Graph-based Label-Efficient Learning: When Graph-Structured Data Meets Limited Labels

Zixing Song

University of Bristol, Bristol, United Kingdom
zixing.song@bristol.ac.uk

The success of deep learning is highly dependent on large-scale labeled data. This presents a formidable challenge in fields like molecular design and materials science, where data annotation is prohibitively expensive. Consequently, developing **label-efficient learning** methods to maximize model performance under limited annotation budgets has recently become more and more critical.

However, most of the current mainstream label-efficient algorithms, like active learning and semi-supervised learning, are primarily designed for Euclidean data, such as images. They cannot effectively process the non-Euclidean **graph-structured data**, thus overlooking the rich topological information embedded within.

In this talk, we aim to bridge this gap through a progressive research path that addresses three core challenges in data annotation for graph-structured data. First, to address the *high cost of annotation*, we adapt active learning and semi-supervised learning from general domains to ***explicit graph data***, enabling the precise labeling of high-value nodes. Second, to address *label scarcity*, we pioneer methods to construct and leverage ***implicit graph structures***, propagating existing labels and generating new information to boost the performance of semi-supervised and self-supervised learning. Finally, to address *label noise*, we perform the fusion of *both explicit and implicit graphs*. By learning an implicit structure from noisy explicit graph data, our methods will identify and mitigate the impact of noise.

Part I: Adapting Label-Efficient Learning to Explicit Graph Data. The talk begins by addressing the foundational challenge of explicit graph data, where the non-IID nature of nodes violates the core assumptions of traditional label-efficient methods. First, we introduced a Bayesian GNN framework that provides a probabilistic interpretation of graph learning, deriving the first closed-form solution for the posterior distribution of node embeddings. This establishes a theoretical foundation that formally captures the topological constraints of graph data. Building on this, I designed a novel graph-based active learning strategy by proving the theoretical equivalence between maximizing the Expected Model Change (EMCM) and minimizing node prediction error (Song, Zhang, and King 2023a). This allows for a principled query strategy that selects the most valuable nodes for annotation, significantly reducing labeling costs.

Part II: Exploiting Implicit Graph Data to Enhance Label-Efficient Learning. The second part of the talk exploits latent relational structures within general data. We introduce a principled framework for constructing implicit graphs to enhance label-efficient learning, effectively turning any semi-supervised setting into a label inference problem over a constructed graph. We first established a unified "graph construction-label propagation" taxonomy (Song et al. 2023). A core innovation is the first precise definition of an optimal implicit graph, which guarantees the most effective propagation of scarce labels (Song, Zhang, and King 2023b). We implement this theory with a novel asymmetric graph construction algorithm using block-wise optimization, which has proven global convergence with a sub-linear rate.

Part III: Combining Explicit and Implicit Graphs for Robust Label-Efficient Learning The final part of this talk confronts the challenge of noisy explicit graphs by fusing them with learned implicit graph structures. We establish a novel theoretical equivalence between graph regularization and graph embedding using Gaussian Markov Random Fields (GMRFs) (Song, Zhang, and King 2022). This enables an asymmetric framework that jointly optimizes graph structure and node classification, using heterophilous edges to identify noise and graph rewiring to mitigate its impact. We also leverage this as a pre-task for self-supervised learning, transferring noise-independent topology to generate denoised labels (Song, Meng, and Hernández-Lobato 2025).

References

Song, Z.; Meng, Z.; and Hernández-Lobato, J. M. 2025. Domain-Adapted Diffusion Model for PROTAC Linker Design Through the Lens of Density Ratio in Chemical Space. In *ICML*.

Song, Z.; Yang, X.; Xu, Z.; and King, I. 2023. Graph-Based Semi-Supervised Learning: A Comprehensive Review. *IEEE Trans. Neural Networks Learn. Syst.*, 34(11): 8174–8194.

Song, Z.; Zhang, Y.; and King, I. 2022. Towards an Optimal Asymmetric Graph Structure for Robust Semi-supervised Node Classification. In *KDD*, 1656–1665. ACM.

Song, Z.; Zhang, Y.; and King, I. 2023a. No Change, No Gain: Empowering Graph Neural Networks with Expected Model Change Maximization for Active Learning. In *NeurIPS*.

Song, Z.; Zhang, Y.; and King, I. 2023b. Optimal Block-wise Asymmetric Graph Construction for Graph-based Semi-supervised Learning. In *NeurIPS*.