Research Statement

Miao Qiao

I am a researcher dedicated to bridging theory and practice in data science. My long-term research goal is to develop theory-informed and practically efficient solutions for data management and analysis, addressing the critical challenges under AI's transformative advancements in representation and prediction. During my six years at the University of Auckland, I have successfully led impactful and well-funded research initiatives. As the sole Principal Investigator, I secured and completed two major grants from prestigious New Zealand funding agencies, totaling 3.3 million NZD. These accomplishments highlight my ability to lead high-impact projects and deliver significant outcomes. Topicwise, my current and future research spans both theoretical and practical domains, with a focus mainly on two directions, *vector data management* and *graph search and analytics*. These endeavors address critical challenges in data science while laying the foundation for exploring pressing future problems in the field.

- Direction 1: Vector Data Management. Models like word2vec and node2vec represent real-world objects (e.g., documents, graphs, and images) in high-dimensional vector spaces, where semantically similar objects are mapped to vectors close to each other in a metric space. Approximate Nearest Neighbor Search (ANNS) identifies data vectors that are close to a given query vector, playing a crucial role in similarity search and powering applications such as Retrieval-Augmented Generation (RAG) and more. Graph-based methods like HNSW [16] and its variations [11, 20] have gained an increasing attention for their superior empirical performance. Our recent work, SeRF [24], tackles the Range-Filtered ANNS problem using graph-based methods. In this setting, each data vector is associated with an attribute value (e.g., timestamps or prices), and queries search for nearest neighbors constrained by a specified attribute range. To the best of our knowledge, SeRF is the first peer-reviewed solution for this problem, achieving an $\Omega(n)$ reduction in memory footprint without compromising search accuracy compared to vanilla approaches. Additionally, we have two works under submission that investigate ANNS for hybrid queries incorporating structural constraints. Despite its advancements, graph-based ANNS remains in its early stages compared to the mature relational database toolset for query processing. This highlights research opportunities to further enhance the scalability, efficiency, and functionality of graph-based ANNS solutions. It is a pressing research direction with massive research questions in both theory and practice in ANNS given the high demand of AI-based data-intensive applications in coming years.
  - Opportunities in practice. A comprehensive suite of techniques has yet to be developed to systematically and efficiently support scalable ANNS query processing with potential point or range filters across a memory hierarchy. Key research directions include i) vector management in external memory, i.e., to developing I/O-efficient techniques for querying and updating ANNS data stored in external memory, ii) multi-model search indexes and iii) real-time, adaptive ANNS systems, i.e., process dynamic, streaming data, that learn and optimize query performance on-the-fly, adapting to evolving datasets and changing user requirements.
  - Opportunities in theory. Theoretical understanding of why HNSW performs so well remains limited, leaving its adaptation to memory hierarchies shrouded in uncertainty. While some progress has been made by Indyk and Xu [10, 21], their work represents only an initial step. This gap presents significant opportunities for further exploration, particularly in developing

foundational insights that can guide the efficient adaptation of graph-based ANNS methods to modern memory architectures.

With my background in theoretical database research [3, 7, 8, 9, 19] and expertise in graph query optimization [2, 3, 12, 13, 14, 15, 17, 18], I shall dedicate a significant portion of my efforts to advancing graph-based ANNS toolbox in both theory and practice.

- Direction 2: Graph Search and Analytics. This research branches two two sub-directions based on the types of the underlying graphs.
  - Direction 2.1: Analyze and search over large graph data. For big graph analytics, we have studied graph clustering, hypergraph clustering, and attribute hypergraph clustering [1, 4, 5, 6] in a scalable and effective way. Our recent work [6] achieved an average of 20% higher F-measure (these are the quality measures for clustering), 24% higher ARI, 26% higher Jaccard Similarity, 10% higher Purity, and runs 5.5× faster than the SOTA method for attribute hypergraph clustering. Apart from clustering we had an in-depth study on the problem of generalized density-based Local Community Search (LCS) [2, 3] where the user provides a set of seed nodes and expects a dense subgraph "around" the seed nodes. Our results published on PODS 2024 [3] unveils the landscape on whether a density-based LCS objective function (a comprehensive family of objective functions have been covered by the results) could possibly be optimized by a "strongly local" algorithm for computation. By "strongly local", we mean the computation complexity is irrelevant to the size of the graph and only related to the user input. Furthermore, we provide a strongly local and practical linear programming based solution that can be easily deployed in practice.

