# Caching revival via information-theory and network theory: novel challenges and breakthroughs in cache-aided wireless networks

George Paschos (Huawei –Paris) and Petros Elia (EURECOM – Sophia Antipolis)

There is currently an urgent need for novel technologies that can partially mitigate the current explosion of wireless traffic volumes. While many existing communication technologies fail to scale with increasing network sizes, recent developments have revealed that caching, when properly transformed and boosted, can come a long way in augmenting the performance and efficiency of wireless networks. Our tutorial will seek to insightfully present the fundamental ingredients behind some recent breakthroughs, as well as describe the key challenges that remain in turning caching into a key ingredient for future wireless networks.

We begin our journey with a discussion of technological limitations that may prohibit the application of caching in wireless. What would it take to have a cache-enabled base station today? By addressing the concerns one by one, we separate them to a) those that are mere contemporary technological limitations and can be easily overcome in the future 5G networks, b) those that are related to business barriers and there is hope for a future resolution, and c) the purely theoretical challenges that may limit the efficiency of caching in wireless networks.

Next we focus on the fundamental research challenges for wireless caching. First we characterize wireless traffic and discuss novel modelling techniques for the item request sequence. These new models uncover temporal and spatial correlations among item popularities, which are crucial to caching efficiency. We show that by efficiently tracking the correlations and combining them with item prefetching, it is possible to improve the caching performance for small population caches using popularity detection.

The optimal amount of memory that we should install in a wireless network is related to the cache/catalogue ratio $\gamma=C/N$. Since wireless networks have limited availability of storage size (small C), we next focus on the technique of "partial caching" whereby the goal is to cache the parts of the content that are most popular. We examine a trace of Youtube videos—as representative traffic mix for wireless traffic—and show that partial caching makes caching feasible for smaller $\gamma$. Another approach to improve caching efficiency for a given $\gamma$ is to enable cache collaboration. The idea is to cache different items in neighboring caches that are jointly reachable by several locations. Nevertheless, due to reachability structure, the optimal item placement often becomes a combinatorial optimization problem. We survey the proposed approaches in this direction and we summarize the first part of the tutorial with a feasibility analysis of caching in wireless networks, whereby combining the above techniques we demonstrate the bandwidth savings for different values of $\gamma$.

We will then focus on a new technique referred to as coded caching, which was recently shown to allow for communication rates that scale with an increasing number of users, irrespective of how many users are in the network. By properly caching a carefully selected coded sequence of `popular' sub-files, the technique allows for coded multi-casting opportunities, even if the users ask for entirely different content. This has the potential to entirely change the way caching is performed, and for this reason the tutorial will highlight some of the powerful coded-caching ingredients that can be applied in different settings, but will also focus on crucial limitations of coded caching, which may potentially be resolved if one takes a joint network-theoretic and information-theoretic approach.

Part of what the tutorial will discuss is the possible new directions that can come from realizing that the problem of coded caching (MAC) and of communication (PHY) are non-separable, in the sense that you cannot just concatenate a powerful caching algorithm, to a powerful communication algorithm, and expect to get the best out of both. We will also discuss the different new directions that are inspired by the transition from the wired to the wireless medium; a transition that seems to entirely change the core of how caching must be performed. Understanding some of the core links between caching and wireless communications, can benefit network theorists which currently hold a somewhat divergent vision from information-theory in how they treat or exploit interference. In the end, this tutorial aims to discuss the key preliminary efforts towards evolving caching, from a tool that changes the volume of the problem (reflecting the old saying `do something today so that you do not have to do it tomorrow'), into a much more powerful ingredient that in fact changes the informational structure of the network.

