Enhancing Performance and Energy Efficiency of Networked Systems by Data-Driven Solutions

Yibo Ma

Abstract

The rapid advancement of new applications, such as autonomous driving, large language models, and smart homes, has led to significant challenges in managing the performance and energy efficiency of networked systems. These systems face bottlenecks due to high connectivity demands, data-intensive operations, and the growing complexity of modern applications. This research aims to address three major challenges: *i*) data scarcity, *ii*) increased system complexity, and *iii*) difficulty in large-scale optimization. By leveraging time-series system measurements and developing high-fidelity digital twins, this research project proposes a scalable, data-driven framework that improves system performance prediction, fault detection, and optimization. The research project seeks to create innovative algorithms and tools that capture the dynamic behaviors of large-scale systems, allowing for comprehensive analysis, multi-level modeling, and collaborative optimization. The findings will advance both academic understanding and industry practices, enabling more efficient and sustainable deployments of wireless networks and mobile systems.

1 Introduction

The exciting advances in machine learning and new applications, such as autonomous driving [3], large language models [6], and smart homes [1], bring more intensive connectivity and higher traffic demands to networked systems [10], which are limited by performance bottlenecks [21, 26, 7]. For example, large-scale deep learning recommender systems continuously collect new data and disseminate gradients via Wide-Area Networks to update models [17]. This process can take minutes to hours, resulting in high latency for delivering new content to clients [27]. Additionally, the increased carbon emissions from loading these computationally demands in systems conflict with the United Nations' sustainable development goals (SDGs) [9], further limiting the societal impact of these technologies [5].

Data-driven frameworks, powered by comprehensive time-series system measurements, make it possible for the creation of high-fidelity digital twins that model the dynamic behaviors of complex systems [20, 18, 15, 14, 16]. These models facilitate performance prediction, fault detection, and risk assessment under varying conditions, while also supporting system optimization through strategy testing and iterative feedback. By integrating expert knowledge and advanced optimization algorithms, such frameworks offer a flexible and interpretable solution for managing and optimizing networked networked systems. However, three main challenges limit the performance of existing data-driven frameworks in real-world networked systems: *i*) **Data scarcity.** Due to cost limitations and privacy concerns, large-scale, multi-metric data collection from real-world systems is difficult [28]. A few studies have collected city-scale data through extensive collaborations with operators [19], but due to security risks, part of the data cannot be open-sourced to benefit the wider research community. *ii*) **Increased system complexity.** To address emerging technologies and applications, system complexity is rising rapidly, with more system layers, complex interactions among components, and dynamic scenario changes [30, 35]. These factors make high-fidelity twin modeling, accurate performance prediction, and large-scale system performance evaluation difficult. *iii*) **Difficulty in large-scale optimization.** The action space for optimizing networked systems grows exponentially with system size. Previous research typically treats large-scale systems as a collection of small-scale subsystems, optimizing each independently [33]. However, as interactions and scenarios in the systems become more complex, this abstraction may oversimplify system dynamics and ignore inter-subsystem interactions, leading to local optima.

By identifying the three challenges outlined, this research project aims to address three key research questions, contributing meaningful algorithms and tools that will advance the application

of data-driven frameworks to overcome these issues:

• How can we capture the distribution of different indicators from networked system measurement data to generate networked system measurement data in common contextual scenarios, particularly focusing on metrics that are important for modeling and optimization?

• How can we conduct a comprehensive analysis of both the micro-mechanisms and macro patterns in networked systems to accurately model the operation mechanisms of components and interactions between them, and construct multi-level digital twin models for networked systems?

• How can we effectively leverage the comprehensive analysis of networked systems to guide the design of global optimization algorithms, enabling their scalable application to the collaborative optimization of various configurations or operations in large-scale networked systems with minimal efficiency cost?

