Scalable Neural Graph Retrieval

(Research Statement)

Indradyumna Roy

My research focuses on developing a unified framework for graph retrieval, which returns a top-K list of relevant graphs or subgraphs in response to a query graph. Graphs serve as a common formalism across various data modalities, such as natural language queries, parse trees, knowledge graphs, and images, facilitating seamless inter-operation during retrieval. To align and connect graphs from these diverse modalities, neural graph models provide superior node and edge representations, transcending traditional schema unification. An effective neural graph retrieval model should ensure (i) *high retrieval accuracy*, (ii) *scalability* while handling large-sizes or large-number of graphs, and (iii) *interpretability* to justify its responses through alignment-based explanations.

Against this backdrop, my current research explores the following directions:

1. **Designing and learning relevance models for graph retrieval:** Designing neural models for graph retrieval and alignment, subgraph matching, designing neural analogs to combinatorial graph optimizations.
2. **ANN Support for Graph Search Models:** Designing Approximate Nearest Neighbor (ANN) techniques compatible with (symmetric and asymmetric) graph similarity functions to enable sublinear time retrieval.

1 Designing and Learning Relevance Models for Graph Retrieval

Defining relevance in graph search is inherently complex, traditionally relying on computationally intensive combinatorial methods. My prior works [14, 20, 21, 23, 25, 26] introduced both early and late interaction neural architectures for Subgraph Isomorphism (SubIso) [20, 21, 26], Maximum Common Subgraph (MCS) [25], and Graph Edit Distance (GED) [14, 23], all trainable under distant supervision using only pairwise relevance judgments.These models are end-to-end differentiable and leverage Sinkhorn-based [8, 18] solvers to optimize transportation objectives, enabling task-specific scoring through expressive cost functions and *interpretable* relevance via soft alignment approximations derived from the resulting transport maps. Crucially, our neural formulation of GED is the first to support variable node and edge edit costs—bridging a longstanding gap between neural models and classical combinatorial solvers—and enables a unified framework [14] for modeling both symmetric and asymmetric graph similarity notions, including SubIso, MCS, and equal-cost GED [3, 4]. Alongside this, we identified [13] a pervasive data leakage issue in widely used graph benchmarks, stemming from GNNs' permutation invariance over structurally isomorphic yet non-identical graphs. Building on a theoretical insight that GED alignments remain stable across a wide range of cost settings, we introduced [24] scalable data augmentation and principled adversarial testing protocols to systematically address this issue.

Future work: Moving forward, I am particularly interested in advancing neural subgraph search in large-scale graphs, where identifying and ranking relevant substructures within a single, massive corpus (*e.g.*, knowledge graphs) poses distinct challenges. A recent step in this direction is our work [22] on clique number estimation, where we propose a neural model for predicting the size of the maximum clique. Building on this, my next goal is to develop differentiable modules capable of retrieving compact subgraphs for a given query—whether a graph or a set of keywords—where different query components may align with distinct substructures across the corpus graph. The objective is to identify a single, coherent subgraph that jointly covers all query components while promoting structural proximity among the matched regions. This opens a rich research direction on matching complex queries to semantically relevant yet minimally dispersed regions of large graphs, with compactness serving as a regularizing prior for both interpretability and relevance.

2 ANN Support for Graph Search Models

Ranking corpus graphs by similarity scores for a given query graph can be prohibitively expensive for large databases. This issue can be mitigated through graph indexing and approximate nearest neighbor search

(ANNS) techniques such as locality-sensitive hashing (LSH). However, most graph matching models require customized scoring functions that are not LSH-compatible. My recent work, FourierHashNet [27], addresses the LSHability of an Order Embedding based asymmetric distance, previously used to detect subgraph isomorphism [17]. FourierHashNet is an asymmetric LSH which transforms the Order Embedding distance into a bounded dominance similarity measure, applies a Fourier transformation, and uses importance-sampled estimates to approximate the expectation of inner products in the frequency domain. This renders Order Embeddings-based relevance measures [29] LSH-compatible.

Future work: In this line of research, I am eager to continue working on three key areas:

• *Extension to shift-invariant scoring functions:* Our proposed asymmetric LSH framework in FourierHashNet [27] can be extended to support any shift-invariant scoring function, potentially enabling LSH compatibility for a broad class of relevance models. This includes, for instance, volume-based scores from Box Embeddings [7] and facility location scores used in ColBERT [16], which offer strong modeling capabilities but remain underutilized in industrial recommender systems due to the absence of efficient indexing mechanisms. To explore this direction, I plan to construct targeted benchmarks and evaluate the effectiveness of our Fourier-based featurization in indexing these shift-invariant scoring functions.

• *LSH for transportation-based graph similarity:* While my earlier work in FourierHashNet [27] enabled sublinear-time retrieval for subgraph isomorphism using order embeddings, extending LSH to more expressive, alignment-driven similarities remains an open problem. Existing LSH methods either target asymmetric distances in Euclidean space [19, 27, 28], or are limited to Earth Mover's Distance (EMD) with symmetric costs [1, 2, 5, 6, 10–12, 15]. No known approach handles EMD with asymmetric cost structures, which is a requirement in many graph retrieval tasks. Motivated by our GED framework, which unifies diverse graph similarities under a generalized transportation score, I plan to develop compatible LSH methods enabling scalable and practical alignment-aware retrieval.

• *Multi-Vector Indexing for Graph Retrieval:* My work explores multi-vector graph representations optimized using transport-based objective, motivating the need for efficient multi-vector indexing strategies that respect nuanced graph similarity. I plan to draw on insights from dense text retrieval like ColBERT [16] and SPLADE [9], which scale through token-level relevance and learned sparse expansions. In a similar vein, I aim to learn a (potentially latent) vocabulary of discriminative subgraph motifs that serve as soft "tokens" with associated relevance weights. These will support posting list-style inverted indexing, combining the accuracy of alignment-based scoring with the scalability of classical IR methods.

