POLITECNICO DI TORINO
Repository ISTITUZIONALE

Content Download in Vehicular Networks in Presence of Noisy Mobility Prediction

Original
Content Download in Vehicular Networks in Presence of Noisy Mobility Prediction / Malandrino, Francesco; Casetti, CLAUDIO ETTORE; Chiasserini, Carla Fabiana; Fiore, Marco. - In: IEEE TRANSACTIONS ON MOBILE COMPUTING. - ISSN 1536-1233. - STAMPA. - 13:5(2014), pp. 1007-1021. [10.1109/TMC.2013.128]

Availability:
This version is available at: 11583/2515712 since:

Publisher:
IEEE / Institute of Electrical and Electronics Engineers Incorporated:445 Hoes Lane:Piscataway, NJ 08854:

Published
DOI:10.1109/TMC.2013.128

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

08 May 2026

Content Download in Vehicular Networks in Presence of Noisy Mobility Prediction

Francesco Malandrino, *Student Member, IEEE,* Claudio Casetti, *Member, IEEE,*
Carla-Fabiana Chiasserini, *Senior Member, IEEE,* and Marco Fiore, *Member, IEEE*

Abstract—Bandwidth availability in the cellular backhaul is challenged by ever-increasing demand by mobile users. Vehicular users, in particular, are likely to retrieve large quantities of data, choking the cellular infrastructure along major thoroughfares and in urban areas. It is envisioned that alternative roadside network connectivity can play an important role in offloading the cellular infrastructure. We investigate the effectiveness of vehicular networks in this task, considering that roadside units can exploit mobility prediction to decide which data they should fetch from the Internet and to schedule transmissions to vehicles. Rather than adopting a specific prediction scheme, we propose a *fog-of-war* model that allows us to express and account for different degrees of prediction accuracy in a simple, yet effective, manner. We show that our fog-of-war model can closely reproduce the prediction accuracy of Markovian techniques. We then provide a probabilistic graph-based representation of the system that includes the prediction information and lets us optimize content prefetching and transmission scheduling. Analytical and simulation results show that our approach to content downloading through vehicular networks can achieve a 70% offload of the cellular network.

Index Terms—Vehicular networks, content downloading, cellular network offloading, time-expanded graphs.

1 Introduction

Thanks to new cellular technologies, spearheaded by the much-vaunted blazing speeds of LTE-Advanced, consumers are lulled into the false conviction that every information content is always readily available onto their tablets or smartphones. While this may be true for home users, mobile users are in for a rude awakening.

Many observers call for the development of alternative communication systems to support and relieve the congested cellular network in areas where the demand by mobile users is expected to be the thickest. In particular, large-sized content downloading accounts for most of the traffic in access networks [1], and is thus a prime candidate for relief efforts. Its requirements are very different from those of information dissemination [2], [3], uploading of user-generated data [4] or content sharing [5]. Solutions designed to offload these kinds of traffic are thus unfit for the scenario we consider.

In the context of vehicular networks, where data recipients are drivers, passengers and on-board vehicle computers, content downloading becomes even more challenging. Several works in the literature have explored the deployment of roadside units (RSUs) that provide spotty radio connectivity to passing-by vehicles. Communication occurs through the 802.11p technology, according to what is commonly referred to as an Infrastructure-to-Vehicle (I2V) paradigm, and also leveraging opportunistic Vehicle-to-Vehicle (V2V) connectivity. Previous work, e.g., [6]–[8], has established that, in order to efficiently support content downloading, (i) RSU deployment should target the areas expected to be the most crowded by vehicles and (ii) I2V content transfer should be complemented by V2V data relaying.

A part of the picture is still missing, though. Given the ability to deliver information to passing-by vehicles through a carefully planned-out RSU deployment, what exactly should be delivered to them, and through which infrastructure? On the one hand, uninterrupted 802.11p coverage of all roads is a requirement that is too hard to come by, since duplicating the existing cellular infrastructure would imply skyrocketing costs. On the other hand, a spotty coverage could meet expectations only on condition that the short time under coverage is fruitful: RSUs should prefetch the content so as to have it promptly available for passing-by vehicles requesting it. Matching between storage at RSUs and demands by vehicles is, however, easier said than done. One possibility is that RSUs have access to the content demand and to predictions of mobility patterns, and exploit them to make prefetching decisions, as in [9]. These decisions however should be taken in view of both direct I2V transfers to downloaders and transmissions to relay vehicles deemed to meet downloaders later on.

To our knowledge, this work is the first to jointly study content prefetching at RSUs, scheduling of I2V transmissions and management of V2V relay transfers for data offloading, given a possibly inaccurate mobility prediction. We describe our system model in Sec. 2, and proceed as follows.

(i) We model the uncertainty affecting a mobility prediction through a *fog-of-war*[1] probabilistic representation of the inter-node contacts. Specifically, given the exact knowledge of the vehicular mobility and inter-node contacts, we add a noise on the contact presence, duration and rate, as described in Sec. 3. By varying the noise level, our fog-of-war model can reflect different degrees of prediction accuracy. As a result,

F. Malandrino, C. Casetti and C.-F. Chiasserini are with Politecnico di Torino, Torino, Italy.
M. Fiore is with CNR–IEIIT, Torino, Italy and INRIA, Lyon, France.

1. The term "fog-of-war" is commonly used by the gaming industry when information is increasingly hidden far away from the player's viewpoint.

the model is not a mobility prediction technique in itself, rather, a convenient way to express the prediction uncertainty and study its effect on the content downloading performance. We verify that the fog-of-war model effectively applies to predictions obtained with practical techniques. Specifically, we show the tight match existing between the output of our fog-of-war model and the accuracy of forecasts obtained through Markovian techniques of different order (Sec. 3.3).

(ii) The output of the fog-of-war model is then used to build a time-expanded graph with probabilistic weights, representing the evolution of the inter-node contacts (Sec. 4.1).

(iii) We exploit such a graph to formulate a non-integer linear programming (LP) optimization problem. The aim is to maximize the amount of data that the system can offload to the vehicular network. By solving the LP problem, each RSU can jointly take content prefetching and scheduling decisions (Sec. 4.2). The data forwarded by RSUs toward relays are then delivered to downloaders, according to different schemes (Sec. 5).

(iv) The offloading efficiency of the system outlined above is compared against benchmark solutions, in the reference scenario introduced in Sec. 6.1. We then assess the capability of a vehicular network to relieve the cellular infrastructure in the presence of location-specific content (Sec. 6.2). Validation results obtained via simulation are also presented in Sec. 6.3, before discussing related work in Sec. 7 and drawing conclusions in Sec. 8.

2 System Model

We consider a 802.11p-based vehicular network composed of mobile users and fixed RSUs, deployed over a road topology that is also covered by a cellular infrastructure. As depicted in Fig. 1, RSUs provide a spotty, yet high-throughput, inexpensive connectivity to vehicles. The cellular network, conversely, guarantees seamless coverage at possibly high connection costs. We consider vehicles to be equipped with a cellular interface. Also, as foreseen by current standardization activities, we assume that RSUs and vehicles have one 802.11p interface only, and that I2V and V2V communications occur on different frequency channels[2]. The system also includes Internet-based servers: beside the server providing data content, we assume the availability of a vehicular traffic manager and of a server handling queries from vehicular users [10], [11].

Users of the vehicular network may become *downloaders*, each possibly wishing to retrieve a different type of data from the Internet-based content server. Since vehicles have both a 802.11p and a cellular radio interface, multiple transfer paradigms for content delivery are possible. More precisely, downloaders can exploit the vehicular network to perform *direct* transfers from the RSUs, or to be assisted by other vehicles acting as *relays*. Such relays forward traffic either through a connected multi-hop path or through a carry-and-forward technique, i.e., vehicles that store and carry the data before delivering them to the target downloader. Alternatively,

2. The extensions to the cases where the nodes have more than one interface, and I2V/V2V communications occur on the same channel are straightforward.

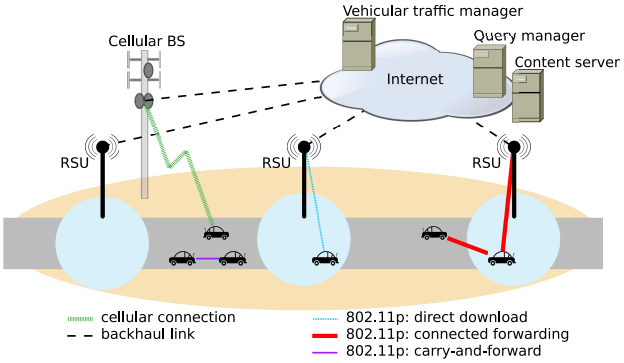


Fig. 1. Network system.

downloaders can resort to *cellular* transfers, in order to retrieve the desired content. Examples of these transfer paradigms are given in Fig. 1.

We model downloader demands by considering *what* they request and *how* they get it, as follows. As far as the *what* is concerned, we address both the cases of location-independent and location-specific content demand [12]. Regarding the *how*, downloaders try at first to obtain the data through inexpensive, opportunistic exchanges with RSUs and relay vehicles. If the desired content cannot be fully retrieved within a timeout T, the downloaders will pay to fetch the remaining portion via a cellular transfer. Note that this model provides an incentive for users to offload the cellular infrastructure through the vehicular network.

Next, we detail the operations that the network and the users undertake during the content downloading process.

A user wishing to retrieve a content generates a request to the query management server, via either an RSU or the cellular network [9]. The query includes the content identifier and the position of the requesting vehicle. The query manager forwards the pending request to the RSUs in the area where the downloader is traveling. RSUs fetch portions of the content from some server storing it. Finally, they deliver the data to the target downloader directly, or to a relay vehicle deemed to meet the downloader later on.

In order to efficiently use network resources over the backbone and the airtime on the wireless medium, RSUs must take timely content prefetching and scheduling decisions. Thus, they try to "foresee" future direct or relayed transfer opportunities that involve downloader vehicles [13]. To this end, the aforementioned traffic manager collects information on the position, speed and heading of cars through a real-time traffic monitoring system, such as those currently implemented by recent navigation solutions [10], [11]. By exploiting such data, the traffic manager predicts the evolution of vehicle movements over a near-future time horizon H, with a time granularity δ (hereinafter referred to as *time step*). The predicted location of cars over the horizon H is then leveraged by the traffic manager to determine future I2V and V2V contacts (i.e., wireless links established by pairs of neighbors within communication range[3]). We stress that the traffic manager can use any technique to predict vehicle mobility and contacts.

