Minimizing peak load from information cascades: Social networks meet cellular networks

Minimizing Peak Load from Information Cascades: Social Networks Meet Cellular Networks

Francesco Malandrino†, Maciej Kurant, *Member, IEEE,* Athina Markopoulou, *Member, IEEE,* Cedric Westphal, *Senior Member, IEEE*, Ulas C. Kozat, *Senior Member, IEEE*

Abstract—Online social networks (OSNs) serve today as a platform for information dissemination. At the same time, mobile devices provide ubiquitous network access through the cellular infrastructure. In this paper, we develop mechanisms for minimizing the peak load of the cellular network due to information cascades spreading on social media.

First, we exploit the social ties for predicting information dissemination and we propose Proactive Seeding– a technique for minimizing the peak load of cellular networks. Much of such a load is due to information cascades spreading in social media, and we address it by proactively pushing ("seeding") content to selected users before they actually request it. We develop a family of algorithms that take as input information primarily about: (i) cascades on the OSN, (ii) the background traffic load in the cellular network, and (iii) the local connectivity among mobiles; the algorithms then select which nodes to seed and when. We prove that Proactive Seeding is optimal when the prediction of information cascades is perfect. We perform simulations driven by traces from Twitter and cellular networks and we find that Proactive Seeding reduces the peak cellular load by 20%-50%. Then, we exploit the fact that there is correlation between social ties and physical proximity and we combine Proactive Seeding with device-to-device communication to further reduce the peak load.

Index Terms—Social network services, Cellular networks, Load management, Wireless networks

1 Introduction

In this paper, we are interested in the interaction of two important types of networks: online social networks (used as an overlay for information dissemination) and cellular networks (used as the underlying communication infrastructure). Both networks have seen explosive growth over the last years and there are several opportunities for synergy and cross-optimization.

† *Correspondence author – email: malandrino@tlc.polito.it; address: c.so Duca degli Abruzzi, 24, 10129 Torino, Italy*

F. Malandrino (malandrino@tlc.polito.it) is with Politecnico di Torino, Torino, Italy. M. Kurant (maciej.kurant@gmail.com) is with Google, Zurich, Switzerland. A. Markopoulou (athina@uci.edu) is with the University of California at Irvine, USA. C. Westphal (cedric.westphal@huawei.com) is with Huawei Innovation Center, Santa Clara, CA, USA. U. C. Kozat (kozat@docomoinnovations.com) is with Ozyegin University, Istanbul, Turkey. When this work was conducted, F. Malandrino was visiting UC Irvine and C. Westphal and U C. Kozat were with DOCOMO USA Labs, Palo Alto.

On one hand, cellular traffic is growing exponentially, tripling every year, with a share of video traffic increasing from 50% now to an expected 66% by 2015 [1]. For example, Credit Suisse reported in [2] that 23% of base stations globally had utilization rates of more than 80 to 85% in busy hours, up from 20% the year before. This dramatic increase in demand is generating serious problems for cellular networks. Since the cellular network is provisioned for *peak traffic*, mechanisms that distribute the network load more evenly over time are of interest to the operators. Essentially all this traffic is represented by data connections [3]; indeed, data traffic is so important that LTE does not foresee dedicated voice connections at all [4].

On the other hand, online social networks (OSNs), are an important way for users to get information. People tend to value highly the content recommended by friends or people with similar interests and are also likely to recommend it further to others. Furthermore, recommendation systems, increasingly used for providing personalized news, take into account social ties. By "OSNs", in this paper, we refer broadly to online information networks that exploit social ties to propagate information to users. Examples include online social networks (Facebook and Twitter), websites with social networking features (such as Digg.com, blogs), email communication, etc.

As mobile devices are becoming the primary way to access the Internet, including OSNs, we see a convergence of social and mobile networks. Most popular online social networks report heavy use from their mobile *apps*. For example, one third of all Facebook users regularly access the service from their mobile devices and they generate twice as much activity than non-mobile users [5]. Therefore, information diffusion over OSNs translates directly into increased cellular traffic. Cellular operators may exploit the knowledge of social ties to alleviate the peak demand in cellular traffic.

Our key observation is that given the vast information often available to the cellular operator and/or the OSN provider, we can, to a certain extent, *predict the future*

demand. Consider, for example, the case of YouTube videos: Google reported that 40% of YouTube videos are delivered to mobile devices in 2013 [6]. Many views of these videos are due to the spread of their URLs over various OSNs. The evolution of such cascades of forwarded URLs depends on the structure of the OSN, similarity of users and other features. With this information, it is possible to predict the diffusion of interest [7,8], and eventually the download of content that increases the cellular load. For example, in [9], the authors apply machine learning techniques to Twitter traces, and predict more than half of URL-based cascades of tweets with only a 15% false positive rate. In summary:

- much of the current (and future) load on cellular networks is represented by data traffic [3], most notably multimedia content [6]: in 2012, videos alone represented 50% of all mobile traffic [1], and this number is expected to reach 66% in 2017.
- the interest in such content increasingly propagates through online social networks [5];
- said interest spreading process has been carefully modeled [7,8] and can be predicted with remarkable accuracy [9].

Making conjectures on the ratio between predictable and non-predictable content would be difficult: on the one hand, not all interest in video content spreads through social media; on the other hand, video is not the only type of predictable content. In our performance evaluation we study a wide range of values for such a ratio, from $1:1$ to $1:6$.

One approach, typically referred to as *traffic shaping*, is to *delay* some of the traffic, *e.g.,* by limiting the diffusion of interest [10] or by using techniques that trade-off user delay for traffic load [11,12]. In other cases, mobile network operators can opt for *dynamic pricing* [13], where users are offered monetary incentives in exchange for a lower download speed [14]. We take a different approach, and aim at serving *impatient* users, *i.e.,* users that expect the content as soon as they become interested in it, and do not tolerate delay.

In this paper, we propose mechanisms for minimizing the peak load in cellular networks due to information cascades on social media. In particular, we propose Proactive Seeding, a technique for reducing the cellular peak load without introducing any additional delay in accessing the content. Proactive Seeding exploits social ties to predict future demand and proactively push ("seed") popular content to users before they request it. This allows to move some cellular traffic from the busiest hours to times with lower load and thus reduce its peaks, as illustrated in Fig. 1.

Our findings are the following. First, we consider the offline case, where the information cascades are assumed perfectly known. We prove that Proactive Seeding is optimal in that case, in the sense that it minimizes the peak load while delivering the content to users no later than they request it. We also show via simulation, driven by traces from Twitter and cellular networks, that Proactive Seeding leads to 20%-50% reduction in the cellular peak load. Second, we consider the more realistic case where prediction of the cascade is imperfect, and we show that Proactive Seeding, based on conservatively underestimating the future demand, brings positive gains. Finally, we exploit the fact that there is correlation between social ties and physical proximity and we combine Proactive Seeding with techniques [15,16] that exploit the local device-to-device (D2D) connectivity (over WiFi or Bluetooth) to deliver content. Proactive Seeding essentially (proactively) spreads the cellular load over time, while device-to-device connections offload cellular traffic over WiFi or BlueTooth. The combination of the two outperforms each individual technique.

Proactive Seeding is not free from potential shortcomings. The most important one is represented by the fact that, since predictions are never perfect, some users may receive a content they did not and will never want. This represents a waste of: (i) network resources; (ii) disk space on mobile phones; (iii) battery on mobile phones.

Since Proactive Seeding takes place during off-peak hours, needlessly consuming some network resources is not a serious issue. Similarly, present-day tablets and smartphones come with multi-gigabyte storage, hence storing an extra video is not a problem – clearly, "seeded" contents would come with an expiration time, so such a storage would anyway be temporary. Battery is potentially the most serious issue: a natural solution is to limit Proactive Seeding to those devices that are connected to a power source (as normally happens at night, which is when most seeding takes place) or anyway exclude devices with low battery.

A good way to make Proactive Seeding attractive for users is to waive the data fees related to seeded contents, whether they end up being requested or not. From the viewpoint of users, this means no extra charge in case the prediction fails, and even some free data if the seeded content ends up being requested. From the viewpoint of the operator, the reduced income is more than compensated by the benefits in terms of reduced peak load, as we see in Section 5. A related trend is represented by *sponsored data plans* [17,18], where content providers (and advertisers) incentivize the consumption of their content by subsidizing the subscribers' data traffic costs, i.e., offering some contents "toll free" from the users' viewpoint. In such a framework, the "seeds" selected by Proactive Seeding could actually be rewarded for their help in improving service rather than charged.

The structure of the rest of the paper is as follows. In Section 2, we formulate the problem. In Section 3, we present the Proactive Seeding algorithms under the assumption that demand can be perfectly predicted. In Section 4, we enhance our framework to allow for imperfect prediction. In Section 5, we present our evaluation results. After reviewing related work in Section 6, we conclude the paper in Section 7.

