PROBABILISTIC KERNEL FUNCTION FOR FAST ANGLE TESTING

Kejing Lu*, Chuan Xiao**, Yoshiharu Ishikawa***
*Yamanashi University, Japan
**Osaka University & Nagoya University, Japan
***Nagoya University, Japan

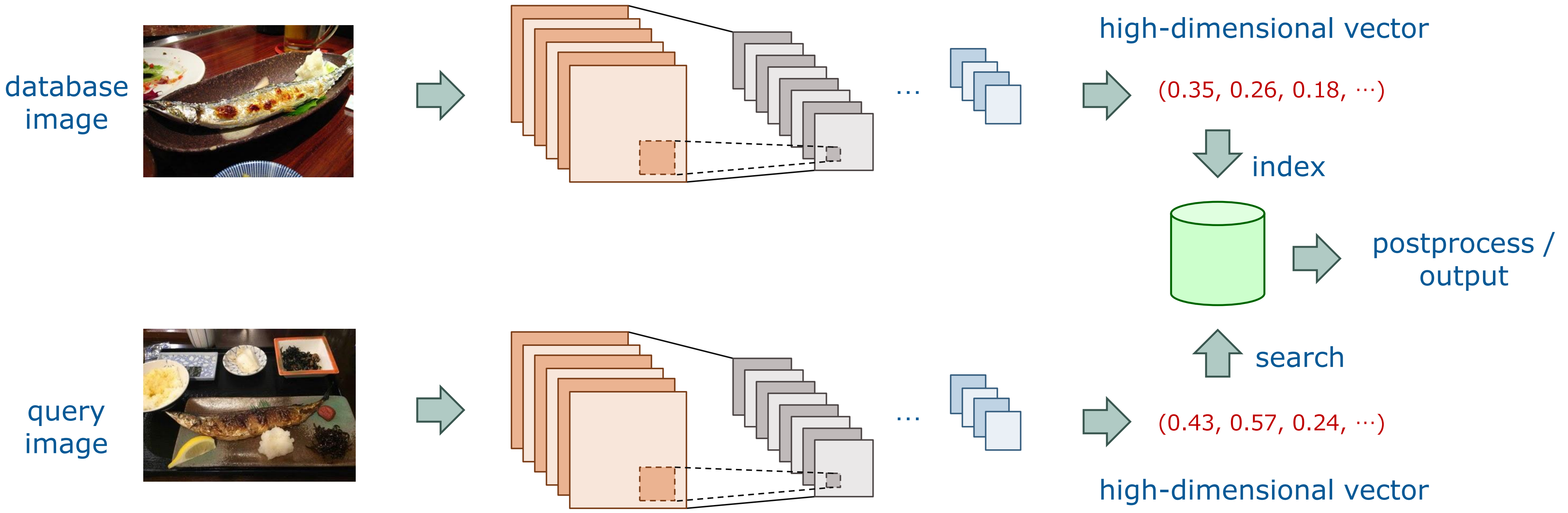
Presenter: Kejing Lu

Outline

- Introduction
- Techniques
- Applications
- Recent Progress & Conclusion

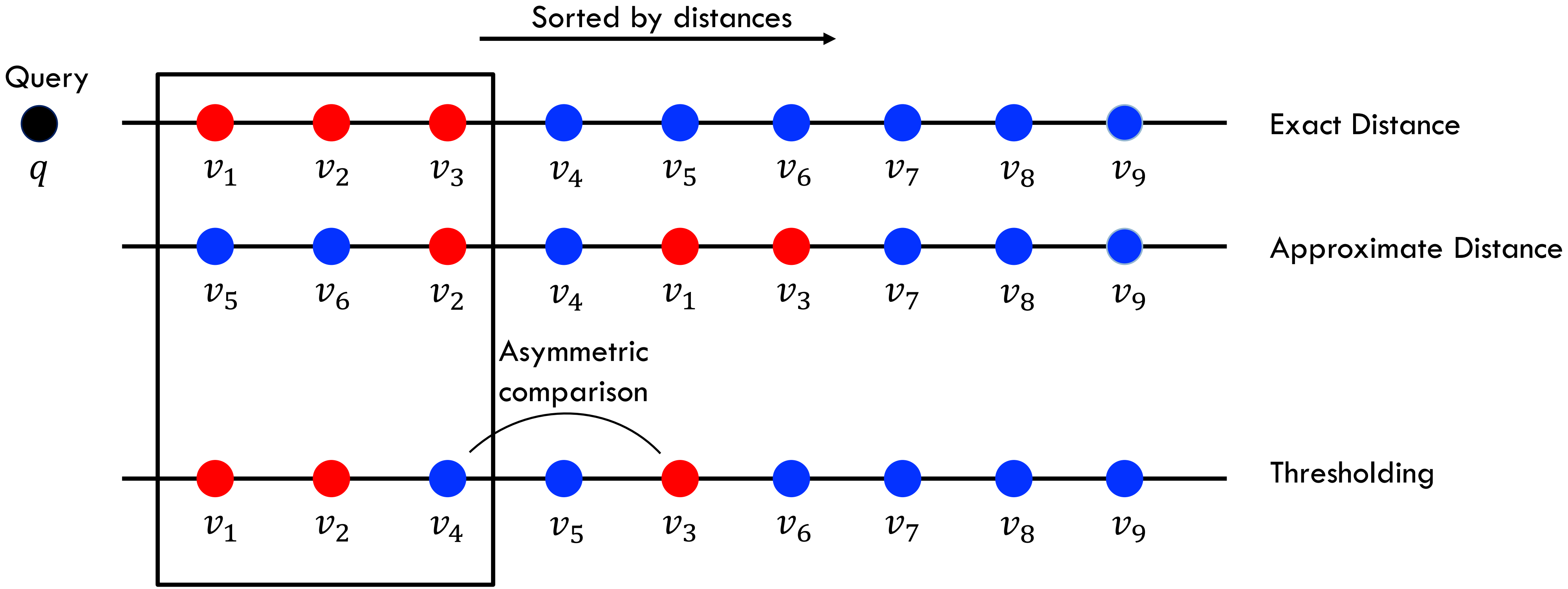
Top-k ANNS in High-Dimensional Spaces

- Downstream Tasks: Recommender Systems, Image Retrieval, RAG, etc.



Challenge: balancing the tradeoff between efficiency and accuracy

Angle-Thresholding Pattern



- The threshold quickly approaches the true top-k boundary.
- **A dedicated function is needed for asymmetric comparison.**

Outline

- Introduction
- Techniques
- Applications
- Recent Progress & Conclusion

Probabilistic Kernel Functions (PKF)

- Design goals
  - We design two probabilistic kernel functions K_S^1 and K_S^2, where K_S^1 is for angle comparison and K_S^2 is for angle thresholding.
    - If $\langle q, v_1 \rangle < \langle q, v_2 \rangle$, then $K_S^1(q, v_1) < K_S^1(q, v_2)$ with high success probability.
    - If $\langle q, v \rangle > \cos\theta$, then $K_S^2(q, v) > \cos\theta$ with high success probability.

- Some requirements for the design
  - The computation of K_S^1 and K_S^2 should be $o(d)$, where d is data dimension.
  - The success probability should be at least 0.5.
  - The computation should be parallel-friendly.

Probabilistic Kernel Functions (PKF)

- Notations
