Network-Assisted Offloading for Mobile Cloud Applications

Claudio Fiandrino, Dzmitry Kliazovich, Pascal Bouvry
University of Luxembourg
6, Rue Richard Coudenhove Kalergi, Luxembourg
{name.surname}@uni.lu

Albert Y. Zomaya
University of Sydney
Sydney, NSW 2006, Australia
albert.zomaya@sydney.edu.au

Abstract—Data traffic from mobile devices experiences unprecedented growth, which current cellular network capacities cannot sustain. Traffic offloading to other type of networks, such as WiFi, can be used to reduce load in cellular networks. In this paper, we propose a novel solution, which unlike other existing methodologies, implements tight cooperation with the cellular network to optimize traffic offloading. The cellular network provides information about channel usage statistics, user mobility patterns, available resources and other parameters. The offloading decisions aim at optimizing the balance between user application requirements and availability of network resources. The validation results, obtained from NS-3 simulations, confirm effectiveness of the proposed solution in balancing cellular traffic load while ensuring QoS.

Keywords—Mobile cloud computing, network awareness, cellular networks, traffic offloading.

I. INTRODUCTION

Mobile cloud applications is one of the fastest growing markets. Currently, more than 7 billion people use mobile devices connected to Internet, of which 4.4 billion will use mobile applications by 2017 [1]. Mobile gadgets and smartphones are already essential in our daily activities [2]. They help to do business, communicate and entertain [3], [4]. In 2013, smart devices accounted for only 21% of the total number of Internet connections originated by mobile devices, but they generated more than 88% of the total mobile data traffic [5]. Currently, mobile data traffic is growing at an unprecedented rate and is projected to generate up to 15 EB per month by 2018 [5], which overloads current cellular networks.

A number of promising solutions have been proposed to address growing traffic demands. With the objective of "bringing network closer to the user", a mix of macro, micro, pico, femto and relay base stations has been proposed in LTE-A [6]. Reducing the cell size helps to increase network capacity and coverage, but comes at the expense of the costs of installing and maintaining base stations denser [7]. Network coding is proved to improve network throughput and is a widely used technique in mobile networks nowadays. Recently, a technique that improves content distribution of mobile cloud applications by combining traffic flows has been proposed for cellular networks [8]. Traffic offloading to other networks can help to avoid having additional cellular network equipment. The offloading is usually performed to WiFi [9]–[11] or to the opportunistic networks [12]–[14]. WiFi operates on unlicensed frequencies and, unlike LTE, which uplink and downlink data rates remain constant to every user, WiFi channels are shared

between the served users. WiFi infrastructure is not expensive and already widespread in many areas. Opportunistic networks do not require any infrastructure, but they operate relying on intermittent contacts of in proximity users.

From the user perspective, with traffic offloading applications can directly contact the server in the cloud without the need to pass by the mobile core network. WiFi and opportunistic communications are more energy efficient than LTE or 3G [15], [16] and commonly available "free of charge". As a result, users do not consume data from the monthly traffic plan they subscribed and for which they have to pay the operators. On the other hand, traffic offloading latency can be an issue affecting performance [17] as both WiFi and opportunistic communications become unreliable with high levels of user mobility. Network performance can change quickly and degrade user experience making it worse than in the case without offloading. In addition, most of the operating systems for smartphones, including Android and iOS, already make preference to WiFi over cellular connectivity for data transmission. However, keeping both interfaces constantly active excessively drains battery power. For these reasons, offloading decisions should be taken carefully and take into account performance and mobility in the offloaded networks.

In this paper, we propose to exploit information from the cellular network about user statistics, application requirements and network resource availability to improve traffic offloading. For this, we introduce a software module called Mobile Cloud Offloading Helper (MCOH) in the mobile operator network, which uses the aforementioned information to take offloading decisions. Wireless service providers have at disposal information about the number and the type of connected devices, the traffic and network statistics, such as available bandwidth, error rates and delay. However, now it is mostly used to understand habits of large groups of users with an objective of predicting general trends. Our contribution, instead, consists of using this information to enable a *"user-application" based* offloading policy. For this, MCOH retrieves and uses the aforementioned information on a per-user and per-application basis, while the application awareness is obtained by inspection of the in transit traffic.

