Exploiting Double Opportunities for Latency-Constrained Content Propagation in Wireless Networks

Han Cai, Irem Koprulu *Student Member, IEEE,*, and Ness B. Shroff *Fellow, IEEE,*

***Abstract*—In this paper, we focus on a mobile wireless network comprising a powerful communication center and a multitude of mobile users. We investigate the propagation of latency-constrained 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. At the same time mobile content sharing applications are becoming increasingly popular, and service providers are interested in supporting multicast communications over wireless networks. However, the resulting increase in traffic has 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. Verizon followed suit in July 2011, and is in the process of ending grandfathered unlimited data plans as of late 2013.

Advances in the communication capabilities of the devices has given rise to wireless networks that can communicate in a *hybrid* format, i.e., nodes can communicate directly with a powerful communication center and with other nodes in a peer-to-peer manner. Many cellular networks, military networks and sensor networks are equipped with this hybrid communication capability. Examples include mobile sensor networks with communication centers as well as cellular networks with pedestrians and vehicles carrying smartphones or tablets.

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-12-1-0385, and NSF grants CNS-1065136 and CNS-1012700.

While presenting a challenge in terms of bandwidth-intensive traffic, the increased density of *mobile* users also gives rise to an abundance of "contact" opportunities in hybrid networks, i.e., opportunities where mobile users are in close enough proximity to each other to communicate. As a result, content sharing through such contacts may occur at a similar time scale as that through a service provider. Resource allocation mechanisms that exploit all available opportunities are critical to serve the efficient usage and successful deployment of wireless systems. We propose to exploit the *double opportunities* of time-varying wireless channel conditions and random contact events among mobile users to facilitate an efficient solution for the downlink multicast scheduling problem.

Traditionally, opportunistic downlink scheduling, mobility, and content distribution have been extensively studied, but often *in isolation*. There are many studies on exploiting channel opportunities for unicast (e.g., [1], [2], [4], [20], [23]) and multicast (e.g., [28], [29]) scheduler design in cellular networks. These works do not exploit 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 (e.g., [7], [8], [10], [11], [27]). These works in mobile ad-hoc networks do not consider the wireless channel's inherent variability.

More recent works combine more aspects of opportunistic scheduling, mobility, and content distribution (e.g., [9], [12], [13], [18]) but to our knowledge, no paper develops a provably optimal solution addressing all three issues. In [9] the authors consider content distribution in a hybrid network model but do not exploit the diversity of the wireless channels. In [12], [13] the authors propose latency-constrained content forwarding exploiting the full potential of user mobility but not channel diversity.

In contrast to the existing literature, in this work we jointly exploit *both* time-varying and user-dependent channel conditions and random contact events among mobile users in order to realize the full potential of performance gains in content distribution. Specifically, our contributions are as follows.

- We develop a class of *double opportunistic* latency constrained multicast scheduling algorithms which *simultaneously* exploit channel opportunities and random mobility. We prove the optimality of our algorithm and show via numerical simulation that the performance gains obtained by jointly exploiting both opportunities are higher than those obtained by independent exploitation.

- We consider different fairness criteria between individual users and groups to enable a *two-layer* (user-and-group) realization of fairness. We employ numerical simulation to assess the trade-off between throughput and fairness.

The rest of the paper is organized as follows. In Section II we introduce our system model; in Sections III and IV we develop our double opportunistic scheduling algorithms for homogeneous and heterogeneous contact processes, respectively; in Section V we describe our simulation setup and results; in Section VI we conclude our work. Proofs are presented in the Appendices.

II. System Model and Problem Statement

We consider a downlink multicast scenario where a single base station (BS) broadcasts independent streams of latency-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 $u_{n,m}$ ($n = 1, \ldots, N$, $m = 1, \ldots, S_n$) represent user m 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, user $u_{n,m}$ can receive up to a *maximum achievable data rate* of $r_{n,m}(t) \in \{0, R_1, \ldots, R_K\}$ at time t. 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 for its decision. 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 $\vec{r}(t) \triangleq (r_{n,m})_{n=1,\ldots,N}^{m=1,\ldots,S_n}$ 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 user 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 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 the lifetime can also be utilized to achieve different trade-offs between throughput and delay, or to control the level of content flooding in the network.

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 two users (e.g., for bluetooth devices, $d \approx 10$m). Since we allow packet forwarding among users of the same group only, we are only concerned with contact events between pairs of users in the same group. 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 $u_{n,m}$ and $u_{n,m'}$ occurs during $[t_0, t_1)$ if $\|x_{n,m}(t_0^-) - x_{n,m'}(t_0^-)\| > d$, $\|x_{n,m}(t) - x_{n,m'}(t)\| \leq d$ for all $t \in [t_0, t_1)$, and $\|x_{n,m}(t_1) - x_{n,m'}(t_1)\| > d$. 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 $d \approx 10$m, 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], [10] under the well-known *i.i.d.* mobility model [19] 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.

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

[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.

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.* □

The motivation for the last two assumptions is based on a separation of time-scale concept. That is, mobility is a relatively slow process compared to exchange of information. Hence, while the contact time may be negligible compared to the inter-contact time, it is still likely to be quite large compared to the transmission times of information. Further, we expect that advances in new technologies will reduce the synchronization and transmission delays, hence improving the applicability of the final assumption.

D. Set of Feasible Schedulers: Recall that, at the t^{th} time slot, a scheduler S chooses a group index $n(t) \in \{1, \ldots, N\}$ and a transmission rate $r^g_{n(t)}(t) \in \{R_1, \ldots, R_K\}$. 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,

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

where $\mathcal{E}_{n,m,k,k+v}$ represents the event that user $u_{n,m}$ receives a copy of the content initially broadcast at time k at time $k+v$. 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 (i.e. packets of content type n do not experience a delay longer than L_n), and contact process dynamics (i.e., there is a contact between user $u_{n,m}$ and another user $u_{n,m'}$ carrying a copy of the content before the content expires). Also, note that the right hand side of the expression in (1) depends on the scheduler S through the choice of the transmission rates $r^g_n(k)$ and reception events $\mathcal{E}_{n,m,k,k+v}$. The throughput of each user is known at the BS: each user keeps track of its throughput and communicates this information via the uplink channel.

