

ABSTRACT

Title of dissertation: JOINT OPTIMIZATION FOR
SOCIAL CONTENT DELIVERY
IN WIRELESS NETWORKS

Xiangnan Weng, Doctor of Philosophy, 2016

Dissertation directed by: Professor John S. Baras
Department of Electrical and Computer
Engineering

Over the last 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, little is done to consider the impacts of the wireless environment, especially the capacity constraints and changing channel.

In this dissertation, we investigate a centralized wireless content delivery system, aiming to optimize overall user experience given the capacity constraints of the wireless networks, by deciding what contents to deliver, when and how. We propose a scheduling framework that incorporates content-based reward and deliverability. Our approach utilizes the broadcast nature of wireless communication and social nature of content, by multicasting and precaching. Results indicate this novel joint optimization approach outperforms existing layered systems that separate recommendation and delivery, especially when the wireless network is operating at maximum capacity. Utilizing limited number of transmission modes, we signif-

icantly reduce the complexity of the optimization. We also introduce the design of a hybrid system to handle transmissions for both system recommended contents (‘push’) and active user requests (‘pull’).

Further, we extend the joint optimization framework to the wireless infrastructure with multiple base stations. The problem becomes much harder in that there are many more system configurations, including but not limited to power allocation and how resources are shared among the base stations (‘out-of-band’ in which base stations transmit with dedicated spectrum resources, thus no interference; and ‘in-band’ in which they share the spectrum and need to mitigate interference). We propose a scalable two-phase scheduling framework: 1) each base station obtains delivery decisions and resource allocation individually; 2) the system consolidates the decisions and allocations, reducing redundant transmissions.

Additionally, if the social network applications could provide the predictions of how the social contents disseminate, the wireless networks could schedule the transmissions accordingly and significantly improve the dissemination performance by reducing the delivery delay. We propose a novel method utilizing: 1) hybrid systems to handle active disseminating requests; and 2) predictions of dissemination dynamics from the social network applications. This method could mitigate the performance degradation for content dissemination due to wireless delivery delay.

Results indicate that our proposed system design is both efficient and easy to implement.

JOINT OPTIMIZATION FOR
SOCIAL CONTENT DELIVERY IN WIRELESS NETWORKS

by

Xiangnan Weng

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:

Professor John S. Baras, Chair/Advisor

Professor Armand M. Makowski

Professor Charalampos (Babis) Papamanthou

Professor Gang Qu

Professor Bruce L. Golden

© Copyright by
Xiangnan Weng
2016

Dedication

I dedicate this dissertation to my family and Matthew, for their constant support and unconditional love. I love you all dearly.

Acknowledgments

I owe my gratitude to all the people who have made this dissertation possible and because of whom my experience in the graduate school has been one that I will cherish forever.

First and foremost, I would like to thank my advisor, Professor John S. Baras, for giving me an invaluable opportunity to work on this challenging and extremely interesting project over the past five years. It has been a pleasure to work with and learn from such an extraordinary individual with broad understanding of systems and insightful mathematical instincts.

Thanks are due to Professor Armand M. Makowski, Professor Charalampos (Babis) Papamanthou, Professor Gang Qu and Professor Bruce L. Golden for agreeing to serve on my dissertation committee and for sparing their invaluable time reviewing the manuscript.

My colleagues at the Baras group have brightened my graduate life in many ways and deserve a special mention. Dr. Xiangyang Liu, my four-year office mate, helped me tremendously with numerous discussions and information sharing. My interactions with Dr. Kaustubh Jain, Dr. Baobing (Brian) Wang, Dr. Anup Menon, Dr. Doo-hyun Sung, Dr. Shalabh Jain, Dr. Tuan (Johnny) Ta, Peixin Gao, Yuchen Zhou, Evripidis Paraskevas, Ladan Rabieekenari, Ren Mao, Wentao Luan, Dipankar Maity have been fruitful and helpful. Also, I would like to thank Ms. Kimberly Edwards for her great administrative support.

I would also like thank my colleagues during my internships at Twitter, Google

and Facebook, in particular Ashish Virmani, Kevin Fei and Alex Nalivko. They provided wonderful industry experience and tremendous help regarding the technical soundness of this dissertation.

I owe my deepest thanks to my family - my mother and father who have always stood by me, my uncle Jason who has provided enormous substantial help. Words cannot express the gratitude I owe them. I would also like to thank Matthew Palmer for his continuous support.

This dissertation was partially supported by National Science Foundation (NSF) grants CNS-1018346, CNS-1035655, National Institute of Standards and Technology (NIST) grant 70NANB11H148, DARPA contract FA8750-15-C-0038, and US Air Force Office of Scientific Research (AFOSR) MURI grants FA9550-09-1-0538, FA9550-10-1-0573.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all!

Table of Contents

List of Figures	vii
List of Abbreviations	ix
1 Introduction	1
1.1 Related Work	5
1.2 Contributions and Organization of the Dissertation	6
2 Joint Optimization for Time-Invariant Rewards with Single Base Station	10
2.1 Overview	10
2.2 Problem Formulation	11
2.2.1 System Model	11
2.2.2 Problem Formulation	15
2.3 Scheduling Framework	16
2.3.1 Decision at Each Time Slot	16
2.3.2 Results and Analysis	17
2.3.2.1 Simulation Setup	18
2.3.2.2 Results	18
2.4 Scaling Up	20
2.5 Real-World User Rewards	23
2.5.1 Performance	24
2.5.2 Sensitivity and Saturation on Wireless Resources	25
2.6 Hybrid Systems	26
2.7 More System-Level Statistics	31
2.7.1 Resource Utilization	34
2.7.2 Contents Scheduled	34
2.8 Device-Side Improvements	35
2.9 Summary	39

3	Joint Optimization for Time-Invariant Rewards with Multiple Base Stations	41
3.1	Overview	41
3.2	Problem Formulation	42
3.2.1	General System Model	42
3.2.2	Two-Phase Scheduling	45
3.2.2.1	Delivery Decisions	45
3.2.2.2	Resource Consolidation	46
3.2.3	Decision Redundancy	47
3.2.4	Wireless Resource	47
3.3	System with ‘Out-of-Band’ Resources	48
3.4	Systems with ‘In-Band’ Resources	53
3.4.1	Coordinated Multi Point Transmission	57
3.5	Simulations and Results	58
3.5.1	Simulation Setup	58
3.5.2	Results	59
3.5.2.1	‘Out-of-Band’ Systems	59
3.5.2.2	‘In-Band’ Systems	60
3.6	Summary	61
4	Joint Optimization for Time-Variant Rewards with Single Base Station	65
4.1	Overview	65
4.2	Problem Formulation	66
4.2.1	General System Model	66
4.2.2	Content Dissemination Without Delivery Delay	69
4.3	Impact of Delivery Delay on Content Dissemination	71
4.4	Look-Ahead Scheduling	75
4.4.1	Runtime	76
4.4.2	Performance	78
4.5	Simulations and Results	81
4.5.1	What Shall We Aim For?	81
4.5.1.1	Different Scheduling Delay Models	82
4.5.1.2	Different Delay Times	83
4.5.2	Scheduling Using Hybrid Systems	85
4.5.3	Look-Ahead Scheduling	87
4.6	Summary	93
5	Conclusions and Future Work	99
5.1	Future Work	100
	Bibliography	102

List of Figures

2.1	System Model: Single Base Station	12
2.2	Comparison of Overall System Rewards ($B = 25$ MHz)	19
2.3	Comparison of Overall System Rewards ($B = 15$ MHz)	19
2.4	Comparison of User Reward ($B = 20$ MHz) with Full Random Reward Matrix	23
2.5	Comparison of User Reward ($B = 20$ MHz) for ML-1M ($\eta = 5\%$)	25
2.6	Comparison of User Reward ($B = 20$ MHz) for Yahoo! Music ($\eta = 2\%$)	26
2.7	Sensitivity of Average User Reward for ML-1M ($\eta = 5\%$)	27
2.8	Sensitivity of Fairness for ML-1M ($\eta = 5\%$)	27
2.9	Sensitivity of Average User Reward for ML-10M ($\eta = 1\%$)	28
2.10	Sensitivity of Fairness for ML-10M ($\eta = 1\%$)	28
2.11	Performance for Hybrid System (ML-100K)	32
2.12	Performance for Hybrid System (ML-1M)	33
2.13	Statistics for Wireless Resource Utilization v	35
2.14	Statistics for Scheduled Contents (Yahoo-Music, $T_d = 10$)	36
2.15	Statistics for Scheduled Contents (ML-1M, $T_d = 10$)	37
3.1	System Model: Multiple Base Stations	44
3.2	Greedy Decision Deduplication Algorithm	52
3.3	Decision Redundancy for ‘Out-of-Band’ Systems.	59
3.4	Saved Wireless Resources for ‘In-Band’ Systems.	61
3.5	Average Saved Wireless Resources for ‘In-Band’ Systems.	62
4.1	System Model: Social Content Dissemination	67
4.2	Illustration of Content Dissemination and Delivery	74
4.3	Monte Carlo Estimation of User Reward	77
4.4	Performance for Monte Carlo Estimation	79
4.5	Performance for Different Delay Models ($\pi_0 = 11s$)	83
4.6	Performance for Different Delay Models ($\pi_0 = 16s$)	84
4.7	Performance for Different Fixed Delay Times π_0	85
4.8	Social Performance for Disseminating Contents in Hybrid System	88
4.9	Delay for Disseminating Contents in Hybrid System	89

4.10	Total Transmissions per Content for Disseminating Contents in Hybrid System	90
4.11	Total Transmission per Content for Disseminating Contents in Hybrid System	91
4.12	Overall System Reward for Hybrid System	92
4.13	Prediction Precision for Hybrid System	94
4.14	Ratio of Unnecessary Transmissions for Hybrid System	95

List of Abbreviations

α	Binary decision variable for transmission
s	Wireless resource allocated
f	Reward
W	Size of content
ϕ	Binary delivery state
API	Application Program Interface
CoMP	Coordinated Multi Point
LP	Linear Programming
MIP	Mixed Integer Programming
QoS	Quality of Service
SNR	Signal-to-Noise Ratio
SINR	Signal-to-Interference-Noise Ratio

Chapter 1: Introduction

Social networks, which enable information sharing and consumption 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 additional social networks, greatly facilitate information exchange for users all across the world. Their success relies 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. Metrics of system utilities are different for different systems. For example, for ads systems, one possible reward metric is the revenue earned from displaying ads to users; for video subscription systems, the time users spent on the video; for general systems, user's

satisfaction.

In the modern era of Internet, the contents consumed most by human users are undoubtedly multimedia (namely pictures and videos). Unlike texts or numerical data, this type of contents are generally large in size, difficult to structuralize, and computationally hard to separate. For example, when dealing with texts, it is generally possible to summarize, categorize, rate, search, highlight, and segment automatically. This is because the texts follow certain grammatical and structural rules; but for pictures and videos, these operations are generally still far from practical. Therefore, content delivery systems, both wired or wireless, are facing tremendous capacity shortages, resulting in the debate of whether system operators can impose certain types of management policies, or more famously, network neutrality as in [1].

Even more challengingly, following the rapid development of the smart phone, most (if not all) social networks are reporting that their users are dominantly mobile, 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 resources 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. This is because the wireless network is a shared-medium communication and becomes congested as it approaches its capacity limits. As a result, quality of service degrades drastically. Hence, we are motivated to provide services in these congested 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.

For the longest time, layered models (e.g. OSI 7-layer model in Table 1.1) have dominated the system design in practice. Undeniably, such models are easy to implement, test and maintain for engineers, because every application program interface (API) is explicitly defined. According to these models, each layer is responsible for their own functions and it only directly invoked APIs of the layer immediately beneath, and provided facility APIs for use by the layer immediately above.

As [3] suggested, such layering is in fact a decomposition of optimization for the complex network architecture. Mathematically, it does not provide optimal performance for the overall system, even if each layer is optimized. Therefore, the layered solutions are inherently bad for overall system performance, especially when the network resources are insufficient.

Existing layered solutions fail to utilize the social nature of content consumption or the broadcast nature of wireless networks. They include two mostly isolated stages:

1. The social network application chooses the best recommended contents for

users, regardless of the size of the contents and the traffic load the users observe;

2. The wireless network allocates resources for each transmission on-demand in a unicast way, unable to group the requests for the same contents or to defer transmissions when the channel is not good.

As stated before, this unicast on-demand approach is viable only when the wireless resources are sufficient and/or when users have absolutely nothing in common. However, if the wireless network is congested, this approach is extremely inefficient.

Our proposed solution tackles the weaknesses of the existing solutions via the following major improvements:

1. Multicast and precache: in light of the preliminary research and industrial practice of social network applications, we observe 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 design that multicasts and precaches contents to groups of users to reduce redundant transmissions.
2. Real-time scheduling to optimize overall performance: the system schedules contents based on real-time channel information. Under congested circumstances, the system delivers contents that maximize overall user experience across the system, but not necessarily optimal for each individual user. In other words, we are more inclined to deliver contents in a collectively optimal way when the wireless networks are congested.

This novel approach requires information from both social and wireless networks to jointly optimize system decisions. However, the APIs proposed in our approach are straightforward and easy to implement. It does not require much modifications of both layers and the information exposed is extremely limited.

1.1 Related Work

Certain preliminary work has been done concerning the joint user experience optimization incorporating social and wireless networks.

A content-based hybrid system was proposed in [4] to handle two types of content-based transmission schemes for satellite communication: unicast ('pull') and multicast ('push'). At the time of the article, it was difficult to predict ahead of time how contents are perceived by the users, thus the scheduling system discussed was essentially reactive to the user requests. However, with the rapid developments of big data, we are now capable to make much more precise prediction regarding content consumption. Therefore, we can design the system to schedule in a proactive manner with the help of precache. We would also want to point out that all wireless transmissions are fundamentally multicast because users with proper clients (including devices, credentials, etc.) could receive the transmission.

In [5], 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 utilize the fact that the contents/videos themselves are also subject to scheduling in social network applications. It is a common practice for social network providers to

shape the demand for contents, though from a different design rationale (increasing user engagements).

In [6] and [7], 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 [8], 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. Additionally, the solution would not scale up well.

In [9], 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 resources in different wireless conditions.

1.2 Contributions and Organization of the Dissertation

In this dissertation, we discussed the content delivery problem with capacity constraints and changing wireless channel in three different scenarios:

1. In Chapter 2, we focus on the basic setup: single base station with time-

invariant reward. This is the basic problem and foundation for later discussions, in that it confirms that our approach is both effective and implementable. Results indicate that myopic optimization is sufficient to obtain high-quality scheduling decisions (what contents to deliver to which users, and how many wireless resources to allocate). We further reduce the complexity of the problem by utilizing the finite multicast modes. After the reduction, we could solve the optimization for optimal almost real-time. To incorporate different types of delivery (system recommendation or active user request), we introduce parameters in the hybrid systems to balance between optimizing overall system reward delivered and satisfying active user requests. Part of this chapter will publish in [10].

2. In Chapter 3, we extend the system design in Chapter 2 to multiple base stations. The wireless infrastructure in this chapter consists of perfectly synchronized heterogeneous base stations, with macro base stations providing signal coverage and pico base stations providing capacity enhancement. We propose a two-phase scheduling solution that is both efficient and scalable. Part of this chapter will publish in [11].
3. In Chapter 4, we discuss how to improve the system design with regard to social dissemination, i.e. time-variant reward. In order to reduce redundant transmissions, we have to utilize predictions of content dissemination obtained from social networks. Results indicate that predictions, even coarse ones, would significantly improve the performance for disseminating contents. Further, the

design of hybrid systems in [2.6](#) provides additional improvements.

We summarize our proposed system design and results in [Chapter 5](#), as well as possible future work.

Table 1.1: OSI 7-Layer Model [2]

Layer		Protocol Data Unit	Function	Examples
Host Layer	7. Application	Data	High-level APIs, including resource sharing, remote file access, directory services and virtual terminals.	
	6. Presentation		Translation of data between a networking service and an application; including character encoding, data compression and encryption/decryption.	
	5. Session		Managing communication sessions, i.e. continuous exchange of information in the form of multiple back-and-forth transmissions between two nodes.	RPC, SCP, NFS, PAP, TLS, FTP, HTTP,HTTPS, SMTP, SSH, Telnet
	4. Transport	Segment(TCP)/ Datagram(UDP)	Reliable transmission of data segments between points on a network, including segmentation, acknowledgement and multiplexing.	NBF, TCP, UDP
Media Layer	3. Network	Packet	Structuring and managing a multi-node network, including addressing, routing and traffic control.	AppleTalk, ICMP, IPsec, IPv4, IPv6
	2. Data Link	Frame	Reliable transmission of data frames between two nodes connected by a physical layer.	IEEE 802.2, L2TP, LLDP, MAC, PPP
	1. Physical	Bit	Transmission and reception of raw bit streams over a physical medium.	DOCSIS, DSL, Ethernet physical layer, ISDN, USB

Chapter 2: Joint Optimization for Time-Invariant Rewards with Single Base Station

2.1 Overview

In this chapter, we introduce the basic setup of the content delivery problem with capacity constraints in a system with centralized wireless infrastructure. We need to decide what contents to transmit to which users and how to transmit.