3. Any propagation model or measurement-based observation can be used to compile the set of I2V and V2V contacts.

Each contact may logically extend over time, hence over multiple steps. A contact is characterized by its probability to occur and its expected data rate, whose value may vary during the contact duration. Information on foreseen I2V and V2V contacts are then periodically issued by the traffic manager to the RSUs, at intervals of duration H. Each RSU can enhance the traffic manager prediction by including the information about the actual contacts it has with passing-by vehicles. The RSU becomes aware of such contacts upon receiving heartbeat messages, e.g., ETSI CAM messages or those described in the SAE J2735 dictionary set.

Based on the overall contact information and taking into account the data retrieval rate B from the content server, RSUs make locally-optimal decisions on which data to prefetch and toward which vehicles (either relays or downloaders) they should be transmitted. V2V transfers occur if RSUs delegate portions of content to relays, and these are in range of (or subsequently meet) a downloader interested in such content. Multi-hop data transmissions, whether of the connected forwarding or of the carry-and-forward type, are limited to two hops from the RSU, since this already allows for nearly optimal performance [8], [14]. For the same reason, we do not compare our approach against multi-hop routing protocols for DTNs, as their complexity is unnecessary in a vehicular network with infrastructure. We also assume that all vehicles are available for traffic relay whenever they are not receiving data from an RSU. Given the storage capabilities of envisioned vehicular network nodes, the memory capacity at RSUs and vehicles is not considered to be an issue.

Finally, errors in the message delivery process due to inaccurate prediction or transmission failure are handled as follows. Each downloader acknowledges correctly-received data to the Internet-based content server through, e.g., the cellular network. That way, the content server keeps track of the data downloaded by each user. When an RSU needs to prefetch data to be delivered to a downloader (as determined by the solution of the optimization problem), it queries the content server for data that such a downloader still has to receive. Portions of content lost during the previous prediction time intervals are thus implicitly rescheduled for transmission in the upcoming prediction interval. Note that a simple timeout mechanism is enough to avoid that on-the-fly data (e.g., packets carried by relays deemed to later meet the downloaders) are rescheduled for transmission. Indeed, a timeout set to H steps suffices to that end, since an RSU only schedules data transfers over such a time horizon.

3 Modeling the Prediction Accuracy

In order to account for the uncertainty of the traffic manager prediction in the downloading process, we propose a fog-of-war model. Representing the forecast accuracy, our model can apply to predictions obtained through any technique.

As detailed in Sec. 3.1, the construction of the fog-of-war model starts from the exact knowledge of vehicular mobility and takes the corresponding inter-node contacts as ground truth. Then, uncertainty is represented by adding a noise that affects presence, duration and rate of the contacts. Sec. 3.2 assesses the impact of the model parameters, while Sec. 3.3 validates our approach when the traffic manager employs Markovian prediction techniques.

3.1 The fog-of-war model

Let $\mathcal{P}(u, H)$ be a contact prediction generated by the traffic manager at step u, covering the following H steps. Specifically, $\mathcal{P}(u, H)$ defines: (i) the presence of each I2V or V2V contact deemed to occur between time steps u and $u + H$, (ii) its duration and (iii) its data transmission rate. We assume the presence, duration and rate of contacts in $\mathcal{P}(u, H)$ to be affected by errors, which we model as follows.

We take the mobility trace and the resulting inter-node contacts as ground truth. For each time step $k \in [u, u+H)$, we process the contacts that start at k. In order to model the error affecting the prediction, we add a noise to the contact presence, duration and rate. The noises affecting the three parameters are assumed i.i.d. Gaussian-distributed variables with zero mean and variance σ_p^2, σ_d^2, σ_r^2, respectively. Noises affecting the parameters of different contacts are also considered to be i.i.d. random variables. The choice of a Gaussian distribution is motivated by the need to describe the effect of several additive sources of error on the prediction (e.g., uncertainty on node positions, propagation conditions, link establishment procedures). This assumption is validated in Sec. 3.3. In addition, in Sec. 6.3, we show the marginal impact of neglecting the correlation that may exist among errors affecting the contact presence, duration and rate.

The variances σ_p^2, σ_d^2, σ_r^2 model the accuracy of the prediction, since the larger the Gaussian noise variance, the less precise the estimation. We let the variance grow linearly [4] with the prediction time advance, as contacts occurring further in time are increasingly hard to forecast. Our results in Sec. 3.3 support such an assumption.

Given a contact, we first extract a realization ν of the Gaussian noise with variance $\sigma_p^2 = \sigma_{0,p}^2(k-u)$ for the presence of I2V contacts and $\sigma_p^2 = 2\sigma_{0,p}^2(k-u)$ for the presence of V2V contacts. Indeed, we assume that the uncertainty affecting the predicted positions of two mobile end-points results in additive noise. Hence, we let the variance value for V2V contacts be twice as large as that for I2V contacts[5]. Such an assumption is validated in Sec. 3.3. If $|\nu| \leq 1$, we assign a probability $1-|\nu|$ to the contact presence that expresses the likelihood with which the traffic manager expects the contact to take place. Otherwise, the contact is evicted and a new, *spurious* one is created and assigned a probability equal to $\min\{|\nu|-1, 1\}$. The nodes sharing the spurious contact are chosen randomly among the network nodes. The spurious contact inherits the duration and data link rate of the true contact that it has replaced. This simple model captures the possibility that prediction techniques underestimate actual contact opportunities when $0 < |\nu| \leq 1$, and wrongly forecast future contacts when $|\nu| > 1$. Clearly, spurious contacts, appearing with the same

4. We assume a linear dependence because it is simple and, as shown later in this section, more accurate than other dependency functions.

5. The sum of two independent Gaussian random variables is still Gaussian-distributed with variance equal to the sum of the variances of the two components.

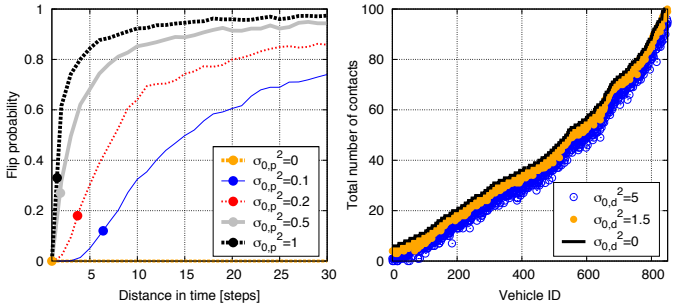


Fig. 2. Left: contact flip probability vs. the prediction time-span, for $\sigma^2_{0,d} = \sigma^2_{0,r} = 0$ and varying $\sigma^2_{0,p}$. Dots represent the average probability. Right: number of contacts for each vehicle, for $\sigma^2_{0,p} = \sigma^2_{0,r} = 0$ and varying $\sigma^2_{0,d}$.

frequency with which actual contacts are evicted, are more frequent if the prediction accuracy is low (i.e., high $\sigma^2_{0,p}$) and the estimation is pushed far ahead in time (i.e., large $k-u$).

For each contact indicated by the above procedure (be it correct or not), we add to the duration of the contact a noise with variance $\sigma^2_d = \sigma^2_{0,d}(k-u)$. If the obtained value is not positive, the contact is evicted. Likewise, a noise with variance $\sigma^2_r = \sigma^2_{0,r}(k-u)$ is extracted once for the whole contact duration and added to the link data rate computed at each step. The resulting value is bounded so that it is neither negative nor greater than the maximum data rate. Note that, by introducing errors in the contact duration and data rate prediction, our fog-of-war model also accounts for wrong estimates of the number of contacts by the traffic manager.

3.2 Impact of the model parameters

We take as reference scenario the real-world road topology and the mobility trace described in Sec. 6.1. We set $\delta = 1$ s and $H = 30$ steps, and we investigate the impact of the fog-of-war model parameters in terms of prediction quality.

The left plot in Fig. 2 presents the probability of *contact flip*, i.e., that an actual contact in the prediction is removed and a spurious one is created. The x-axis shows the time span between the contact inception (step k) and the prediction compilation (step u). The curves are obtained for $\sigma^2_{0,d} = \sigma^2_{0,r} = 0$ and different values of $\sigma^2_{0,p}$, with $\sigma^2_{0,p} = 0$ corresponding to a flawless prediction. As expected, the larger the $\sigma^2_{0,p}$, the higher the probability to predict spurious contacts. Also, the time span $k-u$ has a significant impact, as contacts established further in the future become less predictable and are affected by a higher flip probability. However, contacts already existing at step u have a null distance in time, hence they are always correctly predicted.

In the plot, the dots on the curves represent the flip probability computed over all actual contacts in the mobility trace. Note that for $\sigma^2_{0,p} \geq 0.5$ about 25% of predicted contacts are spurious, while for $\sigma^2_{0,p} = 0.1$ we have a quite reliable prediction (about 9 out of 10 actual contacts are correctly forecast within H).

The right plot in Fig. 2 shows the impact of $\sigma^2_{0,d}$, when $\sigma^2_{0,p} = \sigma^2_{0,r} = 0$. More precisely, we report the total number of contacts per vehicle, over the vehicle trip, as the error on the contact duration ($\sigma^2_{0,d}$) varies. Clearly, $\sigma^2_{0,d} = 0$ corresponds to the actual contact duration statistics. On the x axis, the vehicles are ordered according to the increasing number of actual contacts they have. Note that the larger the $\sigma^2_{0,d}$, the higher the probability that contacts are evicted and do not appear in the prediction at all. The impact of $\sigma^2_{0,d}$ is evident for vehicles with a total number of contacts below 30: a significant percentage is represented by shorter contacts that tend to be evicted. Under moderate vehicle-density conditions, such vehicles are typically those traveling on secondary roads.

Results showing the effect of $\sigma^2_{0,r}$ on the data rate between nodes are omitted because of the marginal impact that this parameter has on content downloading performance (see results in Supplemental Material).

3.3 Model validation

To validate the fog-of-war model, we assume that the traffic manager employs Markov techniques to generate the mobility prediction. This choice is motivated by the fact that such techniques combine limited complexity with good accuracy.

