Cost-Effective Multi-Mode Offloading with peer-assisted communications

CrossMark

Ioannis Komnios [a,*], Fani Tsapeli [a], Sergey Gorinsky [b]

[a] Democritus University of Thrace, Xanthi, Greece
[b] IMDEA Networks Institute, Madrid, Spain

ARTICLE INFO

ABSTRACT

Data offloading through WiFi networks has been identified as a promising solution to cellular network congestion caused by the ongoing explosive growth in mobile data traffic. In this paper, we propose Cost-Effective Multi-Mode Offloading (CEMMO) that enhances offloading with multi-hop peer-assisted communications regardless of content and popularity. CEMMO enables three modes of communication: cellular delivery, delay-tolerant offloading, and peer-assisted offloading. Exploiting user knowledge on mobility and WiFi connectivity, CEMMO assists the cellular operator in selecting the best out of its three modes in order to reduce the overall cost in terms of financial settlement, energy consumption, and user satisfaction. Our simulations with a realistic mobility and connectivity prediction model based on a Markov process show that CEMMO offloads up to 59% of the mobile data traffic and reduces transfer cost per MB up to 16% over delay-tolerant offloading. This paper also discusses practical issues in CEMMO adoption.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Due to the recent proliferation of smartphones and tablets, mobile data traffic has been explosively growing and pushing the capacity limits of cellular networks In particular, smartphones equipped with high-resolution cameras are sold globally at astounding rates, and content generated by mobile users significantly rises fueled by popular social networking and cloud storage applications such as Facebook, and Dropbox. With the global mobile data traffic forecasted to increase 13 times within the next 5 years [1], cellular network providers need to find cost-effective ways to handle the growing traffic demands with expected Quality of Service (QoS) levels. Infrastructure upgrade is the intuitive solution to deal with the congestion problem; however, it is insufficient as a financially sustainable solution. Offloading mobile data through WiFi networks is the most promising solution to tackle the congestion problem because the cost of WiFi access points (APs) required for the offloading is relatively low, WiFi technology uses unlicensed bands and offloading can satisfactorily serve delay-tolerant types of data traffic. So far, research efforts have been focused on either on-the-spot offloading (OTSO) [2], where data is offloaded only over an immediately available WiFi connection, or delay-tolerant offloading (DTO) [3], where the transmission is delayed for some time in case the user encounters an offloading opportunity later. Both offloading types offer significant benefits, mainly in environments with stable WiFi coverage.

In this paper, we target dense urban environments and propose Cost-Effective Multi-Mode Offloading (CEMMO) that enhances OTSO and DTO with a multi-hop peer-assisted offloading mode (PAO), where the offloaded traffic

---

* Corresponding author.
   E-mail addresses: ikomnios@ee.duth.gr (I. Komnios), ttsapeli@ee.duth.gr (F. Tsapeli), sergey.gorinsky@imdea.org (S. Gorinsky).

is delivered through intermediate mobile devices. To the best of our knowledge, PAO is the first peer-assisted offloading solution that is implemented not only for offloading data from the uplink, but also independent of its content or popularity. Given the delay-tolerant nature of a significant portion of the cellular network traffic, an offloading solution that enhances the existing delay-tolerant functionality with peer-assisted offloading (PAO) will leverage the overall offloading capability and, therefore, increase the total amount of data that is offloaded from cellular networks. For PAO, we develop a new data-forwarding scheme that uses limited flooding. By applying a mobility and connectivity prediction model based on a Markov process, CEMMO allows the cellular operator to select the most effective mode for communicating mobile data with different delay constraints based on the estimated costs of delivery through each of the three modes, i.e. cellular delivery, DTO or PAO. OTSO is always exploited when available. The cost reduction leads to lower prices that attract new users, increasing the revenue of the operator. In scenarios with extended WiFi coverage, CEMMO offloads up to 59% of the mobile data traffic and reduces transfer cost per MB up to 16% over DTO. With an energy optimization, CEMMO achieves up to 31% improvement on energy consumption over OTSO.

The novelty of CEMMO lies in the introduction of the first peer-assisted method for offloading traffic from the uplink, regardless of the offloaded content and its popularity, and the selection of the most cost-efficient alternative for data transmission, as defined by each operator. The contributions of the paper can be summarized as follows:

  i. We design and evaluate CEMMO, a cost-effective scheme for mobile data offloading that integrates multiple modes of operation: cellular delivery, DTO, and PAO.
  ii. We propose a mobility and connectivity prediction model based on a Markov process and we develop a forwarding scheme for PAO with low storage and energy overhead.
  iii. CEMMO allows the cellular operators to automatically select the transfer policy (i.e. cellular delivery, DTO or PAO) that provides more gains. The overall cost, as defined by each operator, can include transfer and energy costs, incentives to motivate user participation, etc.
  iv. CEMMO provides cellular operators with knowledge on the amount of data offloaded by each user. Such knowledge is currently unavailable and will help cellular operators design the expansion of their network accordingly.

The rest of the paper is organized as follows. In Section 2, we describe the proposed mobility and connectivity prediction model. Section 3 details DTO and the proposed PAO transfer policy. CEMMO mechanism is presented in Section 4, along with a sample scenario. Section 5 evaluates the performance of CEMMO through simulations. In Sections 6 and 7, we discuss several practical issues on the adoption of CEMMO and related work, respectively. Finally, we conclude the paper in Section 8.

2. Mobility and connectivity prediction model

According to our mobility and connectivity prediction model, a large area is divided into smaller *regions* with unique identifiers, and a time period (e.g. a day) is split into smaller *time intervals* (e.g. 10-min intervals). The region size and the time interval duration affect the spatio-temporal accuracy of the prediction model. Defining small regions and time intervals improves the accuracy of the model at the expense of increased computational complexity. Users store their own mobility information for each time interval. For example, if a one-hour time interval is selected, users store their mobility information for 24 distinct time intervals. In particular, users keep records of their location, their transitions from one region to another and the frequency at which they move between regions. For each user, we define:

- $N(X, t)$ as the number of visits in region X during time interval t (i.e. starts at time t),
- $N(A \rightarrow X, t)$ as the number of transitions from A to a neighboring region X during t,
- t_{duration} as the duration of each time interval and
- $t_{\text{next}} = t_{\text{duration}} + t$ as the start time of the next time interval.

The probability of a user located in region A to move to region X directly within t is:

$$P(X|A, t) = N(A \rightarrow X, t)/N(A, t) \tag{1}$$

The probability that the user stays in region X until the end of time interval t is:

$$P_{\text{stay}}(X, t_{\text{next}}) = P(X|X, t) = N(X \rightarrow X, t)/N(X, t) \tag{2}$$

User mobility during any time interval t is modeled as a Markov process. The next region that will be visited by a user during a time interval is assumed to solely depend on the previous one. Thus, if $NR(A)$ is the set of neighboring regions of A, the probability of a user to move from current region A to region X through any neighboring region i ∈ $NR(A)$ during t is:

$$P(X|A, t) = \sum_{\forall i \in NR(A)} P(i|A, t) \times P(X|i, t) \tag{3}$$

Starting from the current region (L_{current}) and time interval (t_{current}) and based on Eqs. (1) and (3), we estimate the probabilities of a user to visit any other region during any time interval $t \leqslant t_{\text{end}}$, where t_{end} is the time interval until which the user is willing to wait for the data to be offloaded through a WiFi AP. We use Eq. (2) to compute the probabilities of a user to be located in a region at the end of each time interval. We recursively use these regions as potential starting points of user movement during the next time interval and repeat this process to predict user mobility during each time interval, as described in Algorithm 1.