We define user $u_{n,m}$'s *long-term throughput* under the scheduler S as

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

To simplify notation, we will drop the index S and use $\tau_{n,m}$ when there is no ground for confusion. 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 [21].

E. Class of Group and User 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 both between individual users and groups. In other words, we need to ensure fairness *both* between different contents (i.e., groups) and different users subscribed to that content (i.e., individual users). Different choices for the utility functions for individual users have been proposed and studied in the literature (e.g., [15], [16], [22], [24], [25], [26]). 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.

In order to achieve fairness between both groups and individual users, we adopt two separate sets of group utility functions $\{G_n(\cdot)\}_{n=1,\ldots,N}$, and user utility functions $\{U_{n,m}(\cdot)\}^{m=1,\ldots,S_n}_{n=1,\ldots,N}$. We only require that $G_n(\cdot)$ and $U_{n,m}(\cdot)$ are non-decreasing concave functions defined on $[0,\infty)$. Given the group and user utility functions, we want to design a scheduler that maximizes the *total system utility*

$$\sum_n G_n\big(\sum_m U_{n,m}(\tau_{n,m})\big) \qquad (3)$$

where $\tau_{n,m}$ represents user $u_{n,m}$'s long-term throughput as defined in (2).

This forms a very general framework which encompasses several special cases of interest. If we set all group and user utility functions as the identity function, for example, we obtain the so-called MAX scheduler which maximizes the aggregate throughput of all users in the system, but may lead to very unfair resource allocation. As another special case, we derive and examine a scheduler which achieves the so-called $\alpha-$*proportional fairness* [23]. $\alpha-$Proportional Fairness incorporates many well-know fairness principles such as proportional fairness [16], [17] and max-min fairness [3].

For sets of non-negative parameters α, β, $\{w_n\}_{n=1,\ldots,N}$, and $\{v_{n,m}\}^{m=1,\ldots,S_n}_{n=1,\ldots,N}$, we define the proportional fair group utility functions G_n and user utility functions $U_{n,m}$ as follows

$$G_n(y) \triangleq \begin{cases} w_n \frac{y^{1-\alpha}}{1-\alpha}, & \alpha \geq 0,\ \alpha \neq 1 \\ w_n \log(y) & \alpha = 1, \end{cases} \qquad (4)$$

and

$$U_{n,m}(y) \triangleq \begin{cases} v_{n,m} \frac{y^{1-\beta}}{1-\beta}, & \beta \geq 0,\ \beta \neq 1 \\ v_{n,m} \log(y) & \beta = 1. \end{cases} \qquad (5)$$

If we consider the users of a single group and a fixed total allocated rate, then the scheduler solving $\max \sum_m U_{n,m}(\tau_{n,m})$ achieves $(\vec{w}, \alpha)$ proportional fairness [23] among these users. For $\alpha = 1$, the set of optimal throughputs $\{\tau^*_{n,m}\}^{S_n}_{m=1}$ satisfies $\sum_m w_{n,m}(\tau_{n,m} - \tau^*_{n,m})/\tau^*_{n,m} \leq 0$ for any other set of feasible throughputs $\{\tau_{n,m}\}^{S_n}_{m=1}$. In other words, the aggregate proportional changes in $\tau^*_{n,m}$ caused by any other scheduler are non-positive. This is exactly why this criterion is called 'proportional fairness' when $\alpha = 1$.

We say that a scheduler that maximizes the total system utility (3) with the group and user utility functions defined in (4) and (5), respectively, achieves the $(\vec{w}, \vec{v}, \alpha, \beta)$ *group proportional fairness (GPF) criterion*, where $\vec{w} = \{w_n\}_n$ and $\vec{v} = \{v_{n,m}\}_{n,m}$. We call α and β the *group* and *user fairness parameters*, respectively, and show in Section V that these can be adjusted to control the fairness among groups and users.

F. Problem Statement: Given the descriptions of the channel

conditions, content lifetime, contact process dynamics, and utility functions, we are ready to formulate our double opportunistic scheduling problem.

Double Opportunistic Problem (DOP):

$$\max_{S\in\mathcal{S}} \quad \sum_{n=1}^{N} G_n\Big(\sum_{m=1}^{S_n} U_{n,m}(\tau_{n,m}^S)\Big) \tag{6}$$

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

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.

We will first solve this problem under the assumption of statistically homogeneous user mobility in Section III, and then discuss its extension to heterogeneous scenarios in Section IV. In both scenarios, we allow the channel conditions to be statistically heterogeneous across users.

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 for the case of *homogeneous* Poisson contact processes, where the contact rates for all pairs of users in group n are equal to λ_n. This allows us to introduce the optimal algorithm that is extendable to the heterogeneous Poisson contact processes scenario (cf. Section IV) without the cumbersome notation necessary to deal with the heterogeneity.

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

$$X_0(n,t,y) \triangleq \sum_{m=1}^{S_n} 1_{\{y\le r_{n,m}(t)\}}, \tag{7}$$

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.

In order to simplify notation, we set $X_0 = X_0(n,t,y)$, and define two vectors of length $(S_n - X_0 + 1)$ each, as follows

$$\vec{Y_1} \triangleq (X_0, X_0+1, \ldots, S_n), \tag{8}$$

$$\vec{Y_2} \triangleq (1, 0, \ldots, 0). \tag{9}$$

We also define the matrix $\mathbf{A} = [a_{i,j}]_{i,j}$ of size $(S_n - X_0 + 1)\times(S_n - X_0 + 1)$, where

$$a_{i,j} = \begin{cases} (X_0+i-1)(S_n - X_0 - i + 1) & \text{if } i = j, \\ -(X_0+i-2)(S_n - X_0 - i + 2) & \text{if } i = j+1, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The physical meaning of the matrix $\mathbf{A}$ is described in detail in Appendix A. It is part of the infinitesimal generator matrix $\lambda_n\mathbf{A}$ of the continuous-time Markov chain that governs the evolution of the number of users in group n who receive the content over time. Each entry $a_{i,j}$ indicates the rate of transition from a state with j users with content to a state with i users with content.

With these definitions in place, we are ready to state our first result. The following lemma expresses the average number of users that receive the content *directly or indirectly* within its lifetime as a function of the channel dynamics and contact process characteristics .

