angular)	32	0.0347	0.0347	0.0972
IID Normal (L2)	32	$< 10^{-5}$	$< 10^{-5}$	0.8173
IID Normal (Angular)	64	0.0359	0.0417	0.0927
IID Normal (L2)	64	$< 10^{-5}$	$< 10^{-5}$	0.8725
IID Normal (Angular)	128	0.0093	0.0070	0.1316
IID Normal (L2)	128	$< 10^{-5}$	$< 10^{-5}$	0.8428
IID Normal (Angular)	256	$< 10^{-5}$	$< 10^{-5}$	0.3110
IID Normal (L2)	256	$< 10^{-5}$	$< 10^{-5}$	0.8582
IID Normal (Angular)	1024	0.1472	0.1318	0.0598
IID Normal (L2)	1024	$< 10^{-5}$	$< 10^{-5}$	0.8314
IID Normal (Angular)	1536	$< 10^{-5}$	$< 10^{-5}$	0.2356
IID Normal (L2)	1536	$< 10^{-5}$	$< 10^{-5}$	0.8568
GloVe	100	$< 10^{-5}$	$< 10^{-5}$	0.7642
NYTimes	256	$< 10^{-5}$	$< 10^{-5}$	0.9305
GIST	960	$< 10^{-5}$	$< 10^{-5}$	0.6829
Yandex-DEEP	96	0.0013	0.0013	0.1614
Microsoft-SpaceV	100	0.0011	0.0011	0.1644

P99 threshold

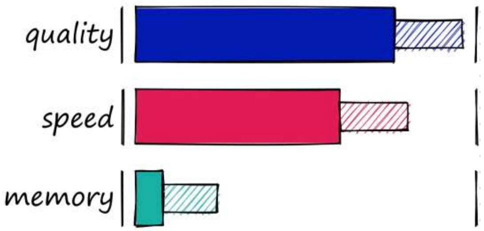
In most cases, non-hypothesis can be rejected

Introduction

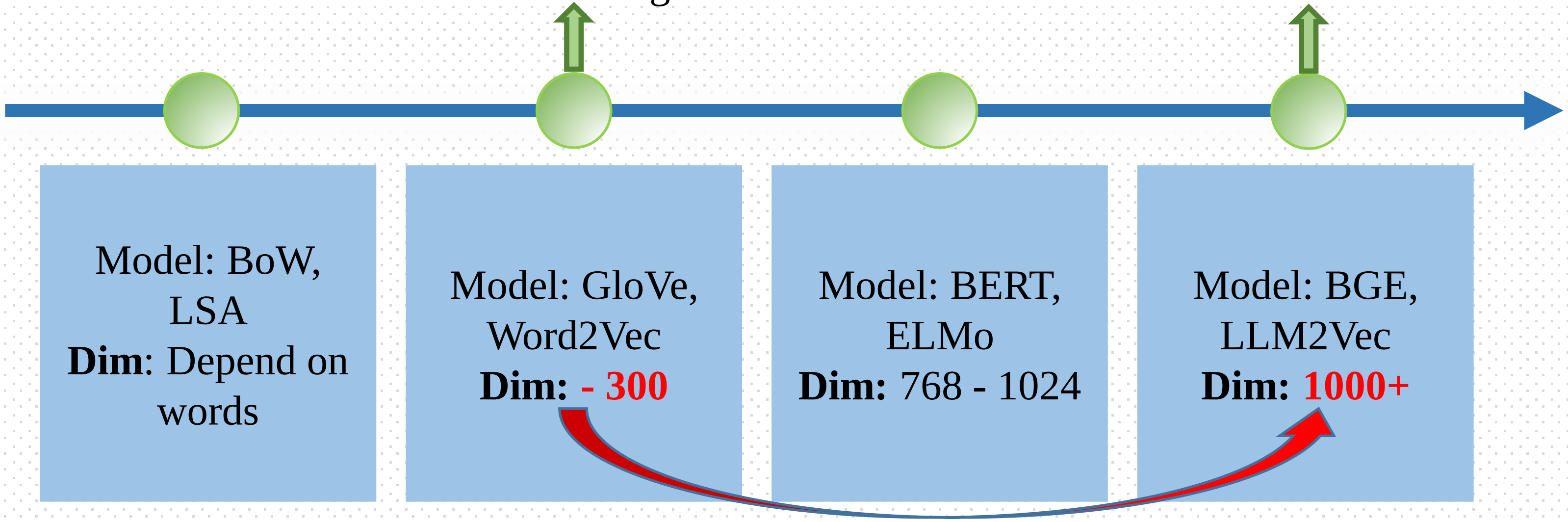
❑Dimensions of vectors become increasingly higher


Greedy search in top levels
NSW with beam search in the bottom level

HNSW Alg.