Each user also keeps statistics regarding WiFi connectivity within each region, since a region may only have partial WiFi connectivity. $NW(X)$ denotes the number of times WiFi connectivity was available in region X. With the selection of small time intervals, we do not need to measure the

duration of the WiFi connectivity. The probability that a user located in region X during time interval t has access to a WiFi access point is defined as:

$$P(\text{WiFi}|X) = NW(X)/N(X, t) \qquad (4)$$

defined as the probability of the source node S to have WiFi connectivity within the specified interval t_w:

$$P_{\text{DTO}_{\text{success}}} = P(E^S_{\text{WiFi},t_w}) \qquad (5)$$

Algorithm 1 Mobility prediction model

//Input: *Probabilities P as extracted from historical data, L_{current}, t_{current}, t_{end}*
//Output: *Probabilities P_{est} to visit any region for each time interval before t_{end}*
Initialize $P_{\text{est}}(\text{Region, Interval}) \leftarrow 0 \ \forall$ region and interval $\leqslant t_{\text{end}}$
$P_{\text{est}}(L_{\text{current}}, t_{\text{current}}) \leftarrow 1$
Add L_{current} to StartingPoints   *//Initial starting point is current location*
$t_i \leftarrow t_{\text{current}}$
$t_{\text{next}} \leftarrow t_i + t_{\text{duration}}$
While $t_i \leqslant t_{\text{end}}$ **do**
   For each $SP \in$ StartingPoints **do**
      For each region R: $P(R|SP, t_i) > 0$ **do**   *//For all possible transitions to another region*
      $P_{\text{est}}(R, t_i) \leftarrow P(R|SP, t_i)$ *//Based on Eq.* (1)
      For each region N: $P(N|R, t_i) > 0$ **do**
        newProb $\leftarrow P(N|R, t_i) \times P(R|SP, t_i)$ *//Based on Eq.* (3)
        $P_{\text{est}}(N, t_i) += $ newProb   *//Transitions to a region through different paths*
      End   *// are mutually exclusive and, therefore, probabilities are added*
     End
   End
   For each region R: $(P_{\text{est}}(R, t_i) > 0$ AND $P(R, t_{\text{next}}) > 0)$ **do**   *//Regions to be used as*
     Add R to StartingPoints   *//starting points during the next time interval*
     $P_{\text{est}}(R, t_{\text{next}}) \leftarrow P_{\text{stay}}(R, t_{\text{next}}) \times P_{\text{est}}(R, t_i)$ *//P_{stay} is based on Eq.* (2)
   End
   $t_i \leftarrow t_{\text{next}}$
   $t_{\text{next}} \leftarrow t_i + t_{\text{duration}}$
End

---

3. Delay-tolerant and peer-assisted offloading transfer policies

In the following subsections, we present the delay-tolerant (DTO) and the peer-assisted offloading (PAO) transfer policies based on the proposed mobility and connectivity prediction model. We assume that each node (i.e. mobile device) is equipped with both IEEE 802.11 and 3G wireless interfaces that can run simultaneously, is capable of generating content for upload and is willing to accept certain delays in applications, such as video upload, without sacrificing too much user satisfaction. We define R as the set of regions that the source node S is likely to visit before t_{end} and U the set of nearby users. We define the following events:

$E^S_{r,t_S} \rightarrow$ source node S visits region r

    $\in R$ during time interval t_S

$E^u_{r,t} \rightarrow$ user $u \in U$ visits region r

    $\in R$ during time interval t

$E^u_{\text{WiFi},t_w} \rightarrow$ user $u \in U$ has WiFi access during t_w

3.1. Delay-tolerant offloading

According to this policy, mobile users attempt to transfer their data directly through a WiFi network within a specific timeframe. The success probability of DTO is

For data upload requests, the time interval t_w needs to be prior to t_{end}. Thus: $t_w \leqslant t_{\text{end}}$.

3.2. Peer-assisted offloading

PAO transfer policy is built around the concept of storing data within a specific region for a short time interval, in a way that any mobile user who enters this region during this interval receives a replica of the data, and utilizing other mobile users as intermediate data carriers to offload data through a WiFi AP. Data storage enables the data exchange between the source node and intermediate users using ad-hoc technologies such as WiFi Direct. The cellular operator divides the overall region into sub-regions and predetermines the limits of each region using Global Positioning System (GPS) coordinates, selects the optimal storing region and time interval based on the probabilities of nearby nodes to visit any region in the near future, and notifies the source node about the GPS coordinates of the flooding region. Users can detect when they have entered the flooding region using location services such as Geofences [4].

When the source node enters the region that has been specified by the operator, replicas of the data are being flooded and all users within the region during the storing interval will receive a replica of the data. The operator can use a control channel, similar to the one used for call

setup or sending text messages, to signal if the WiFi network interface of the mobile device needs to be switched on. The flooding process is regionally and temporally restricted; nodes that have received a replica of the data are only allowed to forward replicas to other users located within the storing region during the storing interval. The geographical and temporal narrowing of this flooding process limits the total number of replicas in the network and, thus, does not create excessive overhead. CEMMO provides the operator with the flexibility to adjust the region size and, therefore, control the overall overhead. Flooding content in a large region (or a small region that is highly dense) will result in high overhead, since many users will act as relays for the data. Flooding to more regions would result in excessive overhead for the operator, making the scheme infeasible.

When users holding a replica leave the storing region, they are only allowed to upload the data directly through a WiFi AP. When an intermediate data carrier successfully uploads a data replica through a WiFi AP, he/she informs the operator, who in turn sends a notification to the source node. In case another intermediate node attempts to transfer another replica, the transfer is aborted, and the node is notified to discard its replica. All other nodes drop their replicas once the delay tolerance threshold expires. In case the source node does not receive any notification of successful upload until t_{end}, data is transferred through the available 3G network. A user u serves an upload request successfully, if he/she manages to transfer the data between the source user and the WiFi AP (i.e. by transferring a replica from the storing region):

$$E_{\text{Success}}^{u,r,t} \rightarrow E_{r,t}^{u} \cap E_{r,t_S}^{S} \cap E_{\text{WiFi},t_w}^{u}$$

For data upload requests, the time interval t, when user u visits region r, needs to be equal to or subsequent of time interval t_S. Moreover, time interval t_w, when node u has access to a WiFi network, needs to be equal to or subsequent of time interval t. Thus:

$$t_S \leqslant t \leqslant t_w \leqslant t_{end}$$

The operator selects to store data in the storing region (R_{storing}) until the expiration of the storing interval (T_{storing}), in order to maximize the probability that data will be successfully transferred within the specified delay threshold. Thus:

$$\{R_{\text{storing}}, T_{\text{storing}}\} = \operatorname{argmax}_{r,t} \ P(\cup_{u \in U} E_{\text{Success}}^{u,r,t}) \tag{6}$$

The success probability of PAO is:

$$P_{\text{PAOsuccess}} = P\left(\cup_{u \in U} E_{\text{Success}}^{u,R_{\text{storing}},T_{\text{storing}}}\right) \tag{7}$$

4. Cost-Effective Multi-Mode Offloading

4.1. CEMMO mechanism

The main assumptions of CEMMO are that all mobile nodes apply the proposed mobility and connectivity prediction model, and that the cellular operator collects data transfer requests from users, performs online predictions on user mobility and connectivity and decides on the transfer policy that minimizes cost. We assume that users perform OTSO when they have direct access to a WiFi network and are willing to accept a delay in the delivery of some non-urgent data, when proper incentives are offered to them, such as reduced service pricing and improved services [5–7]. Web server synchronization, such as Dropbox and Flickr, are typical mobile applications where CEMMO can undertake the responsibility to transfer the data within the specified Delay Tolerance Interval (DTI) (i.e. the time interval that the user is willing to wait for the data transfer), by applying one of the three data transfer policies:

  i. *Delay-Tolerant Offloading (DTO)*. CEMMO evaluates the probability of a user to have access to a WiFi AP before DTI expires.
 ii. *Peer-Assisted Offloading (PAO)*. This technique does not require direct access of the user that generates data to a WiFi network.
iii. *Transfer through the available 3G network*. We assume that 3G connectivity is always available and, therefore, the success probability of this approach is one.

