Economics of WiFi Offloading: Trading Delay for Cellular Capacity

Joohyun Lee†, Yung Yi†, Song Chong†, and Youngmi Jin†

***Abstract*—Cellular networks are facing severe traffic overloads due to the proliferation of smart handheld devices and traffic-hungry applications. A cost-effective and practical solution is to offload cellular data through WiFi. Recent theoretical and experimental studies show that a scheme, referred to as delayed WiFi offloading, can significantly save the cellular capacity by delaying users' data and exploiting mobility and thus increasing chance of meeting WiFi APs (Access Points). Despite a huge potential of WiFi offloading in alleviating mobile data explosion, its success largely depends on the economic incentives provided to users and network providers to deploy and use delayed offloading. In this paper, we study how much economic benefits can be generated due to delayed WiFi offloading, by modeling the interaction between a single provider and users based on a two-stage sequential game. We first analytically prove that WiFi offloading is economically beneficial for both the provider and users. Also, we conduct trace-driven numerical analysis to quantify the practical gain, where the increase ranges from 21 to 152% in the provider's revenue, and from 73 to 319% in the users' surplus.**

I. Introduction

Mobile data traffic is growing enormously, as smart phones/pads equipped with high computing powers and diverse applications are becoming popular. Cisco reported that global mobile data traffic grew 2.3-fold in 2011, more than doubling for the four consecutive years since 2008 [1]. It was also forecast there that the total global mobile data traffic will increase 18-fold between 2011 and 2016, where the average smartphone is projected to generate 1.3 GB per month in 2015 [1]. To cope with such mobile data explosion, upgrading to 4G (e.g., LTE (Long Term Evolution) or WiMAX), may be an immediate solution, but mobile applications are becoming more diverse with larger data consumption and the number of smartphone/pad users are also increasing rapidly. Then, users' traffic demand is expected to exceed the capacity of 4G in the near future, and thus mobile network providers keep seeking other alternatives to efficiently respond to data explosion [2].

WiFi offloading, where users use WiFi prior to 3G/4G whenever they have data to transmit/receive, has been receiving a lot of interest as a practical solution that can be applied without much financial burden in practice. Network providers as well as users can easily and quickly install WiFi access points (APs) with low costs, and in fact many network providers worldwide have already provided WiFi services in hot-spots and residential areas. Recent papers [3]–[5] demonstrate that a huge portion of cellular traffic can be offloaded to WiFi by letting users delay their delay-tolerant data (e.g., movie, software downloads, cloud services), and upload/download data whenever they meet a WiFi AP within a pre-specified delay deadline. We call this *delayed WiFi offloading*, and about 60-80% of cellular traffic can be reduced when 30 mins to 1 hour delay for human mobility [3] and 10 mins of delay for vehicular mobility [4] are allowed. Also, more than 80% of news can be pre-fetched within 700 seconds on a random mobility model in [5]. This remarkable offloading efficiency is due to users' mobility enabling themselves to be under a WiFi AP during a considerable portion of their business time.

However, WiFi offloading's high potential does not always guarantee that users and providers actually adopt it in practice. First, users may be reluctant to delay their traffic without economic incentives, e.g., discounted service fees. For example, if a user pays based on an unlimited data plan, users may have no reason to delay traffic unless WiFi-required services, e.g., services requiring higher bandwidths, are necessary. Also, providers may not always welcome delayed offloading service, since the total cellular traffic to charge may decrease, possibly leading to its revenue reduction. Thus, it is of significant importance to formally address the question on the economic gains of delayed WiFi offloading from the perspective of providers and users, which is the focus of this paper.

In this paper, we model the interaction between a single provider and users based on a two-stage sequential game, where the provider controls the price and users are price-takers. A variety of control knobs will show different economic impacts of delayed WiFi offloading. Our major focus is to understand how and how much users and the provider obtain the economic incentives by adopting delayed WiFi offloading and investigate the effect of different pricing and delay-tolerance. The major features of our model include four different pricing schemes (flat, volume, two-tier, and congestion) and heterogeneous users in terms of traffic demands and willingness to pay.

Using the game model mentioned above, we first conduct analytical studies under flat and volume pricing for the simple cases when the traffic demand follows a certain distribution (obtained from known measurement studies), and users are uniformly distributed among cells. This simplification seems to be unavoidable for mathematical tractability, yet we are able to fundamentally understand how delayed offloading becomes economically beneficial. We formally prove that delayed WiFi offloading indeed generates the economic incentives for the provider as well as users. To obtain more practical messages and quantify the gain of delayed WiFi offloading, we use two

†The authors are with Electrical Engineering, KAIST (Korea Advanced Institute of Science and Technology), e-mails: jhlee@netsys.kaist.ac.kr, {yiyung, songchong}@kaist.edu, youngmi_jin@kaist.ac.kr

This research was supported by the KCC, Korea, under the R&D program supervised by the KCA (KCA-2013-(11-911-05-004)).