Is HNSW still effective?



Model: BoW, LSA
Dim: Depend on words

Model: GloVe, Word2Vec
Dim: - 300

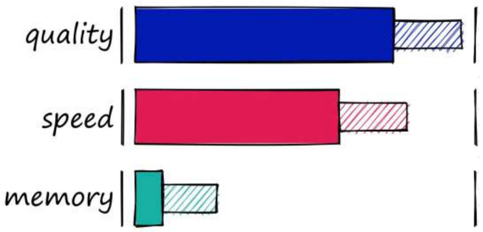
Model: BERT, ELMo
Dim: 768 - 1024

Model: BGE, LLM2Vec
Dim: 1000+

Introduction

❑Brief introduction of HNSW (Hierarchical Navigable Small World)

- ❖With the hierarchical layer, HNSW performs well

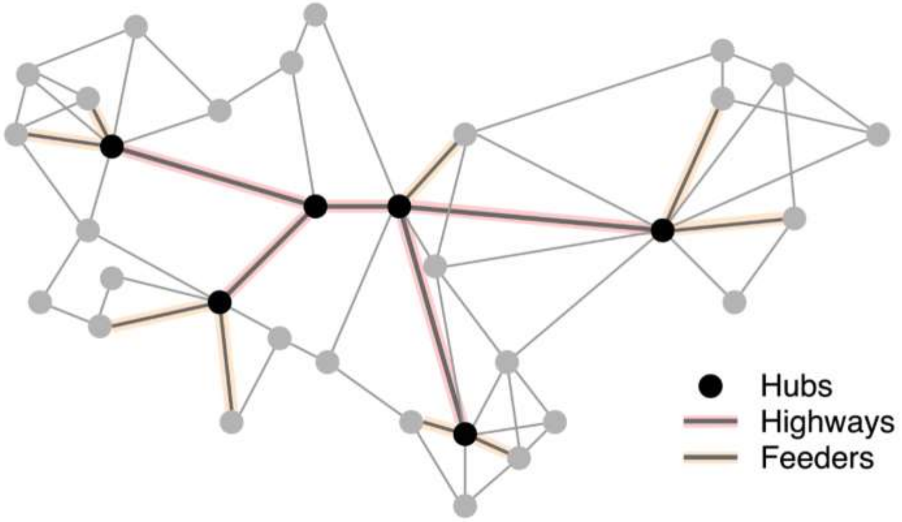


HNSW performs well in ANN search

Hub Highway Hypothesis

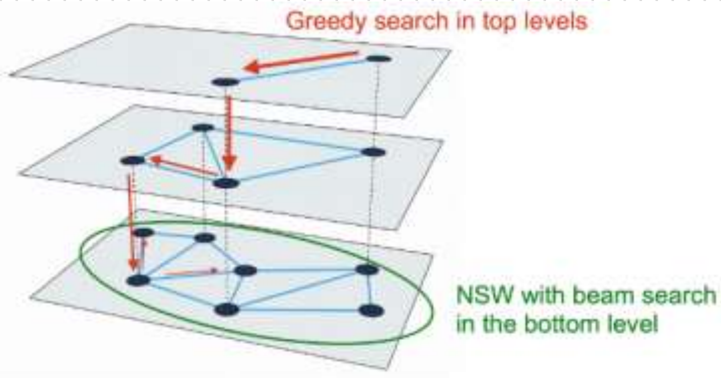
❑Methodology

- ❖Some nodes are visited by queries much more frequently than others
- ❖The hub nodes tend to be connected to each other
- ❖Queries visit many hub nodes early in the search process

