Exploiting Double Opportunities for Deadline Based Content Propagation in Wireless Networks

Han Cai, Irem Koprulu, and Ness B. Shroff

***Abstract*—In this paper, we focus on mobile wireless networks comprising of a powerful communication center and a multitude of mobile users. We investigate the propagation of deadline-based content in the wireless network characterized by heterogeneous (time-varying and user-dependent) wireless channel conditions, heterogeneous user mobility, and where communication could occur in a hybrid format (e.g., directly from the central controller or by exchange with other mobiles in a peer-to-peer manner). We show that exploiting double opportunities, i.e., both time-varying channel conditions and mobility, can result in substantial performance gains. We develop a class of double opportunistic multicast schedulers and prove their optimality in terms of both utility and fairness under heterogeneous channel conditions and user mobility. Extensive simulation results are provided to demonstrate that these algorithms can not only substantially boost the throughput of all users (e.g., by 50% to 150%), but also achieve different consideration of fairness among individual users and groups of users.**

I. Introduction

The last few years have witnessed an enormous growth in the popularity and capabilities of handheld devices such as smartphones, tablets, and laptops. These devices have in turn fueled mobile content sharing applications, which are becoming increasingly popular. However, these devices and the traffic that they generate have put a significant strain on many of today's cellular networks. For example, in June 2010, AT&T had to phase out its unlimited data plans for smartphones in lieu of "metered" data plans with limits on monthly bandwidth. In the same month, iPhone 4 was launched in the U.S. with many Wi-Fi only applications (e.g., FaceTime video calling), that users cannot access over 3G.

In this paper, we focus on wireless networks that comprise of a powerful Communication Center (CC), e.g., a base station, and many mobile users with communication and computation capacity, e.g., pedestrians/soldiers carrying smartphones or tablets, smart robots/sensors, etc. These networks could communicate in a *hybrid* format (e.g., mobiles communicating directly with the CC or with other mobiles in a peer-to-peer manner), could have *heterogeneous* (time-varying and user-dependent) wireless *channel conditions*, and *heterogeneous user mobility*. Examples of such networks are cellular networks, military networks, mobile sensor networks with CC(s), etc. The ever-growing wireless user density will lead to increasing proliferation of these networks, but at the same time generate bandwidth-intensive traffic, which means that appropriate resource allocation mechanisms that exploit all available opportunities, will be critical to the efficient usage and successful deployment of these systems.

The recent unprecedented increase in the density of *mobile* users gives rise to an abundance of "contact" opportunities, i.e., opportunities where mobile users are in close enough proximity of each other to communicate with each other. As a result, content sharing through such contacts may occur at a similar time scale as that through a service provider.

Traditionally, downlink scheduling, mobility, and content distribution have been extensively studied, but often *in isolation*. For example, there have been many studies on the unicast or multicast scheduling problem in cellular networks (to cite, but a few, [1], [2], [4], [21], [24], [30]). These works have not exploited the random mobility of users. Similarly, there is a rich literature on the design and performance analysis of forwarding algorithms by exploiting the opportunistic mobility patterns of mobile users in the system (see [7], [8], [11], [12], [28], [29], among others). These works in mobile ad-hoc networks, as well as a number of recent works on content distribution (e.g. [9], [13], [15], [19]), do not consider the wireless channel's inherent variability. In contrast to the existing literature, in this work we will explicitly consider both time-varying channel conditions and users' random mobility.

To fully realize the performance gains in content distribution by jointly investigating mobility and scheduling, we first develop a class of *doubly opportunistic multicast* algorithms with *heterogeneous* (time varying and user-dependent) wireless channel condition and *homogeneous* contact rates among mobile users. We then extend our class of doubly opportunistic algorithms to scenarios with *heterogeneous* contact rates among mobile users. We refer to such strategies as being *doubly opportunistic* due to two important factors exploited in the design: the time-varying wireless channel conditions and the random contact events among mobile users. These two interacting factors make our study extremely challenging.

We prove that these algorithms achieve a class of Group Proportional Fairness (GPF) criteria, which characterize different fairness considerations among individual users' or user groups' throughput, or equivalently, different intra-group or inter-group tradeoffs. The GPF principles, to be defined in Section II, incorporate many well-know fairness principles such as proportional fairness [17], [18] and max-min fairness [3] as special examples. More importantly, this rich set of fairness principles provides a powerful measure for

Irem Koprulu is with the Department of ECE at The Ohio State University (e-mail: irem.koprulu@gmail.com). Ness B. Shroff holds a joint appointment in the Departments of ECE and CSE at The Ohio State University (e-mail: shroff@ece.osu.edu).

This work has been supported in part by the Army Research Office MURI Award W911NF-08-1-0238, and NSF grants CNS-1065136 and CNS-1012700.

two-layer (user-and-group) views of fairness. In particular, different fairness considerations among individual users (user groups) can be achieved by simply adjusting the *user (group) fairness parameter* we define in the GPF criteria. We conclude our work with numerical simulation results to confirm that the proposed algorithms significantly improve system performance in terms of both throughput and fairness.

II. System Model and Problem Statement

We consider a downlink multicast scenario where a single base station (BS) is broadcasting independent streams of deadline-constrained content to different groups of mobile users. A *group* consists of all users who are interested in receiving the *same* content. For simplicity, we assume that each user belongs to a unique group. Using N to denote the number of groups, and S_n to denote the number of users in group n, we let $u_{n,m}$ ($n = 1, \cdots, N$, $m = 1, \cdots, S_n$) represent the m^{th} user in group n. In addition to communicating with the base station, users in a group can communicate among themselves, and exchange content whenever they come within the communication range of each other. Our objective in this paper is to exploit the *double opportunities* provided by the time-varying channel conditions and mobility of users in order to maximize the amount of content received by users while providing a *fair* distribution of the downlink resources among users and groups.

A. Channel Dynamics: Due to mobility and small-scale fading, each user has time-varying channel conditions. We consider a time-slotted communication system where users' channel conditions remain the same over one time slot. We choose our unit of time as the length of a time slot.

Due to practical limitations, we assume that the BS is capable of broadcasting at a discrete set of rates $\{R_i\}_{i=1}^{K}$ with $0 < R_1 < \ldots < R_K$. Depending on its channel condition, at time t user $u_{n,m}$ can achieve a *maximum achievable data rate* of $r_{n,m}(t) \in \{0, R_1, \ldots, R_K\}$. We assume that at the start of the t^{th} time slot, the BS knows the channel condition and hence $r_{n,m}(t)$ of each user. We make the following mild assumption on $r_{n,m}(t)$.

Assumption 1. *Each user $u_{n,m}$ has stationary and ergodic channel conditions, in particular, the maximum achievable data rate vector $\vec{r}(t) \triangleq \{r_{n,m}\}_{n=1,\cdots,N}^{m=1,\cdots,S_n}$ is stationary and ergodic.* □

Note that Assumption 1 is quite general, and allows for both spatial and temporal correlation of $\vec{r}(t)$, as well as heterogeneity among users' channels (e.g., some users may always have better channel conditions than others).