978-1-4673-1017-8/13/$31.00 ©2013 IEEE

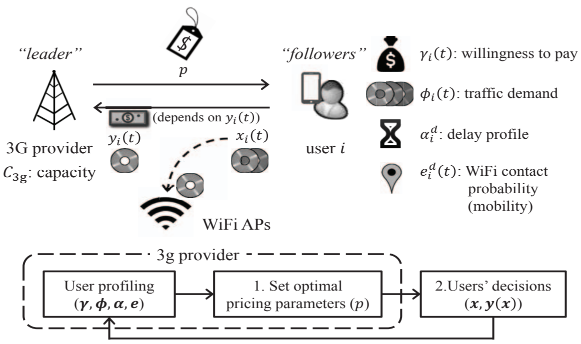


Fig. 1. An illustration of the system model. $x_i(t)$ and $y_i(t)$ are 3G+WiFi and 3G traffic volumes of user i at time t. Note that $x_i(t) \leq \phi_i(t)$.

traces, each of which tells us the information on cellular data usage and WiFi connectivity. We extract the parameters needed by our model from those traces, and obtain numerical results by running a trace-driven simulation.

II. MODEL

We summarize the system model in Fig. 1. We model users with four attributes, *(i)* how much money they can pay (*willingness to pay*, $\boldsymbol{\gamma}$), *(ii)* how much data they want to use (*traffic demand*, $\boldsymbol{\phi}$), *(iii)* how delay-tolerant each user is (*delay profile*, $\boldsymbol{\alpha}$), and *(iv)* how they move (*WiFi contact probability*, $\boldsymbol{e}$). We index users by i, time slots by t, and deadlines by d. Assuming that the provider knows users' attributes and strategies a priori, we model a market based on a two-stage sequential game (e.g. Stackelberg). At the first stage, the provider decides on the pricing parameters, p (for a fixed pricing scheme, which we will describe later), as a *leader*, and at the second stage, each user is a price-taker as a *follower* and chooses the 3G+WiFi traffic volume $\boldsymbol{x}$. Our analysis and numerical results are carried out based on the equilibrium of this game.

A. Network and Traffic Model

1) Network model: We consider a network consisting of cellular base stations (BSs) and WiFi APs, where N users are served by the cellular provider.[1] Users are always guaranteed to be under the coverage of a cellular BS, but not necessarily of a WiFi AP. We consider a one-day time scale whose average analysis over the unit billing cycle, e.g., one month, is presented. A day is divided into equal time slots $t \in \{1, 2, \ldots, T\}$, where T is the last index of one day, depending on the duration of a time slot.[2] Let C_{3g} be the capacity (in volume per slot) provided by a BS.[3] During each day, users move among BSs as well as APs. Let $e_i^d(t)$ be the probability that user i meets any WiFi AP within deadline d at time slot t. For instance, $e_i^{1\text{hour}}(13{:}00) = 0.7$ means that user i meets a WiFi AP from 1 p.m. to 2 p.m. with probability 0.7. The value of $e_i^d(t)$ can be obtained by analyzing user i's mobility trace during, say, a month. We assume that only 3G traffic is charged. Each user has its own set of accessible, free WiFi APs, e.g., ones in home, office, or hotspots, deployed by users, users' companies, providers, or governments. We ignore the cost from offloaded data since the cost of accessing the Internet via a WiFi AP connected to a wired network is considerably lower than that for accessing the cellular network [6].

2) Traffic model: We assume that user i has the average daily traffic demand Φ_i, 3G+WiFi traffic vector $\boldsymbol{x}_i = (x_i(t) : t \in T)$, and 3G traffic vector $\boldsymbol{y}_i(\boldsymbol{x}_i) = (y_i(t) : t \in T)$. The daily traffic demand Φ_i is temporally split into $\boldsymbol{\phi}_i = (\phi_i(t) : t \in T)$, where $\phi_i(t)$ is the traffic demand at slot t, and $\Phi_i = \sum_{t \in T} \phi_i(t)$. We denote $w_i(t) = \phi_i(t)/\Phi_i$ as temporal preference (weight) of user i. The traffic volume of user i generated at slot t is constrained by the traffic demand, i.e., $x_i(t) \leq \phi_i(t)$. Users may not be able to deliver all traffic demand, and the actual transmitted volume depends on the price and user utility, which we will describe later. The traffic volume transfered through 3G, $\boldsymbol{y}_i(\boldsymbol{x}_i)$ which is actually charged, relies on $\boldsymbol{x}_i$ as well as user i's mobility and delay profile, which we define in the following.

We introduce a notion of *delay profile* to model per-user delay-tolerance of traffic. The delay profile is denoted by $\boldsymbol{\alpha} = (\alpha_i^d : i \in N, d \in \{0, 1, ..., D\})$ such that $\sum_{d=0}^{D} \alpha_i^d = 1$, where α_i^d is the portion of user i's traffic demand that allows deadline d, and D is the maximum allowable deadline across all traffic. For example, for a user i's traffic demand 1 GB, if the user has 300 MB and 700 MB, allowing 10 mins and 1 hour, resp., we have $\alpha_i^{10\text{m}} = 0.3$, $\alpha_i^{1\text{h}} = 0.7$. For a given per-user delay profile, each user uses only WiFi connections to deliver some data until the allowable deadline expires, after which the remaining data is immediately transferred through 3G. In particular, when no delay is allowed ($\alpha_i^0 = 1$), we call this regime *on-the-spot* offloading, where a user only uses spontaneous connectivity of WiFi. Most current smartphones support this by default.

