Fair and Optimal Mobile Assisted Offloading

Divya R[§], Amar Prakash Azad [*], Chandramani Singh[§]

Department of Electronic Systems Engineering, IISc, Bangalore-560012[§], IBM Research India, Bangalore-560045[*]
Email:{divya,chandra}@dese.iisc.ernet.in[§], amarazad@in.ibm.com[*]

***Abstract*—We study an offloading mechanism for cellular networks in which mobiles with good cellular links can act as hotspots and can assist other mobiles. We study throughput optimal offloading and also a fair offloading strategy which we call *proportional increment offloading*. We show that the former problem can be reduced to a *capacitated facility location problem* (CFLP) whereas the latter can be solved by solving a sequence of CFLPs. We propose a *belief propagation* based algorithm to solve CFLP, a well known NP complete problem. We primarily consider *point coordination function* (PCF) based WiFi access for offloading but also discuss *distributed coordination function* (DCF) based access and related issues. Further, we argue that all the mobiles benefit through participating in offloading. We perform extensive simulation to evaluate the performance of the proposed algorithms and effectiveness of mobile assisted offloading.**

I. Introduction

Recent advances in cellular network have resulted in dramatic enhancements in wireless capacity. However, the wireless capacity remains in constant deficit due to exploding mobile data traffic usage. Proliferation of smartphones have boosted the data usage through data-hungry mobile apps, video streaming and various social and cloud services. Recently, Cisco has predicted that global mobile data traffic is expected to grow 18 folds between 2011-2018, three time faster than the overall fixed IP traffic in the same period [2], in which 66.5% will be contributed by video content.

Data offloading is becoming the most promising solution either from macrocell to smaller cells (e.g., picocells, femtocells) or over Wireless Fidelity (WiFi) enabled by *heterogeneous network* deployment. These solutions have become integral part of the next generation cellular network such as LTE. Typically, additional infrastructure is deployed to enable offloading in either case, e.g. Femto/Pico base stations (BSs) to offload mobile traffic from macrocells [17] or WiFi hotspots to offload cellular traffic to a backbone or cable network [21]. In some cases, the urban widespread home WiFi access point acts as a hotspot for small range offloading which reduces the additional infrastructure deployment cost [21].

In this work, we study a scheme where mobiles with good cellular links can act as WiFi *hotspots*, which can offload the cellular data of nearby mobiles with poor cellular links. More precisely, the good cellular link of the hotspot is shared with the associated mobiles with poor links. Further, the hotspot sends data to a mobile (or, receives data from it in the case of uplink communication) using WiFi. We focus on the problems of hotspot selection and hotspot mobile association to achieve fair and optimal network wide throughputs. We study a sum throughput maximization problem in which each mobile obtains at least the rate it was originally getting. We see that the aggregate network throughput can be significantly enhanced. Towards fairness, we study a *proportional increment fair* strategy in which each mobile obtains a proportional gain as compared to the *base rate*. While a mobile serving as a hotspot results in overall network throughput gain, incentive for becoming a hotspot is a natural question. To address this, we study a simple incentivizing scheme in which the hotspot is provided an additional amount of throughput fraction. We formulate an optimization problem to capture the incentivizing scheme which has similar structure as that of the *sum throughput maximization problem*. We consider both PCF and DCF based WiFi access for the mobiles to connect to the hotspots.

Interestingly, we show that our problem can be reduced to a capacitated facility location problem (CFLP). A facility location problem consists of a set of candidate facilities and a set of customers. Here, the objective is to open a subset of the facilities so as to minimize the sum of the facility opening costs and the transportation costs. In case of CFLPs the facilities have certain capacities and the associated customers' demands should not exceed these. On the other hand, there are no capacity constraints in the uncapacitated facility location problems (UFLPs). Both UFLP and CFLP are know to be NP complete [24]. Belief Propagation based message passing algorithms for UFLP have been proposed in [13]. In this paper, we propose a message passing based algorithm to solve CFLP, which is also suitable for a distributed setting. We apply this algorithm to our network configuration problem which is shown to perform well through numerical evaluation. In particular, the key contributions of this work are as follows:

- We frame joint problems of BS resource allocation, hotspot selection and association for optimal and fair mobile assisted offloading as constrained *mixed integer nonlinear programming* problems. Then, we reduce these to CFLPs, a widely studied class of mixed integer linear programming problems.
- We propose a belief propagation based algorithm to solve general CFLPs. Observe that CFLPs are rooted in *operation research* and are known to be NP-complete. The proposed algorithm is suitable for distributed implementation. We apply this algorithm to solve optimal (and fair) offloading problems.
- We discuss incentivizing strategies for mobiles to serve as hotspots. We analyse one such strategy and illustrate its impact via numerical evaluation.
- Besides PCF, we also consider DCF based WiFi offloading and discuss its performance and related issues.

A. Related Work

There are several works on offloading cellular data traffic using small cells [5], [8], [17] and WiFi [14], [19].In contrast to our work, they consider offloading the cellular data over independent connection to internet through infrastructure based backhaul or cable based connections. The authors in [10], [21] focus on WLANs wherein the APs also need to serve nonlocal mobiles. They propose collaborative approaches for the APs to share some of their bandwidth to serve these mobiles. The set of APs is known a priori and mobiles are connected to these access points optimally. In [22], the authors focus on placement of a fixed number of Mobile backbone nodes (MBNs) and the assignment of regular nodes to MBNs. Unlike

978-3-901882-90-6/17 ©2017 IFIP

our problem, in their network, the mobiles and the MBNs belong to disjoint sets. In particular, a mobile cannot act as an MBN for other mobiles. Also, we do not a priori fix the number of hotspots as in [22].

A cellular data sharing is proposed in [15] where mobile hosts with poor links use peer-to-peer links to access proxy clients with better channel quality. The paper does not consider the key problem of network reorganization in such cases and how the sharing is done to ensure benefits to all the nodes. The authors in [7], [9] allow a few mobiles to be hotspots and design incentive mechanisms to encourage them to do so. The work in [16] also studies mobile assisted offloading aiming at maximizing the aggregate network throughput. However, they deal with a simplistic model and only provide heuristics.

Facility location problem (FLP) is well studied in operation research context. *Simple FLP* or *uncapacitated FLP (UFLP)* involves locating an undetermined number of facilities to minimize the aggregate cost of serving the demand from these facilities. In [11], authors have established NP-completeness of *UFLP* using *packing-covering-partition* approach. The author in [6] has presented a dual based algorithm for solving UFLP which remains the most efficient technique yet. Message passing based algorithm for UFLP is presented in [13], [20] and is shown to perform well. The *capacitated* version of *FLP*, called as *capacitated facility location problem (CFLP)*, incorporates the capacity limitations on the facilities. Besides several earlier methods, cross-decomposition algorithm of [23] and Lagrangian based approach in [4] are the most effective techniques. The basic idea in [23] is to obtain UFLP structure by dualyzing the capacity constraints.