The conference version of this work appeared in IEEE INFOCOM 2012 [19]. This is as improved and

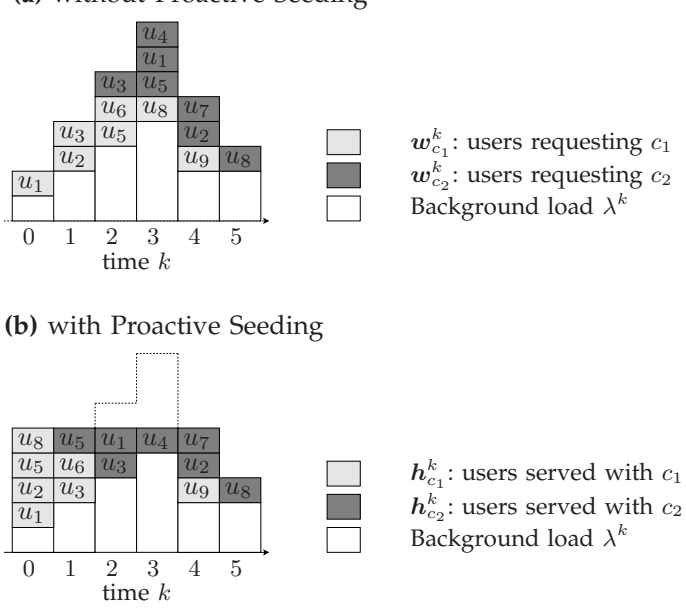

Fig. 1. Illustration of Proactive Seeding in a system with two types of contents $\boldsymbol{C} = \{c_1, c_2\}$ disseminated among 9 users $\boldsymbol{U} = \{u_1 \dots u_9\}$, in presence of the background load λ^k. **(a)** The diffusion of interest between the users in content c_1 (bright gray) and c_2 (dark gray). For example, $u_3 \in \boldsymbol{w}^2_{c_2}$ means that user u_3 becomes interested in content c_2 at time $k = 2$. Without Proactive Seeding, users request and pull the content through cellular right when they get interested in it ($\boldsymbol{h}^k_c \equiv \boldsymbol{w}^k_c$), which results in an uneven total cellular load (the total height of bars). **(b)** Proactive Seeding serves some users before they actually become interested in the content ($\boldsymbol{W}^k_c \subseteq \boldsymbol{H}^k_c$). The total load becomes more even in time and its peaks decrease (here by 3 units).

extended version this includes new materials such as: a chronology-preserving version of Proactive Seeding (Sec. 3.3) and its proof of optimality (Theorem 2); more details about the effectiveness of the prediction, along with a discussion of its impact on the performance of Proactive Seeding (Sec. 5.3.4, Tab. 1); a discussion about the coupling between social links and user mobility, and how it affects Proactive Seeding (Sec. 5.4, Fig. 8); a more detailed review of the related work.

2 Problem Statement

We distinguish between two components of cellular traffic: background load and predictable traffic.

2.1 Background Cellular Load

We denote as background (cellular) load all traffic which is out of our control: its content cannot be predicted and/or served before the actual request occurs. For example, phone conversations and other types of real-time traffic contribute to background load. We denote by λ^k the total amount of background load at time frame k, $0 \leq k \leq K$, with K being the time horizon we consider. Note that, even though the actual traffic (*e.g.,* the phone calls) cannot be predicted, the aggregate amount of traffic *i.e.,* λ^k is known [20] to follow remarkably regular patterns. We illustrate λ^k by white bars in Fig. 1; note that because the content composing it cannot be predicted or served earlier, λ^k remains unchanged in Fig. 1(b).

2.2 Predictable Cellular Traffic

In contrast, the predictable cellular traffic is all the traffic that can somehow be predicted and thus proactively served. As discussed in the Introduction, the bulk of predictable traffic is represented by multimedia contents, e.g., videos, that become popular through social networks. Denote by $\boldsymbol{U}$ the set of all users, and by $\boldsymbol{C}$ the set of all existing pieces of predictable content. We assume that transmitting a single piece $c \in \boldsymbol{C}$ of content to a single user $u \in \boldsymbol{U}$ takes exactly a single unit of cellular traffic.[1] Now, denote by $\boldsymbol{w}^k_c \subset \boldsymbol{U}$ the set of users that demand ("want") the content $c \in \boldsymbol{C}$ exactly at time frame k. In other words, $\boldsymbol{w}^k_c$ describes the diffusion of interest in content c (typically over OSNs). Let

$$\boldsymbol{W}^k_c = \bigcup_{m=0}^{k} \boldsymbol{w}^m_c \quad (\boldsymbol{W}^k_c \subseteq \boldsymbol{U}) \tag{1}$$

be the cumulative version of $\boldsymbol{w}^k_c$, *i.e.,* the set of all users that have requested c until frame k. Finally, we denote by $k(u, c)$ the time when user u demands content c, *i.e.,* such that $u \in \boldsymbol{w}^{k(u,c)}_c$.

In the example in Fig. 1(a), $\boldsymbol{w}^2_{c_1} = \{u_5, u_6\}$ and, consequently, $k(u_5, c_1) = k(u_6, c_1) = 2$.

2.3 Transmission Schedule

In this paper, we decouple the diffusion of interest in the content (*i.e.,* demand) from the actual delivery process. To this end, we denote by $\boldsymbol{h}^k_c \subset \boldsymbol{U}$ the set of users that get ("have") content c over cellular network exactly at frame k. Its cumulative version

$$\boldsymbol{H}^k_c = \bigcup_{m=0}^{k} \boldsymbol{h}^m_c \qquad (\boldsymbol{H}^k_c \subseteq \boldsymbol{U})$$

is the set of all users that have c at frame k. In the other words, $\boldsymbol{h}^k_c$ is a *schedule* that determines when the cellular operator sends content c to which users.

For example, in Fig. 1(b), $\boldsymbol{h}^1_{c_1} = \{u_3, u_6\}$ and $\boldsymbol{h}^1_{c_2} = \{u_5\}$.

2.4 User Impatience

In this work, we consider the case where all users are *impatient*: a user $u \in \boldsymbol{U}$ wants to enjoy content $c \in \boldsymbol{C}$ right after she becomes interested in it. This means that

1. In practice, the content spread over OSNs may greatly vary in size: a ten-minutes-long Youtube movie is orders of magnitude bigger than a photograph. All the equations can be easily modified to reflect heterogeneous content size, at the cost of notation clarity.

u should receive c at time l not larger than $k(u,c)$, *i.e.*, $u \in \boldsymbol{h}_c^l$ such that $l \leq k(u,c)$. This is achieved by guaranteeing that

$$\boldsymbol{W}_c^k \subseteq \boldsymbol{H}_c^k \quad \text{for every } k \text{ and } c. \tag{2}$$

We call such a schedule $\boldsymbol{h}_c^k$ *feasible*.

For example, in Fig. 1(b), we push content c_1 to user u_5 at time $k=0 < k(u_5,c_1)=2$, which is allowed by Eq.(2). In contrast, sending it at time $k>2=k(u_5,c_1)$ would violate the constraint in Eq.(2).

2.5 Objective

Using the notation above, the *total cellular traffic/load* at time k can be decomposed as the sum of background cellular load and total predictable traffic, *i.e.*,

$$\text{total cellular load} \quad = \quad \lambda^k + \sum_{c \in \boldsymbol{C}} |\boldsymbol{h}_c^k|. \tag{3}$$

Our objective is to minimize the peak of total cellular load, *i.e.*,

$$\text{minimize} \quad \max_{0 \leq k \leq K} \left(\lambda^k + \sum_{c \in \boldsymbol{C}} |\boldsymbol{h}_c^k| \right) \tag{4}$$

subject to the user impatience constraint in Eq.(2). Eq.(4) is the maximum (over all time frames, from 0 to our time horizon K) of the load – in other words, the peak load.

Note that because we have no control over the diffusion of interest $\boldsymbol{w}_c^k$, we can affect Eq.(4) only by choosing the schedule $\boldsymbol{h}_c^k$. We give an example of such an optimized schedule in Fig. 1(b). In particular, we (*i.e.*, the cellular operator) predict which users will be interested in content c, and proactively *seed* some of them with c when the cellular load is relatively small, *e.g.*, during the previous night. This allows us to reshape the cellular traffic and reduce its peaks, but not the total traffic.

3 Proactive Seeding Algorithms

In this section, we focus on the *offline* case, where we have perfect knowledge of the future diffusion of interest, *i.e.*, we know $\boldsymbol{w}_c^k$ for all time frames k and pieces of content c. The offline case serves as a baseline for understanding the maximum achievable gains. It also serves as a building block for the more realistic, *online* scenario, where prediction of the future is imperfect, described in Sec. 4.

3.1 Special Case: single content, no background load

Let us first consider the simplest, yet intuitive case: there is only a single content ($\boldsymbol{C}=\{c\}$) and no background load ($\lambda^k=0$). An example of the demand curve corresponding to such a cascade (*e.g.*, a single content flash-crowd) is shown in Fig. 2: the total number of users interested in the content increases until reaches a peak and then decreases.

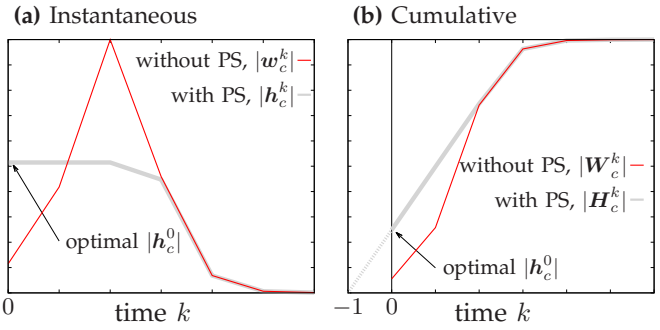

Fig. 2. Geometric interpretation of optimal Proactive Seeding (PS) under a single content cascade ($\boldsymbol{C}=\{c\}$), with no background cellular load ($\lambda^k=0$), as described in Sec. 3.1. The curves represent a typical cascade on the Facebook social graph (see Sec. 5.3). Red lines represent the *demand*, i.e., the number of content requests; gray lines represent the *load*, i.e., the number of copies downloaded by the users. If Proactive Seeding is not in place, demand and load overlap. We minimize the peak instantaneous cellular load in **(a)** while satisfying the impatience constraint Eq.(2), by proactively seeding the users at a constant rate, until the cascade passes. The optimal seeding rate $|\boldsymbol{h}_c^0|$ can be found by studying the cumulative version **(b)** of the time evolution, where a line anchored at point (-1,0) and tangential to $|\boldsymbol{W}_c^k|$, crosses the y-axis at point (0,$|\boldsymbol{h}_c^0|$). This is due to the fact that the value $|\boldsymbol{h}_c^0|$ also represents the slope of the load (gray) line.

In this special case, objective Eq.(4) is equivalent to minimizing $\max_k(|\boldsymbol{h}_c^k|)$ subject to the user impatience constraint Eq.(2). Intuitively, this entails delivering the content more evenly over time. Ideally, we would like to send the content with a constant *seeding rate* $|\boldsymbol{h}_c^k|$ and thus at linear $|\boldsymbol{H}_c^k|$.This rate should be the lowest possible, while still satisfying Eq.(2). Because $\boldsymbol{C}=\{c\}$, Eq.(2) is satisfied if $|\boldsymbol{W}_c^k| \leq |\boldsymbol{H}_c^k|$ for every k. Consequently, $|\boldsymbol{H}_c^k|$ should be linear and never smaller than $|\boldsymbol{W}_c^k|$. This leads to an intuitive geometric solution: Draw a straight line that crosses point (-1,0) and is tangential to $|\boldsymbol{W}_c^k|$. The optimal service rate $|\boldsymbol{h}_c^k|$ is determined by the point where the line crosses the y-axis. We show an example in Fig. 2.

