Joint Optimization for Social Content Delivery in Heterogeneous Wireless Networks

Xiangnan Weng, John S. Baras
Institute for Systems Research and Department of Electrical and Computer Engineering
University of Maryland, College Park, Maryland 20742
Email: {wengxn, baras}@umd.edu

***Abstract*—Over the past decade, success of social networks has significantly reshaped how people consume information. Recommendation of contents based on user profiles is well-received. However, as users become dominantly mobile in content consumption, little is done to consider the optimization regarding the wireless environment. In this paper, we investigate a centralized wireless content delivery system with heterogeneous base stations, aiming to optimize overall user rewards given the capacity constraints of the wireless networks. We propose a scalable two-phase scheduling framework, consisting of: 1) distributed delivery decisions by each base station, and 2) resource consolidation by the system. Results indicate this novel joint optimization approach is both efficient and scalable.**

***Index Terms*—Social network, wireless network, optimization.**

I. Introduction

Social networks, which enable information sharing and consuming among users, have long existed on the Internet in various forms. However, it is not until the past decade that we witnessed their huge commercial and social success. Facebook, Twitter, YouTube, along with innumerable social networks, greatly facilitate information exchange for users all across the world. Their successes rely heavily, if not solely, on content recommendation based on user profiles, including but not limited to social associations, engagement history, etc. On one hand, social network providers are collecting more data about users than ever, to deliver highly relevant contents for users to consume; on the other hand, users are more willing to share personal data with their trusted social network providers, in return for better experience. Therefore, researchers in this area are working diligently to maximize user experience by recommending contents relevant to each individual user so that they are most likely to incur user engagements, with the help of sophisticated yet scalable machine learning and data mining algorithms on big data.

Most (if not all) social networks are reporting that their users are dominantly mobile users, i.e. most of the users access social networks from their mobile devices. Unfortunately, most of the social network (mobile) applications were designed from the root concept of wired connections, assuming unlimited capacity and/or almost no latency or failures in transmission. Unlike wired connections, wireless networks are limited by insufficient radio spectrum resource and ever changing channel characteristics. As multimedia contents de facto dominate contents consumed in social network applications, the disparity, between the capacity constraints of wireless networks and the assumptions of guaranteed delivery of contents, results in poor mobile experience and loss of user engagements, or ultimately users themselves. Even with new generations of access technology (LTE-Advanced and beyond), users are still unable to fetch their contents when the number of active users within the network increases, because the wireless network is a shared-medium communication and becomes congested as it approaches its capacity limits with quality of service degrades drastically. Hence, we are motivated to provide services in these scenarios so that users could still consume contents, even though the contents delivered to them might not be the best contents in the perspective of the social network applications.

Existing solutions fail to utilize the social nature of content consumption or the broadcast nature of wireless networks. In light of the preliminary research and industrial observations of social network applications, we note that users exhibit patterns of temporal, spatial and social correlations: 'similar' users are extremely likely to consume the same contents, though not at the exact same time. This fact leads us towards a novel solution that multicasts and precaches contents to groups of users to reduce redundant transmissions, as presented in [1]. It also suggests that under congested circumstances, the system shall deliver contents that maximize overall user experience across the system, but not necessarily optimal for each individual user. In other words, we are more tempted to deliver contents in a collectively optimal way when the wireless networks are congested. This approach requires information from both social and wireless networks to jointly optimize system decisions.

Certain preliminary work has been done concerning the joint user experience optimization incorporating social and wireless networks. In [1], we investigated the performance gained from a joint optimization approach for a single base station. The result is encouraging: the joint optimization approach outperforms existing layered solutions, especially when the wireless resource is insufficient.

In [2], a multicast pre-caching approach for video-on-demand service is employed to improve energy efficiency of the system. But the work stopped short to incorporate the fact that the contents/videos themselves are also subject to scheduling in social network applications. It is de facto for social network providers to shape the demand for contents,

978-1-5090-1311-1/16/$31.00 ©2016 IEEE

though from a different design rationale (increasing user engagements).

In [3] and [4], a pre-caching scheduler was proposed at social networks given the device profiles. The scheduler decides what contents and when to transmit based on the device profiles. However, the work did not address its implementation in real-world system in terms of the scale and the real-time information exchange of the system. The scheduling decisions rendered are relatively coarse. In the following work [5], a game theoretic approach was proposed to utilize the demand prediction from the devices. Unfortunately, this approach requires substantial effort of the mobile clients, while unable to provide the wireless base stations sufficient information regarding how to optimize overall system performance and the solution would not scale up well.

In [6], an opportunistic protocol based on a wireless ad-hoc system is proposed to forward contents with respect to possible future transmissions that could further disseminate the contents. This protocol considers the physical feasibility of wireless transmission, but does not discuss how to allocate wireless resource in different wireless conditions.

In this paper, we extend the system design in [1] to scenarios with multiple base stations and investigate the content delivery problem with capacity constraints in a system with centralized heterogeneous wireless infrastructure. We need to decide what contents to transmit to which users and how to transmit them, including by which base station(s). In addition, in a multiple base station setup, we need to decide which base station serves the users and how the base stations coordinate.