Remaining of the paper is organized as follows. The network model and optimization problem is described in Section II which includes the cellular and Wifi connection configuration specific to our problem. In Section III, we show the reduction of our problem to CFLP and propose belief propagation based message passing algorithm for CFLP. In Section IV, we study a simple hotspot incentivizing scheme and describe the optimization problem formulation. In Section V, we formulate the hotspot selection problem, in which, the WiFi access happens using DCF. Through numerical evaluation, we show the benefit of optimal and fair mobile assisted offloading in various cases in Section VI. Finally, we conclude with remarks in Section VII.

II. Hotspot Selection Problem

A. Network Model

We consider a single cell of a cellular network where the cell has N mobiles indexed by $i = 1, 2, \ldots, N$. If the BS uses all its resources, e.g., *physical resource blocks* in an OFDMA network or *time slots* in a TDMA network, to serve mobile i, the mobile would obtain a throughput R_i. Assume that the mobiles are indexed such that $R_1 \geq R_2 \geq \cdots \geq R_N$. The BS can arbitrarily divide the resources among the mobiles; if it allocates a fraction α of the resources to serve mobile i, the mobile would obtain throughput αR_i. We assume that, in the base case, the resources are equally shared among the mobiles. Thus, for each i, mobile i receives a throughput R_i/N. However, the following analysis can easily be adapted to an arbitrary allocation of resources.

Mobiles with good cellular links can aid to the network by acting as hotspots and serve other mobiles. For example, in the case of downlink communication, a hotspot can use its cellular

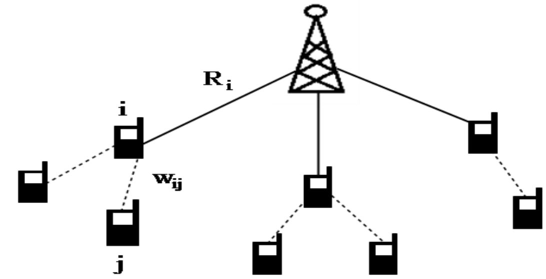


Fig. 1: An illustration of mobile assisted offloading

link to receive data intended for the mobiles that it serves, and can deliver the same to the mobiles using WiFi. So, in our model, each mobile either directly connects to the BS and possibly acts as a hotspot for a few other mobiles or receives data through another hotspot. In the latter case, it can connect to exactly one hotspot. Let $a_{ij} \in \{0, 1\}, i, j = 1, 2, \ldots, N$ be binary variables such that

(a) a_{ii} is one if and only if i directly connects to the BS, and
(b) $a_{ij}, i \neq j$ is one if and only if i serves as a hotspot for j

Clearly,

$$\sum_{i=1}^{N} a_{ij} = 1,\ j = 1, \ldots, N, \tag{1}$$

$$\text{and } a_{ij} \leq a_{ii},\ i, j = 1, \ldots, N. \tag{2}$$

Note that the BS needs to allocate its resources only among the hotspots. Let α_i be the fraction of resources allocated to mobile i; $\alpha_j = 0$ if j is not a hotspot. Let W_{ij} be the maximum possible WiFi rate between mobiles i and j when the former becomes a hotspot and serves the latter. The actual rate depends on the *modulation and coding scheme* and can be less than W_{ij}. Finally, let T_j denote the throughput received by mobile j. Note that, for any hotspot, the sum of its throughput and its client mobiles' throughputs must equal the throughput that it receives from the BS. So, we must have

$$\sum_{j=1}^{N} a_{ij} T_j = \alpha_i R_i,\ i = 1, \ldots, N \tag{3}$$

$$\text{where } \sum_{i=1}^{N} \alpha_i \leq 1. \tag{4}$$

We assume that the hotspots constitute noninterfering WiFi cells. This holds true if the hotspots are sufficiently far apart or if they use disjoint frequency bands. The hotspots can serve mobiles using either PCF or DCF modes of WiFi. In PCF mode, the hotspot can allocate arbitrary fractions of air-times (WiFi slots) to the connected mobiles. These fractions determine the mobiles' throughputs. For a mobile j to get throughput T_j, its hotspot i must give it T_j/W_{ij} fraction of air-time. Clearly,

$$\sum_{j=1:j\neq i}^{N} a_{ij} \frac{T_j}{W_{ij}} \leq 1,\ i = 1, \ldots, N. \tag{5}$$

On the other hand, in DCF mode the nodes use distributed CSMA-CA protocol for medium access. Here, all the mobiles connected through the same hotspot receive equal throughput that is limited by the least WiFi rate among all mobiles. More precisely, for all hotspots i,

$$T_j = \frac{1}{\sum_{l=1}^{N} a_{il}\frac{1}{w_{il}}}, \forall j \neq i \text{ such that } a_{ij} = 1, \qquad (6)$$

where $w_{il} \leq W_{il}$ are the actual WiFi rates between hotspot i and the mobiles.[1]

While DCF mode allows distributed access and is suitable for general hotspots, it is throughput-wise inefficient in case of excessive backoffs. In our offloading setup, mobiles' throughput requirements are prescribed and their association is also regulated (through the optimization problems that we formulate). Hence, a centrally optimized access mode, such as PCF, is preferred. Moreover, forthcoming WiFi standards also advocate centrally controlled medium access (e.g., *Target Wake Time* in IEEE 802.11ah [1]). Hence, we primarily consider PCF mode, i.e., constraints (5), in our analysis. We briefly discuss the analysis of DCF mode and associated complexities in Section V.

B. Problem Formulation

We study the joint problem of BS resource allocation, hotspot selection and mobile-hotspot association which together determine mobiles' throughputs T_is. In general, one may want to optimize some utility function $U(T_1, \ldots, T_N)$ in making decisions. Here, we focus on a particular notion that maximize mobiles' throughputs while ensuring that these are in proportion to mobiles' base throughputs R_i/Ns - we call it *proportional increment optimization*. In particular, we focus on the following optimization

$$\begin{aligned}\mathcal{P}_1 : \text{maximize} \quad & \xi \\ \text{subject to} \quad & T_j = \xi\frac{R_j}{N}, j = 1, \ldots, N, \\ & (1), (2), (3), (4) \text{ and } (5)\end{aligned}$$

The above optimization has $a = (a_{ij}, i, j = 1, \ldots, N), \alpha = (\alpha_i, i = 1, \ldots, N), T = (T_j, j = 1, \ldots, n)$ as decision variables. This is a complex combinatorial optimization problem. Hence, we reduce it to a sequence of problems that are relatively simpler to solve in Section II-C.

