Abstract—The rapid advances in networking, mobile computing, and virtualization, lead to a dramatic increase in the traffic demand. A cost-effective solution for serving it, while maintaining a good quality of service (QoS), would be to offload a part of the traffic originally targeted for cellular base stations (BSs) to a WiFi infrastructure. Related work on the WiFi offloading often considers markets with a single provider and omits parameters, such as the effect of the offloading on the perceived QoS by users, the capacity of the WiFi infrastructure, and competition of providers. In contrast to these approaches, this paper develops a detailed modeling framework for analysing the WiFi offloading using network economics, game theory, and queueing networks. It also proposes a novel network aggregation technique to reduce the computational complexity of the analysis. Using this framework, the performance of WiFi offloading was evaluated under various scenarios with respect to the bandwidth of BSs and APs, coverage of WiFi, and user preferences. Our results highlight that it is not always profitable for providers to invest in a large WiFi infrastructure. The limited capacity of the WiFi APs restricts the benefits of the offloading.

I. INTRODUCTION

According to forecasts, the wireless traffic will exceed the 24.3 exabytes per month worldwide by 2019 [1]. To cope with this explosion of traffic demand, providers aim to increase the capacity of their cellular networks. Traditional solutions for expanding the capacity involve the leasing of additional spectrum or the extension of the network infrastructure. However, such solutions are costly and time consuming. A cost-effective alternative for increasing the capacity that has received considerable attention is the data offloading: Part of the mobile data traffic originally targeted for cellular networks is served by the complementary network infrastructure that has been deployed (e.g., based on a WiFi network, femtocells).

Most approaches that study the WiFi offloading claim that it is beneficial for providers to offload as much traffic as possible to WiFi APs (e.g., [2]–[5]). Their main argument is that a large volume of offloaded traffic alleviates the congestion in cellular networks and reduces the operational costs. However, such studies usually omit the effect of the limited capacity of WiFi APs on the QoS and its long-term impact on the revenue of providers. Only a few studies consider the effect of the offloading on the QoS [6], [7]. However, they only focus on the physical layer (e.g., on the achievable data rate and SINR) and omit economic aspects, such as, the effect of the offloading on the competition of multiple providers and user decisions. In contrast to these approaches, this work develops a detailed modeling framework for evaluating the performance of WiFi offloading considering various technological and economic aspects. It models markets with multiple competing providers that can perform WiFi offloading and a population of users that select their provider considering the offered prices and QoS. Several questions drive this research: What is the optimal percentage of traffic that should be offloaded to a WiFi network? How does the coverage of WiFi, user traffic demand, and user preference affect it? What is its impact on users? Under which cases is it beneficial for providers to invest in WiFi infrastructure? To answer these questions, we have evaluated the performance of WiFi offloading under various scenarios based on the bandwidth of cellular BSs, the coverage of the WiFi network, and the user utility.

Our framework employs detailed queueing-theoretical models of the networks of providers, base stations (BSs), access points (APs), user arrivals, departures, and handovers and provides a methodology to analyse large-scale wireless markets. It is also modular, in that, it can incorporate different user utility functions and traffic demand, mobility patterns, and network topology models. To improve the computational efficiency of the analysis even further, it also proposes a network aggregation method based on the theorem of Norton [8]. This method allows the construction of equivalent network models for a specific region of interest omitting the details of the entire networks of providers achieving significant computational gains.

The analysis highlights that it is not always profitable for a provider to invest in a large number of WiFi APs. The benefits of the offloading increase with the WiFi coverage but with a diminishing rate. Furthermore, it is not beneficial to offload the entire traffic that can be served by APs. From a certain point onwards the APs will become congested. This will result in a QoS degradation reducing the revenue of providers. The paper is structured as follows: Section II overviews the related work. Section III presents our modeling framework, while Section IV discusses the analysis of the WiFi offloading under different scenarios. Section V describes the network aggregation method based on the theorem of Norton and Section VI presents our conclusions and future work plans.

II. RELATED WORK

The problem of WiFi offloading has received considerable attention. The proposed approaches can be classified into two
Fig. 1: Main components of the modeling framework.

A. The network infrastructure

Each provider (e.g., provider \( i \)) has deployed a set of BSs \( (K_i) \) covering a geographical region (e.g., a city). It may also offload a part of its data traffic to a set of APs \( (L_i) \). We assume that the APs are sparsely located within a geographical region of interest. Each AP is associated with a unique BS and may serve a part of the data traffic that is originally targeted for that BS\(^1\). In this work, we do not consider the effect of interference among WiFi APs. We also assume that in all wireless stations (BSs and APs) the available bandwidth is shared equally among connected users (processor-sharing discipline). For LTE cellular BSs, this bandwidth allocation models a scheduler that divides the OFDMA resources fairly among users. Similarly, for WiFi APs, the CSMA protocol guarantees a fair bandwidth sharing among users in the long run\(^2\).

