Cache Optimization Models and Algorithms

Other titles in Foundations and Trends® in Communications and Information Theory

Group Testing: An Information Theory Perspective
Matthew Aldridge, Oliver Johnson and Jonathan Scarlett
ISBN: 978-1-68083-596-0

Sparse Regression Codes
Ramji Venkataramanan, Sekhar Tatikonda and Andrew Barron
ISBN: 978-1-68083-580-9

Fundamentals of Index Coding
Fatemeh Arbabjolfaei and Young-Han Kim
ISBN: 978-1-68083-492-5

Community Detection and Stochastic Block Models
Emmanuel Abbe
ISBN: 978-1-68083-476-5

Cache Optimization Models and Algorithms

Georgios S. Paschos
Amazon.com
gpasxos@gmail.com

George Iosifidis
Trinity College Dublin
george.iosifidis@tcd.ie

Giuseppe Caire
TU Berlin
caire@tu-berlin.de

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

G. Paschos, G. Iosifidis and G. Caire. *Cache Optimization Models and Algorithms*. Foundations and Trends® in Communications and Information Theory, vol. 16, no. 3–4, pp. 156–345, 2020.

ISBN: 978-1-68083-703-2

Foundations and Trends® in Communications and Information Theory
Volume 16, Issue 3–4, 2020

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2020, Volume 16, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

Contents

1 Introduction **2**
1.1 Historical Background and Scope 2
1.2 The Content Delivery Network 4
1.3 Wireless Caching and Beyond 7
1.4 Structure 8

2 Content Popularity **11**
2.1 Introduction to Caching-Related Terms 11
2.2 Power Law Popularity 13
2.3 Request Sequences 25
2.4 Discussion 36

3 Cache Eviction Policies **37**
3.1 Performance Under Arbitrary Requests 37
3.2 Performance Under Stationary Requests 49
3.3 Online Popularity Learning 59
3.4 Discussion of Related Work 75

4 Caching Networks **78**
4.1 Model and Optimization Variables 78
4.2 Deployment of Caching Networks 83
4.3 Bipartite Caching Networks 90

4.4 Hierarchical Caching Networks 109
4.5 General Caching Networks 120

5 Online Bipartite Caching **131**
5.1 Background and Model 132
5.2 Problem Statement 134
5.3 Bipartite Supergradient Caching Algorithm 139
5.4 Extensions and Numerical Evaluation 145
5.5 Discussion of Related Work 149

6 Asymptotic Laws for Caching Networks **151**
6.1 Analysis of Large Caching Networks 152
6.2 Discussion of Related Work 162

Acknowledgments **164**

References **165**

Cache Optimization Models and Algorithms

Georgios Paschos[1], George Iosifidis[2] and Giuseppe Caire[3]

[1] *Amazon.com; gpasxos@gmail.com*
[2] *Trinity College Dublin; george.iosifidis@tcd.ie*
[3] *TU Berlin; caire@tub.edu*

ABSTRACT

Caching refers to the act of replicating information at a faster (or closer) medium with the purpose of improving performance. This deceptively simple idea has given rise to some of the hardest optimization problems in the fields of computer systems, networking, and the Internet, many of which remain unsolved several years after their conception. While a wealth of research contributions exists from the topics of memory systems, data centers, Internet traffic, CDNs, and recently wireless networks, the literature is dispersed and overlapping at times. In this monograph, we take a unifying modeling view: by focusing on the fundamental underlying mathematical models, we re-organize the available material into a powerful framework for performing optimization of caching systems. This way, we aspire to present a solid background for the anticipated explosion in caching research, but also provide a didactic view into how engineers have managed to infuse mathematical models into the study of caching over the last 40 years.

Georgios Paschos, George Iosifidis and Giuseppe Caire (2020), "Cache Optimization Models and Algorithms", Foundations and Trends® in Communications and Information Theory: Vol. 16, No. 3–4, pp 156–345. DOI: 10.1561/0100000104.

1

Introduction

Storage resources and caching techniques permeate almost every area of communication networks today. In the near future, caching is set to play an important role in storage-assisted Internet architectures, information-centric networks and wireless systems, reducing operating and capital expenditures and improving the offered services. In light of the remarkable data traffic growth and the increasing number of rich-media applications, the impact of caching is expected to become even more profound than it is today. Therefore, it is crucial to design these systems in an optimal fashion, ensuring the maximum possible performance and economic benefits from their deployment. To that end, this monograph presents a collection of detailed models and algorithms, which are synthesized to build a powerful analytical framework for caching optimization.

1.1 Historical Background and Scope

The term *cache* was introduced in computer systems in 1970s to describe a memory with very fast access but typically small capacity. In computer applications, memory access often exhibits locality, i.e., most requests are related to memory blocks in a specific area known as *hot spot*.

By replicating these spots on a cache, it is possible to accelerate the performance of the entire memory system. One of the most important first problems in this context was to select which memory blocks to replicate in order to maximize the expected benefits; and several key results, e.g., the oracle MIN policy [26], were developed in that early era of computer systems. Nevertheless, the design of such *caching policies* remains one of the main challenges in caching systems.

The above caching idea was later applied to the Internet. As the population of users was growing fast in 1990s, the client-server connection model became impractical since all content requests (for web pages, in particular) were routed to few central servers. This was creating server and network congestion, and motivated the idea of using *Internet caches*. The latter are deployed closer to end users and host carefully selected web pages. Given the content popularity skewness, i.e., the fact that few web pages attract the majority of requests, even small caches can bring impressive performance benefits. Indeed, it soon became clear that *web caching* can significantly alleviate network congestion and improve the content access time for users. In these interconnected caches the problem of designing optimal caching policies is more intricate, as it requires to decide which files to cache, how to route the content and how to dimension each cache.

The last few years we witness a resurgence of interest in caching in the domain of wireless networks. The expansive growth of mobile video traffic in conjunction with exciting developments — like the use of *coding techniques* — have placed caching at the forefront of research in wireless communications. There is solid theoretical and practical evidence that memory can be a game-changer in our efforts to increase the effective throughput, and there are suggestions for deploying caches at the network core, the base stations or even at the mobile devices. Similarly, novel services that involve in-network computations, require pre-stored information (e.g., machine learning services), or are bounded by low latency constraints, can greatly benefit from caching. In fact, many caching enthusiasts argue that such services can only be deployed if they are supported by intelligent caching techniques.

Amidst these developments, it is more important than ever to model, analyze and optimize the performance of caching systems. Quite

surprisingly, many existing caching solutions, albeit practical, have not been designed using rigorous mathematical tools. Hence, the question of whether they perform optimally remains open. At the same time, the caching literature spans more than 40 years, different systems and even different research communities, and there is lack of a much-needed unified view on caching models and algorithms. This monograph aspires to fill this gap by presenting the theoretical foundations of caching and the latest conceptual and mathematical advances in this area. It provides detailed technical arguments and proofs, aiming to create a stable link between the past and future of caching analysis, and offer a useful starting point for new researchers. In the remainder of this section, we set the ground by discussing key caching systems and ideas, and explain the organization of this monograph.

1.2 The Content Delivery Network

A milestone in the evolution of caching systems was the deployment of *Content Delivery Networks*. CDNs typically consist of: (i) the origin server; (ii) the dispersed caches; (iii) the backbone network; and (iv) the points of ingress user traffic. The origin server is often deployed at a remote location with enormous storage capabilities (e.g., a datacenter) and stores all content files a user of this service might request, i.e., the entire content *catalog*. The caches are smaller servers deployed near the demand points, and are connected with the backbone network. Before CDNs, the users would establish TCP connections with the origin server in order to retrieve the content. In CDNs however, these connections are redirected to caches which serve the requests for these locally replicated files. Modern CDN systems have further evolved to optimize network traffic, offer different levels of Quality of Service (QoS), and increase the robustness of these services by, for example, protecting them from Denial of Service (DoS) attacks. In this monograph we focus on the aspect of content caching, and specifically on the intelligence involved in orchestrating the caching operations.