B. Market Model

We start by explaining the economic metrics of the users and the provider. We assume that the provider and users are rational and try to maximize revenue and net-utility.

1) Users and Provider: We model heterogeneous willingness to pay among users over time slots, which we denote by $\boldsymbol{\gamma}_i = (\gamma_i(t) : t \in T)$. For an average user i, $\gamma_i(t)$ tends to be higher when t is in daytime. We first define user i's utility at time slot t by $\gamma_i(t) x_i(t)^\theta$, where the constant $\theta \in (0, 1)$ is price-sensitivity. The utility function $\gamma_i(t) x_i(t)^\theta$ is called an *iso-elastic* function[4] with the property of an increasing function of traffic volume $x_i(t)$ for all i and t, but of a decreasing marginal payoff. Then, user i's net-utility during a day is:

$$U_i(\boldsymbol{x}_i) = \sum_{t \in T} \gamma_i(t) x_i(t)^\theta - m\big(p, \boldsymbol{y}_i(\boldsymbol{x}_i)\big),$$

where $m(p, \boldsymbol{y}_i(\boldsymbol{x}_i))$ is the daily payment charged by the provider whose price is p. We abuse the notation and use p to refer to the whole parameters of a given pricing scheme, and the function form of m differs across pricing schemes.

[1] Throughout this paper, we use the words 'BS' and 'AP' to refer to a cellular BS and a WiFi AP, respectively.

[2] We also use N and T to refer to a set of all users and time slots to abuse the notation.

[3] We use '3G' to refer to a cellular network (e.g., W-CDMA, LTE).

[4] A function $u(x)$ is said to be *iso-elastic* if for all $k > 0$, $u(kx) = f(k)u(x) + g(k)$ for some functions $f(k), g(k) > 0$.

Given the traffic demand $\boldsymbol{\phi}_i$, mobility pattern $e_i^d(t)$, willingness to pay $\boldsymbol{\gamma}_i$, delay profile α_i^d, and a pricing function, each user i chooses $\boldsymbol{x}_i^\star$ to maximize his/her net-utility, or,

$$\textbf{User}: \quad \max_{x_i(t)\le\phi_i(t),\ \forall t\in T} U_i(\boldsymbol{x}_i), \tag{1}$$

where each user i subscribes to 3G service only if the net-utility is positive, i.e., $U_i(\boldsymbol{x}_i) > 0$. User surplus S is the sum of users' net-utilities, or, $S = \sum_{i\in N} U_i(\boldsymbol{x}_i)$.

Under a given pricing scheme, the provider decides on the price (more precisely, the parameters of the pricing scheme) to maximize its expected revenue, $R(p)$:

$$\textbf{Provider}: \quad \max_{p\in\mathcal{P}} R(p), \tag{2}$$

where $\mathcal{P}$ is the set of all feasible prices such that *(i)* the revenue is positive (*provider rationality*), and *(ii)* the expected 3G traffic volume at each time and cell is smaller than 3G capacity C_{3g} (*capacity constraint*).

The expected revenue $R(p)$ is the total *income* minus *cost*,

$$R(p) = \sum_{i\in N} m\big(p, \boldsymbol{y}_i(\boldsymbol{x}_i)\big) - \sum_{i\in N} c\big(\boldsymbol{y}_i(\boldsymbol{x}_i)\big), \tag{3}$$

where $c(\boldsymbol{y}_i)$ is the network cost to handle the 3G traffic, which we model by a linearly increasing function, $c(\boldsymbol{y}_i) = \eta\sum_{t\in T} y_i(t)$, where η is the cost of the unit 3G volume. The cost term captures the money for operation and maintenance including the electric power cost as well as customer complaints due to congestion. The linearly increasing network cost is commonly used in the analysis of cellular cost [7].

2) Pricing: For a given pricing scheme, the 3G provider fixes a *price parameter* announced to the users. We consider four pricing schemes - *flat, two-tier, volume*, and *congestion* - that are popularly studied in literature. Each pricing scheme has tunable parameters controlled by the provider: $\{p_f\}$, $\{p_t^1, p_t^2, y_{\max}^1\}$, $\{p_v\}$, and $\{p_v(t,s)\}$, which we elaborate shortly. For a given pricing scheme and its price parameters p, a user with 3G traffic volume $\boldsymbol{y}_i$ pays $m\big(p, \boldsymbol{y}_i\big)$ to the provider. Note that if a user does not subscribe to a data plan or generate any traffic, the payment is zero, i.e., $m\big(p, \mathbf{0}\big) = 0$.

Flat. The provider offers unlimited service for users who pay a subscription fee p_f.