Users generate sessions and connect to a wireless station to communicate. Specifically, during a session, a user transmits and receives data via that station. The user session generation follows a Poisson process with a total rate of \( \lambda \). This rate is allocated across providers according to the current market share \( z = (z_0, z_1, \ldots, z_I) \). The ratio of subscribers of the provider \( i \) is indicated by \( z_i \), while \( z_0 \) indicates the ratio of disconnected users. The user mobility in the network of a provider is modeled with a Markov-chain in which a state corresponds to the coverage area of a BS. The total session generation rate of subscribers of the provider \( i \) is further divided among its BSs \( (k = 1, \ldots, |K_i|) \) according to the stationary distribution of this Markov chain. From the new sessions targeted to the BS \( k \), some of them may be offloaded to the WiFi APs associated with this BS. Then, the ratio of traffic that is served by each wireless station \( k \) of the provider \( i \) can be determined. Table I defines the parameters of the

\(^1\)This is a reasonable assumption given that the coverage area of a BS is much larger than that of an AP.

\(^2\)This assumption of a fair allocation of bandwidth among connected users has been also adopted in other studies of WiFi offloading (e.g., [6]).
queueing network of the provider $i$. Note that the horizontal and vertical handover rates are computed based on a fluid-flow mobility model [16]. Such models have been used to describe the user movement among stations in wireless networks. We assume that no horizontal handovers are performed between APs. Let us now focus on a simple case in which all users select the provider $i$ (i.e., $z_i = 1$). The total session arrival rate at a station $k$ ($\gamma_{ik}$) consists of the new sessions ($a_{ik}$) as well as sessions due to horizontal and vertical handovers (e.g., Fig. 2).

$$\gamma_{ik} = a_{ik} + \sum_{m \in K_i \cup L_i} \gamma_{ikm} P_{m,k}^{(i)}$$

The traffic intensity at the station $k$ of the provider $i$ ($\rho_{ik}$) is equal to the ratio of the total session arrival rate at the station $k$ ($\gamma_{ik}$) over the total session departure rate at that station ($\delta_{ik}$).

The queueing network of the provider $i$ is modeled as a Markov chain. Each state corresponds to a vector $n_i = (n_{i1}, ..., n_{i[K_i \cup L_i]})$ indicating the number of connected users at all BSs and APs of the provider $i$. State transitions correspond to various types of events including session arrivals, terminations, and handovers. The stationary distribution of the Markov chain is estimated by solving the global-balance equations. These equations set the arrival rate at each state of the Markov chain equal to the departure rate from that state. Due to the Markovian property of our system and the processor-sharing discipline, the global-balance equations can be simplified into a set of local-balance equations [17]. According to the local-balance equations (Eqs. 2), the rate leaving a state $n_i$ due to the departure of a user at a specific station $k$ is equal to the rate entering that state due to the arrival of a user at the station $k$ either due to a new session or a handover (Eq. 2a).

$$d_{ik}Q_i(n_i) = a_{ik}Q_i(n_i - e_{ik}) + \sum_{m \in K_i \cup L_i} (x_{im} + v_{im}) P_{m,k}^{(i)}Q_i(n_i - e_{ik} + e_{im})$$ (2a)

Furthermore, the rate leaving the state $n_i$ due to the arrival of a new session at a station is equal to the rate entering that state due to the termination of a session at a station (Eq. 2b).

$$\sum_{k \in K_i \cup L_i} a_{ik}Q_i(n_i) = \sum_{k \in K_i \cup L_i} \mu_{ik}Q_i(n_i + e_{ik})$$ (2b)

By substituting Eq. 3 in the local-balance equations (Eqs. 2), we can derive the traffic equations (Eq. 1). This proves the validity of Eq. 3. Given that the stationary distribution is in product form, each station can be viewed as an independent M/M/1 queue with the processor-sharing discipline.

In the general case in which not all users select the provider $i$ (i.e., $z_i < 1$), we can replace $\gamma_{ik}$, $a_{ik}$, and $\rho_{ik}$ with $\gamma_{ik}z_i$, $a_{ik}z_i$, and $\rho_{ik}z_i$, respectively and Eqs. 1-3 still hold. In this case, the average number of connected users at the station $k$ of the provider $i$ is $E[N_{ik}] = \frac{z_i \rho_{ik}}{1 - z_i \rho_{ik}}$ [18]. When a new user arrives at the station $k$, it will share the available bandwidth along with all other currently connected users at that station. Therefore, the amount of bandwidth that a new user can get when it connects to the station $k$ will be $\frac{B_{ik}}{E[N_{ik}]+1} = \frac{B_{ik}(1 - z_i \rho_{ik})}{E[N_{ik}]+1}$, where $B_{ik}$ is the total bandwidth of that station. The average data rate of a user session at the network of the provider $i$ can be now computed as the weighted average of the data rate achieved at each station (Eq. 4).

$$R_i(z_i) = \sum_{k \in K_i \cup L_i} \omega_{ik}B_{ik}(1 - z_i \rho_{ik})$$ (4)

The user service selection employs the average data rate in the decision making process (Fig. 1). The computation of the market equilibriums for users and providers is described in Subsections III-B and III-C, respectively.

### B. User service selection

The user service selection process is modeled by a population game. Each user can choose among $I + 1$ available strategies $H = \{0, 1, ..., I\}$. Strategies 1, 2, ..., $I$ correspond to
substitutions with the providers 1, 2, ..., I, respectively, while strategy 0 denotes the disconnection state. We assume that the user population is homogeneous, and as such, the utility attained when selecting a specific strategy is the same for all users. Therefore, it suffices to describe the service selection of users with a probability distribution over the set of strategies \( H \). This distribution \( z = (z_0, z_1, ..., z_I) \) is the user strategy profile also denoted as market share. All parameters of a wireless market are defined in Table II.