Lemma 1. *If at time t the BS broadcasts to group n at rate y, then the expected number of users in group n who will have a copy of the content by the time it expires is given by*

$$X_L(n,t,y) = \vec{Y_1} e^{-\lambda_n \mathbf{A} L_n} \vec{Y_2}^T. \tag{11}$$

Proof: See Appendix A. ■

Before we introduce the optimal double opportunistic scheduler S^*, let us define a few auxiliary functions to 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.

$$\varphi_n(t) \triangleq \sum_{m=1}^{S_n} U_{n,m}\left(T_{n,m}(t)\right), \tag{12}$$

where $T_{n,m}(t)$ is the throughput of user $u_{n,m}$ at time t. Also, for group n, time slot t and rate y, we define

$$\Delta\varphi_n(t,y) \triangleq \sum_{m=1}^{S_n} U'_{n,m}(T_{n,m}(t))\cdot y\big[1_{\{y\le r_{n,m}(t)\}} + \frac{X_L(n,t,y) - X_0(n,t,y)}{S_n - X_0(n,t,y)} 1_{\{y> r_{n,m}(t)\}}\big]. \tag{13}$$

Note that $\Delta\varphi_n(t,y)$ is a measure of the marginal increase in the aggregate user utility $\varphi_n(t)$ of group n when in slot t the BS broadcasts to that group at rate y.

Now we are ready to formulate our optimal double opportunistic scheduler S^*.

Optimal DO scheduler S^*:

BS part:

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

$$r_n^g(t) \in \underset{y\in\{R_1,\ldots,R_K\}}{\arg\max} \Delta\varphi_n(t,y). \tag{14}$$

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

$$n(t) \in \underset{1\le n\le N}{\arg\max}\, G'(\varphi_n(t))\Delta\varphi_n(t, r_n^g(t)) \tag{15}$$

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.

Computational Complexity: At each time slot, the DO scheduler determines a candidate rate $r_n^g(t)$ for each of the

N groups as in (14). This amounts to finding the maximum among $N \times K$ $\Delta\varphi_n(t,y)$ values. These values are defined in (13) and can be easily calculated by employing a table-lookup for $X_L(n,t,y)$. The computation of this latter quantity as given in (11) involves the exponentiation of a matrix of size up to $(\max_n S_n) \times (\max_n S_n)$ and is computationally expensive. However, at each time slot t, the value of $X_L(n,t,y)$ depends only on the value of $X_0(n,t,y)$ and other fixed system parameters. Thus it can be computed offline and stored in a lookup table. The second decision of the DO scheduler as given in (15) involves the group to be broadcast to. This involves finding the maximum of among N expressions which are in part computed as a result of the first decision. The low complexity of this algorithm makes its real-time implementation feasible even for systems with a larger number of groups N and transmission rates K.

While we give a rigorous proof for the optimality of the DO scheduler S^* in the subsequent theorem, let us also provide the intuition behind its decision making. At each slot t, S^* starts by assigning each group n a potential rate $r_n^g(t)$ which maximizes the marginal increase $\Delta\varphi_n(t,y)$ in the aggregate user utility of that group. Once the scheduler decides the optimal potential rates for each group, it chooses to transmit to the group that will result in the largest increase in the total system utility (6). The expression maximized in (15) is the marginal increase in the group utility of group n if the BS broadcasts to that group at the potential rate $r_n^g(t)$.

We now present the main result in this section:

Theorem 1. *The DO scheduler S^* solves the Double Opportunistic Problem (6) optimally under homogeneous Poisson contact processes.*

Proof: See Appendix B. ■

A. The Double Opportunistic GPF Scheduler under Homogenous Poisson Contact Processes

In this subsection, we present as a special case the double opportunistic *group proportional fair* (GPF) scheduler which solves the double opportunistic problem of (6) for the special set of proportional fair group and user utility functions defined in (4) and (5). For this set of user utility functions, the marginal increase in the aggregate user utility of group n, at time t and rate y is given by

$$\Delta\varphi_n(t,y) \triangleq \sum_{m=1}^{S_n} \frac{v_{n,m}}{\big(\max\{T_{n,m}(t),\epsilon\}\big)^{\beta}} \cdot y\big[1_{\{y\le r_{n,m}(t)\}} + \frac{X_L(n,t,y)-X_0(n,t,y)}{S_n - X_0(n,t,y)} 1_{\{y>r_{n,m}(t)\}}\big], \tag{16}$$

where $\epsilon \to 0^+$ serves to prevent a division by zero.

Then the Double Opportunistic scheduler reduces to the $(\vec{w},\vec{v},\alpha,\beta)$ GPF scheduler.

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

BS part:

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

$$r_n^g(t) \in \underset{y\in\{R_1,\dots,R_K\}}{\arg\max}\ \Delta\varphi_n(t,y). \tag{17}$$

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

$$n(t) \in \underset{1\le n\le N}{\arg\max}\ \frac{w_n \cdot \Delta\varphi_n(t, r_n^g(t))}{\big(\max\{G_n(t),\epsilon\}\big)^{\alpha}}, \tag{18}$$

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 (18), the scheduler chooses the group that maximizes the numerator of the expression in (18).

The optimality of the GPF scheduler described above follows as a special case from Theorem 1. We have described it as a special case since proportional fairness is an important metric and since we use it to conduct our simulations in Section V.

IV. Double Opportunistic Multicast Scheduling under Heterogeneous Poisson Contact Processes

In the previous section, we assume homogeneous Poisson contact processes with the same contact rate between all users within a group. In this section, we extend our results to include scenarios with heterogeneous Poisson contact processes with different contact rates between users within a group. 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 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.