Throughput Optimization: We can consider the objective of maximizing the aggregate throughput of all the mobiles subject to each mobile receiving no less that its base throughput. Such a solution is likely to be unfair but gives an estimate of the maximum benefit of offloading.

$$\begin{aligned}\mathcal{P}_2 : \text{maximize} \quad & \sum_j T_j \\ \text{subject to} \quad & T_j \geq \frac{R_j}{N}, \quad j = 1, \ldots, N. \qquad (7) \\ & (1), (2), (3), (4), (5)\end{aligned}$$

The decision variables in the above optimization problem are a, T and α. The authors in [16] have dealt with a problem similar to ours. However, they do not account for WiFi constraints such as (5) or (6). They also assume that the BS's resources are equally divided among all the mobiles that directly connect to it. This can lead to highly disproportionate resource sharing. To see this, consider a scenario where only two mobiles connect to the BS, first of these act as a hotspot and serves all the other mobiles where as the second one does not serve any. According to [16], both the mobiles get equal share of the BS's resources. The authors in [16] also do not consider any notion of fairness.

Time varying cellular and WiFi rates: We can account for time varying rates by instantiating and solving a new hotspot selection problem, after every few slots, when the rates have changed noticeably. The proposed message passing based solutions are distributed and quickly converge and hence suit our need.

C. Reduction to Simpler Problems

We begin with the throughput optimization problem. We show that it is equivalent to the following problem that has only $a = (a_{ij}, i, j = 1, \ldots, N)$ as decision variables.

$$\begin{aligned}\mathcal{P}_3 : \text{minimize} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\frac{R_j}{NR_i}, \\ \text{subject to} \quad & \sum_{j\neq i} a_{ij}\frac{R_j}{NW_{ij}} \leq 1, \ i = 1, \ldots, N, \qquad (8) \\ & (1), (2)\end{aligned}$$

The optimal objective value of this problem can be interpreted as the least fraction of resources the BS needs to spend to meet the base rate requirements of all the mobiles.

Lemma 2.1: The optimization problems $\mathcal{P}_2$ and $\mathcal{P}_3$ give rise to same sets of solutions $(a_{ij}, i, j = 1, \ldots, N)$. Furthermore, for any solution a^* of $\mathcal{P}_3$, (a^*, T^*, α^*), where

$$\begin{aligned}T_j^* &= \frac{R_j}{N}, j = 2, \ldots, N, \\ \alpha_i^* &= \sum_j \frac{a_{ij}^* R_j}{NR_i}, i = 2, \ldots, N, \\ \alpha_1^* &= 1 - \sum_{i\geq 2} \alpha_i^*, \\ \text{and } T_1^* &= \alpha_1^* R_1 - \sum_{j\geq 2} \frac{a_{1j}^* R_j}{N},\end{aligned}$$

is a solution of $\mathcal{P}_2$.

Please refer [18] for the proof. Now, we consider proportional increment optimization as described by problem $\mathcal{P}_1$. Consider a solution a^* to $\mathcal{P}_3$ and the corresponding optimal value θ^*. Define $T^* = (T_j^* = R_j/N, j = 1, \ldots, N)$ and $\alpha^* = (\alpha_i^* = \sum_{j=1}^{N} a_{ij}^* R_j/NR_i, i = 1, \ldots, N)$. Recall that θ^* is actually the least fraction of the BS resources required to serve all the mobiles' base rates. However, we cannot scale T^* and α^* up by $1/\theta^*$ to get a solution to $\mathcal{P}_1$ as $(a^*, T^*/\theta^*, \alpha^*/\theta^*)$ may violate (8). We now present a way to solve this problem.

Consider $\mathcal{P}_3$ with constraint (8) replaced by $\sum_{j\neq i} a_{ij}R_j/NW_{ij} \leq \theta, i = 1, \ldots, N$, for some $\theta \leq 1$. Let $\mathcal{P}_3^{\theta}$ denote the modified problem. Let $a(\theta)$ and $\gamma(\theta)$ be the solution and objective value, respectively, of $\mathcal{P}_3^{\theta}$. $\gamma(\theta)$ is a stepwise decreasing function and $\gamma(\theta) = \theta$ may or may not have a fixed point. Refer Fig. 3 for an illustration. We propose an iterative algorithm (Algorithm 1) to obtain a solution to $\mathcal{P}_1$.

Let us define

$$R = (R_1, R_2, \ldots, R_N),$$

[1] Actually, (6) gives an upper bound on the throughput of each mobile. This expression well approximates the throughput for very high packet lengths and has been widely used [12].

Algorithm 1

initialize $k = 0, \bar{\theta}_0 = 1, \underline{\theta}_0 = \gamma(1)$,
while $k \leq K$ **do** // *K is the number of iterations.*
$\tilde{\theta}_k = \frac{\underline{\theta}_k + \bar{\theta}_k}{2}$,
$\underline{\theta}_{k+1} = \max\left\{\underline{\theta}_k, \min\{\tilde{\theta}_k, \gamma(\tilde{\theta}_k)\}\right\}$,
$\bar{\theta}_{k+1} = \min\left\{\bar{\theta}_k, \max\{\tilde{\theta}_k, \gamma(\tilde{\theta}_k)\}\right\}$,
$k = k + 1$
end while

$$\alpha_i(\theta) = \sum_{j=1}^{N} \frac{a_{ij}(\theta)R_j}{NR_i\theta}, i = 1, \ldots, N,$$

$$\text{and } S(\theta) = \max_i \sum_{j \neq i} \frac{a_{ij}(\theta)R_j}{NW_{ij}}.$$

Remark 2.1: For any θ, there should be unique $a(\theta)$ for $S(\theta)$ to be well defined. In the following, we assume this to be the case wherever we use $S(\theta)$. If there are multiple $a(\theta)$ for some θ, we can define $S(\theta)$ to be minimum of the above expression across all $a(\theta)$.

Lemma 2.2: The optimal value of $\mathcal{P}_1$ is $1/\theta^*$, where $\theta^* = \min\{\theta : \gamma(\theta) \leq \theta\}$.

Proof: We first show that $1/\theta^*$ is achievable. Recall that $a(\theta^*)$ denotes the solution to $\mathcal{P}_3^{\theta^*}$. Clearly, $\sum_{j \neq i} \frac{a_{ij}(\theta^*)R_j}{NW_{ij}\theta^*} \leq 1$ and $\sum_i \alpha_i(\theta^*) \leq 1$. Also, $a(\theta^*)$ satisfies the constraints (1) and (2). Thus, the optimal value of $\mathcal{P}_1$ is atleast $1/\theta^*$.