Broader Significance and Outlook

My work on scalable graph retrieval and indexing aims to support next-generation retrieval systems that are both structure-aware and system-efficient. Graphs unify diverse domains—molecules and proteins in biomedicine, knowledge graphs in QA, scene graphs in vision, and user-item networks in recommendation—where structural constraints are key to relevance. By designing retrieval models that respect these constraints while scaling to real-world corpora, my research enables richer, more reliable signals than conventional vector similarity.

At the systems level, I develop locality-sensitive hashing, multi-vector indexing, and related techniques that reduce retrieval latency under expressive similarity functions. These advances connect with vector quantization, graph-based ANN, and scalable compression, all critical for low-latency, high-throughput retrieval in practice. The resulting methods not only advance semantic search and recommender systems, but also strengthen retrieval-augmented generation (RAG) pipelines, where multi-modal, token-level retrieval is essential. By bridging structure-aware relevance with efficient indexing, my research contributes to the foundation of **next-generation retrieval models** that are both *domain-adaptive* and *system-efficient*.

References

[1] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the 19th ACM-SIAM Symposium on Discrete Algorithms (SODA '2008)*, pages 343–352, 2008.

[2] Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David Woodruff. Efficient sketches for earth-mover distance, with applications. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS '2009)*, 2009.

[3] Luc Brun, Benoit Gaüzère, and Sébastien Fourey. Relationships between graph edit distance and maximal common unlabeled subgraph. 2012.

[4] Horst Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*, 18(8):689–694, 1997.

[5] Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. An improved analysis of the quadtree for high-dimensional emd. 2020.

[6] Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. New streaming algorithms for high dimensional emd and mst. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 222–233, 2022.

[7] Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. Box embeddings: An open-source library for representation learning using geometric structures. *arXiv preprint arXiv:2109.04997*, 2021. URL https://www.iesl.cs.umass.edu/box-embeddings/main/index.html.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[9] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

[10] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samordinsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[11] Piotr Indyk. Algorithms for dynamic geometric problems over data streams. In *Proceedings of the 36th ACM Symposium on the Theory of Computing (STOC '2004)*, pages 373–380, 2004.

[12] Piotr Indyk and Nitin Thaper. Fast color image retrieval via embeddings. In *Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

[13] Eeshaan Jain, Indradyumna Roy, Saswat Meher, Soumen Chakrabarti, and Abir De. Graph edit distance evaluation datasets: Pitfalls and mitigation. In *The Third Learning on Graphs Conference*.

[14] Eeshaan Jain, Indradyumna Roy, Saswat Meher, Soumen Chakrabarti, and Abir De. Graph edit distance with general costs using neural set divergence. *Advances in Neural Information Processing Systems*, 37:73399–73438, 2024.

[15] Rajesh Jayaram, Erik Waingarten, and Tian Zhang. Data-dependent lsh for the earth mover's distance. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 800–811, 2024.

[16] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020.

[17] Zhaoyu Lou, Jiaxuan You, Chengtao Wen, Arquimedes Canedo, Jure Leskovec, et al. Neural subgraph matching. *arXiv preprint arXiv:2007.03092*, 2020.

[18] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

[19] Behnam Neyshabur and Nathan Srebro. On symmetric and asymmetric lshs for inner product search. In *International Conference on Machine Learning*, pages 1926–1934. PMLR, 2015.

[20] Vaibhav Raj, Indradyumna Roy, Ashwin Ramachandran, Soumen Chakrabarti, and Abir De. Charting the design space of neural graph representations for subgraph matching. In *The Thirteenth International Conference on*

Learning Representations.

[21] Ashwin Ramachandran, Vaibhav Raj, Indradyumna Roy, Soumen Chakrabarti, and Abir De. Iteratively refined early interaction alignment for subgraph matching based graph retrieval. *Advances in Neural Information Processing Systems*, 37:77593–77629, 2024.

[22] Indradyumna Roy, Eeshaan Jain, Soumen Chakrabarti, and Abir De. Clique number estimation via differentiable functions of adjacency matrix permutations. In *The Thirteenth International Conference on Learning Representations*, .

[23] Indradyumna Roy, Eeshaan Jain, Saswat Meher, Soumen Chakrabarti, and Abir De. Graph edit distance with general costs using neural set divergence. In *The Third Learning on Graphs Conference*, .

[24] Indradyumna Roy, Saswat Meher, Eeshaan Jain, Soumen Chakrabarti, and Abir De. Position: Graph matching systems deserve better benchmarks. .

[25] Indradyumna Roy, Soumen Chakrabarti, and Abir De. Maximum common subgraph guided graph retrieval: Late and early interaction networks. *Advances in Neural Information Processing Systems*, 35:32112–32126, 2022.

[26] Indradyumna Roy, Venkata Sai Velugoti, Soumen Chakrabarti, and Abir De. Interpretable neural subgraph matching for graph retrieval. 2022.

[27] Indradyumna Roy, Rishi Agarwal, Soumen Chakrabarti, Anirban Dasgupta, and Abir De. Locality sensitive hashing in fourier frequency domain for soft set containment search. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Anshumali Shrivastava and Ping Li. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). *NeurIPS*, abs/1405.5869, 2014. URL https://arxiv.org/pdf/1405.5869.pdf.

[29] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *ICLR 2016*, 2015. URL https://arxiv.org/pdf/1511.06361.pdf.