The traffic manager first collects the latest available information on vehicle states (in our case corresponding to the data in the mobility trace for the current step). This information is then fed to a Markovian prediction technique so as to compile a prediction of future vehicle positions. From the latter, the inter-node contacts are forecast as described in Sec. 2.

We compute the accuracy of the Markovian prediction by comparing it to the ground truth. Results are used to validate some of our assumptions and to calibrate the fog-of-war model, i.e., to set $\sigma^2_{0,p}$, $\sigma^2_{0,d}$ and $\sigma^2_{0,r}$. Finally, validation tests show that the resulting fog-of-war instances can capture the output of the Markov prediction techniques. A similar validation procedure can be applied to match the outcome of other prediction techniques as well.

To compile the Markovian prediction, we divide our road topology into segments, each of which is a few meters long. Let $\mathcal{S}$ be the set of such segments. At every time step k, each vehicle v_l in the trace is associated with exactly one segment; we denote by r_l^k the variable representing the road segment to which vehicle v_l is matched at step k. For each vehicle we build a Markov chain whose state is given by $(r_l^k, \ldots, r_l^{k-q+1}) \in \mathcal{S}^q$, where q is the chain order. By considering the trajectories of all vehicles in the trace, we compute the transition probabilities $\mathbb{P}(r_l^{k+1} = s_i | r_l^k = s_j, \ldots, r_l^{k-q+1} = s_m)$, with $s_i, s_j, \ldots, s_m \in \mathcal{S}$.

Let u be the time step at which the traffic manager generates its forecast. To obtain the mobility prediction of each vehicle in the trace, we use its associated Markov chain, and compute the probability that the vehicle is at a given position at time step $k \in [u, u+H)$. Clearly, the higher the order of the Markov chain, the higher the prediction accuracy and model complexity. Also, based on the distance between the segments occupied by any two vehicles[6], we determine whether there is a V2V link between them, as well as the link data rate.

6. We map each segment onto its middle point. Since segments are much shorter than the vehicle radio range, the resulting error is negligible.

TABLE 1
Estimated values for the fog-of-war model parameters

Noise variance	$q=1$	$q=2$	$q=3$
$\sigma^2_{0,p}$	1.68	1.22	0.87
$\sigma^2_{0,d}$	4.04	3.07	2.99
$\sigma^2_{0,r}$	1.47	1.33	1.31

By comparing the Markovian contact prediction to the actual trace, we compute the flip probability as well as the distribution of errors in estimating contact duration and contact data rate. Then, using a maximum likelihood estimation technique [19] on the flip probability and on the variance of the estimation errors, we determine the values of $\sigma^2_{0,p}$, $\sigma^2_{0,d}$, $\sigma^2_{0,r}$ that yield the best match between the Markovian prediction and the output of our fog-of-war model. The values are summarized in Table 1. Results clearly indicate that a more accurate prediction is obtained (i.e., the values of $\sigma^2_{0,p}$, $\sigma^2_{0,d}$ and $\sigma^2_{0,r}$ decrease) as the chain order q increases.

Furthermore, we carry out the Kolmogorov-Smirnov test on the distribution of estimation errors to verify that they are Gaussian. As an example, in Table 2 we present the result of the goodness-of-fit test for the error distribution in estimating duration and data rate of V2V contacts, again for different chain orders. In all cases, the p-value is higher than 0.1, which indicates a good fit to the Gaussian distribution.

In Fig. 3, we show the accuracy of the Markovian prediction with respect to the actual trace, as well as the effectiveness of our fog-of-war model in matching the prediction. In particular, Fig. 3(a) and Fig. 3(b) present the flip probability when a third-order Markov chain (i.e., $q=3$) is used, for V2V and I2V links. The two curves match in the case of both V2V and I2V contacts. Also, note that the values of the variance affecting the contact presence, reported in the legend, justify our assumption that the noise variance for V2V contacts is twice that for I2V contacts. Figs. 3(c) and 3(d) depict the cumulative distribution function (CDF) of, respectively, the contact duration and the data rate in the case where a third-order Markov chain is adopted. It can be seen that the fog-of-war model closely matches the output of the Markovian prediction, both when the latter deviates from the actual trace (Fig. 3(c)) and when the two overlap (Fig. 3(d)). The results in Fig. 3(d) confirm that $\sigma^2_{0,r}$ has a limited impact on the prediction quality.

Finally, we validate the assumption on the noise variance being linearly dependent on the time span $(k-u)$. To this end, we focus on the noise on the contact presence and consider two alternative hypotheses about the time dependency of the variance σ^2_p, namely, logarithmic (i.e., $\sigma^2_p = \sigma^2_{0,p}\log(k-u)$) and quadratic (i.e., $\sigma^2_p = \sigma^2_{0,p}(k-u)^2$). For each hypothesis, we compute the value of $\sigma^2_{0,p}$ that minimizes the mean square error on the flip probability between the Markovian prediction and the output of the fog-of-war model. The values of mean square error attained in the linear, logarithmic and quadratic cases are presented in Table 3, for different orders of the Markovian model. Clearly, assuming a linear dependency between σ^2_p and the time span $(k-u)$ provides the best accuracy.

TABLE 2
Error on contact duration and rate: result of the goodness-of-fit test for the Gaussian model

Error	p-value		
	$q=1$	$q=2$	$q=3$
Contact duration	0.18	0.22	0.25
Contact data rate	0.23	0.31	0.38

TABLE 3
Dependence between σ^2_p and $\sigma^2_{0,p}$: mean square error under different assumptions

Hypothesis	Mean square error		
	$q=1$	$q=2$	$q=3$
Logarithmic	0.53	0.58	0.64
Linear	0.14	0.15	0.20
Quadratic	1.49	0.87	1.65

4 Prefetching and scheduling at RSUs

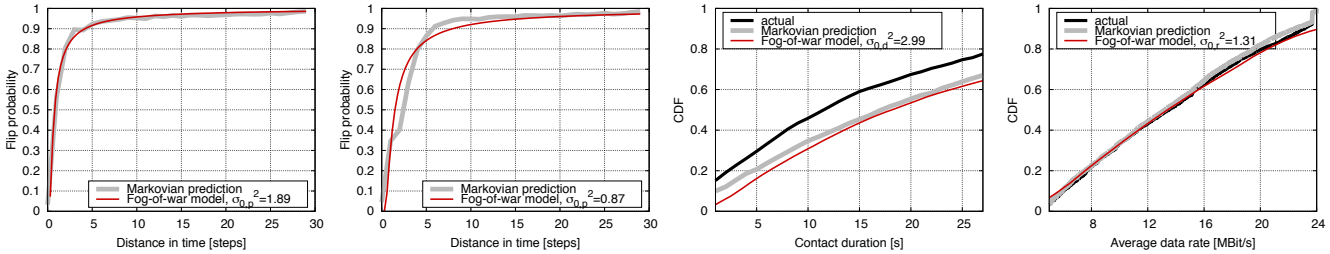
Upon compiling the prediction $\mathcal{P}(u,H)$, the traffic manager forwards it to each RSU. The RSU, in turn, updates it with the I2V contacts with passing-by vehicles it has actually established (whether they were predicted in advance or not). Such contacts are assigned a probability equal to 1, while wrongly predicted I2V contacts involving the RSU are assigned a zero probability. Thus, each RSU R_i has its own prediction $\mathcal{P}_i(u,H)$ and updates it as the time elapses, according to the contacts it observes. The prediction is combined with the information received from the query manager and used to generate a directed time-expanded graph with probabilistic weights (TEG-PW). Using such a graph, the RSU formulates a non-integer LP problem that jointly optimizes data prefetching and scheduling.

4.1 Building the TEG-PW

The prediction $\mathcal{P}_i(u,H)$ allows an RSU R_i to model the time evolution of the contacts between network nodes through a time-expanded graph. Since the prediction is based on discrete time steps of duration δ, the same granularity is used in the construction of the graph.

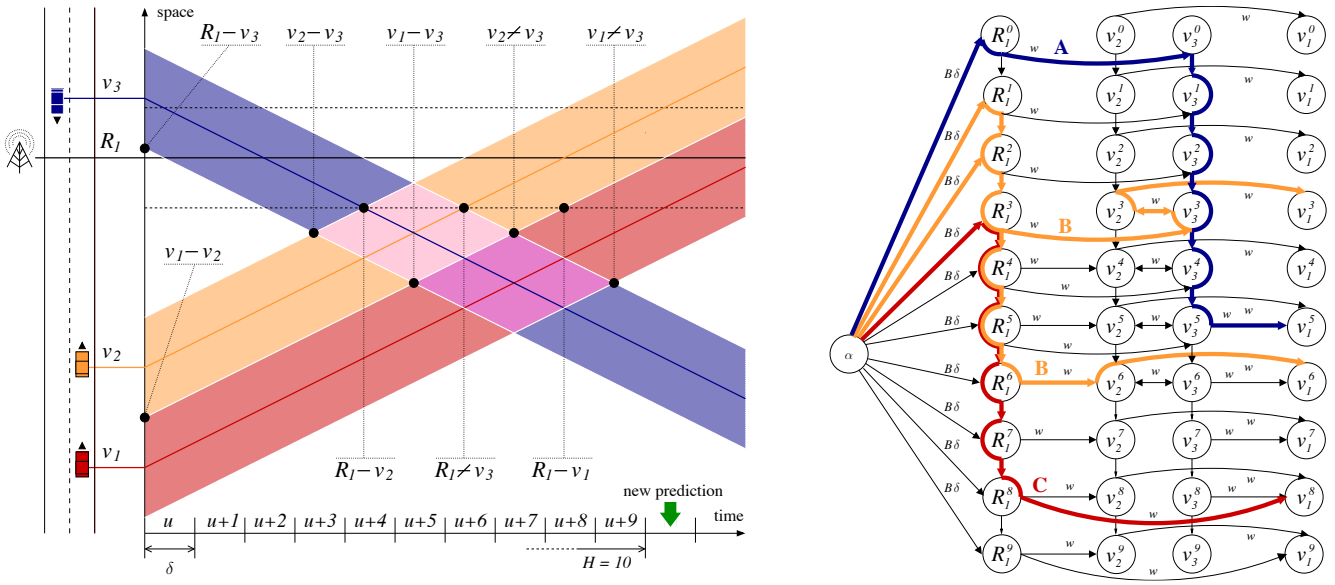
In the graph, each vehicle v_l appearing in the prediction $\mathcal{P}_i(u,H)$ at step $k \in [u, u+H)$ is represented by a vertex v^k_l, whereas each RSU R_i is mapped at each step k onto a vertex R^k_i. We denote by $\mathcal{V}^k$ and $\mathcal{R}^k$ the sets of vertices representing, respectively, the vehicles and the RSUs at step k. At every k, a directed edge connecting two vertices represents the predicted contact between the corresponding pair of nodes. Such edges are referred to as intra-step and correspond either to I2V links, i.e., of the type (R^k_i, v^k_l), or to V2V links, i.e., of the type (v^k_l, v^k_m). The edge direction reflects the way data flow over the network, i.e., I2V edges (R^k_i, v^k_l) point toward vehicle v^k_l while V2V edges (v^k_l, v^k_m) go from relay v^k_l toward downloader v^k_m.