We assume that group n has S_n^1 users in subgroup 1, and S_n^2 users in subgroup 2 with $S_n^1 + S_n^2 = S_n$. Pairs of users in subgroup 1 have contact rate λ_1, pairs of users in subgroup 2 have contact rate λ_2, while two users from different subgroups have contact rate λ_{12}. Similar to the homogeneous case, we define

$$X_0^1(n,t,y) \triangleq \sum_{\{m:u_{n,m}\in\text{Subgroup 1}\}} 1_{\{y\le r_{n,m}(t)\}}, \tag{19}$$

and

$$X_0^2(n,t,y) \triangleq \sum_{\{m:u_{n,m}\in\text{Subgroup 2}\}} 1_{\{y\le r_{n,m}(t)\}}, \tag{20}$$

where $X_0^i(n,t,y)$ represents the number of users in subgroup i ($i=1,2$) of group n receiving content *directly* from the BS if at time t the BS broadcasts to group n at rate y. Next, we want to express the average number of users in each subgroup that receive the content either *directly* or *indirectly* within the content lifetime, as in the homogeneous scenario.

The number of users in subgroups 1 and 2, who have a copy of the content at a given time s forms a continuous-time Markov chain with the two dimensional state $\{(X_s^1, X_s^2)\}_{s\geq 0}$. 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 $i : \{0,1,\ldots,S_n^1\} \times \{0,1,\ldots,S_n^2\} \mapsto \{1,2,\ldots,(S_n^1+1)(S_n^2+1)\}$ 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 $f_1, f_2 : \{1,2,\ldots,(S_n^1+1)(S_n^2+1)\} \mapsto \mathbb{N}$ be the inverse mappings such that $f_j(i(k_1,k_2)) = k_j$ ($j=1,2$). Let $i_0(t,y) = i(X_0^1(n,t,y), X_0^2(n,t,y))$ be the sequence number of the initial state $(X_0^1(n,t,y), X_0^2(n,t,y))$. Define

$$\vec{Y}_1^1 \triangleq (f_1(1), f_1(2), \ldots, f_1((S_n^1+1)(S_n^2+1))), \quad (21)$$

$$\vec{Y}_1^2 \triangleq (f_2(1), f_2(2), \ldots, f_2((S_n^1+1)(S_n^2+1))), \quad (22)$$

and

$$\vec{Y}_2(t,y) \triangleq \vec{e}_{i_0(t,y)} = (0,\ldots,0,1,0,\ldots,0), \quad (23)$$

where $\vec{e}_{i_0(t,y)}$ denotes the $i_0(t,y)^{th}$ unit vector.

The $(S_n^1+1)(S_n^2+1)$ by $(S_n^1+1)(S_n^2+1)$ generator matrix $\mathbf{A} = [a_{i,j}]_{i,j}$ can be described as follows: For $k_1 = 0,\ldots,S_n^1-1$ and $k_2 = 0,\ldots,S_n^2$ let

$$a_{i(k_1,k_2),i(k_1+1,k_2)} = \lambda_1 k_1 (S_n^1 - k_1) + \lambda_{12} k_2 (S_n^1 - k_1), \quad (24)$$

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

$$a_{i(k_1,k_2),i(k_1,k_2+1)} = \lambda_2 k_2 (S_n^2 - k_2) + \lambda_{12} k_1 (S_n^2 - k_2), \quad (25)$$

with all other off-diagonal entries being zero, i.e., $a_{j,k} = 0$, for $j \neq k$, and j,k not covered in (24) or (25). The diagonal entries are chosen as to result in a zero row sum $a_{j,j} = -\sum_k a_{j,k}$ for all j.

Lemma 2. *If at time t the BS broadcasts to group n 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*

$$X_L^1(n,t,y) = \vec{Y}_1^1 e^{\mathbf{A}L_n} \vec{Y}_2^T, \quad (26)$$

and

$$X_L^2(n,t,y) = \vec{Y}_1^2 e^{\mathbf{A}L_n} \vec{Y}_2^T, \quad (27)$$

respectively.

Proof: See Appendix C. ∎

As for the auxiliary functions, we define the aggregate user utility $\varphi_n(t)$ of group n exactly as in (12). To quantify the marginal increase in the aggregate user utility when in slot t the BS broadcasts to group n at rate y we define

$$\Delta\varphi_n(t,y) \triangleq \sum_{m=1}^{S_n} U'_{n,m}(T_{n,m}(t)) \cdot y\Big[1_{\{y\leq r_{n,m}(t)\}} + \frac{X_L^1(n,t,y) - X_0^1(n,t,y)}{S_n^1 - X_0^1(n,t,y)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 1}\}} + \frac{X_L^2(n,t,y) - X_0^2(t,n,y)}{S_n^2 - X_0^2(n,t,y)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 2}\}}\Big]. \quad (28)$$

With all these definitions in place, the description of the optimal DO scheduler S^* for the heterogeneous contact processes scenario remains unmodified except for the use of (28) instead of (13). Also, the optimality of the algorithm continues to hold as shown in the following theorem.

Theorem 2. *The DO scheduler S^* using* (28) *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 D. ∎

A. The Double Opportunistic GPF Scheduler Under Heterogeneous Poisson Contact Processes

In this subsection, we present as a special case the double opportunistic GPF scheduler for the scenario of heterogeneous Poisson contact processes described above. We have for group n, time slot t and rate y,

$$\Delta\varphi_n(t,y) = \sum_{m=1}^{S_n} \frac{v_{n,m}}{\big(\max\{T_{n,m}(t),\epsilon\}\big)^\beta} \cdot y\Big[1_{\{y\leq r_{n,m}(t)\}} + \frac{X_L^1(n,t,y) - X_0^1(n,t,y)}{S_n^1 - X_0^1(n,t,y)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 1}\}} + \frac{X_L^2(n,t,y) - X_0^2(t,n,y)}{S_n^2 - X_0^2(n,t,y)} 1_{\{y>r_{n,m}(t)\}} 1_{\{u_{n,m}\in\text{Subgroup 2}\}}\Big]. \quad (29)$$

where $\epsilon \to 0^+$ serves to prevent a division by zero. ∎

The description of our GPF scheduler S^* for the heterogeneous contact processes scenario remains unmodified except for the use of (29) instead of (16).

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; (iv) examine the effect of group utility function parameters on the fairness and throughput levels achieved by the schedulers; and (vi) demonstrate the need for separate utility functions for groups and individual users.

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 of 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 trade-off 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 $L = 180$ seconds for all groups.