An iconic CDN system is the "Akamai Intelligent Platform", see [181] for a detailed description. Akamai was one of the most prominent CDN providers in the booming Internet era of 2000s, and by 2019 was

responsible for delivering 20% approximately of the Internet traffic. Its 216 K caching servers are dispersed at network edges offering low-latency content access around the globe. The Akamai model was designed to intercept HTTP traffic using a DNS redirect: when a user wants to open a website with the HTTP protocol, it would first contact the local DNS server to retrieve the IP of the origin server. The intelligent platform replaces the DNS entry with the IP of an Akamai cache containing the requested content, and hence the HTTP request is eventually served by that cache. The intelligent operations are handled by the mapping system, which decides where to cache each content file, and accordingly maps DNS entries to caches. Although the mapping system is effectively deciding the placement of content, the local caches are also operated with reactive policies such as the famous LRU and its variants.

Benefits of Caching

The replication of few popular contents can significantly *reduce the traffic* at the backbone network. When a content is available at a nearby cache, the respective requests are redirected to that cache instead of the origin server (an event called *hit*). Therefore, caches are often scattered around the network to minimize the geodesic distance, or network hops, from the users. Previous research has investigated solutions for the optimal placement of servers, e.g., [201], and the sizing of cache storage, called cache dimensioning [135]. Since more hits mean less network traffic, an important criterion for deploying caches is the increase of the cache *hit ratio*. Pertinent optimization problems in this context include the choice of eviction policies, i.e., the dynamic selection of the contents that are evicted from an overflowing cache, and the strategic content placement for enabling cache collaboration [24].

Another benefit of caching is *latency reduction*, i.e., the decrease of elapsed time between the initiation of a request and content delivery. Typically, the latency improvement is attributed to cutting down propagation latency. Packets traversing a transcontinental link, for example, experience latency up to 250 msec due to speed-of-light limitations [218]. Given that each TCP connection involves the exchange of several messages, it might actually take seconds before a requested content is

delivered over such links. These large latencies are very harmful for e-commerce and other real-time applications, and their improvement has been one of the main market-entry advantages of CDNs. Indeed, when users retrieve contents from a nearby cache, the geodesic distance is greatly reduced and so is the propagation time that hinders the content delivery. Nevertheless, latency optimization in caching systems is an intricate task and there are some notable misconceptions.

Firstly, in most applications latency effects smaller than 30 msec do not impact the user experience. Hence, one needs to be cautious in increasing the infrastructure costs in order to deliver content faster than this threshold. In other words, when it comes to latency, a single local cache often suffices to serve a large metropolitan area. Secondly, regarding video content delivery, the latency requirements apply only to the first video chunks and not to the entire file. Delivering fast the first chunks and then exploiting the buffer capacity at the user side is often adequate for ensuring smooth reproduction, even if the later video segments are delivered with higher latency. Finally, several low latency applications, such as reactive virtual reality, vehicular control, or industrial automation, cannot typically benefit from caching since their traffic is not reusable. However, we stress that there are scenarios where one can exploit caching (e.g., using proactive caching policies) in order to boost the performance of such demanding services.

Another important effect of web caching is that it *balances the server load*. For example, the Facebook Photo CDN leverages web browser caches on user devices, edge regional servers, and other caches in order to reduce the traffic reaching the origin servers. Notably, browser caches serve almost 60% of traffic requests, due to the fact that users view the same content multiple times. Edge caches serve 20% of the traffic (i.e., approximately 50% of traffic not served by browser caches), and hence offer important off-network bandwidth savings by serving locally the user sessions. The remaining 20% of content requests are served at the origin, using a combination of slow back-end storage and a fast origin-cache [116]. This CDN functionality shields the main servers from high load and increases the scalability of the architecture. Note that the server load is minimized when the cache hits are maximized, and hence the problem of server load minimization is equivalent to

cache hit maximization. Therefore, in the remaining of this monograph we will focus on hit maximization, as well as bandwidth and latency minimization.

1.3 Wireless Caching and Beyond

Caching has also been considered for improving content delivery in wireless networks, see [191] and references therein. There is consensus that network capacity enhancements by means of improving physical layer access rates or through denser deployment of base stations is a costly approach and outpaced by the fast-increasing data traffic [60]. Caching techniques promise to fill this gap, and several interesting ideas have been suggested to this end: (i) deep caching at the evolved packet core (EPC) in order to reduce content delivery delay [242]; (ii) caching at the base stations to alleviate congestion in their throughput-limited backhaul links [103]; (iii) caching at the mobile devices to leverage device-to-device communications [101]; and (iv) coded caching for accelerating transmissions over a broadcast medium [163].

Furthermore, techniques that combine caching with coding demonstrate revolutionary *goodput* scaling in bandwidth-limited cache-aided networks. This has motivated researchers to revisit the fundamental question of how memory "interacts" with other resources. The topic of *coded caching* started as a powerful tool for broadcast mediums, and led towards establishing an information theory for memory. Similarly, an interesting connection between memory and processing has been identified [154], creating fresh opportunities for improving the performance of distributed and parallel computing systems. These lines of research have re-stirred the interest in joint consideration of bandwidth, processing and memory resources.

At the same time, the advent of technologies such as Software-Defined Networking (SDN) and Network Function Virtualization (NFV) create new opportunities for leveraging caching in wired and wireless networks. Namely, they enable the fine-grained and unified control of storage, computing and network bandwidth, and support the flexible deployment of in-network caching services. This gives rise to the new concept of content-centric network architectures that aim to use storage

and caching as a means to rethink the Internet operation. Similarly, new business models are emerging as new players are entering the content delivery market. Service providers like Facebook are acquiring their own CDNs, network operators deploy in-network cache servers to reduce their bandwidth expenditures, and content providers like Google, Netflix, and Amazon use caches to replicate their content world-wide. Interestingly, smaller content providers can buy caching resources on the cloud market to instantiate their service *just in time and space*. These novel concepts create, unavoidably, new research questions for caching architectures and the caching economic ecosystem, and one of our goals is to provide the fundamental underlying mathematical theories that can support research in these exciting directions.

1.4 Structure

In this subsection we provide a quick summary of the monograph, serving both as a warm-up for reading it, as well as a map with directions to specific information.

We begin in Section 2 with a detailed treatment of content popularity models, a crucial factor shaping the performance of caching policies. We first explain the power-law popularity model and how we can infer its parameters from a dataset, and then use it for the purpose of cache dimensioning. We then define the *Independence Reference Model* (IRM) for describing a request generation process. IRM is a widely used model for caching analysis because of its tractability, but it has limited accuracy since it fails to capture two observed correlation effects of request processes, namely *temporal and spatial locality*. We discuss state-of-the-art mathematical models which are more accurate than IRM in that respect, but also more difficult to analyze. For the case of temporal correlations, we provide the optimal rule for popular/unpopular content classification that maximizes cache performance.

In Section 3 we explore the class of *online eviction policies*, also known as replacement policies. A cache receives requests and a rule must be employed for evicting a content when the cache overflows. The design of such policies is an equally challenging and important problem, and we present the main pertinent results. We begin with the case

of arbitrary requests, for which an oracle policy, known as "Belady", achieves the maximum number of hits under any request sequence. This policy requires knowledge of future requests, and therefore is useful only as a benchmark. Using the Belady policy we prove that the "Least Recently Used" (LRU) rule provides the best competitive performance among all online policies, i.e., those that do not have information about future requests. Then, for stationary IRM requests the "Least Frequently Used" (LFU) rule is the optimal, as it estimates the (assumed static) content popularity using the observed frequencies. We also study the *characteristic time* approximation, with which we obtain the performance of LRU for stationary requests, as well that of *Time To Live* (TTL) caches which allow to optimally tune different content hit probabilities. Last, we depart from the stationary assumption and take a model-free approach inspired by the Machine Learning framework of *Online Convex Optimization* (OCO). We present an adaptation of the Zinkevich's online gradient policy to the caching problem, and show that it achieves the optimal *regret*, i.e., the smallest possible losses with respect to the best static cache configuration with future knowledge.