We denote the set of I2V edges during step k by $\mathcal{E}^k_R$, and that of V2V edges by $\mathcal{E}^k_v$. Every intra-step edge in $\mathcal{E}^k_R$ and $\mathcal{E}^k_v$ is assigned a finite weight, defined as follows. As previously outlined, at the generic $k \in [u, u+H)$, each contact in $\mathcal{P}_i(u,H)$ is characterized by a probability of occurrence and an estimated data rate. We thus include these two aspects in the



(a) $q = 3$, V2V (b) $q = 3$, I2V (c) $q = 3$, duration (d) $q = 3$, rate

Fig. 3. Flip probability predicted through third-order Markovian chain, for V2V (a) and I2V (b) contacts; the corresponding values of $\sigma^2_{0,p}$ are also reported. CDF of the contact duration (c) and data rate (d) for the third-order Markovian chain prediction, with corresponding values of $\sigma^2_{0,d}$, $\sigma^2_{0,r}$.



(a) Space-time representation of the contact events (b) TEG-PW resulting from the contact events

Fig. 4. A sample set of contact events, left, and the corresponding TEG-PW, right, in presence of one RSU R_1 and three vehicles, the first of which (v_1) is a downloader while the others (v_2, v_3) can act as relays. In the left plot, shadowed areas represent *halved* transmission ranges, so that links exist when two shadowed areas touch or overlap, and break when such areas become disjoint. The instants at which links are established or lost are indicated in the figure. Time is fragmented into time steps of duration δ, and the network connectivity during each time step is represented by a row of vertices in the TEG-PW, in the right plot. In the graph, we highlight paths that are representative of the carry-and-forward (A), connected forwarding (B), and direct (C) transfer paradigms

weight of an intra-step edge. As an example, consider a V2V contact between vehicles v_l and v_m at step k. We assign to the edge (v_l^k, v_m^k) a weight $w(v_l^k, v_m^k) = p(v_l^k, v_m^k) \cdot b(v_l^k, v_m^k)$, where $p(v_l^k, v_m^k)$ is the estimated contact probability between the two vehicles at k, and $b(v_l^k, v_m^k)$ is the estimated maximum amount of data that can flow over the link during that time step. An identical discussion applies to I2V contacts.

Also, directed edges of the type (v_l^k, v_l^{k+1}), or (R_i^k, R_i^{k+1}), connect vertices representing the same node at two consecutive steps. Since they model the same node over time, they represent the possibility that vehicles physically carry data during their movement. We refer to these edges as intra-nodal. Since we assume unlimited vehicle storage capabilities, all intra-nodal edges are assigned an infinite weight. Note that accounting for contact duration, rather than considering them as atomic, allows to model critical aspects of real-world communication, like channel contention.

Finally, the content server(s), from which RSUs retrieve the data, are modeled as a vertex named α. The graph is completed with edges (α, R_i^k), from α to any vertex $R_i^k \in \mathcal{R}^k$, with weight equal to $B\delta$, i.e., the amount of data that can be retrieved from the server in one time step.

A graphical example of the resulting TEG-PW is provided in Fig. 4, in presence of one RSU R_1 and three vehicles v_1, v_2, v_3, with v_1 being a downloader and v_2, v_3 possibly acting as relays. Fig. 4(a) depicts the spatio-temporal evolution of vehicle movements, as foreseen by the traffic manager at the first time step, for the following $H = 10$ steps. Such a mobility prediction results in contacts between nodes; the times at which links are established or lost are highlighted in the figure. The TEG-PW built by the RSU R_1 from the presumed contacts is portrayed in Fig. 4(b), where time steps

correspond to rows of vertices. Note that the graph, completed by the vertex α that represents the Internet-based content server(s), allows the modeling of all possible data transfer paradigms. In the example, we observe that the RSU takes scheduling decisions that lead to (i) direct download from the RSU to the downloader, as in path C, (ii) connected forwarding through 3-hops (time step $u+3$) and 2-hops (time step $u+6$), as in path B, and (iii) carry-and-forward through the movement in time of the relay v_3, as in path A.

4.2 Making optimal decisions

At each step, RSU R_i takes its prefetching and scheduling decisions. Specifically, each RSU determines: (i) which data, not already stored, have to be prefetched, in order to be transmitted to vehicles (accounting for the rate at which data can be retrieved from the server); (ii) which data already available[7] at the RSU must be delivered via I2V current contacts, i.e., to downloaders through direct transfers as well to candidate relays deemed to meet downloaders later on.

RSUs take decisions with the aim to maximize the fraction of content that users retrieve through ITS. The retrieval time period is the minimum between the expiration of the timeout T (after T, users fall back to cellular connectivity), and the remaining steps for which a prediction is available. Thus, each RSU formulates an optimization problem based on its TEG-PW, as detailed next.

Let v_m be a generic downloader that sends a request for content c, and $\phi^k_{m,c}$ the fraction of the content that v_m downloads at step k through the vehicular network, i.e.,

$$\phi^k_{m,c} = \frac{1}{S_c}\left[\sum_{(v^k_l,v^k_m)\in\mathcal{E}^k_v} f_c(v^k_l,v^k_m) + \sum_{(R^k_i,v^k_m)\in\mathcal{E}^k_R} f_c(R^k_i,v^k_m)\right]. \quad (1)$$

In (1), S_c is the content size while $f_c(\cdot,\cdot)$ is the expected flow for content c over the edge between two vertices. Thus, $\phi^k_{m,c}$ represents the fraction of data that can be transferred at step k (the flow) over the edges of type (v^k_l, v^k_m) and (R^k_i, v^k_m). Here, v_l and R_i denote, respectively, a relay and an RSU storing at step k part of, or all, content c requested by v_m.

Then, denoting by $t_{m,c}$ the step at which the generic downloader v_m sends the request for content c, each RSU solves the following optimization problem:

$$\max \sum_m \sum_c \sum_{k=t_{m,c}}^{\min(t_{m,c}+T,u+H)} \phi^k_{m,c}\,, \quad (2)$$

where u is the time step at which the RSU received the most recent prediction. The three sums are over all downloaders v_m's, all content c's that v_m has requested, and the time steps k's, respectively. The expected flows f_c's in $\phi^k_{m,c}$ that refer to past time steps (i.e., since $t_{m,c}$ till the current step) have been already determined, thus they are inputs to the problem. On the contrary, f_c's referring to the current and future steps are the problem decision variables.

Clearly, we have to ensure non-negative flows in the TEG-PW. Beside that, from our definitions in Sec. 4.1, the following

7. Data cached at RSUs are modeled by the flow on intra-nodal edges.

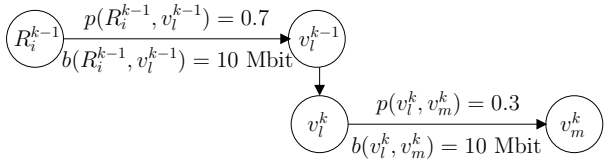


Fig. 5. Flow conservation: an example.

constraints hold:

$$f_c(v^k_l,v^k_m) \le w(v^k_l,v^k_m), \quad f_c(R^k_i,v^k_m) \le w(R^k_i,v^k_m) \quad (3)$$

$$f_c(\alpha, R^k_i) \le B\delta\,. \quad (4)$$

Then, the evaluation of the expected flows must account for the channel contention among network nodes as well as among flows related to different content transfers. It follows that the problem in (2) has to be solved under the additional constraints listed below.

Flow conservation. If each downloader wants to fetch a different content, the total flow for a content on outgoing edges, scaled by the probability of contact occurrence, must be equal to the total incoming flow for the same content. E.g., in the case of a relay vertex v_l, we have:

$$\sum_{(R^k_i,v^k_l)\in\mathcal{E}^k_R} f_c(R^k_i,v^k_l) = \sum_{(v^k_l,v^k_m)\in\mathcal{E}^k_v} \frac{f_c(v^k_l,v^k_m)}{p(v^k_l,v^k_m)} + f_c(v^k_l,v^{k+1}_l)\,. \quad (5)$$

As an example, consider the 2-step evolution in Fig. 5, where v_m is a downloader for content c. Note that the transmissions from R_i to v_l and from v_l to v_m take place at different steps, thus channel access has no effect here. Intuitively, we can try to transfer 10 Mbit from R_i to v_m, and we will succeed with probability $0.7 \cdot 0.3 = 0.21$. Then, the overall flow expected to be delivered to the downloader is $0.21 \cdot 10 = 2.1$ Mbit. However, if only the constraints in (3) were applied on each of the two intra-step edges, the expected flow should not exceed b times the edge probability. Hence, we could incorrectly conclude that the expected flow from R_i to v_m is $\min\{0.7 \cdot 10, 0.3 \cdot 10\} = 3$ Mbit. Instead, imposing (5) for vertices v_l^{k-1} and v_l^k, it correctly results that $f_c(R_i^{k-1}, v_l^{k-1}) = f_c(v_l^k, v_m^k)/p(v_l^k, v_m^k)$, i.e., $f_c(v_l^k, v_m^k) = 2.1$, which is consistent with our intuition.

Flow causality. If multiple downloaders request the same piece of information, the flow conservation constraint in (5) is replaced with the weaker constraint of *causality*. Indeed, while flow conservation implies causality, the vice versa does not hold.