At each time slot t, the BS chooses (i) a group index $n(t)$, and (ii) a transmission rate $r^g_{n(t)}(t) \in \{R_1, \ldots, R_K\}$. If a group n is not chosen for transmission at time t we set $r^g_n(t) = 0$. We assume that at the t^{th} time slot, if the BS chooses to broadcast to group $n(t)$ at rate $r^g_{n(t)}(t)$, then all users $u_{n(t),m}$ that satisfy $r_{n(t),m}(t) \geq r^g_{n(t)}(t)$ can receive and decode the data correctly. After a user $u_{n,m}$ receives data from the BS, it can propagate unexpired data to other users in the *same*[1] group through contact events.

B. Content Lifetime Constraints: We assume that the content of group n (also called *content type* n) expires after $L_n \in (0, \infty)$ units of time. The lifetime L_n of a packet depends primarily on the content's degree of tolerance to delay. But it can also be utilized to achieve different tradeoffs between throughput and delay, or to control the level of content flooding in the network. For simplicity, we consider the case where each content type has the same lifetime, i.e. $L_n = L$ for all n. However, all results in this paper, can be readily extended to the case where L_n are different for different content types.

C. Contact Process Dynamics: A *contact event* between a pair of users occurs when the two users are close enough to communicate and exchange content with each other. We use d to represent the communication range of any two users (e.g., for bluetooth devices, $d \approx$ 10m). If we let $x_{n,m}(t)$ denote the location of user $u_{n,m}$ at (continuous) time t, we say that one contact event between u_{n_1,m_1} and u_{n_2,m_2} occurs during $[t_0, t_1)$ if $\|x_{n_1,m_1}(t_0^-) - x_{n_2,m_2}(t_0^-)\| > d$, $\|x_{n_1,m_1}(t) - x_{n_2,m_2}(t)\| \leq d$ for all $t \in [t_0, t_1)$, and $\|x_{n_1,m_1}(t_1) - x_{n_2,m_2}(t_1)\| > d$. The number of contact events between a pair of users that have occurred up to time t is a counting process called the *contact process*. We will refer to the time between the start of two consecutive contact events between the same pair of users as the *inter-contact time*. For a stationary contact process, the reciprocal of the average inter-contact time is the *contact rate*.

We assume that the length of a contact event's duration is negligible compared to the inter-contact time. This is a reasonable assumption, since the ratio between the average inter-contact time and the average duration of a contact event is approximately the ratio between the area of the mobile domain (the cell) and a single user's communication area (πd^2) [14]. For a cell of radius 500m and a peer-to-peer communication range of $d \approx$ 10m, this ratio would be greater than 6×10^3.

Obtaining complete knowledge of the contact processes can be extremely difficult, and could consume enormous amounts of uplink resources. Also, mathematically characterizing the network performance is intractable for arbitrary contact processes. Thus, we adopt the following assumption for our analytical characterization, but we will allow more general models in the simulations.

Assumption 2. *The contact process between a pair of users is a Poisson process.* □

Poisson contact processes have been shown to be a good approximation [7], [11] under the well-known *i.i.d.* mobility model [20] and Random Waypoint (RWP) mobility model [6]. The RWP model has often been used in protocol design and performance analysis/comparison in mobile ad-hoc networks.

[1] Allowing packet forwarding in *different* groups can further speed up the propagation. However, this raises up additional concerns, e.g., the users' willingness of forwarding copies not in their interest by expending extra energy, and is beyond the scope of this paper.

Our final assumption concerns the nature of the peer-to-peer communication between pairs of users.

Assumption 3. *During a contact event, a pair of users in the same group can exchange all the unexpired content copies, which are absent from each other's list.* □

D. Set of Feasible Schedulers: Recall that, at the t^{th} time slot, a scheduler S chooses a group index $n(t) \in \{1, \cdots, N\}$ and a transmission rate $r^g_{n(t)}(t) \in \{R_1, \cdots, R_K\}$. Before we can define the set of feasible schedulers we need to clarify what we mean by throughput. We define user $u_{n,m}$'s throughput $T^S_{n,m}(t)$ at time slot t under the scheduler S as the running average of the information received by user $u_{n,m}$ until time t either directly through the BS or through contact with peers. Since we are considering a time-slotted communication system and continuous time contact processes, we choose our unit of time as a slot length. Mathematically,

$$T^S_{n,m}(t) \triangleq \frac{1}{t}\sum_{k=1}^{t} r^g_n(k) \sum_{v\in[0,\min\{L,t-k\}]} 1_{\mathcal{E}^S_{n,m,k,k+v}}, \quad (1)$$

where $\mathcal{E}^S_{n,m,k,k+v}$ represents the event that at time $k+v$, user $u_{n,m}$ receives a copy of the content initially broadcast at time k under scheduler S. Note that this event covers both the case of user $u_{n,m}$ receiving the content directly from the BS ($v = 0$) and the case of user $u_{n,m}$ receiving the content from a peer ($0 < v \leq L$). Hence, this event captures the effect of channel dynamics (i.e. $r_{n,m}(t) \geq r^g_{n(t)}(t)$ for successful reception from the BS), content lifetime constraints, and contact process dynamics (i.e. there is a contact between user $u_{n,m}$ and another a user $u_{n,m'}$ carrying a copy of the content before the content expires). We assume that at the start of the t^{th} time slot, the BS knows each user's throughput $T^S_{n,m}(t-1)$ at time $t-1$. We define user $u_{n,m}$'s *long-term throughput* under the scheduler S as

$$\tau^S_{n,m} \triangleq \lim_{t\to\infty} T^S_{n,m}(t). \quad (2)$$

In this work, we consider the set of *feasible schedulers* $\mathcal{S}$ for which this limit exists. This class covers a large range of schedulers including the class of stationary schedulers [22].

E. Class of Group Utility Functions: As in any opportunistic multicast scenario the BS needs to ensure that: (i) the users get as much of their subscribed content either directly from the BS or through contact with peers, and (ii) the downlink resource is shared in a 'fair' way. We adopt sets of group utility functions $\{U^g_n(\cdot)\}_{n=1,\cdots,N}$ and user utility functions $\{U^u_{n,m}(\cdot)\}^{m=1,\cdots,S_n}_{n=1,\cdots,N}$ to characterize fairness among groups and individual users.

We require that $U^g_n(\cdot)$ and $U^u_{n,m}(\cdot)$ are non-decreasing functions defined on $(0,\infty)$. Different choices for the utility functions and their arguments cover a wide range of fairness principles proposed in the literature (e.g. [10], [16], [17], [23], [25], [26], [27]).

The so-called $(\vec{w},\alpha)$ proportional fairness principle [24] among a set of individual users has been widely used in the study of transmission control protocols, unicast scheduling algorithms, etc. However, more general fairness principles need to be developed for a multicast scheduler to characterize fairness among both groups and users, which we do next.