Traditional systems include two separated stages: (i) the social network application chooses the best recommended contents for users, regardless of the size and traffic load; (ii) the wireless network allocates resource for each (unicast) transmission. As stated before, this unicast approach is viable only when the wireless resource is sufficient and/or when users have absolutely nothing in common. However, as shown in [1] if the wireless resource is congested, this approach is extremely inefficient.

We introduce our system model and evaluation framework in Section II. Solutions to two types of system configuration regarding resource allocation are discussed: out-of-band system in Section III, and in-band system in Section IV. The performance is presented in Section V and we summarize our conclusions in Section VI.

II. Problem Formulation

A. General System Model

We consider a centralized system that both

1) selects contents to deliver according to user rewards given wireless capacity constraints; and
2) delivers the contents to users via a wireless network comprising of different types of base stations (as illustrated in Fig.1).

Channel information of all users for all base stations at all time slots $\left\{\text{SINR}_i^{(t),l}\right\}$ are reported to the system. We assume the bandwidth of the wired connections among the base stations and between base stations and content server(s) is sufficient, such that the base station could access contents as if they are stored locally. This assumption is valid in practice because the base stations are generally connected via fiber optic cables. Additionally, with the rapid developments of memory chips, the storage on user devices is sufficient to precache all the contents that users are possibly interested in within the scheduling horizon T_H. Intuitively, the contents that a user could and is willing to consume are bounded in both number and size.

The system is comprised of L base stations and is slotted with perfect synchronization (slot length T). At time slot t, each base station l is allocated bandwidth $B^{(t),l}$ for transmission.

There are M users and N contents to be scheduled. Contents are delivered to users using multicast and each scheduled content is transmitted within one scheduling slot to avoid system complexity. The reward of delivering content j to user i is denoted as f_{ij}, which remains unchanged during the scheduling horizon and could only be claimed in whole at most once (i.e. partial transmission earns no reward and repeated transmissions do not earn additional reward).

The objective of the system is to maximize overall user rewards obtained during the scheduling horizon, subject to the wireless capacity constraints.

$$
\begin{aligned}
\max \quad & \sum_{t=1}^{T_H}\sum_{l}\sum_{i,j}\alpha_{ij}^{(t),l}f_{ij} & \\
\text{s.t.} \quad & \alpha_{ij}^{(t),l}\in\{0,1\} & \forall i,j,t,l \\
& \sum_{t=1}^{T_H}\sum_{l}\alpha_{ij}^{(t),l}\le 1 & \forall i,j \\
& \text{QoS}\left(W_j,\left\{s_j^{(t),l}\right\},\left\{\text{SINR}_i^{(t),l}\right\}\right)=1 & \forall \alpha_{ij}^{(t),\lambda}=1 \\
& \sum_{j}s_j^{(t),l}\le B^{(t),l}T & \forall t,l \\
& s_j^{(t),l}\ge 0 & \forall j,t,l
\end{aligned}
\tag{1}
$$

We render two types of decisions (though the delivery decisions also intertwine with how to transmit):

1) What to transmit: binary delivery decision variable $\alpha_{ij}^{(t),l}$ indicates whether to transmit content j to user i at time slot t using base station l;
2) How to transmit: resource allocation $s_j^{(t),l}$ denotes how much wireless resource of base station l to allocate for content j at time slot t.

Note that (1) is only an offline version. In light of [1], its myopic online version (scheduling at each time slot without lookahead) performs relatively well. Additionally, the Quality of Service (QoS) requirements (binary indicator function with 1 denoting feasible and 0 infeasible) in the formulation are dependent on system configurations (e.g. power allocation, spectrum sharing or not).

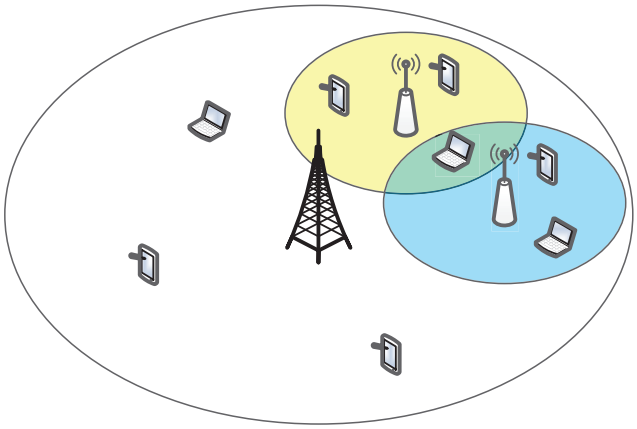

Fig. 1. System Model: Macro and Pico Cells.

In order to increase capacity and improve performance, especially at the edge of the cell, design of heterogeneous cells is introduced in modern cellular systems. There are mostly two types of cells in the systems: 1) macro cells that primarily provide coverage guarantee to ensure user connectivity; 2) pico cells that provide additional capacity to increase system performance at the edge of the macro cells to mitigate poor user Signal-to-Interference-Noise Ratio (SINR). There are generally two types of pico cells: one employing additional wireless resource ('out-of-band'), and the other sharing the same with macro cells ('in-band').

Table I summarizes definitions of parameters.