Two-tier. Multiple price points are provided for several usage options. For example, AT&T has a pricing plan that offers up to 300 MB, 3 GB, and 5 GB for \$20, \$30, and \$50 per month, respectively. In this paper, we consider two price points, where the provider offers maximum daily traffic volume $y_{\max}^1$ for fixed fee p_t^1 and unlimited service for fixed fee p_t^2, or,

$$m\big(p, \boldsymbol{y}_i\big) = \begin{cases} p_t^1, & \text{if} \quad 0 < \sum_{t\in T} y_i(t) \le y_{\max}^1. \\ p_t^2, & \text{if} \quad \sum_{t\in T} y_i(t) > y_{\max}^1. \end{cases}$$

Volume. A user is charged to pay p_v for the unit 3G traffic volume, or, $m\big(p, \boldsymbol{y}_i\big) = \sum_{t\in T} p_v \cdot y_i(t)$.

Congestion. We consider a volume-based congestion pricing, or simply congestion pricing in this paper, where the unit price varies with time and location, or, $m\big(p, \boldsymbol{y}_i\big) = \sum_{t\in T} p_v(t, s_i(t)) \cdot y_i(t)$, where $p_v(t,s)$ is the unit price at slot t and cell s, and $s_i(t)$ is the cell identifier with which user i is associated at t.

III. Analysis of WiFi Offloading Market

In this section, we provide analytical studies of the economic gain of WiFi offloading. Due to complex interplays among pricing parameters, and more importantly users' heterogeneity, our analysis is made under several assumptions. This simplification seems unavoidable for mathematical tractability, yet it helps to understand how offloading becomes economically beneficial. In Section IV, we quantify economic gains of WiFi offloading in more practical settings (heterogeneous cells and willingness to pay of users, and diverse pricing schemes)

A. Assumptions and Definitions

A1. *Homogeneous cells.* User associations are uniformly distributed among cells so that it suffices to consider only a single BS cell, where the number of users in a cell is $\hat{N} = N/$(# of cells). The distribution of users' traffic demand in each cell is identical.

A2. *Traffic demand distribution.* In each cell, the daily traffic demand Φ_i follows a random variable Φ which follows an upper-truncated power-law distribution, given by: $f_\Phi(x) = x^{-\sigma}/Z$, for $0 \le x \le \Phi_{\max}$, where σ is the exponent, $\Phi_{\max}$ is the maximum of Φ, and $Z = \frac{\Phi_{\max}^{1-\sigma}}{1-\sigma}$ with $0 < \sigma < 1$.

A3. *Willingness to pay and temporal preference.* Users are homogeneous in willingness to pay and temporal preference, i.e., $\gamma_i(t) = \gamma(t), w_i(t) = w(t),\ \forall i \in N,\ t\in T$. In regard to willingness to pay, we let $\gamma(t) = w(t)^{1-\theta}$.

A4. *Pricing.* We consider only flat pricing. (see our technical report [10] for volume pricing). Thus, throughout this section, the pricing parameter p refers to the flat fee p_f.

In **A2**, we comment that in recent measurement studies [8], [9], the traffic volume distribution of cellular devices is shown to follow an upper-truncated power-law distribution. Especially, in [8], the adopted pricing policy was flat pricing, so that the measured traffic usage was not affected by pricing. In **A3**, willingness to pay $\gamma(t)$ at time t is set, such that (i) one has larger willingness to pay for larger traffic demand and (ii) utility generated by traffic demand ($\gamma(t)\phi(t)^\theta$) is proportional to the traffic demand ($\phi(t) = w(t)\Phi$).

Remark 3.1: User heterogeneity only comes from traffic demand Φ from our assumptions. For notational simplicity, we omit user subscript i and use subscript Φ to represent the user variables with traffic demand Φ, e.g., $x_\Phi(t), y_\Phi(t)$, etc.

Offloading indicators. We first introduce two indicators to quantify how much 3G data is offloaded: (i) aggregate 3G traffic ratio κ_{avg} and (ii) peak 3G traffic ratio κ_{peak}.

Definition 3.1 (offloading indicators):

$$\kappa_{\text{avg}} \triangleq \frac{\sum_{t\in T} Y(t)}{\sum_{t\in T} X(t)}, \qquad \kappa_{\text{peak}} \triangleq \frac{\max_{t\in T} Y(t)}{\sum_{t\in T} X(t)}, \tag{4}$$

where the transmitted total traffic and 3G traffic over a cell at time t, $X(t)$ and $Y(t)$[5] are: $X(t) = \hat{N}\int_0^{\Phi_{\max}} x_\Phi(t) \mathrm{d}F_\Phi$, $Y(t) = \hat{N}\int_0^{\Phi_{\max}} \sum_{d=0}^{D} \alpha_\Phi^d \big(1 - e_\Phi^d(t-d)\big) x_\Phi(t-d) \mathrm{d}F_\Phi$.

[5] When we emphasize that $Y(t)$ (resp. $X(t)$) depends on a given price p, $Y(t;p)$ will be used instead of $Y(t)$ (resp. $X(t;p)$).

It is clear that as users delay more traffic, the aggregate 3G traffic ratio κ_{avg} provably decreases, since more traffic can be offloaded through WiFi. Also, the peak 3G ratio κ_{peak} decreases as more traffic at *peak* time is offloaded.