For any sets of non-negative parameters $\{w_n\}_{n=1,\cdots,N}$, $\{v_{n,m}\}^{m=1,\cdots,S_n}_{n=1,\cdots,N}$, α, and β, we define the group utility functions U^g_n and user utility functions $U^u_{n,m}$ as follows

$$U^g_n(y) \triangleq \begin{cases} w_n \frac{y^{1-\alpha}}{1-\alpha}, & \alpha \geq 0,\ \alpha \neq 1 \\ w_n \log(y) & \alpha = 1, \end{cases} \quad (3)$$

and

$$U^u_{n,m}(y) \triangleq \begin{cases} v_{n,m} \frac{y^{1-\beta}}{1-\beta}, & \beta \geq 0,\ \beta \neq 1 \\ v_{n,m} \log(y) & \beta = 1. \end{cases} \quad (4)$$

We say that a scheduler that maximizes

$$\sum_n U^g_n\big(\sum_m U^u_{n,m}(\tau_{n,m})\big) \quad (5)$$

achieves the $(\vec{w},\vec{v},\alpha,\beta)$ *group proportional fairness criterion*, where $\vec{w} = \{w_n\}_n$, $\vec{v} = \{v_{n,m}\}_{n,m}$, and $\tau_{n,m}$ represents user $u_{n,m}$'s throughput. We call α and β the *group* and *user fairness parameters*, respectively. When $\alpha = 0$ and $w_n = 1$ for all n, the optimal scheduler solving (5) achieves $(\vec{v},\beta)$ proportional fairness among *individual users*. Similarly, when $\beta = 0$ and $v_{n,m} = 1$ for all n and m, then the solution of (5) achieves $(\vec{w},\alpha)$ proportional fairness among *groups*. When both $\alpha = \beta = 0$, $w_n = 1$, and $v_{n,m} = 1$ for all n and m, (5) reduces to the objective of the so called MAX scheduler, which maximizes the aggregate throughput of all users in the system. We show in Section V how these parameters can be adjusted to control the fairness among groups and users.

F. Problem Statement: Given the descriptions of the channel, content lifetime, and contact process dynamics, we are ready to formulate our double opportunistic scheduling problem.

Double Opportunistic Problem (DOP):

$$\max_{S\in\mathcal{S}} \quad \sum_{n=1}^{N} U^g_n\big(\sum_{m=1}^{S_n} U^u_{n,m}(\tau^S_{n,m})\big) \quad (6)$$

$$\text{s.t.} \quad \tau^S_{n,m} = \lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t} r^g_n(k) \sum_{v\in[0,\min\{L,t-k\}]} 1_{\mathcal{E}^S_{n,m,k,k+v}}.$$

Note that the throughput expression in the constraint accounts for the channel dynamics, content lifetime, and contact process dynamics. The solution to this problem allows for the joint exploitation of both the *channel conditions* and *mobility* to obtain significant performance gains for content distribution. In both scenarios, we allow the channel conditions to be statistically heterogeneous across users.

We will first solve this problem under the assumption of statistically homogeneous user mobility in Section III, and then discuss its extension to the heterogeneous scenario in Section IV.

III. Double Opportunistic Multicast Scheduling under Homogeneous Poisson Contact Processes

In this section, we develop a class of mobility-aware multicast scheduling algorithms that are provably optimal and satisfy the GPF criterion for the case of *homogeneous* Poisson contact processes, where the contact rates for all pairs of users are all equal to λ.[2] This allows us to introduce the optimal algorithm that is extendable to the heterogeneous Poisson contact processes scenario (cf. Section IV), but without the cumbersome notation necessary to deal with the heterogeneity.

We start by characterizing the amount of data received by the mobile users, either directly from the BS or indirectly through mobile peers, as a function of the broadcast rate and the contact process dynamics. To that end, we first define

$$\kappa_n(t,y) \triangleq \sum_{m=1}^{S_n} 1_{\{y \leq r_{n,m}(t)\}}, \tag{7}$$

which gives the number of users in group n receiving content in slot t *directly* from the BS when it broadcasts to group n at rate y. The following lemma uses this information together with the contact process characteristics to express the average number of users that receive the content *directly or indirectly* within its lifetime.

Lemma 1. *Setting $N_0 = \kappa_n(t,y)$, define two $1 \times (S_n - N_0 + 1)$ vectors*

$$\overrightarrow{N}_1 \triangleq [N_0, N_0+1, ...S_n], \quad \overrightarrow{N}_2 \triangleq [1, 0, \ldots, 0], \tag{8}$$

and an $(S_n - N_0 + 1) \times (S_n - N_0 + 1)$ generator matrix $\mathbf{A} = \{a_{i,j}\}$, where

$$a_{i,j} = \begin{cases} (N_0+i-1)(S_n - N_0 - i + 1) & \text{if } i = j, \\ -(N_0+i-2)(S_n - N_0 - i + 2) & \text{if } i = j+1, \\ 0 & \text{otherwise}. \end{cases} \tag{9}$$

Then, given that at time t the BS broadcasts to group n at rate y, the expected number of users in group n who will have a copy of the content by the time it expires is given by

$$\chi_n(t,y) \triangleq \overrightarrow{N}_1 e^{-\lambda \mathbf{A} L} \overrightarrow{N}_2^T. \tag{10}$$

Proof: See Appendix C. ■

Before we provide the optimal scheduler S^*, we need to define a few auxiliary functions that facilitate its description. We define the *aggregate user utility* of group n at time t as the aggregate utility of all users belonging to that group, i.e.

$$G_n(t) \triangleq \sum_{m=1}^{S_n} U^u_{n,m}\left(T_{n,m}(t)\right). \tag{11}$$

[2]The derived scheduler does not require that the contact rate is the same for all users in the system, only within a group. Replacing the system wide contact rate λ with contact rate λ_n for group n covers the latter case. We chose not to do so to simplify exposition.

Also, for group n, time slot t and rate y, we define

$$\varphi_n(t,y) \triangleq \sum_{m=1}^{S_n} v_{n,m} \frac{y}{\left(\max\{T_{n,m}(t), \epsilon\}\right)^{\beta}} \Big[1_{\{y \leq r_{n,m}(t)\}} + \frac{\chi_n(t,y) - \kappa_n(t,y)}{S_n - \kappa_n(t,y)} 1_{\{y > r_{n,m}(t)\}}\Big], \tag{12}$$

where $\epsilon \to 0^+$ serves to prevent a division by zero. This is a measure of the marginal increase in the aggregate user utility $G_n(t)$ when in slot t the BS broadcasts to group n at rate y.

$(\vec{w}, \vec{v}, \alpha, \beta)$ ***GPF scheduler*** S^***:**

BS part:

The BS assigns a rate to each group $n = 1, \cdots, N$:

$$r^g_n(t) \in \underset{y \in \{R_1, \cdots, R_K\}}{\arg\max} \ \varphi_n(t,y). \tag{13}$$

The BS chooses group $n(t)$ to broadcast at the previously assigned rate $r^g_{n(t)}(t)$

$$n(t) \in \underset{1 \leq n \leq N}{\arg\max}\ w_n \frac{\varphi_n(t, r^g_n(t))}{\left(\max\{G_n(t), \epsilon\}\right)^{\alpha}}, \tag{14}$$

where ties are broken uniformly at random.