Each time a mobile user needs to upload data, a request that includes (a) the data size, (b) the delay tolerance indicator (DTI) field as set by the user, (c) his/her current location and (d) his/her probabilities to visit any other region within DTI along with (e) WiFi connectivity probabilities, is sent to the cellular operator. Given the small size of the data transfer requests, this process can be accomplished through the available GSM network in order to reduce the energy consumption of the mobile devices. When the cellular operator receives a request for data transfer, CEMMO exploits the provided information to estimate the success probability of DTO. In order for the operator to estimate the success probability of PAO, a signal is sent to all users who are associated with the same cell tower requesting information on their approximate position and probability to visit other regions within DTI, along with their WiFi connectivity probabilities. Based on this feedback, the operator estimates the success probability of PAO.

Both data transfers through 3G and WiFi networks involve a cost for the operator. In regard to 3G based data transfers, this cost may correspond to the financial cost of the transfer, the associated user dissatisfaction due to increased pricing, network congestion and increased energy consumption that data transfer through 3G incurs. As far as data transfer through WiFi networks is concerned, the cost corresponds to user dissatisfaction due to the delayed transmission, as well as a financial cost for the operator, including any reward the operator needs to pay to users in order to motivate them to delay their transmissions. Data transfer through opportunistic peer-to-peer networks incurs an additional cost, since the limited resources of each user involved in the process are exploited for the transfer of data from other users. We define α as the transfer cost per MB through 3G and β as the transfer cost per MB through WiFi. Since all data that will not be successfully delivered over WiFi will eventually be

transmitted over 3G, CEMMO estimates the transfer cost of DTO and PAO as follows:

$$\text{Est. Cost}_{DTO} = \alpha \times (1 - P_{DTOsuccess}) + \beta \times P_{DTOsuccess} \quad (8)$$

$$\text{Est. Cost}_{PAO} = \alpha \times (1 - P_{PAOsuccess}) + \beta \times N \times P_{PAOsuccess} \quad (9)$$

where N is the number of participants in PAO. CEMMO selects the transfer policy that minimizes the estimated transfer cost. By increasing the β/α ratio, the cost of relaying data to users increases compared to 3G data transfer cost. Thus, less traffic is offloaded through PAO.

Our prediction model models mobility and connectivity within each region during any time interval using first-order Markov models. Markov models are ideal for mobile devices, since their CPU and storage needs are low; they only involve reading and writing individual entries in arrays. Therefore, the complexity and computational overhead from the perspective of users is marginal. From the operator side, the required signaling overhead is limited in comparison to the amount of data that is being offloaded. Moreover, it can be further reduced if all users periodically transmit their probabilities to the operator or the operator holds all collected mobility and connectivity probabilities until their time interval expires. In this case, the operator does not need to request the same probabilities from users for consecutive requests.

4.2. Sample scenario

We consider a region covered by a cellular network, as depicted in Fig. 1. According to our model, the operator divides the greater region into smaller regions (i.e. Regions 1–16). We consider that a user (Node 1), located in Region 7, has recorded a video of 100 MB using the camera of his/her mobile device and wants to upload it to a cloud storage service within the next 30 min. Node 1 sends a request to the cellular operator that includes the data size, the DTI, his/her location, his/her probabilities to visit any other region and the connectivity probabilities before DTI expires. On the reception of the request, the operator requests probabilities from the N nodes currently connected to the same cellular base station, collects them, calculates the success probabilities of both DTO and PAO and estimates the overall transfer cost of each policy. If Est. Cost$_{DTO}$ < Cost$_{PAO}$, the operator opts for DTO and informs Node 1 to hold the data until DTI expires. If Node 1 identifies an available WiFi connection prior to DTI expiration, data is offloaded. If Est. Cost$_{PAO}$ < Est. Cost$_{DTO}$, the operator opts for PAO and informs Node 1 on the storing region (in this case Region 11) and storing interval. When Node 1 enters Region 11, the video is flooded to all nodes located within Region 11 (i.e. Nodes 2, 3 and 4) using an ad-hoc interface of the mobile devices. The operator signals all K nodes ($K \ll N$) within Region 11 to switch on their WiFi interface during the storing interval. Node 2 moves to Region 14, connects to the available WiFi AP and uploads
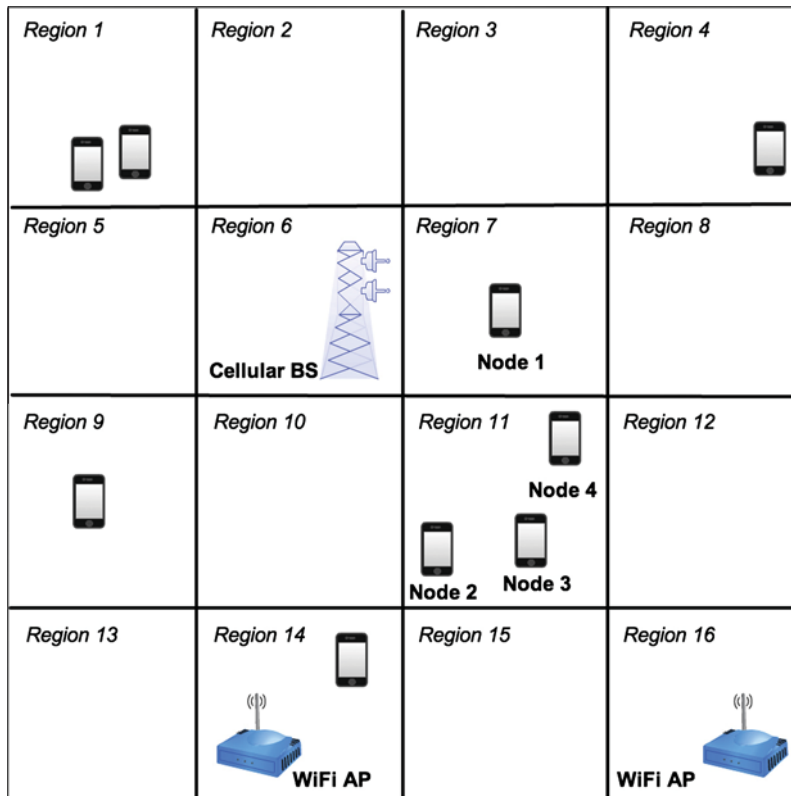

Fig. 1. Sample scenario.

the video to the corresponding server. When the video is successfully uploaded, Node 2 informs the operator that, in turn, informs Node 1. All other nodes drop the video once DTI expires. In case the video has not been successfully uploaded when DTI expires, it is transferred over 3G. The maximum overall signaling overhead associated with CEMMO is not significant compared to the amount of data that is being offloaded; signals only include GPS coordinates and probabilities and, therefore, their size does not exceed a few bytes.

5. Evaluation

In this section, we provide an extensive evaluation of the Cost-Effective Multi-Mode Offloading (CEMMO) mechanism in comparison to pure on-the-spot offloading (OTSO) and pure delay-tolerant offloading (DTO), which are the state-of-the-art approaches in mobile data offloading. All DTO solutions in literature share the same functionality; only the delay-tolerance threshold changes. Our experiments have been conducted using the Opportunistic Network Environment (ONE) simulator [8], which has been designed for terrestrial delay-tolerant communications. In the next subsections, we describe the simulation parameters, define the evaluation metrics and present the simulation results.