Opt-saturated and Opt-unsaturated. We define two notions, *opt-saturated* and *opt-unsaturated*, which characterize the regimes under which *how much traffic is imposed on the network for the equilibrium price.* In general, as traffic demand gets higher compared to the 3G capacity, the network becomes *opt-saturated*, and vice versa. The main reason for introducing those two notions is because the analysis becomes different depending on the volume of network traffic and the market behaves differently, and thus, the way of increasing the revenue and the net-utility can be differently interpreted. For a formal definition, we first recall that $\mathcal{P}$ is the set of all feasible prices, defined by provider rationality and capacity constraint, or,

$$\mathcal{P} \triangleq \{p \mid R(p) > 0, Y(t;p) \leq C_{3\text{g}},\ \forall t \in T\}. \qquad (5)$$

Definition 3.2 (Opt-saturated and Opt-unsaturated): Let $p^\star$ be an *equilibrium price* that maximizes the revenue, i.e., $p^\star \in \arg\max_{p\in\mathcal{P}} R(p)$. The network is said to be *saturated* at p, if $\max_{t\in T} Y(t;p) = C_{3\text{g}}$. For a ***unique*** equilibrium price $p^\star$, the network is said to be *opt-saturated* if the network is saturated at $p^\star$, and *opt-unsaturated* otherwise.

Let p_0 be the *threshold price* above which all the feasible prices lie, i.e., $p_0 = \inf_{p\in\mathcal{P}} p$. Let $A(p) = \max_{t\in T} Y(t;p)$ be the total 3G traffic at *peak* time.

B. Economic gain analysis

Due to space limitation, we present only flat pricing. Refer to [10] for volume pricing. A user pays a flat fee p regardless of its 3G traffic usage if it subscribes to the 3G service. Since there is no incentive to discourage excessive network traffic, the traffic volume generated by a user equals to its traffic demand; $x_\Phi(t) = \phi(t)$. A user with Φ maximizes its net-utility:

$$\sum_{t\in T} \gamma(t) x_\Phi(t)^\theta - p = \sum_{t\in T} w(t)^{1-\theta}\phi(t)^\theta - p = \Phi^\theta - p, \qquad (6)$$

since $\gamma(t) = w(t)^{1-\theta}$ by **A3**, and the temporal preference $w(t) = \phi(t)/\Phi$. From (3), the provider maximizes its revenue:

$$R(p) = \hat{N}p \int_{p^{\frac{1}{\theta}}}^{\Phi_{\max}} \mathrm{d}F_\Phi - \hat{N}\eta \int_{p^{\frac{1}{\theta}}}^{\Phi_{\max}} \Phi \mathrm{d}F_\Phi \qquad (7)$$

where $p^{1/\theta}$ is the lowest traffic demand of a subscriber (i.e., positive net-utility and thus $\Phi^\theta > p$). No users subscribe if the price is too high, i.e., if $p \geq p_{\max}$ from (6), where $p_{\max} = \Phi^\theta_{\max}$. Then, we should have that $\mathcal{P} \subset [0, p_{\max})$.

Our main results, Prop. 3.1 and Theorem 3.1, state how the economic values (e.g. price and revenue) change by offloading.

Proposition 3.1 (Equilibrium Price): If $\eta < (\kappa_{\text{avg}}\Phi^{1-\theta}_{\max})^{-1}$,

(i) $R(p)$ is unimodal[6] over $[0, p_{\max})$, and the feasible price set $\mathcal{P}$ is non-empty and connected.

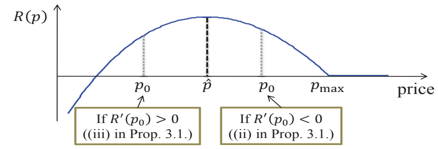


Fig. 2. Revenue function $R(p)$ in flat pricing. The $p_{\max}$ is the highest price above which no user subscribes in flat pricing, $\hat{p}$ is the unique solution of $\partial R(p)/\partial p = 0$, and η is the cost coefficient. The achievable revenue (at equilibrium) is not always at $\hat{p}$, since $\hat{p}$ may not be in the feasible price set.

(ii) The network is *opt-saturated*, if $R'(p_0) < 0$, where the unique equilibrium price $p^\star = p_0$, where $p_0 = \Phi^\theta_{\max}\left(1 - \frac{C_{3g}}{\kappa_{\text{peak}}\hat{N}\mathbb{E}[\Phi]}\right)^{\frac{\theta}{2-\sigma}}$.

(iii) The network is *opt-unsaturated*, if $R'(p_0) > 0$, where the unique equilibrium price $p^\star = p^\star(\kappa_{\text{avg}})$ is such that $\frac{\partial R(p)}{\partial p}\big|_{p=p^\star} = 0$, and $\frac{\partial p^\star(\kappa_{\text{avg}})}{\partial \kappa_{\text{avg}}} > 0$.

Theorem 3.1 (Economic Gain): If $\eta < (\kappa_{\text{avg}}\Phi^{1-\theta}_{\max})^{-1}$, the provider's revenue increases, and the net-utilities of all subscribers increase at equilibrium (thus, the user suplus increases), as (i) κ_{peak} decreases in the opt-saturated case, and (ii) as κ_{avg} decreases in the opt-unsaturated case.