  - q and v are two vectors on the $(d-1)$-dimensional sphere.
  - H is a random rotation matrix and S is a fixed set of m reference points.
  - $Z_S(v) = \arg\max_{u \in S}\langle u, v\rangle$ and $A_s(v) = \langle v, Z_S(v)\rangle$.

Reference angle (cosine)

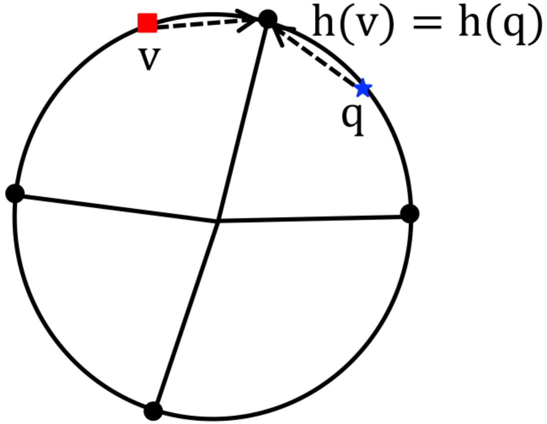
- Definitions of K_s^1 and K_s^2
  - $K_S^1(q, v) = \langle v, Z_{HS}(q)\rangle$
  - $K_S^2(q, v) = \langle Hq, Z_S(Hv)\rangle / A_S(Hv)$
- Why are they efficient
  - $\langle v, Z_{HS}(\cdot)\rangle$ and $A_s(Hv)$ can be precomputed offline.
  - $\langle Hq, Z_S(Hv)\rangle$ can be computed very efficiently.

Properties of K_s^1 and K_s^2

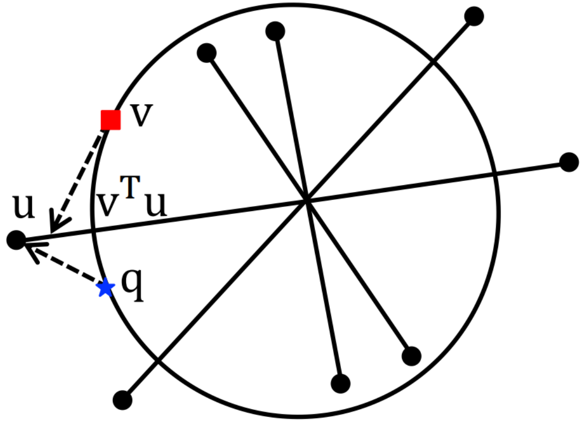
- Angle sensitivity
  - Let $\phi_1, \phi_2 \in (0, \pi)$ and let $\theta \in (0, \pi)$ be an arbitrary angle threshold. A probabilistic kernel function $K(q, v)$ is called angle-sensitive if it satisfies the following properties.
    - If $\cos\theta \leq \cos\phi_1 = \langle q, v \rangle$, then $\Pr[K(q, v) \geq \cos\theta] \geq p_1(\phi_1)$.
    - If $\langle \mathrm{q}, v \rangle = \cos\phi_2 < \cos\theta$, then $\Pr[K(q, v) \geq \cos\theta] < p_2(\phi_2)$.
    - $p_2(\phi_2)$ is strictly decreasing in ϕ_2 and $p_1(\phi_1) > p_2(\phi_2)$ when $\phi_1 < \phi_2$.
- Main results
  - $F_{K_S^1(q,v)}(x \mid A_s(q) = \cos\psi)$ has a closed form.
  - K_S^2 is an angle-sensitive function.
- Key observation
  - The key quantity is the reference angle $A_s(v)$: the smaller it is, the better the estimation.