We adopt the group proportional fair group and user utility functions given in (4) and (5), respectively. 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 time-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 $\lambda_n = 0$. 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 $\lambda_n = 0$, 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 rates $\{\lambda_n\}_{n=1}^N$. 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 three different contact processes. In the first scenario, the contact time between any pair of users is generated according to an actual Poisson process as assumed in our theoretical model. In the two other and more realistic scenarios, 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 these two scenarios, we model the user mobility by the Random Waypoint (RWP) mobility model and Manhattan mobility model , respectively. The RWP model is a widely used mobility model 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], [10], which motivates us to adopt the RWP model. The Manhattan mobility model, on the other hand, more realistically emulates mobility in an urban setting.

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 rates λ_n through numeric simulation. For the RWP mobility model with speed $v = 1$ m/s on the described network, we observe a contact rate of $\lambda \approx 1.39 \times 10^{-4}$.

In the Manhattan mobility model each user is constrained to move along a grid of horizontal and vertical paths that resemble streets in an urban environment. At each intersection, a user either continues straight on its given path with probability 1/2, or takes a right or left turn with probability 1/4 each. In our simulations we impose a rectangular grid with 50 m long blocks, and observe that users moving with speed $v = 1$ m/s on this grid meet at an average rate of $\lambda \approx 1.39 \times 10^{-4}$ - the same as from the RWP mobility model described above. For a fair comparison, we choose the same contact rate when generating the Poisson contact processes. Note that, while the RWP and Manhattan mobility models with the parameters given above result in the same average contact rate between users, the average distance of the users from the BS is different for the two models. We correct for this difference by adjusting the transmission power so that in both cases the users experience comparable achievable data rates in average.

Figure 1 depicts the aggregate throughput of the two groups under the three scheduling schemes with simulated Poisson contact processes. We display the 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

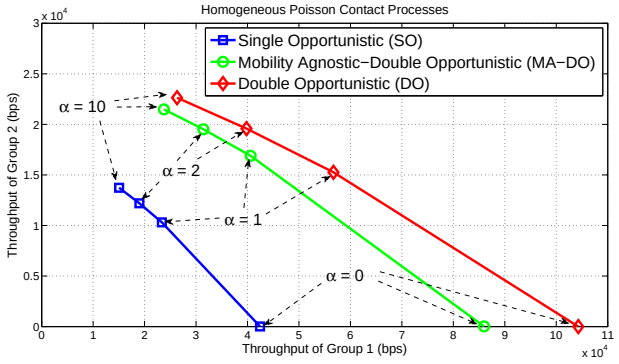


Fig. 1. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with homogeneous Poisson contact processes.

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 in this subsection, 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.

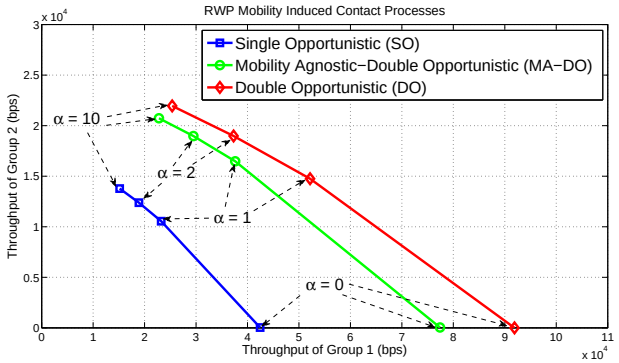


Fig. 2. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with RWP mobility induced contact processes.

Figures 2 and 3 display the aggregate throughputs of the two groups under the three scheduling schemes for *RWP and Manhattan mobility induced contact processes*. Not surprisingly, the baseline SO schedulers achieve 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. The throughputs of both groups are comparable for the RWP and Manhattan mobility models, since both mobility models give rise to contact processes with the same average contact rate ($\lambda \approx 1.39 \times 10^{-4}$). For both mobility models, there is a further increase reaped by the DO scheduler that also utilizes the contact process characteristics. While the performance of the RWP and Manhattan mobility induced contact processes deviates 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. Also note that the performance loss is greater for the Manhattan mobility model than it is for the RWP case. This is expected since the GPF strategy is optimal for Poisson contact processes, and the contact processes arising from RWP mobility resemble Poisson processes better than those arising from Manhattan mobility.

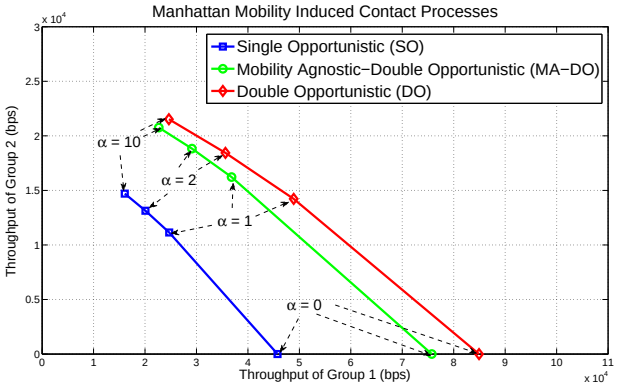


Fig. 3. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with Manhattan mobility induced contact processes.

C. Heterogeneous contact processes

In this 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 ($\lambda_{12} = 10^{-4}$).

Figure 4 visualizes both fairness and performance gains under these scenarios. It 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

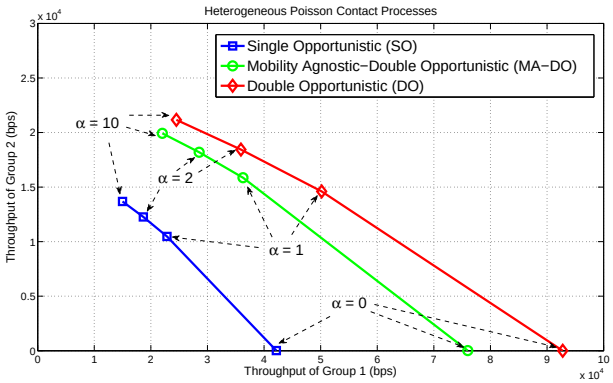


Fig. 4. Aggregate throughputs of the two groups of users (group 1: 70 users; group 2: 30 users) under the three opportunistic schedulers with heterogeneous Poisson contact processes.