    Opportunities: I see the line of work [2, 3] a combinations of theory and practice. Note that "strongly local" is ideal for a graph search algorithm with user input, it is an open problem to explore a wider range of graph queries to develop their strongly local solutions.
  - Direction 2.2: Brain network analytics. This research is conducted under the New Zealand Singapore Data Science Research Programme, focusing on advancing human brain connectivity studies using graph-based data science techniques. A key achievement was the creation of a robust infrastructure for brain network data collection and curation, including preprocessing brain imaging datasets from diverse sources and establishing an open-access repository for benchmarking [23]. On this foundation, we developed novel methods. For example, ContrastPool [22] is a graph pooling approach tailored for brain networks. ContrastPool employs dual attention mechanisms: ROI-wise attention highlights critical brain regions linked to neurological conditions, aligning with neuroscience domain knowledge, while subject-wise attention harmonizes dataset differences, reducing overfitting and enhancing group-level insights. The method outperforms all GNNs designed for brain networks, with improvements of up to 13.6%.

    Opportunities:
    - Multi-Modal Integration. Current challenges in brain connectivity analysis include effectively integrating diverse data modalities such as fMRI, DTI, and patient metadata. One direction is to develop graph-based frameworks that combine multi-modal data using advanced fusion techniques. This involves leveraging shared representations to incorporate the complementary strengths of each modality, enhancing the robustness and depth of connectivity analyses.
    - Tackling Data Scarcity and Overfitting. Brain network analytics frequently suffer from limited datasets and the risk of overfitting in high-dimensional feature spaces, particularly when using sophisticated models. To enrich information and synthesize context from

external sources (e.g., clinical notes, genetic profiles), it could be beneficial to incorporate language models and pre-trained multimodal models, augmenting limited data by incorporating domain knowledge and cross-domain linkages, mitigating overfitting risks.

My research bridges theory and practice in data science, with a current and future focus on advancing scalable and efficient solutions for vector data management and graph search and analytics. These efforts aim to address pressing challenges in AI-driven data-intensive applications, while uncovering new insights. Looking ahead, I am committed to expanding the boundaries of these research areas, fostering interdisciplinary collaborations, and driving innovation that combines theoretical rigor with practical relevance.

References

[1] Lijun Chang and Miao Qiao. Deconstruct densest subgraphs. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2747–2753. ACM / IW3C2, 2020.

[2] Yizhou Dai, Miao Qiao, and Lijun Chang. Anchored densest subgraph. In Zachary G. Ives, Angela Bonifati, and Amr El Abbadi, editors, *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, pages 1200–1213. ACM, 2022.

[3] Yizhou Dai, Miao Qiao, and Rong-Hua Li. On density-based local community search. *Proc. ACM Manag. Data*, 2(2):88, 2024.

[4] Zijin Feng, Miao Qiao, and Hong Cheng. Clustering activation networks. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*, pages 780–792. IEEE, 2022.

[5] Zijin Feng, Miao Qiao, and Hong Cheng. Modularity-based hypergraph clustering: Random hypergraph model, hyperedge-cluster relation, and computation. *Proc. ACM Manag. Data*, 1(3):215:1–215:25, 2023.

[6] Zijin Feng, Miao Qiao, Chengzhi Piao, and Hong Cheng. On graph representation for attributed hypergraph clustering. *To appear, Proc. ACM Manag. Data*, 2025.