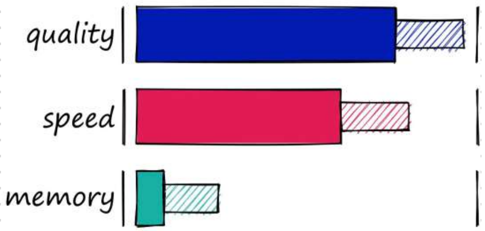


Introduction

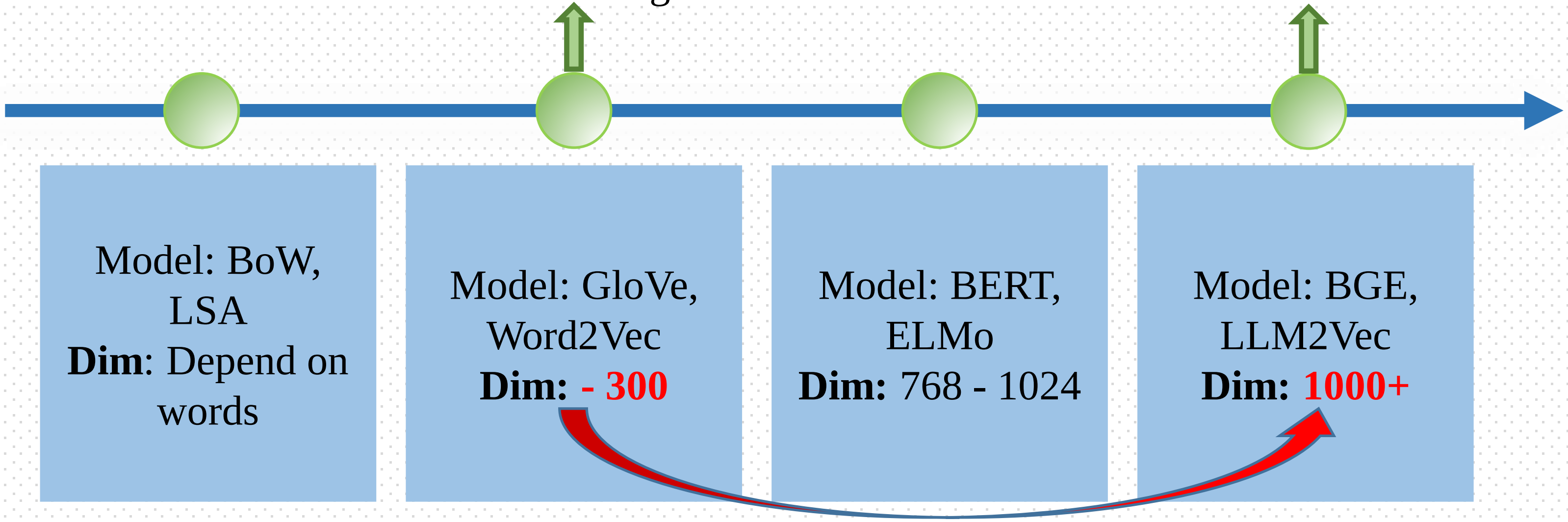
❑Dimensions of vectors become increasingly higher



HNSW Alg.



Is HNSW still effective?



Model: BoW, LSA
Dim: Depend on words

Model: GloVe, Word2Vec
Dim: - 300

Model: BERT, ELMo
Dim: 768 - 1024

Model: BGE, LLM2Vec
Dim: 1000+

Benchmarking Experiments

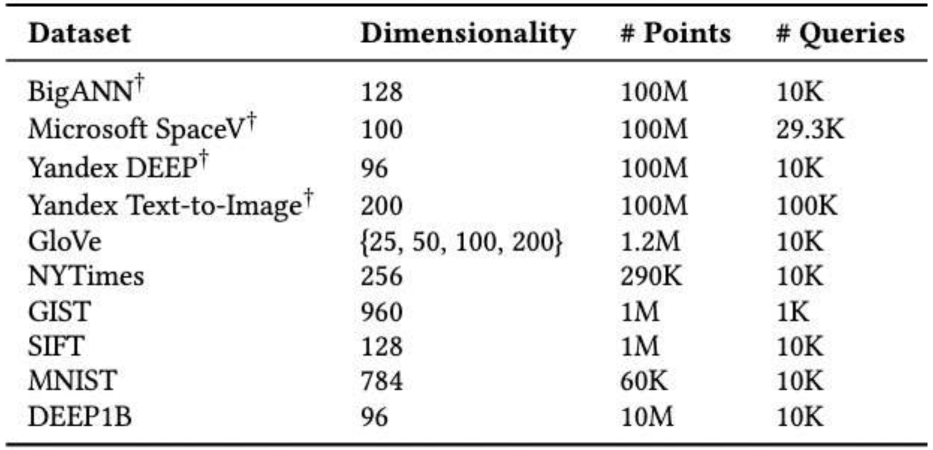
❑Goal

- ❖Evaluate performance of HNSW in high dimensional space

❑Code

- ❖HNSW: hnswlib (open source code from HNSW paper)
- ❖NSW: flatnsw (built from hnswlib)

❑Dataset



Dataset	Dimensionality	# Points	# Queries
BigANN†	128	100M	10K
Microsoft SpaceV†	100	100M	29.3K
Yandex DEEP†	96	100M	10K
Yandex Text-to-Image†	200	100M	100K
GloVe	{25, 50, 100, 200}	1.2M	10K
NYTimes	256	290K	10K
GIST	960	1M	1K
SIFT	128	1M	10K
MNIST	784	60K	10K
DEEP1B	96	10M	10K