5.1. Simulation parameters

5.1.1. Mobility model
We evaluate the proposed mechanism under the Working Day Movement model [9], a realistic mobility model that simulates the daily mobility of average people. We assume that a total of 200 users perform map-based movement within a 4.5 km × 3.4 km section of Helsinki, Finland. Moreover, there exist 5 buses moving on specific routes within this area. People may commute on foot, by car or bus. Mobile users are allocated to a home, a workplace and places of interest (e.g. bars, stores, etc.) that they often visit.

5.1.2. Network availability
We assume that the 3G network is ubiquitous provided by a single operator. Furthermore, there is a private Internet connection available at any home or office; only authorized users utilize private Internet connections. A number of open public WiFi APs is available in popular locations. The uplink data rate of each WiFi AP is 4 Mbps.

5.1.3. Traffic types
We evaluate the impact of traffic with various characteristics to the offloading efficiency of CEMMO, DTO and OTSO. In this work, we emphasize future scenarios where big data uploads become common and CEMMO becomes more relevant. The rise in personal assistance, health monitoring and behavioral intervention applications will lead to an increase in the amount of mobile sensor data (GPS, accelerometer, audio and video) that need to be transferred to remote servers for further processing. We categorize users in three classes based on the amount of data traffic that they generate; low (around 100 MB per day), moderate (around 200 MB per day) and heavy (around 400 MB per day) traffic users. Each user class was generated using a Gaussian distribution and the traffic load values were based on [1]. We have evaluated CEMMO under a variety of scenarios with different percentages of low, moderate and heavy users. When the traffic load that needs to be offloaded is significant, CEMMO outperforms DTO, else CEMMO matches the gains of DTO. Here, we assume that 30% of the users generate low traffic, 40% generate moderate traffic and the remaining 30% generate heavy traffic. The data production rate follows a Pareto distribution. We ignore data that may be produced when a user is at work or home, where an Internet connection is available. We also assume that the storage capacity of the mobile devices is sufficient in all cases, in order to investigate the storage requirements of CEMMO. We consider three traffic types:

- *High priority*. This traffic type comprises 20% of the total mobile traffic and its size varies from 1 KB to 5 MB. Users do not tolerate any delay on the transfer of this traffic type. Webpage requests and chat conversations fall into this traffic type.
- *Medium priority*. 60% of the total mobile traffic is medium priority. Its size varies from 1 MB to 200 MB and users accept moderate delays. Non-urgent e-mails and social networking updates are typical medium priority traffic.
- *Low priority*. This traffic type comprises 20% of the total mobile traffic. Its size varies from 10 MB to 300 MB and users accept increased delays. Cloud storage services and mobile sensor data transfers are examples of low priority traffic.

5.1.4. Other parameters
In all experiments, we set region size equal to 300 m × 340 m, similar to [3], and time interval duration equal to 5 min; this corresponds to 150 regions and 288 distinct time intervals. The total duration of our experiments is 20 days, which corresponds to 10 days of training, in order to build the prediction model, and 10 days of actual evaluation. The results are presented as averages over 10 simulation runs.

5.2. Evaluation metrics

We evaluate performance using the following metrics:

1. *Offloading ratio* expresses the fraction of the total amount of generated data that is offloaded through WiFi networks.

$$Offloading\ Ratio = \text{(Total Data Offloaded through WiFi)} / \text{(Total Data Generated)}$$

2. *Average Cost per MB* is calculated as the sum of the total data (in MB) that was transferred through each transfer policy multiplied by the cost of each policy, divided by the total data transferred.

$$Avg.\ Cost\ per\ MB\ (OTSO) = Avg.\ Cost\ per\ MB\ (DTO)$$
$$= (\alpha \times \text{Data Transmitted 3G}$$
$$+ \beta \times \text{Data Transmitted WiFi})$$
$$1/\text{(Total Data Transmitted)}$$

$Avg.$ Cost per MB $(CEMMO) = Avg.$ Cost per MB (DTO)
$$+ (\beta \times N \times \text{Data Transmitted PAO})$$
$$/(\text{Total Data Transmitted})$$

3. *Average Cache Size* is computed as a ratio of the total storage used in all mobile devices for PAO to the number of the mobile devices involved in PAO.

Avg. *Cache Size* = (Total Storage Used for PAO)
$$/(\text{Mobile Users involved in PAO})$$

5.3. Performance evaluation

5.3.1. Impact of public WiFi APs density

In this experiment, we evaluate the impact of the public WiFi APs density on the performance of CEMMO, DTO and OTSO. We set β/α ratio equal to 0.04 (i.e. $\alpha = 1$, $\beta = 0.04$), which is a moderate value. In Fig. 2, we present the offloading ratio and the normalized data transfer cost per MB for the three transfer policies for an increasing number of public WiFi APs, ranging from 4 to 20. We keep the number of WiFi APs relatively low since measurements have shown that many "open" APs are not accessible as they apply MAC address filter or web-based authentication for access control.

According to Fig. 2, CEMMO is less sensitive to WiFi APs availability and manages to offload around 36% of the mobile traffic even in scenarios with sparse availability. Under the same conditions, DTO offloads 19% of the total mobile traffic, while OTSO only offloads 12% of the traffic. In scenarios with extended WiFi coverage, CEMMO offloads up to 59% of mobile data traffic. CEMMO also outperforms the other mechanisms under investigation in terms of normalized overall data transfer cost per MB. The performance of all mechanisms improves as WiFi availability increases, since more data is offloaded through low-cost WiFi networks. Fig. 3 depicts the distribution of the total offloaded data using CEMMO. When the number of available WiFi APs is low, almost half of the overall traffic is offloaded through PAO. For increased WiFi availability, the probability that a user will encounter a WiFi AP to offload data directly increases and, therefore, more data is offloaded through DTO.
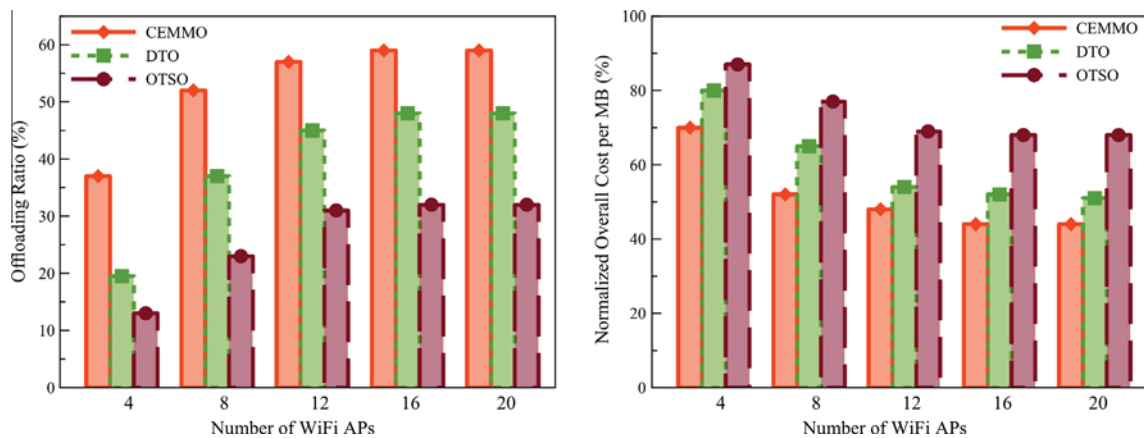
Fig. 4 depicts the relative improvement for the offloading ratio and transfer cost per MB of CEMMO over DTO for the three different traffic types. Users accept delays up to 30 min for medium priority traffic, and delays ranging from 10 to 40 min for low priority traffic, i.e. no low priority traffic is transmitted over 3G prior to the 10-min threshold. Users do not tolerate any delay for high priority traffic; only OTSO is feasible for this traffic type. The performance improvement of CEMMO over DTO for high priority traffic is zero or even slightly negative, since in some cases users may unsuccessfully attempt to transfer data during a short connection to a public WiFi hotspot. CEMMO significantly outperforms DTO for medium and low priority traffic types. We observe that CEMMO offloads more traffic when users are willing to accept larger delays. CEMMO offloads up to 24% more low priority traffic than DTO and reduces the transfer cost per MB up to 16%.