TABLE I
SUMMARY OF VARIABLES

Notations	Definition
M	Number of users.
N	Number of contents.
$\mathcal{L}$	Set of base stations.
f_{ij}	Reward for delivering content j to user i.
$\alpha_{ij}^{(t),l}$	Binary decision variable for base station l whether to transmit content j to user i at time slot t.
$B^{(t),l}$	Total available bandwidth for base station l at time slot t.
$s_j^{(t),l}$	Wireless resource allocated at base station l for content j at time slot t.
$\text{SINR}_i^{(t),l}$	Signal-to-Interference-Noise ratio of user i for base station l at time slot t
W_j	Size of content j in bits

B. Two-Phase Scheduling

The scheduling framework consists of two phases at each scheduling time slot t: delivery decisions and resource consolidation.

1) Delivery Decisions: In this phase, each base station decides individually what contents to deliver to users $\left\{\alpha_{ij}^{(t),l}\right\}$ and resource allocation for contents $\left\{s_j^{(t),l}\right\}$, given dedicated wireless resource allocated to the base station. This phase is distributed.

2) Resource Consolidation: In this phase, the system improves the decisions and resource allocation obtained in the previous phase. It is possible that the system changes both the delivery decisions and configurations of wireless resource to enhance spectrum efficiency. Some resource might be released for new allocation. This phase is centralized.

We could repeat the two phases if needed.

C. Decision Redundancy

Obviously, redundant transmissions are inevitable if each base station makes its decision individually.

Redundant decision is defined as: $\exists i, j, t_1, t_2, l_1, l_2$ and $(t_1, l_1) \neq (t_2, l_2)$, such that $\alpha_{ij}^{(t_1),l_1} = \alpha_{ij}^{(t_2),l_2} = 1$. Due to system design, we could guarantee that if a content has been transmitted to a user before, we do not retransmit in future slots. However, such guarantee does not exist within the same time slot in the delivery decisions phase, to reduce the complexity of the optimization and ensure timeliness.

We denote the percentage of redundant decisions for the system $\rho_t \in [0, 1]$ at time slot t as:

$$\rho_t = \frac{\|\{(i,j) : \sum_l \alpha_{ij}^{(t),l} > 1\}\|}{\|\{(i,j) : \sum_l \alpha_{ij}^{(t),l} \geq 1\}\|} \tag{2}$$

Ideally, we want to achieve zero redundancy in transmissions: $\rho_t = 0, \forall t$. If every base station schedules its own content transmissions without coordination, intuitively, we could encounter arbitrarily high redundancy.

D. Wireless Resource

Generally, there are two types of resource allocated to pico base stations in cellular networks: 1) Out-of-Band: pico base stations employ spectrum exclusively; 2) In-Band: pico base stations share spectrum with other base stations. The difference between the two types is interference. Out-of-band resource results in best signal, but we are unable to utilize the spatial separation of the scheduled users. In-band resource introduces interference, but it is possible to mitigate the negative impact by careful power allocation.

III. SYSTEM WITH 'OUT-OF-BAND' RESOURCE

For 'out-of-band' systems, the base stations do not share spectrum with each other. In these systems, we are scheduling the cells (whether macro or pico) with their dedicated wireless resource to transmit, i.e. no interference. Existing systems usually schedule the cell with which user achieves highest Signal-to-Noise ratio (SNR). However, with regard to our joint optimization system, this might not necessarily be the only solution, if not the worst.

In light of the results in the single cell scenario in [1], we focus on 'push' only systems in this paper, i.e. the system attempts to deliver contents without consideration of user-generated requests. It is easy to employ the same technique in [1] of altering reward values to incorporate the impact of user requests.

In this paper, we focus on one macro cell with several pico cells: $\mathcal{L} = \{\text{macro}\} \cup \mathcal{L}_{\text{pico}}$. With slight abuse of notation,

we drop the dependency on time slot in (1), as results in [1] indicated that myopic optimization performs fairly well and runs significantly faster. At each time slot t, we obtain content delivery decisions $\left\{\alpha_{ij}^{(t),l}\right\}$ and resource allocation $\left\{s_j^{(t),l}\right\}$ based on the solution to the optimization problem, constructed as a mixed integer programming problem as in (3):

$$\begin{aligned}
\underset{\alpha_{ij}^{(t),l}, s_j^{(t),l}}{\text{maximize}} \quad & \sum_l \sum_{i,j} \alpha_{ij}^{(t),l} f_{ij}^{(t)} \\
\text{subject to} \quad & \alpha_{ij}^{(t),l} \in \{0,1\} && \forall i,j,t,l \\
& \sum_l \alpha_{ij}^{(t),l} \leq 1 && \forall i,j \\
& \alpha_{ij}^{(t),l} W_j \leq s_j^{(t),l} \mathcal{R}(\text{SNR}_i^{(t),l}) && \forall i,j,l,t \\
& \sum_j s_j^{(t),l} \leq B^{(t),l} T && \forall t,l \\
& s_j^{(t),l} \geq 0 && \forall j,t,l
\end{aligned} \tag{3}$$

With slight abuse of notation, reward value of delivering content j to user i at time slot t is

$$f_{ij}^{(t)} = f_{ij}^{(t-1)} \left(1 - \max_l \alpha_{ij}^{(t-1),l}\right) \tag{4}$$

with initial values

$$f_{ij}^{(1)} = f_{ij} \tag{5}$$

This approach is different from existing systems, in that it considers the overall system rewards first rather than scheduling users to specific cells before content decision. It is intuitive due to the broadcast nature of wireless communications: all users with adequate channel states could obtain the content, thus reducing redundant transmissions of same contents. An extreme but illustrative example is the scenario where all users in the macro cell are interested in one content: if the cell selection happens first, both macro and pico base stations would transmit it, which might be inefficient in certain circumstances.