In Section 4 we study caching networks (CNs), i.e., systems where multiple caches are interconnected via a network. We focus on *proactive caching* policies which populate the caches based on estimated demand. In CNs, the designer needs to decide where to cache each content file (caching policy), how to route the contents from caches to the requesters (routing policy), and also the capacity of the caches and network links. Therefore, we start the section by explaining this general *CN design and management problem*. This is a notoriously hard problem, that cannot be solved optimally for large CNs and content catalogs. To gain a better understanding into the available solution methodologies, we survey a number of important subproblems: (i) the *cache dimensioning* problem where we decide where to place storage in the network; (ii) the *content caching* in bipartite and tree graphs; and (iii) the *joint content caching and routing* problem in general graphs. Although these are all special cases of the general CN problem, they are governed by significantly different mathematical theories. Therefore, our exposition in this section serves to clarify where each mathematical theory applies best, and how to get a good approximate guarantee for each scenario.

In the following Section 5 we take an approach that combines the two previous sections. We study a CN where the content popularity is unknown, and therefore the goal is to design a joint caching and routing policy which at the same time learns the content popularity, decides in an online manner which files to cache, and in a reactive fashion how to route the contents to requesters. These results generalize the OCO-based policy that was introduced in Section 3 for a single cache. We show that a policy that takes a step in the direction of a subgradient of the previous slot's utility function can provide "no regret" in the CN scenario as well. This means that the designed policy gradually learns to match the performance of the optimal static policy, a theoretical result that is validated through trace-driven numerical experiments. The analysis in this section is general enough to account for changes in the network structure and cache reconfiguration costs.

In the last Section 6 we examine a very large (scaling) network of caches, arranged in a square grid. For this special case, we show that it is possible to relax the original combinatorial problem and obtain a relaxed solution via convex optimization. Then it can be shown that the relaxed solution is of the same order of performance with the actual integral, hence we can use it to understand the scaling performance of a large caching network. The section includes detailed results about the sustainability of networks aiming to deliver content in different regimes of: (i) network size; (ii) catalog size; and (iii) cache size.

References

[1] Abedini, N. and S. Shakkottai (2014). "Content caching and scheduling in wireless networks with elastic and inelastic traffic". *IEEE/ACM Transactions on Networking*. 22(3): 864–874.

[2] Abernethy, J., P. L. Bartlett, A. Rakhlin, and A. Tewari (2008). "Optimal strategies and minimax lower bounds for online convex games". In: *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory (COLT)*. 1–9.

[3] Adamic, L. A. and B. A. Huberman (2002). "Zipf's law and the internet". *Glottometrics*. 3: 143–150.

[4] Ageev, A. A. and M. I. Sviridenko (2004). "Pipage rounding: A new method of constructing algorithms with proven performance guarantee". *Journal of Combinatorial Optimization*. 8(3): 307–328.

[5] Alfano, G., M. Garetto, and E. Leonardi (2014). "Content-centric wireless networks with limited buffers: When mobility hurts". *IEEE/ACM Transactions on Networking*. 24(1): 299–311.

[6] Almeida, V., A. Bestavros, M. Crovella, and A. de Oliveira (1996). "Characterizing reference locality in the WWW". In: *Proceedings of the IEEE Fourth International Conference on Parallel and Distributed Information Systems (PDIS)*. 92–103.

[7] Applegate, D., A. Archer, V. Gopalakrishnan, S. Lee, and K. Ramakrishnan (2016). "Optimal content placement for a large-scale VoD system". *IEEE/ACM Transactions on Networking*. 24(4): 2114–2127.

[8] Arlitt, M. and T. Jin (2000). "A workload characterization study of the 1998 world cup web site". *IEEE Network*. 14(3): 30–37.

[9] Arya, V., N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit (2004). "Local search heuristics for k-median and facility location problems". *SIAM Journal on Computing*. 33(3): 544–562.

[10] Asur, S. and B. A. Huberman (2010). "Predicting the future with social media". In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 492–499.

[11] Athanasiou, G., P. C. Weeraddana, C. Fischione, and L. Tassiulas (2015). "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks". *IEEE/ACM Transactions on Networking*. 23(3): 836–850.

[12] Aven, O. I., E. G. Coffman, and Y. A. Kogan (1987). *Stochastic Analysis of Computer Storage*. Vol. 38. Springer.

[13] Avrachenkov, K., J. Goseling, and B. Serbetci (2017). "A low-complexity approach to distributed cooperative caching with geographic constraints". *Proceedings of the ACM on Measurement and Analysis of Computing Systems*. 1(1): Article 27.

[14] Azimdoost, B., C. Westphal, and H. R. Sadjadpour (2016). "Fundamental limits on throughput capacity in information-centric networks". *IEEE Transactions on Communications*. 64(12): 5037–5049.

[15] Baek, S. K., S. Bernhardsson, and P. Minnhagen (2011). "Zipf's law unzipped". *Tech. rep.* arXiv:1104.1789.

[16] Baev, I. D., R. Rajaraman, and C. Swamy (2008). "Approximation algorithms for data placement problems". *SIAM J. Computing*. 38(4): 1411–1429.

[17] Bansal, S. and D. S. Modha (2004). "CAR: Clock with adaptive replacement". In: *Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST)*. 187–200.

[18] Baştuğ, E., M. Bennis, E. Zeydan, M. Kader, I. Karatepe, A. Er, and M. Debbah (2015a). "Big data meets telcos: A proactive caching perspective". *Journal of Communications and Networks*. 17(6): 549–557.

[19] Baştuğ, E., M. Bennis, and M. Debbah (2014). "Living on the edge: The role of proactive caching in 5G wireless networks". *IEEE Communications Magazine*. 52(8): 82–89.

[20] Baştuğ, E., M. Bennis, and M. Debbah (2015b). "A transfer learning approach for cache-enabled wireless networks". In: *13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE. 161–166.

[21] Bateni, M. and M. Hajiaghayi (2012). "Assignment problem in content distribution networks: Unsplittable hard-capacitated facility location". *ACM Transactions on Algorithms*. 8(3): 20:1–20:19.

[22] Bauke, H. (2007). "Parameter estimation for power-law distributions by maximum likelihood methods". *The European Physical Journal B*. 58(2): 167–173.

[23] Beck, A. and M. Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". *Operations Research Letters*. 31: 167–175.

[24] Bektas, T., O. Oguz, and I. Ouveysi (2007). "Designing cost-effective content distribution networks". *Computers & Operations Research*. 34(8): 2436–2449.

[25] Bektas, T., J.-F. Cordeau, E. Erkut, and G. Laporte (2008). "Exact algorithms for the joint object placement and request routing problem in content distribution networks". *Computers & Operations Research*. 35(12): 3860–3884.

[26] Belady, L. A. (1966). "A study of replacement algorithms for virtual storage computers". *IBM Systems Journal*. 5(2): 78–101.

[27] Belmega, E., P. Mertikopoulos, R. Negrel, and L. Sanguinetti (2018). "Online convex optimization and no-regret learning: Algorithms, guarantees and applications". *Tech. rep.* arXiv:1804.04529.

[28] Benders, J. F. (1962). "Partitioning procedures for solving mixed-variables programming problems". *Numerische Mathematik*. 4: 238–252.

[29] Bennett, J. and S. Lanning (2007). "The netflix prize". In: *Proceedings of the KDD Cup and Workshop*.

[30] Benoit, A., V. Rehn-Sonigo, and Y. Robert (2008). "Replica placement and access policies in tree networks". *IEEE Transactions on Parallel and Distributed Systems*. 19(12): 1614–1627.

[31] Berthet, C. (2016). "Identity of King and Flajolet & al. formulae for LRU miss rate exact computation". *Technical Report*. arXiv: 1607.01283.

[32] Bertsekas, D. P. (1998). *Network Optimization: Continuous and Discrete Models*. Athena Scientific.

[33] Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

[34] Bertsekas, D. P. and N. R. Sandel (1982). "Estimates of the duality gap for large-scale separable non-convex optimization problems". In: *Proceedings of the 21st IEEE Conference on Decision and Control (CDC)*. 782–785.

[35] Bertsimas, D. and J. N. Tsitsiklis (1997). *Introduction to Linear Optimization*. Athena Scientific.

[36] Bhattacharjee, R., S. Banerjee, and A. Sinha (2020). "Fundamental limits on the regret of online network-caching". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)*. 25:1–25:31.