5.3.2. Impact of β to α ratio

In this scenario, we examine the sensitivity of CEMMO to the overall cost as defined by the operator, which involves the financial and energy cost of the transfers and user dissatisfaction due to delayed transmissions, congestion and increased pricing. We assume that there exist 12 open public WiFi APs and we investigate sensitivity of the CEMMO performance to the β/α ratio. The increasing β/α ratio corresponds to an increase in the cost of offloading traffic through WiFi APs compared to the cost of 3G data transfer. It is impossible to make accurate estimations on the overall cost of offloading data through WiFi or 3G networks, since we do not have any knowledge on the costs as seen from the perspective of an operator. With CEMMO, we provide the mechanism to set and adjust α and β values according to the needs of each operator.

Fig. 5 depicts the improvement of CEMMO for the offloading ratio and overall transfer cost over DTO and OTSO. When β is small (i.e. $\beta/\alpha = 0.01$), CEMMO offloads 44% more data than OTSO and 29% more data than DTO and reduces data transfer cost up to 42% and 26%, respectively. For the increasing β/α ratio, we observe that the number of mobile users that are utilized to achieve cost-efficient PAO is reduced and, therefore, the offloading ratio of CEMMO is

Fig. 2. Impact of WiFi APs density on offloading ratio and overall transfer cost.
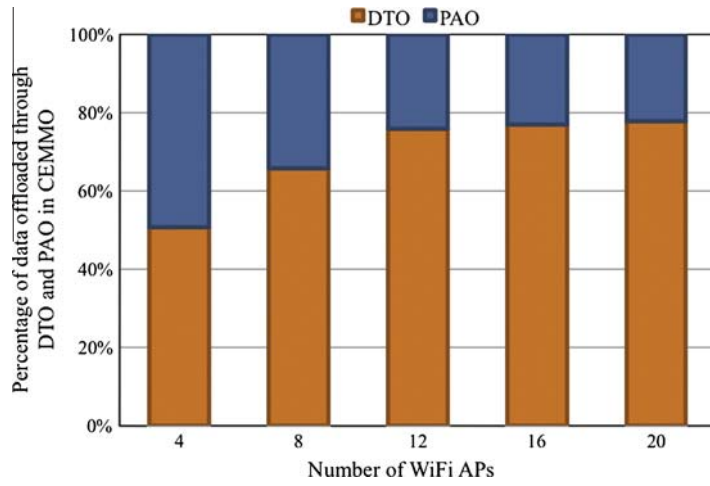
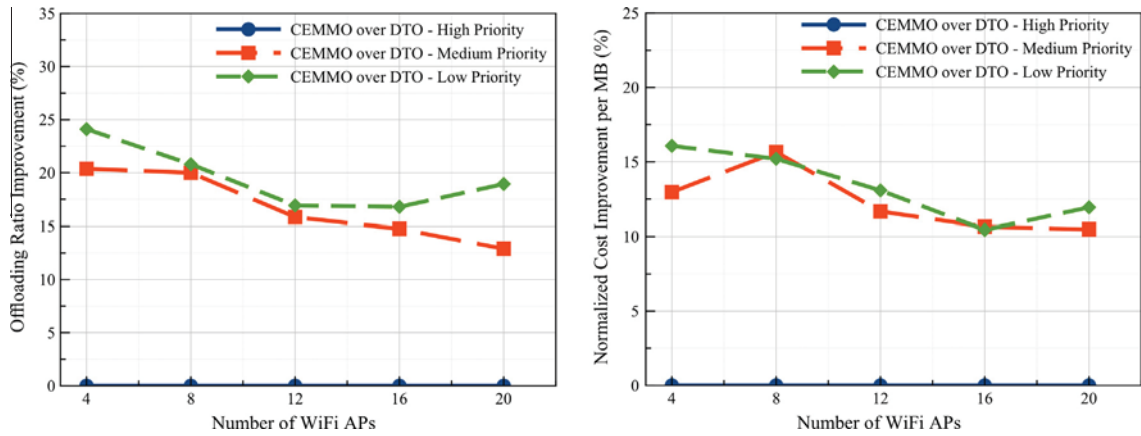Fig. 3. Distribution of the offloaded data in CEMMO.

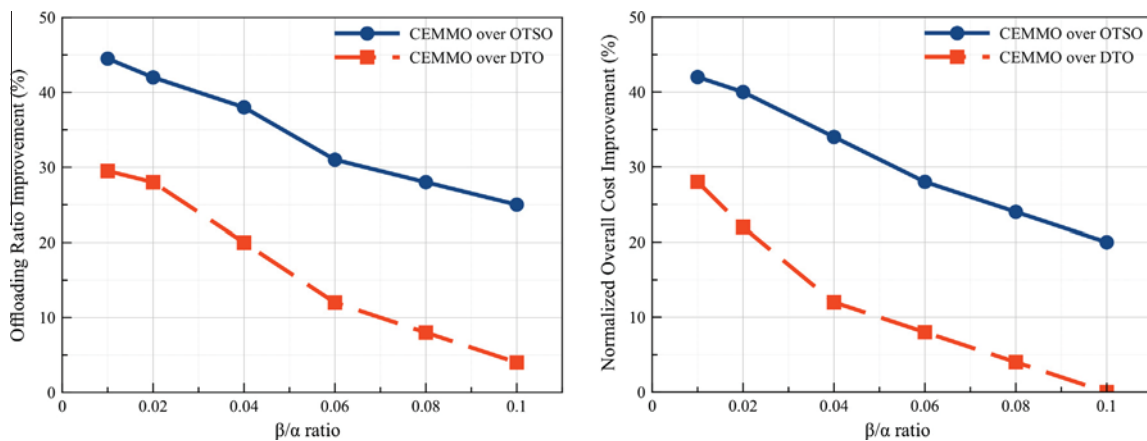Fig. 4. Offloading ratio and cost improvement.

Fig. 5. Offloading ratio and overall cost improvement for the increasing β/α ratio.

reduced. For large β values, CEMMO significantly reduces the amount of traffic that is offloaded through PAO and matches DTO in terms of data transfer cost per MB.

Considering the limited storage resources of mobile devices, it is essential to keep low the amount of data that need to be cached during PAO. In Fig. 6, we evaluate
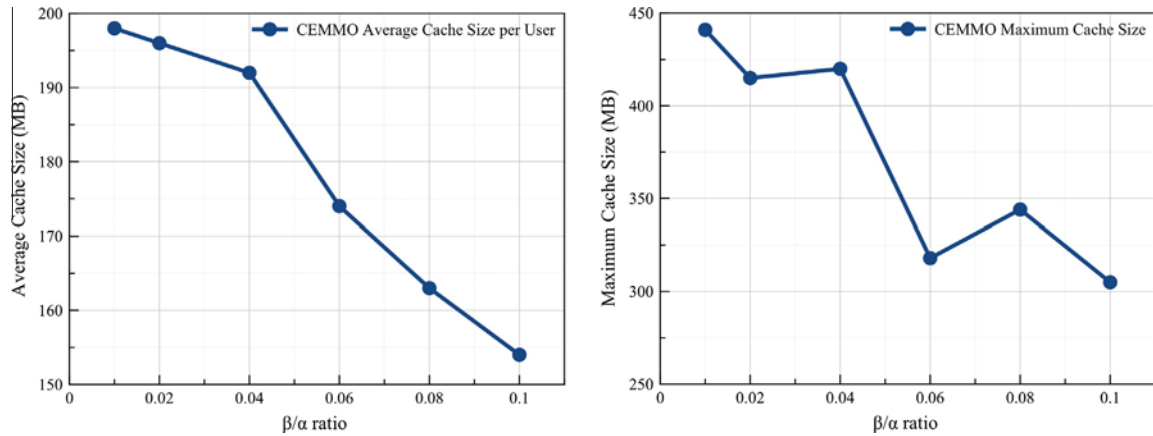
Fig. 6. Average and maximum cache size of CEMMO for the increasing β/α ratio.