In order for a node (be it a vehicle or an RSU) to transmit some data (of any content) at step k, such data must have been already downloaded from some other node at step $h \le k$. In other words, we need to introduce a *causality* constraint, imposing that, at each step k, the data downloaded by node v_m from node v_l until k (as opposed to "during step k alone") are no more than the data v_l obtained until k from other nodes. Thus, for any edge (v^k_l, v^k_m) and content c, we have that:

$$\sum_{h=1}^{k} \frac{f_c(v^h_l,v^h_m)}{p(v^h_l,v^h_m)} \le \sum_{h=1}^{k}\left[\sum_{v^h_n\in\mathcal{V}^h\setminus v^h_m} f_c(v^h_n,v^h_l) + \sum_{R^h_i\in\mathcal{R}^h} f_c(R^h_i,v^h_l)\right].$$

Channel access. We assume that the nodes access the channel using a IEEE 802.11p-based scheme with no hidden terminals, since, as shown by our simulation results, their impact is marginal. Thus, when v_l transmits to v_m, all neighbors of v_l and v_m must be silent. Also, recall that V2V and I2V traffic do not interfere, as they use different frequency channels. Then, the channel access constraint for any v_l at step k is:

$$\sum_{\substack{(v_n^k, v_o^k) \in \mathcal{E}_v^k \\ c \in \mathcal{C}}} \mathbb{1}_{[v_n^k, v_l^k]} \frac{f_c(v_n^k, v_o^k)}{b(v_n^k, v_o^k)} + \sum_{\substack{(v_p^k, v_o^k) \in \mathcal{E}_v^k \\ c \in \mathcal{C}}} \mathbb{1}_{[v_o^k, v_l^k]} \left(1 - \mathbb{1}_{[v_p^k, v_l^k]}\right) \cdot \frac{f_c(v_p^k, v_o^k)}{b(v_p^k, v_o^k)} + \sum_{\substack{(R_i^k, v_o^k) \in \mathcal{E}_R^k \\ c \in \mathcal{C}}} \mathbb{1}_{[R_i^k, v_l^k]} \frac{f_c(R_i^k, v_o^k)}{b(R_i^k, v_o^k)} \leq 1 \,,$$

where $\mathcal{C}$ is the content set, while the indicator function is equal to 1 if the specified vertices either are neighbors or coincide, and it is 0 otherwise. The three sums on the left hand side of the inequality account for the fact that the following events cannot take place at the same time: (i) v_l or a vehicle within range of v_l transmit, (ii) v_l or a vehicle within range of v_l receive, (iii) an RSU that is a neighbor of v_l transmits.

As far as RSUs are concerned, we still have to impose that the total duration of the transmissions by a generic RSU R_i cannot exceed one time step:

$$\sum_{(R_i^k, v_l^k) \in \mathcal{E}_R^k} \sum_{c \in \mathcal{C}} \frac{f_c(R_i^k, v_l^k)}{b(R_i^k, v_l^k)} \leq 1 \,.$$

Summary and problem complexity. In conclusion, at every time step, each RSU R_i formulates an optimization problem as in (2), under the above constraints. The solution of the problem yields the optimal prefetching and scheduling decisions, based on the prediction $\mathcal{P}_i(u, H)$. Since all constraints are linear expressions with respect to the control variables f_c's, which are continuous, the problem falls in the non-integer LP category. Note that non-integer LP problems can be solved in polynomial time and, in particular, our formulation is suitable to be solved in real time [15].

The problem complexity is as follows. Denoting by R the number of RSUs and by V the average number of vehicles in the road layout at a given instant, the number of variables for V2V and I2V flows is $O\left(V^2 H\right)$ and $O(RVH)$, respectively. The number of decision variables representing the intra-nodal flows and the flows from and to the virtual vertices is $O(VH)$. The number of constraints is of the same order of magnitude as the one of the number of variables.

5 Content delivery via V2V relaying

When the solution of the LP problem leads an RSU to schedule transmissions to relays, the latter are in charge of delivering the data to downloaders. We envision two approaches to manage V2V data relaying, as detailed next.

RSU-driven relaying. The solution to the optimization problem formulated by each RSU, as described in Sec. 4.2, implicitly schedules relay-to-downloader data transfers in addition to RSU-to-downloader and RSU-to-relay ones. Such a scheduling is optimal with respect to the contact prediction available at each RSU and the requests it is aware of, and it can be easily leveraged to drive V2V transfers. To that end, it is sufficient that, based on the foreseen contacts, RSUs provide relay vehicles with the identity of the downloaders the data are intended for, as well as with the expected contact times. Relays will then use this information to decide when to establish a V2V connection with a given downloader.

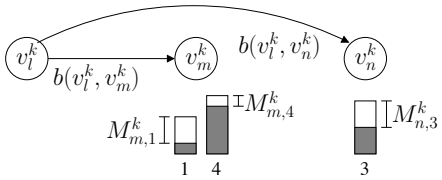


Fig. 6. Greedy relaying example. In phase 1, downloaders v_m and v_n have incomplete content 1, 4 and 3, respectively, and announce the missing data. In phase 2, relay v_l, storing all missing data, allocates its airtime to satisfy the requests by v_m and v_n, adopting a water-filling approach.

Clearly, the performance of this approach highly depends on the prediction accuracy. Uncertainty in contact estimation can lead either to failure in delivering the data if a foreseen V2V link turns out not to be established, or to a waste of opportunities if an exploitable V2V contact is not predicted. Also, the scheduling computed by different RSUs may result to be incompatible since they are generated from different TEG-PWs: this leads to unexpected channel contention and consequent delays, or impossibility to deliver all data.

Greedy relaying. A dual approach to the RSU-driven relaying consists in letting V2V transfers take place in a greedy fashion, by exploiting any opportunity to make incomplete downloads progress. In this case, the LP problem is only employed to take prefetching and I2V transfer decisions at the RSUs, while relays and downloaders autonomously manage V2V transfers. The greedy relaying protocol we adopt involves three phases and is repeated periodically.

In the first phase, each downloader advertises the list of content it is currently downloading, detailing, for each of them, the amount of data it needs to complete the transfer. As shown in Fig. 6, a generic downloader v_m will thus announce at step k the quantity $M_{m,c}^k = S_c \cdot \left(1 - \sum_{i=t_{m,c}}^{k-1} \phi_{m,c}^i\right)$, for each incomplete content c. The information on missing data broadcast by downloaders is received by relays within range. This phase requires loose synchronization (with accuracy of the order of ms) among nearby vehicles. It can be easily obtained through, e.g., GPS, and is already foreseen in the current standards for vehicular networks.

In the second phase, each relay filters missing data requests received from downloaders in its neighborhood, only retaining those for content it actually stores. Then, based on the Signal-to-Noise Ration (SNR) computed on the received broadcast transmission, it estimates the link data rate b, hence the time needed to complete each of the retained transfers. For instance, in Fig. 6, the time computed by relay v_l to complete the transfer to downloader v_m of a content c is $T_{m,c}^k = M_{m,c}^k / b(v_l^k, v_m^k)$.

A relay then decides how to serve the requests by formu-

lating and solving a max-min fairness problem. The rationale behind such a choice is that a max-min fair allocation of the airtime allows downloads to progress evenly, not favoring large downloads over small ones or vice-versa, yet guaranteeing that the medium is fully exploited. Note that, consistently with our system assumptions, incomplete offloaded transfers do not harm users, who can finish their download through the cellular network. However the more content a user downloads through the vehicular network, the lower the cost it incurs.

Denoting the total airtime to be used for data transfer by Δ, the relay assigns a portion of time $0 \leq \tau_{m,c}^k \leq \Delta$ to each downloader, such that the resulting allocation $\mathbb{T} = \{\tau_{m,c}^k\}$ solves the problem:

$$\max \min_{\mathbb{T}} \left(\mathbb{1}_{[\tau_{m,c}^k < T_{m,c}^k]} \tau_{m,c}^k \right), \quad s.t. \sum_{\tau_{m,c}^k \in \mathbb{T}} \tau_{m,c}^k \leq \Delta \,. \quad (6)$$

A water-filling approach is employed to efficiently solve (6). Once the locally-optimal allocation is obtained, in the third phase relays start to transmit their data to target downloaders. If multiple relays are neighbors, they will have to share the medium according to the constraints on channel access defined in the previous section.

6 Performance Evaluation

To evaluate the system performance, we first present the reference scenario in Sec. 6.1. We then discuss our model results in Sec. 6.2, and assess the impact of our model assumptions by comparing analytical and simulation results in Sec. 6.3.

6.1 Reference scenario

We now detail the mobility and communication scenario we used to validate our fog-of-war model and we take as a reference to evaluate the performance of the content downloading system. We consider a real-world road topology representing a 3×3 km^2 section of the urban area of Turin, Italy, as portrayed in Fig. 7. We focus on 30 minutes of road traffic, such that, at any instant, the scenario includes about five hundred vehicles simultaneously traveling over the area and taking part in the vehicular network. The vehicular mobility has been synthetically generated using the SUMO simulator, and it is representative of average traffic conditions in the area [16]. The time granularity of the resulting mobility trace is 1 s, hence we set the granularity of the traffic manager prediction and the periodicity of the execution of the V2V data relaying protocol to $\delta = 1$ s. Results for a real-world vehicular trace are included in the Supplemental Material.

Fig. 7 also depicts the default deployment that we assume for the roadside infrastructure, with 10 RSUs located at the most crowded intersections, represented by green dots. This corresponds to a rather sparse RSU density of 1 RSU/km^2. As for the placement strategy, in [8] it is shown that deploying RSUs at major intersections allows vehicular-based downloading to perform close to the optimum.

With regard to the communication technology, we assume that nodes use the 802.11p protocol with data transmission rate adaptation. It follows that the value of the achievable network-layer rate between any two nodes is set according to their distance. We refer to experimental results in [17, Fig. 5] to derive the values shown in Fig. 7, and we use them as samples of the achievable network-layer rate. Also, we limit the maximum radio range of any node to 200 m since, as stated in [17], this distance allows the establishment of a reliable communication in 80% of the cases.