It is also easy to see that this optimal rate $|\boldsymbol{h}_c^k|$ is also provided by the following formula

$$|\boldsymbol{h}_c^k| \quad = \quad \left\lceil \max_{l=k}^{K} \frac{|\boldsymbol{W}_c^0| - |\boldsymbol{H}_c^l|}{l+1} \right\rceil. \tag{5}$$

3.2 General Case: multiple contents, background traffic

The simple geometric solution from Sec. 3.1 does not directly extend to the general case, *i.e.*, in presence of arbitrary background cellular load $\lambda^k>0$ and multiple contents $|\boldsymbol{C}|>1$. For example, Eq.(5) would not necessarily satisfy the user impatience constraint Eq.(2) for each of the $|\boldsymbol{C}|>1$ contents separately.

Algorithm 1 Proactive Seeding

Require: $\boldsymbol{w}_c^k\ \forall c,k, \quad \lambda^k\ \forall k$ *future demand and load*

1: $\boldsymbol{h}_c^k \leftarrow \emptyset \quad \forall c,k$
2: $L \leftarrow \emptyset$
3: **for all** (u,c) such that $u \in \boldsymbol{W}_c^K$ **do**
4: $\quad L \leftarrow L \cup \{(u,c)\}$
5: **end for**
6: **sort** L by increasing $k(u,c)$
7: **for all** (u,c) in L **do** *water-filling*
8: $\quad k^* \leftarrow \arg\min_{0 \le l \le k(u,c)} (\lambda^l + \sum_c |\boldsymbol{h}_c^l|)$
9: $\quad \boldsymbol{h}_c^{k^*} \leftarrow \boldsymbol{h}_c^{k^*} \cup \{u\}$
10: **end for**
11: **return** $\boldsymbol{h}_c^k \quad \forall c,k$ *optimal*

To address these problems, we propose the Proactive Seeding algorithm, shown in Algorithm 1. We construct the seeding schedule $\boldsymbol{h}_c^k$ iteratively, starting from an empty set (line 1). In lines 2-6, we create a list L of existing user-content pairs (u,c), sorted according to the growing want times $k(u,c)$. Note that user u may appear in L multiple times, *i.e.,* exactly once for each content c she is interested in. Lines 7-9 implement a water-filling type of algorithm, where for each pair (u,c) we find the time frame $k^* \le k(u,c)$ with the smallest total cellular load $\lambda^{k^*} + \sum_c |\boldsymbol{h}_c^{k^*}|$. We then schedule this pair (u,c) at time k^* by adding u to $\boldsymbol{h}_c^{k^*}$ (line 9). Finally, once all existing pairs (u,c) are scheduled, Proactive Seeding returns the seeding schedule $\boldsymbol{h}_c^k$ for all contents c and time frames k.

We illustrate the output of Proactive Seeding in the example of Fig. 1(b). The sorted list L resulting after line 6 is $L = [(u_1,c_1),(u_2,c_1),(u_3,c_1),(u_5,c_1),(u_6,c_1),(u_3,c_2),(u_8,c_1),(u_5,c_2),(u_1,c_2),(u_4,c_2),(u_9,c_1),(u_2,c_2),(u_7,c_2),(u_8,c_2)]$. For pair (u_1,c_1), we have $k(u_1,c_1) = 0$, and therefore lines 8-9 result in $k^* = 0$ and $\boldsymbol{h}_{c_1}^0 = \{u_1\}$, respectively. When processing the second element in L, (u_2,c_1), we have $\lambda^l + \sum_c |\boldsymbol{h}_c^l| = 2$ for both $l = 0$ and $l = 1$. We arbitrarily break this tie by setting $k^* = 0$, which results in $\boldsymbol{h}_{c_1}^0 = \{u_1,u_2\}$. The third pair (u_3,c_1) has now a unique $k^* = 1$, and is scheduled therein. The process continues until L is exhausted.

This schedule $\boldsymbol{h}_c^k$ returned by Proactive Seeding is optimal, as we can show through the following:

Theorem 1 (Optimality of Proactive Seeding)**.** *The seeding schedule $\boldsymbol{h}_c^k, \forall c,k$, created by Proactive Seeding minimizes the peak load (objective in Eq.(4)), while satisfying the user impatience constraint Eq.(2) for each content c separately.*

Proof: First, note that the frame k^* chosen for user u in line 8 is not greater than the time $k(u,c)$ when u actually wants the content. Therefore, by construction, the schedule created by Proactive Seeding always satisfies the user impatience constraint Eq.(2) for every content c separately.

We now have to prove that the objective Eq.(4) is met by Proactive Seeding. Denote by $L(j)$ the set of all pairs (u,c) such that $k(u,c) = j$ and by $L(i,j) = \bigcup_{m=i}^{j} L(m)$. Denote by $\boldsymbol{h}(j)$ the transmission schedule constructed by Proactive Seeding just after processing the pairs $L(j)$ in lines 7-9. In other words, $\boldsymbol{h}(j)$ schedules all contents for all users that want it not later than at time j. Consequently, $\boldsymbol{h}(K)$ denotes the entire schedule, $\boldsymbol{h}(K) \equiv \bigcup_{c,k} \boldsymbol{h}_c^k$. We prove the optimality of Proactive Seeding by induction on j, as follows.

Initialization ($j = 0$): For every pair $(u,c) \in L(0)$, line 8 automatically sets $k^* = 0$. Consequently, $\boldsymbol{h}(0)$ schedules all pairs $L(0)$ at time slot 0. This is the only feasible solution, thus the optimal one.

Induction step: Assume that $\boldsymbol{h}(j)$ is optimal for all pairs $L(0,j)$. We now must prove that $\boldsymbol{h}(j+1)$ is optimal for all pairs $L(0,j+1)$.

Denote by $\max(\boldsymbol{h}(j))$ the peak total cellular load resulting from $\boldsymbol{h}(j)$. Either an optimal allocation will increase the peak rate at $j+1$, or keep it constant. Thus we can distinguish two cases, as follows:

Case 1: It is possible to schedule the pairs $L(j+1)$ such that $\max(\boldsymbol{h}(j+1)) = \max(\boldsymbol{h}(j))$. In this case, lines 7-9 guarantee that this equality holds under Proactive Seeding, by iteratively choosing the least loaded time slots. Now, because $\max(\boldsymbol{h}(j))$ is optimal, it is the smallest value that does not violate the impatience constraint Eq.(2). So $\boldsymbol{h}(j+1)$ cannot be lower than $\max(\boldsymbol{h}(j))$ without violating Eq.(2). Consequently, $\max(\boldsymbol{h}(j+1)) = \max(\boldsymbol{h}(j))$ implies the optimality of $\boldsymbol{h}(j+1)$.

Case 2: It is *not* possible to schedule the pairs $L(j+1)$ such that $\max(\boldsymbol{h}(j+1)) = \max(\boldsymbol{h}(j))$. We can now distinguish two sub-cases, depending of the background load at time $j+1$:

Case 2.1: If $\max(\boldsymbol{h}(j+1)) = \lambda^{j+1}$ is achievable, then lines 7-9 of Proactive Seeding will achieve that by iteratively choosing the least loaded time slots. In this case, the peak load is equal to the background load λ^{j+1}. Such a peak load is optimal, because, by definition, background load cannot be changed.

Case 2.2: If $\max(\boldsymbol{h}(j+1)) = \lambda^{j+1}$ is *not* achievable, then lines 7-9 guarantee that $\max(\boldsymbol{h}(j+1)) - \min(\boldsymbol{h}(j+1)) \le 1$, where $\min(\boldsymbol{h}())$ denotes the minimal total cellular load resulting from $\boldsymbol{h}()$. Consequently, $\max(\boldsymbol{h}(j+1))$ cannot be decreased and $\boldsymbol{h}(j+1)$ is thus optimal. □

3.3 Serving in Chronological Order

Although optimal in the sense of objective Eq.(4), Proactive Seeding does not guarantee that the users will be served in the order they request the content; it may schedule user u_i before user u_j, even if $k(u_i,c) > k(u_j,c)$. For example, in Fig. 1 user u_3 wants content c_1 before user u_5, but is scheduled to receive it after u_5, as we show in Fig. 1(b). Arguably a better solution would be to seed u_3 before u_5, which would give both users one time slot of margin to accommodate potential prediction errors. Fortunately, it is easy to see that reshuffling the users to enforce such "first-want-first-serve" (*i.e.,* chronological)

order, preserves the optimality and feasibility of the resulting schedule $\boldsymbol{h}_c^k$. More specifically:

Theorem 2 (Chronological order)**.** *Let $\boldsymbol{h}_c^k$ be feasible and optimal, and let $\tilde{\boldsymbol{h}}_c^k$ be a version of $\boldsymbol{h}_c^k$ that reshuffles the users interested in c to enforce the "first-want-first-serve" order,* i.e.,
(i) $|\tilde{\boldsymbol{h}}_c^k| = |\boldsymbol{h}_c^k| \quad \forall k, c$.
(ii) If $k(u_i, c) < k(u_j, c)$ then $\tilde{\boldsymbol{h}}_c^k$ schedules u_i before u_j.
Then $\tilde{\boldsymbol{h}}_c^k$ is feasible and optimal too.

Proof: First, because the objective function Eq.(4) depends on cardinality $|\tilde{\boldsymbol{h}}_c^k|$, (i) guarantees the optimality of $\tilde{\boldsymbol{h}}_c^k$.

Second, (ii) implies that for a given content c, users are added to $\tilde{\boldsymbol{H}}_c^k$ in the same order as they appear in $\boldsymbol{W}_c^k$ (*i.e.,* as they want the content). Consequently, the feasibility condition Eq.(2) is reduced to $|\boldsymbol{W}_c^k| \leq |\tilde{\boldsymbol{H}}_c^k|$. The latter is always satisfied, because $|\tilde{\boldsymbol{H}}_c^k| = |\boldsymbol{H}_c^k|$ (implied by (i)), and $|\boldsymbol{W}_c^k| \leq |\boldsymbol{H}_c^k|$ (feasibility of $\boldsymbol{h}_c^k$). □

3.4 Contents of heterogeneous size

For simplicity of presentation, we have so far assumed that all contents have the same size, and that each content can be served in a time frame, as it happens in Fig. 1. It is worth stressing that Proactive Seeding and our algorithms work unmodified even if this is not the case.