Our scheduling framework reduces the dependency among time slots, but it is still difficult to scale up. The major constraint that disables us from distributing the decision process to each individual base station is the single transmission constraint, i.e. each user shall only receive same content once, regardless of the cells selected to transmit. So we remove the single transmission constraints and let each base station render its delivery decisions locally, either by solving the mixed integer programming (3) for the optimal or by employing greedy algorithms. This way, the first phase is fast and distributed.

Therefore, the problem now is to reduce the decision redundancy ρ in the 'out-of-band' systems.

To begin with, we define a partial order for resource allocation decisions $(s^1, \ldots, s^{\mathcal{L}})$ as follows. Allocation decision $\vec{s}_1 = (s_1^1, \ldots, s_1^{|\mathcal{L}|})$ precedes $\vec{s}_2$ with regards to utility function f(denoted as $\vec{s}_1 \preceq_f \vec{s}_2$), iff $s_1^i \leq s_2^i, \forall i = 1, \ldots, |\mathcal{L}|$ and $f(\vec{s}_1) \leq f(\vec{s}_2)$.

Require: User SNR for each base station $\left\{\text{SNR}_i^l\right\}$, delivery decisions of each base station $\{\alpha_i^l\}$, content size W.

procedure GREEDYDECISIONDEDUPLICATION($\vec{s}_0, \hat{\mathcal{L}}$)
  $\vec{s} \leftarrow \vec{0}$
  for $l \leftarrow \{1, \ldots, |\hat{\mathcal{L}}|\}$ **do**
    for $i_0 \leftarrow \{i : \alpha_i^{\hat{\mathcal{L}}_l} = 1\}$ **do**
      $\tilde{\alpha}_{i_0}^{\hat{\mathcal{L}}_l} \leftarrow 1$ ▷ Initialization
      for $l' \leftarrow \{l+1, \ldots |\hat{\mathcal{L}}|\}$ **do**
        if $\alpha_{i_0}^{\hat{\mathcal{L}}_{l'}} == 1$ **then** ▷ If another BS could deliver the content
          $\tilde{\alpha}_{i_0}^{\hat{\mathcal{L}}_l} \leftarrow 0$ ▷ Unload user i to other base stations.
      if $\tilde{\alpha}_{i_0}^{\hat{\mathcal{L}}_l} == 1$ **then**
        $s^{\hat{\mathcal{L}}_l} \leftarrow \max\left(s^{\hat{\mathcal{L}}_l}, \frac{W}{\mathcal{R}\left(\text{SNR}_{i_0}^{\hat{\mathcal{L}}_l}\right)}\right)$ ▷ Recalculate resource allocation given new association.
  return $\vec{s}, \{\tilde{\alpha}_i^l\}$ ▷ Returns the new resource allocation and content delivery decisions.

Fig. 2. Greedy Decision Deduplication Algorithm

We define the overall reward function R for content j given wireless resource allocation $\vec{s}$ at time slot t_0 as follows:

$$R_j^{(t_0)}(\vec{s}) = \sum_i f_{ij}^{(t_0)} \cdot \max_l \mathbf{1}\left(\mathcal{R}\left(\text{SNR}_i^{(t_0),l}\right) \geq \frac{W_j}{s^l}\right) \tag{6}$$

Therefore, given initial resource allocation $\vec{s}_0$, if we cannot find any $\vec{s} \neq \vec{s}_0$, such that $\vec{s} \preceq_{-R} \vec{s}_0$, then we claim such resource allocation $\vec{s}_0$ is *non-improvable*.

For each content j, we consolidate the resource allocation $\vec{s}_{j,0}$ rendered by each individual cell using the Greedy Decision Deduplication Algorithm shown in Fig.2 with respect to an ordered permutation $\hat{\mathcal{L}}$ of base station set $\mathcal{L}$. Note, in the algorithm, we employ the trivial fact that each base station would schedule content if user SNR satisfies QoS requirements and user yields positive reward towards the content. This is due to the additive property of the objective function.

We prove that the new resource allocation $\vec{s}_j$ is non-improvable.

Theorem 1 (Non-improvability of Greedy Decision Deduplication Algorithm)**.** *The result obtained by the Greedy Decision Deduplication Algorithm is non-improvable and without decision redundancy, i.e.* $\rho = 0$.

Proof. The decision redundancy part is straightforward as outlined in the algorithm: all the redundant delivery decisions are removed. At most one base station will serve the user.

Proof of non-improvability by contradiction.