[37] Bielski, M., I. Syrigos, K. Katrinis, D. Syrivelis, A. Reale, D. Theodoropoulos, N. Alachiotis, D. Pnevmatikatos, E. H. Pap, G. Zervas, V. Mishra, A. Saljoghei, A. Rigo, J. Fernando Zazo, S. Lopez-Buedo, M. Torrents, F. Zyulkyarov, M. Enrico, and O. Gonzalez de Dios (2018). "dReDBox: Materializing a full-stack rack-scale system prototype of a next-generation disaggregated datacenter". In: *Proceedings of the IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 1093–1098.

[38] Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley & Sons.

[39] Blasco, P. and D. Gunduz (2014). "Multi-armed bandit optimization of cache content in wireless infostation networks". In: *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. 51–55.

[40] Blaszczyszyn, B. and A. Giovanidis (2015). "Optimal geographic caching in cellular networks". In: *Proceedings of the IEEE International Conference on Communications (ICC)*. 1–6.

[41] Borst, S., V. Gupta, and A. Walid (2010). "Distributed caching algorithms for content distribution networks". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 1–9.

[42] Box, G. E. (1976). "Science and statistics". *Journal of the American Statistical Association*. 71(356): 791–799.

[43] Breslau, L., P. Cue, P. Cao, L. Fan, G. Phillips, and S. Shenker (1999). "Web caching and Zipf-like distributions: Evidence and implications". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 126–134.

[44] Calinescu, G., C. Chekuri, M. Pal, and J. Vondrak (2007). "Maximizing a submodular set function subject to a matroid constraint". M. Fischetti and D. P. Williamson (eds.) *Integer Programming and Combinatorial Optimization, LNCS*. (4513): 182–196.

[45] Cao, P. and S. Irani (1997). "Cost-aware WWW proxy caching algorithms". In: *Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS)*. 1–14.

[46] Cardei, M. and D.-Z. Du (2005). "Improving wireless sensor network lifetime through power aware organization". *Wireless Networks*. 11(3): 333–340.

[47] Carofiglio, G., L. Mekinda, and L. Muscariello (2016). "Joint forwarding and caching with latency awareness in information-centric networking". *Computer Networks*. 110: 133–153.

[48] Carr, R. W. and J. L. Hennessy (1981). "A simple and effective algorithm for virtual memory management". *SIGOPS Operating Systems Review*. 15(5): 87–95.

[49] Challenger, J. R., P. Dantzig, A. Iyengar, M. S. Squillante, and L. Zhang (2004). "Efficiently serving dynamic data at highly accessed web sites". *IEEE/ACM Transactions on Networking*. 12(2): 233–246.

[50] Chankhunthod, A., P. B. Danzig, C. Neerdaels, M. F. Schwartz, and K. J. Worrell (1996). "A hierarchical internet object cache". *Proceedings of the USENIX Technical Conference*. 4(11): 153–163.

[51] Charikar, M. and S. Guha (1999). "Improved combinatorial algorithms for the facility location and k-median problems". In: *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS)*. 378–388.

[52] Charikar, M., S. Guha, E. Tardos, and D. B. Shmoys (2002). "A constant-factor approximation algorithm for the k-median problem". *Journal of Computer and System Sciences*. 65(1): 129–149.

[53] Chater, N. and G. D. Brown (1999). "Scale-invariance as a unifying psychological principle". *Cognition*. 69(3): B17–B24.

[54] Chatzieleftheriou, L. E., M. Karaliopoulos, and I. Koutsopoulos (2017). "Caching-aware recommendations: Nudging user preferences towards better caching performance". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 1–9.

[55] Che, H., Y. Tung, and Z. Wang (2002). "Hierarchical web caching systems: Modeling, design and experimental results". *IEEE Journal on Selected Areas in Communications*. 20(7): 1305–1314.

[56] Cheng, X., C. Dale, and J. Liu (2008). "Statistics and social network of YouTube videos". In: *Proceedings of the IEEE International Workshop on Quality of Service (IWQoS)*. 229–238.

[57] Cheng, X., J. Liu, and C. Dale (2013). "Understanding the characteristics of internet short video sharing: A YouTube-based measurement study". *IEEE Transactions on Multimedia*. 15(5): 1184–1194.

[58] Chiang, M., S. H. Low, R. Calderbank, and J. C. Doyle (2007). "Layering as optimization decomposition: A mathematical theory of network architectures". *Proceedings of the IEEE*. 95(1): 255–312.

[59] Chu, J., K. Labonte, and B. N. Levine (2002). "Availability and popularity measurements of peer-to-peer file systems". In: *Proceedings of SPIE 4868, Scalability and Traffic Control in IP Networks II.*

[60] Cisco (n.d.). "Annual Internet Report (2018–2023)". *Cisco White Paper,* Updated: March 9, 2020.

[61] Clark, C. E. (1961). "The greatest of a finite set of random variables". *Operations Research.* 9(2): 145–162.

[62] Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). "Power-law distributions in empirical data". *SIAM Review.* 51(4): 661–703.

[63] Consuegra, M. E., W. A. Martinez, G. Narasimhan, R. Rangaswami, L. Shao, and G. Vietri (2017). "Analyzing adaptive cache replacement strategies". *arXiv:1503.07624v2.*

[64] Cronin, E., S. Jamin, C. Jin, A. R. Kurc, D. Raz, and Y. Shavitt (2002). "Constrained mirror placement on the internet". *IEEE Journal on Selected Areas in Communications.* 20(7): 1369–1382.

[65] Cui, Y. and D. Jiang (2017). "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks". *IEEE Transactions on Wireless Communications.* 16(1): 250–264.

[66] Cunha, C. R., A. Bestavros, and M. E. Crovella (1995). "Characteristics of WWW client-based traces". Technical Report Nr. TR-95-010, Boston University, MA.

[67] Dai, J., Z. Hu, B. Li, J. Liu, and B. Li (2012). "Collaborative hierarchical caching with dynamic request routing for massive content distribution". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM).* 1158–1169.

[68] Dan, A. and D. Towsley (1990). "An approximate analysis of the LRU and FIFO buffer replacement schemes". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS).* 143–152.

[69] Dán, G. and N. Carlsson (2010). "Power-law revisited: A large scale measurement study of P2P content popularity". In: *Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS)*, San Jose, CA.

[70] Daskin, M. S. (2013). *Network and Discrete Location: Models, Algorithms and Applications*. Wiley.

[71] Dehghan, M., B. Jiang, A. Seetharam, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman (2017). "On the complexity of optimal request routing and content caching in heterogeneous cache networks". *IEEE/ACM Transactions on Networking*. 25(3): 1635–1648.

[72] Dehghan, M., L. Massoulie, D. Towsley, D. Menasche, and Y. C. Tay (2016). "A utility optimization approach to network cache design". In: *IEEE INFOCOM 2016 – The 35th Annual IEEE International Conference on Computer Communications*. IEEE. 1–9.

[73] Duchi, J. C., S. Shalev-Shwartz, Y. Singer, and T. Chandra (2008). "Efficient projections onto the l1-ball for learning in high dimensions". In: *Proceedings of the 25th International Conference on Machine Learning (ICML)*. 272–279.

[74] Eisenbrand, F., F. Grandoni, T. RothvoB, and G. Schafer (2008). "Approximating connected facility location problems via random facility sampling and core detouring". In: *Proceedings of the ACM-SIAM SODA*. ACM. 365–372.

[75] Elayoubi, S.-E. and J. Roberts (2015). "Performance and cost effectiveness of caching in mobile access networks". In: *Proceedings of the 2nd ACM Conference on Information-Centric Networking*. ACM. 79–88.

[76] Englert, M., H. Röglin, J. Spönemann, and B. Vöcking (2013). "Economical caching". *ACM Transactions on Computation Theory*: 4:1–4:21.

[77] Fagin, R. (1977). "Asymptotic miss ratios over independent references". *Journal of Computer and System Sciences*. 14(2): 222–250.

[78] Farmer, J. D. and J. Geanakoplos (2008). "Power laws in economics and elsewhere". *Tech. rep.* Santa Fe Institute.

[79] Figueiredo, F. (2013). "On the prediction of popularity of trends and hits for user generated videos". In: *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. ACM. 741–746.

[80] Fisher, M. L. (2004). "The Lagrangian relaxation method for solving integer programming problems". *Management Science*. 50(12): 1861–1871.