**User utility function.** A user selects a strategy (i.e., a subscription or disconnection) based on the QoS and price:

\[
 u_i(z_i; c) = \begin{cases} 
 f(R_i(z_i)) - w_PC_i & \text{if } i = 1, ..., I \\
 0 & \text{if } i = 0
\end{cases} 
\]  

(5)

The function \( f \) is concave, strictly increasing, and non-negative and defines the impact of the average data rate \( (R_i(z_i)) \) on the user utility. Furthermore, when a user selects the disconnection (i.e., \( i = 0 \)), it attains a utility equal to 0. Every population game has at least one Nash equilibrium (NE) as shown by applying Kakutani's fixed point theorem [19]. Furthermore, the user population game considered in this work has a unique NE. Due to lack of space, the proof has been included in our technical report [20].

**Computation of the user NE**

At a user NE, we can divide the set of strategies \( H \) into two disjoint subsets \( X \) and \( Y \), such that: (i) \( X \) is non-empty, (ii) all strategies in \( X \) correspond to the same utility, (iii) all strategies in \( Y \) correspond to a market share of 0 (proven in our technical report [20]). To compute a NE, we distinguish different cases with respect to the sets \( X \) and \( Y \).

**Case (a).** The subscriptions of all providers correspond to the same utility and there are no disconnected users: \( X = \{1, ..., I\} \) and \( Y = \{\} \).

\[
 u_i(z_i; c) = u_1(z_1; c) \quad \forall i \in \{2, ..., I\}, \quad \sum_{j=1}^{I} z_j = 1 \quad (6)
\]

For the solution of Eqs. 6 \( z^1(c) = (z_0^1(c), z_1^1(c), ..., z_I^1(c)) \) to be a NE, additional conditions should be satisfied (inequalities 7). First, \( z^1(c) \) should be a valid probability distribution and \( c \) should lie in the strategy space of providers (inequalities 7b and 7c, respectively). Furthermore, at the equilibrium, no user should have the incentive to change its strategy.

\[
 u_1(z_1^1(c); c) \geq 0 \quad (7a)
\]

\[
 z_i^1(c) \geq 0, \quad \forall i \in \{1, ..., I\} \quad (7b)
\]

\[
 c_i \geq 0, \quad \forall i \in \{1, ..., I\} \quad (7c)
\]

In general, two types of transitions may happen, namely, (i) a subscriber may change provider, and (ii) a subscriber may become disconnected. However, in this case, none of these can occur: a transition of type (i) is not profitable since all subscriptions have equal utility at the equilibrium (Eq. 6), and a transition of type (ii) reduces the user utility since all subscriptions have higher utility than the disconnection (Eqs. 6 and 7a). Therefore, when the conditions 7 are true, the solution of Eqs. 6 \( (z^1(c)) \) is the unique user NE for the price vector \( c \).

**Case (b).** All strategies, including the disconnection, correspond to the same utility: \( X = H \) and \( Y = \{\} \).

\[
 u_i(z_i; c) = 0 \quad \forall i \in \{1, ..., I\}, \quad \sum_{j=0}^{I} z_j = 1 \quad (8)
\]

For the solution of Eqs. 8 \( z^2(c) = (z_0^2(c), z_1^2(c), ..., z_I^2(c)) \) to be a NE, additional conditions should be satisfied (inequalities 9). The vector \( z^2(c) \) should be a valid probability distribution (inequalities 9a and 9b) and \( c \) should lie in the strategy space of providers (inequality 9c).

\[
 z_0^2(c) \geq 0 \quad (9a)
\]

\[
 z_i^2(c) \geq 0, \quad \forall i \in \{1, ..., I\} \quad (9b)
\]

\[
 c_i \geq 0, \quad \forall i \in \{1, ..., I\} \quad (9c)
\]

When the conditions 9 are true, the solution of Eqs. 8 \( (z^2(c)) \) is the user NE for the price vector \( c \). At that equilibrium, no user has the incentive to change its strategy, since it will not increase its utility by doing so. Except from (a) and (b), other cases can be defined in which the subscriptions of one or more providers belong in the set \( Y \) (i.e., obtain a market share of 0). However, as it will be shown in Section III-C, at the NE of providers, each provider should obtain a strictly positive market share. This can only happen in the cases (a) and (b) and therefore, all other cases are omitted from the analysis.

**C. Competition of providers**

The competition of providers is modeled as a continuous normal-form game \((P, C, \{\sigma_i\}_{i \in P})\). In this game, each provider \( i \) selects a price for its subscription \((c_i)\) belonging in a closed interval \([0, C_{max}^i]\). The strategy space of providers is the set of all possible combinations of prices that can be offered in the market and is a rectangle of the form \( C = [0, C_{max}^1] \times ... \times [0, C_{max}^I] \). Each point of the strategy space \( c = (c_1, ..., c_I) \) is a vector containing a specific price for each provider and corresponds to a unique user NE \( z^*(c) = (z_{0}^*(c), z_1^*(c), ..., z_I^*(c)) \). Based on this equilibrium, the utility function of a provider \( i \) is defined as \( \sigma_i(c) = \sum_{j=1}^{I} z_j^*(c) c_i \) and estimates the total revenue of the provider \( i \) in the market.