The proofs of Prop. 3.1 and Theorem 3.1 are presented in our technical report [10]. Here, we briefly interpret Prop. 3.1 and Theorem 3.1. First, clearly if the network cost is too high, the provider cannot achieve any positive revenue, where the condition of cost coefficient $\eta < (\kappa_{\text{avg}}\Phi^{1-\theta}_{\max})^{-1}$ guarantees the existence of prices under which the revenue is positive. This condition is relaxed as more offloading occurs (i.e., κ_{avg} decreases), resulting in less restricted business condition for the provider. Second, $R(p)$ is unimodal and $\mathcal{P}$ is connected. Thus, at the threshold price, if $R'(p_0) < 0$, then $R(p_0) \geq R(p)$ for all $p \in \mathcal{P}$ (see Fig. 2). Thus, the equilibrium price is unique, and $p^\star = p_0$. Also, the network is opt-saturated, because the peak traffic volume $A(p_0) = C_{3\text{g}}$ (otherwise, there is a smaller feasible price than p_0). Now, if $R'(p_0) > 0$, the equilibrium price $p^\star$ is such that $R'(p^\star) = 0$, as in Prop. 3.1(iii). This case makes the network opt-unsaturated because $A(p^\star) < A(p_0)$ (due to decreasing property of $A(p)$ in p) and $A(p_0) \leq C_{3\text{g}}$.

Using the results of Prop. 3.1, Theorem 3.1 states that offloading is economically beneficial for both the provider and users, where the way of revenue increase is dependent on the opt-saturatedness. In the opt-saturated case, as more 3G traffic is offloaded through WiFi at *peak* time, i.e., κ_{peak} decreases, the provider turns out to have extra 3G capacity. Then, the provider attracts more subscribers by lowering its flat fee, in order to utilize the extra capacity. As the increase in the number of subscribers exceeds the reduced price, the revenue increases. The net-utility increases for all subscribers by price reduction since a subscribing user in flat pricing always generate all the traffic demand. In the opt-unsaturated case, the number of subscribers does not increase drastically even if the provider decreases its flat fee, so that the income does not increase. However, the network cost decreases substantially as the 3G traffic decreases, i.e., κ_{avg} decreases, and the revenue increases. Since the equilibrium price still decreases, as κ_{avg} decreases from Prop. 3.1(iii), the net-utility increases for all subscribers.

[6] A function $f(x)$ is called unimodal, if for some value v, it is monotonically increasing for $x \leq v$ and monotonically decreasing for $x \geq v$.

IV. Trace-driven numerical analysis

A. Setup

In this subsection, we describe the setup for our trace-driven numerical analysis, such as the real traces and the parameter values. The duration of a time slot is set to be an hour, i.e., $T = 24$. The number of BS cells is 31 and the average number of users per cell is 1000,[7] thus total 31000 users. The choice of 31 cells is due to a real trace which will be explained shortly. We test two cellular capacities, 8 Mbps for 3G and 32 Mbps for 4G, where the 4G capacity is projected to be about four times the 3G capacity [13].[8] We set the price sensitivity $\theta = 0.5$ and the cost coefficient $\eta = 0.1$. We use two real traces to get the statistics of users' traffic and WiFi connection probabilities.

Trace 1 (3G traffic usage). The first trace is from a major cellular provider in Korea and includes the information on the number of high speed downlink/uplink packet access (HSDPA/HSUPA) calls, recorded every hour, in each of the 31 BSs in a day. The temporal traffic pattern in each BS depends on the characteristics of the coverage, e.g., residential or office area (see our technical report [10] for the detailed statistics).

Trace 2 (WiFi connection). The second trace is measured by 93 iPhone users from an iPhone user community in Korea, who volunteer and record their time-varying WiFi connectivity and locations, periodically scanned and recorded at every 3 minutes for two weeks (see [3], [10]). We only recorded APs to which users can transmit data by sending a ping packet to our server.

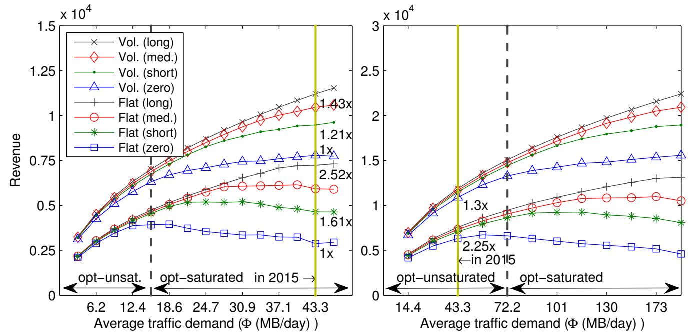
(a) Traffic demand (ϕ_i) and willingness to pay ($\gamma_i(t)$): Measurement studies on mobile data [8], [9], [14] showed that the user traffic volume follows an upper-truncated power-law distribution. Thus, we use an upper-truncated power-law distribution with the exponent $\sigma = 0.57$ (which is observed in [9]), as the distribution of total daily traffic demand Φ_i by scaling the average, so that the per-month average ranges from 93 MB to 5.2 GB. The temporal preference ($w_i(t) = \phi_i(t)/\Phi_i$) of users follows the average temporal *usage pattern* in *trace 1*. Users' willingness to pay is set to include some randomness across users, and it is proportional to temporal preference, i.e., $\gamma_i(t) = \nu_i w(t)^{1-\theta}$, where ν_i is uniformly distributed in $(0, 1)$.