[81] Flajolet, P., D. Gardy, and L. Thimonier (1992). "Birthday paradox, coupon collectors, caching algorithms and self-organizing search". *Discrete Applied Mathematics*. 39(3): 207–229.

[82] Fleischer, L., M. Goemans, V. Mirrokni, and M. Sviridenko (2006). "Tight approximation algorithms for maximum general assignment problems". In: *Proceedings of the ACM SODA*. ACM.

[83] Fofack, N. C., P. Nain, G. Neglia, and D. Towsley (2012). "Analysis of TTL-based cache networks". In: *6th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*. IEEE. 1–10.

[84] Franceschetti, M. and R. Meester (2007). *Random Networks for Communication*. Cambridge Series in Statistical and Probabilistic Mathematics, 24. Cambridge University Press.

[85] Frank, B., I. Poese, G. Smaragdakis, A. Feldmann, B. Maggs, S. Uhlig, V. Aggarwal, and F. Schneider (2013). "Collaboration opportunities for content delivery and network infrastructures". *Recent Advances in Networking*. 1(1). ACM SIGCOMM.

[86] Fricker, C., P. Robert, and J. Roberts (2012a). "A versatile and accurate approximation for LRU cache performance". In: *Proceedings of the 24th International Teletraffic Congress. ITC '12*. IEEE. 8:1–8:8.

[87] Fricker, C., P. Robert, J. Roberts, and N. Sbihi (2012b). "Impact of traffic mix on caching performance in a content-centric network". In: *Proceedings on the 2012 IEEE INFOCOM Conference on Computer Communications Workshops*. IEEE. 310–315.

[88] Furno, A., M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda (2017). "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas". *IEEE Transactions on Mobile Computing*. 6(10): 2682–2696.

[89] Garetto, M., E. Leonardi, and V. Martina (2016). "A unified approach to the performance analysis of caching systems". *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*. 1(3): 12.

[90] Gast, N. and B. Van Houdt (2015). "Transient and steady-state regime of a family of list-based cache replacement algorithms". *ACM SIGMETRICS Performance Evaluation Review*. 43(1): 123–136.

[91] Geoffrion, A. M. (1974). "Lagrangian relaxation and its uses in integer programming". *Mathematical Studies*. 2: 82–114.

[92] Geulen, S., B. Vocking, and M. Winkler (2010). "Regret minimization for online buffering problems using the weighted majority algorithm". In: *Proceedings of the 33rd Annual Conference on Learning Theory (COLT)*. 1–12.

[93] Giannakas, T., P. Sermpezis, and T. Spyropoulos (2018). "Show me the cache: Optimizing cache-friendly recommendations for sequential content access". In: *Proceedings of the IEEE Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*. 1–9.

[94] Gill, P., M. Arlitt, Z. Li, and A. Mahanti (2007). "Youtube traffic characterization: A view from the edge". In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. ACM. 15–28.

[95] Giovanidis, A. and A. Avranas (2016). "Spatial multi-LRU: Distributed caching for wireless networks with coverage overlaps". In: *Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. 403–405.

[96] Gitzenis, S., G. S. Paschos, and L. Tassiulas (2012). "Asymptotic laws for content replication and delivery in wireless networks". In: *Proceedings of the 2012 IEEE INFOCOM Conference on Computer Communications*. IEEE. 126–134.

[97] Gitzenis, S., G. S. Paschos, and L. Tassiulas (2014a). "Enhancing wireless networks with caching: Asymptotic laws, sustainability & tradeoffs". *Computer Networks*. 64: 353–368.

[98] Gitzenis, S., G. Paschos, and L. Tassiulas (2013). "Asymptotic laws for joint content replication and delivery in wireless networks". *IEEE Transactions on Information Theory*. 59(5): 2760–2776.

[99] Gitzenis, S., S. Toumpis, and L. Tassiulas (2014b). "Efficient file replication in large wireless networks with dynamic popularity". In: *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*. 164–168.

[100] Goldstein, M. L., S. A. Morris, and G. G. Yen (2004). "Problems with fitting to the power-law distribution". *The European Physical Journal B – Condensed Matter and Complex Systems*. 41(2): 255–258.

[101] Golrezaei, N., A. Molisch, A. Dimakis, and G. Caire (2013a). "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution". *IEEE Communications Magazine*. 51(4): 142–149.

[102] Golrezaei, N., K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire (2013b). "FemtoCaching: Wireless content delivery through distributed caching helpers". *IEEE Transactions on Information Theory*. 59(12): 8402–8413.

[103] Golrezaei, N., K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire (2012). "Femtocaching: Wireless video content delivery through distributed caching helpers". In: *Proceedings of the 2012 IEEE INFOCOM Conference on Computer Communication*. IEEE. 1107–1115.

[104] Guha, S. and S. Khuller (1999). "Greedy strikes back: Improved facility location algorithms". *Journal of Algorithms*. 31: 228–248.

[105] Guha, S., A. Meyerson, and K. Munagala (2000). "Hierarchical placement and network design problems". In: *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*. 603–612.

[106] Gupta, A., J. Kleinberg, A. Kumar, R. Rastogi, and B. Yener (2001). "Provisioning a virtual private network: A network design problem for multicommodity flow". In: *Proceedings of the ACM STOC*. ACM. 389–398.

[107] Gupta, A., A. Kumar, and T. Roughgarden (2003). "Simple and better approximation algorithms for network design". In: *Proceedings of the ACM STOC*. ACM. 365–372.

[108] Gupta, P. and P. R. Kumar (2000). "The capacity of wireless networks". *IEEE Transactions on Information Theory*. 46(Mar.): 388–404.

[109] Hasslinger, G., K. Ntougias, F. Hasslinger, and O. Hohlfeld (2017). "Performance evaluation for new web caching strategies combining LRU with score based object selection". *Computer Networks*. 125: 172–186.

[110] Hazan, E. (2006). "Efficient algorithms for online convex optimization and their applications". *PhD Thesis*. Princeton University.

[111] Hazan, E. (2016). "Introduction to online convex optimization". *Foundations and Trends in Optimization*. 2(3–4): 157–325.

[112] Hefeeda, M. and O. Saleh (2008). "Traffic modeling and proportional partial caching for peer-to-peer systems". *IEEE/ACM Transactions on Networking (TON)*. 16(6): 1447–1460.

[113] Hilbert, M. (2012). "How much information is there in the 'information society'?" *Significance*. 9(4): 8–12.

[114] Hochbaum, D. S. and D. B. Shmoys (1985). "A best possible heuristic for the k-center problem". *Mathematics of Operations Research*. 10(2): 180–184.

[115] Hochbaum, D. S. and D. B. Shmoys (1986). "A unified approach to approximation algorithms for bottleneck problems". *Journal of the ACM*. 33(3): 533–550.

[116] Huang, Q., K. Birman, R. van Renesse, W. Lloyd, S. Kumar, and H. C. Li (2013). "An analysis of Facebook photo caching". In: *Proceedings of ACM Symposium on Operating Systems Principles. SOSP '13*. ACM. 167–181.

[117] Imbrenda, C., L. Muscariello, and D. Rossi (2014). "Analyzing cacheable traffic in ISP access networks for micro CDN applications via content-centric networking". In: *Proceedings of the 1st International Conference on Information-Centric Networking*. ACM. 57–66.

[118] Ioannidis, S. and E. Yeh (2016). "Adaptive caching networks with optimality guarantees". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)*. 77–87.

[119] Ioannidis, S. and E. Yeh (2017). "Jointly optimal routing and caching for arbitrary network topologies". In: *Proceedings of the IEEE ICN*. IEEE. 77–87.

[120] Ioannidis, S. and E. Yeh (2018). "Jointly optimal routing and caching for arbitrary network topologies". *IEEE Journal on Selected Areas in Communications*. 36(6): 1258–1275.

[121] Iosifidis, G., I. Koutsopoulos, and G. Smaragdakis (2017). "Distributed storage control algorithms for dynamic networks". *IEEE/ACM Transactions on Networking*. 25(3): 1359–1372.

[122] Jain, K., M. Mahdian, E. Markakis, A. Saberi, and V. Vazirani (2003). "Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP". *Journal of the ACM*. 50(6): 795–824.

[123] Jain, K. and V. Vazirani (2001). "Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation". *Journal of the ACM*. 48: 247–296.