Traditional systems include two separated stages:

1. The social network applications choose the best recommended contents for users;
2. The wireless network allocates resources for transmissions of packets.

As stated before, this unicast approach provides optimal results only when the wireless resources are sufficient and/or when users have absolutely nothing in common. However, if the wireless network is congested, this approach fails to work. Worse still, if the users or social network applications cancel or revise the delivery requests (usually due to impatience of long delay), the system would suffer greatly in terms of user experience, yet the network performance metrics (usually throughput) will falsely indicate that the system works great. If we cannot guarantee that a con-

content is delivered to users in whole, it is very hard to evaluate the overall system performance.

We propose a joint optimization framework for delivery requests that are based on recommender system to better utilize the wireless resources and schedule the wireless transmissions to obtain optimal user experience. Further, we embrace the idea of hybrid systems in [4] and extend this framework to handle two types of transmissions: system recommended (‘push’) and active user request (‘pull’).

We introduce our system model and evaluation framework in Section 2.2. The performance with and without look-ahead at the wireless layer is analyzed in 2.3.2. In Section 2.4, we propose a scalable solution by utilizing limited multicast modes and eliminating content limits for users. Then, we analyze its performance and sensitivity against real-world data in Section 2.5. Extension to hybrid systems handling different types of delivery is discussed in Section 2.6. Finally, we present conclusions in Section 2.9.

2.2 Problem Formulation

2.2.1 System Model

We consider a centralized system that both selects contents for users according to rewards given wireless capacity constraints and delivers the contents to users via a wireless network, as illustrated in Fig.2.1. All the users are served using the same base station. We assume the bandwidth of the wired connections between the base station and content server(s) is sufficient enough that the base station could access

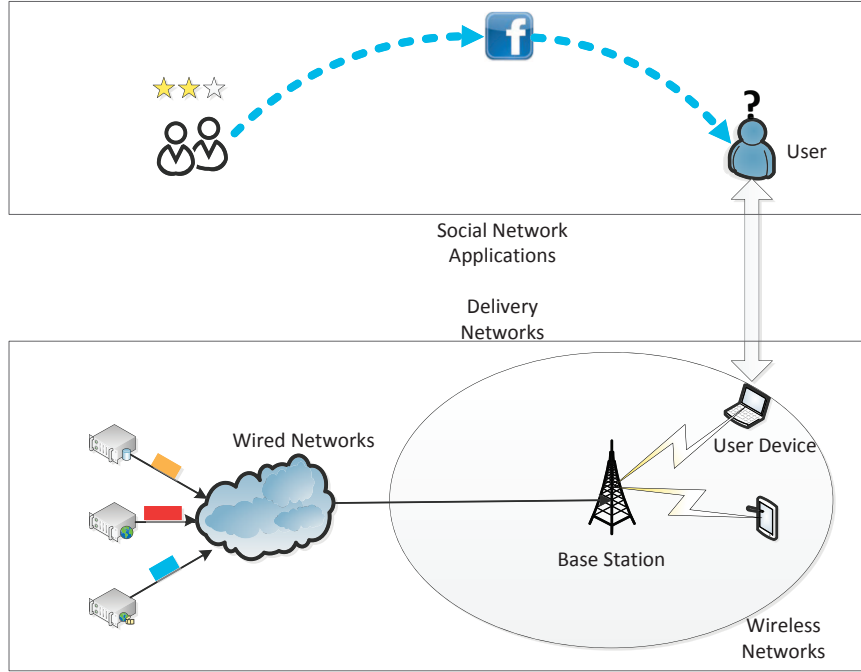


Figure 2.1: System Model: The social network applications are responsible to select contents for users according to their profiles while the delivery networks (including wired backbone and wireless last-hop networks) are responsible to deliver to user devices.

contents as if they are stored locally. This assumption is valid in practice because base stations are generally connected to the Internet via fiber optic cables. Trivially, with modern chips, the storage on user devices is large enough to precache all the contents scheduled for delivery.

(a) The normalized reward (or reward for simplicity) earned from delivering content j to user i is denoted by f_{ij} , $0 \leq f_{ij} \leq 1$. This number is obtained from social network applications via various big data techniques. The reward is earned only after the first successful delivery of the whole content; partial or repeated delivery does not earn (additional) reward(s). It is important to note that the reward matrix

might not be fully filled due to either lack of interest or sufficient data, in which case we simply denote them as no rewards ($f_{ij} = 0$).

(b) The system is time-slotted with slot length T and the scheduling horizon is T_H time slots. This is consistent with modern cellular systems (GSM, UMTS, LTE, LTE-Advanced), which are all time-slot based to simplify overhead of control plane. Generally, the channel state is considered to remain unchanged within the slot. We consider decisions for all the users served by the base station at the beginning of each time slot t , with the bandwidth of the wireless network B .

(c) The wireless channel is slow fading (i.e. it remains unchanged during each scheduling time slot, as stated in (b)) and is described by signal-to-interference-noise ratio $\text{SINR}_i^{(t)}$ for user i at time slot t , where $P_i^{(t)}, I_i^{(t)}, N_i^{(t)}$ are received power strength, interference power and noise power respectively.

$$\text{SINR}_i^{(t)} = \frac{P_i^{(t)}}{I_i^{(t)} + N_i^{(t)}} \quad (2.1)$$

To reduce system complexity, contents shall be transmitted within one time slot; otherwise, management of multicast groups is too complicated to implement, given the changing channel state between slots.

(d) For simplicity, we assume contents could be reliably delivered with given transmission rate not exceeding the Shannon limit. Therefore, given bandwidth B and slot length T , the maximum bits deliverable to a user with SINR in the time slot is denoted by W and it follows Shannon's law:

$$\mathcal{R}(\text{SINR}) = \log_2(1 + \text{SINR}) \quad (2.2)$$

$$W = B \cdot T \cdot \mathcal{R}(\text{SINR}) \quad (2.3)$$

Table 2.1 summarizes definitions of parameters.

Table 2.1: Summary of Variables

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

2.2.2 Problem Formulation

We formulate the content delivery problem as a mixed integer programming (MIP) problem.

$$\begin{aligned}
& \underset{\alpha_{ij}^{(t)}, s_j^{(t)}}{\text{maximize}} && \sum_{t=1}^{T_H} \sum_{i,j} \alpha_{ij}^{(t)} \cdot f_{ij} \\
& \text{subject to} && \alpha_{ij}^{(t)} \in \{0, 1\} && \forall i, j, t \\
& && \sum_{t=1}^{T_H} \alpha_{ij}^{(t)} \leq 1 && \forall i, j \\
& && \sum_j \alpha_{ij}^{(t)} \leq N_0 && \forall i, t \\
& && \alpha_{ij}^{(t)} \cdot W_j \leq s_j^{(t)} \cdot \mathcal{R}(\text{SINR}_i^{(t)}) && \forall i, j \\
& && \sum_j s_j^{(t)} \leq B \cdot T && \forall t \\
& && s_j^{(t)} \geq 0 && \forall j, t
\end{aligned} \tag{2.4}$$

The objective is to maximize the overall system reward, with the constraints of:

1. delivery decisions are binary;
2. each content is delivered to a user at most once;
3. at most N_0 content is delivered to each user in one time slot;
4. quality of service (QoS) is satisfied, i.e. content size (in bits) shall not exceed the channel capacity, given the resource and the state of the wireless channels;
5. resources allocated does not exceed system capacity.

2.3 Scheduling Framework

2.3.1 Decision at Each Time Slot

For general cases, the MIP problem (2.4) is NP-hard and solving it takes exponential time. Moreover, we care more about the online version of the optimization, i.e. at time slot t_0 , we only have access to current and historic channel information $\{\text{SINR}_i^{(t)}, \forall i; t = 1, \dots, t_0\}$. Solving the offline version is of little practical use for scheduling.

Therefore, at each time slot t_0 we introduce the T_L -step lookahead (no lookahead with $T_L = 0$) version of the MIP problem given the wireless channel profile of each user. Scheduling decisions $\{\alpha_{ij}^{(t_0)}\}, \{s_j^{(t_0)}\}$ at each time slot are obtained by solving this lookahead version (2.5). All the auxiliary symbols with tilde are lookahead version of the corresponding parameters in the MIP problem (2.4).

$$\begin{aligned}
& \underset{\tilde{\alpha}_{ij}^{(t)}, \tilde{s}_j^{(t)}}{\text{maximize}} && \sum_{t=t_0}^{t_0+T_L} \sum_{i,j} \tilde{\alpha}_{ij}^{(t)} \cdot \tilde{f}_{ij}^{(t_0)} \\
& \text{subject to} && \tilde{\alpha}_{ij}^{(t)} \in \{0, 1\} && \forall i, j, t \\
& && \sum_{t=t_0}^{t_0+T_L} \tilde{\alpha}_{ij}^{(t)} \leq 1 && \forall i, j \\
& && \sum_j \tilde{\alpha}_{ij}^{(t)} \leq N_0 && \forall i, t \\
& && \tilde{\alpha}_{ij}^{(t)} \cdot W_j \leq \tilde{s}_j^{(t)} \cdot \mathcal{R}(\widetilde{\text{SINR}_i^{(t)}}) && \forall i, j \\
& && \sum_j \tilde{s}_j^{(t)} \leq B \cdot T && \forall t \\
& && \tilde{s}_j^{(t)} \geq 0 && \forall j, t
\end{aligned} \tag{2.5}$$

We rewrite the single transmission constraints in (2.4) by changing reward values at each time slot. Initially,

$$\tilde{f}_{ij}^{(1)} = f_{ij}, \forall i, j \quad (2.6)$$

If content j was successfully transmitted to user i at slot t_0 , the reward drops to zero to avoid future transmission(s).

$$\tilde{f}_{ij}^{(t_0+1)} = \tilde{f}_{ij}^{(t_0)} \cdot \left(1 - \alpha_{ij}^{(t_0)}\right), \forall i, j \quad (2.7)$$

We predict SINR in future scheduling slots for each user using historic channel information:

$$\widetilde{\text{SINR}}_i^{(t)} = \begin{cases} \text{SINR}_i^{(t_0)} & t = t_0 \\ \phi(\text{SINR}_i^{(1)}, \dots, \text{SINR}_i^{(t_0)}) & t = t_0 + 1, \dots, t_0 + T_L \end{cases} \quad (2.8)$$

The actual decisions (what contents to deliver $\{\alpha_{ij}^{(t_0)}\}$ and how to allocate wireless resources $\{s_j^{(t_0)}\}$) taken at time slot t_0 are the decisions at the first slot obtained from the solution of the lookahead version of optimization (2.5). It is trivial to prove that the choice of decisions satisfies all the constraints in MIP (2.4).

$$\alpha_{ij}^{(t_0)} = \tilde{\alpha}_{ij}^{(t_0)}, \forall i, j \quad (2.9)$$

$$s_j^{(t_0)} = \tilde{s}_j^{(t_0)}, \forall j \quad (2.10)$$

2.3.2 Results and Analysis

In this part, we compare our proposed scheduling system with the traditional layered design, where the social network applications attempt to provide users with the most rewarding contents regardless of the status of the wireless networks, while

the wireless networks attempt to deliver the contents with best effort regardless of the rewards of contents.

2.3.2.1 Simulation Setup

There are $M = 30$ users and $N = 20$ contents in the system. Reward values f_{ij} 's are independent and uniformly distributed in $[0, 1]$. Note that this is already the largest scale within which we could obtain the optimal solution to compare the results. Content size is independent and uniformly distributed in $[10, 20]$ Mbits. The scheduling time slot has length of $T = 1s$ and the scheduling horizon is $T_H = 10$. Each user could receive at most $N_0 = 1$ content in every scheduling time slot. System level parameters are shown in Table 2.2 [12].

2.3.2.2 Results

We simulate various instances against different bandwidth, as shown in Fig. 2.2, 2.3. Clearly, joint optimization outperforms the traditional system by a wide margin. The joint optimization gain increases as the available wireless resources decrease.

Surprisingly, one-step lookahead scheduling (whether we use mean or max function as SINR prediction function) does not outperform no-lookahead scheduling. Due to the extra computation cost (it takes at least 10 times of computation time to obtain optimization results), it is sufficient to schedule without looking ahead.

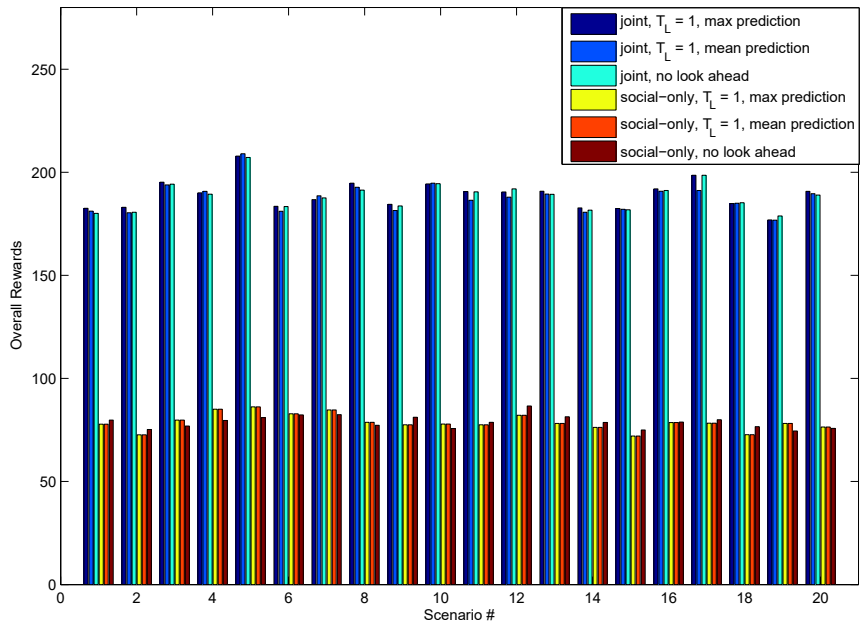


Figure 2.2: Comparison of Overall System Rewards ($B = 25$ MHz)

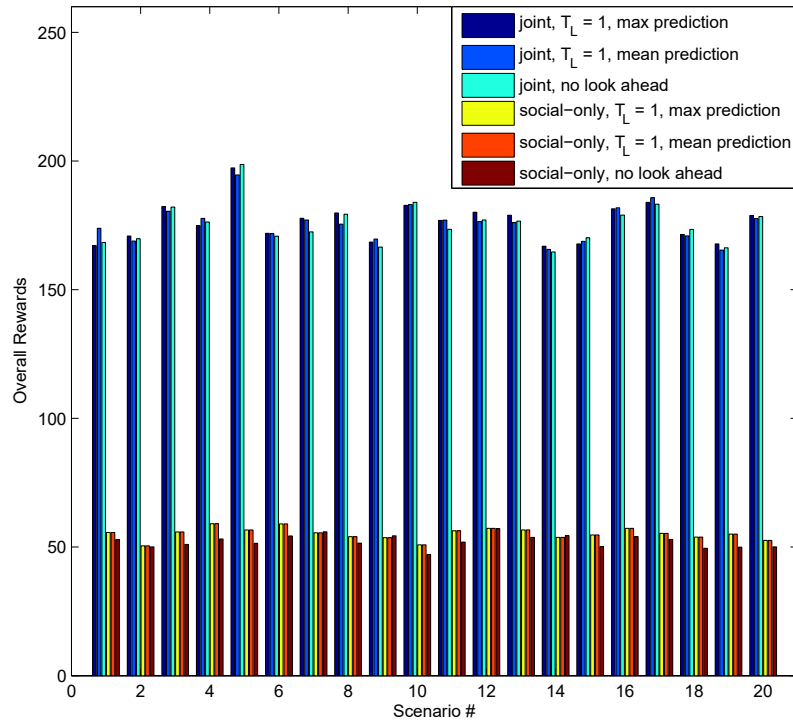


Figure 2.3: Comparison of Overall System Rewards ($B = 15$ MHz)

Table 2.2: System Level Simulation Parameters

Simulation Parameter	Value
UE distribution	UEs dropped with uniform density within the macro coverage area.
Carrier frequency	2.0 GHz
Channel model	Typical Urban (TU)
Inter-site distance	1500 m
Noise power spectral density	-174 dBm/Hz
Macro BS transmit power	40 W (46 dBm)
Macrocell path loss model	$128.1 + 37.6 \log_{10} R$ (R in km)
Macrocell shadowing model	Log normal fading with std. 10 dB
Macro BS antenna gain	15 dBi

2.4 Scaling Up

For real-world systems, the available multicast transmission modes are limited and pre-determined. Assume the system has K available transmission modes, and the associated data transmission rates and minimum channel quality requirements are R_k and SINR_k^{th} , respectively. More formally, we substitute Shannon's Law in the data rate function (2.2) with a step function (slightly abusing notation $R_0 = 0$), denoting K available modes with the order convention $R_{k-1} < R_k, \text{SINR}_{k-1}^{th} <$

$\text{SINR}_k^{th}, \forall k = 1, \dots, K$:

$$\mathcal{R}(\text{SINR}) = \sum_{k=1}^K (R_k - R_{k-1}) \cdot u(\text{SINR} - \text{SINR}_k^{th}) \quad (2.11)$$

Therefore, if we give up the constraints of the maximum number of contents that are allowed to transmit to a user, we could further reduce the scheduling problem to deciding on which wireless transmission mode we use to transmit what contents. In this way, the complexity of the problem is significantly reduced. It would only rely on the number of transmission modes K and contents N , dropping the number of users M .