Empirical Evidence

❑Connectivity between hub nodes

Dataset	Dim	P95 Can non-hypothesis be rejected?	P99 Can non-hypothesis be rejected?
IID Normal(Angular)	16	No	Yes
IID Normal(L2)	16	Yes	Yes
IID Normal(Angular)	32 - 256	Yes	Yes
IID Normal(L2)	32 - 256	Yes	Yes
IID Normal(Angular)	1024	No	No
IID Normal(L2)	1024	Yes	Yes
IID Normal(Angular)	1536	No	Yes
IID Normal(L2)	1536	Yes	Yes

In most cases, non-hypothesis can be rejected

Overview

❑Intro

- ❖趋势：LLM等应用让用到的向量维度越来越高，但是大家用的方法还是遵循着过去的惯性 – 在高维场景下一些低维的算法可能不适用
- ❖简单介绍HNSW算法与NSW算法之间的区别

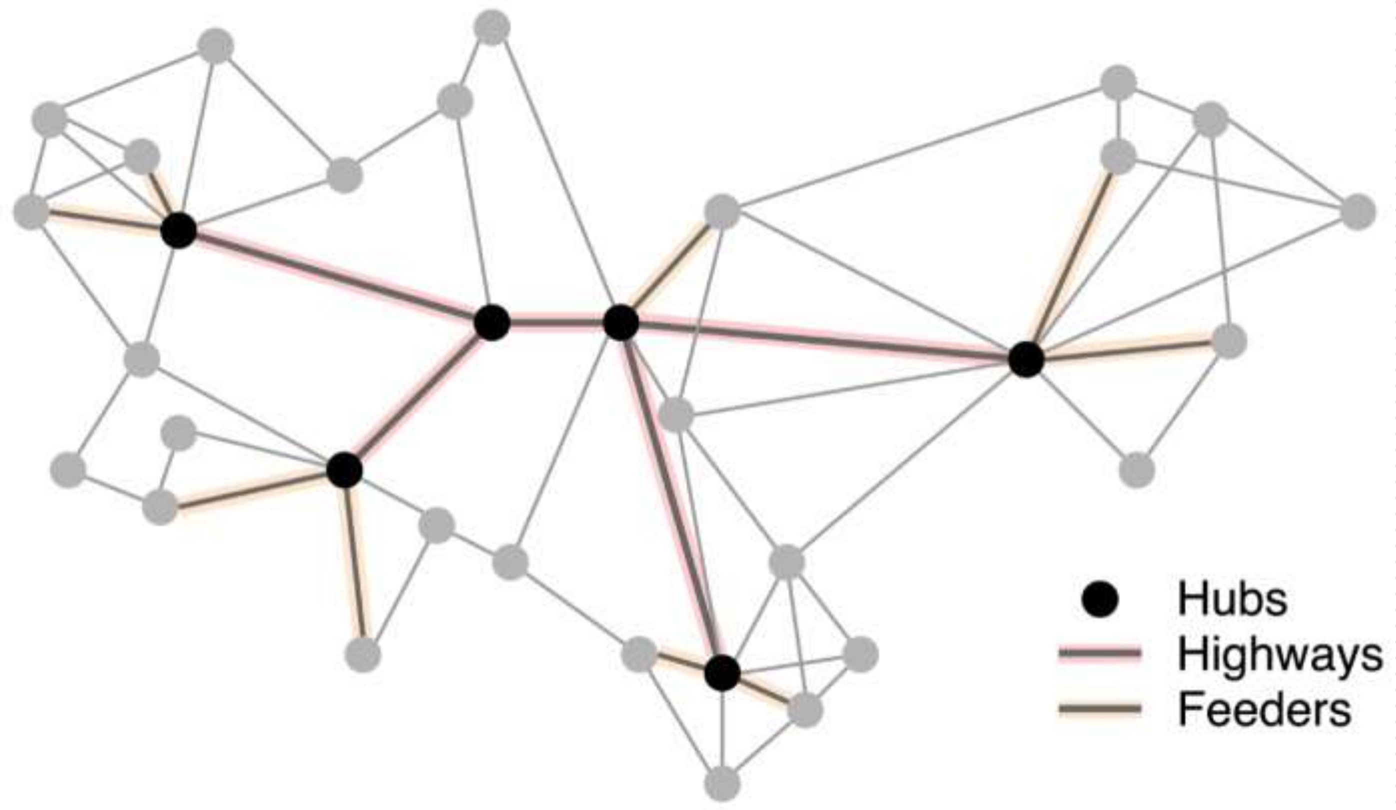
❑解释原因 – Hub

- ❖实验证明Hub存在
- ❖实验证明Hub之间的联通性很高(不直观，可以略过)
- ❖实验证明搜索时先搜索到Hub向量

❑展示结果

- ❖NSW在高维情况下确实和HNSW相差不大

总结与讨论


● Hubs
Highways
Feeders