User part:

Whenever any two users of the same group meet each other, they share each other's content such that each will have the union of their sets of unexpired copies after the contact.

The inclusion of the parameter $\epsilon \to 0^+$ in the formulation of the GPF scheduler is to simplify mathematical notation. This parameter can be omitted if we adopt the conventions that $0^0 = 1$ and $1/0 = \infty$. In case there exist several groups attaining infinite value in (14), the scheduler chooses the group that maximizes the numerator of the expression in (14). In what follows, we proceed by dropping ϵ and adopting the above cited conventions.

While optimality of the GPF scheduler is more rigorously proven in the subsequent theorem, let us provide the intuition behind its decision making. It assigns each group n rate $r^g_n(t)$ that maximizes the increase in the aggregate user utility of that group, since $\varphi_n(t,y)$ is the marginal increase in the aggregate user utility $G_n(t)$ when in slot t the BS broadcasts to group n at rate y. Once the scheduler decides the optimal potential rates for each group, it chooses the group that will result in the largest increase in the objective function (6). Given the group utility functions in (3), the expression in the RHS of (14) is the marginal increase in the group utility of group n if the BS broadcasts to that group at the potential rate $r^g_n(t)$. We now present the main result in this section:

Theorem 1. *The above $(\vec{w}, \vec{v}, \alpha, \beta)$ GPF scheduler S^* solves the Double Opportunistic Problem (6) optimally under homogeneous Poisson contact processes.*

Proof: See Appendix A. ■

IV. Double Opportunistic Multicast Scheduling under Heterogeneous Poisson Contact Processes

In the previous section, we assumed homogeneous Poisson contact processes among the set of all users within each group. In this section, we extend our results to include scenarios with heterogeneous Poisson contact processes. We consider a model where each group of users is divided into further *subgroups* with different mobility characteristics, leading to heterogeneous contact behavior. Such a model is well-motivated by real world examples, e.g., a network with both vehicular and pedestrian users. In order to keep notation relatively simple, we consider the case of two subgroups, but the results can be readily extended to an arbitrary number of subgroups.

Let us assume that group n has S_n^1 users in subgroup 1 and S_n^2 users in subgroup 2 with $S_n^1 + S_n^2 = S_n$. Let users within subgroup 1 have contact rate λ_1, users within subgroup 2 have contact rate λ_2, and two users of different subgroups have contact rate λ_{12}. Similar to the homogeneous case, let us define

$$\kappa_n^1(t,y) \triangleq \sum_{\{m:u_{n,m}\in\text{Subgroup 1}\}} 1_{\{y\le r_{n,m}(t)\}},$$
$$\kappa_n^2(t,y) \triangleq \sum_{\{m:u_{n,m}\in\text{Subgroup 2}\}} 1_{\{y\le r_{n,m}(t)\}}, \quad (15)$$

where $\kappa_n^i(t,y)$ represents the number of users in subgroup i ($i = 1, 2$) receiving content *directly* from the BS if the BS broadcasts to group n at rate y. We are now ready to express the average number of users in each subgroup that receive the content either *directly or indirectly*, as in the homogeneous scenario.

Before we describe the generator matrix in this scenario, we need to map the two dimensional state space to one dimension. To that end, let $i : \{0,1,\cdots,S_n^1\} \times \{0,1,\cdots,S_n^2\} \mapsto \{1,2,\cdots,(S_n^1+1)(S_n^2+1)\}$ be an enumeration of all possible states (k_1,k_2). One example of such an enumeration would be $i(k_1,k_2) = k_1(S_n^2+1) + k_2 + 1$. Also, let $f_1, f_2 : \{1,2,\cdots,(S_n^1+1)(S_n^2+1)\} \mapsto \mathbb{N}$ be the inverse mappings such that $f_j(i(k_1,k_2)) = k_j$ ($j = 1,2$). Let $i_0(t,y) = i(\kappa_n^1(t,y), \kappa_n^2(t,y))$ be the sequence number of the initial state $(\kappa_n^1(t,y), \kappa_n^2(t,y))$. Define

$$\overrightarrow{N}_1^1 \triangleq [f_1(1), f_1(2), \cdots, f_1((S_n^1+1)(S_n^2+1))], \quad (16)$$
$$\overrightarrow{N}_1^2 \triangleq [f_2(1), f_2(2), \cdots, f_2((S_n^1+1)(S_n^2+1))], \quad (17)$$
$$\text{and} \quad \overrightarrow{N}_2(t,y) \triangleq \vec{e}_{i_0(t,y)} = [0,\cdots,0,1,0,\cdots,0], \quad (18)$$

where $\vec{e}_{i_0(t,y)}$ denotes the $i_0(t,y)^{th}$ unit vector. Then, we can construct the $(S_n^1+1)(S_n^2+1)$ by $(S_n^1+1)(S_n^2+1)$ generator matrix $\mathbf{A} = \{a_{i,j}\}$ as follows:
For $k_1 = 0,\cdots,S_n^1 - 1$ and $k_2 = 0,\cdots,S_n^2$ let

$$a_{i(k_1,k_2),i(k_1+1,k_2)} = \lambda_1 k_1(S_n^1 - k_1) + \lambda_{12} k_2(S_n^1 - k_1),$$

for $k_1 = 0,\cdots,S_n^1$ and $k_2 = 0,\cdots,S_n^2 - 1$ let

$$a_{i(k_1,k_2),i(k_1,k_2+1)} = \lambda_2 k_2(S_n^2 - k_2) + \lambda_{12} k_1(S_n^2 - k_2),$$

and for all other entries let $a_{i,j} = 0$, for $i \neq j$, and $a_{i,i} = -\sum_j a_{i,j}$ for all i.

Lemma 2. *If at time t the BS broadcasts the n^{th} content at rate y, the average number of users in subgroups 1 and 2 that will have a copy of the content at the end of its lifetime are*

$$\chi_n^1(t,y) \triangleq \overrightarrow{N}_1^1 e^{\mathbf{A}L} \overrightarrow{N}_2^T \text{and} \quad \chi_n^2(t,y) \triangleq \overrightarrow{N}_1^2 e^{\mathbf{A}L} \overrightarrow{N}_2^T, \quad (19)$$

respectively.

Proof: See Appendix D. ■

As for the auxiliary functions, we define the aggregate user utility $G_n(t)$ exactly as in (11). Its marginal increase when in slot t the BS broadcasts to group n at rate y is given by

$$\varphi_n(t,y) \triangleq \sum_{m=1}^{S_n} v_{n,m} \frac{y}{\big(T_{n,m}(t)\big)^\beta} \Big[1_{\{y\le r_{n,m}(t)\}} + \frac{\chi_n^1(t,y) - \kappa_n^1(t,y)}{S_n^1 - \kappa_n^1(t,y)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 1}\}} + \frac{\chi_n^2(t,y) - \kappa_n^2(t,y)}{\big(S_n^2 - \kappa_n^2(t,y)\big)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 2}\}}\Big]. \quad (20)$$

With all these definitions in place, the description of our GPF scheduler S^* for the heterogeneous contact processes scenario remains unmodified except for the use of (20) instead of (12). Also, the optimality of the algorithm continues to hold with minor modifications as shown in the following theorem.