This research project aims to develop a scalable, data-driven framework to enhance the performance and energy efficiency of networked systems. By addressing challenges of data scarcity, system complexity, and large-scale optimization, the framework will offer valuable insights and algorithms for real-world networked systems. Key findings will be disseminated through peer-reviewed publications in top conferences on networking, systems, and machine learning. Additionally, this research project can benefit the academic community and Industry. Through the generation of synthetic datasets, high-fidelity digital twins, and advanced optimization strategies, this work helps researchers to create customized models, support system benchmarking, and optimize real-world networked systems to improve the efficiency of large-scale applications.

2 Literature Review

2.1 Data Generation for Networked Systems

Measurement data in networked systems, typically represented as time series, reflects the operational patterns of the system. Early research on generating such data focuses on capturing implicit distributions and temporal dependencies. For instance, Wang et al. [31] propose using recurrent neural networks (RNNs) combined with generative adversarial networks (GANs) to generate traffic data. RNNs [4], as powerful models for sequence modeling, effectively capture temporal dependencies, while GANs [8] employ an adversarial learning architecture to iteratively improve data generation quality, capturing the implicit data distribution. However, such studies often overlook the potential correlations between contextual factors and system measurement data.

With advances in conditional generative models, recent works have incorporated contextual embeddings as conditions for data generation. For example, Xu et al. [32] propose the SpectraGAN model, which considers data such as points of interest (POI) and land use as conditional embeddings. Hui et al. [13] use device functionality to generate mobile traffic for IoT devices. Zhang et al. [34] introduce a transfer learning paradigm based on Urban Knowledge Graphs (UKGs) to generate corresponding traffic data for cities lacking sufficient data. However, these approaches often rely on non-public datasets, such as constructing complex POI relationships in UKGs or leveraging service-level data sources in AppShot [29], making their reproduction and application challenging. Identifying and effectively utilizing easily accessible contextual conditions to enhance data generation quality remains a critical research question.

Moreover, GANs are notorious for training instability and susceptibility to mode collapse, which can result in generating fixed data patterns unrelated to the conditions. Diffusion models predict noise through a noise prediction network and gradually recover the original data from sampled noise. These models demonstrate substantial potential for data generation and are well-suited for incorporating contextual conditions. While diffusion models [11] have been highly successful in image generation tasks—leveraging semantic or contextual conditions to control outputs—they require further exploration and adaptation for capturing spatiotemporal data distributions and introducing conditions in large-scale networked systems.

2.2 Performance Optimization for networked Systems

Existing studies have optimized the network performance of distributed networked systems from multiple aspects. Some have proposed solutions such as model compression and parallelism, effectively reducing the data volume for communication between computing nodes, but without optimizing the system's operational mechanisms and resource allocation. Other research has focused on improving communication efficiency from the perspective of distributed network protocols.

Recht et al. [23] proposed the Asynchronous Parallel (ASP) protocol, which enhances parameter synchronization efficiency during model training by eliminating network communication bottlenecks in the critical path of distributed nodes. Sapio et al. [25] reduced traffic within the network on a hop-by-hop basis. However, as the number of system nodes increases, the ratio of communication to computation demands grows, leading to significant performance degradation and limited scalability for these methods.

Google introduces a new distributed machine learning architecture, federated learning [22], which facilitates participants to collaboratively build models without sharing local data. However, the communication overhead in federated learning far exceeds computational costs. Due to bandwidth limitations and inconsistent connection quality across clients, federated learning relies on communication compression techniques to reduce bandwidth usage, such as gradient compression [24] and parameter compression [2].

Recently, many studies have recognized that data-driven simulation and optimization can achieve more efficient system operations and resource scheduling. For instance, Huang et al. [12] simulate the operations of geographically distributed data centers in a virtual domain and optimize data table placement and task scheduling to minimize network bandwidth usage under limited resource constraints and real-time response requirements. Li et al. [19] model the energy consumption of the units in mobile networks, base stations, to simulate city-scale energy consumption patterns. By analyzing the impact of various traffic scenarios on energy consumption, they develop energy consumption models for 5G networks in different provinces of China and test data-driven energy-saving optimization algorithms in a virtual domain to obtain performance feedback and train models. These data-driven optimization strategies require an in-depth analysis of system operations to gain solid insights for high-fidelity system modeling. However, due to the frequent updates and deployments of networked systems, such modeling and optimization methods may oversimplify interactions between components in increasingly complex systems. Further research is needed to achieve high-precision modeling and large-scale optimization in more complex systems with diverse requirements.