Next, suppose $\xi' > 1/\theta^*$ is the optimal value of $\mathcal{P}_1$ and (a', T', α', ξ') is an optimal solution. Clearly, a' is also a feasible solution to $\mathcal{P}_3$ with the right hand side of (8) replaced by $1/\xi'$ and $\gamma(1/\xi') \leq 1/\xi'$. But, since $1/\xi' < \theta^*$, $\gamma(1/\xi') > 1/\xi'$ leading to a contradiction. Therefore, $1/\theta^*$ is the optimal value of $\mathcal{P}_1$. ■

Remark 2.2: Define $\theta' := \sup\{\theta : \gamma(\theta) \geq \theta\}$. In general, θ^* can be strictly less than $\gamma(\theta')$. In that case $1/\gamma(\theta')$ is not the optimal value of $\mathcal{P}_1$.

Consider Fig. 2. Let $b_k, k \geq 1$ be successive jump points as shown in the figure. Then, we have the following:

Lemma 2.3: If $\mathcal{P}_3^\theta$ has unique solution for all $\theta \in [b_k, b_{k+1})$, then $S(\theta) = b_k, \forall\, \theta \in [b_k, b_{k+1})$.
Please refer [18] for the proof.

Theorem 2.1: (i) For all $k \geq 0$, $\bar{\theta}_k \geq \underline{\theta}_k, \gamma(\underline{\theta}_k) \geq \underline{\theta}_k$ and $\gamma(\bar{\theta}_k) \leq \bar{\theta}_k$.
(ii) $\bar{\theta}_k - \underline{\theta}_k, k \geq 1$ is strictly decreasing until it becomes zero. However, there exists a K such that both $\gamma(\bar{\theta}_k)$ and $\gamma(\underline{\theta}_k)$ converge in K steps.
(iii) For any $k \geq K$, if $a(\bar{\theta}_k)$ is unique, then $\theta^* = \max\{S(\bar{\theta}_k), \gamma(\bar{\theta}_k)\}$, and $a(\bar{\theta}_k)$ is an optimal association for $\mathcal{P}_1$. Moreover, if $a_{ij}(\bar{\theta}_k) = 1$, hotspot i serves mobile j $\frac{R_j}{NW_{ij}\theta^*}$ fraction of time.

Proof: (i) We prove the inequalities via induction. Clearly, $\bar{\theta}_0 = 1 \geq \gamma(1) = \underline{\theta}_0$. Suppose $\bar{\theta}_k \geq \underline{\theta}_k$ for some $k \geq 0$. Then

$$\max\left\{\underline{\theta}_k, \min\{\tilde{\theta}_k, \gamma(\tilde{\theta}_k)\}\right\} \leq \min\left\{\bar{\theta}_k, \max\{\tilde{\theta}_k, \gamma(\tilde{\theta}_k)\}\right\},$$

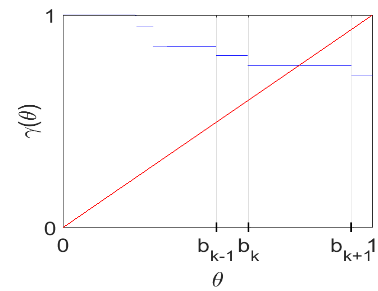
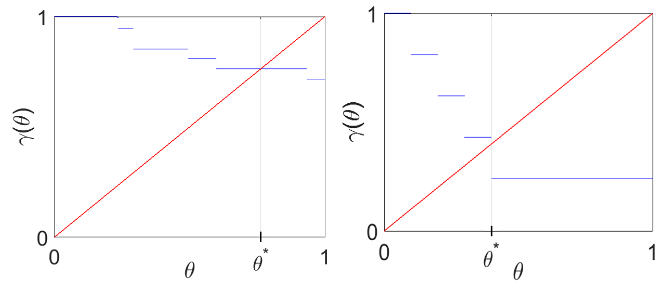


Fig. 2: $S(\theta) = b_k\ \forall\theta \in [b_k, b_{k+1})$, where $b_k, k \geq 1$ are successive jump points.



(a) (b)

Fig. 3: $\gamma(\theta)$ vs θ. In (a) for $k \geq K, S(\bar{\theta}_k) < \gamma(\bar{\theta}_k)$ and $\theta^* = \gamma(\bar{\theta}_k)$. In (b) for $k \geq K, S(\bar{\theta}_k) > \gamma(\bar{\theta}_k)$ and $\theta^* = S(\bar{\theta}_k)$.

which proves the inequality for $k + 1$, thus completing the induction step.

Next, since $\gamma(\theta)$ is nonincreasing, $\gamma(\underline{\theta}_0) \geq \gamma(1) = \underline{\theta}_0$. Suppose $\gamma(\underline{\theta}_k) \geq \underline{\theta}_k$ for some $k \geq 0$. Also note that $\tilde{\theta}_k \geq \underline{\theta}_k$. Now there can be three cases:

(a) $\gamma(\tilde{\theta}_k) \geq \tilde{\theta}_k$: In this case $\underline{\theta}_{k+1} = \tilde{\theta}_k$, and so, the desired inequality holds for $k + 1$.
(b) $\underline{\theta}_k \leq \gamma(\tilde{\theta}_k) \leq \tilde{\theta}_k$: In this case $\underline{\theta}_{k+1} = \gamma(\tilde{\theta}_k)$. Again, using the fact that $\gamma(\theta)$ is nonincreasing, $\gamma(\underline{\theta}_{k+1}) = \gamma(\gamma(\tilde{\theta}_k)) \geq \gamma(\tilde{\theta}_k) = \underline{\theta}_{k+1}$, which is the desired inequality for $k + 1$.
(c) $\gamma(\tilde{\theta}_k) \leq \underline{\theta}_k \leq \tilde{\theta}_k$: In this case, $\underline{\theta}_{k+1} = \underline{\theta}_k$. So there is nothing to prove.

Thus the inequality holds for all k. The last inequality can also be shown similarly.

(ii) Clearly, $\underline{\theta}_{k+1} \geq \underline{\theta}_k$ and $\bar{\theta}_{k+1} \leq \bar{\theta}_k$ for all $k \geq 0$. We show that at least one of the inequalities is strict for each k until $\bar{\theta}_k - \underline{\theta}_k = 0$. For any such k, $\underline{\theta}_k < \tilde{\theta}_k < \bar{\theta}_k$. Now, there can be cases:

(a) $\gamma(\tilde{\theta}_k) \geq \tilde{\theta}_k$: In this case, $\underline{\theta}_{k+1} = \tilde{\theta}_k > \underline{\theta}_k$.
(b) $\gamma(\tilde{\theta}_k) \leq \tilde{\theta}_k$: In this case, $\bar{\theta}_{k+1} = \tilde{\theta}_k < \bar{\theta}_k$.

Thus $\Delta\theta_k := \bar{\theta}_k - \underline{\theta}_k$ goes arbitrarily close to zero. However, notice that $\gamma(\theta)$ is a stepwise decreasing function. Hence, depending on the smallest step width (this is positive since there are finitely many positive step widths) there exists a threshold on $\Delta\theta_k$, and hence on k, beyond which variation in $\underline{\theta}_k$ and $\bar{\theta}_k$ do not affect $\gamma(\underline{\theta}_k)$ and $\gamma(\bar{\theta}_k)$. This threshold on k is referred to as K.