Without loss of gennerality, we start with $\hat{\mathcal{L}} = \{1, \ldots, |\mathcal{L}|\}$. Denote $\vec{s} = GreedyDecisionDeduplication(\vec{s}_0)$. Trivially, if we use less wireless resource, we will achieve no better performance, or $\forall \vec{s'} \preceq \vec{s}$, we have $R_j^{(t_0)}\left(\vec{s'}\right) \leq R_j^{(t_0)}(\vec{s})$. The loop invariance of greedy algorithm ensures that whether a user is

served or not remains unchanged before or after the algorithm; therefore $R_j^{(t_0)}(\vec{s}) = R_j^{(t_0)}(\vec{s}_0)$. Hence, we only need to prove $\nexists \vec{s'} \preceq \vec{s}, \vec{s'} \neq \vec{s}$, such that $R_j^{(t_0)}\left(\vec{s'}\right) = R_j^{(t_0)}(\vec{s})$.

Assume the contrary and denote the index of first discrepancy as l_0, i.e. $s_l = s'_l, \forall l \in \{1, \ldots, l_0 - 1\}, s_{l_0} < s'_{l_0}$. After each loop, all the users that could be served by other cells are offloaded and only the users that could not be served by any other cell will remain within the cell. Since $s'_{l_0} < s_{l_0}$, there exists at least one user that could be offloaded to other cells, which is contradictory to what the algorithm dictates. Therefore, the contrary assumption could not hold. □

Theorem 2 (Complexity of the Greedy Decision Deduplication Algorithm)**.** *The complexity of the Greedy Decision Deduplication Algorithm is $\mathcal{O}(M \cdot |\mathcal{L}|)$, and can be improved to $\mathcal{O}(M \cdot L_{\max})$, where $L_{\max}$ is the maximum number of cells a user could associate with:*

$$L_{\max} = \max_i \sum_l \mathbf{1}\left(SNR_i^l \geq SINR^{th}\right) \tag{7}$$

Therefore, using the Greedy Decision Deduplication Algorithm, we could efficiently and quickly figure out the overall resource allocation among different cells to deliver the same content without redundancy ($\rho = 0$). This essentially breaks down the large optimization problem containing multiple cells into a set of small optimization problems for each individual cell, hence reducing the complexity for scheduling.

IV. Systems with 'In-Band' Resource

For 'in-band' systems, different base stations could utilize the same wireless resource. It is not guaranteed to perform better, because simultaneous transmissions introduces undesirable interference at the receiver. However, if the users (receivers) are spatially separated, such interference might not degrade quality of service for intended transmissions and therefore might save wireless resource system-wide.

This is a much harder problem, because we could change user SINR by adjusting the transmit power of each base station. To reduce scheduling complexity, we could initially treat 'in-band' systems just like 'out-of-band' systems, as discussed in Section III, by allocating dedicated wireless resource to individual cell and applying the Greedy Decision Deduplication Algorithm to reduce redundant delivery decisions from different base stations. Afterwards, we determine whether we could further consolidate the wireless resource, by deciding whether or not to share the spectrum to transmit, in order to achieve better spectrum efficiency.

Therefore, with in-band wireless resource, we need to decide whether sharing the spectrum would be efficient for the base station set $\mathcal{L}$. Denote the set of users scheduled for transmission at base station l as $U_l = \{i : \alpha_i^l = 1\}$. Denote the SINR for spectrum sharing given a power allocation vector $\vec{P}$ as $\widetilde{\text{SINR}}(\vec{P})$. We have $\forall l \in \mathcal{L}$, the minimum SINR for scheduled users is:

$$\widetilde{\text{SINR}}^l(\vec{P}) = \min_{i \in U_l} \widetilde{\text{SINR}}_i^l(\vec{P}) \tag{8}$$

where the SINR from base station l to user i given the power allocation $\vec{P} \in \mathbb{R}_+^{|\mathcal{L}|}$ is denoted as:

$$\widetilde{\text{SINR}}_i^l(\vec{P}) = \frac{h_{l,i} P_l}{N_0 + \sum_{q \neq l} h_{q,i} P_q} \tag{9}$$

We denote $h_{q,i}$ as channel gain from base station q to user i.

Specifically, the SINR for transmission without spectrum sharing is SNR.

$$\begin{aligned} \text{SNR}_i^l &= \frac{h_{l,i} P_{l,\max}}{N_0} \\ &= \widetilde{\text{SINR}}_i^l(0, \ldots, P_{l,\max}, \ldots, 0) \end{aligned} \tag{10}$$

Trivially,

$$\widetilde{\text{SINR}}^l(\vec{P}) < \text{SNR}^l = \min_{i \in U_l} \text{SNR}_i^l \tag{11}$$

With slight abuse of notation, we omit $\vec{P}$ in the following discussions for simplicity, but any variables with tilde imply their dependencies on the power allocation vector $\vec{P}$.