For this complexity-reduction version, at slot t_0 , the reward $\hat{f}_{ij}^{(t_0)}$ for transmitting content j in wireless mode k is induced from summing up all the rewards of the users that meet the quality of service requirement of this mode.

$$\hat{f}_{jk}^{(t_0)} = \sum_{i: \text{SINR}_i^{(t_0)} \geq \text{SINR}_k^{th}} \tilde{f}_{ij}^{(t_0)} \quad (2.12)$$

The new scheduling formulation is thus:

$$\begin{aligned} & \underset{\hat{\alpha}_{jk}^{(t_0)}, s_j^{(t_0)}}{\text{maximize}} && \sum_{j,k} \hat{\alpha}_{jk}^{(t_0)} \cdot \hat{f}_{jk}^{(t_0)} \\ & \text{subject to} && \hat{\alpha}_{jk}^{(t_0)} \in \{0, 1\} && \forall j, k \\ & && \sum_{k=1}^K \hat{\alpha}_{jk}^{(t_0)} \leq 1 && \forall j \\ & && \hat{\alpha}_{jk}^{(t_0)} \cdot W_j \leq s_j^{(t_0)} \cdot R_k && \forall j, k \\ & && \sum_j s_j^{(t_0)} \leq B \cdot T \\ & && s_j^{(t_0)} \geq 0 && \forall j \end{aligned} \quad (2.13)$$

The reverse mapping between system decision and decisions for each individual user is:

$$\alpha_{ij}^{(t_0)} = \sum_{\substack{i,j: \hat{f}_{ij}^{(t_0)} > 0 \\ k: \text{SINR}_k^{th} \leq \text{SINR}_i^{(t_0)}}} \hat{\alpha}_{jk}^{(t_0)} \quad (2.14)$$

Apparently, this user-aggregated version of scheduling framework only relies on N (the number of contents to be scheduled) and K (the number of multicast transmission modes in the system), thus it is robust against increase in the number of users in the system.

However, when we relax the constraints on the number of contents delivered to each individual user, it is important to consider fairness among users. The system is now evaluated on both average overall reward delivered to each user and fairness among users.

For each user, the overall reward delivered during the scheduling horizon $[0, T_H]$ is calculated as:

$$u_i = \sum_{t=1}^{T_H} \sum_j \alpha_{ij}^{(t)} f_{ij} \quad (2.15)$$

Denote the user average reward and its variation as \bar{u}, σ_u^2 respectively.

$$\bar{u} = \frac{1}{M} \sum_{i=1}^M u_i \quad (2.16)$$

$$\sigma_u^2 = \frac{1}{M} \sum_{i=1}^M (u_i - \bar{u})^2 \quad (2.17)$$

We use Jain's fairness index [13] as a fairness metric.

$$J(\vec{u}) = \frac{\bar{u}^2}{\bar{u}^2 + \sigma_u^2} \quad (2.18)$$

where $\vec{u} = (u_1, \dots, u_M)$. The larger the fairness index, the more fare for a scheduling result.

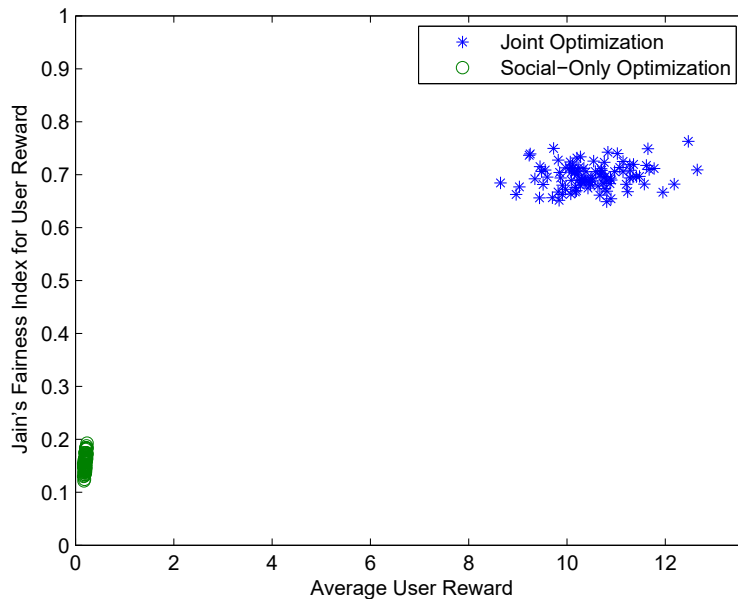


Figure 2.4: Comparison of User Reward ($B = 20$ MHz) with Full Random Reward Matrix

Fig.2.4 shows the performance comparison of our proposed scheduling framework and traditional layered design, on a full random reward matrix. Each point in the graph represents one simulation instance. Clearly, the joint optimization framework is better in terms of average overall rewards delivered per user and Jain’s fairness metric.

2.5 Real-World User Rewards

In real-world applications, the reward matrix $F = [f_{ij}]$ is usually sparse due to multiple factors: (1) naturally, users exhibit diverse interests towards different contents; (2) technically, it is extremely difficult, if not impossible, to gather enough information about users, in order to obtain accurate and comprehensive prediction.

In our system, we do not distinguish between unknown reward or lack of interests, by assigning zero reward value, i.e. $f_{ij} = 0$, to avoid such transmissions (proof is trivial due to the formulation of the optimization problem).

We further define the sparseness of the reward matrix as follows:

$$\eta = \frac{|\{f_{ij} : f_{ij} > 0\}|}{M \cdot N} \quad (2.19)$$

Obviously, for the full random reward matrix, $\mathbb{P}[\eta = 1] = 1$.

In light of this observation, we need to examine the performance of our scheduling framework on sparse real-world data sets.

Our empirical analysis is based on two data sets:

1. MovieLens [14]: users' ratings of different movies. This data set has two sub data sets. (i) ML-1M, with sparseness of $\eta = 5\%$, number of total users 6000 and number of total contents 4000. (ii) ML-10M, with sparseness of $\eta = 1\%$, number of total users 72,000 and number of total contents 10,000.
2. The Yahoo! Music ratings for User-Selected and Randomly Selected Songs, version 1.0 data set, which is available through the Yahoo! Webscope data sharing program. This data set has sparseness of $\eta = 2\%$, number of total users 15,400 and number of total contents 1000.

2.5.1 Performance

From Fig.2.5 and Fig.2.6, it is clear that our scheduling framework still works well for real-world data sets. However, as sparsity of reward matrix increases ($\eta \downarrow$),

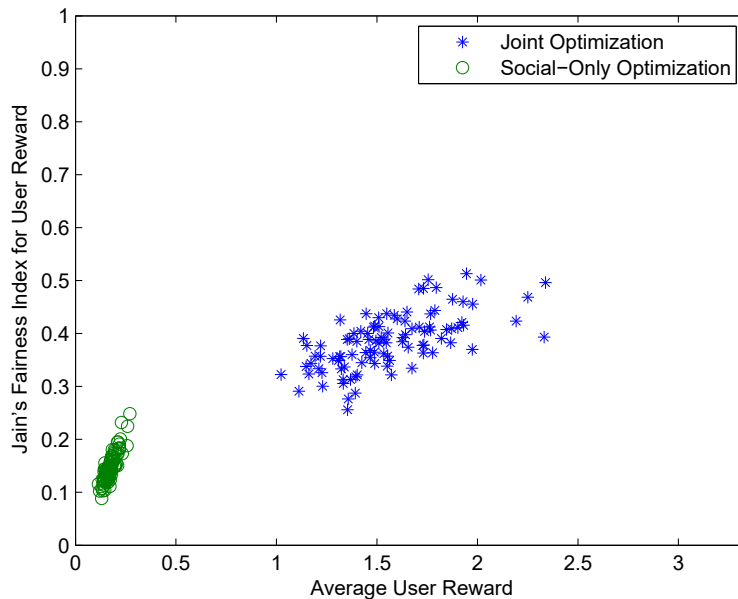


Figure 2.5: Comparison of User Reward ($B = 20$ MHz) for ML-1M ($\eta = 5\%$)

the performance gain compared to the existing system diminishes. This result is intuitive, as multicast works best only when the number of users that are interested in the same contents is large enough.

2.5.2 Sensitivity and Saturation on Wireless Resources

In this part, we demonstrate the performance sensitivity with respect to available wireless resources (in our system, bandwidth B). Intuitively, when the wireless resources are sufficient, the performance solely relies on the numerical reward values obtained from the social network, as we could deliver all the contents. However, if the wireless resources are insufficient, the performance of the system shall be reduced due to its incapability to deliver the contents.

We plot the user average reward and fairness index against bandwidth in

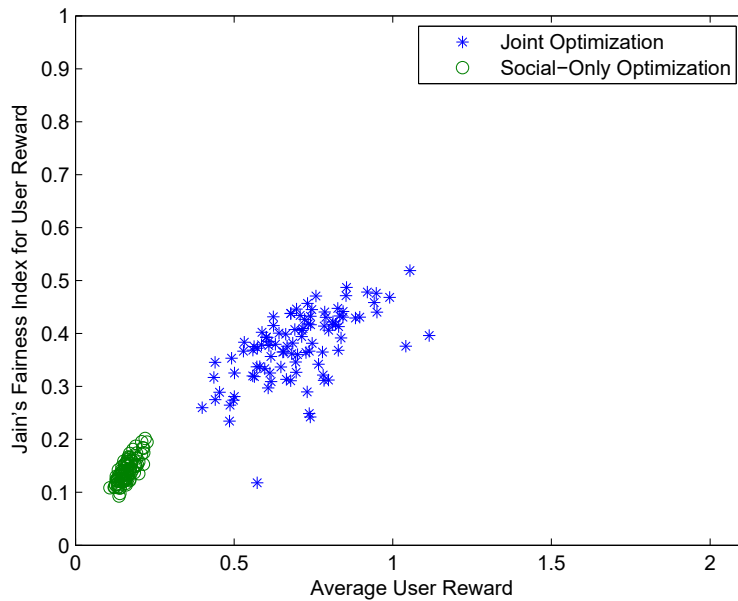


Figure 2.6: Comparison of User Reward ($B = 20$ MHz) for Yahoo! Music ($\eta = 2\%$)

Fig.2.7,2.8 respectively, to analyze the benefits of adding more wireless resources to the system. Fig.2.9,2.10 shows the performance saturation for our joint optimization approach when the delivery demand is low compared to the bandwidth.

2.6 Hybrid Systems

Until now, we have been discussing systems that involve no active user requests for delivery. This assumption is generally valid for a pure recommender system, in which users passively consume contents delivered to their devices. We define this delivery method as the ‘push’ operation. However, practical systems are also required to handle active user requests within a designated timeframe, which we define as the ‘pull’ operation. The hybrid systems shall handle both ‘push’ and ‘pull’ functions smoothly.

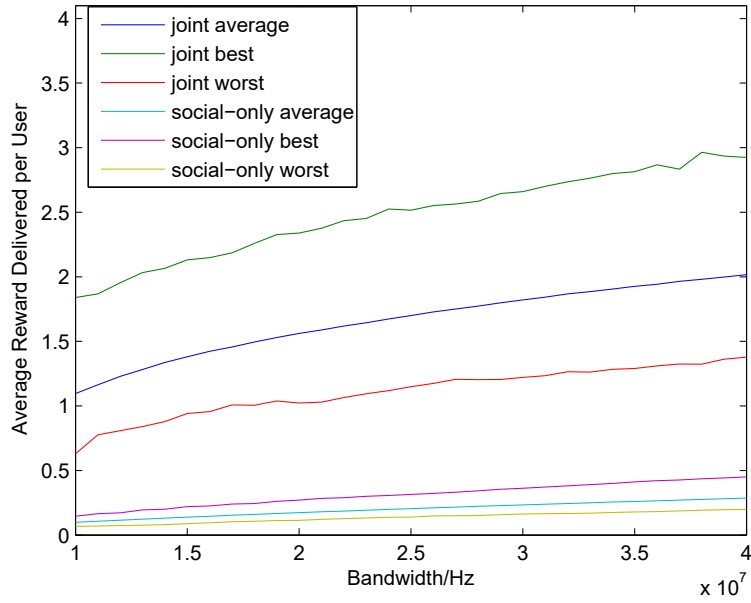


Figure 2.7: Sensitivity of Average User Reward for ML-1M ($\eta = 5\%$)

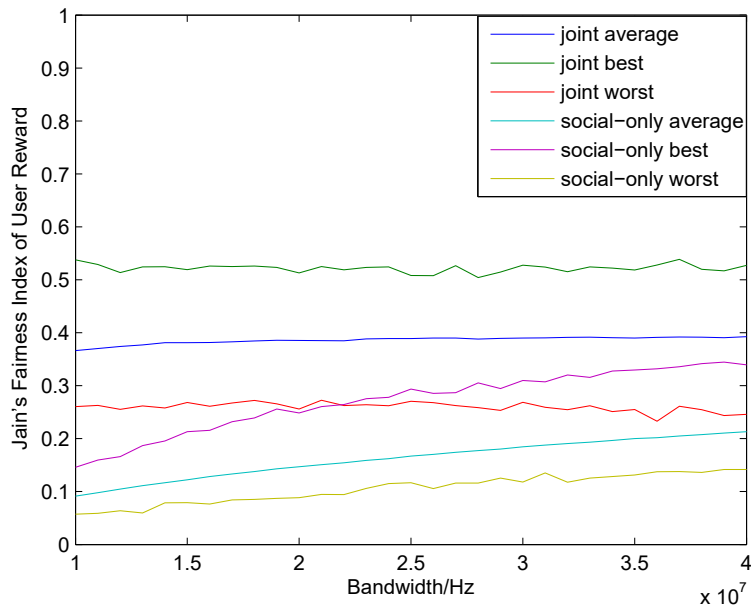


Figure 2.8: Sensitivity of Fairness for ML-1M ($\eta = 5\%$)

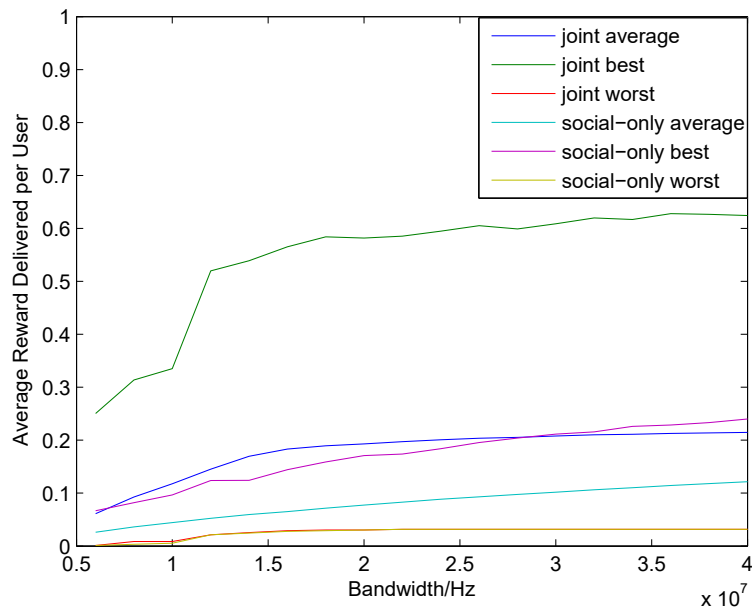


Figure 2.9: Sensitivity of Average User Reward for ML-10M ($\eta = 1\%$)