[124] Jain, K., M. Mahdian, and A. Saberi (2002). "A new greedy approach for facility location problems". In: *Proceedings of the 29th ACM Symposium on Theory of Computing (STOC)*. 731–740.

[125] Jamin, S., C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang (2000). "On the placement of internet instrumentation". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 295–304.

[126] Jelenkovic, P. R. (1999). "Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities". *The Annals of Applied Probability*. 9(2): 430–464.

[127] Ji, M., G. Caire, and A. Molisch (2016a). "Wireless device-to-device caching networks: Basic principles and system performance". *IEEE Journal on Selected Areas in Communication*. 34(1): 176–189.

[128] Ji, M., G. Caire, and A. F. Molisch (2016b). "Fundamental limits of caching in wireless D2D networks". *IEEE Transactions on Information Theory*. 62(2): 849–869.

[129] Jia, X., D. Li, H. Du, and J. Cao (2005). "On optimal replication of data object at hierarchical and transparent web proxies". *IEEE Transactions on Parallel and Distributed Systems*. 16(8): 673–685.

[130] Jung, J., A. W. Berger, and H. Balakrishnan (2003). "Modeling TTL-based internet caches". In: *Proceedings of IEEE INFOCOM*. Vol. 1. IEEE. 417–426.

[131] Kang, X., H. Zhang, G. Jiang, H. Chen, X. Meng, and K. Yoshihira (2010). "Understanding internet video sharing site workload: A view from data center design". *Journal of Visual Communication and Image Representation*. 21(2): 129–138.

[132] Kangasharju, J., J. Roberts, and K. Ross (2002). "Object replication strategies in content distribution networks". *Computer Communications*. 25(4): 376–383.

[133] Kao, M.-Y. (2008). *Encyclopedia of Algorithms*. Springer.

[134] Kastanakis, S., P. Sermpezis, V. Kotronis, and X. Dimitropoulos (2018). "CABaRet: Leveraging recommendation systems for mobile edge caching". In: *Proceedings of the Workshop on Mobile Edge Communications (MECOMM)*. 19–24.

[135] Kelly, T. and D. Reeves (2000). "Optimal web cache sizing: Scalable methods for exact solutions". *Computer Communications*. 24(July): 163–173.

[136] King III, W. F. (1971). "Analysis of demand paging algorithms". In: *IFIP Congress*. North-Holland Publishing Company.

[137] Korupolu, M. R., C. G. Plaxton, and R. Rajaraman (2001). "Placement algorithms for hierarchical cooperative caching". *Journal of Algorithms*. 38(1): 260–302.

[138] Krashakov, S. A., A. B. Teslyuk, and L. N. Shchur (2006). "On the universality of rank distributions of website popularity". *Computer Networks*. 50(11): 1769–1780.

[139] Krishnan, P., D. Raz, and Y. Shavitt (2000). "The cache location problem". *IEEE/ACM Transactions on Networking*. 8: 568–582.

[140] Krolikowski, J., A. Giovanidis, and M. D. Renzo (2018). "A decomposition framework for optimal edge-cache leasing". *IEEE Journal on Selected Areas in Communications*. 6(36): 1345–1359.

[141] Laoutaris, N., M. Sirivianos, X. Yang, and P. Rodriguez (2011). "Inter-datacenter bulk transfers with Netstitcher". In: *ACM SIGCOMM*. ACM. 74–85.

[142] Laoutaris, N., G. Smaragdakis, K. Oikonomou, I. Stavrakakis, and A. Bestavros (2007). "Distributed placement of service facilities in large-scale networks". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 2144–2152.

[143] Laoutaris, N., V. Zissimopoulos, and I. Stavrakakis (2005). "On the optimization of storage capacity allocation for content distribution". *Computer Networks*. 47: 409–428.

[144] Leconte, M., G. S. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas (2016). "Placing dynamic content in caches with small population". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 1–9.

[145] Lee, M.-K., P. Michaud, J. S. Sim, and D. Nyang (2015). "A simple proof of optimality for the MIN cache replacement policy". *Information Processing Letters*. 116(2): 168–170.

[146] Leff, A., J. Wolf, and P. Yu (1993). "Replication algorithms in a remote caching architecture". *IEEE Transactions on Parallel and Distributed Systems*. 4(11): 1185–1204.

[147] Lelarge, M., L. Massoulié, and J. Xu (2013). "Reconstruction in the labeled stochastic block model". In: *IEEE Information Theory Workshop (ITW)*. 1–5.

[148] Leonardi, E. and G. Neglia (2018). "Implicit coordination of caches in small cell networks under unknown popularity profiles". *IEEE Journal on Selected Areas in Communications*. 36(6): 1276–1285.

[149] Levi, R., D. Shmoys, and C. Swamy (2012). "LP-based approximation algorithms for capacitated facility location". *Math. Program.* 131: 365–379.

[150] Li, B., M. Golin, G. Italiano, X. Deng, and K. Sohraby (1999). "On the optimal placement of web proxies in the internet". In: *IEEE INFOCOM*. IEEE. 1282–1290.

[151] Li, J., S. Shakkottai, J. C. Lui, and V. Subramanian (2018a). "Accurate learning or fast mixing? dynamic adaptability of caching algorithms". *IEEE Journal on Selected Areas in Communications*. 36(6): 1314–1330.

[152] Li, K., H. Shen, F. Y. L. Chin, and S. Q. Zheng (2005). "Optimal methods for coordinated enroute web caching for tree networks". *ACM Transactions on Internet Technology*. 5(3): 480–507.

[153] Li, S. (2013). "A 1.488 approximation algorithm for the uncapacitated facility location problem". *Information and Computation*. 31: 45–58.

[154] Li, S., M. A. Maddah-Ali, Q. Yu, and A. S. Avestimehr (2018b). "A fundamental tradeoff between computation and communication in distributed computing". *IEEE Transactions on Information Theory*. 64(1): 109–128.

[155] Li, S. and O. Svensson (2013). "Approximating k-median via pseudo-approximation". In: *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*. 901–910.

[156] Li, S., J. Xu, M. Schaar, and W. Li (2016). "Trend-aware video caching through online learning". *IEEE Transactions on Multimedia*. 18(12): 2503–2516.

[157] Li, W., E. Chan, G. Feng, D. Chen, and S. Lu (2010). "Analysis and performance study for coordinated hierarchical cache placement strategies". *Computer Communications*. 33(15): 1834–1842.

[158] Liakopoulos, N., A. Destounis, G. S. Paschos, T. Spyropoulos, and P. Mertikopoulos (2019). "Cautious regret minimization: Online optimization with long-term budget constraints". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 1–9.

[159] Lim, K., J. Chang, T. Mudge, P. Ranganathan, S. Reinhardt, and T. Wenisch (2009). "Disaggregated memory for expansion and sharing in blade servers". In: *Proceedings of the 36th Annual International Symposium on Computer Architecture (ISCA)*. 267–278.

[160] Liu, A. and V. K. N. Lau (2016). "Asymptotic scaling laws of wireless ad hoc network with physical layer caching". *IEEE Transactions on Wireless Communications*. 15(3): 1657–1664.

[161] Lykouris, T. and S. Vassilvitskii (2018). "Competitive caching with machine learning advice". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 1–10.

[162] Maculan, N., C. P. Santiago, E. M. Macambira, and M. H. C. Jardim (2003). "An $O(n)$ algorithm for projecting a vector on the intersection of a hyperplane and a box in R^n". *Journal of Optimization Theory and Applications*. 117: 553–574.

[163] Maddah-Ali, M. and U. Niesen (2014). "Fundamental limits of caching". *IEEE Transactions on Information Theory*. 60(5): 2856–2867.

[164] Maggi, L., L. Gkatzikis, G. Paschos, and J. Leguay (2018). "Adapting caching to audience retention rate". *Computer Communications*. 116: 159–171.

[165] Mahanti, A., C. Williamson, and D. Eager (2000). "Traffic analysis of a web proxy caching hierarchy". *IEEE Network*. 14(3): 16–23.

[166] Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

[167] Maskin, M. S. (1990). *Network and Discrete Location: Models, Algorithms, and Applications*. John Wiley and Sons.