3 Methodology

Depending on more comprehensive and granular data collection, the data-driven framework has become a feasible solution for managing and optimizing networked systems. Specifically, by utilizing time-series system measurement data collected from real-world systems, high-fidelity digital twins can be established in virtual domains to accurately model the dynamic behaviors and interactions of complex systems, reflecting their operational patterns. Analyzing a combination of configurations and diverse scenarios allows for performance estimations under varying demands and facilitates proactive fault detection and risk assessment using time-series forecasting algorithms. For system optimization, translating system operations and strategies into actions within the virtual twin can predict the outcomes of these strategies, providing feedback for iteration and optimization. This process is akin to reward estimation of model-based reinforcement learning but with better interpretability and scene generalization. As a flexible framework, data-driven frameworks can continuously integrate expert knowledge and advanced optimization algorithms.

This research project aims to address the three challenges outlined above and develop more robust and reliable data-driven solutions for optimizing resource utilization and energy efficiency in networked systems, particularly in edge and mobile computing environments. Specifically, to tackle the first challenge, data scarcity, this research project involves custom deep generative models to generate high-fidelity system measurement data. These conditional models, such as Variational Autoencoders, Generative Adversarial Networks, and Diffusion Models, can generate time-series data with the operational scenarios of the system serving as conditions to control the generation process. This approach generates the system measurement data that closely resembles real-world data patterns when given contextual features, thus supporting system performance analysis, modeling, and optimization. To address the second challenge, system complexity, this study leverages extensive data analysis to extract key insights for a deeper understanding of system operations, such as business change patterns, causal dependencies between performance metrics, system component interactions, and resource scheduling rules. This allows for the targeted application of high-fidelity simulation methods to replicate network operation. To address the third challenge, large-scale optimization, this study applies advanced optimization methods for large-scale networked system optimization. Graph Neural Networks (GNNs) are used to extract correlation features between components or subsystems, effectively capturing complex dependencies of system structure, and

thereby overcoming the challenges of optimization in large-scale systems due to the exponential growth of the action space. Additionally, Large Language Models (LLMs) are employed as intelligent agents to understand global optimization goals and local contextual features through carefully designed prompts and reasoning chains. The rich prior knowledge from pre-trained models, combined with expert knowledge introduced in the prompts and reasoning processes, allows for a more interpretable reduction of the optimization action space, thereby addressing the issue of excessive action space in large-scale systems.

4 Dissemination and Timeline

This research project aims to develop a reliable, scalable, and deployable data-driven framework to enhance the performance and energy efficiency of real-world networked systems through extensive analysis and continuous experimentation. The framework, designed to address the three outlined challenges, incorporates a deep understanding of the operation of networked systems in real-world scenarios, with carefully designed components and algorithms. These insights and algorithms will be shared with the international academic community through peer-reviewed publications in high-quality conferences such as *ACM SIGCOMM*, *USENIX NSDI*, *NeurIPS*, and others.

During the first year of the Ph.D., the focus will be on foundational coursework in machine learning, deep generative models, network systems, and large-scale optimization. Initial experiments will explore methods to address data scarcity, specifically through the design and application of conditional generative models to generate realistic system measurement data for mobile computing environments. In the second year, the research will deepen with extensive data analysis to understand system operations, key performance metrics, and component interactions. This will help the development of high-fidelity simulations for resource optimization. Collaboration with peers in academic institutes will be a key aspect of the research, especially in the context of system complexity. By the third year, the focus will shift to the design and application of advanced optimization algorithms, addressing large-scale system challenges by improving efficiency and performance in resource allocation. Fine-tuning optimization strategies using expert knowledge and system data will enhance the precision of these algorithms. The work will be presented at workshops and conferences, with efforts aimed at producing high-impact journal publications. The final year will be dedicated to completing the thesis, synthesizing the research on data scarcity, system complexity, and optimization, followed by the defense of the dissertation. An additional publication will focus on the application of these optimization algorithms in real-world network environments. Career planning will focus on opportunities for postdoctor careers in academia or industry, particularly in large-scale and AI-driven systems.