Theorem 2. *The $(\vec{w},\vec{v},\alpha,\beta)$ GPF scheduler S^* using* (20) *for $\varphi_n(t,y)$ solves the Double Opportunistic Problem (6) optimally under the class of heterogeneous Poisson contact processes described above.*

Proof: See Appendix B. ■

V. Simulation Results

In this section, we present simulation results that: (i) validate our theoretical results both under homogeneous and heterogeneous contact processes; (ii) investigate the influence of relaxing the Poisson contact process assumption to more realistic contact processes; (iii) quantitatively compare three main classes of scheduling strategies with varying degrees of opportunistic features and with varying degrees of awareness of user mobility; and (iv) examine the effect of group utility function parameters on the fairness and throughput levels achieved by the schedulers.

Our investigations in this section not only help to quantify the performance improvement achieved by progressively more mobility-cognizant schedulers over the baseline opportunistic one, but also to indicate that the percentage gains achieved by our optimal GPF scheduler (designed under Poisson contact assumptions) are observed under more realistic mobility patterns. Such insensitivity provides a strong promise for the effective use our GPF scheduler under real life conditions.

A. Basic Setup

We consider a square-shaped network area Ω of size $(500\,m)^2$ with a BS located at the center. We examine two

asymmetrically sized groups with 70 and 30 users in order to illustrate the effects of the group fairness parameter α on the tradeoff between fairness and throughput. The channel gains of individual users are composed of two independent components: a slow fading gain determined by the users' distance from the BS (with a power loss exponent of 1.5), and a fast fading gain drawn according to a unit mode Rayleigh distribution independently and identically across users and time slots. We have chosen the downlink rates of the BS following the CDMA2000 1xEV-DO specification as $\{38.4, 76.8, 153.6, 307.2, 614.4, 921.6, 1228.8, 1843.2, 2457.6\}$ kbps. We fix a content lifetime of 180 seconds.

For the user and group utility functions, we set $w_n = 1$ and $v_{n,m} = 1$ for all n, m. Different w_n (resp. $v_{n,m}$) can be interpreted as different prices that each group (resp. user) is willing to pay for a given amount of data. Fixing the unit price of data as such allows us to isolate and illustrate the effect of the group fairness parameter on the fairness and system throughput.

In order to make a fair assessment of the performance gains associated with our GPF scheduler, we compare three different *opportunistic* scheduling strategies, each achieving fairness among groups and users, but with different degrees of opportunistic capabilities:

► **Single Opportunistic** (SO) scheduler, where the BS only takes advantage of varying channel conditions to schedule its transmissions, but there is no peer-to-peer content propagation. Thus, under the SO scheduler, mobility is exploited only indirectly through its effect on channel conditions. This is the current wireless cellular systems.

► **Mobility-Agnostic Double Opportunistic** (MA-DO) scheduler, which corresponds to the special case of our GPF scheduler with $\lambda = 0$. Accordingly, under the MA-DO scheduler, not only does the BS exploit the channel variations (as in SO) but also the users exploit mobility through peer-to-peer content propagation. However, since $\lambda = 0$, the scheduler has no knowledge of the contact processes (hence the name mobility-agnostic), and does not incorporate the future effect of mobility in its decision making.

► **Double Opportunistic** (DO) scheduler, which the same as our GPF scheduler with knowledge of the actual contact rate λ. We refer to the GPF scheduler with this new name to differentiate it from the MA-DO scheduler and to highlight the two degrees of opportunism it utilizes, both in channel variations and in the contact process statistics.

B. Homogeneous Contact Processes

In this subsection, we illustrate and compare the performance of the three opportunistic schedulers introduced above under homogeneous contact processes (cf. Section III). We also relax the Poisson contact process assumption to study the impact of implementing the opportunistic schedulers under more realistic mobility induced contact processes.

We examine two different contact processes. In the first scenario, contact time between any pair of users is generated according to an actual Poisson process as assumed in our theoretical model. In the second and more realistic scenario, we simulate the motion of users in the network, and declare a contact when two users actually fall within their peer-to-peer communication range ($d = 10$m). In this second scenario, we model the user mobility by the Random Waypoint (RWP) mobility model, which is one of the most widely used mobility models in protocol design and performance analysis/comparison in mobile ad-hoc networks [6]. As we have noted earlier, the contact processes arising from the RWP model have been shown to approximate homogeneous Poisson processes [7], [11], which motivates us to adopt the RWP model.

In the RWP model, each user chooses a random destination within the network area Ω, and moves towards its chosen destination on a straight line at a given speed $v > 0$. The entire procedure is repeated once the user arrives at its destination. In order to implement our GPF scheduler proposed in Section III, we need to obtain an estimate of the contact rate λ through numeric simulation. For the RWP mobility model with speed $v = 1$ m/s on the described network, we observe a contact rate of $\lambda \approx 1.39 \times 10^{-4}$. For a fair comparison, we choose this contact rate when generating the Poisson contact processes.

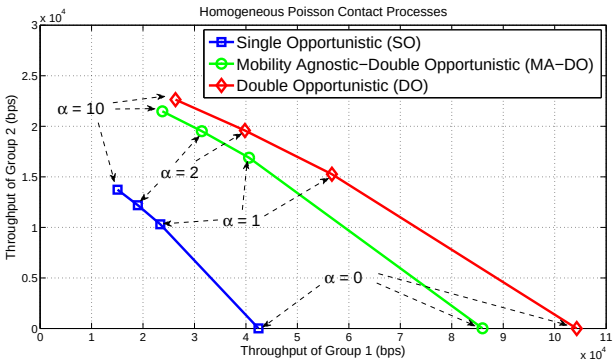

Fig. 1. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with homogeneous Poisson contact processes.

Figure 1 depicts the aggregate throughput of the two groups under the three scheduling schemes with simulated Poisson contact processes. We display results for the three scheduling scenarios (SO, MA-DO, and DO) for different group fairness parameters α. The results clearly reveal significant percentage gains (ranging from 50% to 100%) achieved by the MA-DO scheduler over the SO scheduler due its use of peer-to-peer forwarding capability. Also, we see that the DO scheduler provides another non-negligible level of improvement over the MA-DO scheduler due to its knowledge and effective use of contact process characteristics. When compared to the baseline SO scheduler, the full-fetched DO scheduler can observe a percentage gain between 75% and 150% in its aggregate throughput performance!

Homogeneous mobility among users results in homogeneous channel conditions, and as a result throughput is fairly equal across users. For this reason, we do not investigate fairness among users, and set the user fairness parameter $\beta = 0$.

For all three scheduling schemes, we observe that increasing α has the effect of equalizing the aggregate throughput of the two groups. The schedulers with $\alpha = 0$, corresponding to linear group utility functions, strive to maximize the sum total throughput of the two groups, and serve to the larger group exclusively. Adopting a larger group fairness parameter α increases the throughput of the smaller group at the cost of the total throughput. Another effect of increasing α is the narrowing gap between the throughput curves of the different scheduling schemes: schedulers must forego opportunities in order to meet stricter fairness constraints.

