Research Statement

Yanxiao Liu
Postdoctoral Research Associate, Imperial College London
yanxiaoliu@link.cuhk.edu.hk

I am trained as an information theorist. My research lies broadly in the area of **information science**, with a focus on **information theory**, **network science**, **distributed learning**, and **privacy**. In the era of artificial intelligence (AI), the scale of data required to sustain progress is rapidly outpacing the hardware capabilities available to support it. All the collected data must be transmitted for training and analytics, but this process faces constraints in bandwidth and storage. It is therefore crucial to develop new methods to reduce the memory, computational, and communication burdens of large-scale data and networks, while minimally impacting the performance of the algorithms that rely on them. At the same time, much of these valuable data inherently contain sensitive information, making it vulnerable to privacy breaches during acquisition, transmission, or use. Thus, ensuring the privacy and security of data is also a central concern. Information theory provides powerful tools that address **both** issues.

My long-term goal is to develop and apply information-theoretic techniques to large data-driven problems across and beyond computer science, statistics and machine learning. In my research, I seek mathematically rigorous solutions and aim to determine **fundamental limits** on learning, communication, and privacy in data science problems. To uncover these limits, it is essential to understand the underlying structure of the problems. I derive explicit and efficient algorithms and analyses to gain deeper insights into those structures. Figure 1 presents a diagram that illustrates the scope of my research.

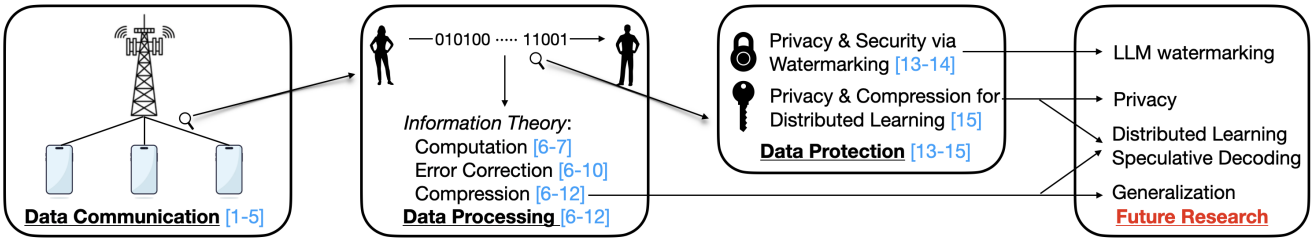

Figure 1: My research roadmap. The blue brackets highlight my papers related to different themes.

In Figure 1, my past research has been divided to three parts:

1. **Data Communication over Mobile Networks**: This part involves gathering data from vast numbers of edge devices over large, unreliable networks. Key challenges include ensuring reliable collection amid massive random access, unstable signals, and high latency, common in challenging environments like remote sensors, industrial IoT and underwater/deep-space networks. Scalability and energy efficiency further complicate coordination across such networks.

2. **Data Processing via Information Theory**: This part employs information-theoretic coding schemes to compress and quantize large-scale data efficiently while handling variable-length inputs and multi-stage processing, requiring universal schemes that adapt to arbitrary data sizes with low complexity yet strong theoretical guarantees, without prior knowledge of data statistics.

3. **Data Protection: Privacy, Security and Watermarking**: This part ensures robust protection of sensitive data while maintaining high utility for remote analytics, addressing challenges like

privacy-preserving computation, secure multi-party aggregation, and adversarial resilience. Moreover, along with the success of generative AI, it becomes crucial to embed imperceptible yet traceable watermarks to deter unauthorized use of the data, for compliance, federated learning, distributed AI systems and large language models (LLM).

I will now highlight several of my works and describe how I addressed the specific challenges in each case. In the final section, I will discuss my future research plans.

Data Communication over Mobile Networks

In modern mobile networks, a large number of edge devices from multi-dimensional platforms are integrated, and a massive amount of data is collected from these devices for storage, analytics, and centralized model training. Ensuring reliable data collection has become increasingly difficult due to several challenges: the large scale of the system, where the number of devices is enormous; unreliable data transmission caused by noise and limited communication among edge devices; and the latency and collisions that may occur when devices operate in challenging environments, such as underwater acoustic networks or deep space networks.