5 My Preparation

My research journey in the past has been deeply rooted in large-scale networks, energy optimization, and the development of digital twins, particularly in the context of mobile networks and sustainable energy. During my master's program, I worked on estimating carbon emissions of 5G networks, contributing to publication in *Nature Sustainability*. In a work published in *IEEE TNSM*, I proposed an energy-saving algorithm, REDEEM, which involves time-series prediction and energy efficiency analysis, laying a strong foundation for my future research. This experience sharpened my problem-solving skills and deepened my interest in large-scale network measurement and optimization. I then expanded my research focus with a project on 5G network usage and energy efficiency, where I developed an accurate energy consumption model and identified key factors influencing network efficiency. The resulting insights and energy-saving strategies were published in *ACM CoNEXT 2024* and *IEEE TMC*, emphasizing my ability to conduct comprehensive, data-driven analyses and propose actionable solutions for network operators. Another significant project involved deep simulation techniques, where I used Variational Autoencoders to model user mobility and app usage, addressing the environmental impact of 5G network adoption on residential power systems. This work honed my skills in bridging micro and macro-level systems modeling and simulation, essential for tackling complex challenges in networked systems. My most recent project focused on developing a high-fidelity digital twin for optimizing energy-saving strategies in mobile networks. I proposed the SoftControl framework, which addressed practical deployment challenges and demonstrated measurable energy savings, highlighting my ability to translate theoretical models into real-world applications. These experiences have not only enhanced my technical expertise

in energy optimization, system modeling, and deep learning but also fueled my ambition to drive advancements in large-scale networked systems. They provide a solid foundation for my future research, where I aim to develop more efficient, scalable, and sustainable data-driven solutions in the rapidly evolving networked systems landscape.

6 Ethical Considerations

This research project involves the collection and analysis of large-scale networked system data. Given the potential for privacy concerns and the need for responsible data usage, ethical considerations must be at the forefront of this work. First, the research project may involve the use of real-world data, including user interactions with networked systems. To protect user privacy, data will be anonymized and aggregated to ensure that no personally identifiable information is used in the analysis. Furthermore, any data collected will strictly adhere to data protection regulations. Secure data storage, encryption, and access control measures will be implemented to prevent unauthorized access. Users will be informed about the nature of the data being collected, the purpose of its use, and their rights in terms of data access and withdrawal. In cases where data is collected from third-party sources, the data-sharing practices and policies of those organizations will be reviewed to ensure compliance with ethical standards. The use of data for generating synthetic datasets, developing digital twins, and performing system optimizations will prioritize fairness and transparency. In summary, this research project will prioritize user privacy, fairness, and security of networked systems. Ethical considerations will be integrated throughout the research process to ensure that the outcomes contribute positively to both academic and societal goals while safeguarding the rights and interests of individuals involved in the data collection and analysis.

7 Discussion

The results of this research project can benefit a broader community. By addressing the challenge of data scarcity, this research project generates and releases datasets for regions where data collection is difficult, reflecting the operational patterns of real-world systems without security risks. Scholars and researchers can directly use these datasets to develop additional models and algorithms, or apply the proposed data generation methods to create customized datasets, further advancing research in the networked systems field. By addressing system complexity, the high-fidelity digital twin models established by this research project can help developers benchmark their products and assess the performance in real-world systems. Additionally, the deep insights derived from system operation analysis, such as spatiotemporal patterns, causal dependencies, and interaction relationships, can assist system maintenance and tuning teams in designing tailored solutions to meet highly heterogeneous demands for various scenarios and needs. By tackling the challenge of large-scale optimization, the proposed optimization strategies overcome the difficulties posed by the exponential growth of system size and action space. These strategies have vast application potential in emerging fields such as ultra-dense 5G networks, IoT, and robotics networks, improving resource utilization and energy efficiency of large-scale systems, and minimizing the barriers to deploying these promising systems, thus promoting the efficient use of emerging technologies.