Cellular and WiFi technologies provide different performance levels according to the application currently in use. Mobile cloud applications have very different requirements from the networking point of view. For example, multimedia streaming applications like VoIP impose severe constraints on the latency. On the other hand, retrieving large objects
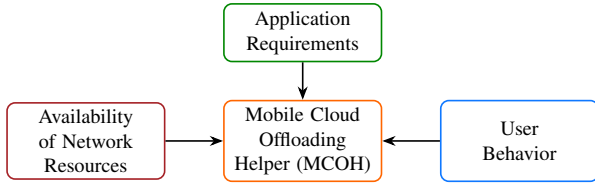
Fig. 1. Input parameters

APPLICATION CLASS	COEFFICIENTS		
	μ	ρ	ω
Bandwidth intensive	0.15	0.70	0.15
Delay constrained	0.425	0.150	0.425
Bandwidth intensive and Delay constrained	0.34	0.33	0.33

like geographic maps requires bandwidth availability, but the application is delay tolerant. Mobile cloud gaming is an example of application which necessitates of high traffic load and requires tight delay constraints to keep synchronization among the players. Hence, WiFi technology appears to be more suitable for services demanding high-bandwidth like navigation through online maps because of its capability to achieve higher bit rates, but due to higher loss ratios it might not be suitable for VoIP applications.

II. THE MOBILE CLOUD OFFLOADING HELPER

Mobile Cloud Offloading Helper (MCOH) is a software module, which resides in mobile operator network and uses feedback from the network, application requirements and user behavior to assist in taking traffic offloading decisions (see Fig. 1). Channel quality, data rate and latency are the main performance characteristics affecting availability of network. From user behavior, MCOH considers mobility and data traffic plan. Offloading decisions aim at balancing user behavior and application requirements with available network resources. Traffic is offloaded when the aforementioned requirements exceed a threshold called *offloading-sensitivity*, which allows the mobile network operators to tune the amount of offloaded traffic.

A. Network Architecture

Fig. 2 shows the position of MCOH in mobile operator network featuring both LTE[1] and WiFi connectivity. Dual connectivity is currently very common, as more and more operators deploy proprietary WiFi networks [10], [14].

Packet Data Network Gateway (P-GW) bridges cellular, WiFi and wide-area networks. It allocates network resources to UEs and handles QoS and data encapsulation. Current P-GWs usually run Deep Packet Inspection (DPI), which can also be used by MCOH to acquire knowledge about the application requirements. In addition, with the help of the Policy Control and Charging Rules Function (PCRF), the P-GW also provides flow-based charging control decisions. Indeed, the PCRF is in charge of providing to the P-GW the QoS parameters and the bit rates according to subscription profiles of the users. This information permits the MCOH to be aware of the amount of data consumed by each user.

The Serving Gateway (S-GW) handles network traffic before it is delivered to eNodeBs and UEs. The S-GW is responsible of packet forwarding and collecting statistics for billing, i.e., the amount of traffic sent and received by each user. The S-GW also helps in user authentication and interacts with the Home

Subscriber Server (HSS). The latter contains user information and subscription profiles containing QoS parameters, such as traffic classes and allowed bit rates. The Mobility Management Entity (MME) controls signaling between UEs and the core network including, for example, procedures for S-GW selection for initial UE attachment and during horizontal handovers. The MCOH obtains information on user mobility from this module.

B. The User and Application Based Offloading Policy

Unlike traditional offloading approaches, the MCOH offloads traffic taking into account user application requirements and availability of network resources. The MCOH relies on a set of impact factors to model both application and user requirements. The set of impact factors can be extended with respect to the representative ones considered in this paper for more fine models.