[168] Mattson, R. L., J. Gecsei, D. R. Slutz, and I. L. Traiger (1970). "Evaluation techniques for storage hierarchies". *IBM Systems Journal*. 9(2): 78–117.

[169] Megiddo, N. and D. S. Modha (2003). "ARC: A self-tuning, low overhead replacement cache". In: *Proceedings of the 2nd Usenix Conference on File and Storage Technologies (FAST'03)*. Vol. 3. No. 2003. 115–130.

[170] Meiss, M., F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani (2008). "Ranking web sites with real user traffic". In: *Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM)*. 65–75.

[171] Mirchandani, P. and R. Francis (1990). *Discrete Location Theory*. John Wiley and Sons.

[172] Mitzenmacher, M. (2004). "A brief history of generative models for power law and lognormal distributions". *Internet Mathematics*. 1(2): 226–251.

[173] Mo, J. and J. Walrand (2000). "Fair end-to-end window-based congestion control". *IEEE/ACM Transactions on Networking*. 8(5): 556–567.

[174] Nair, J., A. Wierman, and B. Zwart (2013). "The fundamentals of heavy-tails: properties, emergence, and identification". In: *ACM SIGMETRICS Performance Evaluation Review*. Vol. 41. No. 1. ACM. 387–388.

[175] Naveen, K., L. Massoulie, E. Baccelli, A. Viana, and D. Towsley (2015). "On the interaction between content caching and request assignment in cellular cache networks". In: *Proceedings of the 5th Workshop on All Things Cellular (ATC)*. ACM. 37–42.

[176] Nedic, A. and A. Ozdaglar (2009). "Approximate primal solutions and rate analysis for dual subgradient methods". *SIAM Journal on Optimization*. 19(4): 1757–1780.

[177] Nemhauser, G. and L. Wolsey (1988). *Integer and Combinatorial Optimization*. Wiley Publishing.

[178] Nemhauser, G., L. Wolsey, and M. Fisher (1978). "An analysis of approximations for maximizing submodular set functions". *Mathematical Programming*. 14(1): 265–294.

[179] Newman, M. E. J. (2005). "Power laws, Pareto distributions and Zipf's law". *Contemporary Physics*. 46(Sept.): 323–351.

[180] Niesen, U., D. Shah, and G. Wornell (2009). "Caching in wireless networks". In: *Proceedings IEEE International Symposium on Information Theory*. IEEE. 2111–2115.

[181] Nygren, E., R. K. Sitaraman, and J. Sun (2010). "The Akamai network: A platform for high-performance Internet applications". *SIGOPS Operating Systems Review*. 44(3): 2–19.

[182] Olmos, F., B. Kauffmann, A. Simonian, and Y. Carlinet (2014). "Catalog dynamics: Impact of content publishing and perishing on the performance of a LRU cache". In: *26th International Teletraffic Congress (ITC), 2014*. IEEE. 1–9.

[183] Olmos, F. and B. Kauffmann (2015). "An inverse problem approach for content popularity estimation". *EAI Endorsed Trans. Scalable Inf. Syst.* 3(9): e3.

[184] O'Neil, E. J., P. E. O'Neil, and G. Weikum (1999). "An optimality proof of the LRU-K page replacement algorithm". *Journal of the ACM (JACM)*. 46(1): 92–112.

[185] Pacifici, V., S. Josilo, and G. Dan (2016). "Distributed algorithms for content caching in mobile Backhaul networks". In: *Proceedings of International Teletraffic Congress*. 313–321.

[186] Padmanabhan, V. N. and L. Qiu (2000). "The content and access dynamics of a busy web site: Findings and implications". *ACM SIGCOMM Computer Communication Review*. 30(4): 111–123.

[187] Pappas, S. (2016). "How big is the Internet, really?" *Live Science*. URL: https://www.livescience.com/54094-how-big-is-the-internet.html.

[188] Paschos, G. S., A. Destounis, and G. Iosifidis (2019a). "Learning to cooperate in D2D caching networks". In: *Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 1–5.

[189] Paschos, G., A. Destounis, and G. Iosifidis (2020). "Online convex optimization for caching networks". *IEEE/ACM Transactions on Networking*. 28(2): 625–638.

[190] Paschos, G., A. Destounis, L. Vignieri, and G. Iosifidis (2019b). "Learning to cache with no regret". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 235–243.

[191] Paschos, G. S., E. Bastug, I. Land, G. Caire, and M. Debbah (2016). "Wireless caching: Technical misconceptions and business barriers". *IEEE Communications Magazine*. 54(8): 16–22.

[192] Paschos, G. S., S. Gitzenis, and L. Tassiulas (2012). "The effect of caching in sustainability of large wireless networks". In: *Proceedings of the 10th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt 2012)*. IEEE. 355–360.

[193] Paschos, G., G. Iosifidis, M. Tao, D. Towsley, and G. Caire (2018). "The role of caching in future communication systems and networks". *IEEE Journal on Selected Areas of Communication (Editorial)*. 36(6): 1111–1125.

[194] Piantadosi, S. (2014). "Zipf's word frequency law in natural language: A critical review and future directions". *Psychonomic Bulletin and Review*. 21: 1112–1130.

[195] Poularakis, K., G. Iosifidis, A. Argyriou, and L. Tassiulas (2014a). "Video delivery over heterogeneous cellular networks: Optimizing cost and performance". In: *Proceedings of IEEE INFOCOM*. IEEE. 1078–1086.

[196] Poularakis, K., G. Iosifidis, I. Pefkianakis, L. Tassiulas, and M. May (2016a). "Mobile data offloading through caching in residential 802.11 wireless networks". *IEEE Transactions on Network and Service Management*. 13(11): 71–84.

[197] Poularakis, K., G. Iosifidis, V. Sourlas, and L. Tassiulas (2016b). "Exploiting caching and multicast for 5G wireless networks". *IEEE Transactions on Wireless Communications*. 15(4): 2995–3007.

[198] Poularakis, K. and L. Tassiulas (2012). "Optimal cooperative content placement algorithms in hierarchical cache topologies". In: *Proceedings of the 46th Annual IEEE Conference on Information Sciences and Systems (CISS)*. 1–6.

[199] Poularakis, K., G. Iosifidis, and L. Tassiulas (2014b). "Approximation algorithms for mobile data caching in small cell networks". *IEEE Transactions on Communications*. 62(10): 3665–3677.

[200] Poularakis, K. and L. Tassiulas (2016). "On the complexity of optimal content placement in hierarchical caching networks". *IEEE Transactions on Communications*. 64(5): 2092–2103.

[201] Qiu, L., V. N. Padmanabhan, and G. M. Voelker (2001). "On the placement of web server replicas". In: *IEEE INFOCOM*. IEEE. 1587–1596.

[202] Rahmaniani, R., T. C. Crainic, M. Gendreaua, and W. Rei (2017). "The benders decomposition algorithm: A literature review". *European Journal of Operational Research*. 259(3): 801–817.

[203] Rao, C. R. (1992). "Information and the accuracy attainable in the estimation of statistical parameters". In: *Breakthroughs in Statistics*. Springer. 235–247.

[204] Rappaport, A. and D. Raz (2013). "Update aware replica placement". In: *Proceedings of the 9th IEEE International Conference on Network and Service Management (CNSM)*. 92–99.

[205] Roadknight, C., I. Marshall, and D. Vearer (2000). "File popularity characterisation". *ACM Sigmetrics Performance Evaluation Review*. 27(4): 45–50.

[206] Roberts, J. and N. Sbihi (2013). "Exploring the memory-bandwidth tradeoff in an information-centric network". In: *Proceedings of the 25th IEEE International Teletraffic Congress (ITC)*. 1–9.

[207] Rodolakis, G., S. Siachalou, and L. Georgiadis (2017). "Replicated server placement with QoS constraints". *IEEE Transactions on Parallel Distributed Systems*. 17(10): 1151–1162.

[208] Rodriguez, P., C. Spanner, and E. W. Biersack (2001). "Analysis of web caching architectures hierarchical and distributed caching". *IEEE/ACM Transactions on Networking*. 9(4): 404–418.

[209] Roy, B. V. (2007). "A short proof of optimality for the MIN cache replacement algorithm". *Information Processing Letters*. 102(2–3): 72–73.