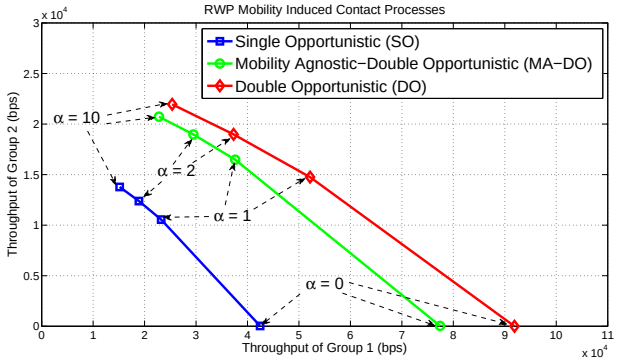

Fig. 2. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with RWP mobility induced contact processes.

Figure 2 displays the aggregate throughput of the two groups under the three scheduling schemes for *RWP mobility induced contact processes*. Not surprisingly, the baseline SO scheduler achieves the same throughput as with Poisson contact processes, since contact processes have no significance in the single opportunistic scheduling scenario. The aggregate throughput of both groups increases significantly once peer-to-peer communication is enabled by the MA-DO scheduler. Again, there is a further increase reaped by the DO scheduler that also utilizes the contact process characteristics. While the performance of the RWP mobility induced contact process deviates slightly from the simulated Poisson contact processes, the performance gains exhibit almost the same characteristics in both scenarios. This is a reassuring result that promotes the use of GPF strategy in more realistic mobility models.

C. Heterogeneous contact processes

In this last subsection, we assess the performance of the three opportunistic schedulers (the SO, MA-DO, and DO schedulers) under heterogeneous Poisson contact processes. We recall that the DO scheduler implements the GPF scheduler proposed in Section IV. As in the previous subsection, we consider two groups with 70 and 30 users, respectively, but also assume that both groups are further divided into two subgroups of fast and slow users (comprising 10% and 90% of the total number of users, respectively). We simulate Poisson contact processes of three different rates between two fast users ($\lambda_1 = 10^{-3}$), two slow users ($\lambda_2 = 10^{-5}$), and a fast and a slow user ($\lambda_{12} = 10^{-4}$).

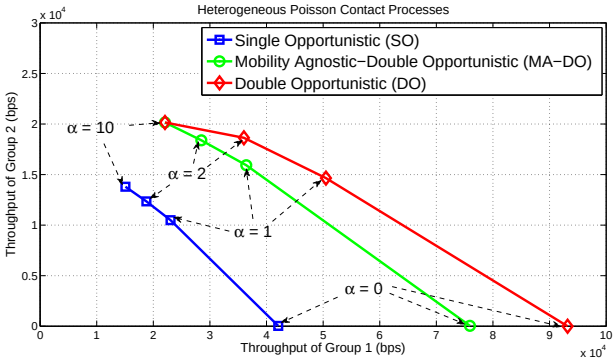

Fig. 3. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with heterogeneous Poisson contact processes.

Figure 3 displays the aggregate throughput of the two groups under the three scheduling schemes for heterogeneous Poisson contact processes. The baseline SO scheduler performance shows the same throughput as with Poisson contact processes, since contact processes have no significance in the single opportunistic scheduling scenario. The MA-DO and DO schedulers, again, provide significant performance improvements by effectively utilizing the peer-to-peer dissemination and contact process knowledge, respectively. These results validate both the fairness and efficiency aspects of our GPF design under the heterogenous contact processes.

Overall, the numerical investigations under both the homogenous and the heterogeneous mobility scenarios show significant and consistent gains that the class of GPF schedulers achieves through its opportunistic use of peer-to-peer data dissemination capabilities and its knowledge of contact statistics among users.

VI. Conclusion

In this paper we studied the propagation of deadline-based content in wireless network characterized by *heterogeneous* (time-varying and user-dependent) wireless *channel conditions*, *heterogeneous user mobility*, and where communication could occur in a *hybrid* format (e.g., directly from the central controller or by exchange with other mobiles in a peer-to-peer manner). For this 3H wireless system, we showed that by exploiting double opportunities of channel condition and mobility afforded us substantial performance gains. We introduced a set of Group Proportional Fairness (GPF) criteria to characterize different considerations of fairness and performance tradeoffs. We developed a class of double opportunistic multicast schedulers and proved their optimality in terms of both utility and fairness. Simulation results confirmed that the proposed algorithms significantly improved system performance in terms of both throughput and fairness. Our work provides the key first steps and guideline on how to *appropriately* exploit multiple opportunities in the design for content sharing in future wireless systems.

REFERENCES

[1] P. Agashe, R. Rezaiifar, and P. Bender, "Cdma2000 high rate broadcast packet data air interface design", in *IEEE Communications Magazine*, pages 83–89, Feb 2004.

[2] R. Agrawal, A. Bedekar, R. La, R. Pazhyannur, and V. Subramanian, "A class and channel-condition based weighted proportionally fair scheduler for edge/gprs", in *ITCOM'01*, Denver, CO, August 2001.

[3] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, Englewood Cliffs, NJ, second edition, 1992.

[4] S. Borst, "User-level performance of channel-aware acheduling algorithms in wireless data networks", in *IEEE/ACM Trans. Networking*, 13(3):636–647, 2005.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[6] J. Broch, D. Maltz, D. Johnson, Y. Hu, and J. Jetcheva, "Multi-hop wireless ad hoc networking routing protocols", in *ACM Mobicom*, Dallas, TX, March 1998.

[7] H. Cai and D. Y. Eun, "Aging Rules: What Does the Past Tell About the Future in Mobile Ad-Hoc Networks?" in *ACM MobiHoc*, New Orleans, LA, May 2009.

[8] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms", in *IEEE INFOCOM*, Barcelona, Catalunya, Spain, 2006.

[9] A. Chaintreau, J.-Y. Le Boudec, and N. Ristanovic, "The age of gossip: spatial mean field regime", in *SIGMETRICS '09: Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pages 109–120, New York, NY, USA, 2009.

[10] A. Eryilmaz and R. Srikant, "Joint Congestion Control, Routing and MAC for Stability and Fairness in Wireless Networks", in *IEEE Journal on Selected Areas in Communications*, pages 1514–1524, August 2006.

[11] R. Groenevelt, P. Nain, and G. Koole, "Message delay in MANET", in *Proceedings of ACM SIGMETRICS*, New York, NY, June 2004.

[12] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of Ad Hoc wireless networks", in *IEEE/ACM Transactions on Networking*, 4:477–486, August 2002.

[13] B. Han, P. Hui, M. Marathe, G. Pei, A. Srinivasan, and A. Vullikanti, "Cellular Traffic Offloading through Opportunistic Communications: A Case Study", in *CHANTS'10*, Chicago, Illinois, USA, Sep 2010.