1) Application Requirements: Different types of cloud applications use network resources differently [18]. Therefore, considering their requirements is a key point for taking proper offloading decisions. In this paper, we consider three classes of applications: (a) delay constrained (e.g., VoIP), (b) bandwidth intensive (e.g., navigation) and (c) both delay constrained and bandwidth intensive (e.g., gaming). Such classification reflects in a general way the most important network resources that applications need for their operation. Packet inspection tools like DPI, commonly installed in P-GW, can be used to identify type applications and their requirements. From them, MCOH computes *application impact factor* as a weighted sum of channel quality (f_{A_C}), bit rate (f_{A_R}) and latency (f_{A_L}):

$$f_A = \mu \cdot f_{A_C} + \rho \cdot f_{A_R} + \omega \cdot f_{A_L}. \qquad (1)$$

Coefficients μ, ρ and ω define the impact of the corresponding components. Higher values of μ and ω are assigned to delay-constrained applications, while higher values of ρ characterize bandwidth intensive applications. The sum of μ, ρ and ω must equal unity. Table I shows representative setups obtained from the evaluation experiments. The parameters f_{A_C}, f_{A_R}, and f_{A_L} are all in the range $[0, 1]$.

To track channel quality we rely on the Received Signal Strength Indicator (RSSI) that is of easy detection since every wireless card measures it. Although being measured on the reception of Physical Layer Convergence Protocol (PCLP) preamble and header that are always transmitted at basic rate [19], low values of RSSI usually lead to low Mean Opinion Score (MOS) values [20] in delay constrained applications. Ideally, Signal to Interference plus Noise Ratio (SINR) would be a better option in estimating channel quality. However, SINR can not be measured efficiently [21]. We define the *channel quality impact factor* f_{A_C} as:

$$f_{A_C} = 1 - \frac{|\text{RSSI}_{high} - \text{RSSI}|}{|\text{RSSI}_{high} - \text{RSSI}_{low}|}, \qquad (2)$$
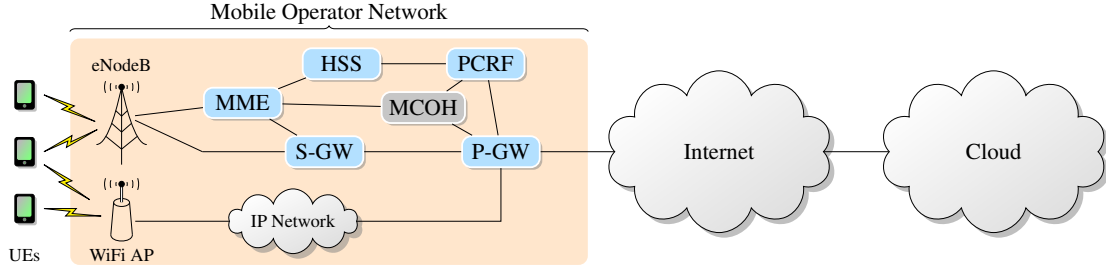
---

[1]http://www.3gpp.org/LTE/

Fig. 2.   System scenario

where RSSI is the last RSSI value detected for the user, RSSI_{high} corresponds to the RSSI determined as the user was at one meter distant from the AP being in line of sight and RSSI_{low} is the lowest RSSI detected for the card.

Signal-to-Noise (SNR) measured at the receiver side should guide the selection of proper transmission rate [22]. However, in practice the transmission rate is selected based on channel probing [22], which is obtained by testing the effectiveness of individual rates and measuring losses. High loss ratios force the selection of lower rates and, vice versa, low loss ratios allow transmission at higher rates. It should be noted that whenever collisions are detected, packet retransmissions occur at lower rates. We define the *rate impact factor* f_{A_R} as:

$$f_{A_R} = \frac{R}{R_{\max}}, \tag{3}$$

where R is the last selected transmission rate and $R_{\max}$ is the maximum transmission rate available to the user.