8 Conclusion

This research project aims to tackle the challenges faced by real-world networked systems, particularly in terms of data scarcity, increasing system complexity, and the difficulties associated with large-scale optimization. By developing a comprehensive, data-driven framework, the project addresses critical gaps in capturing system measurements, accurately modeling complex interactions, and enabling scalable optimization. Through the creation of high-fidelity digital twins and advanced optimization algorithms, this framework offers the potential to enhance both the performance and energy efficiency of networked systems. The results of this research project will provide meaningful contributions to the academic community and industry, enabling more efficient, sustainable, and scalable networked system deployments. By focusing on these key areas, this research project aims to significantly improve the practical application of data-driven frameworks and their ability to overcome current limitations, ultimately advancing the broader field of networked systems.

References

[1] Asem Alzoubi. Machine learning for intelligent energy consumption in smart homes. *International Journal of Computations, Information and Manufacturing (IJCIM)*, 2(1), 2022.

[2] Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[3] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015.

[4] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[5] Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Prateek Sharma, Fan Chen, and Lei Jiang. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. *arXiv preprint arXiv:2309.14393*, 2023.

[6] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023.

[7] Görkem Giray. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software*, 180:111031, 2021.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] David Griggs, Mark Stafford-Smith, Owen Gaffney, Johan Rockström, Marcus C Öhman, Priya Shyamsundar, Will Steffen, Gisbert Glaser, Norichika Kanie, and Ian Noble. Sustainable development goals for people and planet. *Nature*, 495(7441):305–307, 2013.

[10] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE international symposium on high performance computer architecture (HPCA)*, pages 620–629. IEEE, 2018.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Yuzhen Huang, Yingjie Shi, Zheng Zhong, Yihui Feng, James Cheng, Jiwei Li, Haochuan Fan, Chao Li, Tao Guan, and Jingren Zhou. Yugong: Geo-distributed data and job placement at scale. *Proceedings of the VLDB Endowment*, 12(12):2155–2169, 2019.

[13] Shuodi Hui, Huandong Wang, Tong Li, Xinghao Yang, Xing Wang, Junlan Feng, Lin Zhu, Chao Deng, Pan Hui, Depeng Jin, et al. Large-scale urban cellular traffic generation via knowledge-enhanced gans with multi-periodic patterns. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4195–4206, 2023.

[14] Shuowei Jin, Francis Y Yan, Cheng Tan, Anuj Kalia, Xenofon Foukas, and Z Morley Mao. Autospec: Automated generation of neural network specifications. *arXiv preprint arXiv:2409.10897*, 2024.

[15] Siva Kesava Reddy Kakarla, Francis Y Yan, and Ryan Beckett. Diffy: Data-driven bug finding for configurations. *Proceedings of the ACM on Programming Languages*, 8(PLDI):199–222, 2024.

[16] Caner Kilinc, Mahesh K Marina, Muhammad Usama, Salih Ergut, Jon Crowcroft, Tugrul Gundogdu, and Ilhan Akinci. Jade: Data-driven automated jammer detection framework for operational mobile networks. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1139–1148. IEEE, 2022.

[17] Fan Lai, Wei Zhang, Rui Liu, William Tsai, Xiaohan Wei, Yuxi Hu, Sabin Devkota, Jianyu Huang, Jongsoo Park, Xing Liu, et al. {AdaEmbed}: Adaptive embedding for {Large-Scale} recommendation models. In *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pages 817–831, 2023.