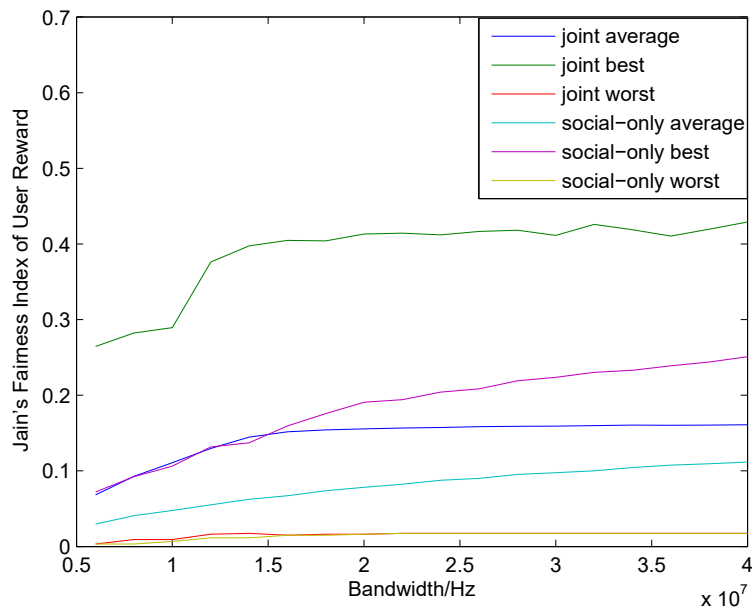


Figure 2.10: Sensitivity of Fairness for ML-10M ($\eta = 1\%$)

Now we formally define the ‘pull’ operation. The user request q could be described using the quadruple $q = (t_0, i, j, t_{\max}) \in \mathcal{N}^4$, denoting the request originates at slot t_0 and demands the system to deliver content j to user i no later than t_{\max} (deadline). Obviously, in real systems, $t_0 \leq t_{\max}$. We further assert that all requests are valid, assuming all requests fulfilled in the past shall not occur again. This assumption could easily be implemented by serving the transmitted content(s) from the cache of the user’s device. Denote the set of new and active requests at slot t as $Q_t^{\text{new}}, Q_t^{\text{active}}$, respectively, $Q_t^{\text{served}} \subseteq Q_t^{\text{active}}$ as the set of requests served and $Q_t^{\text{expired}} \subseteq Q_t^{\text{active}}$ as the set of requests expired at the end of the slot. Then,

$$Q_t^{\text{served}} = \{q \in Q_t^{\text{active}} : \alpha_{q_2 q_3}^{(t)} = 1\} \quad (2.20)$$

$$Q_t^{\text{expired}} = \{q \in Q_t^{\text{active}} : q_4 = t\} \quad (2.21)$$

$$Q_{t+1}^{\text{active}} = (Q_{t+1}^{\text{new}} \cup Q_t^{\text{active}}) \setminus (Q_t^{\text{served}} \cup Q_t^{\text{expired}}) \quad (2.22)$$

with the trivial convention $Q_0^{\text{active}} = Q_0^{\text{served}} = Q_0^{\text{expired}} = \phi$.

The validity of requests is described as: $\forall q \in Q_{t_0}^{\text{new}}$, we have $q_1 = t_0$ and $\nexists q' \in \bigcup_{t=1}^{t_0-1} Q_t^{\text{served}}$, such that $q_2 = q'_2, q_3 = q'_3$.

The objective for serving the ‘pull’ operation is to serve as many requests as possible. We could write the optimization problem for ‘pull’ system as:

$$\begin{aligned} & \underset{Q_t^{\text{served}}}{\text{maximize}} && \sum_{t=1}^{T_H} |Q_t^{\text{served}}| \\ & \text{subject to} && W_{q_3} \leq s_{q_3}^{(t)} \cdot \mathcal{R} \left(\text{SINR}_{q_2}^{(t)} \right) \quad \forall q \in Q_t^{\text{served}} \\ & && \sum_j s_j^{(t)} \leq B \cdot T \quad \forall t \\ & && s_j^{(t)} \geq 0 \quad \forall j, t \end{aligned} \quad (2.23)$$

To integrate this ‘pull’ system with the existing ‘push’ system in Section 2.4, we add additional reward(s) for requests with respect to their expiration time. Reward transition in (2.7) is rewritten in part for the ‘pull’ request as:

$$\tilde{f}_{ij}^{(t_0+1)} = \begin{cases} f_{ij} \cdot (1 - \alpha_{ij}^{(t_0)}) & \exists(*, i, j, t_0) \in Q_{t_0}^{\text{expired}} \\ \left(\tilde{f}_{ij}^{(t_0)} + \lambda \Gamma(i, j, t_0) \right) \cdot (1 - \alpha_{ij}^{(t_0)}) & \text{otherwise} \end{cases} \quad (2.24)$$

where the incentive function Γ is exclusively awarded to unexpired active user requests:

$$\Gamma(i, j, t_0) = \begin{cases} \gamma(t_d - t_0) & \exists(*, i, j, t_d) \in Q_{t_0}^{\text{active}}, t_0 < t_d \\ 0 & \text{otherwise} \end{cases} \quad (2.25)$$

To balance between the two types of operations, we introduce a ‘push’ weight coefficient $\lambda \in \mathbb{R}_+$ to configure the bias of hybrid system with respect to user requests, which reduces to pure ‘push’ system (no active user requests) when $\lambda = 0$, or pure ‘pull’ system (best-effort to serve all active user requests) when $\lambda = \infty$.

In this way, we seamlessly integrate the recommender system and active user requests together to form a hybrid delivery system that could deal with both types of content delivery. Essentially, if we did not yet schedule transmission for the request(s), we shall add additional reward(s) as time goes by. In doing so, we steer the system in the direction to accomplish such requests before deadlines for the active user request(s).

To capture the urgency when approaching the request deadline, we could wisely choose the γ function to properly prioritize these requests. One possible function is:

$$\gamma(t) = \frac{1}{t} \quad (2.26)$$

The maximum additional award for a request is thus

$$\Gamma_{\max} = \sum_{t=1}^{T_d} \frac{1}{t} \quad (2.27)$$

In hybrid systems, we are evaluating the performance of both the overall user rewards delivered by the ‘push’ operation and total served requests coming from the ‘pull’ operation.

In the simulation, new user requests are generated at the beginning of each scheduling time slot with same expiration time T_d .

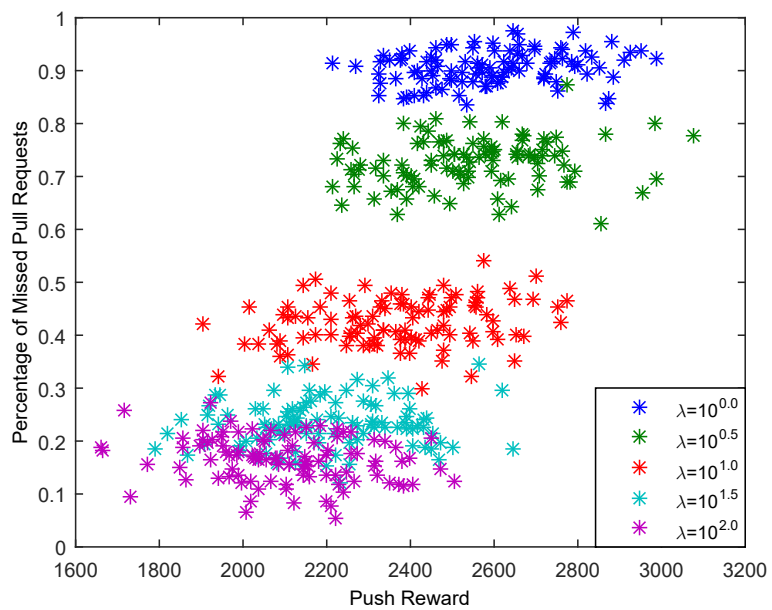
$$q_4 = q_1 + T_d, \forall q \quad (2.28)$$

The number of requests per slot follows independent and identical uniform distribution $U[0, 3]$. The requests at slot t_0 are uniformly selected in random from an unfulfilled set of user-content pairs $\{(i, j) : \tilde{f}_{ij}^{(t_0)} > 0\}$ so that the transmission has not been scheduled before slot t_0 .

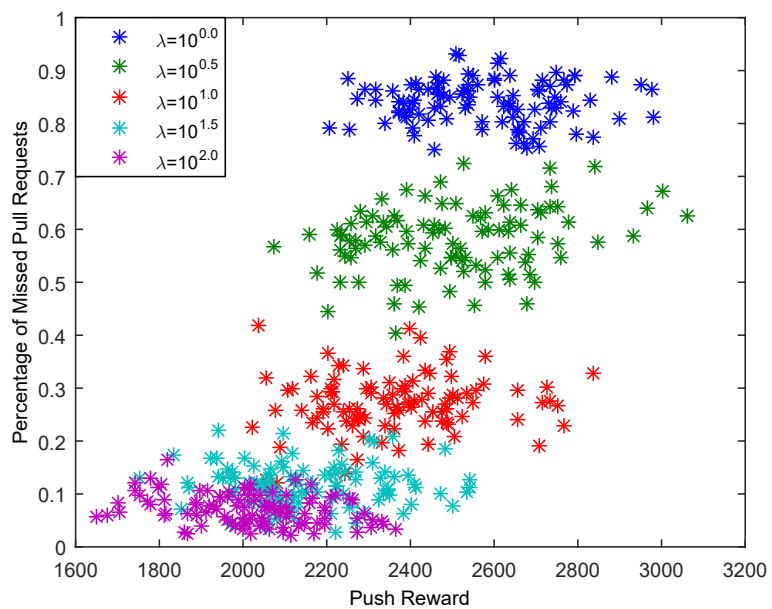
We plot the performance (percentage of missed ‘pull’ requests and overall user rewards delivered) of our hybrid system with scheduling horizon $T_H = 60s$ and different deadlines T_d in Fig.2.11(a),2.12(a) (with $T_d = 5$) and Fig.2.11(b),2.12(b) (with $T_d = 10$) for comparison. Obviously, the larger the ‘push’ weight coefficient λ , the fewer missed user requests, but the less overall system reward.

2.7 More System-Level Statistics

System operators need to consider multiple aspects in order to optimize their operations and return of investment. In this section, we provide more statistics of

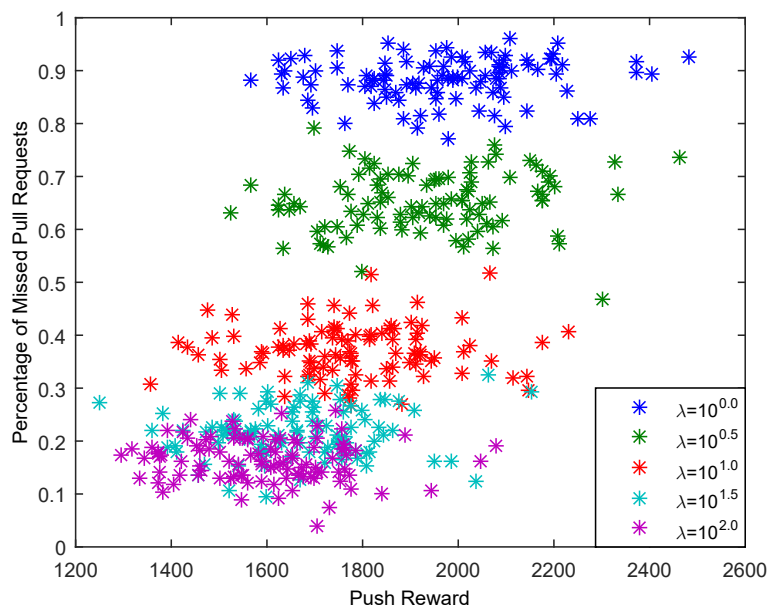


(a) $T_d = 5$

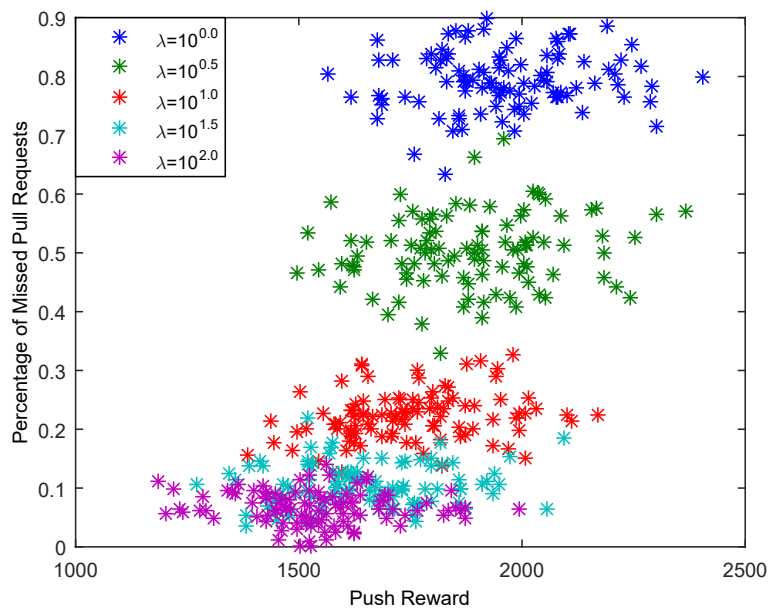


(b) $T_d = 10$

Figure 2.11: Performance for Hybrid System (ML-100K)



(a) $T_d = 5$



(b) $T_d = 10$

Figure 2.12: Performance for Hybrid System (ML-1M)

our scheduling framework to demonstrate the advantages of cross-layer joint optimization.

2.7.1 Resource Utilization

Since spectrum is the most valuable resource in the system, it is a requirement for the system design to ensure that the resource utilization v is high.

$$v_t = \frac{1}{B^{(t)} \cdot T} \sum_j s_j^{(t)} \quad (2.29)$$

Our scheduling framework achieves very high resource utilization per slot, as shown in Fig.2.13. Obviously, the concern for the worst-case scenarios in the MIP problem, in which a substantial portion of the capacity remains unused, is not present in the system and thus the concern is well addressed. Further, unused wireless resources, if any, could be released for other purposes (e.g. normal one-to-one unicast).

2.7.2 Contents Scheduled

In this part, we present the statistics regarding the scheduled transmissions in terms of the rewards delivered and wireless resources used. Intuitively, from the perspective of system operators, it is desirable to transmit contents with higher values using fewer resources, i.e. higher reward per unit resource.

This is confirmed by Fig.2.14,2.15, in which we provided the heat map of the normalized intensity of scheduled contents, for both the wireless resource scheduled ($s_j^{(t)}$) on the x-axis and reward ($\sum_i \alpha_{ij}^{(t)} \cdot \tilde{f}_{ij}^{(t)}$), which includes the additional reward

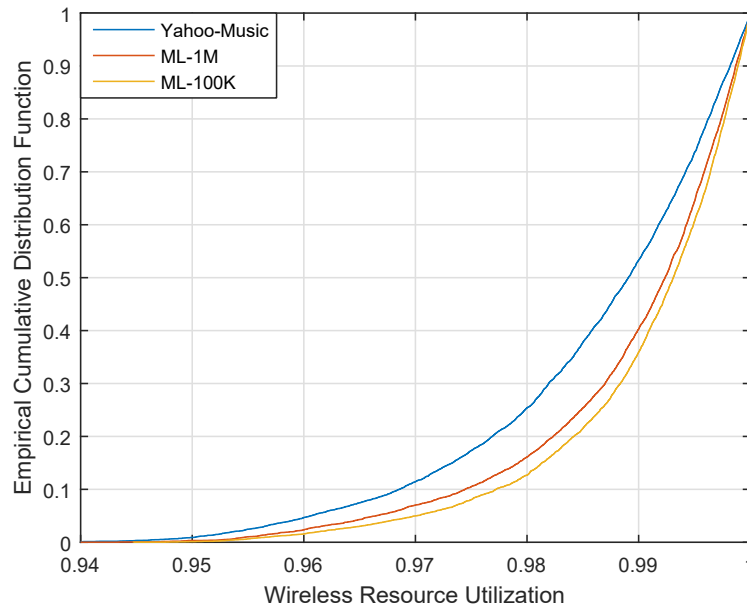
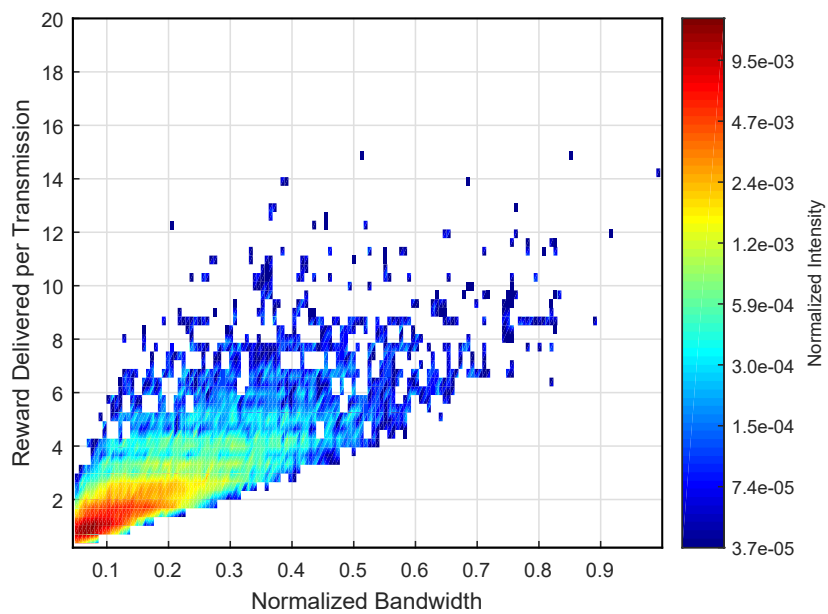


