Lecture-style Tutorial: Towards Graph Foundation Models

Chuan Shi[*]
Beijing University of Posts and Telecommunications
Haidian, Beijing, China
shichuan@bupt.edu.cn

Cheng Yang
Beijing University of Posts and Telecommunications
Haidian, Beijing, China
yangcheng@bupt.edu.cn

Yuan Fang
Singapore Management University
Singapore
yfang@smu.edu.sg

Lichao Sun
Lehigh University
Bethlehem, Pennsylvania, USA
lis221@lehigh.edu

Philip S. Yu
University of Illinois Chicago
Chicago, Illinois, USA
psyu@uic.edu

ABSTRACT

Emerging as fundamental building blocks for diverse artificial intelligence applications, foundation models have achieved notable success across natural language processing and many other domains. Concurrently, graph machine learning has gradually evolved from shallow methods to deep models to leverage the abundant graph-structured data that constitute an important pillar in the data ecosystem for artificial intelligence. Naturally, the emergence and homogenization capabilities of foundation models have piqued the interest of graph machine learning researchers. This has sparked discussions about developing a next-generation graph learning paradigm, one that is pre-trained on broad graph data and can be adapted to a wide range of downstream graph-based tasks. However, there is currently no clear definition or systematic analysis for this type of work.

In this tutorial, we will introduce the concept of *graph foundation models* (GFMs), and provide a comprehensive exposition on their key characteristics and underpinning technologies. Subsequently, we will thoroughly review existing works that lay the groundwork towards GFMs, which are summarized into three primary categories based on their roots in graph neural networks, large language models, or a hybrid of both. Beyond providing a comprehensive overview and in-depth analysis of the current landscape and progress towards graph foundation models, this tutorial will also explore potential avenues for future research in this important and dynamic field. Finally, to help the audience gain a systematic understanding of the topics covered in this tutorial, we present further details in our recent preprint paper, "Towards Graph Foundation Models: A Survey and Beyond" [4], available at https://arxiv.org/pdf/2310.11829.pdf.

[*]Corresponding tutor

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05…$15.00
https://doi.org/10.1145/3589335.3641246

CCS CONCEPTS

• **Information systems** → **Data mining**; **World Wide Web**; • **Computing methodologies** → **Learning paradigms**; **Artificial intelligence**.

KEYWORDS

Graph Foundation Models, Large Language Models, Graph Neural Networks, Deep Learning

1 GENERAL INFORMATION

Title. Towards Graph Foundation Models.

Contributors. There are five contributors to this tutorial. Prof. Chuan Shi is the primary point of contact, and the contact information of all contributors are listed in the author information block.

- ***Chuan Shi*** is professor and PhD supervisor in School of Computer Sciences of Beijing University of Posts and Telecommunications, deputy director of Beijing Key Lab of Intelligent Telecommunication Software and Multimedia. The main research interests include data mining, machine learning, artificial intelligence and big data analysis. He has published more than 100 refereed papers, including top journals and conferences in data mining and machine learning, such as IEEE TKDE, ACM TKDD, KDD, WWW, NeurIPS, AAAI and IJCAI. He has been honored as the best paper award in ADMA 2011 and ADMA 2018, and the best paper nomination in the Web Conference 2021. He also won several awards, such as the second prize of Natural Science of Beijing/CCF (1st), the first prize of artificial intelligence science and technology progress of Wu Wenjun (3rd) and the first prize of science and technology progress of the Chinese Institute of Electronics (1st).
- ***Cheng Yang*** is an Associate Professor of Computer Science at Beijing University of Posts and Telecommunications (BUPT). He received his Bachelor and Ph.D. degrees from Tsinghua University in 2014 and 2019, respectively. Cheng's research interests include data mining, natural language processing and social computing. He has published 50+ papers in top journals and conferences, such as IEEE TKDE, ACM TOIS, KDD, WWW, Neurips and ACL. His work has got more than 8,000 citations as shown by Google Scholar, and named by Baidu as one of the Top 100 Chinese Young Scholars in Artificial Intelligence.