(iii) From *Lemma* 2.3, $S(\theta) = \min\{\theta' : \gamma(\theta') = \gamma(\theta)\}$. From Fig. 3 if $\gamma(\theta) = \theta$ has a fixed point, then for $k \geq K, S(\bar{\theta}_k) < \gamma(\bar{\theta}_k)$ and $\theta^* = \gamma(\bar{\theta}_k)$. On the other hand, when $\gamma(\theta) = \theta$ does not have a fixed point, for $k \geq K, S(\bar{\theta}_k) >$

$\gamma(\bar{\theta}_k)$ and $\theta^* = S(\bar{\theta}_k)$. Thus, $\theta^* = \max\{S(\bar{\theta}_k), \gamma(\bar{\theta}_k)\}$. Also, from part (ii) for $k \geq K$, $\gamma(\bar{\theta}_k) = \gamma(\theta^*)$. Hence, $a(\bar{\theta}_k)$, which is feasible for $\mathcal{P}_3^{\theta^*}$ is also optimal. Further, it is also an optimal association for $\mathcal{P}_1$. Also, from *Lemma* 2.2, each mobile j gets a throughput of $\frac{R_j}{N\theta^*}$, which is obtained by hotspot i serving mobile j for which $a_{ij}(\bar{\theta}_k) = 1$ for $\frac{R_j}{N\theta^* W_{ij}}$ fraction of time. ■

Remark 2.3: Observe that we do not know K in practice. Let $\theta_k = \max(S(\bar{\theta}_k), \gamma(\bar{\theta}_k))$, for $k \geq 1$. Then, $(a(\bar{\theta}_k), R/\theta_k, \alpha(\theta_k))$ is a feasible solution to $\mathcal{P}_1$ and yields a value $1/\theta_k$ for all $k \geq 1$. In fact, numerical results suggest that $K \leq 5$ for the parameter values of interest, i.e., we get the optimal solution of $\mathcal{P}_1$ in at most 5 iterations.

III. Belief Propagation Algorithm

Observe that $\mathcal{P}_3$ is an instance of CFLP. Consider a CFLP consisting of M facilities, $F_1, F_2, \ldots, F_M$, and N customers, $C_1, C_2, \ldots, C_N$. The facilities have associated opening costs, $f_1, \ldots, f_M$, and supply constraints (i.e., capacities), $S_1, \ldots, S_M$. The customers have demands denoted as $d_1, \ldots, d_N$. A facility, if opened, can serve a set of customers whose aggregate demand does not exceed the facility's capacity. Finally, if a facility i serves a customer j, a service cost g_{ij} is also incurred. The CFLP deals with the joint problem of facility opening and allocation to the customers, with its aim being minimizing the total cost. It can be expressed as the following optimization problem.

$$\begin{aligned}
\mathcal{P}_4 : \text{minimize} \quad & \sum_i \sum_j x_{ij} g_{ij} + \sum_i y_i f_i \\
\text{subject to} \quad & \sum_i x_{ij} = 1,\ j = 1, \ldots, N \\
& x_{ij} \leq y_i,\ i = 1, \ldots, M,\ j = 1, \ldots, N \qquad (9) \\
& \sum_j x_{ij} d_{ij} \leq S_i,\ i = 1, \ldots, M \qquad (10) \\
& x_{ij}, y_i \in \{0, 1\},\ i = 1, \ldots, M,\ j = 1, \ldots, N. \qquad (11)
\end{aligned}$$

In $\mathcal{P}_3$, the facilities and the customers come from the same set (in particular, $M = N$),

$$g_{ij} = \begin{cases} \frac{R_j}{NR_i} & \text{if } j \neq i \\ 0 & \text{if } j = i, \end{cases} \qquad f_i = \frac{1}{N}, i = 1, \ldots, N,$$

$$d_{ij} = \begin{cases} \frac{R_j}{NW_{ij}} & \text{if } j \neq i \\ 0 & \text{if } j = i, \end{cases} \quad \text{and} \qquad S_i = 1, i = 1, \ldots, N.$$

Next, we present a belief propagation based algorithm for the general CFLP, which we also use to solve $\mathcal{P}_3$.

Remark 3.1: In the absence of constraints (9), $\mathcal{P}_4$ becomes the *generalized assignment problem* [27], whereas in the absence of (10), it becomes UFLP [13]. The authors in [27] and [13] have proposed belief propagation based algorithms for the respective problems. Those algorithms can be deduced from our algorithm for CFLP.

A. The Graphical Model

We pose $\mathcal{P}_3$ as a joint probability maximization problem in a *pairwise Markov random field*. A pairwise Markov random field consists of a graph, where each node is associated with a random variable (please see [26] for the general model and algorithm). For example, in CFLP,

(a) the facilities and the customers constitute the nodes,
(b) there is an edge between every facility and every customer,
(c) for each facility F_i, the associated random variable is the set valued variable representing the set of customers served by F_i.
(d) for each customer C_j, the associated random variable represents the facility serving C_j, i.e., it takes values in $\{1, \ldots, M\}$.

The joint distribution of the variables can be factored into terms consisting of one or two variables (in the latter cases the variables correspond to nodes that are connected in the underlying graph). Further, the joint probability maximization (or, minimization) can be decomposed into optimization of marginal probabilities, also called *beliefs*. An iterative message passing algorithm, called *belief propagation* is used for efficient computation of these beliefs. The algorithm provably converges to the correct beliefs in singly connected graphs but has empirically shown excellent performance for many general graphs also. The graph underlying CFLP clearly has many loops. In this case, the joint distribution takes the following form. For $\psi_1, \ldots, \psi_M \in 2^{[N]}$ and $l_1, \ldots, l_N \in [M]$,

$$\begin{aligned}
p(&\psi_1, \ldots, \psi_M, l_1, \ldots, l_N) \\
&= Z \prod_{i,j} \delta_{C_j, F_i}(l_j, \psi_i) \prod_j \beta_j(l_j) \prod_i \phi_i(\psi_i)
\end{aligned}$$

where

$$\delta_{C_j, F_i}(l_j, \psi_i) = \begin{cases} 1, & \text{if } l_j = i, C_j \in \psi_i \\ 1, & \text{if } l_j \neq i, C_j \notin \psi_i \\ \infty, & \text{otherwise} \end{cases}$$

$$\beta_j(l_j) = e^{g_{j l_j}}$$

$$\phi_i(\psi_i) = \begin{cases} e^{f_i}, & \text{if } 0 < \sum_{q \in \psi_i} d_{iq} \leq S_i \\ 1, & \text{if } \sum_{q \in \psi_i} d_{iq} = 0 \\ \infty, & \text{otherwise} \end{cases}$$

and Z is the normalizing constant.