Figure 2.13: Statistics for Wireless Resource Utilization v

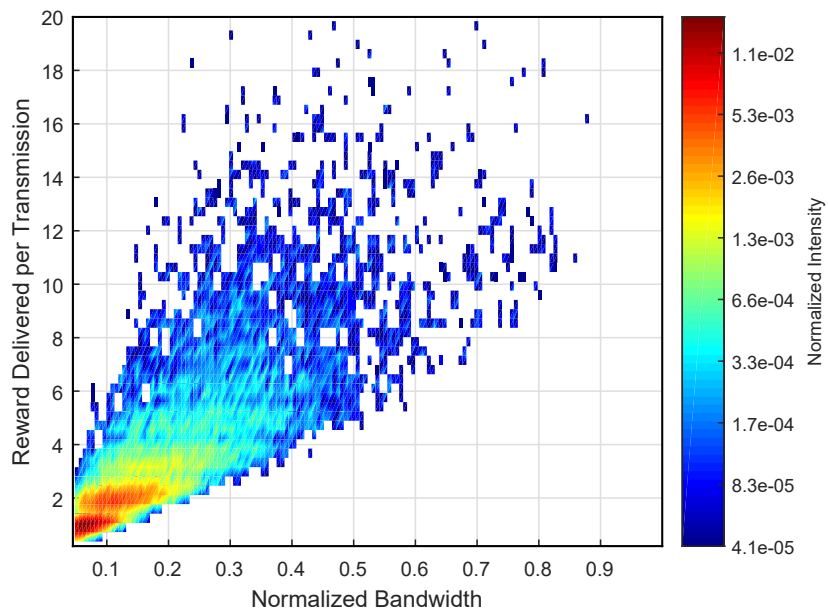
for ‘pull’ requests, since we consider the actual input for the scheduler) on the y-axis for each content transmission. As illustrated, most of the scheduled contents employ only a small portion of the overall wireless resources. It appears that a lower bound exists for the reward delivered per unit wireless resource (or the slope in the figures), though such lower bounds are different for different data sets. The lower bounds will be helpful in further reducing the runtime of solving the MIP problem for the optimal.

2.8 Device-Side Improvements

Up to now, we are investigating the efforts at base station (or server) side. Yet, our optimization mechanism would be incomplete until we investigate how to improve the system performance at the device side as well.

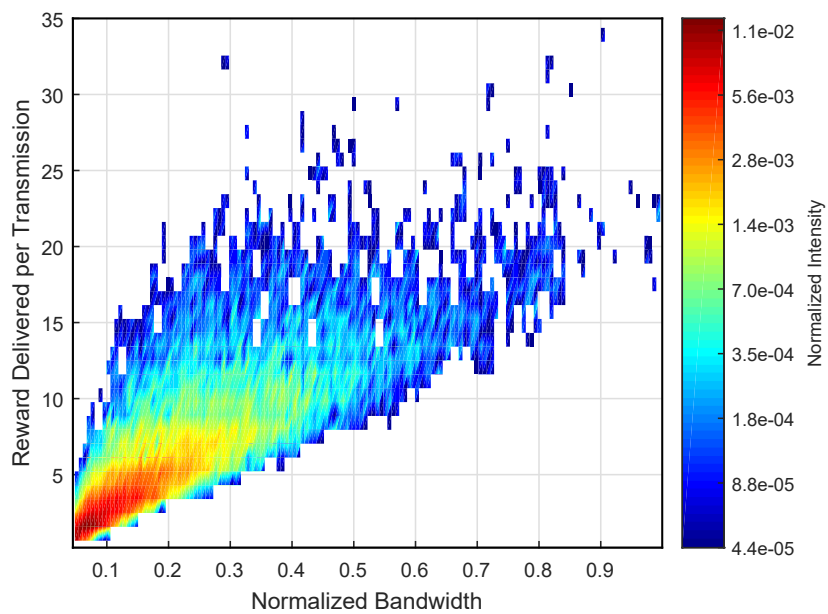


(a) 'Push' weight coefficient $\lambda = 0$ (Pure 'pull' system)

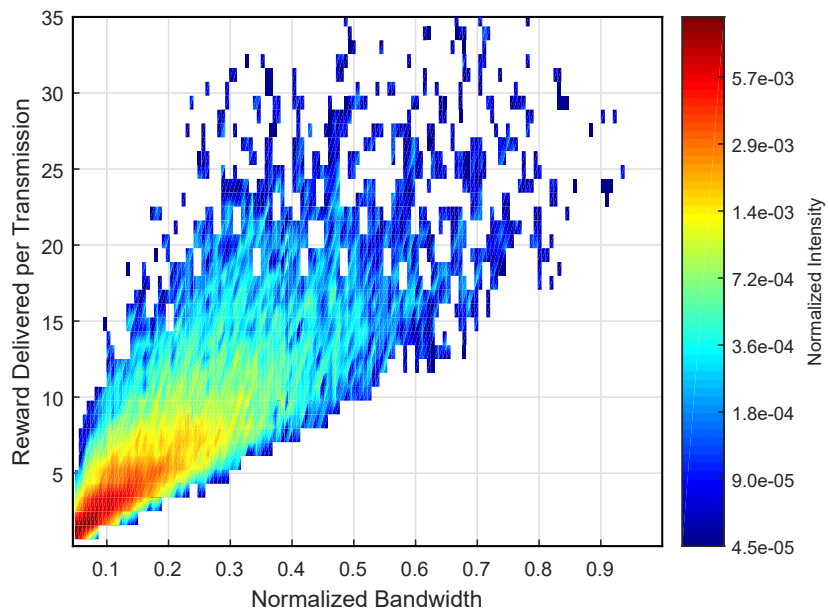


(b) 'Push' weight coefficient $\lambda = 10$

Figure 2.14: Statistics for Scheduled Contents (Yahoo-Music, $T_d = 10$)



(a) 'Push' weight coefficient $\lambda = 0$ (Pure 'pull' system)



(b) 'Push' weight coefficient $\lambda = 10$

Figure 2.15: Statistics for Scheduled Contents (ML-1M, $T_d = 10$)

We observe the following facts:

1. Different users have different device profiles (including usage, storage, remaining battery, etc.), hence it is important that the base stations schedule transmissions accordingly.
2. Most users possess multiple mobile devices (e.g. tablets, smart phones) and they frequently use these devices interchangeably. These modern devices are equipped with near-field communication (NFC) capability, i.e. transmissions between nearby devices consume almost negligible battery and wireless resources. We seldom need to worry about scheduling these NFC transmissions. Better still, users are universally willing to accept coordination between their devices without hesitation, while it is generally much harder to convince them to help other users.

Therefore, we incorporate the following improvements to the system:

1. Devices periodically report their reward threshold $\varepsilon_i^{(t)}$, based on profiles. The system would only schedule for the transmissions of the contents with reward values greater than the reported threshold. This is done by revising (2.12) to (2.30), and (2.14) to (2.31) correspondingly.
2. The system treats the devices belonging to the same individual user as one and schedules transmissions only to the device with the best channel, rather than solely transmitting to the device users are actively using. In other words, the device with the best channel serves as a relay for all other devices of the user.

$$\hat{f}_{jk}^{(t_0)} = \sum_{i: \text{SINR}_i^{(t_0)} \geq \text{SINR}_k^{th}} \mathbf{1} \left(\tilde{f}_{ij}^{(t_0)} > \varepsilon_i^{(t_0)} \right) \cdot \tilde{f}_{ij}^{(t_0)} \quad (2.30)$$

$$\alpha_{ij}^{(t_0)} = \sum_{\substack{i, j: \tilde{f}_{ij}^{(t_0)} > \varepsilon_i^{(t_0)} \\ k: \text{SINR}_k^{th} \leq \text{SINR}_i^{(t_0)}}} \hat{\alpha}_{jk}^{(t_0)} \quad (2.31)$$

2.9 Summary

In this chapter, we investigate the basic social content delivery system with single base station and time-invariant reward, given the constraints from the wireless networks. We propose a framework that evaluates the performance of the system in terms of overall delivered rewards. To optimize system performance, we need to schedule contents and wireless resources according to the solution of a MIP problem, which requires exponential time to obtain the optimal solution. Results show that myopic optimization (without lookahead) yields to sufficiently good performance. We further reduce the complexity of the joint optimization approach to rely only on the number of candidate contents and system transmission modes, regardless of the number of users, by aggregating user rewards at each supported multicast transmission mode.

The simulation results indicate that our joint optimization approach provides better system performance than the traditional layered systems that schedule contents without considering constraints of wireless networks. The joint optimization gain is significant when the resources of the wireless network are comparatively insufficient, either due to the fact that (i) the number of users is large, and/or (ii) the

conditions of the wireless channel are not good for a large number of users. The scheduling is also fair among users in terms of overall rewards received by each user during the scheduling horizon.

Moreover, we investigate the performance of a hybrid system, by introducing additional rewards for user generated requests ('pull' operation) with deadlines. This hybrid system provides a natural way to balance the resource allocation between suggested contents generated by a recommender system and actual user requests. It further proves that our joint optimization framework is a suitable scheduling solution for social content delivery.

Chapter 3: Joint Optimization for Time-Invariant Rewards with Multiple Base Stations

3.1 Overview

In this chapter, we extend the system design in Chapter 2 to scenarios with multiple base stations and investigate the content delivery problem with capacity constraints in a system with centralized heterogeneous wireless infrastructure. In addition to Chapter 2, in a system with multiple base stations, we need to decide which base station serves the users and how the base stations coordinate. It is not a simple extension: the complexity of the system builds up significantly, due to the added freedoms and constraints introduced by new configurations, including but not limited to power control and interference management. However, the timeliness of the online scheduling framework remains essential to the system.

We propose a scalable two-phase scheduling framework, consisting of:

1. distributed delivery decisions by each base station;
2. resource consolidation by the system.

The first step localizes and reduces the problem to several instances of simple single base station scheduling problem in Chapter 2. The second step reduces the

transmission redundancy of the system brought by the first step.

We introduce our system model and evaluation framework in Section 3.2. Solutions to two types of system configuration regarding resource allocation are discussed: ‘out-of-band’ system in Section 3.3, and ‘in-band’ system in Section 3.4. The performance is presented in Section 3.5 and we summarize our conclusions in Section 3.6.

3.2 Problem Formulation

3.2.1 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.3.1).

Channel information of all users for all base stations at all time slots $\{\text{SINR}_i^{(t,l)}\}$ 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. The specified reward could only be claimed in whole at most once (i.e. partial transmission earns no reward and repeated transmissions do not earn additional rewards).

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

$$\begin{aligned}
& \max \sum_{t=1}^{T_H} \sum_l \sum_{i,j} \alpha_{ij}^{(t),l} f_{ij} \\
& \text{s.t.} \quad \alpha_{ij}^{(t),l} \in \{0, 1\} && \forall i, j, t, l \\
& \quad \sum_{t=1}^{T_H} \sum_l \alpha_{ij}^{(t),l} \leq 1 && \forall i, j \\
& \quad \text{QoS} \left(W_j, \{s_j^{(t),l}\}, \{\text{SINR}_i^{(t),l}\} \right) = 1 && \forall \alpha_{ij}^{(t),\lambda} = 1 \\
& \quad \sum_j s_j^{(t),l} \leq B^{(t),l} T && \forall t, l \\
& \quad s_j^{(t),l} \geq 0 && \forall j, t, l
\end{aligned} \tag{3.1}$$

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

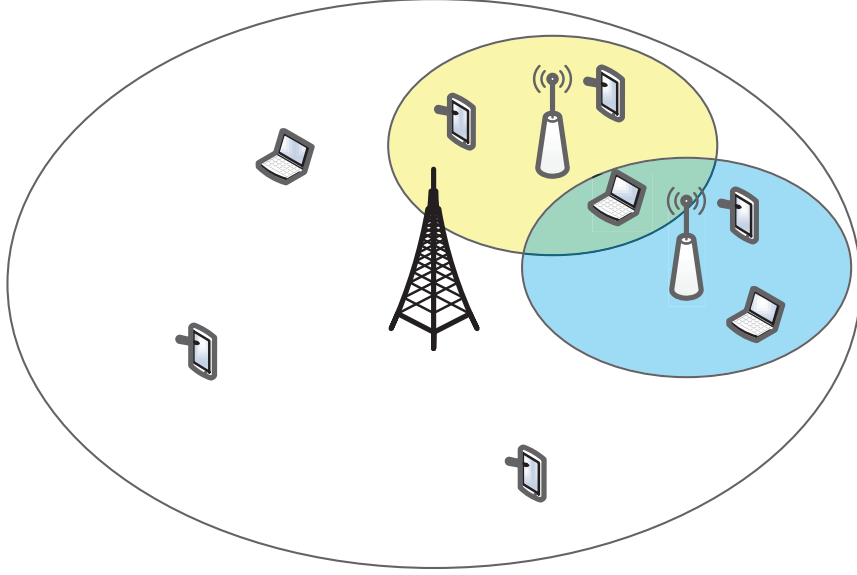


Figure 3.1: System Model: Macro and Pico Cells. Macro cells are generally responsible for signal coverage, while pico cells are responsible for increasing capacity in the area with heavy traffic.

1. What to transmit: binary delivery decision variable $\alpha_{ij}^{(t),l}$ indicates whether or not 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 many wireless resources of base station l to allocate for content j at time slot t .

Note that (3.1) is only an offline version. In light of Chapter 2, 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).

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 resources ('out-of-band'), and the other sharing the same with macro cells ('in-band').

Table 3.1 summarizes definitions of parameters.

3.2.2 Two-Phase Scheduling

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

3.2.2.1 Delivery Decisions

In this phase, given the dedicated wireless resources allocated to it, each base station individually and independently decides the following:

1. what contents to deliver to users $\{\alpha_{ij}^{(t),l}\}$; and
2. resource allocation for the scheduled contents $\{s_j^{(t),l}\}$.

This phase is naturally distributed.

Table 3.1: 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

3.2.2.2 Resource Consolidation

In this phase, the system improves the decisions and resource allocation obtained in the previous phase, e.g. by reducing duplicate transmissions. It is possible that the system changes both the delivery decisions and configurations of wireless resources to enhance the overall spectrum efficiency. Some resources might be released for new allocation. This phase is centralized.

The two phases could be repeated if needed.

3.2.3 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 specific content has been transmitted to a user before, it is not retransmitted in future slots. However, such a guarantee does not exist within the same time slot in the delivery decisions phase. Consequently, we are able 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:

$$\rho_t = \frac{\|\{(i, j) : \sum_l \alpha_{ij}^{(t), l} > 1\}\|}{\|\{(i, j) : \sum_l \alpha_{ij}^{(t), l} \geq 1\}\|} \quad (3.2)$$

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.

3.2.4 Wireless Resource

Generally, there are two types of resources 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 differ-

ence between these two types of configurations lies in their respective interference. ‘Out-of-band’ resource results in an optimal 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.

3.3 System with ‘Out-of-Band’ Resources

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 Chapter 2, we focus on ‘push’ only systems in this chapter, i.e. the system attempts to deliver contents without consideration of user-generated requests. It is easy to employ the same technique in Chapter 2 of altering reward values to incorporate the impact of user requests.

In this chapter, 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 (3.1), as results in Chapter 2 indicated that myopic optimization performs fairly well and runs significantly faster. At each time slot t , we obtain content delivery decisions $\{\alpha_{ij}^{(t),l}\}$ and resource allocation $\{s_j^{(t),l}\}$ based on the solution to the optimization problem, constructed as a mixed integer programming problem as in

(3.3):

$$\begin{aligned}
& \underset{\alpha_{ij}^{(t),l}, s_j^{(t),l}}{\text{maximize}} && \sum_l \sum_{i,j} \alpha_{ij}^{(t),l} f_{ij}^{(t)} \\
& \text{subject to} && \alpha_{ij}^{(t),l} \in \{0, 1\} && \forall i, j, t, l \\
& && \sum_l \alpha_{ij}^{(t),l} \leq 1 && \forall i, j \\
& && \alpha_{ij}^{(t),l} W_j \leq s_j^{(t),l} \mathcal{R}(\text{SNR}_i^{(t),l}) && \forall i, j, l, t \\
& && \sum_j s_j^{(t),l} \leq B^{(t),l} T && \forall t, l \\
& && s_j^{(t),l} \geq 0 && \forall j, t, l
\end{aligned} \tag{3.3}$$

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

$$f_{ij}^{(t)} = f_{ij}^{(t-1)} \left(1 - \max_l \alpha_{ij}^{(t-1),l} \right) \tag{3.4}$$

with initial values

$$f_{ij}^{(1)} = f_{ij} \tag{3.5}$$

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 identical content. 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 the 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 (3.3) 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 $(s^1, \dots, s^{\mathcal{L}})$ as follows. Allocation decision $\vec{s}_1 = (s_1^1, \dots, s_1^{|\mathcal{L}|})$ precedes \vec{s}_2 with regards to utility function f (denoted as $\vec{s}_1 \preceq_f \vec{s}_2$), iff $s_1^i \leq s_2^i, \forall i = 1, \dots, |\mathcal{L}|$ and $f(\vec{s}_1) \leq f(\vec{s}_2)$.