(b) WiFi connection probability ($e_i^d(t)$): We use the *trace 2* to obtain the values of $(e_i^d(t) : i \in N, t \in T)$. Since *trace 2* includes only 93 users, we repeatedly use their individual traces to generate N users' data, i.e., about $N/93$ users have the same $e_i^d(t)$. We refer the readers to [3] to know how often users meet WiFi in the experiment. For the 10 minutes and 6 hours deadlines, the average WiFi contact probabilities are 0.7 and 0.88, and the medians are 0.87 and 0.97, respectively.

(c) BS association ($s_i(t)$): Since *Trace 1* does not include users' cell-level mobility mainly because of privacy, we combine *trace 1* with the handover statistics of users per day from [8] to obtain the cell-level mobility $s_i(t)$ (see our technical report for the detail [10]).

[7] This is a typical number of users in a macro BS. For example, Sprint has 66,000 BSs and 55 million subscribers at the end of 2011 [11], [12].

[8] Yet, we still use the notation C_{3g} for notational simplicity.



(a) C_{3g} = 3.6 GB/hour (8 Mbps)   (b) C_{3g} = 14.4 GB/hour (32 Mbps)

Fig. 3. Flat and volume pricing: revenue for various delay profiles in Table I and traffic demand, with different cellular capacities, where users experience a single scenario in Table I. The numbers (-x) represent the increase of revenue by (a) the delayed offloading or (b) the network upgrade (from 3G to 4G). Opt-saturatedness is determined by the traffic demand and cellular capacity, where the dotted line shows the threshold in the on-the-spot offloading.

(d) Delay profile (α_i^d): To model the delay profile of users, we use various scenarios, such as *no-deadline*, *short*, *medium*, and *long*, where each scenario consists of four different classes (Video, Data, P2P, and Audio), as classified by Cisco [1]. The details are described in Table I. We consider different scenarios where all users are uniformly given a scenario, e.g., *medium*.

TABLE I
Traffic classification projected in year 2015 from Cisco [1] and assigned deadlines for each traffic class. SC: Scenario.

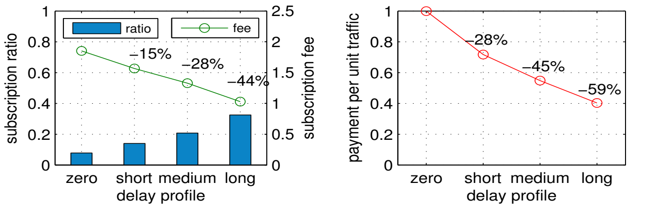
	Video	Data	P2P	Audio (VoIP)	Total
Ratio	66.4 %	20.9 %	6.1 %	6.6 %	100 %
SC:zero	0 sec.	0 sec.	0 sec.	0 sec.	-
SC:short	10 min.	30 min.	10 min.	0 sec.	-
SC:medium	30 min.	1 hour	30 min.	0 sec.	-
SC:long	2 hours	6 hours	2 hours	0 sec.	-

B. Results

Our numerical results quantify the benefits of delayed WiFi offloading in various aspects. We present our results by summarizing the key observations.

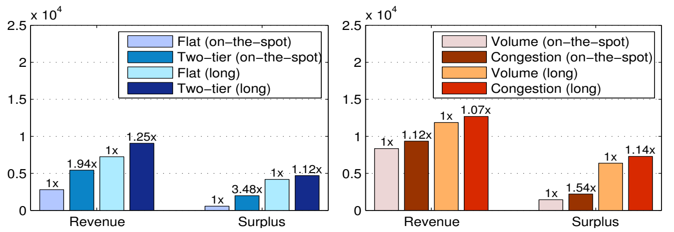
1) Revenue in volume pricing exceeds that in flat pricing by applying delayed WiFi offloading, but the revenue increase is higher in flat pricing than volume pricing: Fig. 3(a) depicts the revenue of flat and volume for various traffic demand and delay profiles. The revenue in volume exceeds that in flat in all cases, because in flat, a subscriber with high traffic demand generates heavy traffic and dominates the network resources without paying more fees to the provider, whereas in volume, the user payment is proportional to the traffic volume, so that if a subscriber generates heavy traffic, the payment is high. However, the revenue *increase* of delayed offloading, compared to on-the-spot offloading, is higher in flat. The revenue increase in flat (volume) is about 61-152% (21-43%), when $\mathbb{E}[\Phi_i]$ =43.3 MB/day. This is because in flat, 3G traffic reduction does not affect the provider's income, whereas, in volume, 3G traffic reduction decreases the income.

2) The revenue gain from on-the-spot to delayed offloading is similar to that generated by the network upgrade from 3G to 4G: Fig. 3 depicts revenue of flat and volume pricing for various traffic demand, with different cellular capacities. If a provider



(a) Flat: fee and subscription ratio (b) Volume: payment per unit traffic

Fig. 4. Change of flat price and subscription ratio in flat pricing, and payment per unit 3G+WiFi traffic in volume pricing. The average traffic demand is 43.3 MB/day (1.5GB/month) and 3G capacity is 3.6 GB/hour (opt-saturated).