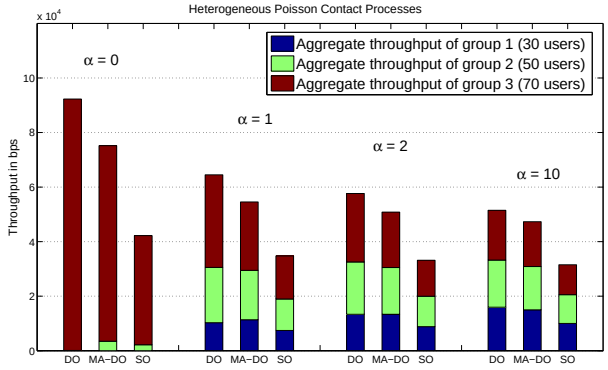


Fig. 5. Aggregate throughputs of three groups of users (group 1: 30 users; group 2: 50 users; group 3: 70 users) under the three opportunistic schedulers with heterogeneous Poisson contact processes.

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.

In this subsection, we also assess the performance of the three opportunistic schedulers (the SO, MA-DO, and DO schedulers) in a scenario with three groups. In particular, we consider groups with $30, 50$ and 70 users, where the fast and slow users comprise 10% and 90% of the total group sizes, respectively. As before, 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 ($\lambda_{12} = 10^{-4}$).

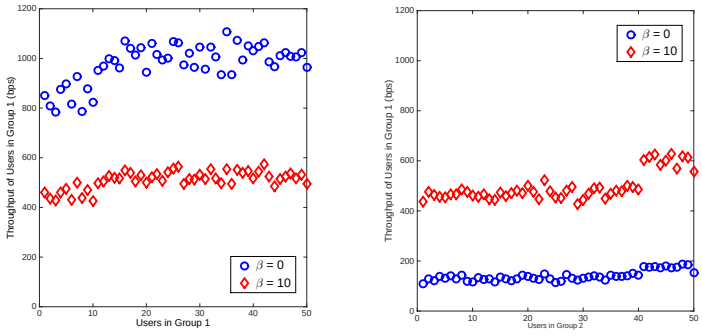
Figure 5 displays the aggregate throughput of the three groups achieved under the SO, MA-DO, and DO schedulers for a range of group fairness parameters α. As before, we observe how the MA-DO and DO schedulers provide significant performance improvements over the baseline SO scheduler, with the DO scheduler outperforming all. We also observe how the total system throughput is divided more evenly between the three differently sized groups as the group fairness parame α increases. This increase in fairness, however, comes at cost of lower total throughput for all three schedulers. The results validate both the fairness and efficiency aspects of GPF design under the heterogeneous contact processes.

D. Fairness trade-offs between users and groups

In this last subsection, we illustrate the need for util functions both for individual users and groups. In the previc subsections, we have already seen how a choice of differ group utility functions, in particular, the group fairness para eter α, determines the trade-off between the total throughput and fairness between two asymmetrically sized groups. In this subsection, we consider in addition an asymmetric scenario between different individual users to assess the interplay between fairness and throughput for individual users as well as groups.

Although subscribed to the same content, users in the same group may not receive an equal amount of data due to the variations in their channel conditions. To study the trade-off between individual users' throughput and the aggregate throughput of all users in the group, we consider a scenario in which some users' channels are consistently worse than other users' channels. We consider two groups of 50 users each on the same square-shaped network area Ω of size $(500\,m)^2$ with the BS located at the center. In addition to the slow and fast fading gain components as described in subsection V-A, some users' channel gains are degraded by a constant factor $0 < c < 1$ to model different reception capabilities of devices. We continue to use the group proportional fair group and user utility functions given in (4) and (5), respectively, with unit w_n and $v_{n,m}$ for all n, m.

Figure 6 shows how we can change the *user* fairness parameter β to shape the trade-off between fairness among users and aggregate throughput. In particular, a larger β can be used to give higher priority to the users with the consistently worse channel conditions when allocating the downlink bandwidth. However, doing so sacrifices some overall throughput, as well as throughput provided to users that see consistently better channels.



(a) Group 1 (50 users) (b) Group 2 (50 users)

Fig. 6. Adjusting the *user* fairness parameter β ($\{0, 10\}$) achieves fairness among individual users at the cost of low aggregate throughput. Group 1 has 10 users with consistently bad reception, while group 2 has 40.

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 latency-constrained content in a 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 a single base-station 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 trade-offs. 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. There are several interesting directions for extending our work, including wireless systems with multiple base stations, non-Poisson contact processes and incomplete contact information at the base-station(s).

Appendix A
Proof of Lemma 1

Suppose that at time t the BS transmits to group n at rate y and let $X_0 = X_0(n,t,y)$, be the number of users who receive the content directly from the BS at the time of the broadcast. Let $\{X_s\}_{s\geq 0}$ denote the number of users in group n who have a copy of the content at time $t+s$. Note that under Poisson contact processes $\{X_s\}_{s\geq 0}$ is a continuous-time Markov chain with initial state X_0. The only non-zero transition probabilities are

$$\mathbb{P}\{X_{s+\delta s} = i+1 \mid X_s = i\} = \lambda_n i(S_n - i)\delta s + o(\delta s) \quad (30)$$

and

$$\mathbb{P}\{X_{s+\delta s} = i \mid X_s = i\} = 1 - \lambda_n i(S_n - i)\delta s - o(\delta s) \quad (31)$$

for all $i \in \{X_0, \ldots, S_n - 1\}$.

Let us define $p_i(s) \triangleq \mathbb{P}\{X_s = i \mid X_0\}$, i.e., the probability that at time s there are i users with content when initially X_0 users received the content from the BS. Then, we can write the forward Kolmogorov equations as

$$\begin{aligned}
&\dot{p}_{X_0}(s) = -\lambda_n X_0(S_n - X_0)p_{X_0}(s),\\
&\ldots\\
&\dot{p}_i(s) = \lambda_n(i-1)(S_n - i + 1)p_{i-1}(s) - \lambda_n i(S_n - i)p_i(s),\\
&\ldots\\
&\dot{p}_{S_n}(s) = \lambda_n(S_n - 1)p_{S_n - 1}(s).
\end{aligned} \quad (32)$$

Letting $\vec{P}(s) \triangleq [p_{X_0}(s), p_{X_0+1}(s), \ldots, p_{S_n(t)}]^T$, the set of equations in (32) can be rewritten as

$$\frac{d\vec{P}(s)}{ds} = -\lambda_n \mathbf{A}\vec{P}(s) \quad (33)$$

where $\mathbf{A}$ is the infinitesimal generator defined in (10). Thus we have $\vec{P}(s) = e^{-\lambda_n \mathbf{A}s}\vec{Y}_2^T$, where $\vec{Y}_2$ is defined in (9). Finally, the average number of users with content at the end of the content lifetime L_n can be expressed as

$$\mathbb{E}[X_{L_n}] = \sum_{i=X_0}^{S_n} i \cdot p_i(L_n) = \vec{Y}_1 e^{-\lambda_n \mathbf{A}L_n}\vec{Y}_2^T, \quad (34)$$

where $\vec{Y}_1$ is as defined in (8).