In this part of my research, I focus on network throughput optimization and scheduling, aiming to ensure reliable data collection from edge devices, even in large-scale and unreliable networks. In [1], we investigated a very general network model (without any assumptions on network topology, size and coding schemes, etc.) to study the fundamental limits of large-scale multi-hop networks in terms of optimized network multiflow, with network coding employed at intermediate nodes. Using a novel decomposition method and a joint optimization framework, we developed explicit algorithms that are *exponentially faster* than conventional approaches. In [2]–[4], we studied the scheduling of data transmissions over large-scale networks, possibly in challenging environments (e.g., underwater or deep-space communication), where large propagation delays make the scheduling problem exponentially more difficult. We provided the first efficient algorithm that can explicitly characterize the *fundamental limits* (i.e., the entire scheduling rate region) of such networks. Furthermore, in [5], we addressed massive random access in networks without feedback, where packet collisions are common and data transmissions must still meet guaranteed performance levels. We proposed constructive schemes that overcome these constraints.

For these works, I was very fortunate to collaborate with leading scholars in the information theory and communication communities (Raymond W. Yeung, Shenghao Yang and Yi Chen). These challenges are frequently encountered in modern mobile systems, and a deeper understanding of the theoretical limits, together with explicit algorithmic solutions, can greatly enhance the design and analysis of downstream data analytics tasks. For example, in scenarios where massive edge devices are deployed to transmit collected data back to central servers for model training (e.g., in federated learning), reliable communication through noisy and unreliable environments is essential.

Data Processing via Information Theory

For data transmission and compression, an important stage is to encode the raw data, either to ensure robustness against noise and errors during transmission, or to reduce its size (possibly at the cost of an acceptable level of distortion, i.e., quantization noise). In information theory, these processes are known as *channel coding* and *source coding*, respectively. Both have been extensively studied over the past 70 years, beginning with the foundational work of Shannon. In modern data science, information theory plays a critical role by providing the theoretical foundation for various phenomena, as well as offering efficient solutions of data analytics problems in unreliable or communication-constrained environments.

In this part of my research, I investigate the fundamental limits of canonical information theory problems. A key insight driving my work is that many long standing assumptions in the literature, though historically influential, can often be relaxed without sacrificing performance, leading to more precise benchmarks for contemporary applications. Conventional information theory relies on several

assumptions: the signal to be transmitted or compressed has a blocklength approaching infinity (hence the law of large numbers can be employed to derive the asymptotic behaviors); coding schemes are tailored to specific network settings; the roles of encoders and decoders are separated; and the constructions of channel coding schemes and source coding schemes are treated as fundamentally different. In [6], [7], we demonstrated that these assumptions are not always necessary. We designed a unified *one-shot* coding scheme that can encode signals with *arbitrary* blocklength and is applicable to *any* combination of source coding, channel coding, and coding for computing tasks, over *general* noisy networks that may contain an *arbitrary* number of nodes and *arbitrary* topology. This scheme unifies the roles of encoders and decoders, as well as channel and source coding, significantly simplifies the analysis, and yields various novel information theoretic insights, while also recovering most of the existing results in the literature.

Following a soft-coding idea in [6], [7], we designed novel linear coding schemes that achieve state-of-the-art performance with short blocklengths on channels with side information [8]–[10]. This setting commonly arises in federated learning, where the encoder can access certain side information to enable more effective coding, while the transmitted packets, which often contain model updates from edge devices, are typically short. In such cases, conventional coding schemes tend to perform poorly, making tailored short blocklength designs particularly important.

In [11], [12], we studied the information bottleneck channel, which is a compression task that is tightly related to remote lossy source coding, both of which related to the information bottleneck in deep learning. We provided a novel scheme based on the remote source generation, and provided the first nonasymptotic analyses through explicit mechanism designs on this problem.