High latencies can seriously affect performance of the delay-constrained applications. To understand whether it is beneficial to offload traffic or not, the MCOH compares the one way delay (L) of WiFi (L_{WiFi}) and LTE (L_{LTE}) networks measured between the P-GW and the user. The *latency impact factor* is then defined as:

$$f_{A_L} = \max\left(\frac{L_{\text{LTE}} - L_{\text{WiFi}}}{L_{\text{LTE}}}, 0\right). \tag{4}$$

2) User Behavior: User mobility and employed data plan can affect offloading decisions considerably. High velocities commonly lead to higher variations in the received signal, fast fading and connection termination. Under such hypothesis, offloading delay-constrained application traffic is risky because multiple losses can impact QoS dramatically.

To compute the *mobility impact factor* f_M the MCOH considers the records of users velocity and determines the average velocity $\overline{V}_t$ of user u at time t through the Exponential Weighted Moving Average (EWMA):

$$\overline{V}_t = \sigma \cdot V_t + (1-\sigma) \cdot \overline{V}_{t-1}, \tag{5}$$

where V_t is the observed velocity at time t, $\overline{V}_{t-1}$ corresponds to the previous value of the average velocity and parameter σ is exponential weighting coefficient. Higher values of σ limit the contribution of older values.

Low values of $\overline{V}_t$ do not compromise user QoS. On the other hand, higher values of $\overline{V}_t$ are a serious obstacle to user

QoS. These observations suggest that f_M can be defined as a sigmoid function:

$$f_M = \frac{1}{1 + e^{-\frac{\psi_M}{\varepsilon_M} \cdot \left(-v + \left(1 - \frac{\varepsilon_M}{2}\right)\right)}}, \tag{6}$$

where parameters ψ_M and $1 - \varepsilon_M/2$ represent the incline and the center of the curve illustrated in Fig. 3(a) respectively, while v is the normalized velocity defined as ratio between $\overline{V}_{u|t}$ and a reference maximum velocity $V_{\max}$.

Users subscribe to a data plan with their operator, which defines the amount of data traffic they have at disposal. When the amount of data consumed is close to a limit, users are very interested in *not* receiving traffic from cellular interface. On the other hand, when the amount of data left to consume is high, users may not care whether the traffic is offloaded or not. From these observations, we define the *data plan impact factor* f_D as:

$$f_D = \frac{1}{1 + e^{-\psi_D \cdot (d - \varepsilon_D)}}. \tag{7}$$

where parameters ψ_D and ε_D represent the incline and the center of the curve respectively illustrated in Fig. 3(b), while d is the normalized utilized user data defined as ratio between the utilized user data and the nominal value.

Having at disposal information on user behaviour, application requirements and network resources, the MCOH can decide whether to offload traffic. The user-application based offloading policy of the MCOH takes into account the aforementioned information through a combination of application requirements (f_A), user velocity (f_M) and user data plan (f_D) factors:

$$f_A \cdot f_M \cdot f_D \geq \gamma. \tag{8}$$

Parameter γ is the *offloading sensitivity* which can be tuned by the mobile network operator to increase or decrease the amount of the offloaded traffic. Low values of γ will increase the amount of the offloaded traffic, which means that the user behavior and application requirements are fully satisfied. On the other hand, higher values of γ impose tighter constrains for offloading, which may decrease the amount of the offloaded traffic if the requirements are not met.

III.   PERFORMANCE EVALUATION

The effectiveness of the MCOH module performance has been evaluated using the NS-3 network simulator[2], extended with LTE functionality from LENA project framework.
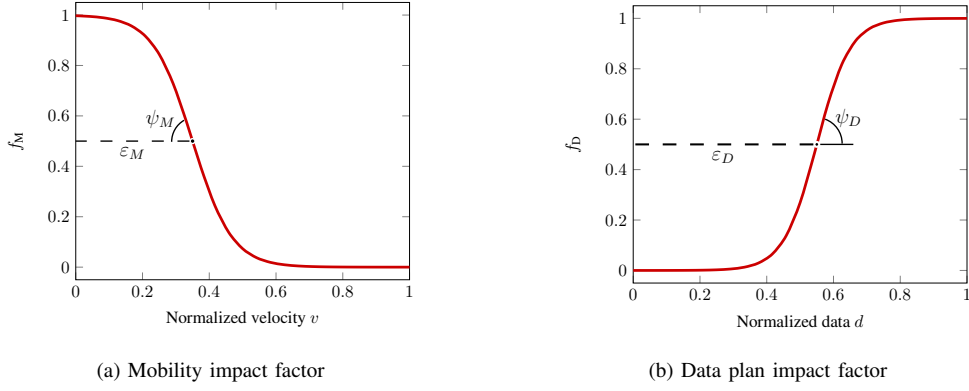
---

