Personal statement

With strong interest and motivation, I am determined to apply to your PhD program in Computing and Data Sciences at your prestigious university and would like to use this brief personal statement to highlight my academic background and research interests. Before that, I should say how much I respect Columbia University, both for its well-known reputation and for her great contributions to scientific research. My academic goals are properly met by the competent program design and great faculty team there. I have solid grounds for believing that this Ph.D. program will help me grow professionally so that I can achieve my career goals of becoming an engineer, a senior researcher in a technology company, or an academia professor in the future, which is exactly in line with the program's design goal and convinces me that I made the right choice.

Since 2021, I have been enrolled in the master's program of Data Science at Tufts University with an educational background in Statistics from Guangdong University of Foreign Studies. Therefore, it is easy to see from my past academic performance that I have established a solid foundation of knowledge in the two areas mentioned above, particularly in regard to subjects like Mathematical Statistics, Probability Theory, Advanced Mathematics, Applied Statistics, Algorithms, Statistics Pattern Recognition, Data Mining, Deep learning, Machine Learning and so forth.

In my past study, I have placed a special emphasis on the courses of Machine Learning Principles and its framework programming like Python & Java programming, in addition to my proficiency in Pytorch, and Scikit-learn machine learning framework. My PhD stage's primary focus will also be on machine learning. I intend to learn more about Federal learning, Deep graph learning, Natural language processing, Generative model diffusion, Application of graph data, and Graph machine learning, and learn some mathematics courses in Machine Learning, such as Stochastic process, Information theory, Bayesian machine learning, etc., to improve my mathematical derivation and my capacity to turn objective problems into mathematical problems. Besides, I want to conduct more project programming, complete the project from scratch, increase my proficiency at translating mathematical issues into computer language, work on generative models/algorithms, and improve the inadequacies of current graph generative models.

What affected me most was the course Statistic Pattern Recognition taught by Dr. Isaac Lage. This class taught me many probabilistic models, such as the Hidden Markov Model, Gaussian mixture model, and k-means algorithm. Also, I learned how to perform inference through MCMC and variational inference. I think the class paved a solid foundation for my statistical learning background, which helped me a lot in my future machine learning research.

In the deep learning course taught by Prof. Liping Liu, I found myself very into the class on generative modeling. In this class, I comprehensively learned about various neural networks such as RNN, transformer, and CNN. What excited me the most was the course project I

worked on, in which I did style transformation by using cycleGAN. In this project, I tried to innovate the loss function and discriminator architecture. Specifically, I reweighted adversarial loss, cycle consistent loss, and identity loss to see how the image style would change accordingly. Besides, I simplified the discriminator to make the optimization of the generator easier. Though my ideas did not improve over the original model, I became more interested in generative models and deep learning.

To better have hands-on experience, I have worked on two projects related to graph generation, one was to benchmark the generative performance over one-shot and sequential graph generative models, and the second was to develop graph generative models that meet differential privacy.

The goal of the first project was to provide a fair comparison of two kinds of generative models in terms of the likelihood metrics. Since one-shot generative models are usually permutation-invariant and sequential generative model depends on the generating orders, these two models are not directly comparable due to the marginalization over the permutation. In this project, my first duty was to derive the log-likelihood for the one-shot generative models, by taking the permutation into consideration. This leads me to get familiar with the mechanism of flow-based generative models and how to model discrete data. While the log-likelihood calculation is well-established in sequential generative models, my second duty is to run extensive experiments to compare these two kinds of methods, thus bridging the two models over this more universal metric. I am fortunate to be advised by Prof. Liu Li-ping on this project. I learned a lot about the basic techniques of Variational Inference, data dequantization, and a broad class of Graph Neural Networks. Furthermore, I learn how to code with graph data structure, which is more complicated than image and text data as its irregular data representation.

The second project was accepted by ICML2023. Its main goal is to generate graphs that preserve the graph statistics without leaking the users' privacy in the training graphs. However, current deep graph generative models cannot generate large networks with satisfying graph statistics. The first goal of this project is to develop a scalable and powerful graph generative model, which leads us to first turn our attention to the diffusion model. During the development of the new diffusion graph model, I am responsible for implementing a batching training procedure, which speeds up the training and reduces the gradient variance. Besides, I derived the posterior of the non-uniform diffusion forward kernel. When working on the derivation, I had a deep understanding of how the diffusion models are derived, which [made me gain more understanding generative model. Besides, I read plenty of papers on differential privacy with graph data and papers about how to use differential privacy and create data. After developing such a model, I will focus on generating large networks with privacy guarantees.

Objectively speaking, I choose machine learning as the major topic of my Ph.D. research primarily due to my advantages in statistics, mathematics, and computer knowledge and skills. Considering statistics, as an illustration, Machine Learning uses techniques like

Markov Monte Carlo Sampling, Bayesian inference, conditional probability in diffusion models, and ELBO derivation in variational auto-encoders, etc., which makes statistics applications extremely valuable and further piques my interest. Recently, machine learning has developed a number of extremely potent capabilities, such as the chatGPT3 that was just released, which has made me realize the enormous potential and development space in this field of study, and greatly inspired me to invest in further research and exploration of machine learning.
Becoming an expert in the domains of graph machine learning generative models, and natural language is my professional life's ambition.

In a sense, my research supervisor and Feynman are the two people who have had the biggest influences on my intense interest and motivation in scientific research. My mentor is a serious thinker, whose dedication to research and analytical approach has had a significant impact on me. He always challenges me to speak with mathematics, articulate ideas with mathematics, and approach issues with rigorous and abstract mathematical thinking during our academic sessions. My knowledge of the issue gradually deepened after discovering this sophisticated mathematical statement, and I started to pay closer attention to my mathematical prowess. The other person is Feynman, who had a greater spiritual influence on me, and whose quote I use as my motto. In an interview, the host questioned Feynman about the significance of astronomical research. He responded that in truth, it may not have much significance or significantly improve people's lives, but learning itself is the best purpose since it allows us to better understand the world and explore the unknown. Yes, seeking knowledge is the greatest purpose and joy, thus I believe conducting research calls for this kind of mindset.