B. Message Passing Algorithms

Now we describe the message passing algorithm for computation and optimization of beliefs. In each iteration, all the facilities send messages to all the customers and viceversa. Let $m^k_{C_j \to F_i}$ be the message from customer C_j to facility F_i at kth iteration:

$$m^k_{C_j \to F_i} = \left(m^k_{C_j \to F_i}(\psi), \psi \subset [N]\right).$$

Similarly, let $m^k_{F_i \to C_j}$ be the message from facility F_i to customer C_j at kth iteration:

$$m^k_{F_i \to C_j} = \left(m^k_{F_i \to C_j}(l), l = 1, \ldots, M\right)$$

Let $b^k_{F_i}$ and $b^k_{C_j}$ be facility F_i's and customer C_j's belief vectors, respectively, and $a^k_{C_j}$ be customer C_j's choice at the end of kth iteration. Finally, for each facility F_i, let $\mathcal{F}_i$ be the collection of sets of customers that can be served by F_i:

$$\mathcal{F}_i = \left\{ \psi \subset [N] : \sum_{j \in \psi} a_{ij} d_{ij} \leq S_i \right\}.$$

BP Algorithm

1) *Initialization:*
$$m^0_{C_j\to F_i} = m^0_{F_i\to C_j} = 0$$
2) *Messages at kth iteration:*
$$m^k_{C_j\to F_i}(\psi) = \min_l \delta_{C_j,F_i}(l,\psi)\left[\sum_{p\neq i} m^{k-1}_{F_p\to C_j}(l) + g_{lj}\right]$$
$$m^k_{F_i\to C_j}(l) = \min_{\psi\in\mathcal{F}_i} \delta_{C_j,F_i}(l,\psi)\left[\sum_{p\neq j} m^{k-1}_{C_p\to F_i}(\psi) + f_i \mathbb{1}_{\{\psi\neq\emptyset\}}\right]$$
3) *Belief at kth iteration:*
$$b^k_{F_i}(\psi) = f_i \mathbb{1}_{\{\psi\neq\emptyset\}} + \sum_p m^k_{C_p\to F_i}(\psi)$$
$$b^k_{C_j}(l) = g_{lj} + \sum_p m^k_{F_p\to C_j}(l)$$
4) *Assignment at the end of kth iteration:*
$$a^k_{C_j} = \operatorname*{argmin}_l \{b^k_{C_j}(l)\}$$

1) A Simplified Algorithm: Note that the dimension of the vector $m^k_{C_j\to F_i}$ in the above algorithm is 2^N (i.e., exponential in N). Also, to compute each entry for the message $m^k_{F_i\to C_j}$, one needs to compare up to 2^N subsets of customers if facility i has enough capacity to serve all the customers. Consequently, the algorithm has exponential running time.

We now simplify the previous message passing algorithm to a pseudo-polynomial one. In particular, the facilities and customers send scalar messages in this algorithm.

Simplified BP Algorithm

1) *Initialization:*
$$\mu^0_{C_j\to F_i} = \mu^0_{F_i\to C_j} = 0$$
2) *Messages at kth iteration:*
$$\mu^k_{C_j\to F_i} = -g_{ij} - \max_{l\neq i}(\mu^{k-1}_{F_l\to C_j} - g_{lj})$$
$$\mu^k_{F_i\to C_j} = \max_{\substack{\psi\in\mathcal{F}_i:\\ j\in\psi}}\left[\sum_{p\in\psi:p\neq j} \mu^{k-1}_{C_p\to F_i}\right] - \max\left\{\max_{\substack{\psi\in\mathcal{F}_i:\\ j\notin\psi,\psi\neq\emptyset}}\left[\sum_{p\in\psi}\mu^{k-1}_{C_p\to F_i}\right], f_i\right\}$$
3) *Belief at kth iteration:*
$$\tilde{b}^k_{C_j}(l) = g_{lj} - \mu^k_{F_l\to C_j}$$
4) *Assignment at the end of kth iteration:*
$$\tilde{a}^k_{C_j} = \operatorname*{argmin}_l \{\tilde{b}^k_{C_j}(l)\}$$

For brevity, we omit the proof of equivalence of this algorithm with the original one.

2) Damped Message Passing: Note that the above message passing algorithms are not guaranteed to converge. A common approach to deal with message oscillations is to use damped messages. In damped message passing, the updates at each iteration are calculated from the messages of the previous iteration exactly as before. But the new messages are weighted averages of the old messages and updates. More specifically, the damped version of the above simplified message passing algorithm works as follows. Suppose $\{\mu^{k-1}_{F_i\to C_j}, \mu^{k-1}_{C_j\to F_i}, i = 1,\ldots,M, j = 1,\ldots,N\}$ are the messages passed at $(k-1)$th iteration ($k \geq 1$). Also, let $\{\tilde{\mu}^k_{F_i\to C_j}, \tilde{\mu}^k_{C_j\to F_i}, i = 1,\ldots,M, j = 1,\ldots,N\}$ denote the updates at kth iteration. These are functions of the messages at $(k-1)$th iteration and are given by the same expressions as in Simplified BP Algorithm. However, the new messages from any node to any other node are computed as follows
$$\mu^k_{\cdot\to\cdot} = \lambda\mu^{k-1}_{\cdot\to\cdot} + (1-\lambda)\tilde{\mu}^k_{\cdot\to\cdot},$$
where $\lambda \in [0,1)$ is the dampness parameter. A larger value of λ increases computational stability (i.e., instances of convergence) but yields inferior solutions.

IV. Offloading in Presence of Rational Mobiles

When mobiles become hotspots and serve other mobiles, they spend resources (e.g., energy, memory etc.) Therefore mobiles should have a rational basis for acting as hotspots. This aspect can be modeled in several ways; following are a few examples.

Each mobile that acts as a hotspot can be given added throughput where the extra throughput is proportional to the throughput that the tagged mobile offers to other mobiles. We consider this scheme in Section IV-A. We show that the resulting optimization problems have similar structure as those in Section II-B.

In yet another framework, each mobile that becomes a hotspot can be paid off in terms of free data usage in proportion to the throughput that it delivers to other mobiles. On the other hand, the mobiles may have utility for extra throughput and may even pay of the same. Here, the mobiles would be interested in optimizing their net payoffs. However, we do not incorporate utilities for data usage and throughputs in our framework, and do not pursue this viewpoint.