[14] E. Hyytiä and J. Virtamo, "Random waypoint mobility model in cellular networks", in *Wirel. Netw.*, 13(2):177–188, 2007.

[15] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network", in *IEEE INFOCOM*, Rio de Janeiro, Brazil, April 2009.

[16] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system", in *Vehicular Technology Conference (VTC2000-Spring)*, pages 1854–1858, Tokyo, Japan, May 2000.

[17] F. Kelly, "Charging and rate control for elastic traffic", in *European Transactions on Telecommunications*, 8:33–37, Jan 1997.

[18] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability", in *Journal of the Operational Research Society*, pages 237–252, 1998.

[19] K. W. Kwong, A. Chaintreau, and R. Guerin, "Quantifying content consistency improvements through opportunistic contacts", in *CHANTS '09: Proceedings of the 4th ACM workshop on Challenged networks*, pages 43–50, 2009.

[20] X. Lin and N. B. Shroff, "The fundamental capacity-delay tradeoff in large mobile ad hoc networks", in *Third Annual Mediterranean Ad Hoc Networking Workshop*, 2004.

[21] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks", in *IEEE Journal on Selected Areas in Communications*, 24:1452–1463, 2006.

[22] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks", in *Comput. Netw.*, 41(4):451–474, 2003.

[23] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks", in *IEEE/ACM Transactions on Networking*, 7(4), 1999.

[24] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control", in *IEEE/ACM Trans. Netw.*, 8(5):556–567, 2000.

[25] T. Nandagopal, S. Lu, and V. Bharghavan, "A united architecture for the design and evaluation of wireless fair queueing algorithms", in *ACM MobiCom*, Seattle, Washington, Aug. 1999.

[26] T. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors", in *Proceedings of IEEE INFOCOM*, San Francisco, CA, 1998.

[27] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel", in *Proceedings of ACM Workshop on Wireless and Mobile Multimedia*, Seattle, WA, August 1999.

[28] T. Spyropoulos, K. Psounis, and C. Raghavendra, "Efficient Routing in Intermittently Connected Mobile Networks: The multi-copy case", in *IEEE/ACM Transactions on Networking*, Feb. 2008.

[29] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. de Amorim, "Relieving the Wireless Infrastructure: When Opportunistic Networks Meet Guaranteed Delays", in *IEEE WoWMoM*, 2011.

[30] H. Won, H. Cai, and A. Netravali K. Sabnani I. Rhee D. Y. Eun, K. Guo, "Multicast Scheduling in Cellular Data Networks", in *IEEE Transactions on Wireless Communications*, 8(9):4540–4549, 2009.

APPENDIX A
PROOF OF THEOREM 1

Recall that the scheduler depends on $\vec{r}(t)$ to make its decision. Next, we study the performance of the scheduler for each fixed maximum achievable rate vector $\vec{r}$ to show its optimality. For each such $\vec{r}$, we let $f_{n,R_i}^{S,\vec{r}}$ denote the frequency that a scheduler S chooses to broadcast to group n at rate R_i when the maximum achievable data rate vector is $\vec{r}$. The existence of this frequency is guaranteed by our definition of the feasible scheduler set $\mathcal{S}$. To simplify notation in what follows, let us define

$$R^{\vec{r}}(i,n,m) \triangleq$$

$$R_i \left[1_{\{r_{n,m} \geq R_i\}} + \frac{\chi_n(t,R_i) - \kappa_n(t,R_i)}{S_n - \kappa_n(t,R_i)} 1_{\{r_{n,m} < R_i\}} \right], \quad (21)$$

which gives the expected contribution to the throughput of user $u_{n,m}$ when the scheduler broadcasts to group n at rate R_i given $\vec{r}(t) = \vec{r}$ (cf. (7) and (10) for the definitions of $\kappa_n(\cdot,\cdot)$ and $\chi_n(\cdot,\cdot)$).

Then, the throughput user $u_{n,m}$ would get under scheduler S when the maximum achievable data rate vector were fixed to be $\vec{r}$ can be written as

$$\tau_{n,m}^{S,\vec{r}} = \sum_{i=1}^{K} f_{n,R_i}^{S,\vec{r}} R^{\vec{r}}(i,n,m). \quad (22)$$

Consequently, user $u_{n,m}$'s total throughput under scheduler S can be expressed as

$$\tau_{n,m}^{S} = \sum_{\text{all } \vec{r}} \pi(\vec{r}) \tau_{n,m}^{S,\vec{r}}, \quad (23)$$

where $\pi(\vec{r})$ is the probability of observing the maximum achievable rate vector $\vec{r}$, and the summation is carried out over the finite set of all possible maximum achievable rate vectors. Note that by the stationarity and periodicity of the maximum achievable data rate vector (Assumption 1), $\pi(\vec{r})$ corresponds to the fraction of time that $\vec{r}$ is in effect.

In the following, we compare our optimal scheduler S^* and any arbitrary feasible scheduler $S \in \mathcal{S}$. Let $\tau_{n,m}^{S^*}$ and $\tau_{n,m}^{S}$ denote the long-term throughputs user $u_{n,m}$ would get under schedulers S^* and S, respectively. Also, let $\gamma_n^{S^*}$ and γ_n^{S} denote the long-term aggregate user utilities of group n under schedulers S^* and S, respectively.

Given the concave and non-decreasing group and user utility functions defined in (3) and (4), the objective function in (6) is a concave function of the user throughput for any set of non-negative parameters $\{w_n\}_{n=1,\cdots,N}$, $\{v_{n,m}\}_{n=1,\cdots,N}^{m=1,\cdots,S_n}$, α and β. Thus, in order to prove the optimality of S^*, it suffices to show that the global optimality criterion for convex optimization ([5]) is satisfied, i.e.,

$$\sum_{n=1}^{N}\sum_{m=1}^{S_n}\frac{w_n}{\left(\gamma_n^{S*}\right)^{\alpha}}\cdot\frac{v_{n,m}}{(\tau_{n,m}^{*})^{\beta}}\cdot(\tau_{n,m}^{S}-\tau_{n,m}^{S^*})\leq 0. \tag{24}$$

Note that (24) is a legal expression if we adopt the convention that $0^0=1$. If there exists some n such that $\gamma_n^{=}0$, or some n,m such that $\tau_{n,m}^{S^*}=0$, then we must have $\alpha=0$ or $\beta=0$, respectively.