CEMMO in terms of storage requirements. In particular, we depict the average cache size per user required to serve data relaying requests. On average, CEMMO requires up to 200 MB of cache for small β/α ratio, while the required cache size is reduced up to 45 MB for larger β values, since less data is offloaded through PAO. We observe that CEMMO has low storage requirements due to the spatio-temporal restrictions of the flooding process; in the worst case scenario CEMMO only requires 440 MB of cache for its operation.

5.3.3. Energy optimization

The energy efficiency of the proposed offloading mechanism is important for user participation. If an offloading scheme only aims to increase the offloading traffic ratio without considering energy consumption on mobile devices, it can quickly drain the battery of the devices. In this set of experiments, we investigate the capability of CEMMO to adapt its operation based on energy criteria that correspond to the energy consumption of each transfer policy. Table 1 summarizes the energy consumption values for WiFi-based and 3G-based communications based on [10]. The cellular network interface on mobile devices is typically always on, so the total energy consumption for scanning and connection is zero. While the energy consumption per MB of data transfer through WiFi is significantly lower than the corresponding energy consumption of 3G, WiFi-based communications consume additional energy for network discovery and connection. When WiFi operates in the power save mode (PSM), we assume that the total energy required for scanning and connection is, on average, 20 J for each data transfer.

The estimated energy cost for the data transfer of a single message of size M is:

$$\text{Est. Cost}_{\text{DTO}} = M \times (130 \times (1 - P_{\text{DTOsuccess}}) + 5 \times P_{\text{DTOsuccess}}) + 20$$

$$\begin{aligned}\text{Est. Cost}_{\text{PAO}} = M \times (&130 \times (1 - P_{\text{PAOsuccess}}) \\ &+ 2 \times 5 \times N \times P_{\text{PAOsuccess}}) + 2 \times N \times 20\end{aligned}$$

We note that energy consumption is doubled for PAO since any two peers both consume energy for the restricted flooding of data in the storing region during the storing interval.

We assume that there exist 12 open public WiFi APs and we evaluate CEMMO, DTO and OTSO for three distinct delay tolerance profiles: small, medium and large. Table 2 summarizes the time restrictions for each of the three traffic types. Delays are random and distributed uniformly between the two thresholds, while the lower threshold in each profile signals the threshold prior to which no traffic will be transmitted over 3G. We do not use excessive delay tolerance intervals; such intervals would significantly affect user satisfaction.

In Fig. 7, we present the relative improvement of CEMMO and DTO over OTSO, in terms of offloading ratio and energy. CEMMO and DTO improve the offloading ratio up to 47% and 30%, respectively, over OTSO for the large delay tolerance profile. While PAO consumes additional energy for peer-to-peer transfers, CEMMO consumes less total energy than OTSO and DTO, since less data needs to be transferred over energy-intensive 3G communications. CEMMO achieves up to 31% improvement on energy consumption over OTSO for the large delay tolerance profile, while the corresponding improvement for DTO over OTSO is 28%. Similar conclusions apply for small and medium delay tolerance profiles.

Next, we investigate the energy consumption separately for each user. In Fig. 8, we present the Complementary Cumulative Distribution Function (CCDF) of the energy improvement of CEMMO and DTO over OTSO for each delay tolerance profile. According to our simulation

Table 1
Energy consumption.

	Data transfer (J/MB)	Scanning-connection (J)
3G	130	0
WiFi	5	20

Table 2
Delay tolerance profiles.

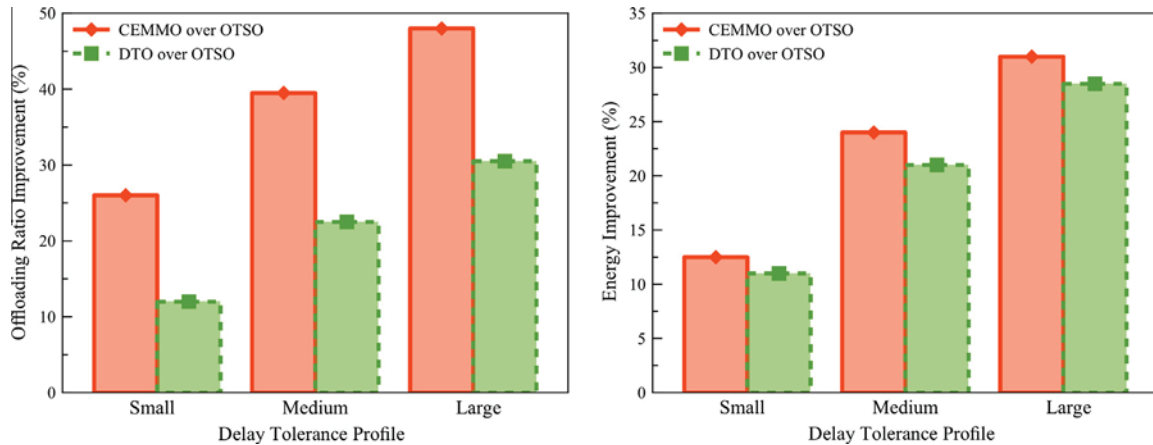|	High priority (s)	Medium priority (min)	Low priority (min)
Small	0	0–10	10–20
Medium	0	10–20	20–30
Large	0	20–30	30–40

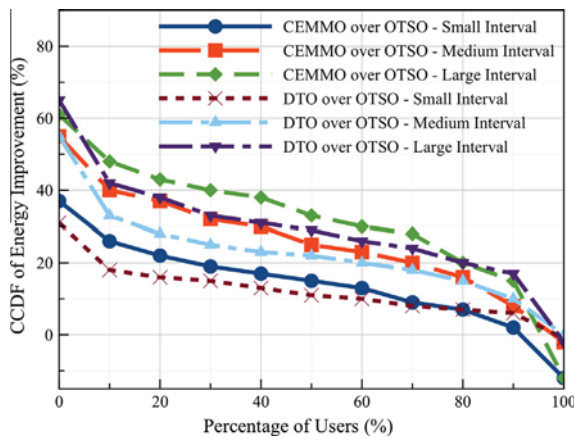**Fig. 7.** Offloading ratio and energy improvement.

Fig. 8. CCDF of energy improvement.

results, CEMMO consumes less energy than DTO and OTSO for around 80% of the users. For the large delay tolerance profile, around 30% of the users consume more than 40% less energy, when CEMMO is used. Moreover, around 95% of the users observe improved battery lifetime compared to OTSO. We conclude that CEMMO not only achieves increased offloading ratio, but also improves energy consumption for a majority of users.

6. Adopting CEMMO

In this section, we discuss some issues related to the adoption of CEMMO. First, there exists a need for a mechanism that allows users to set the desired delay tolerance threshold for each application according to its requirements. This can be implemented by exposing a simple application programming interface (API) per application, similar to e-mail refresh rate settings in mobile devices. Alternatively, CEMMO can use application port information to infer a predefined delay tolerance to specific ports. Intermediate carriers also have to be able to opt-in and opt-out of the information dissemination process anytime

they want and set their own battery power thresholds, in order to tune their participation in PAO. For example, when their battery level is high they can fully collaborate; when their battery level is intermediate they can partially participate, while below a specific battery power threshold they no longer participate in the offloading process.