Appendix B
Proof of Theorem 1

In this section we prove the optimality of the DO scheduler S^* from the set of feasible schedulers $\mathcal{S}$. Recall that a scheduler depends on the maximum achievable data rate vector $\vec{r}(t)$ to make its decision. First, we study the performance of a scheduler S for each fixed maximum achievable rate vector $\vec{r}$. For each such $\vec{r}$, we let $f_{n,R_i}^{S,\vec{r}}$ denote the fraction of time 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 $f_{n,R_i}^{S,\vec{r}}$ is guaranteed by our definition of the feasible scheduler set $\mathcal{S}$.

To simplify notation in what follows, let us define $R^{\vec{r}}(n,m,i) \triangleq$

$$R_i\left[1_{\{r_{n,m}\geq R_i\}} + \frac{X_L(t,n,R_i) - X_0(t,n,R_i)}{S_n - X_0(t,n,R_i)} 1_{\{r_{n,m}<R_i\}}\right], \quad (35)$$

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 (11) for the definitions of $X_0(\cdot,\cdot,\cdot)$ and $X_L(\cdot,\cdot,\cdot)$).

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

$$\tau_{n,m}^{S,\vec{r}} = \sum_{i=1}^{K} f_{n,R_i}^{S,\vec{r}} R^{\vec{r}}(n,m,i). \quad (36)$$

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

$$\tau_{n,m}^{S} = \sum_{\text{all } \vec{r}} \pi(\vec{r})\tau_{n,m}^{S,\vec{r}}, \quad (37)$$

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 DO 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 of user $u_{n,m}$ under schedulers S^* and S, respectively. Also, let $\varphi_n^{S^*}$ and φ_n^{S} denote the long-

term aggregate user utilities of group n under schedulers S^* and S, respectively.

Given the concave and non-decreasing nature of the group and user utility functions, the total system utility in (6) is a concave function of the user throughputs. Thus, in order to prove the optimality of S^*, it suffices to show that the global optimality criterion for convex optimization is satisfied [5], i.e.,

$$\sum_{n=1}^{N}\sum_{m=1}^{S_n} G_n'\big(\big(\varphi_n^{S*}\big)\big)\cdot U_{n,m}'(\tau_{n,m}^*)\cdot(\tau_{n,m}^{S}-\tau_{n,m}^{S^*})\le 0. \quad (38)$$

In light of (37), it suffices to show that for any given maximum achievable data rate vector $\vec{r}$,

$$\sum_{n=1}^{N}\sum_{m=1}^{S_n} G_n'\big(\big(\varphi_n^{S*}\big)\big)\cdot U_{n,m}'(\tau_{n,m}^*)\cdot(\tau_{n,m}^{S,\vec{r}}-\tau_{n,m}^{S^*,\vec{r}})\le 0. \quad (39)$$

To show that (39) holds for any scheduler S, we first define $f^{S^*,S}_{n,n',R_i,R_j}$ as the joint frequency under the maximum achievable rate vector $\vec{r}$ that 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

$$\tau_{n,m}^{S^*,\vec{r}}=\sum_{n'=1}^{N}\sum_{i,j=1}^{K} f^{S^*,S}_{n,n',R_i,R_j}R^{\vec{r}}(n,m,i), \quad (40)$$

$$\tau_{n',m'}^{S,\vec{r}}=\sum_{n=1}^{N}\sum_{i,j=1}^{K} f^{S^*,S}_{n,n',R_i,R_j}R^{\vec{r}}(j,n',m'). \quad (41)$$

Then, using (12), (13), (14) and (15), we have

$$\sum_{m'=1}^{S_{n'}} G_{n'}'\big(\big(\varphi_{n'}^{S*}\big)\big)\cdot U_{n',m'}'(\tau_{n',m'}^*)f^{S^*,S,\vec{r}}_{n,n',R_i,R_j}R^{\vec{r}}(j,n',m') \le \sum_{m=1}^{S_n} G_n'\big(\big(\varphi_n^{S*}\big)\big)\cdot U_{n,m}'(\tau_{n,m}^*)f^{S^*,S,\vec{r}}_{n,n',i,j}R^{\vec{r}}(n,m,i). \quad (42)$$

From (40), (41) and (42), we have

$$\begin{aligned}
&\sum_{n'=1}^{N}\sum_{m'=1}^{S_{n'}} G_{n'}'\big(\big(\varphi_{n'}^{S*}\big)\big)\cdot U_{n',m'}'(\tau_{n',m'}^*)\cdot\tau_{n',m'}^{S,\vec{r}}\\
&=\sum_{n'=1}^{N}\sum_{m'=1}^{S_{n'}}\sum_{n=1}^{N}\sum_{i,j=1}^{K} G_{n'}'\big(\big(\varphi_{n'}^{S*}\big)\big)\cdot U_{n',m'}'(\tau_{n',m'}^*)\\
&\cdot f^{S^*,S,\vec{r}}_{n,n',R_i,R_j}\cdot R^{\vec{r}}(j,n',m') \quad (43)\\
&\le\sum_{n=1}^{N}\sum_{m=1}^{S_n}\sum_{n'=1}^{N}\sum_{i,j=1}^{K} G_n'\big(\big(\varphi_n^{S*}\big)\big)\cdot U_{n,m}'(\tau_{n,m}^*)\\
&\cdot f^{S^*,S,\vec{r}}_{n,n',R_i,R_j}\cdot R^{\vec{r}}(n,m,i) \quad (44)\\
&=\sum_{n=1}^{N}\sum_{m=1}^{S_n} G_n'\big(\big(\varphi_n^{S*}\big)\big)\cdot U_{n,m}'(\tau_{n,m}^*)\cdot\tau_{n,m}^{S^*,\vec{r}}, \quad (45)
\end{aligned}$$

where (43) and (45) follow from (41) and (40), respectively, and (44) follows from (42). This completes the proof of (39), which immediately yields the desired optimality criterion (38).