A. Incentivizing Hotspots

Suppose each hotspot gets additional throughput that is a factor η of the throughput that it delivers to other mobiles. We can study both, throughput maximization and proportional increment problems under this strategy. For example, the throughput maximization problem will be same as problem $\mathcal{P}_2$ with constraint (7) replaced by
$$T_i \geq \frac{R_i}{N} + \eta\sum_{j\neq i} a_{ij}T_j.$$

We can argue that it is equivalent to the following problem that has only ($a = (a_{ij}, i,j = 1,\ldots,N)$ as decision variables.
$$\mathcal{P}_5 : \text{minimize} \quad (1+\eta)\sum_i\sum_{j\neq i} a_{ij}\frac{R_j}{NR_i} + \sum_i a_{ii}\frac{1}{N}$$
$$\text{subject to} \quad (1),(2),(8)$$

In particular, a solution a^* to this problem yields a solution to the throughput maximization problem. This can be argued as in Lemma 2.1. We omit the details for brevity.

V. Offloading using DCF

Solving for throughput optimal (or, fair) offloading using DCF is a more complex problem. We focus on the throughput maximization to illustrate the difficulty. We also show how the problem can be converted to a relatively simpler mixed-integer linear integer problem.

The throughput optimization problem can be expressed as:

$$\begin{aligned} \mathcal{P}_6 : \text{maximize} \quad & \sum_j T_j \\ \text{subject to} \quad & T_j \geq \frac{R_j}{N}, \ j = 1, \ldots, N \\ & (1), (2), (3), (4), (6) \end{aligned}$$

Notice that the rate constraints for DCF based access, given by (6), are more complex than the corresponding constraints for PCF based access. We now develop equivalent mixed-integer linear constraints. First, observe that the mobiles' rates must satisfy

$$a_{ij} T_j \leq \frac{1}{\sum_{l=1}^{N} a_{il} \frac{1}{W_{il}}}, \forall i, j, i \neq j \tag{12}$$

which can be also written as

$$\sum_l a_{il} \frac{1}{W_{il}} \leq \frac{a_{ij}}{T_j} + (1 - a_{ij}) \sum_l \frac{1}{W_{il}} \tag{13}$$

Next, following the arguments similar to Lemma 2.1, we can show that all the mobiles but the first one should be given their least throughput. But under DCF, all the mobiles connecting to the same hotspot must get equal throughput. Thus the optimization problem can be written as

$$\begin{aligned} \text{minimize} \quad & \sum_i \sum_{j \neq i} \frac{a_{ij}}{R_i} \max_{l \neq i} \left\{ \frac{a_{il} R_l}{N} \right\} + \sum_i \frac{a_{ii}}{N} \\ \text{subject to} \quad & (1), (2), (13), \end{aligned}$$

where we set $T_j = R_j/N$ in (13). Here, the objective function is a nonlinear function. Let θ_i be the aggregate throughput delivered by i to other mobiles. Then, the above objective can also be expressed as

$$\begin{aligned} \text{minimize} \quad & \sum_i \frac{\theta_i}{R_i} + \sum_i \frac{a_{ii}}{N} \\ \text{subject to} \quad & \theta_i \geq \left(\sum_{l \neq i} a_{il} \right) \frac{R_j}{N} - (1 - a_{ij}) B, \forall i, j, i \neq j, \\ & (1), (2), (13), \end{aligned}$$

where B is large constant and $T_j = R_j/N$ in (13).

Discussion: Clearly, optimal offloading using DCF is a much more difficult problem than optimal offloading using PCF. Moreover, notice that rate constraints (12) could be strict inequalities for a few hotspots in the optimal solution. For any such hotspot i, in view of (6), $w_{ij} < W_{ij}$ for at least a few mobiles connecting to i. In other words, at least a few hotspots and mobiles operate at less than their maximum feasible rates (they can do so via appropriate choices of modulation and coding schemes). This reemphasizes why DCF is not the preferred mode of access for optimal offloading.

VI. Performance Evaluation

We now illustrate performance of the proposed algorithms and benefits of mobile assisted offloading through numerical evaluation. We assume that the mobiles' base cellular rates ($R_i, i = 1, \ldots, N$) are in the range $100 - 250$Mbps. On the other hand, WiFi rates between the mobiles and the hotspots (W_{ij}s) are considered to be in the range $50 -$ 100Mbps. We perform numerical evaluation for several values of N, namely, $N = 5, 10, 15, 20$ and 25. For each N, we consider 100 sets of values of rates, and evaluate the performance of the proposed algorithms and offloading for each set of parameters. The rates for different mobiles and for different evaluations are chosen to be independent and uniformly distributed in the respective ranges. We now report the average performance measures.

We first present the performance of the proposed message passing algorithm. We use two values of the dampness parameter, $\lambda = 0.7$ and $\lambda = 0.8$. Table I shows the number of times the damped message passing algorithm converges (out of 100 runs) and also shows the average errors in the objective value at the limiting point compared to the optimal objective value. Expectedly, the higher value of λ (i.e., $\lambda = 0.8$) increases the chances of convergence and still gives satisfactory results.

In practice, one may not want to wait until the message passing algorithm converges. We can rather terminate the algorithm if the objective function does not improve substantially in a few successive iterations - this requires defining suitable thresholds on differences of successive objective values. We have found that such heuristics lead to quicker termination and still yield satisfactory performance. In most of the instances the message passing algorithm took less than 20-50 iterations (depending on the values of N and λ) to converge and less than 10 iterations to give satisfactory results.

We next focus on the benefits of offloading. Table II shows the improvement in the aggregate throughput when we aim at maximizing it. It also shows the average throughput improvement of each mobile when pursuing proportional increment optimization. An interesting observation is that one does not lose much in terms of the aggregate throughput while being fair - per mobile throughput improvement in the proportional increment case are of the same order as the maximum possible throughput enhancement. We also see that, in either case, the benefit of offloading increases with the number of mobiles.

Finally, we evaluate the mechanism of Section IV-A where we incentivize the mobiles for acting as hotspots. Here, we focus on aggregate throughput optimization. We see that as the incentivization factor η increases, the improvement in the average percentage gain diminishes but this degradation is not substantial (see Table III). Even for $\eta = 0.2$, we observe marked improvement in the average percentage gain.

N		5	10	15	20	25
$\lambda = 0.7$	No. of times converged	100	98	90	71	68
	% error	2.84	3.18	2.45	1.24	0.85
$\lambda = 0.8$	No. of times converged	100	99	96	92	90
	% error	5.43	4.07	3.74	2.83	2.06

TABLE I: Performance of the message passing algorithm.

N	5	10	15	20	25
(a) % gain in aggregate throughput	25.47	33.42	36.59	38.45	38.62
(b) % gain in mobiles' throughputs	22.40	29.48	33.47	35.62	37.98

TABLE II: Benefit of offloading in (a) throughput optimization and (b) proportional increment optimization.

N	5	10	15	20	25
$\eta = 0.1$	24.86	32.61	35.68	37.58	38.15
$\eta = 0.2$	23.89	31.03	33.75	35.55	36.66

TABLE III: Benefit of offloading while incentivizing hotspots.

