mance is fixed at either optimizing bandwidth cost or cache-miss probability, without regarding for the other metric.

In the thus far implemented joint-price schemes, we have been assigning requests to download chunks *strictly* from the set of min-priced servers. We now relax this strict consideration by regarding any BS, whose price is within a *tolerance value* of τ from the minimum price, as *eligible* to serve chunks for an incoming request (provided as before the eligible BSs are connected to the request's location); chunks are however downloaded by sorting the eligible BSs in the increasing order of their prices. In Figure 4 we have depicted the performance of the joint-price scheme for different values of τ, where the earlier strict min-price approach now corresponds to $\tau = 0$; the case where all connected BSs (irrespective of their prices) are eligible corresponds to $\tau = \infty$.

We observe that as τ increases the trade-off curves shift downwards, implying that it is possible to achieve a lower cache-miss probability for a given target bandwidth cost. However, the range of trade-off that is possible reduces as τ increases. For instance, using $\tau = 100$ it is not possible to reduce the bandwidth cost to less than 40; In fact, $\tau = \infty$ case achieves the lowest cache-miss probability, however at the expense of being constrained to operate at a higher range of bandwidth cost. This shift in performance is because, as τ increases, more BSs become eligible to serve a request so that the cache-miss probability decreases; however, more BSs are now loaded, resulting in a bandwidth scarce system. Thus, further trade-off in performance can be achieved by suitably choosing the value of τ.

In summary, the joint-price scheme, in conjunction with the CLRU policy for cache management, can be a practical candidate for request assignments in futuristic cellular networks where a dense deployment of heterogeneous femtocell base-stations are expected. The cache-miss cost b, along with the tolerance value τ, can be used to serve as "tunable knobs" to trade-off one metric for the other.

5. CONCLUSION

We proposed an optimization framework to study the problem of cost minimization in cellular cache networks. Towards this direction, we proposed an online algorithm for caching and request assignments which is shown to be optimal and stable in a limiting regime that is obtained by scaling the arrival rates and the content chunking. Based on the online algorithm, we proposed a light-weight joint-price scheme for request assignment that can be used in conjunction with the LRU cache management strategy. Through simulations we found that our joint-price based request assignment strategy outperforms the common practices of routing purely based on either load or content availability. Our proposed joint-price routing mechanisms are thus an appealing candidate combining sound theoretical guarantees with good experimental performance while being simple to implement.

6. REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019. February 2015.

[2] N. Abedini and S. Shakkottai. Content Caching and Scheduling in Wireless Networks With Elastic and Inelastic Traffic. *IEEE/ACM Transactions on Networking*, 2013.

[3] M. Amble, P. Parag, S. Shakkottai, and L. Ying. Content-Aware Caching and Traffic Management in Content Distribution Networks. In *INFOCOM '11*.

[4] I. Baev, R. Rajaraman, and C. Swamy. Approximation Algorithms for Data Placement Problems. *SIAM J. Comput.*, 38(4):1411–1429, August 2008.

[5] G. Bianchi, A. Detti, A. Caponi, and N. Blefari Melazzi. Check Before Storing: What is the Performance Price of Content Integrity Verification in LRU Caching? *SIGCOMM CCR*, July 2013.

[6] T. Bonald, M. Jonckheere, and A. Proutiére. Insensitive Load Balancing. In *ACM SIGMETRICS /PERFORMANCE '04*.

[7] S. Borst, V. Gupta, and A. Walid. Distributed Caching Algorithms for Content Distribution Networks. In *INFOCOM '10*.

[8] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li. Collaborative Hierarchical Caching with Dynamic Request Routing for Massive Content Distribution. In *INFOCOM '12*.

[9] A. Dan and D. Towsley. An Approximate Analysis of the LRU and FIFO Buffer Replacement Schemes. In *ACM SIGMETRICS '90*.

[10] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman. On the Complexity of Optimal Routing and Content Caching in Heterogeneous Networks. In *to appear at INFOCOM 2015*.

[11] C. Fricker, P. Robert, and J. Roberts. A Versatile and Accurate Approximation for LRU Cache Performance. In *Proceedings of the 24th International Teletraffic Congress, ITC '12*.

[12] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire. FemtoCaching: Wireless Video Content Delivery Through Distributed Caching Helpers. In *INFOCOM '12*.

[13] E. Jaho, M. Karaliopoulos, and I. Stavrakakis. Social Similarity Favors Cooperation: The Distributed Content Replication Case. *IEEE Transactions on Parallel and Distributed Systems*, March 2013.

[14] W. Jiang, S. Ioannidis, L. Massoulié, and F. Picconi. Orchestrating Massively Distributed CDNs. In *ACM CoNEXT '12*.

[15] M. Jonckheere and J. Virtamo. Optimal Insensitive Routing and Bandwidth Sharing in Simple Data Networks. In *ACM SIGMETRICS '05*.

[16] M. Leconte, M. Lelarge, and L. Massoulié. Bipartite Graph Structures for Efficient Balancing of Heterogeneous Loads. In *ACM SIGMETRICS/ PERFORMANCE '12*.

[17] K. P. Naveen, L. Massoulie, E. Baccelli, A. Carneiro Viana, and D. Towsley. On the Interaction between Content Caching and Request Assignment in Cellular Cache Networks. Research Report RR-8707, INRIA Saclay, March 2015.

[18] K. Poularakis and L. Tassiulas. Exploiting User Mobility for Wireless Content Delivery. In *IEEE International Symposium on Information Theory Proceedings, ISIT '13*.

[19] B. Tan and L. Massoulié. Optimal Content Placement for Peer-to-Peer Video-on-Demand Systems. *IEEE/ACM Transactions on Networking*, April 2013.