- ***Yuan Fang*** is an Assistant Professor at the School of Computing and Information Systems, Singapore Management University (SMU). Prior to joining SMU, he was a data scientist at DBS Bank, and a research scientist at A*STAR. He obtained a PhD Degree in Computer Science from the University of Illinois at Urbana-Champaign in 2014 on a fully funded scholarship from A*STAR, and Bachelor of Computing with First Class Honors from National University of Singapore in 2009 as the top student in Computer Science. His research focuses on graph-based learning and mining, as well as its applications in recommendation systems, social network analysis and bioinformatics.
- ***Lichao Sun*** is an Assistant Professor in Computer Science and Engineering at Lehigh University. Meanwhile, he is also an Adjunct Professor at Mayo Clinc. Before that, he received his Ph.D. degree in Computer Science at University of Illinois, Chicago in 2020, under the supervision of Prof. Philip S. Yu. Further before, he obtained M.S. and B.S. from University of Nebraska Lincoln. He also have long-term research intern experience in the industrial research labs, including Microsoft Research, Salesforce Research, Samsung Research America, Amazon Web Service, and LinkedIn.
- ***Philip S. Yu***'s main research interests include big data, data mining (especially on graph/network mining), social network, privacy preserving data publishing, data stream, database systems, and Internet applications and technologies. He is a Disthinguished Professor in the Department of Computer Science at UIC and also holds the Wexler Chair in Information and Technology. Before joining UIC, he was with IBM Thomas J. Watson Research Center, where he was manager of the Software Tools and Techniques department. Dr. Yu has published more than 970 papers in refereed journals and conferences with more than 74,500 citations and an H-index of 127. He holds or has applied for more than 300 US patents. Dr. Yu is a Fellow of the ACM and the IEEE. He is the recepient of ACM SIGKDD 2016 Innovation Award for his influential research and scientific contributions on mining, fusion and anonymization of big data, the IEEE Computer Society's 2013 Technical Achievement Award for "pioneering and fundamentally innovative contributions to scalable indexing, querying, searching, mining and anonymization of big data", and the Research Contributions Award from IEEE Intl. Conference on Data Mining (ICDM) in 2003 for his pioneering contributions to the field of data mining. He also received an IEEE Region 1 Award for "promoting and perpetuating numerous new electrical engineering concepts" in 1999. He had received several UIC honors, including Research of the Year at 2013 and UI Faculty Scholar at 2014. He also received many IBM honors including 2 IBM Outstanding Innovation Awards, an Outstanding Technical Achievement Award, 2 Research Division Awards and the 94th plateau of Invention Achievement Awards. He was an IBM Master Inventor.

2 TOPIC AND RELEVANCE

Scope. The scope of the tutorial topic covers the following aspects.

- An introduction on the background and preliminaries of deep graph learning and language foundation models, from fundamental concepts to the cutting-edge, on-going research status.
  - Graph deep learning: Graph data, graph neural networks (GNN), and learning paradigms;
  - Language foundation models: Language data, large language mdoels (LLM), and learning paradigms.
- A discussion on the concept of the Graph Foundation Model (GFM), including the following.
  - The definition of GFM and its key characteristics of emergence and homogenization;
  - Underpinning technologies of GFM involving the pre-training and adaptation of graph models;
  - The impact from graph data and graph tasks;
  - A comparison to traditional deep graph learning models, as illustrated in Figure 1 [4];
  - A comparison to language foundation models in terms of the similarities and differences.
- A review of existing works that establish the groundwork and advance the research frontier towards GFM. They can be primarily summarized into three types based on their roots in different technologies, including the following.
  - GNN-based methods, which are based on message-passing or transformer-based architectures;
  - LLM-based methods, which are based on graph-to-token or graph-to-text architectures;
  - GNN+LLM-based methods, which are based on GNN-centric, LLM-centric, or GNN-LLM symmetric architectures.
- Future directions in this area, where we will present our perspectives and insights as summarized into challenges about three dimensions:
  - Data and evaluation, including data quantity and quality, as well as the evaluation on the performance, robustness, and trustworthiness;
  - Model architectures and training paradigms beyond current backbones, pre-training tasks and adaptation strategies;
  - Killer applications, as well as the safety issues associated with these applications.

Importance and Timeliness. Foundation models have achieved remarkable success in natural language processing, computer vision, and beyond. While graph machine learning has also evolved from shallow methods to deep models, the emergence and homogenization capabilities of foundation models have sparked discussions about the next-generation graph learning paradigm, or as we call "graph foundation models", to benefit from broad graph data and accomplish a wide range of graph tasks. However, there is currently no clear definition or systematic analysis for this type of work. Hence, this tutorial is important and timely, which could inspire and fuel high-impact applications involving graph data.

Relevance to the Web Conference. The Web has increasingly become a giant repository that interconnect vast entities, leading to various Web graphs such as social networks, e-commerce graphs and research collaboration graphs. Modeling and analysis of these large-scale Web graphs are crucial to understand human behaviours and interactions on the Web, capture the dynamic and involving nature of the Web, as well as generate insights and knowledge from the Web graphs. In fact, the Web Conference has a research track