In order for CEMMO to handle download requests, the cellular operator has to be the operator of the WiFi APs as well, in order to be able to push data to the intermediate carriers that will deliver it to the source node that issued a request. The gains of CEMMO will be even greater when downloading content in regions where 3G coverage is not strong and mobile devices switch to 2G communications. Downloading large volumes of content through CEMMO, instead of 2G or low-rate 3G networks, can lead to faster download times and reduced costs for operators. The implementation of CEMMO in the downlink will provide further flexibility to operators and CEMMO can become the enabler for new services. The performance of CEMMO when serving download requests will be studied in future work.

In regard to the proposed mobility and connectivity prediction model, changes in a user's schedule may take long time to be reflected in the model. Users should be able to reset their history when their schedule changes (e.g. between semesters when the everyday schedule of students completely changes). Moreover, CEMMO could also measure the throughput of WiFi networks and include this aspect in the offloading decision, so that users select the most efficient connection. In order to convince other clients to participate in PAO and share their handheld resources, it is fundamental that every participant node perceives some level of fairness and profit sharing within the network in the long term. Economic frameworks for peer-assisted services that create the right incentives for both users and providers to participate have been proposed [5–7].

Resolving any security and privacy issues that CEMMO may involve is essential for engaging their participation and maintaining their trust. CEMMO requires transmission of private data about user mobility to the cellular operator. Users may be reluctant to disclose sensitive information

even though such information is nowadays open-handedly provided to a vast amount of popular apps for smartphones. In order to preserve privacy, providers store this information temporarily, utilize it only to decide on the most efficient data forwarding strategy, and subsequently discard it. In case offloading is performed through unmanaged, untrusted WiFi APs, IPsec or selective IP traffic offload (SIPTO) can provide effective security solutions. Several solutions have been proposed for making WiFi APs as secure and easy to use as cellular networks [11,12]. Moreover, all messages in PAO need to be encrypted, so that no intermediate carrier can gain access to them.

7. Related work

The concept of data offloading through femtocells was the first approach to alleviate the cellular burden by using an alternative access network, however it contends for limited licensed spectrum with cellular networks [13]. The recent advancements in the APIs of mobile operating systems enable seamless transition from 3G to WiFi networks transparently to the user [14,15], have led to the adoption of OTSO [2,16] due to its financial and energy efficiency. Solutions that utilize multiple wireless interfaces simultaneously [17,18] and methods that intelligently switch on the WiFi interface [19,20] have also been proposed. Such solutions can be leveraged as enablers for CEMMO. The notion of DTO was introduced as a method to achieve increased data offloading. In [3], Balasubramanian et al. introduce _Wiffler_, a system that enables fast switching from 3G to WiFi and utilizes predictions about node connectivity through WiFi hotspots in order to perform offloading of traffic that tolerates delays up to 100 s. The authors of [21] studied the optimal placement of the WiFi APs in order to achieve increased offloading. The DTO operation of CEMMO and the pure DTO mechanism are similar to these approaches.

Mobile data offloading through opportunistic communications has also been introduced in the context of disseminating popular content [22,23]. One of the first concepts is 7DS [24], a peer-to-peer dissemination and sharing system for mobile devices, aiming at increasing the communication availability of users with intermittent connectivity. A similar approach is presented in [25], where predictions about inter-device connectivity are utilized, in order to define the optimal subset of users. The authors of [26] extend this idea by introducing acknowledgements as feedback in order for the source node to adjust the number of replicas that are propagated to users. The dissemination of popular data through frequently visited public WiFi hotspots is also presented in [27]; users retain this content in cache and participate in the data propagation through opportunistic device-to-device communications. To the best of our knowledge, CEMMO is the first offloading mechanism that supports data offloading from the uplink through opportunistic communications regardless of its content and popularity. CEMMO introduces a device-to-device forwarding mechanism that relies on the notion of data storage in a specific region

for a restricted time interval. This mechanism is partially inspired by the idea of floating content [28,29]. In CEMMO, we utilize data storage as a means of bridging the connectivity gap between the source user and the WiFi AP.

There has been a lot of prior work on mobility and connectivity prediction. In the context of mobile handoffs, Ref. [30] presents a spatio-temporal handoff predictor based on Markov models; however a sequence of cell associations rather than a record of client position and velocity is considered as movement, restricting the accuracy of the model. Along the same lines, Ref. [31] proposes the use of nonlinear time series for the prediction of the future location of users, as well as their arrival and residence time. This work focuses on the predictability of users when they visit their most important places, rather than on the transitions between different locations. Ref. [32] applies a first-order Markov model to predict user transitions between significant places only focusing on location prediction, irrespectively of time. Ref. [33] presents Breadcrumbs, a comparative evaluation of different mobility predictors, and concludes that sufficiently accurate predictions can be achieved by applying first- or second-order Markov models. The mobility and connectivity prediction model used in our work follows a similar approach. In particular, we model user mobility and the existence of WiFi connectivity within each region during any time interval using first-order Markov models. Markov models are ideal for mobile devices, since their CPU and storage needs are low; Markov models only involve reading and writing individual entries in arrays. Our mobility model also considers temporal accuracy. In order to capture temporal user behavior, we model user mobility separately during small time intervals. Several solutions in the area of mobility prediction are based on social contacts [34] or transitions only between important places for a user [35]. CEMMO utilizes users that do not necessarily have social interaction with the source node as intermediate carriers and estimates the mobility and connectivity probabilities for all regions, not only important places.

8. Conclusions

Cellular networks will not be able to handle the upcoming explosion in mobile data growth mainly due to economic reasons that pose barriers to wide infrastructure upgrades. As an alternative solution, cellular operators seek to offload delay-tolerant traffic from the cellular network utilizing low-cost WiFi connections. We argue that the total amount of offloaded traffic can significantly increase through peer-assisted offloading; intermediate nodes act as relays between the source node and the WiFi AP. In this context, we introduced CEMMO, a Cost-Effective Multi-Mode Offloading mechanism that offloads mobile traffic from the uplink regardless of its content and popularity and incorporates three modes of operation: cellular delivery, delay-tolerant and peer-assisted offloading. On-the-spot offloading is always exploited when available. Users implement a mobility and connectivity prediction model based on a Markov process and CEMMO utilizes this information to estimate the most efficient offloading

approach in terms of overall cost as defined by the operator. Overall cost may include financial transfer cost, energy consumption and user satisfaction. We evaluated the performance of CEMMO in scenarios with multiple traffic types and delay tolerance profiles and concluded that CEMMO increases the offloading ratio and reduces the cost compared to on-the-spot and delay-tolerant offloading, by switching between transfer policies according to the needs of each transfer. Our results also showed that CEMMO reduces the total energy consumption of the mobile devices involved in the offloading process, since less data is offloaded through energy-intensive 3G communications.