Appendix C
Proof of Lemma 2

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

Appendix D
Proof of Theorem 2

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

$$\begin{aligned}
R^{\vec{r}}(n,m,i)\triangleq\sum_{m=1}^{S_n} v_{n,m}\frac{y}{\big(T_{n,m}(t)\big)^{\beta}}\big[&1_{\{y\le r_{n,m}(t)\}}\\
&+\frac{X_L^1(n,t,y)-X_0^1(n,t,y)}{S_n^1-X_0^1(n,t,y)}1_{\{y>r_{n,m}(t)\}}1_{\{u_{n,m}\in\text{Subgroup 1}\}}\\
&+\frac{X_L^2(n,t,y)-X_0^2(n,t,y)}{S_n^2-X_0^2(n,t,y)}1_{\{y>r_{n,m}(t)\}}1_{\{u_{n,m}\in\text{Subgroup 2}\}}\big].
\end{aligned}$$

References

[1] P. Agashe, R. Rezaiifar, and P. Bender. Cdma2000 high rate broadcast packet data air interface design. *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. *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] R. Groenevelt, P. Nain, and G. Koole. Message delay in MANET. In *Proceedings of ACM SIGMETRICS*, New York, NY, June 2004.

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

[12] 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.

[13] B. Han, P. Hui, V.S.A. Kumar, M.V. Marathe, J. Shao and A. Srinivasan. Mobile Data Offloading through Opportunistic Communications and Social Participation. In *Mobile Computing, IEEE Transactions on*, 11(5):821–834, 2012.

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

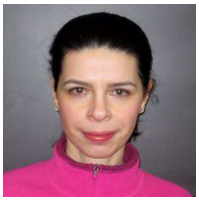
[15] A. Jalali, R. Padovani, and R. Pankaj. Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In *Vehicular Technology Conference (VTC2000-Spring)*, pages 1854–1858, Tokyo, Japan, May 2000.

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

[17] F. P. Kelly, A. Maulloo, and D. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, pages 237–252, 1998.
[18] 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.
[19] 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.
[20] X. Lin, N. B. Shroff, and R. Srikant. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected Areas in Communications*, 24:1452–1463, 2006.
[21] X. Liu, E. K. P. Chong, and N. B. Shroff. A framework for opportunistic scheduling in wireless networks. *Comput. Netw.*, 41(4):451–474, 2003.
[22] S. Lu, V. Bharghavan, and R. Srikant. Fair scheduling in wireless packet networks. *IEEE/ACM Transactions on Networking*, 7(4), 1999.
[23] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, 2000.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] M.O. Sunay and A. Eksim. Wireless multicast with multi-user diversity. In *Vehicular Technology Conference*, May 2004.
[29] 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.

Han Cai received her B.E. degree in Electrical Engineering from Chongqing University, China, in 2001, and her M.E. degree in Electrical Engineering from Beijing University, China, in 2004. After receiving her Ph.D. degree in Computer Engineering from North Carolina State University, Raleigh, NC, in 2009, she had been working with Professor Ness B. Shroff of ECE and CSE departments at The Ohio State University as a Postdoctoral Researcher till 2011.

Her research interests include performance analysis and design of MANET, stochastic approach for network performance analysis, and the effect of asynchronism in networks. She received the Best Student Paper Award in ACM MobiCom 2007.



Irem Koprulu received her B.S. degree in Electrical and Electronics Engineering from Boğaziçi University, Istanbul, in 1999, and her M.S. degrees in Electrical and Computer Engineering and Mathematics from the University of Illinois at UrbanaChampaign in 2005. She is currently working toward the Ph.D. degree in Electrical and Computer Engineering at The Ohio State University, Columbus.

Ness B. Shroff (S91-M93-SM01-F07) received his Ph.D. degree in Electrical Engineering from Columbia University in 1994. He joined Purdue university immediately thereafter as an Assistant Professor in the school of Electrical and Computer Engineering. At Purdue, he became Full Professor of ECE in 2003 and director of CWSA in 2004, a university-wide center on wireless systems and applications. In July 2007, he joined The Ohio State University, where he holds the Ohio Eminent Scholar endowed chair in Networking and Communications, in the departments of ECE and CSE. From 2009-2012, he served as a Guest Chaired professor of Wireless Communications at Tsinghua University, Beijing, China, and currently holds an honorary Guest professor at Shanghai Jiaotong University in China.

His research interests span the areas of communication, social, and cyberphysical networks. He is especially interested in fundamental problems in the design, control, performance, pricing, and security of these networks. Dr. Shroff is a past editor for IEEE/ACM Trans. on Networking and the IEEE Communication Letters. He currently serves on the editorial board of the Computer Networks Journal, IEEE Network Magazine, and the Networking Science journal. He has chaired various conferences and workshops, and co-organized workshops for the NSF to chart the future of communication networks. Dr. Shroff is a Fellow of the IEEE and an NSF CAREER awardee. He has received numerous best paper awards for his research, e.g., at IEEE INFOCOM 2008, IEEE INFOCOM 2006, Journal of Communication and Networking 2005, Computer Networks 2003 (two of his papers also received runner-up awards at IEEE INFOCOM 2005 and INFOCOM 2013), and also student best paper awards (from all papers whose first author is a student) at IEEE WiOPT 2013, IEEE WiOPT 2012 and IEEE IWQoS 2006. Dr. Shroff is among the list of highly cited researchers from Thomson Reuters (formerly ISI web of Knowledge) and in Thomson Reuters Book on The World's Most Influential Scientific Minds in 2014. In 2014, he received the IEEE INFOCOM achievement award for seminal contributions to scheduling and resource allocation in wireless networks.