Suppose that a certain content is bigger than the others, and fills two of the blocks we show in Fig. 1. We can split that content in two chunks, each of which can be served in a time frame, and consider those chunks as two separate contents. Each of these "virtual contents" will have the same deadline of the original content, and will be served – either seeded or fetched – in time.

3.5 Extension: D2D-aware Proactive Seeding

In addition to their cellular connections, it is often the case that some users are within physical proximity of each other and can establish direct device-to-device (or D2D [21]) connections between them, *e.g.,* via ad-hoc 802.11 or Bluetooth. If these users are interested in the same content, they can exploit their D2D connectivity, and thus offload the cellular network. Several variants of this idea have been studied in the past, *e.g.,* in [15,16,22,23]. What makes this particularly promising, in our context, is the fact that there is a correlation between proximity on the social graph and geographical proximity, at both medium [24] and small [25] scale. We show below (and later, in simulations) that these techniques can be combined with Proactive Seeding, and address two complementary aspects: using the D2D connections helps to offload the total aggregated cellular load, while Proactive Seeding helps to smooth the load over time.

The D2D connectivity graph changes over time. We denote by $\boldsymbol{N}^k(u)$ all D2D neighbors of user u at time k. Consider time $k(u, c)$ when user u becomes interested in content c. We will assume that each mobile user behaves as follows:

1) If u has been seeded with c before, no action is needed.
2) Otherwise, u attempts to pull c from its current local neighbors $\boldsymbol{N}^{k(u,c)}(u)$. This is possible only if at least one of these neighbors has c, *i.e.,* if $\boldsymbol{N}^{k(u,c)}(u) \cap \boldsymbol{H}_c^{k(u,c)} \neq \emptyset$.
3) Otherwise, u fetches c through the cellular network.

Depending on the extent to which the operator is aware of D2D connectivity, different optimizations are possible:

3.5.1 D2D-unaware Proactive Seeding

In this simplest scenario, the operator does not have information about the location of users and thus performs Proactive Seeding without taking proximity into account. Consequently, user u can benefit from D2D, in an opportunistic way, *i.e.,* only if u has not been seeded earlier (*i.e.,* if $u \in \boldsymbol{h}_c^{k(u,c)} \cap \boldsymbol{w}_c^{k(u,c)}$), which results in

$$\boldsymbol{h}_c^k \leftarrow \boldsymbol{h}_c^k \setminus \left\{ u \in \boldsymbol{h}_c^k \cap \boldsymbol{w}_c^k : \quad \boldsymbol{N}^{k(u,c)}(u) \cap \boldsymbol{H}_c^{k(u,c)} \neq \emptyset \right\}.$$

In the example of Fig. 1, user u_4 will pull content c_2 from its D2D neighbors $\boldsymbol{N}^3(u_4)$ at time $k = 3$ if at least one of them is in $\{u_1, u_3, u_5\} = \boldsymbol{H}_{c_2}^2$ (*i.e.,* already has c_2).

3.5.2 D2D-aware Proactive Seeding

In this scenario, the operator has information about location and thus proximity of users and takes it into account while seeding. In particular, it applies Proactive Seeding but avoids seeding user u if u will be able to get the content from its neighbors. This can be achieved by the following refinement of schedule $\boldsymbol{h}_c^k$:

$$\boldsymbol{h}_c^k \leftarrow \boldsymbol{h}_c^k \setminus \left\{ u \in \boldsymbol{h}_c^k : \quad \boldsymbol{N}^{k(u,c)}(u) \cap \boldsymbol{H}_c^{k(u,c)} \neq \emptyset \right\}. \quad (6)$$

In the example of Fig. 1, we will seed user u_5 with content c_2 at time $k = 1$. If we know that $u_5 \in \boldsymbol{N}^3(u_1)$, *i.e.,* that u_1 and u_5 will form a D2D connection at time $k = 3$ (*i.e.,* when u_1 wants c_2) then then we can exclude u_1 from $\boldsymbol{h}_{c_2}^2$.

Notice that even in the D2D-aware case, the operator has information about the *current* position and connectivity of the users, not the future ones. Recall that, owing to the feedback mechanism described in Fig. 3, the refined schedule in Eq.(6) is computed at time k and not before.

4 Dealing with Uncertainty

In Sec. 3, we developed an optimal seeding strategy given the full and precise knowledge of the future (i) cellular background load, and (ii) predictable traffic pattern. Clearly, the performance of Proactive Seeding will strongly depend on the quality of our estimation of the predictable traffic $\boldsymbol{w}_c^k$. Many prediction techniques

have been proposed in the literature[2] and developing new ones is out of the scope of this paper. Instead, in this section, we review some existing techniques, and we show how they can be incorporated in Proactive Seeding.

4.1 Interest diffusion on OSNs

In this paper, we are interested in the content that becomes popular through social ties. One can exploit the structure of the social network and information about interest diffusion, in order to predict information cascades. Such a prediction can then serve as input (instead of the offline knowledge) to our predictive seeding algorithms.

There is a rich literature on predicting the diffusion of interest in social networks, see *e.g.*, [7,8]. In our context, predicting the future progress of a cascade related to content c, can be modeled as finding the probability

$$\mathbb{P}\left(\boldsymbol{w}_c^{k+1}, \boldsymbol{w}_c^{k+2}, ... \mid \boldsymbol{w}_c^{k}, \boldsymbol{w}_c^{k-1}, ..., \boldsymbol{w}_c^{0},\ I_{other}\right), \quad (7)$$

where $\boldsymbol{w}_c^{k}, \boldsymbol{w}_c^{k-1}, ..., \boldsymbol{w}_c^{0}$ is the observed history at the current time k, and I_{other} represents any other available piece of information. Below, we comment on how some of the existing approaches translate into the Eq.(7) probabilities.

4.1.1 The threshold model

In the threshold model [7], each user u is associated with a threshold $0 \leq \theta_u \leq 1$. u becomes interested in the content at time $k+1$ if at least a (weighted) fraction of θ_u of her neighbors are interested in it at time k. This model is deterministic, *i.e.*, the probabilities in Eq.(7) are either 0 or 1.

4.1.2 The cascade model

In the cascade model [7,8], each edge (u, w) of the social graph is associated with an activation probability $q_{u,w}$. If user u gets interested in the content at time k, then the edge (u, w) is used exactly once to determine whether user w will become interested in the content at frame $k+1$, which happens with probability $q_{u,w}$. In other words, given the activation probabilities $q_{u,w}$ (*i.e.*, I_{other}) and the history $\boldsymbol{w}_c^{k}, \boldsymbol{w}_c^{k-1}, ..., \boldsymbol{w}_c^{0}$, the cascade model gives us the following probabilities, concerning the next time frame:

$$\mathbb{P}\left(\boldsymbol{w}_c^{k+1} \mid \boldsymbol{w}_c^{k}, \boldsymbol{w}_c^{k-1}, ..., \boldsymbol{w}_c^{0},\ I_{other}\right), \quad (8)$$

which is a special case of Eq.(7).

2. Clearly, the performance of these techniques depends on the amount of information available to train them. Thankfully, such information is plentiful. The cellular operator has already access to activity on the phone, such as address books, session logs, location history. Furthermore, operators may partner with OSNs to obtain either raw information about the social graph and user activity, the results of cascade prediction performed on the OSN side. Finally, the users themselves could voluntarily disclose their information, *e.g.*, by running an app directly on their phones, in exchange for faster access to content and cheaper data plans.

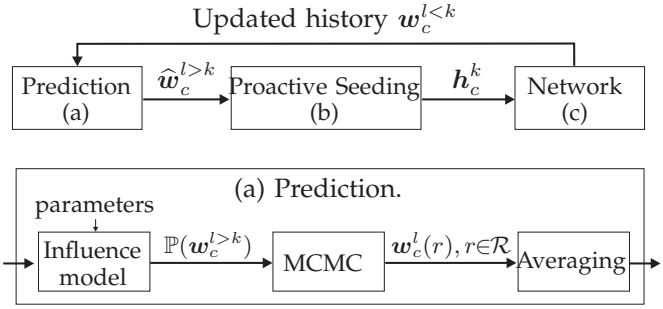

Fig. 3. Adaptive Proactive Seeding. (top) High-level overview. (bottom) The "Prediction" block.

4.1.3 Machine learning

Another line of research focuses on machine learning techniques that make use of all the available information. For example, in [9], the authors, based on the observed history, manage to accurately predict more than half of future re-tweets (of URL links) with 15% false positives.

4.2 From probabilities to Proactive Seeding

Given the knowledge of probabilities in Eq.(7), we follow the procedure presented in Fig. 3. First, at the current time k, we use Eq.(7) to calculate the most likely future $\widehat{\boldsymbol{w}}_c^{l>k}$ (Fig. 3(a)). Next, we plug $\widehat{\boldsymbol{w}}_c^{l>k}$ into Proactive Seeding (Fig. 3(b)), which returns us the schedule $\boldsymbol{h}_c^k$ for the current time frame. Finally, we implement $\boldsymbol{h}_c^k$ and collect the actual evolution of demand $\boldsymbol{w}_c^k$ that is used to refine our calculations in the next time frame (Fig. 3(c)). This means that our scheme is *adaptive* – at every iteration it updates the history by the current state of the network and recalculates $\boldsymbol{h}_c^k$. Our prediction includes all times l between the current time k and our time horizon K.

For instance, for the cascade influence model, K is upper-bounded by the total number of users, *i.e.*, $K \leq |\boldsymbol{U}|$. Indeed, if no new users are activated at a frame $\hat{k}$, no users will be activated at any frames $k > \hat{k}$. Therefore, as long as the cascade lasts there is at least a new activation per frame, which limits the number of frames a cascade can last to $K \leq |\boldsymbol{U}|$.