Part 1
Introduction

Challenges

- Security and user privacy
- Protocol availability
- Time-varying popularity
- Memory availability

Popularity estimation in deep caching

- Novel popularity models with temporal and spatal correlations
- Popularity estimation
- LRU-techniques
- Prefetching
- Economical analysis of deep caching

Promising caching techniques for limited memory

Partial caching
Collaborative caching (aka femtocaching)

Part 2

- Simple description of coded caching.
- Insight on (single-stream) coded caching (Italian table example - first part).
- Measures of performance
  - Motivate these measures (delay, throughput, DoF - power consumption)
  - Slow build up to how delay increases and throughput decreases when we add more users.
  - basic examples of coded caching in single stream case.
- Gains over traditional caching

  - Natural multicasting vs coded multicasting (clearing misconceptions)
  - Coded multicasting with distinct file popularities (uncoded caching cannot scale and cannot match any coded version)
- Limitations of coded caching (can be perceived as open problems)
  - The exponential refinement problem (exponential caching problem - possible alleviations)
  - Synchronization
  - Topology (weakest guy brings down everyone, unless …).
  - Linear barrier of coded caching (microscopic gains for modest-sized caches).

Wireless coded caching

- Basics of wireless channel (start simple - BC)
- Basics on feedback (in a nutshell)
  - Range of performance in wireless communications
- Revisit coded-caching insight, now in wireless communications (Italian table example - second part).

Fundamental differences between wired and wires coded caching – non separability

- Interesting (competing + synergistic) duality between feedback and memory (caching)
  - Simple example on need to alter caching and transmission when going from wired to wireless (Allerton)
- Handling interference: give example of topology
- Centralized multi-server problem: ripping the gains of both words
- Structural connections
- Exponential performance gains in performance (addressing the small size concern)
- Caching as a means to simplify communications (reducing feedback and network size)
- Caching at both transmitters and receivers (Interference channel)

- Outer bounds – very basic exposition of their usefulness and of the bounding techniques

Future directions.

- Most settings still suffer from linear barrier
- Multi-server problem with different data.
- Combining with other network resources (complex but necessary task)
- Cache updates - what is the load for updating caches
- Combining transmit and receive caches
- Complexity (many problems have extremely high comp complexity)

<u>CVs</u>

**Petros Elia** received the B.Sc. degree from the Illinois Institute of Technology, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Southern California (USC), Los Angeles, in 2001 and 2006 respectively. Since February 2008 he has been an Associate Professor with the Department of Mobile Communications at EURECOM in Sophia Antipolis, France. His latest research deals with the intersection of coded caching and feedback-aided communications in multiuser settings. He has also considered different problems in the area of complexity-constrained communications, MIMO, cooperative and multiple access protocols and transceivers, complexity of communication, as well as with isolation and connectivity in dense networks, queueing theory and cross-layer design, coding theory, information theoretic limits in cooperative communications, and surveillance networks. He is a Fulbright scholar, the co-recipient of the SPAWC-2011 best student paper award on the topic of reduced complexity bidirectional communication with limited feedback, and of the NEWCOM++ distinguished achievement award 2008-2011 for a sequence of publications on the topic of reduced complexity multimode communications in the presence of little or no feedback.

**Georgios Paschos** is a principal researcher at Huawei Technologies, Paris, France, leading the Network Control and Resource Allocation team since Nov 2014. Previously, he conducted his research at MIT in the team of Prof. Eytan Modiano. For the period June 2008-Nov 2014 he was affiliated with “The Center of Research and Technology Hellas - Informatics & Telematics Institute”CERTH-ITI, Greece, working with Prof. Leandros Tassiulas. He also taught in the University of Thessaly, Dept. of Electrical and Computer Engineering as an adjunct lecturer for the period 2009-2011. In 2007-2008 he was an ERCIM Postdoc Fellow in VTT, Finland, working on the team of Prof. Norros. He received his diploma in Electrical and Computer Engineering (2002) from Aristotle University of Thessaloniki, and his PhD degree in Wireless Networks (2006) from ECE dept. University of Patras (supervisor Prof. Stavros Kotsopoulos), both in Greece. Two of his papers won the best paper award, in GLOBECOM 07’ and IFIP Wireless Days 09’ respectively. He serves as an associate editor for IEEE/ACM Trans. on Networking, and as a TPC member of IEEE INFOCOM. See also http://gpasxos.pagesperso-orange.fr/