[210] Saavedra, A. G., X. Costa-Perez, D. J. Leith, and G. Iosifidis (2019). "FluidRAN: Optimized vRAN/MEC orchestration". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 2366–2374.

[211] Sadeghi, A., F. Sheikholeslami, and G. B. Giannakis (2018). "Optimal and scalable caching for 5G using reinforcement learning of space-time popularities". *IEEE Journal of Selected Topics in Signal Processing*. 12(1): 180–190.

[212] Sahoo, J., M. A. Salahuddin, R. Glitho, H. Elbiaze, and W. Ajib (2017). "A survey on replica server placement algorithms for content delivery networks". *IEEE Communications Surveys and Tutorials*. 19(2): 1002–1026.

[213] Scellato, S., C. Mascolo, M. Musolesi, and J. Crowcroft (2011). "Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades". In: *Proceedings of the 20th International Conference on World Wide Web. WWW '11*. Hyderabad, India: ACM. 457–466.

[214] Schmidt, M., E. Berg, M. Friedlander, and K. Murphy (2009). "Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm". In: *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR*. Vol. 5. 456–463.

[215] Shalev-Shwartz, S. (2012). *Online Learning and Online Convex Optimization*. Now Publishers Inc.

[216] Shmoys, D. B., C. Swamy, and R. Levi (2004). "Facility location with service installation costs". In: *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 1088–1097.

[217] Shmoys, D. B., E. Tardos, and K. Aardal (1997). "Approximation algorithms for facility location problems". In: *Proceedings of the 29th ACM Symposium on Theory of Computing (STOC)*. 265–274.

[218] Singla, A., B. Chandrasekaran, P. Godfrey, and B. Maggs (2014). "The internet at the speed of light". In: *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*. ACM. 1.

[219] Sleator, D. D. and R. E. Tarjan (1985). "Amortized efficiency of list update and paging rules". *Communications of the ACM*. 28(2): 202–208.

[220] Somuyiwa, S. O., A. Gyorgy, and D. Gunduz (2018). "A reinforcement-learning approach to proactive caching in wireless networks". *IEEE Journal on Selected Areas in Communications*. 36(6): 1331–1344.

[221] Sourlas, V., P. Flegkas, G. S. Paschos, D. Katsaros, and L. Tassiulas (2011). "Storage planning and replica assignment in content-centric publish/subscribe networks". *Computer Networks*. 55(18): 4021–4032.

[222] Starobinski, D. and D. Tse (2001). "Probabilistic methods for web caching". *Performance Evaluation*. 46(2–3): 125–137.

[223] Su, H., A. W. Yu, and L. Fei-Fei (2012). "Efficient Euclidean projections onto the intersection of norm balls". In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*. 1563–1570.

[224] Swamy, C. and Kumar (2004). "Primal-dual algorithms for connected facility location problems". *Algorithmica*. 40(4): 182–196.

[225] Tanenbaum, A. S. (2009). *Modern Operating System*. Pearson Education, Inc.

[226] Tortelli, M., D. Rossi, and E. Leonardi (2016). "ModelGraft: Accurate, scalable, and flexible performance evaluation of general cache networks". In: *Proceedings of the 28th International Teletraffic Congress*. Vol. 01. IEEE. 304–312.

[227] Tran, T. X., A. Hajisami, and D. Pompili (2017). "Cooperative hierarchical caching in 5G cloud radio access networks". *IEEE Network*. 31(4): 35–41.

[228] Traverso, S., M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini (2013). "Temporal locality in today's content caching: Why it matters and how to model it". *ACM Computer Communication Review*. 43(5): 5–12.

[229] Tuholukova, A., G. Neglia, and T. Spyropoulos (2017). "Optimal cache allocation for femto helpers with joint transmission capabilities". In: *Proceedings of the IEEE International Conference on Communications*. 1–7.

[230] Urdaneta, G., G. Pierre, and M. Van Steen (2009). "Wikipedia workload analysis for decentralized hosting". *Computer Networks*. 53(11): 1830–1845.

[231] Valls, V., G. Iosifids, D. Leith, and L. Tassiulas (2020). "Online convex optimization with perturbed constraints: Optimal rates against stronger benchmarks". In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR*. Vol. 108. 2885–2895.

[232] Valls, V. and D. J. Leith (2019). "A convex optimization approach to discrete optimal control". *IEEE Transactions on Automatic Control*. 64(1): 35–50.

[233] Voelker, G. M., N. Sharma, N. Cardwell, A. Karlin, and H. M. Levy (1999). "On the scale and performance of cooperative web proxy caching". In: *Proceedings of the 17th ACM Symposium on Operating System Principles (SOSP)*. 16–31.

[234] Vogler, W. (2008). "Another short proof of optimality for the MIN cache replacement algorithm". *Information Processing Letters*. 106(5): 219–220.

[235] Von Luxburg, U. (2007). "A tutorial on spectral clustering". *Statistics and Computing*. 17(4): 395–416.

[236] Wang, L., S. Bayhan, and J. Kangasharju (2015). "Optimal chunking and partial caching in information-centric networks". *Computer Communications*. 61: 48–57.

[237] Wang, M. (2017). "Vanishing price of anarchy in large coordinate nonconvex optimization". *SIAM Journal on Optimization*. 27(3): 1977–2009.

[238] Wang, M., C. W. Tan, W. Xu, and A. Tang (2011). "Cost of not splitting in routing: Characterization and estimation". *IEEE/ACM Transactions on Networking*. 19(6): 836–850.

[239] Wang, W. and C. Lu (2015). "Projection onto the capped simplex". *arXiv preprint arXiv:1503.01002*.

[240] Wierzbicki, A., N. Leibowitz, M. Ripeanu, and R. Woźniak (2004). "Cache replacement policies for P2P file sharing protocols". *Transactions on Emerging Telecommunications Technologies*. 15(6): 559–569.

[241] Wikipedia statistics (n.d.). *Top-1000 request statistics for English Wikipedia pages*. URL: https://wikitech.wikimedia.org/wiki/Pageviews_API.

[242] Woo, S., E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park (2013). "Comparison of caching strategies in modern cellular Backhaul networks". In: *Proceeding of the 11th ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 319–332.

[243] Xie, H., G. Shi, and P. Wang (2012). "Tecc: Towards collaborative in-network caching guided by traffic engineering". In: *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*. 2546–2550.

[244] Yamakami, T. (2006). "A Zipf-like distribution of popularity and hits in the mobile web pages with short life time". In: *Proceedings of Parallel and Distributed Computing, Applications and Technologies, PDCAT '06*. IEEE. 240–243.

[245] Young, N. E. (1994). "The k-server dual and loose competitiveness for paging". *Algorithmica*. 11(6): 525–541.

[246] Young, N. E. (2002). "On-line file caching". *Algorithmica*. 33(3): 371–383.

[247] Yu, H., D. Zheng, B. Y. Zhao, and W. Zheng (2006). "Understanding user behavior in large-scale video-on-demand systems". *ACM SIGOPS Operating Systems Review*. 40(4): 333–344.

[248] Zhang, J., B. Chen, and Y. Ye (2005). "A multiexchange local search algorithm for the capacitated facility location problem". *Mathematics of Operations Research*. 30(2): 389–403.

[249] Zhao, S., D. Stutzbach, and R. Rejaie (2006). "Characterizing files in the modern Gnutella network: A measurement study". In: *Proceedings of SPIE 6071, Multimedia Computing and Networking*. 1–13.

[250] Zhao, T., I.-H. Hou, S. Wang, and K. S. Chan (2018). "ReD/LeD: An asymptotically optimal and scalable online algorithm for service caching at the edge". *IEEE Journal on Selected Areas in Communications*. 36(8): 1857–1870.

[251] Zink, M., K. Suh, Y. Gu, and J. Kurose (2009). "Characteristics of YouTube network traffic at a campus network – Measurements, models, and implications". *Computer Networks*. 53(4): 501–514.

[252] Zinkevich, M. (2003). "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th International Conference on Machine Learning (ICML)*. 928–935.

[253] Zotano, M. G., J. G. Sanz, and J. Pavón (2015). "Analysis of web objects distribution". In: *Distributed Computing and Artificial Intelligence, 12th International Conference*. Springer. 105–112.