N	5	10	15	20	25
Using MATLAB **intlinprog** solver	20.20	30.41	—	—	—
Using Local search heuristic	18.93	25.98	28.81	29.91	30.58

TABLE IV: % gain in aggregate throughput for offloading using DCF.

The performance of offloading using DCF is given in Table IV. The corresponding optimization problem can be solved using MATLAB **intlinprog** solver. We have shown results for $N = 5, 10$, in Table IV but the solver takes excessively long time for higher values of N. Thus we use a local search heuristic to obtain an approximate solution to this optimization problem [25]. Such a heuristic is used in [3] to solve a similar but simpler access point association problem. This iterative method is much faster than the MATLAB **intliprog** solver and gives satisfactory results[2]. As we expect, the average percentage gain in the aggregate throughput is lesser when compared to the offloading using PCF. However, the performance gain is still noticeable.

VII. Conclusion

We have studied mobile assisted offloading in cellular networks where mobiles are capable of acting as hotposts and serving other mobiles. We have studied throughput optimal offloading and also a fair offloading strategy which we call *proportional increment offloading*.We show that the former problem can be reduced to a *capacitated facility location problem* (CFLP) whereas the latter can be solved by solving a sequence of CFLPs (Theorem 2.1). Then, we have proposed a belief propagation based algorithm to solve general CFLPs (Section III-B). This generalizes analogous algorithms for generalized assignment problems and UFLPs. We have also argued why DCF based WiFi access is not preferred for the proposed offloading mechanism.

Our future work consists of investigating other notions of fairness in the context of offloading. In particular, we will focus on solution concepts from the theory of coalitional games. We also plan to study properties of the proposed belief propagation based algorithm.

References

[1] IEEE 802.11ah — Wikipedia, the free encyclopedia, 2014.

[2] Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020. White Paper *(Available online)*, Feb 2016.

[2]For instance, for 20 nodes, **intlinprog** takes a few days whereas the local search heuristic gives an approximate solution in a few seconds.

[3] Mohammed Amer, Anthony Busson, and Isabelle Guérin Lassous. Association optimization in wi-fi networks: Use of an access-based fairness. In *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 119–126. ACM, 2016.

[4] J.E. Beasley. An algorithm for solving large capacitated warehouse location problems. *European Journal of Operational Research*, 33(3):314 – 325, 1988.

[5] X. Chen, J. Wu, Y. Cai, H. Zhang, and T. Chen. Energy-efficiency oriented traffic offloading in wireless networks: A brief survey and a learning approach for heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(4):627–640, April 2015.

[6] Donald Erlenkotter. A dual-based procedure for uncapacitated facility location. *Operations Research*, 26(6):992–1009, 1978.

[7] Lin Gao, George Iosifidis, Jianwei Huang, and Leandros Tassiulas. Hybrid data pricing for network-assisted user-provided connectivity. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 682–690. IEEE, 2014.

[8] G. Iosifidis, L. Gao, J. Huang, and L. Tassiulas. An iterative double auction for mobile data offloading. In *International Symposium on Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt)*, pages 154–161, May 2013.

[9] George Iosifidis, Lin Gao, Jianwei Huang, and Leandros Tassiulas. Enabling crowd-sourced mobile internet access. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 451–459. IEEE, 2014.

[10] O. B. Karimi, J. Liu, and J. Rexford. Optimal collaborative access point association in wireless networks. In *Proceedings of IEEE INFOCOM*, pages 1141–1149, April 2014.

[11] Jakob Krarup and Peter Mark Pruzan. The simple plant location problem: Survey and synthesis. *European Journal of Operational Research*, 12(1):36–57, 1983.

[12] Anurag Kumar and Vinod Kumar. Optimal association of stations and APs in an IEEE 802.11 WLAN. In *11th National Conference on Communications*, pages 145–149, Jan 2005.

[13] Nevena Lazic, Brendan J Frey, and Parham Aarabi. Solving the uncapacitated facility location problem using message passing algorithms. In *Proceedings of AISTATS*, pages 429–436, 2010.

[14] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong. Mobile data offloading: How much can wifi deliver? *Proceedings of IEEE/ACM Transactions on Networking*, 21(2):536–550, April 2013.

[15] Haiyun Luo, Ramachandran Ramjee, Prasun Sinha, Li Erran Li, and Songwu Lu. Ucan: a unified cellular and ad-hoc network architecture. In *Proceedings of international conference on Mobile computing and networking*, pages 353–367. ACM, 2003.

[16] V. Mittal, S. K. Kaul, and S. Roy. On optimal hotspot selection and offloading. In *Proceedings of IEEE International Conference on Communications (ICC)*, pages 1–6, May 2016.

[17] M. H. Qutqut, F. M. Al-Turjman, and H. S. Hassanein. Mfw: Mobile femtocells utilizing wifi: A data offloading framework for cellular networks using mobile femtocells. In *2013 IEEE International Conference on Communications (ICC)*, pages 6427–6431, June 2013.

[18] Divya R., Amar Prakash Azad, and Chandramani Singh. Fair and optimal mobile assisted offloading. Technical report, 2017. http://www.dese.iisc.ernet.in/people/chandra/publications/divya-etal17mobile-assisted-offloading.pdf.

[19] F. Rebecchi, M. Dias de Amorim, V. Conan, A. Passarella, R. Bruno, and M. Conti. Data offloading techniques in cellular networks: A survey. *IEEE Communications Surveys Tutorials*, 17(2):580–603, 2015.

[20] Sujay Sanghavi, Dmitry Malioutov, and Alan Willsky. Networking sensors using belief propagation. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 384–391. IEEE, 2008.

[21] D. Shehadeh, N. Montavont, T. Kerdoncuff, and A. Blanc. Minimal access point set in urban area wifi networks. In *International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 221–228, May 2015.

[22] A. Srinivas and E. Modiano. Joint node placement and assignment for throughput optimization in mobile backbone networks. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, 2008.

[23] Tony J. van Roy. A cross decomposition algorithm for capacitated facility location. *Operations Research*, 34(1):145–163, 1986.

[24] Vedat Verter. *Uncapacitated and Capacitated Facility Location Problems*, pages 25–37. Springer US, Boston, MA, 2011.

[25] Joachim Paul Walser. *Integer optimization by local search: a domain-independent approach*. Springer-Verlag, 1999.

[26] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Exploring artificial intelligence in the new millennium. chapter Understanding Belief Propagation and Its Generalizations, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[27] Mindi Yuan, Chong Jiang, Shen Li, Wei Shen, Yannis Pavlidis, and Jun Li. *Message Passing Algorithm for the Generalized Assignment Problem*, pages 423–434. Springer Berlin Heidelberg, 2014.