Why Our Projection Structure Is Better

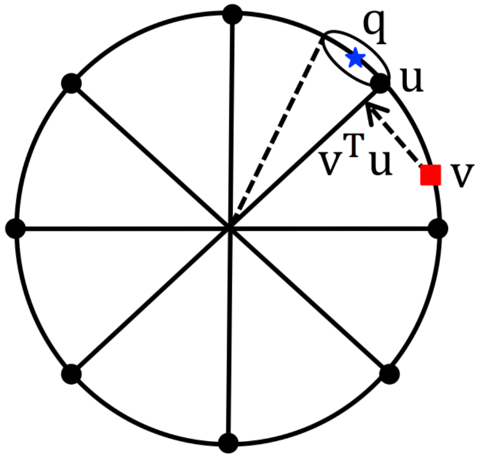
- Key insights
  - Our guarantees are exact and non-asymptotic.
  - Gaussian projections are suboptimal because they do not minimize the reference angle.
  - The optimal configuration of S corresponds to the best covering of the sphere.



Cross-polytope LSH



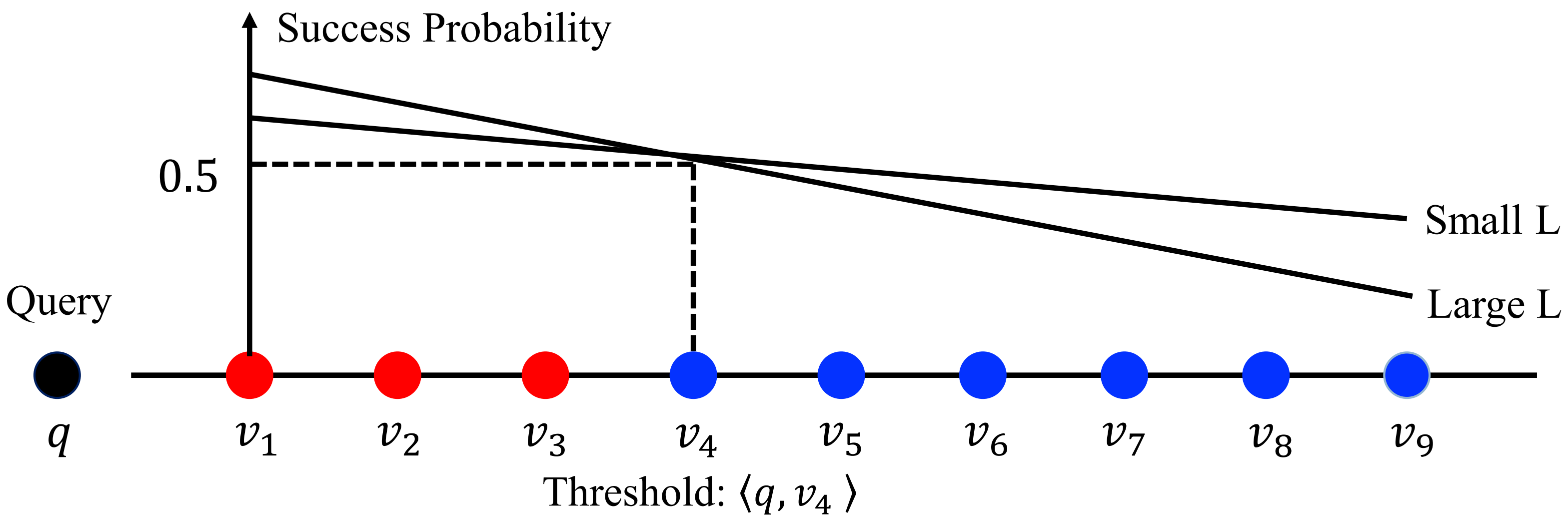
Gaussian-based Projection



Our Projection (KS1)

Structure of Reference Points

- Key insights
  - The more projection vectors there are, the more accurate our estimation becomes.
  - The original space is partitioned into L subspaces, each of which contains m sub-vectors.
  - The sub-vectors are expected to be uniformly distributed, and a configuration based on multiple randomly rotated cross-polytopes is preferred.



Comparison with LSH/VQ

- Comparison with Locality Sensitive Hashing (LSH)
  - vs. SimHash
    - SimHash aggregates multiple random projections while PKF uses one selected projection.
    - It can be shown that, under the same space budget, PKF provides a more accurate estimation.
  - vs. Cross-polytope LSH
    - PKF shows that FJLT is not merely a practical transform, but also has some theoretical justification.