Obviously, the resource allocation is based on the scheduled user(s) with worst wireless channel state, with or without spectrum sharing:

$$s_l = \max_{i \in U_l} \frac{W_l}{\mathcal{R}\left(\text{SNR}_i^l\right)} = \frac{W_l}{\mathcal{R}\left(\text{SNR}^l\right)} \tag{12}$$

$$\tilde{s}_l = \frac{W_l}{\mathcal{R}\left(\widetilde{\text{SINR}}^l\right)} = s_l \cdot \frac{\mathcal{R}\left(\text{SNR}^l\right)}{\mathcal{R}\left(\widetilde{\text{SINR}}^l\right)} \tag{13}$$

Denote

$$\tilde{s}_{\min} = \min_l \tilde{s}_l = \min_l \left(\frac{s_l}{\tilde{c}_l}\right) \tag{14}$$

with rate decay ratio $\tilde{c}_l$ for base station l defined as

$$\tilde{c}_l = \frac{s_l}{\tilde{s}_l} = \frac{\mathcal{R}\left(\widetilde{\text{SINR}}^l\right)}{\mathcal{R}\left(\text{SNR}^l\right)}, \forall l \in \mathcal{L} \tag{15}$$

Apparently, $0 \leq \tilde{c}_l \leq 1$.

The overall resource allocated for spectrum sharing among the base station set $\mathcal{L}$ is thus comprised of interfered and non-interfered parts:

$$\tilde{s} = \tilde{s}_{\min} + \sum_{l \in \mathcal{L}} \left(1 - \frac{\tilde{s}_{\min}}{\tilde{s}_l}\right) \cdot s_l \tag{16}$$

We could decide whether to share spectrum to transmit depending on: $\sum_l s_l \gtrless \tilde{s}$

The wireless resource saved by sharing spectrum $\tilde{\Delta}$ could be written as

$$\begin{aligned} \tilde{\Delta} &= \sum_l s_l - \tilde{s} \\ &= \tilde{s}_{\min} \left(\sum_{l \in \mathcal{L}} \tilde{c}_l - 1\right) \end{aligned} \tag{17}$$

Therefore, we have improvement condition:

Theorem 3 (Spectrum-Sharing Criterion)**.** *Spectrum sharing uses less resource for the rate decay ratio vector* $\tilde{\mathbf{c}} = (\tilde{c}_1, \ldots, \tilde{c}_{|\mathcal{L}|}) \in [0,1]^{|\mathcal{L}|}$*, iff*

$$\sum_{l \in \mathcal{L}} \tilde{c}_l > 1 \tag{18}$$

Note that the discussion above is independent of which contents are being transmitted, therefore, it could be applied to different and/or same contents, rather than deduplicating delivery decisions for the same contents, as in the 'out-of-band' discussion.

If the available transmission modes are limited, as in practical systems (e.g. LTE [7]), we can transform the problem into feasibility problems with different parameters. The basic formulation is: given the target SINR level $\text{SINR}_{th}^{k_l}$ for transmission mode $k_l \in \{1, \ldots, K\}$ at each base station $l \in \mathcal{L}$, determine whether a power allocation vector $\vec{P}$ for spectrum sharing exists such that,

$$\widetilde{\text{SINR}}_i^l(\vec{P}) \geq \text{SINR}_{th}^{k_l}, \forall i \in U_l \tag{19}$$

Trivially, we require mutual exclusiveness (20):

$$U_{l_1} \cap U_{l_2} = \emptyset, \forall l_1 \neq l_2 \in \mathcal{L} \tag{20}$$

Essentially, it is equivalent to a feasibility problem with respect to the linear constraint set:

$$\begin{array}{cc} -\frac{h_{l,i}}{\text{SINR}_{th}^{k_l}} P_l + \sum_{q \neq l} h_{q,i} P_q + N_0 \leq 0 & \forall i \in U_l \\ 0 \leq P_l \leq P_l^{\max} & \forall l \in \mathcal{L} \end{array} \tag{21}$$

We could further normalize the constraint set to

$$\begin{array}{cc} -\frac{p_{il}}{\text{SINR}_{th}^{k_l}} \xi_l + \sum_{q \neq l} p_{iq} \xi_q \leq -1 & \forall i \in U_l \\ 0 \leq \xi_l \leq 1 & \forall l \in \mathcal{L} \end{array} \tag{22}$$

where p_{il} is the maximum receiver SNR of user i for base station l

$$p_{il} = \frac{h_{l,i}}{N_0} P_l^{\max} \tag{23}$$

The power allocation for each base station is thus a feasible solution to the linear programming constraint set

$$P_l = \xi_l P_l^{\max} \tag{24}$$

The feasibility problem with respect to linear constraint set is a special form of linear programming problem, which could be solved efficiently and fast in most practical problems by the Simplex algorithm. Therefore, the problem of wireless resource consolidation among base stations is reduced to a solvable form.

For each candidate decision of transmission modes $(k_1, \ldots, k_{|\mathcal{L}|}) \in \{1, \ldots, K\}^{|\mathcal{L}|}$ that satisfies (25), we run feasibility test for constraint set (22) to determine if a power allocation solution is available for such decision

$$\sum_{l \in \mathcal{L}} \frac{R_{k_l}}{\mathcal{R}\left(\text{SINR}^l\right)} > 1 \tag{25}$$

V. Simulations and Results

A. Simulation Setup

There are $M = 300$ users and $N = 600$ contents in the system. As far as we understand, there are no generative models available that could mimic real-world data, so we settle on data-driven simulations. Reward values f_{ij}'s are taken from datasets of Yahoo [8] and MovieLens [9], and normalized to $[0, 1]$.

Content size is independent and uniformly distributed in $[5, 40]$ Mbits. The scheduling time slot has length of $T = 1s$ and the scheduling horizon is $T_H = 10$. System level parameters are shown in Table II [10].