[7] Xiaocheng Hu, Miao Qiao, and Yufei Tao. Independent range sampling. In Richard Hull and Martin Grohe, editors, *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS'14, Snowbird, UT, USA, June 22-27, 2014*, pages 246–255. ACM, 2014.

[8] Xiaocheng Hu, Miao Qiao, and Yufei Tao. Independent range sampling on a RAM. *IEEE Data Eng. Bull.*, 38(3):76–83, 2015.

[9] Xiaocheng Hu, Miao Qiao, and Yufei Tao. Join dependency testing, loomis-whitney join, and triangle enumeration. In Tova Milo and Diego Calvanese, editors, *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 291–301. ACM, 2015.

[10] Piotr Indyk and Haike Xu. Worst-case performance of popular approximate nearest neighbor search implementations: Guarantees and limitations. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[11] Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. Diskann: Fast accurate billion-point nearest neighbor search on a single node. In *NeurIPS*, volume 32, 2019.

[12] Wentao Li, Miao Qiao, Lu Qin, Lijun Chang, Ying Zhang, and Xuemin Lin. On scalable computation of graph eccentricities. In Zachary G. Ives, Angela Bonifati, and Amr El Abbadi, editors, *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, pages 904–916. ACM, 2022.

[13] Wentao Li, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. Scaling distance labeling on small-world networks. In Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 1060–1077. ACM, 2019.

[14] Wentao Li, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. Scaling up distance labeling on graphs with core-periphery properties. In David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo, editors, *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 1367–1381. ACM, 2020.

[15] Wentao Li, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. Distance labeling: on parallelism, compression, and ordering. *VLDB J.*, 31(1):129–155, 2022.

[16] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *TPAMI*, 42(4):824–836, 2018.

[17] Miao Qiao, Hong Cheng, Lijun Chang, and Jeffrey Xu Yu. Approximate shortest distance computing: A query-dependent local landmark scheme. In Anastasios Kementsietsidis and Marcos Antonio Vaz Salles, editors, *IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012*, pages 462–473. IEEE Computer Society, 2012.

[18] Miao Qiao, Lu Qin, Hong Cheng, Jeffrey Xu Yu, and Wentao Tian. Top-k nearest keyword search on large graphs. *Proc. VLDB Endow.*, 6(10):901–912, 2013.

[19] Miao Qiao and Yufei Tao. Two-attribute skew free, isolated CP theorem, and massively parallel joins. In Leonid Libkin, Reinhard Pichler, and Paolo Guagliardo, editors, *PODS'21: Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Virtual Event, China, June 20-25, 2021*, pages 166–180. ACM, 2021.

[20] Aditi Singh, Suhas Jayaram Subramanya, Ravishankar Krishnaswamy, and Harsha Vardhan Simhadri. Freshdiskann: A fast and accurate graph-based ann index for streaming similarity search, 2021.

[21] Haike Xu, Sandeep Silwal, and Piotr Indyk. A bi-metric framework for fast similarity search, 2024.

[22] Jiaxing Xu, Qingtian Bian, Xinhang Li, Aihu Zhang, Yiping Ke, Miao Qiao, Wei Zhang, Wei Khang Jeremy Sim, and Balázs Gulyás. Contrastive graph pooling for explainable classification of brain networks. *IEEE Trans. Medical Imaging*, 43(9):3292–3305, 2024.

[23] Jiaxing Xu, Yunhan Yang, David Tse Jung Huang, Sophi Shilpa Gururajapathy, Yiping Ke, Miao Qiao, Alan Wang, Haribalan Kumar, Josh McGeown, and Eryn Kwon. Data-driven network neuroscience: On data collection and benchmark. In Alice Oh, Tristan Naumann, Amir Globerson,

Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[24] Chaoji Zuo, Miao Qiao, Wenchao Zhou, Feifei Li, and Dong Deng. Serf: Segment graph for range-filtering approximate nearest neighbor search. *Proc. ACM Manag. Data*, 2(1):69:1–69:26, 2024.