In light of (23), it suffices to show that for any given maximum achievable data rate vector $\vec{r}$

$$\sum_{n=1}^{N}\sum_{m=1}^{S_n}\frac{w_n}{\left(\gamma_n^{S^*}\right)^{\alpha}}\cdot\frac{v_{n,m}}{(\tau_{n,m}^{S^*})^{\beta}}\cdot(\tau_{n,m}^{S,\vec{r}}-\tau_{n,m}^{S^*,\vec{r}})\leq 0. \tag{25}$$

To show that (25) holds for any scheduler S, we first define $f_{n,n',R_i,R_j}^{S^*,S,\vec{r}}$ as the joint frequency that given the maximum achievable rate vector $\vec{r}$, scheduler S^* chooses to broadcast to group n at rate R_i *and* scheduler S chooses to broadcast to group n' at rate R_j. Therefore, we have

$$\tau_{n,m}^{S^*,\vec{r}}=\sum_{n'=1}^{N}\sum_{i,j=1}^{K}f_{n,n',R_i,R_j}^{S^*,S,\vec{r}}R^{\vec{r}}(i,n,m), \tag{26}$$

$$\tau_{n',m'}^{S,\vec{r}}=\sum_{n=1}^{N}\sum_{i,j=1}^{K}f_{n,n',R_i,R_j}^{S^*,S,\vec{r}}R^{\vec{r}}(j,n',m'). \tag{27}$$

Then, using (19), (11), (13) and (14), we have

$$\sum_{m'=1}^{S_{n'}}\frac{w_{n'}v_{n',m'}f_{n,n',R_i,R_j,\vec{r}}^{S^*,S}R^{\vec{r}}(j,n',m')}{\left(\gamma_{n'}^{S^*}\right)^{\alpha}(\tau_{n',m'}^{S^*})^{\beta}} \leq\sum_{m=1}^{S_n}\frac{w_nv_{n,m}f_{n,n',R_i,R_j,\vec{r}}^{S^*,S}R^{\vec{r}}(i,n,m)}{\left(\gamma_n^{S^*}\right)^{\alpha}(\tau_{n,m}^{S^*})^{\beta}}. \tag{28}$$

From (26), (27) and (28), we have

$$\sum_{n'=1}^{N}\sum_{m'=1}^{S_{n'}}\frac{w_{n'}v_{n',m'}\tau_{n',m'}^{S,\vec{r}}}{\left(\gamma_{n'}^{S^*}\right)^{\alpha}(\tau_{n',m'}^{S^*})^{\beta}} =\sum_{n'=1}^{N}\sum_{m'=1}^{S_{n'}}\sum_{n=1}^{N}\sum_{i,j=1}^{K}\frac{w_{n'}v_{n',m'}f_{n,n',R_i,R_j}^{S^*,S,\vec{r}}R^{\vec{r}}(j,n',m')}{\left(\gamma_{n'}^{S^*}\right)^{\alpha}(\tau_{n',m'}^{S^*})^{\beta}} \tag{29}$$

$$\leq\sum_{n=1}^{N}\sum_{m=1}^{S_n}\sum_{n'=1}^{N}\sum_{i,j=1}^{K}\frac{w_nv_{n,m}f_{n,n',R_i,R_j}^{S^*,S,\vec{r}}R^{\vec{r}}(i,n,m)}{\left(\gamma_n^{S^*}\right)^{\alpha}(\tau_{n,m}^{S^*})^{\beta}} \tag{30}$$

$$=\sum_{n=1}^{N}\sum_{m=1}^{S_n}\frac{w_nv_{n,m}\tau_{n,m}^{S^*,\vec{r}}}{\left(\gamma_n^{S^*}\right)^{\alpha}(\tau_{n,m}^{S^*})^{\beta}}, \tag{31}$$

where (29) and (31) follow from (27) and (26), respectively, and (30) follows from (28). This completes the proof of (25), which immediately yields the desired optimality criterion (24).

Appendix B
Proof of Theorem 2

The proof follows from the same line of argument as in the proof of Theorem 1 (cf. Appendix A), once we redefine

$$\begin{aligned}R^{\vec{r}}(i,n,m)\triangleq\sum_{m=1}^{S_n}v_{n,m}\frac{y}{\left(T_{n,m}(t)\right)^{\beta}}\Big[&1_{\{y\leq r_{n,m}(t)\}}\\&+\frac{\chi_n^1(t,y)-\kappa_n^1(t,y)}{S_n^1-\kappa_n^1(t,y)}1_{\{y>r_{n,m}(t)\}}1_{\{u_{n,m}\in\text{Subgroup 1}\}}\\&+\frac{\chi_n^2(t,y)-\kappa_n^2(t,y)}{\left(S_n^2-\kappa_n^2(t,y)\right.}1_{\{y>r_{n,m}(t)\}}1_{\{u_{n,m}\in\text{Subgroup 2}\}}\Big].\end{aligned}$$

Appendix C
Proof of Lemma 1

Let $\{X(s)\}_{s\geq 0}$ denote the number of users in group n who have a copy of the content at time s. Here, we measure the time s starting from the initial broadcast of the content. Note that $\{X(s)\}_{s\geq 0}$ is a continuous-time Markov chain with initial state $X(0)=N_0$, where $N_0\triangleq\kappa_n(t,y)$ is the number of users who receive the content directly from the BS at the time of the broadcast. Furthermore, the only non-zero transition probabilities are $\mathbb{P}\{X(s+\delta s)=i+1\mid X(s)=i\}=\lambda i(S_n-i)\delta s+o(\delta s)$ and $\mathbb{P}\{X(s+\delta s)=i\mid X(s)=i\}=1-\lambda i(S_n-i)\delta s-o(\delta s)$ for all $i\in\{N_0,\cdots,S_n-1\}$. Let us define $p_i(s)\triangleq\mathbb{P}\{X(s)=i\mid X(0)=N_0\}$, i.e., the probability that at time s there are i users with content when initially N_0 users received the content from the BS. Then, we can write the forward Kolmogorov equations as

$$\begin{aligned}&\dot{p}_{N_0}(s)=-\lambda N_0(S_n-N_0)p_{N_0}(s),\\&\cdots\\&\dot{p}_i(s)=\lambda(i-1)(S_n-i+1)p_{i-1}(s)-\lambda i(S_n-i)p_i(s),\\&\cdots\\&\dot{p}_{S_n}(s)=\lambda(S_n-1)p_{S_n-1}(s),\end{aligned} \tag{32}$$

Letting $\vec{P}(s)=[p_{N_0}(s),p_{N_0+1}(s),\cdots,p_{S_n(t)}]^T$, the set of equations in (32) can be rewritten as $\frac{d\vec{P}(s)}{ds}=-\lambda\mathbf{A}\vec{P}(s)$ where $\mathbf{A}$ is the infinitesimal generator defined in (9). Thus we have $\vec{P}(s)=e^{-\lambda\mathbf{A}s}\overrightarrow{N}_2^T$, where $\overrightarrow{N}_2$ is defined in (8). Finally, the average number of users with content at the end of the content lifetime L can be expressed as $\mathbb{E}[X(L)]=\sum_{i=N_0}^{S_n}i\cdot p_i(L)=\overrightarrow{N}_1e^{-\lambda\mathbf{A}L}\overrightarrow{N}_2^T$, where $\overrightarrow{N}_2$ is defined in (8).

Appendix D
Proof of Lemma 2

The proof follows from the same line of argument as in the proof of Lemma 1 (cf. Appendix C), when we consider the continuous-time Markov chain with state $\{(X_1(s),X_2(s))\}_{s\geq 0}$, where $X_1(s)$ and $X_2(s)$ denote the number of users in subgroups 1 and 2, respectively, who have a copy of the content at time s.