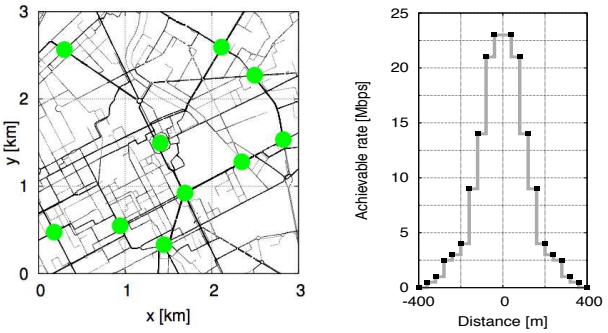


Fig. 7. Road topology (left) and network-layer rate (right).

As for the cellular network, we assume that full cellular coverage of the area is available. A user can always complete its download through the cellular infrastructure if it could not retrieve the whole content through the vehicular network within T seconds. Unless otherwise specified, we set $T = 120$ s, a value suitable for delay-tolerant applications. Results for different T's, including values suitable for quasi real-time services, are available in the Supplemental Material.

User content demand is modeled by assuming that 100 content items are available and have the same size $S_c = 10$ MBytes. The per-user request rate is Poisson distributed with rate $\lambda = 0.005$. When location-specific content is considered, we identify the vehicular flows[8] in the mobility trace and assign identical demands to cars in the same flow.

Finally, we assume that the traffic manager generates its predictions every 30 s, forecasting contacts in the next 30 seconds. Since $\delta = 1$ s, this implies $H = 30$ in the following.

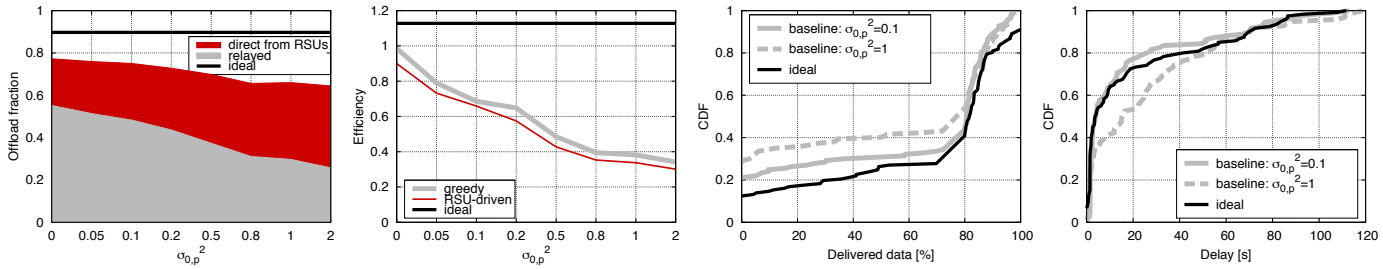
6.2 Performance of content downloading

We evaluate the effectiveness of offloading content download from cellular to vehicular networks, in the reference scenario previously described.

We first assume (i) a content demand process where each content is requested by vehicles with equal probability, (ii) unlimited time validity for content, and (iii) $B = 100$ Mbit/s, i.e., high-bandwidth links connecting the RSUs with the content servers. Note that this essentially implies ideal vehicular network operation, as RSUs need to download content only once, thanks to their unlimited cache size and the infinite content validity. We refer to this system configuration as our *baseline* scenario. The rationale is that it allows us to study the wireless portion of the system, while avoiding bias due to the demand distribution or to backbone limitations.

As a second step, we relax the assumptions on the RSU content retrieval operation, content demand and validity, and

8. We run the κ-means clustering algorithm [18] on the mobility trace, and consider clusters, detected in consecutive steps and having the closest centroids, as snapshots of the same flow. We use $\kappa = 5$ so as to track the 5 largest vehicle groups (each turns out to include at least 10 vehicles).



(a) Offload fraction (b) Offload efficiency (c) Offload distribution (d) Delay

Fig. 8. Performance of cellular network offloading via vehicular communication, in the baseline system scenario.

investigate: (i) a *constrained* system configuration, where the content validity is limited in time and the RSU backbone bandwidth is reduced; (ii) a *location-specific* system configuration, where the content requested by vehicles is influenced by the traffic flow they belong to. The latter configuration allows us to compare the offloading performance of a prediction-based scheme to that of a push-based scheduling based on content popularity only [20].

6.2.1 Baseline scenario

The performance of the offloading process in the baseline scenario is presented in Fig. 8(a), which portrays the average fraction of requested content that a vehicle can successfully download through the vehicular network before the expiration of the timeout T. In the following, unless otherwise specified, the results have been obtained as $\sigma^2_{0,p}$ varies, under the greedy relaying scheme and for $\sigma^2_{0,d} = \sigma^2_{0,r} = 0$.

The offload fraction is broken down into content retrieved directly from RSUs and content obtained from relays through V2V communication, and it is compared against the ideal offload performance. The latter is derived by solving the optimization problem for $\sigma^2_{0,p} = 0$, a very large prediction horizon ($H = 300$) and assuming that future user requests are known a priori; this enables perfect I2V and V2V scheduling.

Firstly, we observe that vehicular networks can relieve the cellular infrastructure of 70-80% of the cost of content download. Secondly, a sizable contribution comes from V2V relaying, bearing between 30 and 60% of the content transfer effort. This confirms that opportunistic transfers are highly beneficial in the offload process. Thirdly, the overall performance is not too far from the ideal one, which would allow a 90% offload.

The impact of the accuracy of the contact prediction is shown by varying $\sigma^2_{0,p}$. Quite surprisingly, very accurate predictions (low values on the x axis) result in a performance that is just slightly better than that scored by almost random contact estimations (high values of $\sigma^2_{0,p}$). Inaccurate predictions lead however to a reduced contribution of V2V with respect to I2V transfers, as the former drops from 60% to less than 30%.

The actual cost of an imprecise contact prediction is revealed by Fig. 8(b), which shows the offload efficiency, i.e., the ratio of the amount of data delivered to a downloader to that transmitted by the RSUs (to either downloaders or relays). A low efficiency implies a waste of wireless resources at the RSUs, while a high efficiency means that only useful vehicular-based transfers are performed. The efficiency can be higher than 1, since a relay can download some content (or part of it) and then provide it to multiple downloaders interested in the content. The plot clearly shows that, in order to maintain high offload fractions, the less precise the information on future contacts, the larger the amount of data the RSUs have to transfer to relays.

Another interesting fact underscored by Fig. 8(b) is that RSU-driven relaying consistently performs worse than the greedy approach. The reason for such a behavior is that the amount of data transmitted by RSUs is the same in either case, but the former is unable to exploit data transfers to future downloaders (of which RSUs are unaware). This is an important contribution to the performance, unlike the optimized RSU-driven scheduling that is beneficial only in the rare case of multiple, simultaneous relay-downloader transfers. As a consequence, the greedy approach is to be preferred and we will focus only on it in the following.

Fig. 8(c) further details the offload performance, showing the CDF of the fraction of content that each downloader can retrieve through the vehicular network. Results are shown for quite accurate ($\sigma^2_{0,p} = 0.1$) and rather imprecise ($\sigma^2_{0,p} = 1$) predictions, and benchmarked against the ideal case. The CDFs clearly identify two larger classes of downloaders: those that can get a very small percentage (possibly zero) of the data they request, and those (over 50% of the total) that can obtain almost all (80% or more) of the data through the vehicular network. These two categories correspond to users traveling, respectively, on secondary roads and main thoroughfares. The former are seldom under RSU coverage and experience fewer contacts with (relay) vehicles. Interestingly, the latter do not seem to be affected by $\sigma^2_{0,p}$, as curves are very close for high values on the x axis. On the contrary, the percentage of downloaders unable to get any data is sensibly reduced as the contact estimation precision grows. We can thus conclude that an accurate prediction is most useful to offload traffic destined to for hard-to-reach users.

Finally, Fig. 8(d) portrays the CDF of the delay in content delivery through the vehicular network. A large data portion, amounting to 70% of the content size, can be obtained within a short time span (approximately 20 s). Results are similar under ideal and precise contact predictions. However, in the ideal case, the higher fraction of downloaded content leads to an increased latency for users on unfavorable routes. An inaccurate contact prediction, instead, yields higher delays.

Tab. 4 shows the offload fraction for varying $\sigma^2_{0,p}$ and number of deployed RSUs. As expected, increasing the number

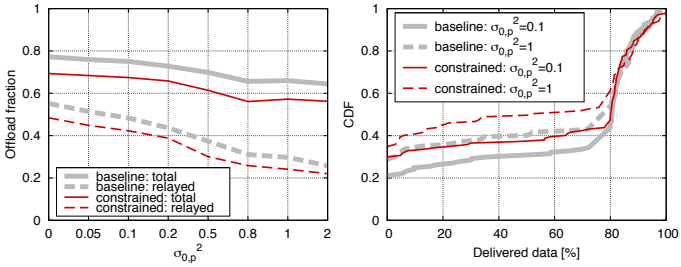


Fig. 9. Vehicular-based download performance in the baseline and constrained scenarios.

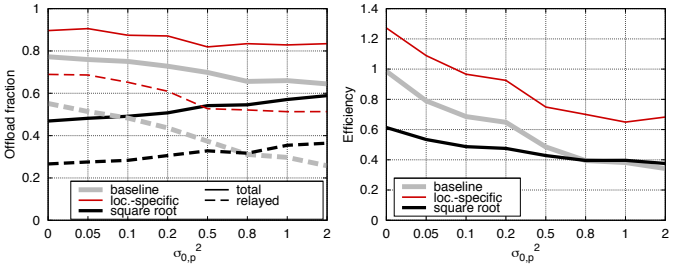


Fig. 10. Impact of location-specific content on download performance and comparison between prediction- and content popularity-based approaches.

of RSUs favors the offloading process. However, improving future contacts estimation can compensate for a less pervasive RSU coverage. Indeed, by cross-checking similar offload fractions over different columns, we note that an accurate prediction requires between 20 and 30% fewer RSUs, while maintaining similar performance.

The benefits of an accurate prediction are also shown in Tab. 5, which reports the offload fraction for different values of $\sigma^2_{0,p}$ and T (the time after which users start retrieving data from the cellular network). Indeed, the higher the T, the larger the amount of data downloaded through the vehicular network. However, improving the forecast reliability pays significantly more than delaying the use of the cellular network.