In Fig. 3(bottom), we show in more detail the "Prediction" block from Fig. 3. Given the knowledge of Eq.(7), we are, in principle, able to calculate exactly the expected future demand $\mathbb{E}[\boldsymbol{w}_c^{l>k}]$. In practice, however, the solution space is too big (especially if the number $|\boldsymbol{U}|$ of users or the final time K are large) to do it precisely. Instead, we run an MCMC (Monte Carlo Markov Chain) simulation, *i.e.*, we use Eq.(7) to generate a number of realizations $\boldsymbol{w}_c^{l>k}(r)$, $r \in \mathcal{R}$. This step is illustrated by the middle block in Fig. 3(bottom). Next, we average over all $|\mathcal{R}|$ realizations (right-most block in Fig. 3, bottom), as follows.

First, we estimate the *number of users* $|\widehat{\boldsymbol{W}}_c^K|$ that eventually become interested content c, by the average over

all the realizations:

$$|\widehat{\boldsymbol{W}}_c^K| = \frac{1}{|\mathcal{R}|} \cdot \sum_{r \in \mathcal{R}} |\boldsymbol{W}_c^K(r)|.$$

Next, we decide *which users* will become interested in the content, by taking $|\widehat{\boldsymbol{W}}_c^K|$ users with the highest observed probabilities $\widehat{\mathbb{P}}(u \in \boldsymbol{W}_c^K) = \frac{1}{|\mathcal{R}|} \cdot |\{r \in \mathcal{R} : u \in \boldsymbol{W}_c^K(r)\}|$ to request it. Finally, we interpret as $k(u,c)$ the time that is the most frequent across the realizations in R:

$$\widehat{k}(u,c) = \arg \max_{0 \le k \le K} |\{r \in \mathcal{R} : u \in \boldsymbol{w}_c^k(r)\}|.$$

The above process provides an estimate $\widehat{\boldsymbol{w}}_c^k$ of the future demand, which we use as input to Proactive Seeding, as in Fig. 3(b).

5 Evaluation

In this section, we evaluate the performance of Proactive Seeding through simulation.

5.1 Performance Metric

Without Proactive Seeding, user u fetches the content c over cellular when she wants it, which yields $\boldsymbol{h}_c^k \equiv \boldsymbol{w}_c^k$ and the peak cellular load equal to $\max_k \left(\lambda^k + \sum_c |\boldsymbol{w}_c^k|\right)$. In contrast, with Proactive Seeding, the peak cellular load drops to $\max_k \left(\lambda^k + \sum_c |\boldsymbol{h}_c^k|\right)$. Our main performance metric is the relative *gain* in peak cellular load, defined as

$$\gamma \;=\; \frac{\max_k \left(\lambda^k + \sum_c |\boldsymbol{w}_c^k|\right) - \max_k \left(\lambda^k + \sum_c |\boldsymbol{h}_c^k|\right)}{\max_k \left(\lambda^k + \sum_c |\boldsymbol{w}_c^k|\right)}.$$

Clearly, the larger the amount of the predictable traffic, the bigger gain γ we can expect. We therefore denote by ρ the ratio of the unpredictable traffic (aggregate over all contents) over the aggregate predictable traffic, *i.e.*,

$$\rho \;=\; \frac{\text{aggregated unpredictable traffic}}{\text{aggregated predictable traffic}} \;=\; \frac{\sum_k \lambda^k}{\sum_k \sum_c |\boldsymbol{w}_c^k|}. \quad (9)$$

5.2 Offline Scenario

First, we consider the offline case, with large-scale simulations fed by real traces of (a) interest diffusion process in Twitter [9], (b) background traffic from a US cellular operator [20], and (c) mobility [26]. This allows us to evaluate Proactive Seeding in presence of cellular background load and techniques that exploit D2D connectivity. We assume a priori knowledge of (a), (b), (c), and we evaluate how much gain γ is achieved by Proactive Seeding.

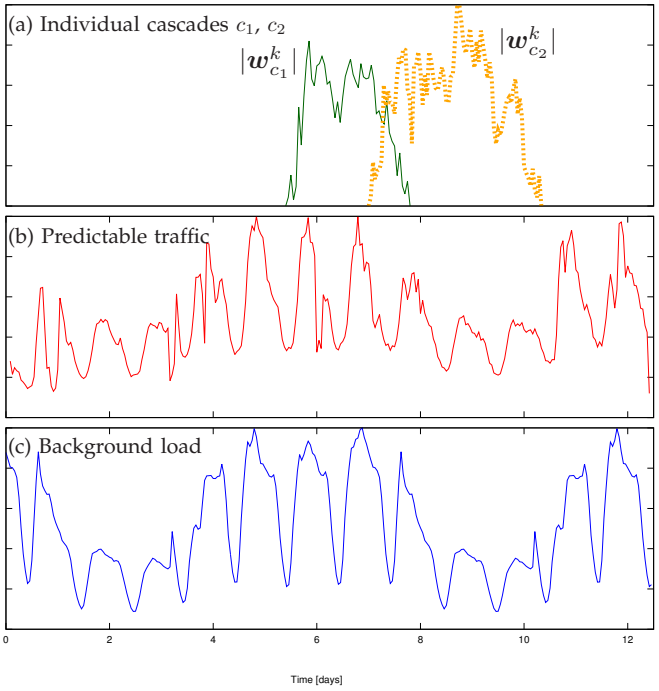

Fig. 4. Traces used in offline simulations. **(a)** Example of two individual Twitter cascades; **(b)** All 9000 Twitter cascades together [9]; **(c)** Background cellular load from a US operator [20]. For the sake of readability, all figures are normalized with respect to the peak value of the data they represent (*i.e.*, they do not have the same scale).

5.2.1 Description of Datasets

(a) Predictable traffic π^k: We use the Twitter trace from [9], where the authors collected the tweets that carry a URL (which defines our content), over a period of 300 hours (12.5 days). For our simulations, we kept only the "retweets" (indicated by an `RT` tag), which allows us to directly follow the cascades of interests in valuable (non-spam) content on Twitter (see also RT-cascades in [9]). Furthermore, in order to be able to observe the full evolution of such cascades, we exclude the URLs that appear in the first three or the last three hours of the trace. This leaves us with around 2.5M of tweets from 554K different users, sharing about 9000 contents (URLs). In Fig. 4(a), we show the evolution of two typical cascades from that trace. The "cascade" behavior is easy to see: the URL's popularity quickly increases over time, reaches a peak, and then declines. However, when we aggregate all the 9000 cascades together in Fig. 4(b), the individual cascade shapes are not visible anymore; instead, the aggregated predictable traffic π^k clearly follows the daily pattern.[3]

(b) Background cellular load λ^k: As background load λ^k, we take a cellular traffic trace coming from a major

3. Recall, however, that our constraint Eq.(2) is defined for each content, not for the aggregated traffic.

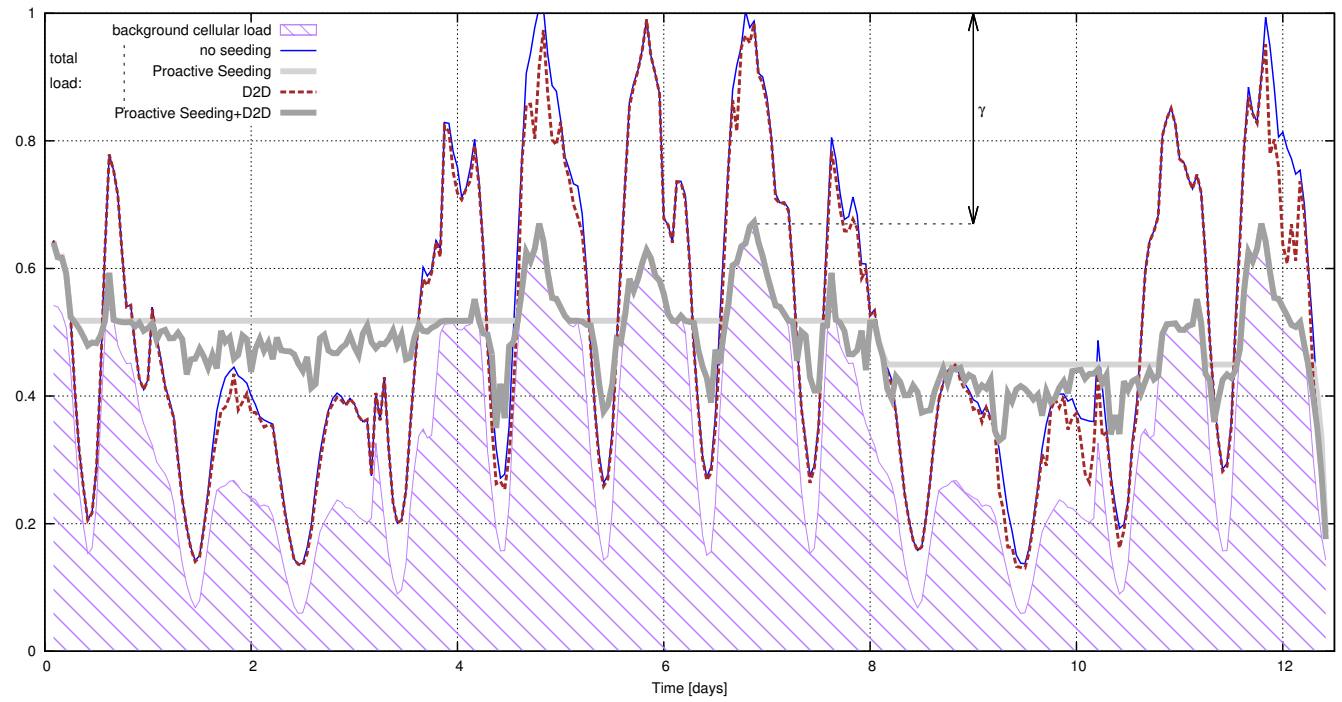

Fig. 5. Offline simulations driven by traces of (i) Twitter cascades (predictable traffic), (ii) background cellular load, and (iii) mobility: per-hour time evolution of the total cellular load $\lambda^k + \pi^k$ under various scenarios, for a traffic ratio of $\rho = 2$.

operator in one US state [20].[4] Because this trace covers one full week (at a resolution of 1 hour), we replicate it, concatenate, and shift to match the 12.5 days of the Twitter trace. The result is presented in Fig. 4(c). Similarly to Twitter, the cellular background load follows weekly and daily patterns.