Moreover, in [13], [14], we extended our one-shot coding schemes to *secrecy* problems, where signal transmissions are unreliable. That is, the transmission channels have uncertainties, and the transmitted data may be *attacked* through noise or distortion. The former led to a novel watermarking scheme that can efficiently embed secret information into a host source, and this technique has the potential to address the AI-safety concerns arising alongside the success of generative AI: the detection and protection of AI generated content. We will elaborate on this in the next section.

Data Protection: Privacy, Security and Watermarking

Due to the growing dependence on large volumes of high quality data that are often generated by edge devices such as photos and videos captured by smartphones, or messages hosted by social networks, it is crucial to protect such data from breaches during acquisition, collection, or utilization. Although adding noise to the data (e.g., via *differential privacy (DP)*) enhances security and privacy, it also increases the entropy of data. As a result, the communication of local data from edge devices to a central server becomes a bottleneck in the system pipeline, especially when dealing with high-dimensional data, which is common in machine learning applications.

In [15], we address the fundamental question that in privacy-preserving distributed learning: **how can we efficiently communicate privatized data?** We employ information theoretic techniques, specifically channel simulation (also known as remote source generation), to compress DP mechanisms. The idea is intuitive: by *interpreting DP mechanisms as channels*, we design a novel algorithm leveraging Poisson processes to *simulate* those channels. Unlike existing approaches, which are either limited to specific mechanisms (such as those using additive noise) or distort the output distribution, our algorithm *universally* compress DP mechanisms to near-optimal sizes, while *exactly* preserving the joint distribution of the input data and the output of the original local randomizer. This ensures that all desirable statistical properties (e.g., unbiasedness and Gaussianity) are retained, and start-or-the-art empirical performance in distributed mean estimation (which is the canonical task in distributed learnin) can be achieved. For this work, I am fortunate to collaborate with leading scholars in federated learning and differential privacy communities (Ayfer Özgür and Wei-Ning Chen).

Moreover, with the rapid success of generative AI and LLM, it is critical not only to protect but also to detect content generated by generative models, and the design of *watermarking* tools becomes

essential. In [13], [14], we proposed a novel watermarking scheme that uses one-shot codes based on the Poisson processes, deriving a *distributionally robust* scheme against external attacks on the watermarked content. Although this scheme was not originally designed for LLM watermarking, we have discussed this potential applications in [13], [14]. Given the recent advancements in information-theoretic watermarking techniques for LLMs, this direction is promisingr. Moving forward, we plan to further investigate this direction soon.

Future Works

The relevance of information theory continues to grow as it increasingly intersects with fields such as machine learning, data analysis, security, and privacy. Looking ahead, I will continue to advance the boundaries of information theory and develop new new techniques applicable to the important problems in real-word, in the era of data science. In addition to furthering the directions I have explored, I am also passionate to investigate several new and promising research directions, as described below.

- **Compression, Privacy and Distributed Learning**: An emerging interplay between information theoretic quantization techniques, privacy mechanisms, and machine learning applications has been observed in recent years, driven by the success of distributed learning. Dithered quantization and channel simulation, which I studied during my doctoral research, are popular techniques in lossy compression, and the resulting quantization noise can be carefully manipulated to provide privacy guarantees [15]. I am particularly interested in further advancing our understanding of the tradeoff between communication, privacy, and utility, and provide *exact-optimal* (rather than order-optimal) results on the tradeoff. A deeper understanding of this tradeoff can enable novel designs for neural network architectures.

- **Generalization via Compression**: Differential privacy is closely related to the generalization guarantee, or what is sometimes referred to as the *stability*, of stochastic algorithms. Recent studies have shown that private algorithms can generalize well, and one interesting direction would be investigating the privacy and utility of the compressed private algorithms, considering lossy compression can even improve the generalization performance. Moreover, rate-distortion results can help characterize the generalization error bounds in terms of the "compressibility" of the algorithms. We have briefly explored how the one-shot lossy source coding results (studied in [6], [7], [11], [12], [15]) provide better guarantees on the generalization error, and it is promising to advance current theories on generalization guarantees following this direction.