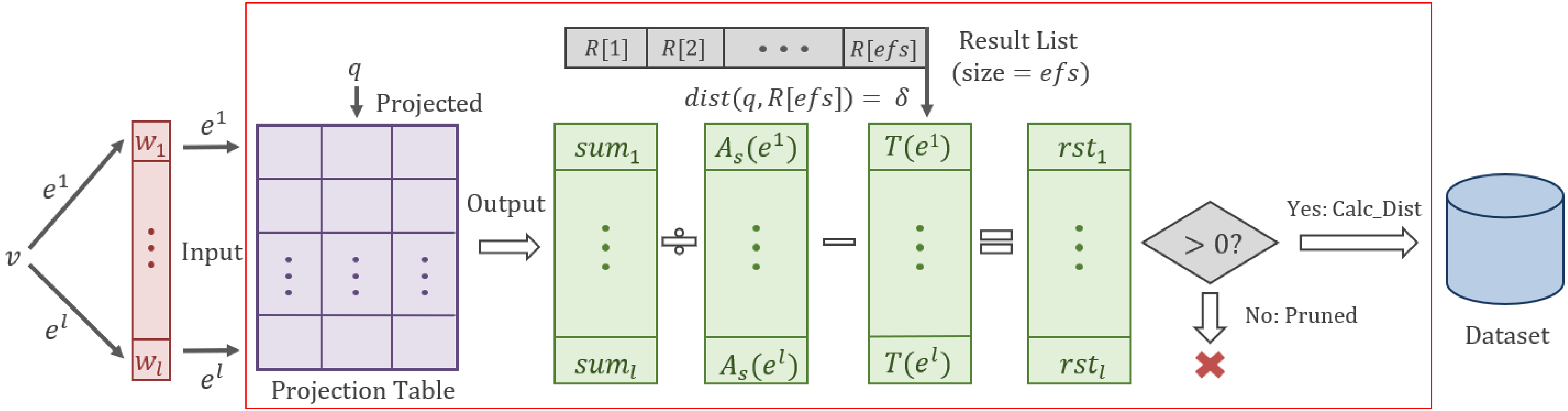
- Comparison with vector quantization (VQ)
  - Key differences
    - Small quantization error does not necessarily lead to a more accurate test.
    - PKF is not as sensitive to the number of partitions as VQ.

Outline

- Introduction
- Techniques
- Applications
- Recent Progress & Conclusion

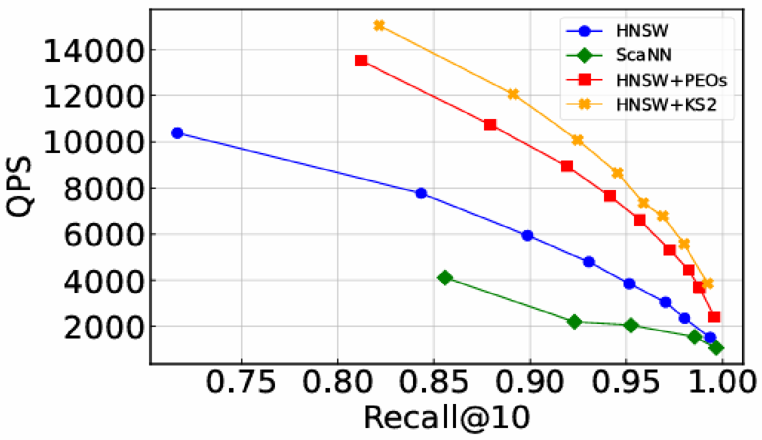
Application to Similarity Graphs

- Why KS2 is suited for similarity graphs
  - The threshold induced by the current result list quickly approaches the true top-k boundary
  - The KS2 test can be vectorized across neighbors and fits graph traversal well.