(a) Flat vs. Two-tier (b) Volume vs. Congestion

Fig. 5. The revenue and surplus in flat, tiered, volume and congestion pricing. The number (-x) above each bar represents the increase compared to the revenue in (a) flat or (b) volume pricing. The average traffic demand is set to be 43.3MB/day (1.5GB/month) and the 3G capacity is 3.6GB/hour.

upgrades the network from 3G to 4G the revenue increases by 115% in flat and 30% in volume, when $\mathbb{E}[\Phi_i] = 43.3$ MB/day, which is similar to the gain from delayed offloading. Thus, if the traffic demand is high compared to capacity, adopting delayed offloading can be a good solution to increase revenue, where the network upgrade induces huge installation costs.

3) As more traffic is offloaded, the flat price decreases in flat pricing, and the payment per unit traffic decreases in volume pricing: As shown in Fig. 4, the flat fee decreases by 15-44% and the subscription ratio increases accordingly in flat pricing, and the payment per unit 3G+WiFi traffic decreases by 28-59% in volume pricing, when $\mathbb{E}[\Phi_i] = 43.3$ MB/day (opt-saturated). In the opt-unsaturated case, price reduction is not drastic both in flat and volume pricing, since the *income* does not increase by price reduction due to the low traffic demand. The reduction in flat price and payment per unit traffic in volume pricing induces increase in user surplus, as shown in Fig. 5.

4) Two-tier and congestion pricing increase the revenue, compared to flat and volume pricing, but such gains become smaller, as more traffic is offloaded: Fig. 5 shows the change of revenue and surplus in four pricing schemes. It is intuitive that as pricing granularity increases in terms of price (from flat to tiered) or space/time (from volume to congestion), the revenue increases, because the provider has more degree of freedom to control the market. However, the rate of increase diminishes as more traffic is offloaded. The revenue in two-tier pricing is greater than that in flat pricing by 94% and 25% in on-the-spot and delayed offloading, where the revenue in congestion pricing is greater than that in volume pricing by 12% and 7% in on-the-spot and delayed offloading, respectively. We find that delayed offloading reduces spatiotemporal imbalance by dispersing traffic to other time and locations, so that the effect of space/time-varying price is reduced (see Fig. 8. in [10]).

V. Concluding Remark and Future Work

In this paper, we model a game-theoretic framework to study the economic aspects of WiFi offloading, where we drew the following messages from the analytical and numerical studies: (a) WiFi offloading is economically beneficial for both the provider and users, where the increase ranges from 21% to 152% in the provider revenue and from 73% to 319% in the users' surplus. (b) The revenue gain from on-the-spot to delayed offloading is similar to that generated by the network upgrade from 3G to 4G. (c) The revenue increase of complex pricing schemes (such as two-tier from flat, and congestion from volume), becomes smaller for higher offloading chances, which is true as of now and in the future, when more WiFi APs are expected to be deployed.

Another well-known benefit from WiFi offloading is transmission energy saving. It is shown in [3] that 50-60% of transmission energy can be reduced for 1-hour delay. Even if we do not consider the energy benefit in this paper, it is obvious that most users benefit from the increased battery lifetime.

There are a few limitations in our work. Our results rely on the assumption that the network traffic has a degree of delay tolerance and users can tolerate some amount of delay, where delay depends on the class of traffic. Thus, our results can sometimes be regarded as an upper-bound on the economic benefits of WiFi offloading. A simple way of reflecting those limitations would be to design a net-utility function which jointly captures the happiness by data transmission and the disutility by delay, which we leave as a future work. Another future work is to consider multiple providers competing for users, where they have different plans to overcome the mobile data explosion (e.g. delayed offloading or network upgrade).

References

[1] Cisco Systems Inc., "Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016," 2012.
[2] CNN, "4g won't solve 3g's problems," March 2011.
[3] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can wifi deliver?" in *Proc. of ACM CoNEXT*, 2010.
[4] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proc. of ACM MobiSys*, 2010.
[5] J. Hultell, P. Lungaro, and J. Zander, "Service provisioning with ad-hoc deployed high-speed access points in urban environments," in *Proc. of IEEE PIMRC*, 2005.
[6] Gigaom, "Forget wireless bandwidth hogs, lets talk solutions," 2012.
[7] K. Johansson, J. Zander, and A. Furuskar, "Modelling the cost of heterogeneous wireless access networks," *International Journal of Mobile Network Design and Innovation*, vol. 2, no. 1, 2007.
[8] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. of IEEE INFOCOM*, 2011.
[9] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proc. of ACM SIGMETRICS*, 2010.
[10] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of wifi offloading: Trading delay for cellular capacity," in *Technical Report*, 2012, *available at* http://arxiv.org/abs/1207.6607.
[11] Nadine Manjaro, "Sprint made the best choice in selecting to deploy its own fdd lte network," 2011.
[12] Sprint, "Sprint nextel reports - fourth quarter and full year 2011 results," 2012.
[13] Ofcom, "4g capacity gains," January 2011.
[14] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin, "A first look at traffic on smartphones," in *Proc. of ACM IMC*, 2010.