(c) D2D connectivity: We use the Dartmouth/Campus contact trace from the Crawdad repository [26] to simulate the device-to-device (D2D) connectivity. The trace logs the activity of over 25,000 users for a period of eleven weeks. In particular, it includes association logs for 476 APs over 161 buildings.

For each content c, we randomly map the users $\boldsymbol{H}_c^K$ (*i.e.,* eventually requesting c) to the users in the trace. We assume that two users can exchange data in a D2D fashion if they are associated to the same AP.

This ensures that the D2D connectivity we observe for the 554K users of simulation has the same features, e.g., node degree and contact duration, of the original trace.

4. Strictly speaking, the trace [20] represents the total cellular traffic. For simplicity of presentation (*e.g.,* independence of ρ), we interpret this trace as the background cellular load λ^k. We have also considered in simulations this trace as the total load, subtracting π^k to get the background load. The results in both cases are very similar.

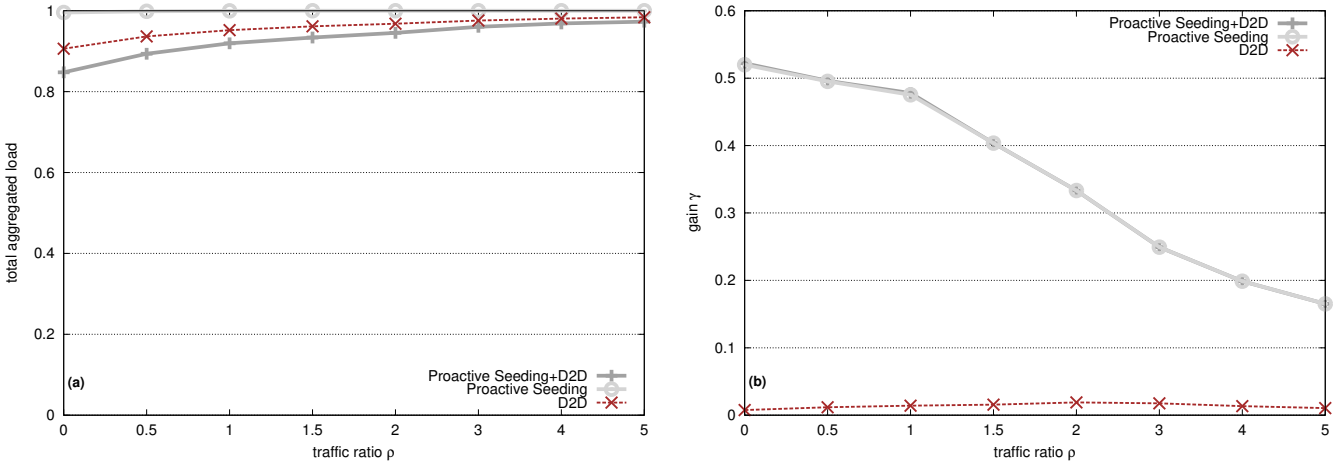

Fig. 6. Aggregated cellular traffic (a) and gain γ (b) as a function of the traffic ratio ρ.

The above mapping matches users $\boldsymbol{U}$ with nodes in the mobility trace in a purely random way. Indeed, the *coupling* between mobility and social relationships has a paramount importance for the performance of Proactive Seeding, and we carry out a more thorough investigation on this issue in Sec. 5.4.

5.2.2 Results

In Fig. 5 we focus on a case when $\rho = 2$, *i.e.*, the background load is twice the predictable traffic, and depict the time evolution of the total load on the 3G network in the following cases:

- no seeding: All users get the content they are interested in through the cellular network (*i.e.*, $\boldsymbol{h}_c^k = \boldsymbol{w}_c^k, \forall c, k$).
- Proactive Seeding: Proactive Seeding algorithm is used to schedule predictable traffic. D2D is disabled.
- D2D: Users exploit the D2D connectivity as explained in Sec. 3.5, but Proactive Seeding is disabled.
- D2D-unaware Proactive Seeding: predictable traffic is scheduled using Proactive Seeding *and* users exploit D2D links if available.

The no-seeding scenario results in a cellular load that is very uneven over time, with high peaks and periods of very low usage. Under D2D, we observe a slight reduction in the network load, with the peaks almost unchanged. In contrast, Proactive Seeding effectively reshapes the total cellular traffic, reducing the peaks by exploiting the less busy periods. Note that the peak load (around day 9) corresponds to a peak in the *background* load, which confirms that Proactive Seeding is optimal with respect to objective Eq.(4) (as we proved in Theorem 1). Finally, when we combine Proactive Seeding and D2D, we observe a further reduction in the network load.

Fig. 6(a) and Fig. 6(b) show how the aggregated (*i.e.*, over the whole trace duration) load and the gain γ depend on the ratio ρ between predictable and background load. Unsurprisingly, the higher ρ, the less beneficial Proactive Seeding becomes. Proactive Seeding effectively reduces the peak load (Fig. 6(b)), but has no impact on the aggregated load (Fig. 6(a)). The effect of D2D is quite the opposite. Applying both Proactive Seeding and D2D, we get the best of both worlds: *i.e.*, a significant reduction in both the peak and the aggregated load.

5.3 The Online Case (using Diffusion Models on OSNs)

Sec. 5.2 assumed full knowledge of the entire traces. In this section, we consider the case where the future can be predicted only with some amount of uncertainty, as described in Sec. 4. For ease of explanation, we assume no background load and a single content c and we focus on evaluating the effect of uncertainty on the results.

5.3.1 Social Graphs (Datasets)

We use datasets from two different graphs, each capturing a different type of social tie.

- *Facebook*: The New Orleans network of the Facebook social graph [27], consisting of 63K vertices and 816K edges. The rationale for using this data set is that friends in Facebook share links and thus participate in spreading information about content.
- *Email*: a trace [28] of e-mail contacts, collected within the Enron company in 2004, consisting of 1133 nodes and 5452 edges. The rationale behind using this datasets is that emails often contain links that propagate in a viral way, leading to information cascades.

5.3.2 Social Influence (Models)

Using each of the previous graphs, we simulate interest diffusion through the cascade model [7,8] described in Sec. 4.1.2. We assume that 5% of users are interested in the content at time $k = 0$. The activation probability for each edge (u, w) is set to $q_{u,w} = 0.1$.

By setting such a high value, we set ourselves in the most interesting (and challenging) scenario: on the one hand, the value is high enough for social cascades to happen; on the other, it is low enough to avoid that eventually all users become activated – which would make it trivial to foresee who the activated users are.

5.3.3 Uncertainty about the model and its parameters

Although the cascade model provides us with a probabilistic output, there are several other major sources of uncertainty about the future, which naturally lead to errors in the prediction. In particular, in practice, (i) we can never know exactly the model driving the spread of information and (ii) we can never know precisely the parameters of such a model. We capture these two effects in our simulations by introducing a multiplicative noise ν to the probabilities Eq.(7), *i.e.*, we set $\mathbb{P}() \leftarrow \min(1, \nu\,\mathbb{P}())$. For example, $\nu = 1.2$ results in a systematic overestimation of the future demand by 20%, and $\nu = 0.8$ underestimates it by 20%.

5.3.4 Prediction performance

Tab. 1 shows the prediction performance for the Facebook graph. Clearly, the prediction is far from ideal - the rates of false positives and false negatives are similar to those reported in [9]. Also, as expected, a higher ν implies a higher rate of false positives and a lower rate of false negatives, and vice versa.

There is a significant number of users for which the time $\hat{k}(u, c)$ at which they are expected to become interested in the content is later than the actual time $k(u, c)$. This particular kind of prediction error can potentially have the same effect of a false negative, because if user u is not served before time $k(u, c)$, she will anyway fetch content c from the cellular network. Fortunately, as shown in the fourth line of the table, this effect is negligible with respect to other sources of error. Indeed, users which are predicted to request the content are served, in general, before the time $\hat{k}(u, c)$ at which the request is expected, and such a time is in most cases earlier than the actual request time $k(u, c)$.

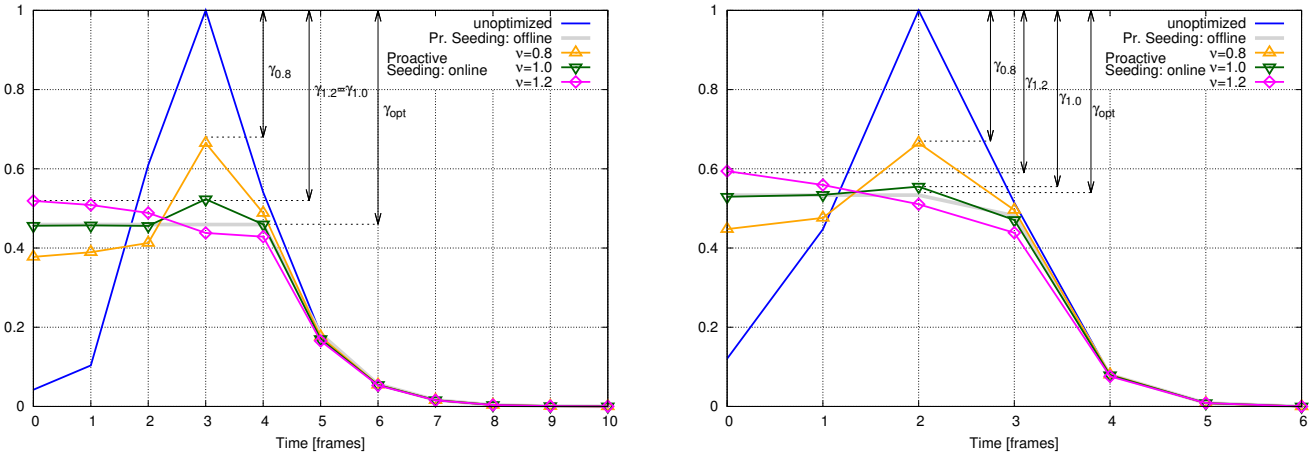

Fig. 7. Online simulations: normalized load for the Facebook (left) and Email (right) graphs. As in Fig. 2, in the "unoptimized" case demand and load coincide.