[2]www.nsnam.org

(a) Mobility impact factor

(b) Data plan impact factor

Fig. 3.   Impact of user requirements

TABLE II.   SIMULATION PARAMETERS

PARAMETER	VALUE
μ	0.45
ρ	0.10
ω	0.45
ε_D	18.00
ψ_D	0.25
ε_M	9.00
ψ_M	0.65

Fig. 2 illustrates simulation scenario. A set of 4 users has been equipped to support LTE and WiFi functionalities. The users move at different average speeds on a square region with WiFi AP and eNodeB placed in the center. The maximum velocity $V_{\max}$ has been set to 10 m/s, which corresponds to the velocity of an average human run. User terminals constantly signal their location, which is used to determine their distance d from the WiFi AP. Knowing d it becomes possible to estimate RSSI [23]:

$$\text{RSSI} = -(10 \cdot n \cdot \log_{10} d + A), \tag{9}$$

where n is the loss exponent ($n = 2$ for propagation in free space, $n = 2.7$ in the simulations) and A is the received signal strength when $d = 1$ m.

At the beginning of the simulation, an available data plan uniformly distributed between $[0, 1]$ GB is associated to each user. During simulation run, the data plan is reduced accordingly to the amount of traffic utilized.

At the beginning of the simulation, the amount of data left on the data plan unconsumed is uniformly distributed between $[0, 1]$ Gb for each user. All the users run a VoIP application modeled according to G.711 codec standard[3]. On and off voice periods are randomly chosen from exponential distribution. The VoIP application transmits packets at 64 Kbit/s with a 160 Byte long payload. To compute the application impact factor the weighting coefficients have been selected to be sensitive to channel quality and delay to satisfy the profile of delay-constrained applications. Table II summarizes the values of simulation parameters.

Fig. 4 shows the amount of data sent to the users through WiFi and LTE links with increasing values of the *offloading*
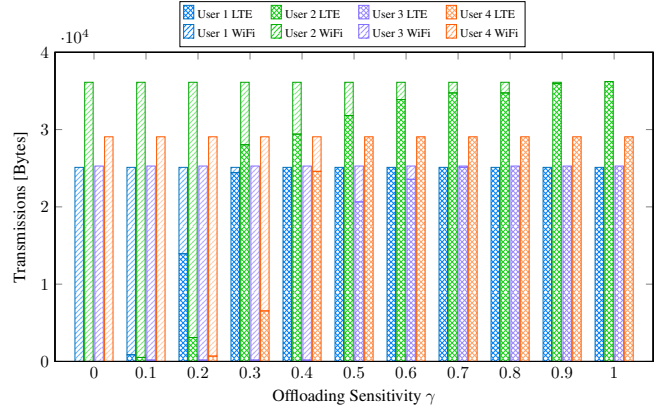

Fig. 4.   Data transmission on WiFi and LTE links

sensitivity γ for simulation duration of 60 s. In the extreme cases, when γ is equal to 0 or 1, all traffic is sent completely through the WiFi or LTE interface respectively. User 1 is the farthest from the AP, moves at a higher average speed than the other users and has utilized the major part of the available data. For these reasons, only low values of offloading sensitivity ($\gamma < 0.2$) permit to offload more than 50% of the traffic. User 2 moves less quickly than User 1. In addition, the user is not very close to the AP and has utilized a small part of her available data. Consequently, with high values of the offloading sensitivity part of her traffic can still be offloaded. Users 3 and 4 have intermediate parameters with respect to Users 1 and 2. Hence, they receive most of the traffic through the WiFi link with values of $\gamma < 0.5$. User 3 moves very slow being the closest user to the AP. However, the incentive to offloading given by good channel quality is mitigated by the fact that she utilized a small part of his available data. On the other hand, User 4 moves faster than User 3. In addition, the user is relatively distant from the AP, which penalizes offloading. However, having utilized most of her available traffic, the data plan impact factor does not inflate the offloading disincentive caused by mobility.