In general, continuous games can be analysed efficiently provided that they have a rectangular strategy space and twice continuously differentiable utility functions [21, 22]. However, in our case, there exist a finite set of surfaces in the strategy space, at which, the derivatives of the utility functions of providers are discontinuous. Those surfaces divide the strategy space into a finite number of regions. At the interior of each region, the set of user strategies that obtain

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>Number of providers</td>
</tr>
<tr>
<td>( N )</td>
<td>Number of users</td>
</tr>
<tr>
<td>( c )</td>
<td>Vector with the prices of all providers</td>
</tr>
<tr>
<td>( H )</td>
<td>Set of user strategies</td>
</tr>
<tr>
<td>( I(R_i(z_i)) )</td>
<td>Impact of average data rate on user utility</td>
</tr>
<tr>
<td>( w_P )</td>
<td>Weight of price</td>
</tr>
<tr>
<td>( u_i(z_i; c) )</td>
<td>User utility function</td>
</tr>
<tr>
<td>( z_i(c) )</td>
<td>User NE</td>
</tr>
<tr>
<td>( \sigma_i(c) )</td>
<td>Utility function of provider ( i )</td>
</tr>
<tr>
<td>( \sigma_i^c(c) )</td>
<td>Utility function of provider ( i ) restricted in the region ( r ) of the strategy space of providers</td>
</tr>
<tr>
<td>( g_j(c) \leq 0 )</td>
<td>j-th constraint used to define the region ( r ) of the strategy space of providers</td>
</tr>
<tr>
<td>( A_r )</td>
<td>Set of price vectors corresponding to the region ( r ) of the strategy space of providers</td>
</tr>
</tbody>
</table>

**TABLE II: Main parameters of a wireless market**
a strictly positive market share at the user NE is fixed. Fig. 3 depicts two examples of the strategy space of providers in a simple case of a duopoly under large and small user traffic demand (Figs. 3a and 3b, respectively). These figures have been constructed by computing the user NE over a set of prices. At the interior of each region, all strategies that correspond to a strictly positive market share at the user NE are listed. The region 1 (region 2) is the set of price vectors that satisfy the conditions of case (a) (case (b)) of Section II-B, respectively. The region 2 is larger in Fig. 3a compared to Fig. 3b, since, as the traffic demand increases, the average data rate drops along with the user utility resulting in more disconnected users. In the case of low traffic demand (Fig. 3b), there are two additional regions (the regions 6 and 7) where all users become subscribers of one provider.

The segmentation of the strategy space appears in markets with multiple providers each offering a unique service for a price and with users that make rational decisions. In such markets, a NE can be narrowed down at the interiors of the regions 1 and 2 and at the surface that separates those two regions. This can be easily proven by contradiction. If a NE existed outside those sets of points, then at least one provider would obtain a market share of 0. However, this provider would have the incentive to reduce its price and obtain a strictly positive market share and revenue. This contradicts the definition of a NE. For example, if a NE existed in the region 3 of Fig. 3a (e.g., at the point \(X\)), then the provider 2 would obtain a market share of 0. However, this provider would have this incentive to reduce its price to obtain a strictly positive market share and revenue.

To compute the NEs of providers, we propose a novel algorithm (illustrated in Fig. 4). First, the strategy space is split into the different regions. Then, two separate games are defined for the regions 1 and 2. The problem of computing the NEs of a game with its strategy space restricted in a single region is a generalized Nash equilibrium problem (GNEP) [23], [24]. The final step checks whether the NEs corresponding to the regions 1 and 2 are also global NEs of the game of providers. Let us now describe the algorithm in more detail.

**Computation of NEs in the region 1.** At a NE, the price of a provider is a best response to the prices of its competitors. To compute its best response, a provider solves an optimization problem to select the price that maximizes its revenue. To restrict the solution of this problem in the region 1, we add the constraints defined by the inequalities 7. As it will be shown bellow, all NEs lie either in the interior of the region 1 or at the set of points at which the constraint 7a is active. The local maxima of the utility functions of providers at these sets of points satisfy the linear independence constraint qualification (LICQ)\(^2\), and thus, at a NE, the Karush-Kuhn-Tucker (KKT) conditions of the optimization problems of individual providers should be satisfied [25]. Table III defines a system that combines the KKT conditions of these problems. The formulation of this system is described in our technical report [20].

To compute a NE, we distinguish various cases with respect to the location of that equilibrium: A point at which at least

---

1. Split strategy space
2. Compute the NEs of providers in region 2 (GNEP)
   2a. Compute the NEs of providers in region 1 (GNEP)
   3. Compute global NEs

Fig. 4: The main idea for computing the NEs of providers.

---

\(^1\)A constraint \(g(c) \leq 0\) is active when the equality holds.

\(^2\)Consider a local maximum \(x^*\) of an optimization problem with continuously differentiable objective and constraint functions. If the gradients of the active inequality constraints and the gradients of the equality constraints are linearly independent at \(x^*\), the KKT conditions should be satisfied at \(x^*\).
TABLE III: Combined KKT conditions of the optimization problems of providers for computing their best response