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

$$R_j^{(t_0)}(\vec{s}) = \sum_i f_{ij}^{(t_0)} \cdot \max_l \mathbf{1} \left(\mathcal{R} \left(\text{SNR}_i^{(t_0),l} \right) \geq \frac{W_j}{s^l} \right) \quad (3.6)$$

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.3.2 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 generality, we start with $\hat{\mathcal{L}} = \{1, \dots, |\mathcal{L}|\}$. Denote $\vec{s} = \text{GreedyDecisionDeduplication}(\vec{s}_0)$. Trivially, if we use fewer wireless resources, we will achieve no better performance, or $\forall \vec{s}' \preceq \vec{s}$, we have $R_j^{(t_0)}(\vec{s}') \leq R_j^{(t_0)}(\vec{s})$. The loop invariance of the 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 $R_j^{(t_0)}(\vec{s}') = R_j^{(t_0)}(\vec{s})$.

Assume the contrary and denote the index of first discrepancy as l_0 , i.e. $s_l = s'_l, \forall l \in \{1, \dots, l_0 - 1\}, s_{l_0} < s'_{l_0}$. 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. \square

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

```

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

```

Figure 3.2: Greedy Decision Deduplication Algorithm

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:*

$$L_{\max} = \max_i \sum_l \mathbf{1}(SNR_i^l \geq SINR^{th}) \quad (3.7)$$

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.

3.4 Systems with ‘In-Band’ Resources

For ‘in-band’ systems, different base stations could utilize the same wireless resources. It is not guaranteed to perform better, because simultaneous transmissions introduce 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 resources 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 3.3, by allocating dedicated wireless resources to individual cell and ap-

plying the Greedy Decision Deduplication Algorithm to reduce redundant delivery decisions from different base stations. Afterwards, we determine whether we could further consolidate the wireless resources, by deciding whether or not to share the spectrum to transmit, in order to achieve better spectrum efficiency.

Therefore, with ‘in-band’ wireless resources, 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:

$$\widetilde{\text{SINR}}^l(\vec{P}) = \min_{i \in U_l} \widetilde{\text{SINR}}_i^l(\vec{P}) \quad (3.8)$$

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

$$\widetilde{\text{SINR}}_i^l(\vec{P}) = \frac{h_{l,i}P_l}{N_0 + \sum_{q \neq l} h_{q,i}P_q} \quad (3.9)$$

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.

$$\begin{aligned} \text{SNR}_i^l &= \frac{h_{l,i}P_{l,\max}}{N_0} \\ &= \widetilde{\text{SINR}}_i^l(0, \dots, P_{l,\max}, \dots, 0) \end{aligned} \quad (3.10)$$

Trivially,

$$\widetilde{\text{SINR}}^l(\vec{P}) < \text{SNR}^l = \min_{i \in U_l} \text{SNR}_i^l \quad (3.11)$$

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:

$$s_l = \max_{i \in U_l} \frac{W_l}{\mathcal{R}(\text{SNR}_i^l)} = \frac{W_l}{\mathcal{R}(\text{SNR}^l)} \quad (3.12)$$

$$\tilde{s}_l = \frac{W_l}{\mathcal{R}(\widetilde{\text{SINR}}^l)} = s_l \cdot \frac{\mathcal{R}(\text{SNR}^l)}{\mathcal{R}(\widetilde{\text{SINR}}^l)} \quad (3.13)$$

Denote

$$\tilde{s}_{\min} = \min_l \tilde{s}_l = \min_l \left(\frac{s_l}{\tilde{c}_l} \right) \quad (3.14)$$

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

$$\tilde{c}_l = \frac{s_l}{\tilde{s}_l} = \frac{\mathcal{R}(\widetilde{\text{SINR}}^l)}{\mathcal{R}(\text{SNR}^l)}, \forall l \in \mathcal{L} \quad (3.15)$$

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:

$$\tilde{s} = \tilde{s}_{\min} + \sum_{l \in \mathcal{L}} \left(1 - \frac{\tilde{s}_{\min}}{\tilde{s}_l} \right) \cdot s_l \quad (3.16)$$

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

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

$$\begin{aligned} \tilde{\Delta} &= \sum_l s_l - \tilde{s} \\ &= \tilde{s}_{\min} \left(\sum_{l \in \mathcal{L}} \tilde{c}_l - 1 \right) \end{aligned} \quad (3.17)$$

Therefore, we have the improvement condition:

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

$$\sum_{l \in \mathcal{L}} \tilde{c}_l > 1 \quad (3.18)$$

Note that the discussion above is independent of which contents are being transmitted, therefore, it could be applied to different and/or identical 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 [15]), we can transform the problem into feasibility problems with different parameters. The basic formulation is as follows: given the target SINR level $\text{SINR}_{th}^{k_l}$ for transmission mode $k_l \in \{1, \dots, K\}$ at each base station $l \in \mathcal{L}$, determine whether a power allocation vector \vec{P} for spectrum sharing exists such that,

$$\widetilde{\text{SINR}}_i^l(\vec{P}) \geq \text{SINR}_{th}^{k_l}, \forall i \in U_l \quad (3.19)$$

Trivially, we require mutual exclusiveness (3.20):

$$U_{l_1} \cap U_{l_2} = \emptyset, \forall l_1 \neq l_2 \in \mathcal{L} \quad (3.20)$$

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

$$\begin{aligned} -\frac{h_{l,i}}{\text{SINR}_{th}^{k_l}} P_l + \sum_{q \neq l} h_{q,i} P_q + N_0 &\leq 0 \quad \forall i \in U_l \\ 0 &\leq P_l \leq P_l^{\max} \quad \forall l \in \mathcal{L} \end{aligned} \quad (3.21)$$

We could further normalize the constraint set to

$$\begin{aligned} -\frac{p_{il}}{\text{SINR}_{th}^{k_l}} \xi_l + \sum_{q \neq l} p_{iq} \xi_q &\leq -1 \quad \forall i \in U_l \\ 0 &\leq \xi_l \leq 1 \quad \forall l \in \mathcal{L} \end{aligned} \quad (3.22)$$

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

$$p_{il} = \frac{h_{l,i}}{N_0} P_l^{\max} \quad (3.23)$$

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

$$P_l = \xi_l P_l^{\max} \quad (3.24)$$

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 $(k_1, \dots, k_{|\mathcal{L}|}) \in \{1, \dots, K\}^{|\mathcal{L}|}$ that satisfies (3.25), we run a feasibility test for constraint set (3.22) to determine if a power allocation solution is available for such decision

$$\sum_{l \in \mathcal{L}} \frac{R_{k_l}}{\mathcal{R}(\text{SINR}^l)} > 1 \quad (3.25)$$

3.4.1 Coordinated Multi Point Transmission

As [16] suggested, multiple base stations, if fully connected and synchronized, could cooperate and transmit the same signal simultaneously to increase the SINR by transforming interference into a useful signal. Obviously, it is an ideal addition in our application scenario, as long as it could actually be implemented.

It is easy to incorporate such technology in our scheduling framework. As long as the base stations are delivering the same content(s), we need to obtain the power

configuration for all the base stations \mathcal{L} based on the feasibility test of multicast mode k for the base station cooperation set \mathcal{L}^c :

$$\begin{aligned}
-\frac{1}{\text{SINR}_{th}^k} \sum_{l \in \mathcal{L}^c} p_{il} \xi_l + \sum_{l \notin \mathcal{L}^c} p_{il} \xi_l &\leq -1 \quad \forall i \in \bigcup_{l \in \mathcal{L}^c} U_l \\
0 \leq \xi_l &\leq 1 \quad \forall l \in \mathcal{L}
\end{aligned} \tag{3.26}$$

3.5 Simulations and Results

3.5.1 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 data sets of Yahoo [14] and MovieLens [17], 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 3.2 [12].

There is one macro base station and two pico base stations. The two pico base stations are located with a 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 resources 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 (3.3) for the optimal.

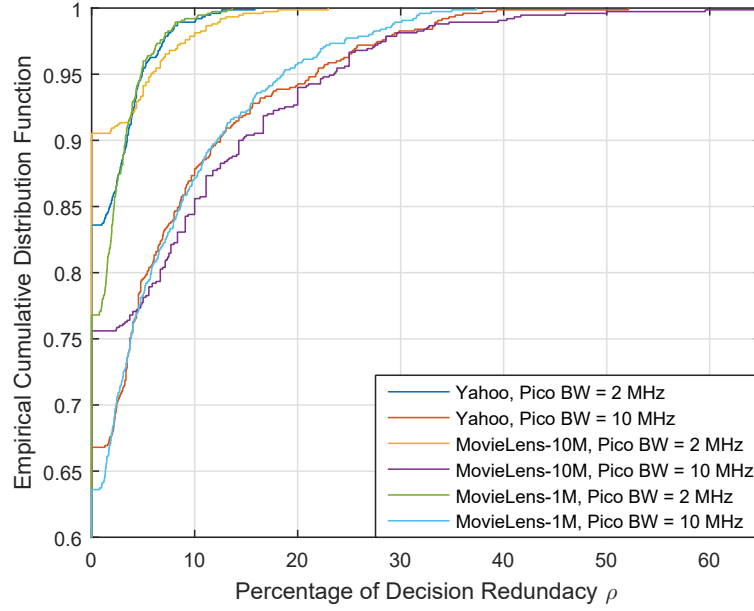


Figure 3.3: Decision Redundancy for ‘Out-of-Band’ Systems.

3.5.2 Results

3.5.2.1 ‘Out-of-Band’ Systems

We first present the distribution of decision redundancy ρ for the ‘out-of-band’ systems in Fig. 3.3. The redundancy is higher when pico base stations are allocated with more resources. This is intuitive because with more resources, 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 reasonably 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.

3.5.2.2 ‘In-Band’ Systems

The results of ‘in-band’ systems are presented in Fig.3.4, 3.5. 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 resources, or 76% of the wireless resources 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 (3.22), we could potentially save more wireless resources, but at the expense of losing user rewards.

We can conclude from Fig.3.4 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 com-

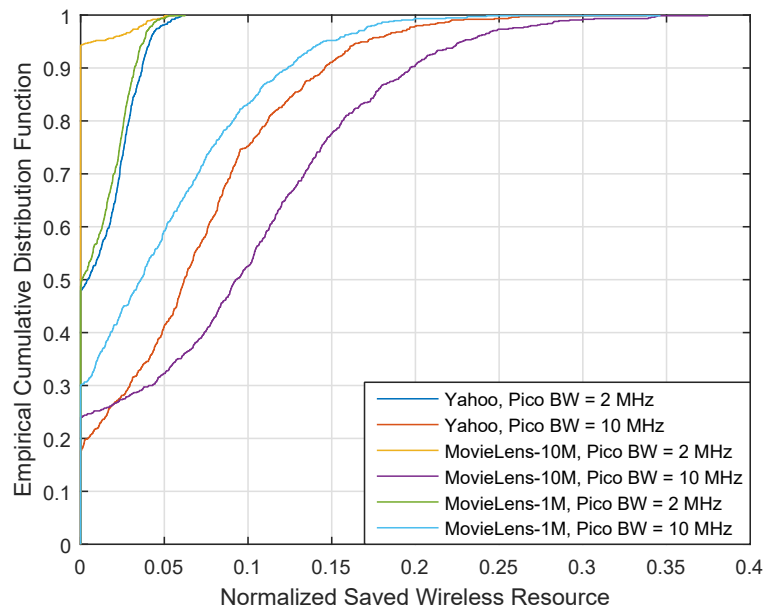


Figure 3.4: Saved Wireless Resources for ‘In-Band’ Systems.

peting transmissions. The latter has been inherently reduced in the system design because the first phase of the scheduling framework executes with dedicated wireless resources.

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

3.6 Summary

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 data sets and the results indicate this novel approach is both efficient

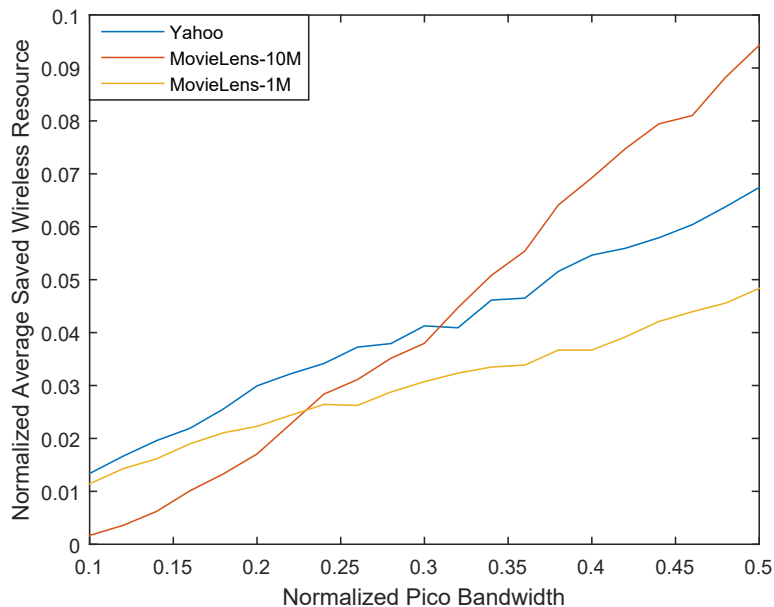


Figure 3.5: Average Saved Wireless Resources for ‘In-Band’ Systems.

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

There are no explicit references to the hierarchy of the base stations in this chapter, 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 chapter 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).

Table 3.2: 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

Chapter 4: Joint Optimization for Time-Variant Rewards with Single Base Station

4.1 Overview

In the real world, user rewards are hardly static since they evolve with time, though the dynamics governing such evolutions still remain an open question. Intuitively, just like any contagious disease, the evolutions are ‘viral’, in that they require certain forms of interactions between the users, be it ‘share’, ‘like’, rate or comment. Particularly, researchers have already confirmed the effectiveness of friend’s recommendations compared to those generated by the algorithm [18]. In fact, as the commercial success of Facebook demonstrates, users are more vulnerable to the influence of other users than that of machine suggestions. In other words, users are highly likely to consume contents (even ads) that are disseminated by actual human users, especially by their friends. Modern social networks significantly reduce the degree of separations between users [19] and ‘infection’ time using attention-grabbing mobile notifications, which in turn makes these interactions more pervasive.

Unlike Chapter 2 and Chapter 3, in this chapter, we are taking the dynamics of social networks into account: the reward values are time-variant due to the in-

teractions between users. An intuitive example of the importance of the dynamics is that the contents only become more rewarding and viral after certain users (e.g. celebrities, public figures) consume and spread them.

According to [20], content dissemination (or ‘cascade’) in social network generally follows a diffusion process. The cascades are predictable with high precision based on user profiles. Fast algorithms are developed in [21] to infer network diffusion parameters for a continuous time diffusion model. However, the models used at social network layer failed to consider the fact that mobile delivery can be the bottleneck of social content dissemination.

In this chapter, we investigate the content dissemination problem with capacity constraints in a system with centralized wireless infrastructure. We need to decide what contents to transmit to which users and how to transmit them.

We introduce our system model and evaluation framework in Section 4.2. The impact of delivery delay is presented in Section 4.3 and we propose look-ahead scheduling based on predictions of social dynamics in Section 4.4. The performance is analyzed in Section 4.5 and we summarize our conclusions in Section 4.6.