[18] Franck Le, Mudhakar Srivatsa, Raghu Ganti, and Vyas Sekar. Rethinking data-driven networking with foundation models: challenges and opportunities. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks*, pages 188–197, 2022.

[19] Tong Li, Li Yu, Yibo Ma, Tong Duan, Wenzhen Huang, Yan Zhou, Depeng Jin, Yong Li, and Tao Jiang. Carbon emissions of 5g mobile networks in china. *Nature Sustainability*, 6(12):1620–1631, 2023.

[20] Jiachen Liu, Fan Lai, Yinwei Dai, Aditya Akella, Harsha Madhyastha, and Mosharaf Chowdhury. Auxo: Heterogeneity-mitigating federated learning via scalable client clustering. *arXiv preprint arXiv:2210.16656*, 2022.

[21] Lucy Ellen Lwakatare, Aiswarya Raj, Ivica Crnkovic, Jan Bosch, and Helena Holmström Olsson. Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. *Information and software technology*, 127:106368, 2020.

[22] H Brendan McMahan, FX Yu, P Richtarik, AT Suresh, D Bacon, et al. Federated learning: Strategies for improving communication efficiency. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain*, pages 5–10, 2016.

[23] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

[24] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR, 2020.

[25] Amedeo Sapio, Ibrahim Abdelaziz, Abdulla Aldilaijan, Marco Canini, and Panos Kalnis. In-network computation is a dumb idea whose time has come. In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, pages 150–156, 2017.

[26] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

[27] Chijun Sima, Yao Fu, Man-Kit Sit, Liyi Guo, Xuri Gong, Feng Lin, Junyu Wu, Yongsheng Li, Haidong Rong, Pierre-Louis Aublin, et al. Ekko: A {Large-Scale} deep learning recommender system with {Low-Latency} model update. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 821–839, 2022.

[28] Chuanhao Sun, Kai Xu, Gianni Antichi, and Mahesh K Marina. Netgsr: Towards efficient and reliable network monitoring with generative super resolution. *Proceedings of the ACM on Networking*, 2(CoNEXT4):1–27, 2024.

[29] Chuanhao Sun, Kai Xu, Marco Fiore, Mahesh K Marina, Yue Wang, and Cezary Ziemlicki. Appshot: a conditional deep generative model for synthesizing service-level mobile traffic snapshots at city scale. *IEEE Transactions on Network and Service Management*, 19(4):4136–4150, 2022.

[30] Sandeep Tata, Alexandrin Popescul, Marc Najork, Mike Colagrosso, Julian Gibbons, Alan Green, Alexandre Mah, Michael Smith, Divanshu Garg, Cayden Meyer, et al. Quick access: building a smart experience for google drive. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1643–1651, 2017.

[31] Zi Wang, Jia Hu, Geyong Min, Zhiwei Zhao, and Jin Wang. Data-augmentation-based cellular traffic prediction in edge-computing-enabled smart city. *IEEE Transactions on Industrial Informatics*, 17(6):4179–4187, 2020.

[32] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, pages 243–258, 2021.

[33] Zhiying Xu, Francis Y Yan, and Minlan Yu. Zeal: Rethinking large-scale resource allocation with" decouple and decompose". *arXiv preprint arXiv:2412.11447*, 2024.

[34] Shiyuan Zhang, Tong Li, Shuodi Hui, Guangyu Li, Yanping Liang, Li Yu, Depeng Jin, and Yong Li. Deep transfer learning for city-scale cellular traffic generation through urban knowledge graph. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4842–4851, 2023.

[35] Yifan Zhang, Bo Liu, Yulu Gong, Jiaxin Huang, Jingyu Xu, and Weixiang Wan. Application of machine learning optimization in cloud computing resource scheduling and management. In *Proceedings of the 5th International Conference on Computer Information and Big Data Applications*, CIBDA '24, page 171–175, New York, NY, USA, 2024. Association for Computing Machinery.