<table>
<thead>
<tr>
<th>Condition name</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equate to 0 the derivative of the Lagrange fun.</td>
<td>$-\frac{\partial \sigma^1_i(c)}{\partial c_i} + \sum_{j=1}^{2I+1} \lambda_{ij} \frac{\partial g^j_i(c)}{\partial c_i} = 0, \ i = 1, \ldots, I$ (10a)</td>
<td>The term $\sigma^1_i(c)$ corresponds to the utility function of the provider $i$ restricted in the region 1.</td>
</tr>
<tr>
<td>Primal feasibility</td>
<td>$g^j_i(c) \leq 0, \ j = 1, \ldots, 2I+1$ (10b)</td>
<td>Define the region 1: $g^j_i(c) \leq 0$ corresponds to the constraint 7a. $g^j_{i+1}(c), \ldots, g^j_{2I+1}(c) \leq 0$ correspond to the constraints 7b.</td>
</tr>
<tr>
<td>Dual feasibility</td>
<td>$\lambda_{ij} \geq 0, \ i = 1, \ldots, I$ and $j = 1, \ldots, 2I+1$ (10c)</td>
<td>Set the Lagrange multipliers greater than or equal to 0.</td>
</tr>
<tr>
<td>Complementary slackness</td>
<td>$\lambda_{ij} g^j_i(c) = 0, \ i = 1, \ldots, I$ and $j = 1, \ldots, 2I+1$ (10d)</td>
<td>Set the product of the Lagrange multipliers with their corresponding constraints equal to 0.</td>
</tr>
</tbody>
</table>

One of the constraints $7c$ is active can not be a NE. At such a point, there is always a provider with price equal to 0 that has the incentive to increase its price and attain a strictly positive utility (e.g., point $D$ in Fig. 3b). Similarly, a point at which a constraint 7b is active can not be a NE. At such a point, there is always a provider with 0 market share that has the incentive to reduce its price and attain a strictly positive utility (e.g., point $E$ in Fig. 3b). Therefore, all NEs lie either in the interior of the region 1 or at the set of points at which only the constraint 7a is active. This set of points corresponds to the surface separating the regions 1 and 2 (e.g., Fig. 3b). In the interior of the region 1, all inequalities 10b are strict and therefore, based on the complementary slackness KKT conditions (Eq. 10d), all Lagrange multipliers are equal to 0. This reduces the system 10 into the following system:

$$\frac{\partial \sigma^1_i(c)}{\partial c_i} = 0, \text{ for all } i = 1, \ldots, I$$ (11)

Standard numerical analysis methods are used to solve the above system. If the solution satisfies the constraints 10b (i.e., the inequalities 7) and corresponds to a global maximum of the utility functions of providers, it is a NE. Let us now focus on the set of points at which only the constraint 7a is active. Based on the complementary slackness KKT conditions (Eq. 10d) and the fact that all other constraints except 7a are not active, we derive that all Lagrange multipliers are equal to 0 except those corresponding to the constraint 7a. This reduces the system 10 into the following system:

$$\frac{\partial \sigma^1_i(c)}{\partial c_i} - \lambda_{i1} \frac{\partial g^1_i(c)}{\partial c_i} = 0, \ i = 1, \ldots, I$$ (12a)

$$\lambda_{i1} \geq 0, \ i = 1, \ldots, I$$ (12b)

The derivative $\frac{\partial \sigma^1_i(c)}{\partial c_i}$ is always positive, for all mathematical models of the user utility function considered in this work. Therefore, the system of Eqs. 12 is reduced to the following inequalities:

$$\frac{\partial \sigma^1_i(c)}{\partial c_i} \geq 0, \ i = 1, \ldots, I$$ (13)

This system of inequalities restricted at the points in which the constraint 7a is active corresponds to a feasibility problem that can be solved efficiently. Such a problem may have uncountable solutions, and therefore, the set of NEs on the surface separating the regions 1 and 2 may be infinite.

**Computation of NEs in the region 2.** To compute the NEs of the game of providers in the region 2, we follow a similar procedure. The region 2 is the set of price vectors that satisfy the constraints 9. As in the case of the region 1, a point at which at least one of the constraints $9c$ or $9b$ is active can not be a NE (e.g., points $D$ and $E$ in Fig. 3a). Therefore, the NEs could either lie in the interior of the region 2 or at the set points at which the constraint 9a is active. To search for a NE in the interior of the region 2, the following system of equations should be solved:

$$\frac{\partial \sigma^2_i(c)}{\partial c_i} = 0, \text{ for all } i = 1, \ldots, I$$ (14)

The term $\sigma^2_i(c)$ corresponds to the utility function of the provider $i$ restricted in the region 2. If the solution of Eq. 14 satisfies the constraints 9, then it is a NE of the game of providers restricted in the region 2. Furthermore, a point $c$ at which the constraint 9a is active is a NE if the following conditions hold.

$$\frac{\partial \sigma^2_i(c)}{\partial c_i} \leq 0, \ i = 1, \ldots, I$$ (15)

Again, solving the inequalities 15 restricted at the points at which the constraint 9a is active corresponds to a feasibility problem with potentially uncountable solutions.

**Computation of global NEs.** Let us denote the sets of price vectors corresponding to the regions 1 and 2 as $A_1$ and $A_2$, respectively. The games restricted in these regions can be then defined as $\Gamma_1 = \{P, A_1, \{\sigma^1_i\}_{i \in P}\}$ and $\Gamma_2 = \{P, A_2, \{\sigma^2_i\}_{i \in P}\}$, respectively. A more general game $\Gamma = \{P, A_1 \cup A_2, \{\sigma_i\}_{i \in P}\}$ that is restricted on the union of the regions 1 and 2 can now be formed. The following set of theorems proven in our technical report [20] relate the NEs of the game $\Gamma$ with the NEs of the games $\Gamma_1$ and $\Gamma_2$.