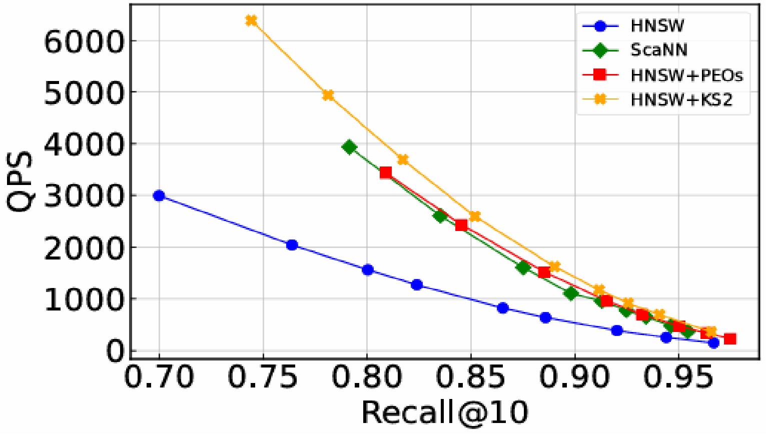


KS2 test

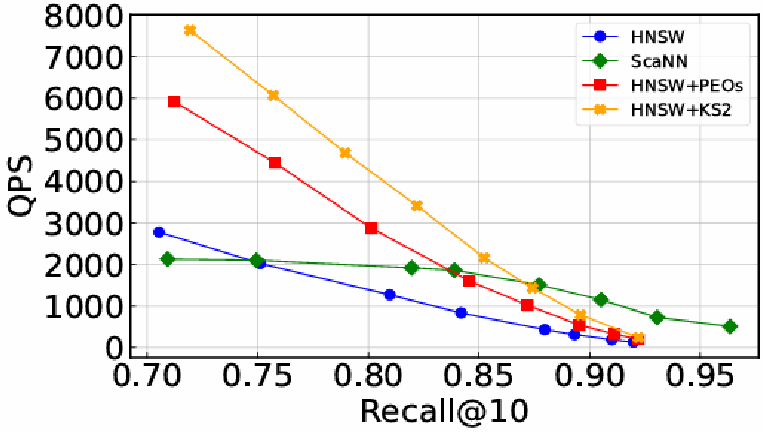
ANNS Performance



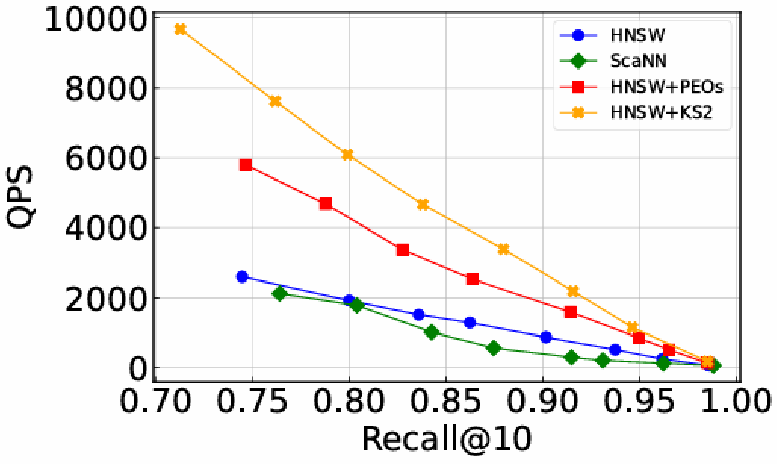
SIFT



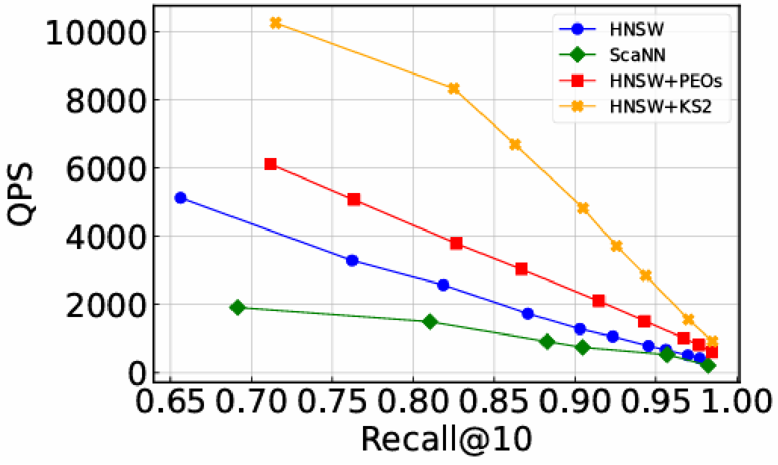
GloVe1M



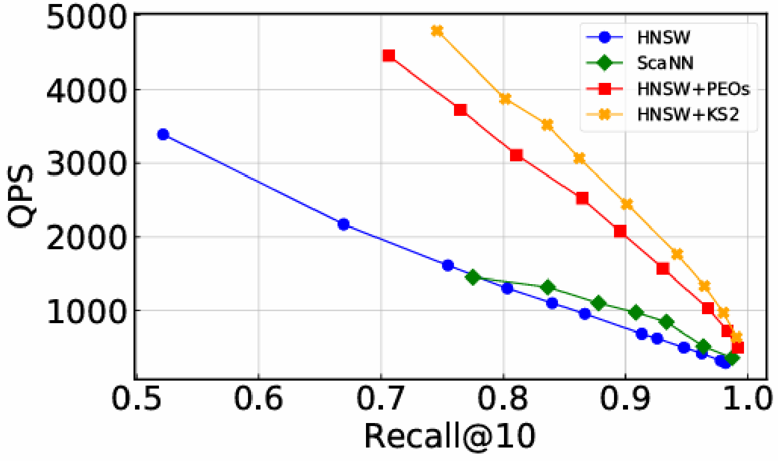
Word



GloVe2M

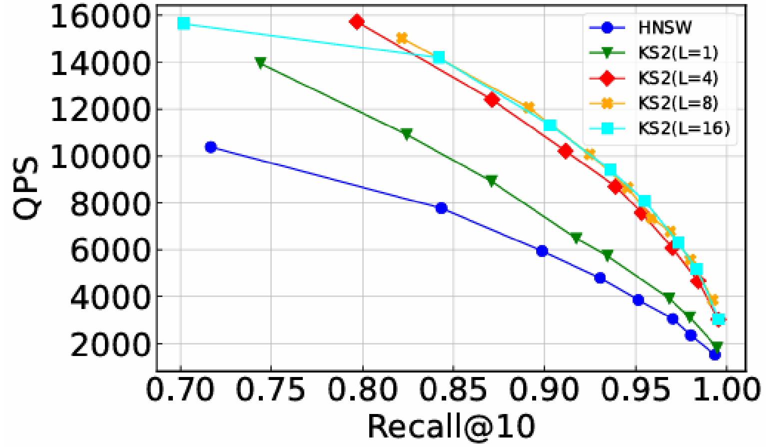


Tiny

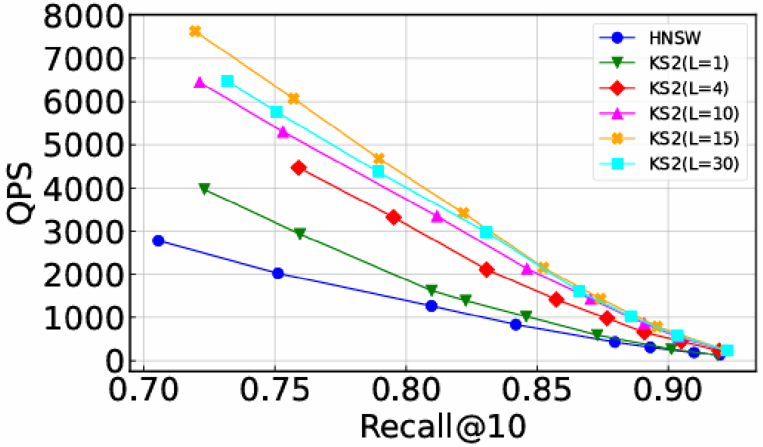


