Description of the topic for the doctoral thesis of ADEEL MALIK

Title of the Thesis: Fundamental limits of machine-learning aided caching

Description of the topic:

The proposed thesis aims to develop the theoretical foundations of coded caching, when this is assisted by advanced artificial intelligence methods. Malik will seek to explore the practical ramifications of such approaches in wireless communication networks.

Motivated by the long-lasting open challenge to invent a communication technology that scales with the network size, we have recently discovered early indications of how preemptive use of distributed data-storage at the receiving communication nodes (well before transmission), can offer unprecedented throughput gains by surprisingly bypassing the dreaded bottleneck of real-time channel-feedback. For an exploratory downlink configuration, we unearthed a hidden duality between feedback and preemptive use of memory, which managed to doubly-exponentially reduce the needed memory size, and consequently offered unbounded throughput gains compared to all existing solutions with the same resources. This was surprising because feedback and memory were thought to be mostly disconnected; one is used on the wireless PHY layer, the other on the wired MAC.

At a time of substantial improvements in machine learning methods, this development of ours prompts our key scientific challenge which is to pursue the mathematical convergence between advanced caching and machine learning, and to then design ultra-fast memory-aided communication algorithms, that pass a battery of tests for real-life validation.

This is a structurally new approach, which promises to reveal deep links between caching and machine learning approaches, for a variety of envisioned wireless-network architectures of exceptional promise. In doing so, our new proposed theory stands to identify the basic principles of how a splash of memory can surgically alter the informational structure of these networks, rendering them faster, simpler and more efficient. In the end, this study has the potential to directly translate the continuously increasing data-storage capabilities, and the continuously improving machine learning approaches, into gains of wireless network capacity, and to ultimately avert the looming network-overload caused by these same indefinite increases of data volumes.

This proposal is about applying machine learning approaches in memory aided large wireless communication networks, and what we propose is to develop the theoretical and practical foundations of how we can apply preemptive use of storage capacity at the nodes, and advanced artificial intelligence methods, to surgically alter the informational structure of communication networks, making them faster, simpler and more efficient.

Current communication solutions do not successfully scale in large networks, because as the number of users increases, these solutions cannot fully separate all users' signals. This problem in turn can gradually leave each user with near-zero communication rates, and it comes at a time when wireless data traffic is expected to increase by 10 times in just 5 years. If data volumes continue to expand so rapidly, it is foreseen that wireless networks will soon come to a halt, thus inevitably compromising the informational foundations of society. This looming network-overflow can also have severe environmental consequences, because current systems would require exponential increases in transmit-power to accommodate future data volumes; Telecommunications has in fact already a higher carbon footprint than aviation. Despite the imminence of this overload, the consensus is that there is no existing or currently envisioned technology that resolves this, not even with brute force increase of bandwidth, or of the number and size of base-stations.

The problem in addressing the above overload, mainly originates from the inherently real-time nature of communications, where data must be `served' on time. This entails having to continuously adapt communications – in real-time – to the rapidly fluctuating states of a large wireless network. `Feedback' in this context, refers to the action of disseminating large amounts of overhead information (referred to as channel-state information, or CSI) about the instantaneous strengths of each propagation-path between the different nodes. These channels fluctuate up to hundreds of times per second, hence as the network size increases, the overhead consumes more and more resources, eventually leaving no room for actual data. It is well understood that this overhead can provably bring all current systems and envisioned methods – which are based on the fundamentals of feedback information theory – to a halt.

Recently, an upper-layer solution was proposed that relied on caching network-coded data at the receiving devices. The proposed solution was motivated by the fact that wireless traffic is heavily video or audio on-demand (over 60%), which entails an ability to predict data-requests in advance (`the night before'). This approach was based on caching content from an existing library of many popular files. Each user would pre-store (without knowing next day's requests) a carefully selected sequence of sub-files from the library, specifically designed to speed up (next day's) communications. This promising approach did not fully scale though, this time because memory size at the receiving devices had to scale exponentially

in the number of users. Even if this issue was resolved, the gains would have still been very small.

These two approaches (feedback information theory and coded-caching) were thought to be disconnected; one uses feedback on the PHY layer, the other uses memory on MAC. We have evidence of a powerful duality between the two, which we believe can be harnessed by solving advanced optimization problems that fall in the area of machine learning and artificial intelligence.

This thesis addresses the issue of understanding how machine learning can boost communications performance by algorithmically telling us what is the best content that we should be placing in user storage, given different user's preferences, and viewing habits. The student will seek to eluminate the degree to which machine learning can help in understanding the different effects of caching in the performance and complexity of wireless networks, and the student will also seek to study the approximate impact of proper machine-learning based prediction of content (think of youtube) on the revenue of different companies such as operators as well as content providers. The student will suggest different machine learning approaches that yield data placement mechanisms to support different business models.

Our aim is to create a new theory that reveals how, with a splash of feedback and ML-based caching of a micro fraction of popular content, we can adequately handle the anticipated extreme increase in users and demand. This structurally new approach has the potential to unearth deep and surprising connections between distributed storage and machine learning.

Date Supervisor

---------- ----------