	$\nu = 0.8$	$\nu = 1$	$\nu = 1.2$
False positives (among all users)	11%	14%	18%
False negatives (among all users)	12%	8%	6%
Late true positives $\hat{k}(u,c) > k(u,c)$ (among wanters)	32%	28%	26%
True positives which had to fetch (among wanters)	6%	6%	5%

TABLE 1

Prediction performance for different values of the multiplicative noise ν in the Facebook graph with 48K users out of which 40K (83%) eventually want the content. Results are averaged over 50 independent simulation runs. The first and second line show the number of false positives (*i.e.*, users that are predicted to request the content while they do not) and false negatives (*i.e.*, users that are not predicted to request the content, while they do). The third line shows the number of users that are correctly predicted to request the content, but for which the predicted request time is later than the actual one, and the fourth line shows how many of such users are scheduled for service too late (*i.e.*, they have to fetch the content).

With respect to the concern we raised in the introduction, i.e., that some users may be pushed a content they will never want, the first line of Tab. 1 shows that they represent more than 18% of the total downloads, even when the prediction is severely flawed.

Pushing an unrequested content to users once every five times may not sound like a very effective strategy: considering the whole graph and the average size of a YouTube video [29], it translates into almost 60 GByte of extra data transfer throughout the whole network, *i.e.*, scarcely more than one megabyte per user. Recall however that, as discussed in the Introduction, true positives result in a significant benefit for the operator, *i.e.*, a reduction of the peak load on its infrastructure, while false positives have mild consequences for both the operator and the users.

5.3.5 Results

In Fig. 7, we present results for the Facebook (left) and Email (right) graphs. Although the two networks are very different in size and structure, they exhibit the same qualitative behavior, with a clear cascade evolution. The way Proactive Seeding works is easy to observe: the users known (or assumed) to request the content during the peak time are served during earlier frames, thus reducing the peak load.

For both networks, we compare the ideal (*i.e.*, offline) performance with the adaptive (*i.e.*, online) case, in which the demand is not known a priori. In the latter, we consider three values of the noise ν. If our prediction is not systematically biased ($\nu = 1$), the online performance of Proactive Seeding is close to the optimal (offline). In contrast, systematically overestimating ($\nu > 1$) or underestimating ($\nu < 1$) the future demand leads to less gain γ, but with qualitatively different effects. *Overestimating* the demand means serving users that will never need the content, thus wasting network and user resources. In the extreme case, it may even lead to a negative gain, *i.e.*, a peak load $\max_k |\boldsymbol{h}_c^k|$ greater than the peak demand $\max_k |\boldsymbol{w}_c^k|$. On the other hand, *underestimating* the demand is conservative, as moves towards the no-seeding case. The gain γ can decrease, but is still above zero. Therefore, as a practical take-away from our online evaluation, we can recommend to tune the prediction parameters so as to underestimate rather than overestimation the demand.

Fig. 7 also allows us to see how the adaptiveness, *i.e.*, the fact that at each time frame k we feed the actual

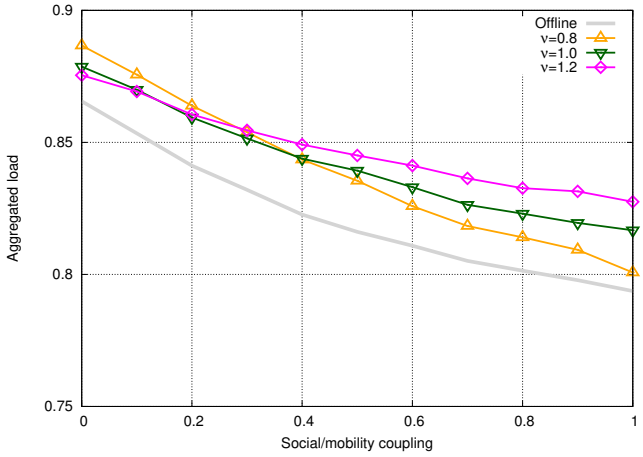

Fig. 8. Aggregated load under Proactive Seeding with D2D, normalized by pure Proactive Seeding load, for different values of the social/mobility coupling f (*i.e.*, the link-wise overlap between the friendship and D2D graphs).

set $\boldsymbol{W}_c^k$ of users interested in the content back to the prediction algorithm, allows us to recover from prediction errors. If $\nu > 1$, we tend to overestimate the number of users interested in the content at the begin of the cascade. However, as we observe the actual number of interested users, we are able to correct the error, and schedule fewer users in the subsequent frames. Conversely, if $\nu < 1$, we start seeding fewer users than we should, and we make it up for this error later. Notice however that both such cases imply a peak load that is higher than the ideal (*i.e.*, offline) one.

5.4 On the coupling of friendship and mobility

In Sec. 5.2, the 3G users where randomly mapped to the D2D connectivity traces. However, in practice the D2D connectivity is naturally correlated with the OSN friendship graph [27], which may potentially help exploit the D2D links.

We study this effect in the Facebook, as follows. First, we generate a random D2D connectivity graph. Then, we map the Facebook users to the D2D users such that a specific fraction f of links in the two graphs coincide. Finally, we apply Proactive Seeding with D2D, and study the resulting aggregated load.

Fig. 8 shows that the aggregated load indeed decreases with f. However, this effect is rather limited (note the scale on y-axis), which means that the performance of Proactive Seeding does not rely on high correlation between D2D and OSN.

It is worth stressing that the results summarized in Fig. 8 have general validity, and do not depend on a specific mobility trace.

6 Related work

Proactive Seeding is related to three broad research areas: offloading of cellular networks; traffic shaping; social-supported wireless networking; and influence spreading in social networks. They are discussed in separate subsections below.

6.1 Cellular network offloading

The core idea of works such as [11,12,22] is that when several users need to get the same content (*e.g.*, a file), a portion them may get it through opportunistic (*i.e.*, device-to-device) contacts instead of a cellular connection.

[11] targets the problem of propagating content updates. The objective is to optimize a user-satisfaction metric, linked to the delay affecting such updates, while ensuring system scalability. This is achieved by combining device-to-device and cellular connections, using social information to decide which users should act as relays.

The work in [22] sets in a similar scenario, and presents a set of algorithms exploiting device-to-device connectivity to offload the cellular (3G in that case) network. The authors assume that the social relationships among the users (*e.g.*, friendship or common interest) are strongly correlated with both their mobility and the contents they are going to request. [12] addresses another aspect of the same problem, *i.e.*, offloading the cellular network through vehicular connections, while still meeting (comparatively) strict delivery deadlines. Based on the feedbacks received from the users, a central authority adaptively decides (i) how many more copies of the content shall be injected in the network and (ii) which vehicles are most suitable (*e.g.*, due to their connectivity) to receive it.

6.2 Traffic shaping and dynamic pricing

When the traffic load threatens to exceed the network capacity, there are two main options: increase network capacity, or decrease the peak load. The latter approach is commonly known as *traffic shaping*.

The most straightforward way of performing traffic shaping is to refuse service to some QoS classes if the network is overloaded, as in [30,31]. More recent works, e.g., [10], exploit social networks: the authors of [10] envision to delay the spread of interest over social networks, effectively reducing the peak bandwidth demand.

Another approach is represented by *dynamic pricing*: no requests are denied, but the same request can be subject to different charges at different times [13] – similar to what happens with cars in congested cities. More closely related to our work, [14], assumes that users will accept some delay in their download in exchange of such benefits as a reduced rate. The same concept, i.e., that the value (hence the price) of bandwidth change over time, is applied to heterogeneous networks based on dynamic spectrum access [32].

Our approach is in a way similar to traffic shaping: traditional traffic shaping approaches move some load

forward in time, while Proactive Seeding moves it back in time – and without the explicit cooperation of users. Indeed, in our vision users are *impatient*: they have grown accustomed to obtain any content they need, anytime, anywhere, and we deem such a behavior unlikely to change.

6.3 Social-supported wireless networking

Many works [33]–[37] exploit the principle that interest and (in a wide sense) "social" links drive the mobility of pedestrian and vehicular users. Knowing such links, it is possible to predict the users' mobility, or at least some of its key features (*e.g.,* encounter frequency or community structure), and thus optimize routing and content delivery.

[33] presents a routing protocol for DTNs, based on the concepts of *community* and *centrality*. User terminals detect the community their owner belongs to, and the centrality she has inside it. This information is subsequently propagated and used to identify the best next hop.

Other works [35] use social information to optimize content discovery in a publish/subscribe setting. More "social" users are elected as brokers, or otherwise given a special role in the delivery process.

Finally, the works [34,36,37] exploit social information to route the queries, as well as to decide which information items are most relevant and should thus be cached or duplicated.

6.4 Influence spreading on social networks

The correlation between personal relationships and individual behaviors has been studied long before the Internet. More recent works such as [8] propose a set of graph-based models to study this phenomenon. Directly connected to online social networks, [7] proposes a greedy algorithm to maximize the spread of influence, given the social structure. It also discusses and reviews several influence models.

Other than modeling, several more works aim at analyzing the spread of influence on actual social networks. [38] identifies and studies several cascades on the Flickr social network, and assesses to which extent friendship drives them. Related to this, [24] shows that there is a significant correlation between interest and geographical proximity, for four different OSNs (BrightKite, FourSquare, LiveJournal and Twitter), offering different amount of location information. [39] analyzes 1.5 million YouTube videos, showing that not all popular videos are "social", and that highly social videos rise to, and fall from, their peak popularity more quickly than less social videos. It also studies the referrals coming from other social networks, with the Twitter referrals resulting in a much higher relative growth than the Facebook ones. Other works [40,41] take a networking viewpoint and investigate the specific features of OSN-driven traffic.

More closely related to our work is [10], which sets in a scenario in which social network-triggered bandwidth demand peaks can overload a cellular network. The authors assume that the mobile operator has the possibility to delay the spread of the interest by (temporarily) inhibiting some social links, and present a way to optimally choose such links.

Forecasting demand and popularity of contents is another active research area. [42] presents methods to predict the popularity of specific content items given the history of such content access (thus without modeling the influence spreading over a social graph), for the YouTube and Digg social networks. [9] collects a dataset of 22M tweets, containing 15M URLs. The authors also present a methodology (based on influence spreading models) which is able to predict more than half of the tweets in the dataset with only 15% false positive rate.