**Theorem 1:** A point $c^* \in A_1 \cap A_2$ is a NE of the game $\Gamma$, if and only if, it is a NE of the games $\Gamma_1$ and $\Gamma_2$.

**Theorem 2:** A point $c^* \in A_1 \setminus A_2$ is a NE of the game $\Gamma$, if and only if, it is a NE of the game $\Gamma_1$ and the following conditions are true.

$$\sigma^1_i(c^*_i, c^*_{-i}) \geq \sigma^2_i(c_i, c^*_{-i}), \forall c_i : (c_i, c^*_{-i}) \in A_2, \forall i \in P$$ (16)

**Theorem 3:** A point $c^* \in A_2 \setminus A_1$ is a NE of the game $\Gamma$, if and only if, it is a NE of the game $\Gamma_2$ and the following

---

3The derivatives of the provider utility functions are computed numerically according to the method described in our technical report [20].
conditions are true.

\[ \sigma^2_i(c^*_i, c^{* -}_i) \geq \sigma^1_i(c^*_i, c^{* -}_i), \ \forall c_i : (c_i, c^{* -}_i) \in A_1, \forall i \in P \]  

(17)

In the inequalities 16 and 17, the point \( c^* \) is also denoted as \( (c^*_i, c^{* -}_i) \), where \( c^*_i \) is the price of the provider \( i \) and \( c^{* -}_i \) is a vector containing the prices of all other providers except \( i \), at \( c^* \). Theorem 2 implies that if there exists a NE in the interior of the region 1 (i.e., solution of Eqs. 11) and if the conditions of NE hold for points lying in the region 2 (inequalities 16), then it is also a global NE. For example, in Fig. 4, a NE in the interior of the region 1 is global if conditions of NE hold for the points on the dotted lines in the region 2. These dotted lines correspond to the points \((c_i, c^{* -}_i) \in A_2\) considered in the inequalities 16. Similarly, Theorem 3 implies that if there exists a NE in the interior of the region 2 (i.e., provided by Eqs. 14), then it is also a global NE if the conditions of inequalities 17 are satisfied. Furthermore, according to the Theorem 1, the set of NEs at the surface separating the regions 1 and 2 is the intersection of the sets of NEs of the games restricted in the regions 1 and 2, respectively (Fig. 4).

1) An algorithm for the computation of a NE: The KKT system defined in Table III is a set of necessary conditions for a point to be a NE of the game of providers restricted in the region 1. These conditions are also sufficient only if the utility functions of providers are concave in their prices [26]. In such a case, by applying the methodology described above, we are guaranteed to compute a global NE if one exists. However, there are scenarios in which the utility functions of providers are not concave in the region 1. In these cases, while the KKT conditions of Table III are still necessary for a point to be a NE, there are not sufficient. Therefore, when computing a solution of these conditions, we should verify if it is a NE.

Our algorithm proceeds as follows: First, it attempts to compute a NE at the interior of the region 2 by solving Eqs. 14 and checking the conditions 17. If a global NE is computed, the algorithm returns it; otherwise, it attempts to compute a NE at the interior of the region 1 by solving Eqs. 11. If a solution is computed, the algorithm verifies whether it corresponds to a global maximum of the utility functions of providers. In such a case, it is a global NE and is returned. If it corresponds to a local maximum of the utility functions of providers, the algorithm reports it as a “local NE”. Finally, if the solution corresponds to a local minimum for at least one of the providers or if no solution was computed for Eqs. 11, the algorithm searches for a NE at the surface separating the regions 1 and 2 by solving the inequalities 13 and 15. Note that if the utility functions of providers are not concave at the interior of the region 1, there may be scenarios in which there is no pure-strategy NE of the game of providers. In such cases, our algorithm will not report a NE.

IV. PERFORMANCE EVALUATION

A. Experiment settings and objectives

We implemented the modeling framework in Matlab and instantiated a wireless oligopoly of a small city of 180 \( km^2 \) with 4 providers and a population of 300,000 users.

Network infrastructure. Each provider has deployed a cellular network covering the entire city. The BSs at each network are placed on the sites of a triangular grid, with a distance between two neighbouring sites of 1.6 km. The city is divided into 9 equally sized rectangular areas. We assume that a WiFi AP infrastructure has been deployed at different areas in the city. A provider with access to the APs of a specific area may offload a part of its mobile data traffic to these APs. A number of the new sessions that are generated within the coverage of an AP (say AP \( q \)) may be served by that AP. Furthermore, when a user moves into the coverage of that AP during a session, a vertical handover can be performed to that AP. From all the above sessions, a provider offloads to the AP \( q \) a certain percentage denoted as offloading percentage. Note that the offloading percentage is the same for all APs of a provider. When a user who is being served by an AP moves out of the coverage of that AP, a vertical handover is performed back to the cellular network of its provider. The goal of a provider is to select the optimal offloading percentage that will maximize its revenue in the market. Apart from the traffic of cellular BSs, the WiFi APs may also serve their own WiFi customers. This reduces the effective capacity of APs which is available for offloading. We have performed an extensive set of experiments in which the optimal offloading percentage of a provider was estimated under different cases with respect to: (a) the bandwidth of cellular BSs, (b) the coverage of the WiFi network, and (c) the user utility function.