4.2 Problem Formulation

4.2.1 General System Model

We consider a centralized system that both selects contents to deliver according to user rewards given wireless capacity constraints and delivers the contents to users via a wireless network comprising different transmission modes, as illustrated in

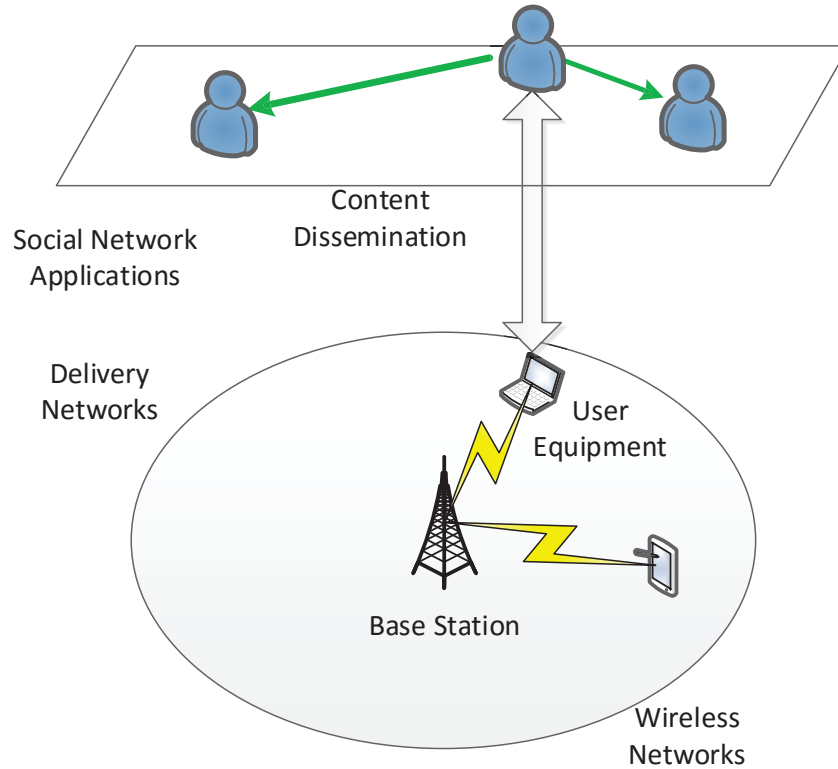


Figure 4.1: System Model: The disseminating contents are propagated from influencer(s) to influencee(s) based on influence graphs (the boldness of the edge demonstrate its ‘strength’), while the wireless networks are responsible to deliver them to user devices.

Fig.4.1. Channel information of all users for all base stations are known to the system and we assume the bandwidth of the wired connections among the base stations and between base stations and content server(s) is sufficient that the base station could access contents as if they are stored locally. This assumption is valid in practice because base stations are generally connected via fiber optic cables. Trivially, with developments of memory chips, the storage on user devices is sufficiently large to precache all the contents scheduled for delivery.

In this chapter, we focus on a slotted single base station scenario with K multicast modes and slot length T . The same as Chapter 2, we complete transmitting any content within one scheduling slot to avoid management complexity of multicast groups. The objective is to maximize overall user rewards of the system subject to capacity constraints.

$$\begin{aligned}
& \max \quad \sum_{t=1}^{T_H} \sum_k \sum_{i,j} \alpha_{ij}^{(t),k} f_{ij}^{(t)} \\
& \text{s.t.} \quad \alpha_{ij}^{(t),k} \in \{0, 1\} \quad \forall i, j, k, t \\
& \quad \sum_{t=1}^{T_H} \sum_k \alpha_{ij}^{(t),k} \leq 1 \quad \forall i, j \\
& \quad s_j^{(t)} \cdot R_k \geq W_j \quad \forall \alpha_{ij}^{(t),k} = 1 \\
& \quad \text{SNR}_i^{(t)} \geq \text{SNR}_k^{th} \quad \forall \alpha_{ij}^{(t),k} = 1 \\
& \quad \sum_j s_j^{(t)} \leq B^{(t)}T \quad \forall t \\
& \quad s_j^{(t)} \geq 0 \quad \forall j, t
\end{aligned} \tag{4.1}$$

Reward for transmitting content j to user i is time-variant $f_{ij}^{(t)} \in [0, 1]$ (unlike Chapter 2 and 3). Given the optimization formulation in (4.1), only contents with positive reward values would be scheduled for transmission.

The time-variance of reward usually comes from the following two aspects:

1. social dynamics: users become more (or less) interested in a content than before, due to other users (either their own behaviors or interactions); and/or
2. information revelation: after gathering sufficient information regarding users (or their peer), the social networks may adjust previous predictions.

In this chapter, we model the time-variance based on social dynamics. Specifically, we focus on the content dissemination with the help of influence graphs.

4.2.2 Content Dissemination Without Delivery Delay

At social layer, for a given content without delivery delay, we model the continuous-time reward change as activation process: reward for user i to consume the content (j omitted in notation for simplicity) at time slot t is dependent on reward value f_i and its binary activation state $\gamma_i(t)$.

$$\tilde{f}_i^{(t)} = f_i \cdot \tilde{\gamma}_i(t) \quad (4.2)$$

The activation process is based on directed graph $G = (V, E)$. Binary activation state of user u is dependent on previous activation states of influencer set I_u^{\leftarrow} .

$$I_u^{\leftarrow} = \{v : \lambda_{vu} > 0\} \quad (4.3)$$

Conversely, the influencee set is I_u^{\rightarrow} .

$$I_u^{\rightarrow} = \{v : \lambda_{uv} > 0\} \quad (4.4)$$

For generality, we do not confine the influencer/influencee set to include only the explicit neighbors of the user, but due to system complexity, the influence propagation is usually calculated based on a sparse set of users.

$$\tilde{\gamma}_u(t) = \mathcal{A}_u \left(\left\{ \tilde{\gamma}_v(\tau) \right\}_{\substack{v \in I_u^{\leftarrow} \\ \tau < t}} \right) \quad (4.5)$$

For simplicity, we assume that once users are activated, they will stay acti-

vated, or

$$\tilde{\gamma}_u(t_1) \leq \tilde{\gamma}_u(t_2), \forall t_1 < t_2 \quad (4.6)$$

We denote the activation time $\tilde{\mathcal{T}}_u$ as

$$\tilde{\gamma}_u(t) = \begin{cases} 0 & \forall t < \tilde{\mathcal{T}}_u \\ 1 & \forall t \geq \tilde{\mathcal{T}}_u \end{cases} \quad (4.7)$$

Based on the assumptions of independence and time-shift-invariance as in [21], we rewrite the state transition as

$$\tilde{\gamma}_u(t) = \max_{v \in I_u^{\leftarrow}} \mathcal{A}_{vu} \left(t - \tilde{\mathcal{T}}_v \right) \quad (4.8)$$

Mappings $\{\mathcal{A}(\cdot)\}$ are obtained from social network applications based on user profiles, and they are subject to different models.

The mappings satisfy causality

$$\mathcal{A}_{vu}(t) = 0, \forall t < 0 \quad (4.9)$$

In general cases, mappings $\mathcal{A}(\cdot)$ are defined on a random space. In this chapter, we assume that the pairwise activation time τ_{vu} for user v to ‘infect’ user u is independent and follows exponential distribution $\tau_{vu} \sim \text{Exp}(\lambda_{vu})$. Diffusion parameter λ_{vu} reflects the influence rate of the influence link $v \rightarrow u$, the larger the more quickly user u is activated due to user v . User v has no influence on user u iff $\lambda_{uv} = 0$.

Then we could obtain the activation time of user u in (4.10).

$$\tilde{\mathcal{T}}_u = \min_{v \in I_u^{\leftarrow}} \left(\tilde{\mathcal{T}}_v + \tau_{vu} \right) \quad (4.10)$$

Denote the social dissemination graph for a specific content as $G(V, E)$, with users as its nodes V , directed influence link as its edges $E = \{(u, v) : \lambda_{uv} > 0\}$. For time t , the binary user activation state vector $\tilde{\gamma}_G(t) \in \{0, 1\}^{|V|}$ in the graph is:

$$\tilde{\gamma}_G(t) = (\tilde{\gamma}_u(t))_{u \in V} \quad (4.11)$$

Then, the state transition process is essentially a Markov Chain with transition rate:

$$\lambda_{\tilde{\gamma}_G} = \sum_{(u,v) \in E} \tilde{\gamma}_u \cdot (1 - \tilde{\gamma}_v) \cdot \lambda_{uv} \quad (4.12)$$

and transition probability

$$\mathbb{P}[\tilde{\gamma}_G + \mathbf{e}_v \mid \tilde{\gamma}_G] = \frac{1}{\lambda_{\tilde{\gamma}_G}} \left(\sum_{u \in I_v^+} \mathbf{1}(\tilde{\gamma}_u) \cdot [1 - \mathbf{1}(\tilde{\gamma}_v)] \lambda_{uv} \right) \quad (4.13)$$

Note that modern social network applications significantly increase the influence rate λ 's (in a selective way), resulting in much shorter activation time for certain social connections. This is due to the facts that notifications on mobile devices are more visible to the users, and at the same time users are generally more attentive when they use mobile devices. But this is way beyond our discussions for this chapter, or even dissertation.

Table 4.1 summarizes definitions of parameters.

4.3 Impact of Delivery Delay on Content Dissemination

Due to capacity constraint of wireless networks, we might not be able to schedule the contents immediately. However, a user is activated after the content is available at the device. Therefore, the content dissemination starts after the content

itself is delivered and consumable to the target users. For simplicity, we assume users will wait for the content rather than moving on to the next content. Denote π_v as the delivery delay for user v , then the diffusion process is no longer Markovian:

$$\mathcal{T}_u = \min_{v \in I_u} (\mathcal{T}_v + \pi_v + \tau_{vu}) \quad (4.14)$$

with \mathcal{T}_u the activation time of user u in the new process.

As evident, this new process of dissemination is not quite the same as the original diffusion process except when $\pi_u \ll \mathcal{T}_u, \forall u$. Intuitively, we might want to minimize each π_u so that the dissemination faithfully follows the prediction and we could obtain a higher reward (as shown in [Theorem 4](#)).

There are two possible solutions:

1. On-demand: traditional systems schedule transmissions only after users are activated, hoping to achieve low latency for each individual transmission. Such goal is extremely hard to achieve without sacrificing overall system performance (as the results indicated in [Chapter 2](#)), especially when the wireless capacity is insufficient. Note that $\pi_u > 0, \forall u$.

2. Precache: the system transmits content to the users before they request it. This approach makes it possible for some users to enjoy no delivery delay $\pi_u = 0$, if not for all, provided that the delivery happens before user activation happens.

Therefore, we need flexibility and insight in design to obtain a better system.

We observe two facts:

(i) We are not obliged to provide universal delivery guarantee for all users any time, i.e. certain users at certain social network activation states could receive different quality of service.

(ii) With multicast available, we might be able to eliminate the delivery delay for a group of users at the same time.

Theorem 4 (Worse Performance with Delay). *With delivery delay, the expected rewards delivered by the system until scheduling horizon \mathcal{T}_H is no better than the system without delay.*

Proof. The expected rewards collected is

$$\mathcal{R}_{\gamma_0}(\mathcal{T}_H) = \sum_{u \in V} f_u \cdot \mathbb{P}[\mathcal{T}_u \leq \mathcal{T}_H \mid \gamma_0] \quad (4.15)$$

and the no-delay version is

$$\tilde{\mathcal{R}}_{\gamma_0}(\mathcal{T}_H) = \sum_{u \in V} f_u \cdot \mathbb{P}[\tilde{\mathcal{T}}_u \leq \mathcal{T}_H \mid \gamma_0] \quad (4.16)$$

Yet with conditional probability

$$\mathbb{P}[\mathcal{T}_u \leq \mathcal{T}_H \mid \gamma_0] = \int \mathbf{1}[T_u \leq \mathcal{T}_H \mid \gamma_0, \{\tau_{\mu\nu}\}] dF(\{\tau_{\mu\nu}\}) \quad (4.17)$$

Given $\{\tau_{\mu\nu}\}$ and $\{\pi_\mu\}$, we have

$$T_u \geq \tilde{T}_u, \forall u \quad (4.18)$$

Therefore, we have

$$\mathcal{R}_{\gamma_0}(\mathcal{T}_H) \leq \tilde{\mathcal{R}}_{\gamma_0}(\mathcal{T}_H) \quad (4.19)$$

□

4.4 Look-Ahead Scheduling

Obviously, if we could leverage the information from the social network on how contents disseminate, the system might be able to precache smartly, utilize wireless resource fully, and provide better user experience.

Unfortunately, for general graphs, calculating the exact probability of activation before social scheduling horizon \mathcal{T}_H is not quite easy. For a directed forest graph (in which each node has at most one parent, i.e. in-degree of each node is at most one $|\{v : \tau_{vu} > 0\}| \leq 1, \forall u$), it yields to analytical form [22], provided that a node u 's root ancestor is activated at time 0:

$$\mathbb{P}[\gamma_u(\mathcal{T}_H) = 1] = \sum_{\omega \in \theta_u} \left[(1 - e^{-\lambda_\omega \mathcal{T}_H}) \cdot \prod_{\substack{\omega' \in \theta_u \\ \omega' \neq \omega}} \frac{\lambda_{\omega'}}{\lambda_{\omega'} - \lambda_\omega} \right] \quad (4.20)$$

where θ_u is the path (set of directed edges) from node u 's root ancestor to itself.

However, if the graph is moderately complicated, e.g. the graph has rings or alternative paths, it is computationally expensive to calculate the exact probability in its analytic form. To the best of our knowledge, there are no reported results regarding analytical forms of the solution.

To resolve the disadvantage, we employ the Monte Carlo method to approximate the reward. The basic idea is to generate a large number of instances and use instance average to substitute expectation. For each instance, we scan the dissemination path with breadth first search, using min-heap H (ordered by activation time) to store the nodes to be activated before scheduling horizon \mathcal{T}_H and their respective activation times. The complexity of each heap operation is $\mathcal{O}(\log |H|)$,

hence the overall complexity for running a simulation instance is $\mathcal{O}(|E| + |V| \log |V|)$. The algorithm is stated in Fig.4.3. Note that if the content has been delivered before (denoted by binary user delivery state ϕ_u , 0 not delivered and 1 delivered), we do not attribute the reward to current transmission, but rather previous transmission. Given the fact that the activation graph is sparse, we could run each instance relatively fast and different instances could run in parallel.

Obviously, we need to run sufficient number M_s of instances and use the instance result to predict activation states for each disseminating content given its current user activation states.

$$\hat{\gamma} = \Psi(\hat{\gamma}^1, \dots, \hat{\gamma}^{M_s}) \quad (4.21)$$

where Ψ is a chosen detector function.

4.4.1 Runtime

The runtime for each simulation instance depends on the number of edges evaluated. We present the statistics of runtimes for 100,000 simulation instances running on a machine with 2.2 GHz CPU in Table 4.2, for graph parameters used in Section 4.5. Apparently, the simulations are fast and hence could be run real-time in actual systems.

Require: User activation state γ , directed and weighted diffusion graph $G = (V, E)$

with diffusion parameter λ_{uv} as the weight of the edge from u to v .

```

1: procedure ESTIMATEUSERACTIVATION( $\mathcal{T}_H$ )
2:    $H \leftarrow \emptyset$   $\triangleright H$  is a min-heap for tuple  $(i, \hat{\mathcal{T}}_i)$  ordered on ascending estimated
   activation time  $\hat{\mathcal{T}}_i$ .
3:    $\hat{\gamma} \leftarrow \gamma$ 
4:   for  $i \leftarrow \{i : \gamma_i = 1\}$  do
5:      $H.insert((i, 0))$ 
6:   while  $H \neq \emptyset$  do
7:      $u, \hat{\mathcal{T}}_u \leftarrow H.poll()$   $\triangleright$  Next activated node.
8:      $\hat{\gamma}_u \leftarrow 1$ 
9:     for  $v \leftarrow \{v : v \in I_u^{\rightarrow}, \gamma'_v = 0\}$  do
10:       $\hat{\tau}_{uv} \leftarrow exp(\lambda_{uv})$ 
11:       $\hat{\mathcal{T}}'_v = \hat{\mathcal{T}}_u + \hat{\tau}_{uv}$ 
12:      if  $\hat{\mathcal{T}}'_v \leq \mathcal{T}_H$  then  $\triangleright$  Only update when the activation is within
      scheduling horizon.
13:        if  $v \notin H$  then
14:           $H.insert((v, \hat{\mathcal{T}}'_v))$ 
15:        else
16:           $H.update(v, \min(\hat{\mathcal{T}}_v, \hat{\mathcal{T}}'_v))$ 
17:   return  $\{\hat{\gamma}\}$ 

```

Figure 4.3: Monte Carlo Estimation of User Reward

4.4.2 Performance

The major metrics about the performance of the estimations are false prediction ratio P_F and missed prediction ratio P_M .

$$P_F = \frac{|\{i : \gamma_i = 0 \wedge \hat{\gamma}_i = 1\}|}{|\{i : \hat{\gamma}_i = 1\}|} \quad (4.22)$$

$$P_M = \frac{|\{i : \gamma_i = 1 \wedge \hat{\gamma}_i = 0\}|}{|\{i : \gamma_i = 1\}|} \quad (4.23)$$

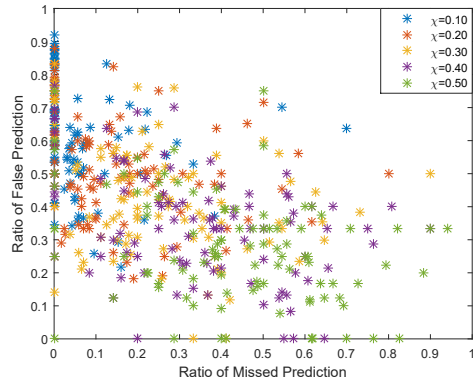
We use a simple independent and identical threshold hard-detector in (4.24) for all users. The number of estimations per instance is $M_s = 50$. We plot the performance for different relative threshold ratio χ in Fig.4.4 with different parameters in each sub graph (number of initial activated users $\|\gamma_0\|_1$ and dissemination horizon \mathcal{T}_H).

$$\Psi_i(\{\hat{\gamma}^q\}) = \mathbf{1} \left(\sum_q \hat{\gamma}_i^q > \chi \cdot M_s \right) \quad (4.24)$$

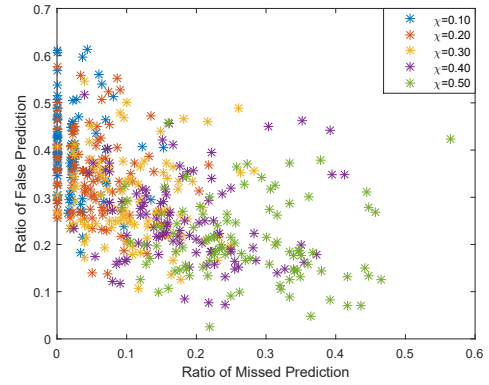
We could conclude that:

1. the performance of Monte Carlo Estimation based on a simple detector is somewhat acceptable;
2. the more initial activated users (denoted by $\|\gamma_0\|_1$) and/or the longer the dissemination horizon (\mathcal{T}_H), the more precise Monte Carlo Estimation is;
3. trivially, when the threshold χ is higher, missed prediction is higher but false prediction is lower.