Fig. 5(a) and Fig. 5(b) illustrate the effect of the user parameters on the amount of offloaded traffic. Values close to 1 correspond to having all the traffic sent through the WiFi interface while values close to -1 denote that all the traffic is delivered through the LTE interface. Focusing on User 4 and

---

[3]http://www.cisco.com/c/en/us/support/docs/voice/voice-quality/ 7934-bwidth-consume.html

(a) Effect of mobility impact factor
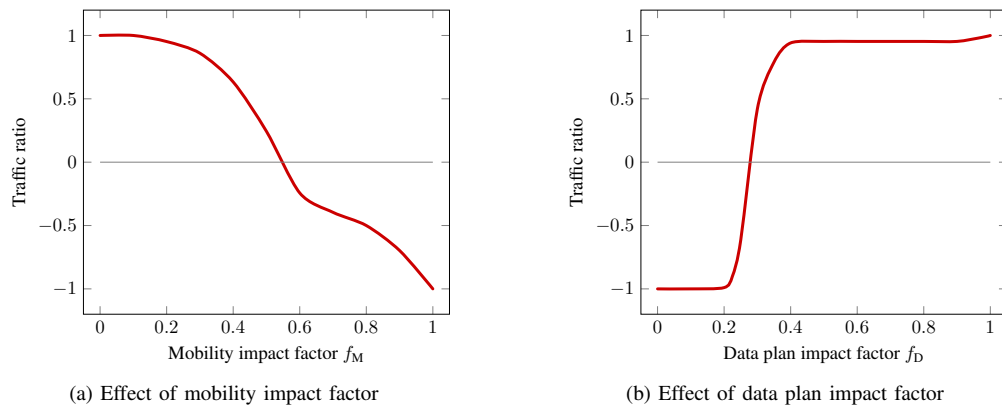

(b) Effect of data plan impact factor

Fig. 5. Effect of user parameters on the amount of offloaded traffic

having fixed the initial data plan at user disposal for all the experiments, the user was allowed to move with increasingly average velocity during the simulation runs to analyze the effect of the mobility. On the other hand, having fixed the user average velocity for all the experiments, increasing values of data plan have been set at the beginning of each simulation run. As a result, the mobility and the data plan impact factors in Fig. 5 correspond to averages computed over the entire simulation period. As expected, the results trace a sigmoid shape. Only the curve in Fig. 5(b) is very close to the one presented in Fig 3(b). With the offloading decisions taken on a per-packet basis, the instantaneous velocity is important and not the average over the entire simulation period velocity. Thus, the system can still offload traffic even when the average mobility impact factor approaches the value of 1. This behavior is not shown by the analysis of data plan impact factor because data utilized is progressively increased during the simulation period and VOIP application does not require significant bandwidth availability. Consequently, instantaneous variation of the utilized data has no relevant effect on the data plan impact factor.

IV. CONCLUSION AND FUTURE WORK

In this paper we propose MCOH, a software module that takes offloading decisions with the aim of optimizing balance between user behavior and application requirements with availability of network resources. For this, the MCOH module uses information already available in mobile operator networks including channel usage statistics and user mobility patterns among the others.

We validated the performance of the MCOH module through simulations performed with the NS-3 network simulator. The results highlight the importance of the aforementioned information in taking offloading decisions able to guarantee user QoS.

Future work will focus on extending the concept of the MCOH to allow using mobile operator network information to help not only the offloading process, but in a broader range of cloud networking applications.