Bandwidth of wireless stations. We distinguish two scenarios with respect to the capabilities of the BSs and APs: the 3G scenario with BSs of lower bandwidth than that of APs (e.g., [4]) as well as the LTE scenario with BSs that have bandwidth larger than that of WiFi APs. In this analysis, the bandwidth of an AP is 6 Mbps. In 3G, the maximum data rate with which a BS can serve sessions is 5, 4.5, 4, and 3.5 Mbps for the providers 1, 2, 3, and 4, respectively, while in LTE, the maximum data rates of providers are 25, 22, 19, and 16 Mbps, respectively.

Wi-Fi infrastructure. We assumed that in the city of interest, there exists a WiFi infrastructure. In our experiments, either the strongest provider (provider 1) or the weakest one (provider 4) may perform offloading. With respect to the WiFi coverage, we have considered cases in which APs are located in one, two, or three different areas close to the city center. We have also defined different cases in which the number of APs that correspond to a single BS varies from 1 up to 15. It is assumed that the APs associated with a BS are located within its cell and their coverage areas do not overlap. 16 APs are required to cover a cell, and therefore, the larger the number of APs per BS, the larger the WiFi coverage. Furthermore, an AP serves sessions from its own WiFi customers with a total arrival rate of 2 sessions per minute. Due to the traffic demand of WiFi customers, the effective bandwidth of APs that can be used for offloading is reduced from 6 to around 3.3 Mbps.

User utility function. We assumed exponential user utility functions (Eq. 5) that depend on the average data rate \( f(x) = w_R (t - e^{-hx}) \), where \( w_R \), \( t \) and \( h \) are equal to 30, 1, and 0.6, respectively. The exponential utility function has a diminishing derivative with respect to the data rate. For each case, we computed the market equilibria given that no offloading is performed, and then, we computed the optimal offloading percentage for a provider, its additional revenue and the decrease in disconnected users.

B. The LTE scenario

We evaluated the performance of the WiFi offloading in the case of LTE. In this scenario, the average traffic demand of a user varied from 0.9 up to 1.5 sessions/hour. Fig. 5 presents the
benefits of WiFi offloading. In general, as the number of APs associated with a BS increases, the additional revenue from the offloading also increases but in a diminishing manner (Figs. 5a and 5d). By increasing the number of APs per BS up to 6, the optimal offloading percentage is equal to 100% (Figs. 5b and 5e) which results in an increased revenue. However, an increase of the number of APs above this threshold has a diminishing “return” due to the decrease of the optimal offloading percentage (Figs. 5b and 5e). In other words, it is profitable for a provider to invest in WiFi infrastructure up to a certain threshold (e.g., around 6 APs per BS in this scenario). Above this threshold, the investment would not be beneficial. Under large WiFi coverage, the low optimal offloading percentage (e.g., Figs. 5b and 5e) is due to two opposing trends: The increased offloading percentage alleviates the congestion at BSs allowing for higher data rates for users. On the other hand, it also results in a larger number of users being served by APs, which have lower bandwidth compared to that of BSs. Therefore, the data rate of these users decreases. In general, it is not always profitable for a provider to offload all traffic that can be served by APs. The optimal offloading percentage achieves the best load balancing among APs and BSs. It is selected in such a way to improve the average data rate, attract more users, and maximize the revenue of the provider. As expected, an increase in the number of APs per BS results in a larger decrease in the percentage of disconnected users (Figs. 5c and 5f). Furthermore, when the number of areas with a WiFi infrastructure expands, the revenue of a provider grows and the percentage of disconnected users drops. Surprisingly, it also results in a slight decrease of the optimal offloading percentage (Figs. 5b and 5e). This is due to the phenomenon discussed earlier. In general, an increase in the WiFi coverage results in a decrease of the optimal offloading percentage. Let us now discuss the case in which the provider 4 performs the offloading. Although the percentage increase of its revenue is higher compared to that of the provider 1 (Figs. 5a and 5d), the corresponding absolute increase of its revenue and the reduction of disconnected users are lower compared to the provider 1 (Figs. 5c and 5f). Therefore, from the perspective of the overall market performance, it is more beneficial for the provider with the largest capacity to perform the offloading.

C. The 3G scenario

We repeated the analysis of Section IV-B for the 3G scenario. In this scenario, the traffic demand of a user varied from 0.15 up to 0.25 sessions/hour. Fig. 6 presents the results. The percentage increase of the revenue of providers is higher in the 3G scenario compared to the LTE one (Fig. 5). There is also a more prominent decrease in the percentage of disconnected users. The bandwidth of BSs in the 3G scenario is closer to the one of APs compared to the LTE scenario increasing the benefits of the offloading. The larger effective bandwidth of the BSs of the provider 1 (i.e., 5 Mbps) compared to that of APs (i.e., 3.3 Mbps) results in a reduction of the optimal offloading percentage of the provider 1 above a certain threshold (Fig. 6b). On the contrary, the bandwidth of the BSs of the provider 4 is very close to that of APs (i.e., 3.5 Mbps) resulting in an optimal offloading percentage equal to 100% for almost all cases (Fig. 6e). In general, the closer the effective bandwidth of APs to that of BSs, the more traffic the BSs can offload to these APs without negatively affecting the QoS.