In conclusion, our results show that vehicular networks are a viable alternative, or a complementary solution, to cellular networks for content downloading by mobile users. In particular, if a relatively reliable mobility prediction is available, the offload of the cellular infrastructure can be achieved by sparing wireless resources, better serving downloaders on secondary roads, reducing the download latency, and lowering the RSU deployment cost. Furthermore, an imprecise mobility prediction has a relatively small impact on the actual offload fraction, but it significantly impairs the system efficiency.

6.2.2 Constrained scenario

Here, we focus on the case of RSU backbone links with bandwidth limited to $B = 10$ Mbps and content expiring after an exponentially distributed time with mean equal to 200 s. The latter condition forces, upon expiration of a content, both RSUs and downloaders to discard any portion of the content they have obtained, and restart the download from scratch (if they have not completed it).

TABLE 4
Offload fraction as the number of RSUs and $\sigma^2_{0,p}$ vary

$\sigma^2_{0,p}$ \ No. RSUs	6	8	10	12	14	16
0.1	0.55	0.67	0.76	0.79	0.92	0.94
1	0.48	0.57	0.66	0.71	0.82	0.84

TABLE 5
Offload fraction as the timeout T and $\sigma^2_{0,p}$ vary

$\sigma^2_{0,p}$ \ T [s]	60	120	180	240
0.1	0.69	0.75	0.78	0.80
1	0.58	0.66	0.71	0.72

The offload fraction obtained in such a constrained configuration is presented and compared to our baseline in Fig. 9. More precisely, the left plot shows the average offloading fraction as $\sigma^2_{0,p}$ varies, while the right one details the per-downloader CDF of the offload fraction. The first plot clearly highlights that the introduction of the constraints leads to an unchanged trend with respect to the contact prediction accuracy, at the cost of a performance reduction. Interestingly, the performance drop mainly concerns the download via V2V relaying, since, upon expiration of the content, relays have to discard the data and cannot help in the delivery any longer. In the second plot, we can once more observe how a constrained scenario affects downloaders on unfavorable routes (e.g., traveling on secondary roads).

6.2.3 Location-specific content scenario

We now evaluate the offload performance in presence of location-specific content. Users belonging to the same vehicular flow request content according to the same Zipf's distribution with exponent 2, and users belonging to different vehicular flows request items within disjoint sets of content. Vehicles not belonging to any vehicular flow request all content with equal probability.

Fig. 10 shows that, in presence of location-specific content, the amount of data the downloaders can retrieve through the vehicular network significantly increases, mostly due to V2V relaying. Indeed, thanks to the tighter correlation between vehicles' routes and the content they request, it is likely that the desired information is obtained from nearby vehicles. This also explains the very high efficiency of relayed traffic in the right plot of Fig. 10.

The plots also portray the performance of an approach based on content popularity that exploits knowledge of popularity distribution, instead of mobility forecast. Specifically, it lets RSUs select the content to be pushed towards a relay, with a probability proportional to the square root of the content popularity [20]. For the sake of a fair comparison, we force the amount of data sent by RSUs to match what is observed in our prediction-based scheme with location-specific content.

Results show that predicting contacts yields significantly better performance than the knowledge of content popularity. This is due to the high mobility of our scenario: either the content is delivered to the right vehicular flow, or retrieving the content from a vehicle carrying the data becomes very hard. Such an observation is confirmed by the curve referring to

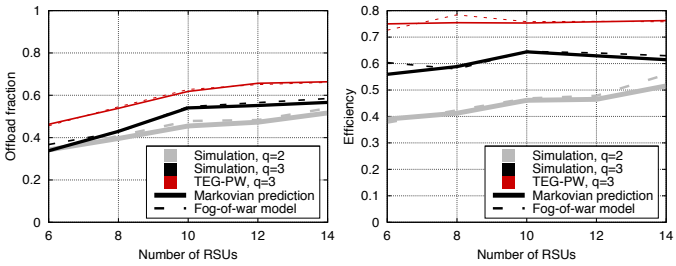


Fig. 11. Comparing simulation and optimization results: offload fraction (left) and efficiency (right).

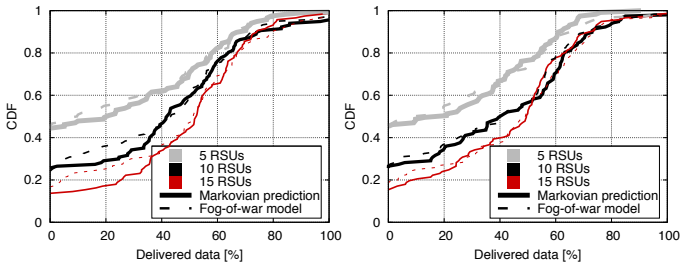


Fig. 12. Simulation: CDF of the fraction of data delivered to downloaders for $q = 2$ (left) and $q = 3$ (right).

relayed traffic in the left plot, which is significantly lower for the square root approach than in the prediction-based scheme. As a last remark, the offload fraction in the square root case grows with the increase of $\sigma^2_{0,p}$, since, for fair comparison, the RSUs inject more data in the network to match the amount observed in the location-specific case. Nevertheless, such an increase in the delivered data does not make up for the higher radio resource consumption, thus leading to a lower efficiency.

6.3 Validation through simulation

Thanks to its limited computational complexity, solving the optimization problem allows a comprehensive evaluation of the performance of the vehicular download framework, as detailed in the previous section.

However, the optimization problem formulation builds on simplifying assumptions that grant its mathematical tractability. More precisely, ideal physical (PHY) and Medium Access Control (MAC) layers are considered, i.e., lossless channel and perfect channel contention with no additional overhead.

In order to assess the impact of more realistic PHY and MAC layer modeling on the system performance, we compare the optimization problem outcome with that of a complete network simulation of the framework.

To that end, we employ ns-2 as simulation environment including the IEEE 802.11p PHY and MAC layers. At the PHY layer, we adopt a log-distance propagation loss model with exponent 3.0, and set the transmit output power to 16 dBm. Since no rate adaptation algorithm is specified in the IEEE 802.11p standard, we use the Adaptive Auto Rate Fallback to set data transmission rates at the MAC layer.

We solve the optimization problem by using either the Markovian prediction or the fog-of-war model, and feed the resulting scheduling to the simulator. We then simulate I2V and V2V exchanges and record the download performance at each user. Note that, within a time step, an RSU (relay) orders data packet transfers based on the content identifier.

An overview of the comparison between optimization and simulation results is provided in Fig. 11, where the vehicular download performance is summarized in terms of offload fraction and efficiency. The plots portray the outcome of ns-2 simulations for two values of the order q of the Markovian prediction technique, i.e., 2 and 3, as well as the optimization result for $q = 3$. For each of such cases, we show both the performance obtained with the Markovian technique itself (solid lines) and the fitted fog-of-war model (dotted lines). A greedy V2V content delivery is adopted in all settings.

We first comment on the impact of the PHY and MAC layer simplifying assumptions we made in the optimization problem. Compare the two curves for $q = 3$, which refer to simulation and optimization, respectively. It is clear that ideal signal propagation and channel access result in better performance. While this was easily expected, we remark that the relative performance difference is reasonable, as the optimization framework overestimates the actual performance by 20%. More importantly, the trend in either curve is the same. The offload fraction grows as more RSUs are deployed, although 10 RSUs are already enough to grant a good download performance to most downloaders. The efficiency instead is almost unaffected by the roadside infrastructure coverage. In light of such results, the optimization problem solution appears as a good indicator of the real-world performance of the system. The latter observation is especially interesting when considering that ns-2 simulations are very time consuming and cannot be used for an extensive performance evaluation of content download in large-scale vehicular networks.

As a second interesting aspect, we observe the impact that the prediction accuracy has on download performance, when a realistic simulation environment is employed. A higher-order Markovian technique yields higher prediction accuracy: in turn, more precise contact forecasts lead to slightly smaller offload improvement but significantly higher efficiency. These observations are consistent with those derived via the optimization framework on the impact of the prediction accuracy (Fig. 8(a) and Fig. 8(b)).

Finally, we stress that the Markov prediction technique and the fitted fog-of-war model result in nearly identical performance in all combinations of network settings and chain order. This is a further proof of how our approach to modeling the prediction accuracy can closely mimic different practical forecast techniques.

Further details on simulation results are provided in Fig. 12, depicting the CDF of the amount of content delivered to each downloader, for a Markovian prediction of order $q = 2$ and $q = 3$. Once around, solid lines refer to results obtained by employing the actual Markovian prediction, while dotted lines show the performance under the fitted fog-of-war model. Here, colors correspond to different numbers of deployed RSUs.

The plots indicate how more pervasive deployments significantly help downloaders traveling on secondary roads, while they have marginal impact on the experience of onboard users already traveling along major traffic thoroughfares. This effect is similar to that obtained with higher mobility prediction

accuracy or less constrained backbone capacity (see Sec. 6.2), and it confirms that a more efficient framework also tends to behave more fairly toward users.

Finally, Fig. 12 shows once more the flexibility of the fog-of-war model, faithfully reproducing the distributions obtained with prediction techniques characterized by different accuracy.

7 Related work

A few works have studied scenarios where opportunistic transfers and cellular technologies coexist, so as to offload the infrastructure through user-to-user communication. However, the problem that most of them address is not that of content downloading, but the dissemination of some data items to all mobile users. Thus, rather than scheduling the transfer of heterogeneous content, the problem becomes that of determining how many copies of an item shall be injected in the network and which users are most suitable to receive them. Clearly, solutions designed for dissemination offloading cannot be applied to the concurrent downloading of different content. Among these works, [2] considers vehicular users but it significantly differs in scope from our work. Specifically, [2] adopts a push approach and aims at injecting the right amount of content copies in the network. Instead, we assume a pull approach and we leverage mobility prediction for optimal data prefetching and scheduling, in the case of content downloading. As for [3], the scope is similar to the one in [2] and different from ours; also it deals with generic smartphone users rather than vehicular ones.

Content downloading is instead the target of [9], [12]. The work in [12] explicitly takes into account the link between vehicle location and requested content, considering that nearby vehicles are more likely to request the same content item. In [9], only I2V direct transfers are considered, and the focus is on the prefetching of content at RSUs, which are assumed to have high-latency, low-bandwidth links to the Internet. The objective is then to optimize the usage of such links, by estimating the amount of traffic the vehicles will be able to download from each RSU. Moreover, in [9] the use of the cellular infrastructure is limited to signaling purposes. Prefetching is also exploited in [14], where experimental and analytical results show the contribution of V2V and I2V communications to the system performance. Unlike ours, the study in [14] deals with interactive applications, such as Web search and browsing, in a DTN composed of access points and buses, and the focus is mainly on web page prioritization and mobile-to-mobile routing.