GIST

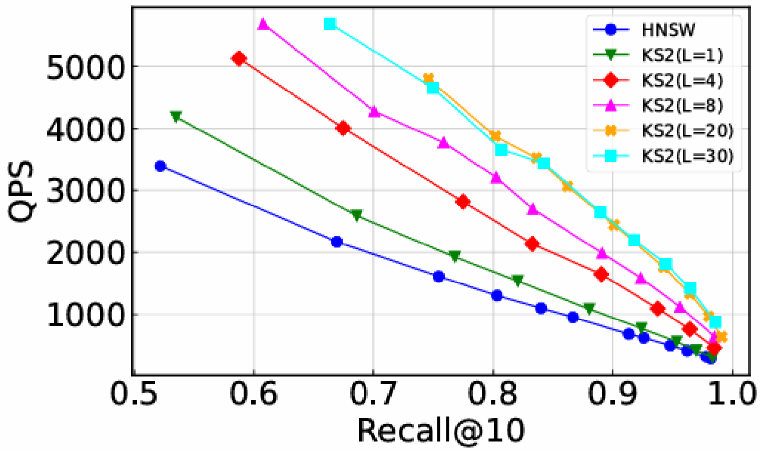
Effect of L



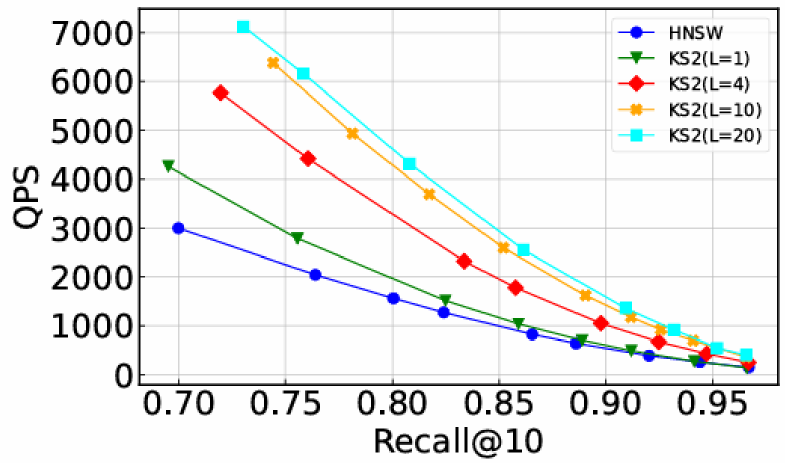
SIFT



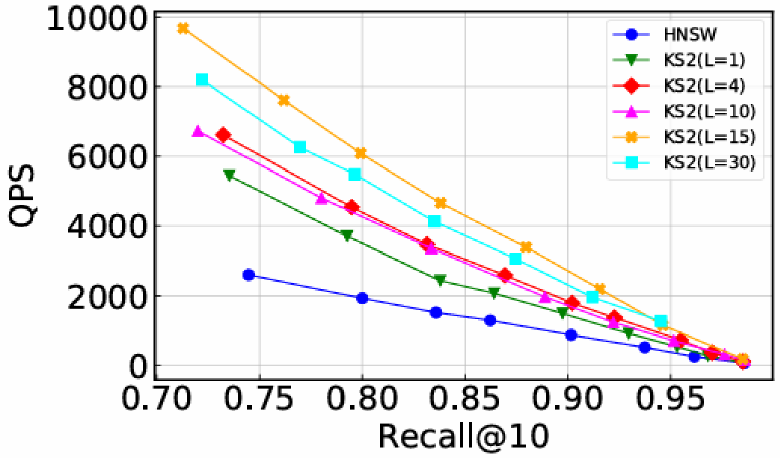
Word



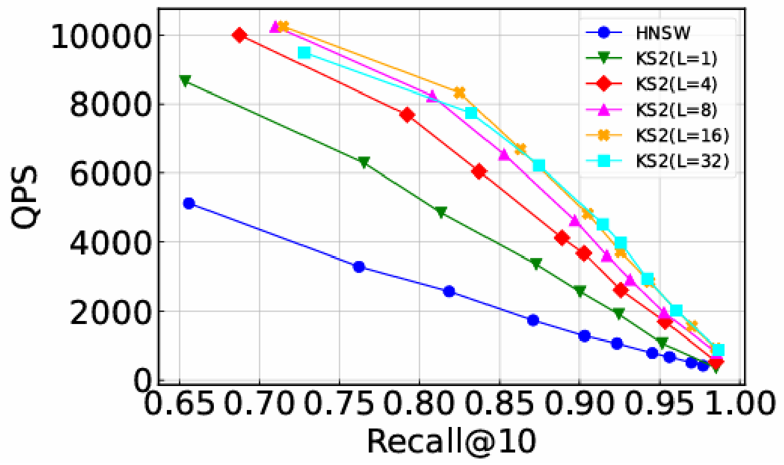
GIST



GloVe1M



GloVe2M

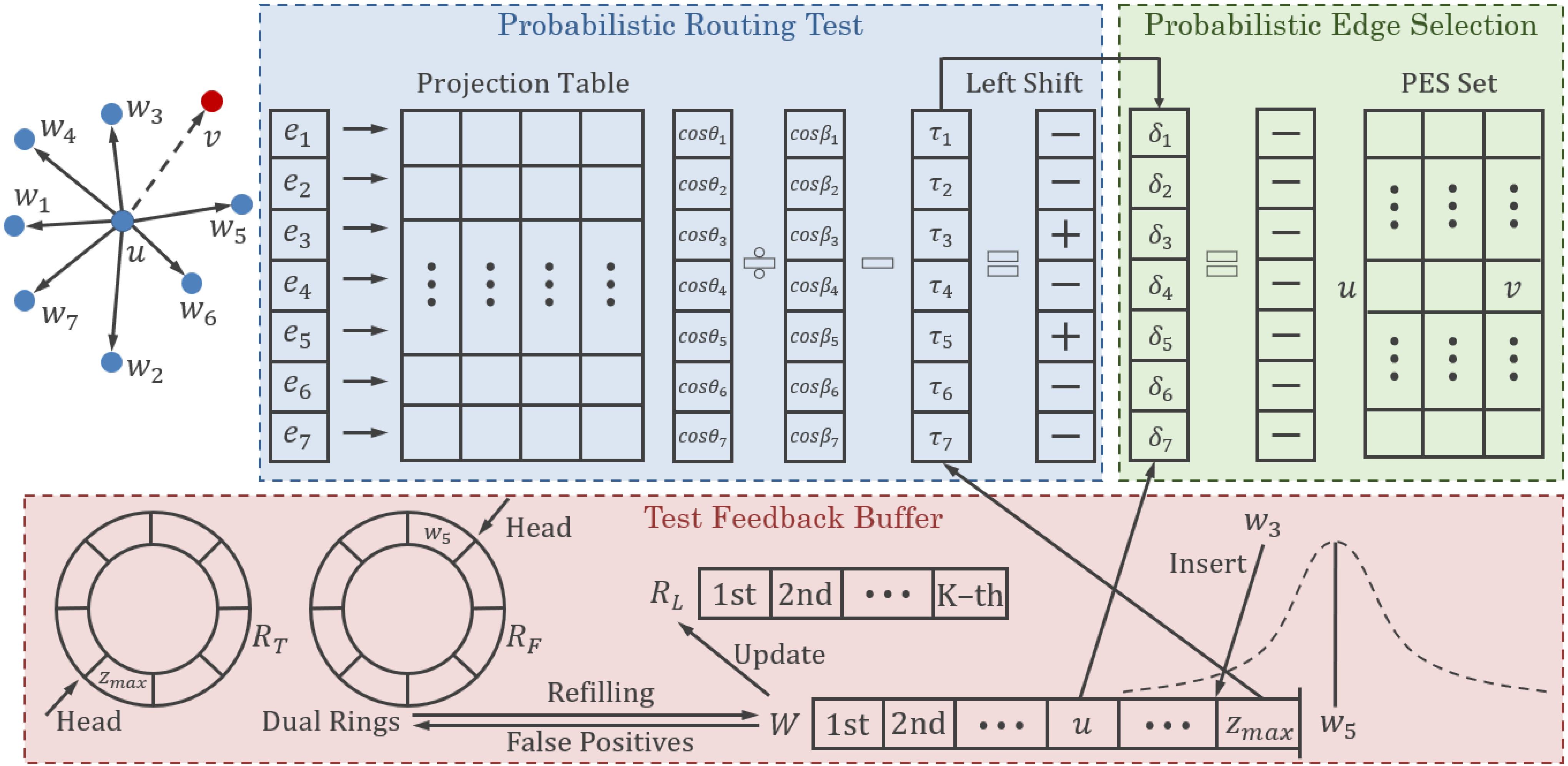


Tiny

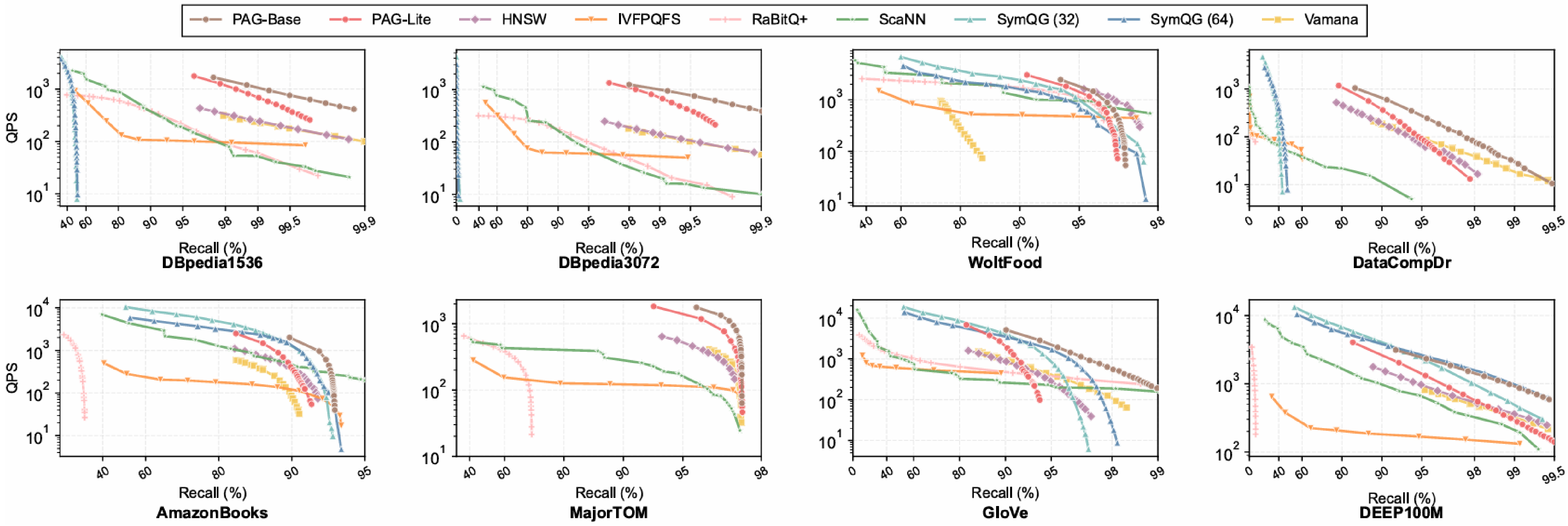
Outline

- Introduction
- Techniques
- Applications