V. NETWORK AGGREGATION

In the case of a provider with access to a WiFi infrastructure, the provider should estimate the optimal percentage of traffic to offload to this infrastructure to maximize its revenue. The estimation of the optimal offloading percentage is a computationally complex process. It requires to solve the traffic equations (Eq. 1) and compute the NEs of users and providers multiple times, each time for a different value of the offloading percentage. In such cases, to reduce the computational complexity, a queueing network aggregation methodology based on the theorem of Norton [8] can be
Fig. 6: Benefits of offloading in the 3G scenario. Only the provider 1 (4) performs offloading in top (bottom) figures, respectively.

applied. For the case with offloading percentage equal to 100%, this methodology constructs two separate queueing networks, one for the region of interest with the WiFi infrastructure and one for the remaining area. By solving the traffic equations of those two networks, we can compute the average data rate (according to Eq. 4). When the offloading percentage changes, we modify the queueing network of the region of interest accordingly. Instead of solving the traffic equations of the entire network of the provider, we only solve the traffic equations corresponding to the queueing network of the region of interest and keep the queueing network of the remaining area unchanged. This results in significant computational gains. Bellow this procedure is described in more detail.

1. Construct the equivalent queueing network of the region of interest. Consider the network of a provider and let \( s = \{M_1, ..., M_n\} \) be the subset that contains the stations of the region of interest and \( \hat{s} = \{1, ..., K\} \) be the subset that contains all stations in the remaining area. The first step of the algorithm constructs a reduced queueing network of the remaining area in which all stations of the region of interest are removed one by one by a “shortening” process. Each time a station is removed, the service rate of its corresponding queue is set to equal to infinity (i.e., \( \gamma_{tk} \)). To estimate the average data rate, we modify the traffic from the remaining area has a rate of \( \sum_{k \in \hat{s}} \gamma_{tk} p_k M_t \). An example of the construction of the reduced network is shown in Fig. 7. At the first phase, the station 1 is removed (in the middle). The new transition and session arrival rates \( (1) \), \( (1) \) are computed based on Eqs. 18 and 19. At the second phase, the station 2 is removed and the reduced network is formed (at the right).

The construction of the equivalent queueing network of the region of interest is performed as follows: A subgraph of the original queueing network of the provider is selected that contains only the BSs and APs of the region of interest. Then, at each station of the region of interest, the corresponding Poisson source that models the input traffic from the remaining area (computed earlier) is added to the total input traffic of that station.

2. Construct the equivalent queueing network of the remaining area. The procedure is exactly the same as the one presented in the previous paragraph for the construction of the equivalent queueing network of the region of interest. The only difference is that the sets \( s \) and \( \hat{s} \) are interchanged.

When the value of the offloading percentage \( (p_o) \) becomes lower than 100%, we modify the equivalent queueing network of the region of interest accordingly and keep the equivalent queueing network of the remaining area unchanged. Specifically, the new session arrival rate of each AP \( q \) in the region of interest is reduced \( a_{0q}^2 = a_{0q} p_o / 100 \) and the subtracted session arrival rate (i.e., \( a_{0q} (1 - p_o / 100) \)) is added back to the BS associated with the AP \( q \). The vertical handover rate of each station \( m \) in the region of interest is also reduced \( x_{im}^n = x_{im} p_o / 100 \) and the corresponding handover probabilities change accordingly. To estimate the average data rate, we solve the traffic equations of the modified queueing network of the region of interest and keep the traffic intensities...
corresponding to the stations of the remaining area unchanged. The above process significantly reduces the computational complexity of estimating the average data rate compared to solving the traffic equations of the entire queuing network of the provider but introduces an error. This error is due to the assumption that the queuing network of the remaining area is completely unaffected by the change of the offloading percentage. Our experiments indicate that in a scenario in which the WiFi infrastructure covers 22% of the total area of the network and each BS is associated with 9 APs, this process reduces the execution time of the solution of the traffic equations by 94% with a maximum mean absolute error of 0.38 Mbps for the estimated average data rate (considering all values of the offloading percentage).

VI. CONCLUSIONS AND FUTURE WORK

Our framework provides detailed models for several economic and technological aspects of WiFi offloading in wireless markets. In contrast to other approaches, it models the effect of offloading on the perceived QoS by users, the competition of providers, and user decisions. This work solved several technical difficulties which arise due to the detailed modeling of the markets (e.g., various discontinuities appear in the derivatives of the utility functions of providers) and proposed a novel algorithm that computes the NEs of providers under such discontinuities. Moreover, to reduce the computational complexity of the analysis of large-scale markets, it also proposed an innovative queuing network aggregation algorithm based on the theorem of Norton.

Based on our framework, we performed an extensive evaluation of the WiFi offloading under various scenarios. Our results highlight the benefits of the WiFi offloading to users and providers. It is not always beneficial for providers to invest in a large number of APs. The additional revenue of a provider from the offloading increases with the WiFi coverage but with a diminishing rate. Our framework can be extended to enable providers to design their business plan for offloading. Specifically, it can incorporate the capital (e.g., investment for WiFi, backhaul equipments, installation) and operational (e.g., WiFi and backhaul site rentals, maintenance) expenditures for supporting the offloading. Based on that, it can perform a cost-benefit analysis to estimate whether offloading is profitable.

In our ongoing work, the modeling framework has been generalized to model heterogeneous user populations with multiple customer segments, each with its own utility function, characteristics, and preferences. Our framework can be also extended to study a variety of business cases in wireless markets including, partnerships between providers, MVNOs, and femtocells, and the problem of capacity planning. This work sets a methodological basis for performing such studies.

REFERENCES