Statement of Purpose

Rishabh Ranjan, Ph.D. Applicant

I embarked on my research journey in the summer of 2020, during my undergrad at IIT Delhi. By the time I graduated (summer of 2022), my inclination towards research had blossomed into a strong resolve for a **career in academia**. In this period, I produced 3 first author papers under the mentorship of eminent advisors and collaborators. My research output is complemented by my academic performance – recognized with the **President's Gold Medal** for highest CGPA (9.904) in the batch – with the highest letter grade in numerous graduate-level courses. My undergrad experiences culminated with a broad interest in **Machine Learning (ML)** research, which I am pursuing further in my internship at **Carnegie Mellon University (CMU)**, under **Prof. Zachary Lipton**. With the research and technical skills I have amassed, I feel confident and excited to undertake graduate studies.

Research in Graph ML. In my junior year, I undertook my first project in ML under the guidance of **Prof. Sayan Ranu**, in collaboration with researchers from **University of Illinois Chicago (UIC)** and **IBM Research India**. We worked on learning graph similarities via **Graph Neural Networks (GNNs)** with the goal of answering queries on large databases. This has applications in *drug discovery*, *protein-protein interaction*, *knowledge graph analogy reasoning*, *social network analytics*, etc. Apart from crucial advisory inputs from collaborators, I worked independently during the development phase, later enlisting the help of a fellow undergrad for benchmarking.

Prior works focus solely on pairwise graph similarity prediction, and do not scale to real-world querying scenarios. In response, I came up with a framework which supports *pre-computing* and *indexing* graph embeddings for efficient retrieval at query-time. The key idea was to map graphs into a learned embedding space with *rich geometric structure* reminiscent of the graph distance measures we intended to predict. I proved guarantees such as *triangle inequality*, which make our work unique among the contemporaries. The proposed framework achieves significantly better prediction quality, with up to **1000 times faster** range and k-NN queries, scaling to large networks with millions of nodes. This has resulted in a publication at **NeurIPS 2022** [5].

This work experienced a few re-submissions. Seeing the work improve after each round of feedback has imbued in me the mindset to sail through rejections without getting discouraged. Further, the payoff has been great. It was an absolute delight to see great interest at our NeurIPS poster, including from people directly involved in the drug discovery industry, who reaffirmed the motivations and applicability of our work. A recent work [2] which builds on our released code[1] motivates me to make my research easily accessible.

Research in Neuro-Symbolic AI. For my **thesis project**, I worked under the guidance of **Prof. Mausam** and **Prof. Parag Singla** on the broad goal of integrating combinatorial solvers into deep learning pipelines to augment neural networks with better reasoning and planning capabilities. Over brainstorming sessions with my advisors and a graduate student collaborator, backed by extensive literature review, I identified shortcomings in a SOTA technique for training neural networks along with an **Integer Linear Programming (ILP)** layer.

Guided by the geometric intuition that an ILP instance corresponds to a *polytope* in high-dimensional space, I proposed a novel margin-based loss function for training neural ILP architectures. Our approach allows bypassing the ILP solver in the forward pass, thereby alleviating the major bottleneck in prior work. The key idea is to transform the *combinatorial search* problem of ILP into a *binary classification* problem, by leveraging training supervision. Our framework outperforms all the baselines with significant margins on a number of tasks, such as **visual sudoku** where we achieve **98%** accuracy compared to the purely neural baseline's 71%, and the best neuro-symbolic baseline's 18%. This work was received warmly at **NeurIPS 2022** [3].

For this project, I was awarded the **Suresh Chandra Memorial Trust Award for the best undergraduate thesis**. I have released a light-weight python library[2], allowing anyone to train neural ILP architectures with ease. Through this project, I have learned to appreciate the intricate relationship between optimizing a suitable objective, learnability, and the resulting capability of the trained model.

Research in Distributed Systems. My first research project was on verifying distributed programs for communication deadlocks, with **Prof. Subodh Sharma**. While the deadlock detection problem is NP-hard in general, real-world programs exhibit redundancies which can be exploited for faster verification.

I developed a two-fold framework to (i) decompose the program into independently-verifiable communication *epochs* and (ii) prune away *symmetric* parts of the search space when looking for deadlocks (via a SAT solver). I proved *soundness* and *completeness* results, establishing the correctness of the approach. We obtained orders-of-magnitude reduction in verification times and were able to scale to larger programs and higher degrees of parallelism than prior art. This work was accepted at **ASE 2022** [4], a top venue in software engineering. Earlier, I presented it at the **IARCS SAT+SMT Workshop 2020**.

I implemented our techniques in a prototype tool called Simian[3], capable of verifying C/C++ programs written using the **Message Passing Interface (MPI)**. This project gave me an opportunity to mentor a junior student, who is the second-author. We have extended the work to a broader class of programs, and are preparing for journal

[1] `https://github.com/dair-iitd/greed`
[2] `https://github.com/rishabh-ranjan/ilploss`
[3] `https://github.com/rishabh-ranjan/simian`

submission.