ACKNOWLEDGMENT

REFERENCES

[1] "Global mobile statistics 2014," mobiThinking, 2014. [Online]. Available: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats

[2] A. Persaud and I. Azhar, "Innovative mobile marketing via smartphones," *Marketing Intelligence & Planning*, vol. 30, no. 4, pp. 418–443, 2012. [Online]. Available: http://www.emeraldinsight.com/doi/abs/10.1108/02634501211231883

[3] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 281–287. [Online]. Available: http://doi.acm.org/10.1145/1879141.1879176

[4] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, 2013. [Online]. Available: http://dx.doi.org/10.1002/wcm.1203

[5] "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," Cisco White paper, 2014. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

[6] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *2010 European Wireless Conference (EW)*, April 2010, pp. 978–982.

[7] M. Siegrist, T. C. Earle, H. Gutscher, and C. Keller, "Perception of mobile phone and base station risks," *Risk Analysis*, vol. 25, no. 5, pp. 1253–1264, 2005.

[8] C. Fiandrino, D. Kliazovich, P. Bouvry, and A. Y. Zomaya, "NC-CELL: Network coding-based content distribution in cellular networks for cloud applications," in *2014 IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 1205–1210.

[9] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 536–550, April 2013.

[10] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys'10. New York, NY, USA: ACM, 2010, pp. 209–222. [Online]. Available: http://doi.acm.org/10.1145/1814433.1814456

[11] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *In Proceedings of IEEE MASS*, 2011, pp. 192–201.

[12] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Picu, "Offloading cellular traffic through opportunistic communications: Analysis and optimization," *CoRR*, vol. abs/1405.3548, 2014.

[13] Y. Zhu, C. Zhang, and Y. Wang, "Mobile data delivery through opportunistic communications among cellular users: A case study for the D4D challenge," *Third International Conference on the Analysis of Mobile Phone Datasets (NetMob 2013)*, May 2013.

[14] B. Han, P. Hui, V. Kumar, M. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social

participation," *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 821–834, May 2012.

[15] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '12. New York, NY, USA: ACM, 2012, pp. 225–238. [Online]. Available: http://doi.acm.org/10.1145/2307636.2307658

[16] R. Friedman, A. Kogan, and Y. Krivolapov, "On power and throughput tradeoffs of WiFi and Bluetooth in smartphones," *IEEE Transactions on Mobile Computing*, vol. 12, no. 7, pp. 1363–1376, July 2013.

[17] X. Zhuo, W. Gao, G. Cao, and S. Hua, "An incentive framework for cellular traffic offloading," *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 541–555, March 2014.

[18] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "Profiling resource usage for mobile applications: A cross-layer approach," in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '11. New York, NY, USA: ACM, 2011, pp. 321–334. [Online]. Available: http://doi.acm.org/10.1145/1999995.2000026

[19] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," in *Proceedings of the ACM SIGCOMM 2010 Conference*, ser. SIGCOMM '10. New York, NY, USA: ACM, 2010, pp. 159–170. [Online]. Available: http://doi.acm.org/10.1145/1851182.1851203

[20] J. Zhu, "On traffic characteristics and user experience of skype video call," in *2011 IEEE 19th International Workshop on Quality of Service (IWQoS)*, June 2011, pp. 1–3.

[21] A. Vlavianos, L. Law, I. Broustis, S. Krishnamurthy, and M. Faloutsos, "Assessing link quality in IEEE 802.11 wireless networks: Which is the right metric?" in *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008.*, Sept 2008, pp. 1–6.

[22] D. Halperin, "Simplifying the configuration of 802.11 wireless networks with effective SNR," *CoRR*, vol. abs/1301.6644, 2013. [Online]. Available: http://arxiv.org/abs/1301.6644

[23] E.-E.-L. Lau and W.-Y. Chung, "Enhanced RSSI-based real-time user location tracking system for indoor and outdoor environments," in *International Conference on Convergence Information Technology, 2007*, Nov 2007, pp. 1213–1218.