Works such as [21] investigate content downloading through publicly available WiFi hotspots, and the link between acceptable delay and offloading fraction. Note that, unlike previous works, we study content downloading in vehicular networks accounting for all communication methodologies, i.e., I2V, V2V, and cellular-based, at a time. This allows us to jointly investigate the problems of content prefetching at RSUs, scheduling of I2V transfers, and management of opportunistic V2V transfers. We do that by formulating an optimization problem and by accounting for inaccurate mobility prediction. This makes our study different also from [22], where software modules for content downloading and offloading are presented and evaluated through trace-driven simulation.

The approach we adopt relates our work to the problem of transmission scheduling in wireless networks, which has been widely studied. However, most works address the case of connected multi-hop networks, e.g., [23], or social delay-tolerant networks, e.g., [24]. The vehicular environment mixes elements of both, thus solutions that assume full reachability or contacts periodicity [25] in the order of hours or days do not apply to our context. In [26], vehicular movement prediction is leveraged for data prefetching at RSUs and handoff between them. The experimental results show the practical viability of exploiting mobility prediction to take full advantage of I2V contacts. Our work extends such a principle to a large-scale architecture involving V2V contacts as well, and it investigates the full potential of the system by formulating an optimization problem. A scheduling and prefetching scheme for content downloading in vehicular networks is presented in [7]. This work, however, employs simplistic mobility models and does not consider the presence of a cellular infrastructure. As further additions to the literature on transmission scheduling to vehicles, we take into account, for the first time, the role that mobility-based communities have in the generation of content demand, and evaluate the impact of uncertainty in the estimation of future I2V and V2V contacts.

Concerning the latter aspect, there are several ongoing efforts on inferring future vehicular contacts, given the current position and past car trajectories [13], [27]. The two main approaches consist in modeling the vehicle location through a Markovian process, and in studying the time and duration of contacts. Thanks to our fog-of-war model, our system can use any of these techniques, or future ones, as an input.

Finally, the representation of a time-varying network as a time-expanded graph has also been employed in our previous work [8]. Beside the different scope, the time-expanded graph we propose here significantly differs from the above representation as we introduce probabilistic edge weights, so as to model uncertainty in the prediction of inter-node contacts.

8 Conclusion

After arguing in favor of cellular network offloading through vehicular network content download, we addressed content prefetching and data transmissions scheduling to passing-by vehicles in the realistic case of finite-horizon, inaccurate mobility prediction.

Our first contribution consists in a fog-of-war model that represents in a convenient, yet accurate, way the uncertainty of mobility prediction. We validated such a model by showing that it can closely match the output of widely-used Markovian prediction techniques.

Then, we integrated our fog-of-war model in a time-expanded graph representation of network dynamics. Through such a graph-based model, we formulate a non-integer LP problem that is solved at each RSU in order to make optimal data prefetching and transmission scheduling decisions.The amount of data offloaded from the cellular to the vehicular network is thus maximized through the predicted mobility.

Our results yield the following findings.
(i) Overall, vehicular networks are a viable, complementary solution to cellular networks for content downloading. Indeed, 70% of the data can be offloaded to the vehicular network in presence of moderate vehicular traffic conditions, and even under a quite sparse RSU deployment.
(ii) Concerning the mobility prediction, its accuracy has a surprisingly low impact on the fraction of offloaded data. However, a more accurate prediction implies both a higher offloading efficiency, i.e., a better utilization of radio resources, and a higher throughput and lower delay for users traveling on secondary roads. Finally, a better prediction can also make up for a sparser RSU deployment.
(iii) As for the exploitation of V2V contacts, adopting a simple, greedy approach leads to similar, or even better, performance as that obtained when V2V relaying follows the transmission scheduling provided by RSUs.
(iv) Limitations on the backhaul bandwidth and on the time validity of the content cached at RSUs and vehicles only impair the performance of users on secondary roads.
(v) The offload efficiency achieved through a prediction-based approach is significantly better than that of a content popularity-based solution. Additional benefits can be expected for location-specific content, which can be locally cached by RSUs and vehicles and subsequently requested by users traveling on the same roads.
(vi) Simulation results derived through ns-2 show that the qualitative performance behavior is the same as the one obtained through our graph model, and even the quantitative difference is limited. Simulations further prove that our approach in modeling the prediction accuracy mimics well different forecast techniques.

9 Acknowledgments

We thank Prof. Malnati and Dr. Barberis for providing the mobility trace, and Regione Piemonte for supporting this work through the IoT_|_ToI project (POR F.E.S.R. 2007/2013).

References

[1] R. Pries, F. Wamser, D. Staehle, K. Heck, P. Tran-Gia, "Traffic measurement and analysis of a broadband wireless Internet access," *VTC09 Spring*, 2009.
[2] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, M. Dias de Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," *Pervasive and Mobile Computing*, vol. 8, no. 5, 2012.
[3] B. Han, P. Hui, V.S.A. Kumar, M.V. Marathe, J. Shao, A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. on Mob. Comp.*, vol. 11, no. 5, 2012.
[4] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, "Taming user-generated content in mobile networks via drop zones," *Infocom*, 2011.
[5] T.H. Luan, L.X. Cai , J. Chen, S. Shen, F. Bai, "VTube: Towards the Media Rich City Life with Autonomous Vehicular Content Distribution," *Secon*, 2011.
[6] Z. Zheng, Z. Lu, P. Sinha, S. Kumar, "Maximizing the contact opportunity for vehicular Internet access," *Infocom,* 2010.
[7] B. B. Chen, M. C. Chan, "MobTorrent: A framework for mobile Internet access from vehicles," *Infocom*, 2009.
[8] F. Malandrino, C. Casetti, C.-F. Chiasserini, M. Fiore, "Optimal content downloading in vehicular networks," *IEEE Trans. on Mob. Comp.*, vol. 12, no. 5, 2013.
[9] S. Yoon, D. T. Ha, H. Q. Ngo, C. Qiao, "MoPADS: A mobility profile aided file downloading service in vehicular networks," *IEEE Trans. on Veh. Tech.*, vol. 58, no. 9, 2009.
[10] TomTom, "How TomTom's HD TrafficTM and IQ RoutesTM data provides the very best routing," *White paper*, 2010.
[11] Waze, http://www.waze.com [Accessed on Nov. 2012].
[12] N. Lu, T.H. Luan, M. Wang, X. Shen, and F. Bai, "Capacity and Delay Analysis for Social-Proximity Urban Vehicular Networks", *Infocom*, 2012.
[13] H. Zhu, S. Chang, M. Li, S. Naik, S. Shen, "Exploiting temporal dependency for opportunistic forwarding in urban vehicular networks," *Infocom*, 2011.
[14] A. Balasubramanian, B.N. Levine, A. Venkataramani, "Enhancing interactive Web applications in hybrid networks," *Mobicom*, 2008.
[15] J. Mattingley, S. Boyd, "Real-time convex optimization in signal processing," *IEEE Sig. Proc. Mag.*, no. 27, vol. 3, 2010.
[16] C. Barberis, G. Malnati, "Epidemic information diffusion in realistic vehicular network mobility scenarios," *ICUMT,* 2009.
[17] D. Hadaller, S. Keshav, T. Brecht, S. Agarwal, "Vehicular opportunistic communication under the microscope," *MobySys,* 2007.
[18] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations," *5th Symposium on Math, Statistics, and Probability,* Berkeley, CA, 1967.
[19] W. Gould, J. Pitblado, W. Sribne, *Maximum likelihood estimation with stata*, Stata Press, 2006.
[20] E. Cohen, S. Shenker, "Replication strategies in unstructured peer-to-peer networks," *Sigcomm,* 2002.
[21] A. Balasubramanian, R. Mahajan, A. Venkataramani, "Augmenting mobile 3G using WiFi," *MobiSys*, 2010.
[22] S. Dimatteo, P. Hui, B. Han, V.O.K. Li, "Cellular traffic offloading through WiFi networks," *Mass*, 2011.
[23] R. L., Cruz, A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," *Infocom*, 2003.
[24] W. Gao, G. Cao, "User-centric data dissemination in disruption tolerant networks," *Infocom*, 2011.
[25] U. G. Acer, P. Giaccone, D. Hay, G. Neglia, S. Tarapiah, "Timely data delivery in a realistic bus network," *Infocom Mini-Conference*, 2011.
[26] P. Deshpande , A. Kashyap , C. Sung , S.R. Das, "Predictive Methods for Improved Vehicular WiFi Access," *Mobisys*, 2009.
[27] X. Li, X. Yu, A. Wagh, C. Qiao, "Human factors-aware service scheduling in vehicular cyber-physical systems," *Infocom*, 2011.

Francesco Malandrino (S'10) graduated (summa cum laude) in Computer Engineering from Politecnico di Torino in 2008. He received his PhD from Politecnico di Torino in 2011. From 2010 till 2011, he has been a visiting researcher at the University of California at Irvine. His interests focus on wireless and vehicular networks and infrastructure management.

Claudio Casetti (M'05) graduated from Politecnico di Torino in 1992 and received his PhD in Electronic Engineering from the same institution in 1997. He is an Assistant Professor at Politecnico di Torino. He has coauthored more than 130 papers in the fields of networking and holds three patents. His interests focus on ad hoc wireless networks and vehicular networks.

Carla-Fabiana Chiasserini (M'98, SM'09) received her Ph.D. in 2000 from Politecnico di Torino, where she is currently an Associate Professor. Her research interests include architectures, protocols, and performance analysis of wireless networks. Dr. Chiasserini has published over 200 papers at major venues, and serves as Associated Editor of several journals.

Marco Fiore (S'05, M'09) is a researcher at CNR–IEIIT, Italy, and at INRIA within the UrbaNet team hosted by the CITI Lab of INSA Lyon, France. He received M.Sc degrees from the University of Illinois at Chicago and Politecnico di Torino, in 2003 and 2004, respectively, and a PhD degree from Politecnico di Torino, in 2008. His research interests are in the field of mobile and vehicular networking.