- **LLM Watermarking**: Due to the success of large language models, watermarking [13], [14] has become a critical issue, both for protecting AI generated output and for detecting whether certain content was produced by AI. Compared to classical watermarking, several new challenges arise: the source distribution is unknown to the watermark detector, and watermarking must be performed at the per-token level. We are interested in developing novel LLM watermarking tools based on our information-theoretic understanding and schemes of this problem. It would play an important role in designing responsible and safe AI models.

- **LLM Decoding**: The Gumbel-Max trick has been a widely known technique in machine learning and differential privacy, and it has been recently been applied to LLM decoding as well. Interestingly, recently we found that the one-shot coding schemes that I proposed [6], [7], [11]–[15] can be readily viewed as a nontrivial generalization of the Gumbel-Max trick with better properties! Therefore, I am eager to explore applying the coding schemes I proposed to LLM decoding, which could potentially result in an LLM decoder with a better design, faster decoding, and a higher success rate. This line of research could potentially open a new door from information theory to LLM design.

References

[1] **Yanxiao Liu**, Shenghao Yang, and Cheuk Ting Li, "Joint scheduling and multiflow maximization in wireless networks," *Submitted to IEEE Transactions on Networking*, 2025.

[2] Shenghao Yang, Jun Ma, and **Yanxiao Liu**, "Wireless network scheduling with discrete propagation delays: Theorems and algorithms," *IEEE Transactions on Information Theory*, vol. 70, no. 3, pp. 1852–1875, 2023.

[3] Yijun Fan, **Yanxiao Liu**, and Shenghao Yang, "Continuity of link scheduling rate region for wireless networks with propagation delays," in *2022 IEEE International Symposium on Information Theory (**ISIT**)*, 2022.

[4] Jun Ma, **Yanxiao Liu**, and Shenghao Yang, "Rate region of scheduling a wireless network with discrete propagation delays," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications (**INFOCOM**)*, 2021.

[5] Yijun Fan, **Yanxiao Liu**, Yi Chen, Shenghao Yang, and Raymond W. Yeung, "Reliable throughput of generalized collision channel without synchronization," in *2023 IEEE International Symposium on Information Theory (**ISIT**)*, 2023.

[6] **Yanxiao Liu** and Cheuk Ting Li, "One-shot coding over general noisy networks," in *2024 IEEE International Symposium on Information Theory (**ISIT**)*, 2024.

[7] **Yanxiao Liu** and Cheuk Ting Li, "One-shot coding over general noisy networks," *IEEE Transactions on Information Theory*, 2025.

[8] Chih Wei Ling, **Yanxiao Liu**, and Cheuk Ting Li, "Weighted parity-check codes for channels with state and asymmetric channels," in *2022 IEEE International Symposium on Information Theory (**ISIT**)*, 2022.

[9] Chih Wei Ling*, **Yanxiao Liu*(co-first author)**, and Cheuk Ting Li, "Weighted parity-check codes for channels with state and asymmetric channels," *IEEE Transactions on Information Theory*, vol. 70, no. 8, pp. 5573–5588, 2024.

[10] **Yanxiao Liu**, Chih Wei Ling, and Cheuk Ting Li, "Weighted polar codes for channels with state,"

[11] **Yanxiao Liu**, Sepehr Heidari Advary, and Cheuk Ting Li, "Nonasymptotic oblivious relaying and variable-length noisy lossy source coding," in *2025 IEEE International Symposium on Information Theory (**ISIT**)*, 2025.

[12] **Yanxiao Liu**, Sepehr Heidari Advary, and Cheuk Ting Li, "Nonasymptotic oblivious relaying and variable-length noisy lossy source coding," *Submitted to IEEE Transactions on Information Theory*, 2025.

[13] **Yanxiao Liu** and Cheuk Ting Li, "One-shot information hiding," in *2024 IEEE Information Theory Workshop (**ITW**)*, 2024.

[14] **Yanxiao Liu** and Cheuk Ting Li, "One-shot information hiding and compound wiretap channels," *Submitted to IEEE Transactions on Information Theory*, 2025.

[15] **Yanxiao Liu**, Wei-Ning Chen, Ayfer Özgür, and Cheuk Ting Li, "Universal exact compression of differentially private mechanisms," in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (**NeurIPS**)*, 2024.