On Reliability and Controllability of Machine Learning Models for Structured Data

Kartik Sharma

Artificial Intelligence (AI) systems face many challenges for their successful adoption in society. This arises primarily due to a gap in the existing solutions and the users' expectations and needs. Real-world data has a definite structure that can be efficiently exploited to narrow this gap in a scalable manner. Through my research, my goal is to ***study and build machine learning (ML) models for structured data, which are reliable and can be easily controlled by the users***. I have divided my research agenda into three directions – **(1)** Measuring the reliability and alignment of ML models, **(2)** Designing reliable and interpretable predictive models, and **(3)** Enabling user-controllable generation. During my Ph.D. journey so far, I have explored specific research questions within these areas and am extremely enthusiastic to continue my exploration in the future. Below I give an overview of my ongoing projects and my plan for the future.

(1) Measuring the reliability and alignment of current models. In a work published at KDD 2023, we probed the vulnerabilities of learning models for dynamic graph structures, which are trained to predict future interactions based on past interactions. Our goal was to assess the *reliability* of these models against practical perturbations that would theoretically remain undetected by anomaly detectors. To this end, we came up with a formulation that our perturbations at each time should be at most a small fraction times the actual changes at that time in the graph. Then, using a Projected Gradient Descent-based approach, we showed that our state-of-the-art models are highly vulnerable to such attacks without being flagged by the anomaly detectors. This implies that these models, in their current state, are not ideal for making real-world predictions on dynamic graph-structured data, especially for security-critical applications such as financial services.
In an ongoing work, our objective is to efficiently find the *alignment* between large language models (LLMs) and large knowledge graphs. LLMs have been shown to suffer from hallucinations of unnecessary content that can be harmful. Thus, it is extremely important to pre-emptively identify and flag these failure cases before deployment. Knowledge graphs provide us with a benchmark to verify the language models at scale. These graphs can be domain-specific for industry-focused LLMs (such as a set of financial clauses to be followed by a finance LLM) or general knowledge for general-purpose LLMs (such as Wikipedia for GPT-3.5). In this work, we aim to find the missing facts in an LLM by asking as few yes/no queries as possible. To this end, we have designed a Graph Neural Network (GNN)-based architecture to approximate any LLM's learned knowledge of the facts in a given knowledge graph based on the previous queries. This allows us to efficiently find missing facts in any off-the-shelf LLMs even over billion-sized databases by asking a minimal number of queries.

(2) Designing reliable and interpretable predictive models. In a study published at CIKM 2023, we developed a method to predict the dynamics in signed social networks that are interpretable using social-structural theories of signed networks. Dynamic signed social networks represent evolving positive and negative sentiments among individuals, enterprises, or groups. Predicting the link structure in these networks allows us to anticipate and mitigate the risks of toxic online discourse, unhealthy industrial relations, etc. We thus designed a dynamic GNN that implicitly imposes the theory of structure balance, i.e., a friend's friend is likely to be a friend (positive) and an enemy's friend is an enemy (negative). In particular, we separately encoded data from the neighbors of positive and negative social

connections, as they are made in continuous time. Through comprehensive experiments on various datasets and tasks, we then show that a model grounded in social theories is not only *self-interpretable* but can also efficiently forecast the nature and likelihood of future interactions between users, achieving better results than conventional models.
In another work done during the internship at Visa Research, we developed a strategy to train GNN models to classify a node using a personalized number of hops or scope of its local neighborhood. This allows us to predict not just the node label but also the optimal number of hops of local information needed to make the prediction. Different types of fraud can thus be differentiated in terms of the neighborhood scope. Using our plug-and-play strategy, we make the node-level predictions of any GNN more *interpretable, robust, faster-converging,* and *reliable in heterophilic settings*.

(3) Enabling user-controllable generation. In a study presented at ICML 2023's SPI-GM workshop (full version under review), we explored the challenge of generating graphs that satisfy user-defined hard constraints. For example, in AI-driven drug design, a designer is likely to be interested in molecules whose molecular weight is not more than some number, W. Graph diffusion models do not allow user-specified constraints without relying on retraining or uninterpretable probabilistic guidance. We thus developed a plug-and-play solution that manipulates a pre-trained diffusion model's sampling process to meet the exact sample-time requirements, such as specific molecular weight. Our technique is based on moving the data sample after each denoising phase to *align* with the imposed constraints. It efficiently accommodates a wide range of constraints, like edge count, molecular weight, valency, etc. The generated graphs give up to 100% satisfaction of the given constraints while remaining close to the underlying data distribution.

Future plan: I plan to soon work on controlling text generation models to satisfy certain constraints like a given graph structure or a given linguistic structure for the generations. Thus, for the remaining years of my Ph.D., I want to work on the alignment and controllability of LLM generation. I am also interested in developing security protocols that allow us to distinguish human and machine-generated texts without sharing the model parameters. This allows to increase the reliability of these models by reducing their potential misuse due to easy and secure checking. In the future, I also plan to extend my experience in other types of structured data, particularly, relational tabular databases, and documents, due to their wide availability in the real-world infrastructure. My experience in general graph structures can prove extremely useful for my planned investigations into these structures.

Interests in JPMC and financial services: An exciting application of my research exists for *financial services* due to the security-critical and structured nature of the underlying data. Thus, both components of my research, i.e., reliability and structure, are directly applicable to the financial sector. Thus, I would be quite interested in doing a summer internship at JPMC as it provides me with an exciting opportunity to work on real-world challenges in ML applications around the reliability and controllability of structured data. I would ideally prefer my internship to be motivated by the problems that are faced or will likely be faced by the intended users of ML applications of JPMC. For example, due to the extreme security requirement of the financial business, it is crucial to build automated chatbots that are free of any hallucinations for domain-specific queries. I would thus be interested in working on the scalable evaluation of LLMs against a set of these structured banking clauses and controlling the text generation models to follow these clauses faithfully.