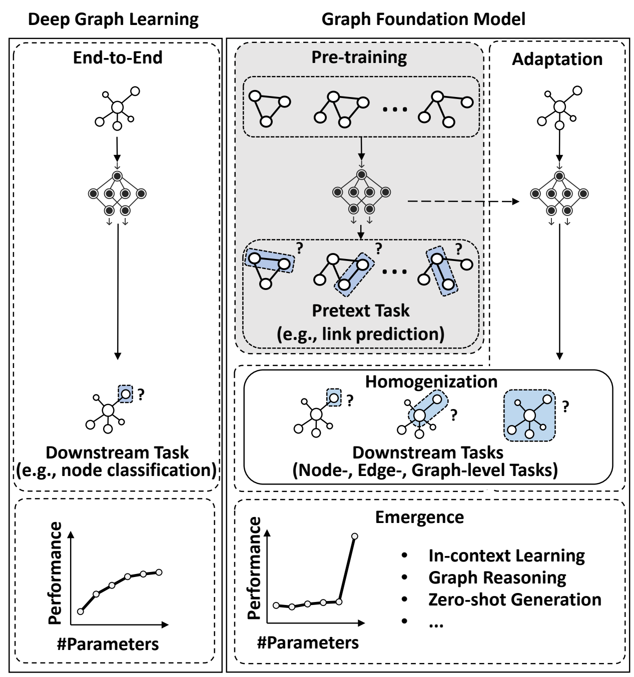

Figure 1: A comparison between deep graph learning models and graph foundation models (GFM). Traditional graph deep models address specific tasks on specific datasets through end-to-end training, typically in a semi-supervised or supervised fashion. In contrast, GFMs are first pre-trained using self-supervised pretext tasks on broad graph data, and subsequently adapted to new data or tasks.

dedicated to graph algorithms and modelling for the Web[1] in the main conference. Hence, this tutorial is relevant to the conference, and can complement the graph track by envisioning challenges and directions for future research.

Qualification of Presenters. The contributors/presenters are all experts who have been working on graph-related research in recent years, and have published extensively in many subtopics covered by this tutorial.

3 STYLE

This will be a lecture-style tutorial, giving its goal of providing a comprehensive overview and systematic analysis of this novel and fast growing direction.

4 SCHEDULE

This is a half-day tutorial lasting 3 hours. A summary of the schedule is listed in Table 1, while more details of each section are given in §4.1–4.4.

4.1 Introduction and Overview (40 mins)

- Presenters: Dr. Chuan Shi, Dr. Philip S. Yu

[1]https://www2024.thewebconf.org/calls/research-tracks/

Table 1: Summary of tutorial schedule.

Part	Presenter(s)	Duration
Introduction & Overview (§4.1)	Chuan Shi, Philip S. Yu	40 mins
GNN-based Methods (§4.2)	Cheng Yang	45 mins
LLM-based Methods (§4.3)	Yuan Fang	45 mins
GNN+LLM-based Methods (§4.4)	Lichao Sun	45 mins

- Overview of Graph Foundation Models
- Perspectives and insights

With the rapid rise in computational power and breakthroughs in deep learning techniques particularly on the Transformer architectures, the artificial intelligence community has introduced the notion of "foundation models". As defined earlier, *a foundation model is any model that is trained on broad data and can be adapted to a wide range of downstream tasks [1].* Inspired by the success of large language models (LLMs) as foundation models in natural language processing, researchers have explored the possibilities of Graph Foundation Models towards the emergence and homogenization capabilities on and for graph data and graph tasks. In this section, we will first define the concepts of graph foundation models. Then, we will discuss the impact from graph data and graph tasks on graph foundation models. Next, we will discuss the similarities and differences between graph foundation models and language foundation models. Finally, we will delve into the avenues for future exploration in this research area.

4.2 GNN-Based Methods (45 mins)

- Presenter: Dr. Cheng Yang
- Using GNNs as backbones towards GFMs
- Review and summarize
- Model details and strategies

Thanks to effective model architectures and training paradigms, language models have achieved remarkable performance in natural language processing tasks. The backbone, pre-training and adaptation techniques employed in language models have inspired a series of corresponding efforts in the field of graph-based tasks. In this section, we will delve into GNN-based methods, which draw inspiration from the model architectures or training paradigms used in the language domain to apply them to graph data and graph-related tasks. Importantly, unlike the LLM-based models and GNN+LLM-based models to be introduced in the following sections, GNN-based models do not explicitly model text data in their pipeline.

4.3 LLM-Based Methods (45 mins)

- Presenter: Dr. Yuan Fang
- Using LLMs as backbones towards GFMs
- Review and summarize
- Model details and strategies

Researchers are actively exploring ways to leverage LLMs as the core or even sole backbone for graph learning, given the following advantages. First, transformer-based models show a remarkable ability to seamlessly integrate textual information into graph data. Second, employing an LLM-like backbone empowers the model to unify diverse graph learning tasks, as these tasks can be described using natural language. Furthermore, recent advancements, such as NLGraph [8] and GPT4Graph [2], showcase the power of LLMs in preliminary graph reasoning. These advantages mark a highly promising direction for the development of such models. To discover the potential of integrating LLMs into graph learning, these works involve both graph-based properties and textual information as input for the backbone architectures.