- Recent Progress & Conclusion

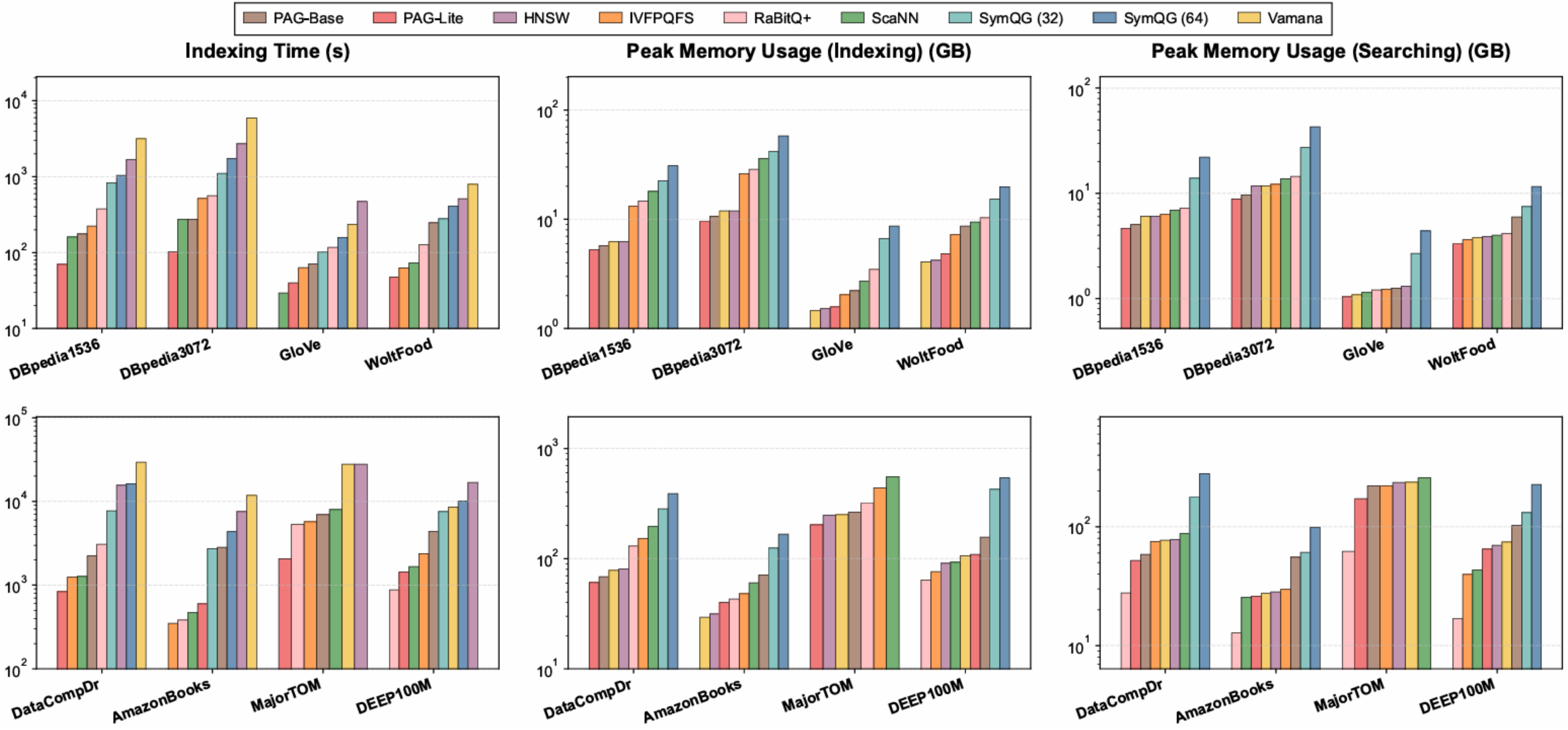
Projection-Augmented Graph (PAG)



ANNS Performance



Indexing Time and Memory Footprint



Conclusion & Future Work

- Methodological Contribution
  - We propose two kernel functions for angle comparison and angle thresholding.
  - Both provide probabilistic guarantees.
- Practical Utility
  - We apply KS2 to the routing test in similarity graphs.
  - Combining graph search with KS2 significantly improves search performance.
- Future Work
  - We will extend the proposed functions to other metrics, such as inner product.
  - We are also seeking practical deployment opportunities and welcome collaboration with industry partners.

THANK YOU!