There is one macro base station and two pico base stations. The two pico base stations are located with distance of $1.9r$ (where r is the designed range for a pico base station), ensuring there is certain but not major coverage overlap of the two pico base stations. The pico base stations are assigned equal spectrum resource in the delivery decisions phase and such assignment is time-invariant, i.e. $B^{(t),l} = B^l$. Each base station makes its delivery decisions by solving the mixed integer programming problem (3) for the optimal.

TABLE II
System Level Simulation Parameters

Simulation Parameter	Value
UE distribution	Uniformly dropped within respective cells. Macro: 25%, Pico: 75%.
Carrier frequency	2.0 GHz
Bandwidth	20 MHz
Channel model	Typical Urban (TU)
Inter-site distance	750 m
Noise power spectral density	-174 dBm/Hz
Macro BS transmit power	40 W (46 dBm)
Macro cell path loss model	128.1 + 37.6 $\log_{10} R$ (R in km)
Macro cell shadowing model	Log normal fading with std. 10 dB
Macro BS antenna gain	15 dBi
Pico BS transmit power	250 mW (24 dBm)
Pico cell path loss model	140.7 + 36.7 $\log_{10} R$ (R in km)
Pico cell shadowing model	Log normal fading with std. 6 dB
Pico BS antenna gain	5 dBi

B. Results

1) 'Out-of-Band' Systems: We first present the distribution of decision redundancy ρ for the 'out-of-band' systems in Fig.3. The redundancy is higher when pico base stations are allocated with more resource. This is intuitive because with more resource, pico base stations could deliver more contents, and thus introduce redundancy. We also observe that in more than 60% of the scheduling instances, there would be no decision redundancy, indicating that the distributed delivery decisions phase of our proposed scheduling framework works pretty well for real-world data. At the same time, decision redundancy in certain instances could be as high as 65%, indicating that decision deduplication must be more than optional.

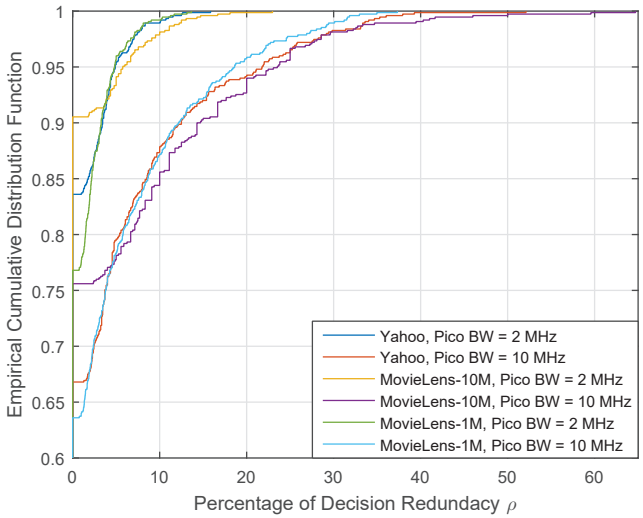

Fig. 3. Decision Redundancy for 'Out-of-Band' Systems.

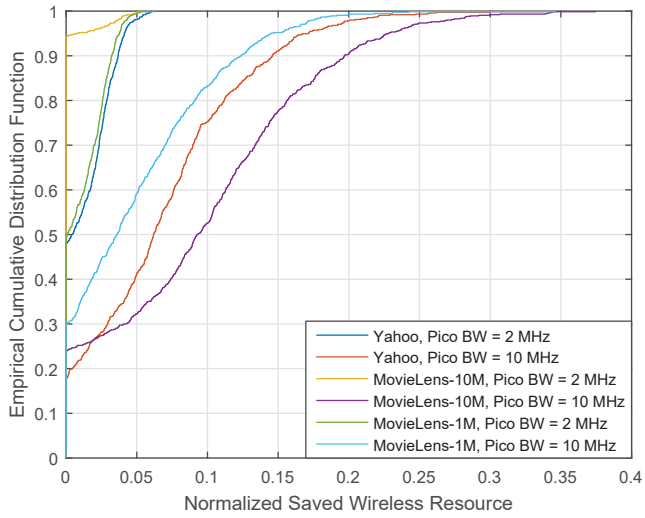

Fig. 4. Saved Wireless Resource for 'In-Band' Systems.

2) 'In-Band' Systems: The results of 'in-band' systems are presented in Fig.4, 5. In the figures, we plot the maximum resource saved by spectrum sharing, normalized with respect to overall spectrum available, among all base stations. The results are even more impressive if we choose to normalize against overall resource used (rather than available) for all base stations before resource consolidation, but it is not a fair comparison and might elude the big picture.

As the bandwidth dedicated to pico base stations increases, the resource consolidation phase saves more, illustrated in both the cumulative distribution function and the average. In certain scheduling instances, it could save as much as 38% of the overall available resource, or 76% of the wireless resource allocated to the pico base stations. We could either reapply the saved resource in the two-phase scheduling framework or release it for other purposes. Note that we are only plotting (one) maximum saving scheme. It is possible to employ non-exclusive saving schemes to save even more. Further, if we are allowed to drop certain 'difficult' users, or mathematically, users with contradictory constraints in (22), we could potentially save more wireless resource, but at the expense of losing user rewards.