4.4 GNN+LLM-Based Methods (45 mins)

- Presenter: Dr. Lichao Sun
- Using GNN+LLM as backbones towards GFMs
- Review and summarize
- Model details and strategies

GNN-based models lack the ability to process text and thus cannot directly make predictions based on textual data. Additionally, they cannot make predictions based on natural language instructions provided by users. On the other hand, LLM-based models for graph learning have their share of limitations. These limitations include the inability of LLMs to to explicit handle graph structures and multi-hop logical reasoning, etc. These shortcomings underline the necessity for further research and innovation in this domain. To overcome these limitations and harness the strengths of both language understanding from LLMs and structural analysis from GNNs, integrating LLMs and GNNs can potentially lead to a more comprehensive and powerful backbone.

5 AUDIENCE

Target Audience: Conference attendees interested in deep learning on graphs, large language models, and foundational models in general. This tutorial is designed to suit various backgrounds, whether the attendee is new to these subjects and seeking a solid introduction, or an experienced professional looking to delve deeper and gain new insights. It is designed to provide a comprehensive and systematic understanding of these subjects as well as their applications, making it accessible and informative for attendees from both academic and industrial backgrounds.

Prerequisites: While there is no specific prerequisite, the audience is expected to have some basic or general understanding of learning on graphs and language models. However, this tutorial will provide a detailed yet accessible introduction, and we will begin by explaining this background. Therefore, both researchers from academia and professionals from industry can easily follow our pace.

Potential Learning Outcomes: After this tutorial, the audience would benefit from the following potential outcomes.

- Understanding GFMs: Participants will gain a solid understanding of what Graph Foundation Models are, including their key characteristics, technological underpinnings, and their role in the field of artificial intelligence.
- Categorizing Research: Participants will learn to categorize existing research towards GFMs into three distinct categories based on their reliance on graph neural networks and large language models.
- Comprehensive Overview: Attendees will obtain a comprehensive overview of the current landscape of graph foundation models, allowing them to grasp the state of the art in the field.
- Potential Research Directions: The tutorial will spark discussions on potential research directions in the evolving field of Graph Foundation Models, enabling participants to identify new avenues for exploration.
- Access to Additional Resources: Attendees will be informed about additional resources, such as our preprint survey paper [4] and slides of this tutorial, which they can use for further study and research in this area.
- Engagement in Discussions: The tutorial will encourage participants to engage in discussions, share their perspectives, and ask questions, fostering a deeper understanding and exchange of ideas on the topic.

6 PREVIOUS EDITIONS

As this tutorial covers a novel direction, it has not been presented before. It will be the first time we introduce this topic at a conference, and the Web Conference 2024 will be a timely venue.

7 TUTORIAL MATERIALS

Tentatively, the participants will have access to various material, such as but not limited to the following:

- Our preprint survey paper "Towards Graph Foundation Models: A Survey and Beyond" [4], and any related resources;
- An organized list of papers referenced in the tutorial, available at https://github.com/BUPT-GAMMA/GFMPapers;
- Presentation slides of the tutorial.

8 VIDEO TEASER

A video teaser is available[2] at https://t.ly/lqPmI

KEY REFERENCES IN THIS TUTORIAL

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[2] Jiayan Guo, Lun Du, and Hengyu Liu. 2023. GPT4Graph: Can Large Language Models Understand Graph Structured Data? An Empirical Evaluation and Benchmarking. *arXiv preprint arXiv:2305.15066* (2023).

[3] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open* (2022).

[4] Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. 2023. Towards Graph Foundation Models: A Survey and Beyond. *arXiv preprint arXiv:2310.11829* (2023).

[5] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. 2022. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 6 (2022), 5879–5900.

[6] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63, 10 (2020), 1872–1897.

[7] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2016), 17–37.

[8] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can Language Models Solve Graph Problems in Natural Language? *arXiv preprint arXiv:2305.10037* (2023).

[9] Yang Yuan. 2023. On the power of foundation models. In *International Conference on Machine Learning*. PMLR, 40519–40530.

[10] Ziwei Zhang, Haoyang Li, Zeyang Zhang, Yijian Qin, Xin Wang, and Wenwu Zhu. 2023. Large Graph Models: A Perspective. *arXiv preprint arXiv:2308.14522* (2023).

[2]In case the short url doesn't work https://www.dropbox.com/scl/fi/ecsdvj9d5pnovheg684ga/GFM-Tutorial-20231109.mp4?rlkey=qix43ex770sxxn84zqhfhzw9l&dl=0