6.5 Mobility traces and models

The effectiveness of many solutions combining OSNs and wireless networking depends on the extent to which social relationships influence the mobility of users. To study this important aspect, [43] logged the device-to-device contacts of 78 users participating to the INFOCOM 2006 conference, complementing such information with additional data about each user's interests and nationality.

The trace [25] includes call, contact and cellular BS association logs for one hundred people, over a period of nine months. Based on the high-level features observed in [25], the authors of [44] propose a mobility model accounting for the presence of both "friend" and "stranger" nodes.

Among the many traces hosted in the Crawdad repository, particular relevant to us is the Dartmouth/Campus one [26]. It includes association information for over 25,000 users in a period of eleven weeks. This makes it possible to perform a one-by-one association between users in our social cascades and users in the trace, as explained in Sec. 5.2.1.

7 Discussion and Conclusion

We addressed the phenomenon of *flash-crowds*, *i.e.*, groups of mobile users suddenly becoming interested (through a social network) in the same content. This represents a challenge for cellular network, as they must be provisioned in order to meet such abrupt demand peaks. With our Proactive Seeding technique, we leveraged the fact that interest spreading can be modeled and predicted with remarkable accuracy, and framed our problem as the minimization of the peak network load, given knowledge of the future demand and subject to user delay constraints.

In the special case of single content with no background load, the optimal solution has an intuitive interpretation. In the special case of multiple contents with

known background traffic, we provide a greedy algorithm and prove its optimality, in the offline case. In the online case, we evaluated our algorithms by replacing the actual future demand by the predicted demand; we found that they are robust, especially when conservatively underestimating the demand. We also extended our algorithm to take into account opportunistic mobile-to-mobile communication, thus offloading cellular traffic and further reducing the peak load up to 50%. We further investigated the impact of the correlation between mobility and social links on the performance of our proposal, and found that Proactive Seeding is effective even when this correlation is weak.

ACKNOWLEDGMENTS

We would like to thank the authors of [9] and [20] for providing the Twitter and 3G traffic traces, respectively.

This work has been partially supported NSF Award 1028394, as well as by the following grants: DOCOMO contract DCL-49538, EU FP7 project FIGARO (grant no. 258378), Swiss SNF grant PBELP2-130871, NSF CDI award #1028394, AFOSR award FA9550-10-1-0310.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019," 2014.

[2] Credit Suisse, "U.S. wireless networks running at 80% of capacity," http://www.fiercewireless.com/story/credit-suisse-report-us-wireless-networks-running-80-total-capacity/2011-07-18?utm_medium=nl&utm_source=internal, 2011.

[3] S. Landström, A. Furuskär, K. Johansson, L. Falconetti, and F. Kronestedt, "Heterogeneous networks–increasing cellular capacity," *The data boom: opportunities and challenges*, p. 4, 2011.

[4] S. Stefania, T. Issam, and B. Matthew, "Lte-the umts long term evolution: from theory to practice," *A John Wiley and Sons, Ltd*, vol. 6, pp. 136–144, 2009.

[5] Facebook, "Facebook statistics," http://www.facebook.com/press/info.php?statistics, 2011.

[6] J. Constine, "40% Of YouTube Traffic Now Mobile, Up From 25% In 2012, 6% In 2011," http://techcrunch.com/2013/10/17/youtube-goes-mobile/, 2013.

[7] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *ACM SIGKDD*, 2003.

[8] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, no. 3, pp. 211–223, 2001.

[9] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs," in *WOSN*, 2010.

[10] H. Sharara, C. Westphal, S. Radosavac, and U. C. Kozat, "Utilizing Social Influence in Content Distribution Networks," in *IEEE INFOCOM*, 2011.

[11] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and Scalable Distribution of Content Updates over a Mobile Social Network," in *IEEE INFOCOM*, 2009.

[12] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. De Amorim, "Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays," 2010.

[13] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Pricing data: A look at past proposals, current plans, and future trends," *CoRR, abs/1201.4197*, 2012.

[14] S. Sengupta, S. Anand, M. Chatterjee, and R. Chandramouli, "Dynamic pricing for service provisioning and network selection in heterogeneous networks," *Physical Communication*, vol. 2, no. 1, pp. 138–150, 2009.

[15] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Transactionsctions on Mobile Computing on Mobile Computing*, vol. 5, no. 1, pp. 77–89, 2006.

[16] H. Gupta and S. Das, "Benefit-Based Data Caching in Ad Hoc Networks," *IEEE Transactionsctions on Mobile Computing*, vol. 7, no. 3, pp. 289–304, 2008.

[17] "Cisco Use Case: Sponsored Data," 2015.

[18] C. Joe-Wong, S. Ha, and M. Chiang, "Sponsoring mobile data: An economic analysis of the impact on users and content providers," 2015.

[19] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. Kozat, "Proactive Seeding for Information Cascades in Cellular Networks," in *IEEE INFOCOM*, 2012.

[20] M. Shafiq, L. Ji, and A. Liu, "Characterizing and modeling internet traffic dynamics of cellular devices," in *ACM SIGMETRICS*, 2011.

[21] 3GPP Work Item Description TSG-RAN, "Study on LTE device to device discovery and communication," RP-110707 LTE-D2D RAN, 2011.

[22] B. Han, P. Hui, V. Kumar, M. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: a case study," in *ACM CHANTS*, 2010.

[23] Qualcomm, "Flashlinq," http://www.qualcomm.com/stories/2011/02/08/ phone-conversations-minus-people, 2011.

[24] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora, "Distance matters: Geo-social metrics for online social networks," in *WOSN*, 2010.

[25] N. Eagle and A. (Sandy) Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2005.

[26] D. Kotz and K. Essien, "Analysis of a campus-wide wireless network," http://snap.stanford.edu/data/email-Enron.html, 2002.

[27] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi, "On the evolution of user interaction in facebook," in *WOSN*, Barcelona, Spain, 2009.

[28] S. Project, "Enron email trace," 2004.

[29] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 2008, pp. 229–238.

[30] J. Elias, F. Martignon, A. Capone, and G. Pujolle, "A new approach to dynamic bandwidth allocation in quality of service networks: Performance and bounds," *Computer Networks*, vol. 51, no. 10, pp. 2833–2853, 2007.

[31] L. Georgiadis, R. Guérin, V. Peris, and K. N. Sivarajan, "Efficient network qos provisioning based on per node traffic shaping," *IEEE/ACM Trans. Netw.*, vol. 4, no. 4, pp. 482–501, Aug. 1996. [Online]. Available: http://dx.doi.org/10.1109/90.532860

[32] S. Dixit, S. Periyalwar, and H. Yanikomeroglu, "Secondary user access in lte architecture based on a base-station-centric framework with dynamic pricing," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 1, pp. 284–296, Jan 2013.

[33] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," *IEEE Transactionsctions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1374652

[34] C. Boldrini, M. Conti, and A. Passarella, "ContentPlace: social-aware data dissemination in opportunistic networks," in *ACM MSWiM*, 2008.

[35] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft, "A socio-aware overlay for publish/subscribe communication in delay tolerant networks," in *ACM MSWiM*, 2007.

[36] E. Jaho and I. Stavrakakis, "Joint interest- and locality-aware content dissemination in social networks," in *IEEE/IFIP WONS*, 2009.

[37] A. J. Mashhadi, S. B. Mokhtar, and L. Capra, "Habit: Leveraging human mobility and social network for efficient content dissemination in Delay Tolerant Networks," in *IEEE WoWMoM*, 2009.

[38] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *WWW*, New York, New York, USA, 2009.

[39] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, "Catching a Viral Video," in *IEEE ICDM Workshops*, 2010.

[40] A. Nazir, S. Raza, D. Gupta, C.-N. Chuah, and B. Krishnamurthy, "Network level footprints of facebook applications," in *ACM IMC*, New York, New York, USA, 2009.

[41] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding online social network usage from a network perspective," in *ACM IMC*, 2009.

[42] G. Szabó and B. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2008.
[43] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "CRAWDAD data set cambridge/haggle (v. 2009-05-29)," Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, 2009.
[44] A. Miklas, K. Gollu, K. Chan, S. Saroiu, K. Gummadi, and E. De Lara, "Exploiting social interactions in mobile systems," in *ACM UbiComp*, 2007.

Francesco Malandrino earned his Ph.D. in 2011 from Politecnico di Torino, where he is currently a post-doctoral fellow. His interests focus on wireless and cellular networks, optimization and infrastructure management.

Maciej Kurant got his Ph.D. in 2009 EPFL Lausanne. He held short-term appointments at the University of California, Irvine, and ETH, Zurich. Since 2012, he is at Google. His main interests are graph sampling and complex networks.

Athina Markopoulou (S98-M02-SM13) received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens, Greece, in 1996, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University in 1998 and 2003, respectively. She is currently an Associate Professor in the EECS Department, at the University of California, Irvine. She has held short-term/visiting positions at SprintLabs (2003), Arista Networks (2005) and IT University of Copenhagen (2013). Her research interests are in the broad area of computer networks and include network measurement and modeling, mobile and online social networks, network security and privacy, and network coding. She received the Henry Samueli School of Engineering Faculty Midcareer Award for Research (2014) and the NSF CAREER Award (2008). She has been an Associate Editor for IEEE/ACM Transactions on Networking (2013-2015).

Cedric Westphal is currently a Principal Architect with Huawei Innovations, and an an adjunct assistant professor with the University of California, Santa Cruz. Prior to that, he worked at DOCOMO Labs and Nokia. His main interests are Content-Centric Networking infrastructure, and SDN/OpenFlow.

Ulas C. Kozat received his Ph.D. in Electrical and Computer Engineering from the University of Maryland, College Park. He is an adjunct associate professor at Ozyegin University, Istanbul, Turkey. He has worked as project manager and principal researcher at DOCOMO Innovations (formerly DOCOMO USA Labs). Prior to that he was a research assistant at Institute for Systems Research in University of Maryland. He also worked at Telcordia Technologies Applied Research and HRL Laboratories during his graduate studies. He has conducted research in the broad areas of wireless communications and computer/communication networks.