Current Research. In my internship with **Prof. Zachary Lipton** at **CMU**, I am looking at *non-separable data* settings where the ground-truth $p(y|x)$ is not 1-hot. The technical challenge arises from observing only hard labels in the train set. Such scenarios abound in *industrial*, *medical* and *web* domains due to incomplete information or instance-dependent label noise. However, they have received little interest in academia. To bridge this gap, I am curating a compelling benchmark of real-world datasets exhibiting non-separability and developing principled approaches to address the challenges posed. Also, I am implementing and evaluating simple baselines and existing techniques on these settings.

Working on this, I have observed that training neural networks under noisy regimes uncovers fascinating phenomenon relating to learning dynamics, especially when training beyond the *interpolation regime*. This, I seek to actively investigate and understand. This has sparked in me a fundamental curiosity about the hidden capabilities and limitations of neural networks, which I'd love to explore both *theoretically* and *empirically* in my future work.

Preparation. While I have touched on diverse fields in my past research, I have been able to contribute effectively to them. I believe this has been possible because I am a fast and diligent learner, and because the research skills involved are highly transferable, to the extent that different perspectives afforded by different backgrounds can prove highly effective in tackling the problem at hand.

Each of my experiences has made me more mature as a researcher, and has improved my understanding of the research process. Interactions with my advisors has shaped my research outlook and developed my ability to identify and attempt to close the gaps between formulations studied in the literature and application requirements, which I now view as an uncompromising grounding for research. Persevering through dead-ends and insurmountable-looking roadblocks has taught me the vital skill of recognizing the shortcomings in my ideas and using them as seeds for new and better ideas, while being on the lookout for faint signals pointing to potential breakthroughs. This has given me the confidence to undertake ambitious projects which are compelling in the *problem space* with only secondary regard to the perceived difficulty in the *solution space*.

Further, I have learned to *multi-task* and effectively handle periods of high workload, as I was faced with writing, experimentation, rebuttal, presentation, etc. of two papers simultaneously at NeurIPS 2022, and another paper at ASE 2022 at around the same time, all at first author capacity. This skill was also reinforced when balancing the course workload with research.

I have had the opportunity to contrast my research experience with industrial experience, having interned at **Samsung Electronics, South Korea**. My takeaway has been that my temperament is more suited to thrive in the uncertain environment of research, where we constantly ask fundamental questions and take deep dives into the unknown looking for the answer, much more so than in the epistemic comforts of industry. I enjoy all aspects of academic life such as *writing*, *presenting*, *teaching*, *mentoring* and *reviewing*, which further upscales my attraction to academia.

At Stanford. I'd be excited to continue working on machine learning for graphs with **Prof. Jure Leskovec**. His research has been a major influence in shaping the entire field, and has heavily impacted my own work with Graph Neural Networks. With Prof. Leskovec, I'd be interested in investigating the theoretical foundations of GNNs as well as their applications to domains like computational biology and social networks.

I'd also love to work with **Prof. Tengyu Ma**. His theoretically-grounded work sheds light on various aspects of Machine Learning that I'm interested in, such as out-of-distribution generalization, optimization, and uncertainty quantification. With prior experience in distributed systems, exploring distributed / federated ML is also an exciting prospect.

In a similar vein, I'm also interested in **Prof. John Duchi's** work on statistical machine learning and optimization, and would love to explore further by working with him.

Looking ahead. My personal experiences and the exposure I have received at IIT Delhi and CMU have informed my research interests and refined my tastes. Going forward, I seek to (i) **understand** the foundations, properties, strengths and limitations of *models* and *learning algorithms* both *theoretically* and *empirically*, and (ii) **build** learning systems better capable of system 2 [1] tasks like *reasoning* and *planning*, and which can naturally achieve systematic generalization to *out-of-distribution* settings by incorporating *common-sense*, a *causal* understanding of the world and/or suitable *inductive biases*. I believe both (i) and (ii) go hand in hand, with better understanding lending itself to more principled innovations, and better techniques posing questions worthy of deeper investigation.

With the potential to open up endless possibilities in all fields of scientific and human endeavor, solving intelligence is the ultimate meta-problem faced by the human species. In the pursuit of this overarching goal, I aspire to work at the frontiers of cutting-edge AI research in the company of brilliant and inspiring minds. Through a PhD I seek to gain and consolidate the requisite skills to realize this dream. With its deep-rooted academic traditions and an erudite pool of researchers, Stanford would be an ideal place to shape my intellectual growth and help me uncover my true potential as a researcher.

References

[1] Daniel Kahneman. *Thinking, fast and slow.* Macmillan, 2011.

[2] Linfeng Liu, Xu Han, Dawei Zhou, and Li-Ping Liu. Towards accurate subgraph similarity computation via neural graph pruning. *arXiv preprint arXiv:2210.10643*, 2022.

[3] Yatin Nandwani*, **Rishabh Ranjan***, Mausam, and Parag Singla. A solver-free framework for scalable learning in neural ILP architectures. In *Advances in Neural Information Processing Systems*, 2022.

[4] **Rishabh Ranjan**, Ishita Agrawal, and Subodh Sharma. Exploiting epochs and symmetries in analysing MPI programs. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, October 2022.

[5] **Rishabh Ranjan**, Siddharth Grover, Sourav Medya, Venkatesan Chakaravarthy, Yogish Sabharwal, and Sayan Ranu. GREED: a neural framework for learning graph distance functions. In *Advances in Neural Information Processing Systems*, 2022.