We can conclude from Fig.4 that there is a nontrivial portion of situations in which the resource consolidation does not help. One extreme example is that all base stations decide to transmit to the same set of users with different contents. In this case, no spectrum sharing is obviously the best solution. In our two-phase framework, we are only aiming to reduce redundant transmissions, rather than competing transmissions. The latter has been inherently reduced in the system design because the first phase of the scheduling framework executes with dedicated wireless resource.

The P90 of runtime of feasibility tests is $623\mu s$.

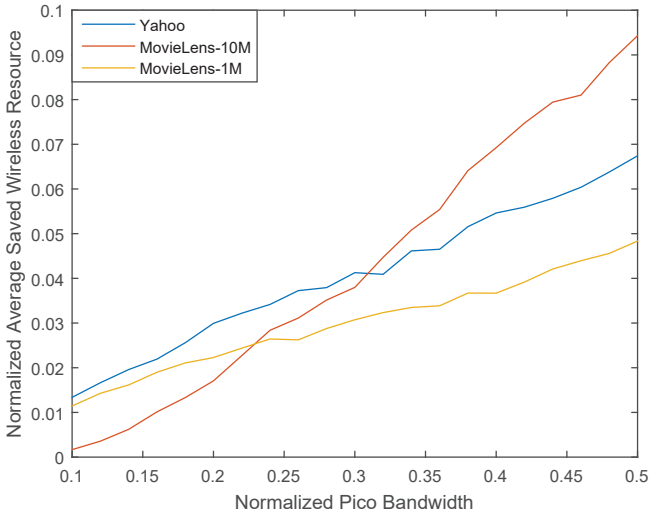

Fig. 5. Average Saved Wireless Resource for 'In-Band' Systems.

VI. Conclusions

We investigated a centralized wireless content delivery system with heterogeneous base stations, aiming to optimize overall user experience given the capacity constraints of the wireless networks. We proposed a scalable two-phase scheduling framework, consisting of: 1) distributed delivery decisions by each base station, and 2) centralized resource consolidation by the system. We tested the design using real-world rating datasets and the results indicate this novel approach is both efficient and scalable. The scheduling framework is able to incorporate both the objective of social networks to harvest more overall user rewards and the capacity constraints of the wireless networks. More importantly, this framework is scalable and requires minimal information exchange between social networks and wireless networks and among base sta-

tions. With a resource consolidation phase, we could further utilize spectrum sharing and power allocation to use less wireless resource, hence increasing system efficiency.

There are no explicit references to the hierarchy of the base stations in this paper, except for the priority order of different base stations in the Greedy Decision Deduplication Algorithm. Therefore, this work could be naturally extended to scenarios with multiple macro base stations.

Future work for this paper includes extending this framework by relaxing the commitments to the delivery decisions made by each base station in scheduling phase 1. Such commitments are expensive, because in certain situations, the edge users prohibit coordination of base stations, due to the conflicting requirements (one user's signal is another's interference).

ACKNOWLEDGMENT

Research partially supported by grants US AFOSR MURI FA9550-09-1-0538, AFOSR MURI FA-9550-10-1-0573, NSF CNS-1035655, NIST 70NANB11H148, and DARPA contract FA8750-14-C-0019 Phase 2.

REFERENCES

[1] X. Weng and J. Baras, "Joint optimization for social content delivery in wireless networks," in *IEEE ICC 2016 - Communication QoS, Reliability and Modeling Symposium (ICC'16 CQRM)*, May 2016.

[2] Y. Bao, X. Wang, S. Zhou, and Z. Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," in *Communications (APCC), 2012 18th Asia-Pacific Conference on*, pp. 566–571, IEEE, 2012.

[3] O. Shoukry, M. Abd El-Mohsen, J. Tadrous, H. El Gamal, T. ElBatt, N. Wanas, Y. Elnakieb, and M. Khairy, "Proactive scheduling for content pre-fetching in mobile networks," in *Communications (ICC), 2014 IEEE International Conference on*, pp. 2848–2854, IEEE, 2014.

[4] O. K. Shoukry and M. B. Fayek, "Evolutionary scheduler for content pre-fetching in mobile networks," in *2013 AAAI Fall Symposium Series*, 2013.

[5] F. Alotaibi, S. Hosny, H. El Gamal, and A. Eryilmaz, "A game theoretic approach to content trading in proactive wireless networks," in *Information Theory (ISIT), 2015 IEEE International Symposium on*, pp. 2216–2220, IEEE, 2015.

[6] K. Lin, C. Chen, and C. Chou, "Preference-aware content dissemination in opportunistic mobile social networks," in *INFOCOM, 2012 Proceedings IEEE*, pp. 1960–1968, 2012.

[7] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," TS 36.213, 3rd Generation Partnership Project (3GPP), 12 2015.

[8] "MovieLens dataset." http://www.grouplens.org/data/, 2003.

[9] "Yahoo! webscope dataset, ydata-ymusic-rating-study-v1_0-train." http://research.yahoo.com/Academic_Relations, 2009.

[10] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) requirements for LTE Pico Node B," TR 36.931, 3rd Generation Partnership Project (3GPP), 9 2014.