Acknowledgments

References

[1] Cisco Visual Networking Index, Global Mobile Data Traffic Forecast Update, 2012–2017.
[2] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, Mobile data offloading: how much can WiFi deliver?, IEEE Trans Netw. 21 (2) (2013) 536–551.
[3] A. Balasubramanian, R. Mahajan, A. Venkataramani, Augmenting mobile 3G using WiFi, in: ACM MobiSys, 2010.
[4] Android Geofences. <http://developer.android.com/training/location/geofencing.html>.
[5] V. Misra, S. Ioannidis, A. Chaintreau, L. Massouli, Incentivizing peer-assisted services: a fluid Shapley value approach, in: SIGMETRICS'10, ACM, NY, USA, 2010, pp. 215–226.
[6] J. Lee, Y. Yi, S. Chong, Y. Jin, Economics of WiFi offloading: trading delay for cellular capacity, in: IEEE INFOCOM, 2013.
[7] X. Zhuo, W. Gao, G. Cao, S. Hua, An incentive framework for cellular traffic offloading, IEEE Trans. Mobile Comput. 13 (3) (2014) 541–555.
[8] A. Keränen, J. Ott, T. Kärkkäinen, The ONE simulator for DTN protocol evaluation, in: SIMUTools'09, Rome, 2009.
[9] F. Ekman, A. Keränen, J. Karvo, J. Ott, Working day movement model, in: 1st ACM/SIGMOBILE Workshop on Mobility Models for Networking Research, 2008.
[10] A. Rahmati, L. Zhong, Context-for-wireless: context-sensitive energy-efficient wireless data transfer, in: MobiSys'07, ACM, NY, USA, 2007, pp. 165–178.
[11] D. Leroy et al., SWISH: secure WiFi sharing, Elsevier Comput. Netw. 55 (7) (2011) 1614–1630.
[12] CISCO, The Future of Hotspots: Making Wi-Fi as Secure and Easy to Use as Cellular, White Paper, 2012.
[13] J. Roh, Y. Ji, Y.G. Lee, T. Ahn, Femtocell traffic offload scheme for core networks, in: NTMS, Paris, France, 2011, pp. 1–5.
[14] S.M.S. Nirjon, A. Nicoara, C.-H. Hsu, J. Singh, J. Stankovi, MultiNets: policy oriented real-time switching of wireless interfaces on mobile devices, in: RTAS, 2012.
[15] T. Pering, Y. Agarwal, R. Gupta, R. Want, CoolSpots: reducing the power consumption of wireless mobile devices with multiple radio interfaces, in: MobiSys, 2006, pp. 220–232.
[16] S. Liu, A. Striegel, Casting doubts on the viability of WiFi offloading, in: ACM SIGCOMM Workshop on Cellular Networks, 2012, pp. 25–30.
[17] S. Sharafeddine, K. Jahed, N. Abbas, E. Yaacoub, Z. Dawy, Exploiting multiple wireless interfaces in smartphones for traffic offloading, in: BlackSeaCom, 2013.
[18] A. Bar, D. Mimran, L. Chekina, Y. Elovici, B. Shapira, Nesto – network selection and traffic offloading system for android mobile devices, in: IWCMC, 2013, pp. 337–342.

[19] A.Y. Ding, B. Han, Y. Xiao, P. Hui, A. Srinivasan, M. Kojo, S. Tarkoma, Enabling energy-aware collaborative mobile data offloading for smartphones, in: SECON, 2013.
[20] G. Ananthanarayanan, I. Stoica, Blue-Fi: enhancing Wi-Fi performance using bluetooth signals, in: MobiSys'09, ACM, New York, NY, USA, 2009, pp. 249–262.
[21] S. Dimatteo, P. Hui, B. Han, V. Li, Cellular traffic offloading through wi-fi Networks, in: IEEE MASS, 2011.
[22] Y. Li et al., Multiple mobile data offloading through delay tolerant networks, IEEE Trans. Mobile Comput. 13 (7) (2014) 1579–1596.
[23] B. Han et al., Mobile data offloading through opportunistic communications and social participation, IEEE Trans. Mobile Comput. 11 (5) (2012).
[24] M. Papadopouli, H.A. Schulzrinne, Performance analysis of 7DS a peer-to-peer data dissemination and prefetching tool for mobile users, in: IEEE Sarnoff Symposium, 2001.
[25] P. Baier, F. Durr, K. Rothermel, Tomp: opportunistic traffic offloading using movement predictions, in: LCN 2012, 2012, pp. 50–58.
[26] J. Whitbeck et al., Push-and-track: saving infrastructure bandwidth through opportunistic forwarding, Pervas. Mobile Comput. 8 (5) (2012) 682–697.
[27] N. Ristanovic, J. Le Boudec, A. Chaintreau, V. Erramilli, Energy efficient offloading of 3G networks, in: IEEE MASS, 2011.
[28] E. Hyytiä et al., When does content float? Characterizing availability of anchored information in opportunistic content sharing, in: IEEE Infocom, 2011.
[29] J. Ott et al., Floating content: information sharing in urban areas, in: Percom, 2011.
[30] L. Song, U. Deshpande, U.C. Kozat, D. Kotz, R. Jain, Predictability of WLAN mobility and its effects on bandwidth provisioning, in: INFOCOM'06, 2006.
[31] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, A.T. Campbell, NextPlace: a spatio-temporal prediction framework for pervasive systems, in: Pervasive'11, USA, 2011.
[32] D. Ashbrook, T. Starner, Using GPS to learn significant locations and predict movement across multiple users, J. Pers. Ubiq. Comp. 7 (5) (2003).
[33] A.J. Nicholson, B.D. Noble, BreadCrumbs: forecasting mobile connectivity, in: MobiCom '08, ACM, New York, NY, USA, 2008, pp. 46–57.
[34] W. Moreira, P. Mendes, S. Sargento, Opportunistic routing based on daily routines, in: IEEE WoWMoM, 2012.
[35] D. Fang, Moving Object Trajectory Based Spatio-Temporal Mobility Prediction, Master's of Science Thesis in Geoinformatics, KTH, Stockholm, Sweden, 2012.
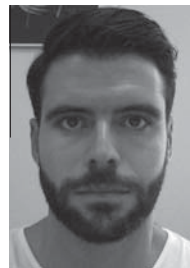
Ioannis Komnios graduated and received his MSc degree on Computer Networks from the Department of Electrical and Computer Engineering, Democritus University of Thrace, in 2007 and 2009, respectively. Currently he is a PhD candidate at the same department and a researcher at Space Internetworking Center in Xanthi, Greece. Ioannis has a long experience in internetworking in delay/disruptive tolerant environments and routing protocols. His current research interests lie in the area of data offloading and less-than-best-effort services.

Fani Tsapeli received a diploma in Electronic and Computer Engineering from the Technical University of Crete, Greece, in 2009 and an MSc in Computer Networks from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2012 under the advisory of Prof. Vassilis Tsaoussidis. From 2010 to 2013, Fani was working as a researcher at the Space Internetworking Center (http://www.spice-center.org/).

Sergey Gorinsky received an Engineer degree from Moscow Institute of Electronic Technology, Zelenograd, Russia in 1994 and M.S. and Ph.D. degrees from the University of Texas at Austin, USA in 1999 and 2003 respectively. From 2003 to 2009, he served on the tenure-track faculty at Washington University in St. Louis, USA. Dr. Gorinsky currently works as a tenured Research Associate Professor at IMDEA Networks Institute, Madrid, Spain. The areas of his primary research interests are computer networking, distributed systems, and network economics. His research contributions include multicast congestion control resilient to receiver misbehavior, analysis of binary adjustment algorithms, efficient fair transfer of bulk data, network service differentiation based on performance incentives, and economic perspectives on Internet interconnections and routing. His work appeared at top conferences and journals such as ACM SIGCOMM, IEEE INFOCOM, ACM CoNEXT, IEEE/ACM Transactions on Networking, and IEEE Journal on Selected Areas in Communications. Sergey Gorinsky delivered keynote addresses at NPSec 2013 and RAIT 2012. He has served on the TPCs (technical program committees) of SIGCOMM (2012), INFOCOM (2006–2014), ICNP (2008, 2010–2014), and other networking conferences. Prof. Gorinsky co-chaired the TPCs of COMSNETS 2013, NetSciCom 2014, E6 2012, HSN 2008, FIAP 2008, served as a TPC vice-chair for ICCCN 2009 and TPC area chair for ICNP 2013.