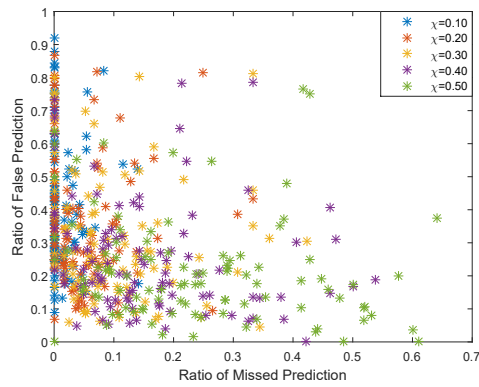
The predicted look-ahead reward for delivering content j to user i at slot t is



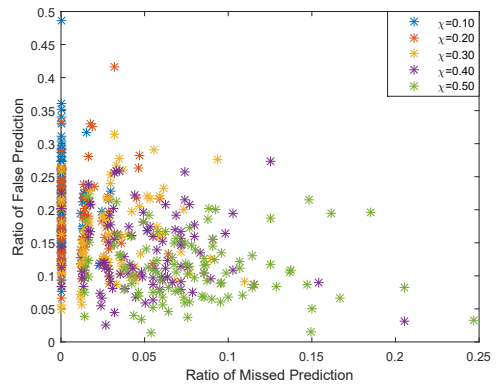
(a) $\|\gamma_0\|_1 = 1, \mathcal{T}_H = 60s$



(b) $\|\gamma_0\|_1 = 5, \mathcal{T}_H = 60s$



(c) $\|\gamma_0\|_1 = 1, \mathcal{T}_H = 120s$



(d) $\|\gamma_0\|_1 = 5, \mathcal{T}_H = 120s$

Figure 4.4: Performance for Monte Carlo Estimation

thus:

$$\hat{f}_{ij}(t) = \hat{\gamma}_{ij}^{(t)} \cdot \left(1 - \phi_{ij}^{(t)}\right) \cdot f_{ij} \quad (4.25)$$

even though the reward might not be immediately claimed at the same slot of delivery.

Note that we only need to run these Monte Carlo simulations once new user activations occur.

Given the computation complexity and in light of Chapter 2, we obtain the scheduling decisions for each time slot t by solving mixed integer programming problem (4.26) for the optimal. We could easily confirm that the scheduling solution satisfies all constraints in (4.1).

$$\begin{aligned} \max \quad & \sum_k \sum_{i,j} \bar{\alpha}_{jk}^{(t)} \cdot \bar{f}_{jk}^{(t)} \\ \text{s.t.} \quad & \bar{\alpha}_{jk}^{(t)} \in \{0, 1\} \quad \forall j, k, t \\ & \sum_k \bar{\alpha}_{jk}^{(t)} \leq 1 \quad \forall j, k \\ & s_j^{(t)} \cdot R_k \geq \alpha_{jk}^{(t)} \cdot W_j \quad \forall j, k, t \\ & \sum_j s_j^{(t)} \leq B^{(t)}T \quad \forall t \\ & s_j^{(t)} \geq 0 \quad \forall j, t \end{aligned} \quad (4.26)$$

with look-ahead reward $\bar{f}_{jk}^{(t)}$ for content j transmitted at mode k based on predicted user rewards in (4.27).

$$\bar{f}_{jk}^{(t)} = \sum_{i: \text{SNR}_i^{(t)} \geq \text{SNR}_k^{th}} \hat{f}_{ij}^{(t)} \quad (4.27)$$

Then we have the reverse mapping and state transitioning

$$\alpha_{ij}^{(t)} = \sum_{\substack{i,j:\hat{f}_{ij}^{(t)}>0 \\ k:\text{SINR}_k^{th}\leq\text{SINR}_i^{(t)}}} \bar{\alpha}_{jk}^{(t)} \quad (4.28)$$

$$\phi_{ij}^{(t+1)} = \max\left(\phi_{ij}^{(t)}, \alpha_{ij}^{(t)}\right), \forall i, j, t \quad (4.29)$$

4.5 Simulations and Results

In this part, we conduct simulations to demonstrate the relationship between delay, total rewards obtained for content cascade and overall system reward delivered.

We simulate the performance based on synthetic Kronecker graphs [23], generated by SNAP [24]. In the simulations, we use the default parameters (with seed matrix $[0.9, 0.5; 0.5, 1]$) to generate Kronecker synthetic graph instances of size 256 (or 2^8) nodes. The set of initial activated users are chosen as follows: 1) choose number of activated users M_A uniformly from 1 to 10; 2) choose M_a users uniformly random from the top 20 influencers (i.e. users with largest out-degree $|I_i^{-\rightarrow}|$). For simplicity, we assume that the reward for an activated user to consume a disseminating content is 1.

System level parameters are the same as Chapter 2, as shown in Table 2.2.

4.5.1 What Shall We Aim For?

We fix the horizon $\mathcal{T}_H = 120s$. For each instance Ω , consisted of activation times $\{\tau_{uv}\}$ and delay times $\{\pi_u\}$, we focus on the ratio $\kappa \in [0, 1]$ of activated and

delivered users (essentially the users that have collected the reward) normalized against no delay.

$$\kappa(\Omega) = \frac{\mathcal{N}(\{\tau_{uv}\}, \{\pi_u\}; \gamma, G)}{\mathcal{N}(\{\tau_{uv}\}, \mathbf{0}; \gamma, G)} \quad (4.30)$$

Delivery delay is consisted of two parts: actual transmission time and scheduling delay. Since we use a slotted system, the actual transmission time T_0 is the same as the time slot length $T_0 = T = 1s$.

4.5.1.1 Different Scheduling Delay Models

We use three scheduling delay models with the same expected total delay time π_0 (i.e. scheduling delay is $\pi_0 - T_0$):

1. Fixed:

$$f_{\text{fixed}}(\pi) = \delta(\pi - \pi_0) \quad (4.31)$$

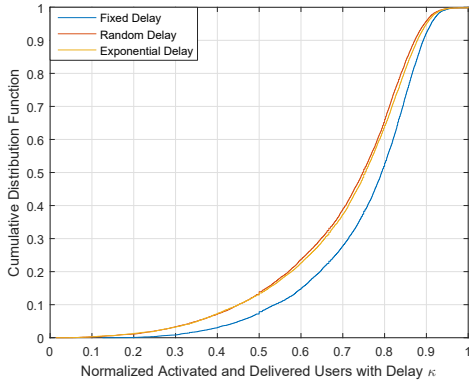
2. Exponential:

$$f_{\text{exp}}(\pi) = \frac{1}{\pi_0 - T_0} e^{-\frac{\pi}{\pi_0 - T_0}} \cdot \mathbf{1}(\pi > T_0) \quad (4.32)$$

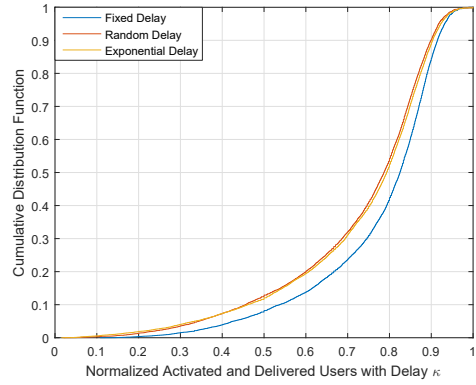
3. Uniformly random:

$$f_{\text{uniform}}(\pi) = \frac{1}{2(\pi_0 - T_0)} \cdot \mathbf{1}(T_0 < \pi < 2\pi_0 - T_0) \quad (4.33)$$

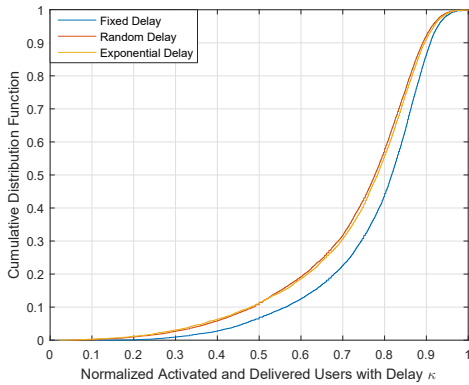
We plot the distribution of κ with different π_0 in Fig.4.5,4.6. Ideally, we want the distribution to be close to the step function $u(t - 1)$, because it essentially means that the content disseminates as if no delay is present. From the figures, we could tell that fixed delay performs better in distribution than the other two



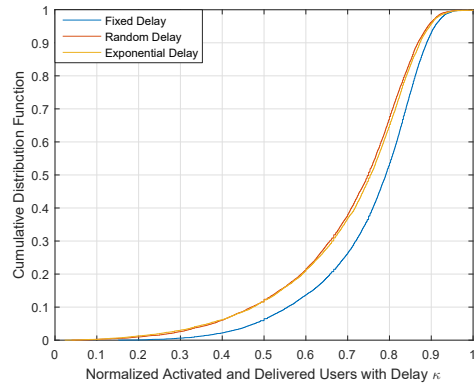
(a) Graph Instance #1



(b) Graph Instance #2



(c) Graph Instance #3



(d) Graph Instance #4

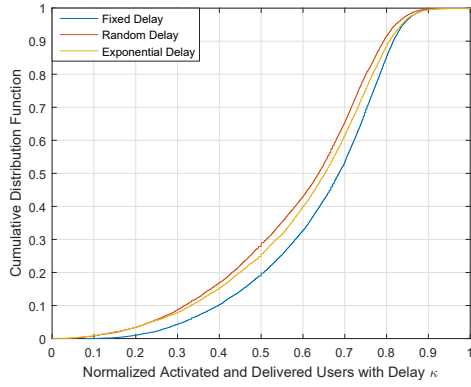
Figure 4.5: Performance for Different Delay Models ($\pi_0 = 11s$)

delay models. This suggests that our multicast and precache approach might help in content dissemination because it elevates the priority of the content in the scheduling system while reducing delays for all the users in the multicast group.

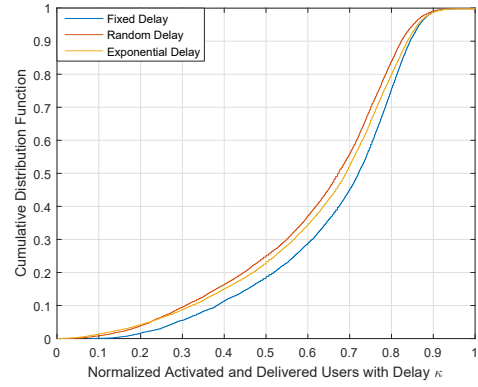
4.5.1.2 Different Delay Times

Given the results in 4.5.1.1, we focus on the performance for different fixed delays. We plot the cumulative distribution function of κ with different π_0 in Fig.4.7.

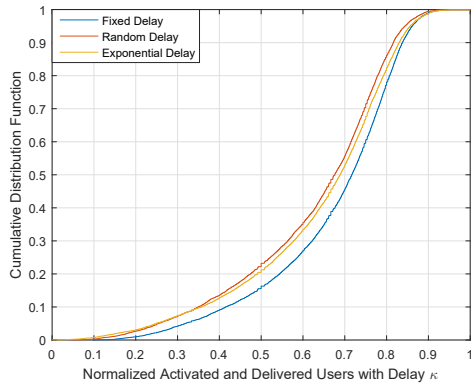
Apparently, without scheduling delay, the system achieves almost identical



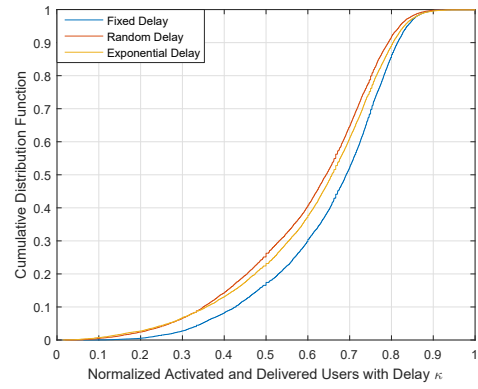
(a) Graph Instance #1



(b) Graph Instance #2

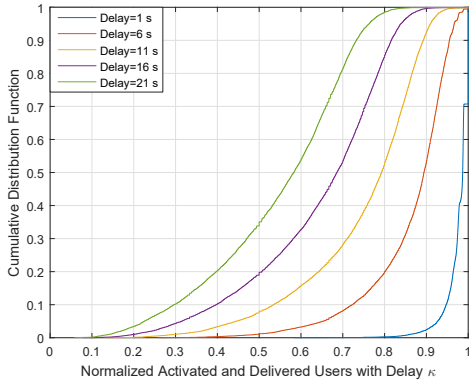


(c) Graph Instance #3

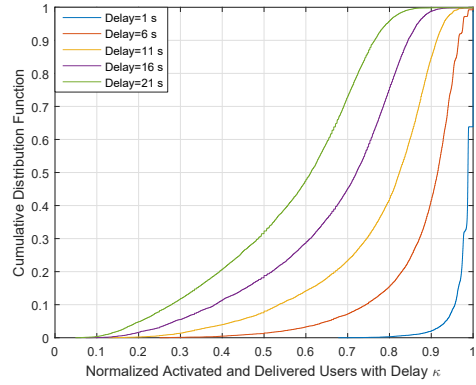


(d) Graph Instance #4

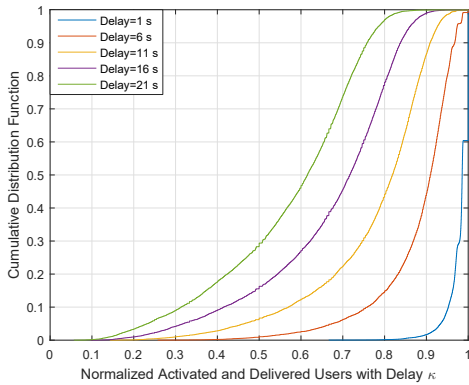
Figure 4.6: Performance for Different Delay Models ($\pi_0 = 16s$)



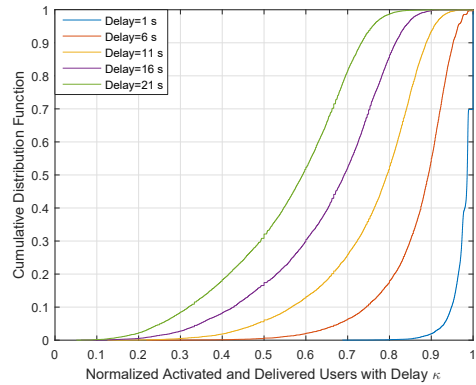
(a) Graph Instance #1



(b) Graph Instance #2



(c) Graph Instance #3



(d) Graph Instance #4

Figure 4.7: Performance for Different Fixed Delay Times π_0

performance ($\mathbb{P}[\kappa > 0.9] \approx 1$) as the ideal dissemination model without any delay.

As delay grows larger, the performance quickly starts to degrade significantly.

4.5.2 Scheduling Using Hybrid Systems

In this part, we investigate the performance for social content dissemination in our joint optimization scheduling framework for hybrid requests (‘push’ for system recommendation and ‘pull’ for active user requests). Unlike Chapter 2, we impose no deadlines, but rather introduce unbounded additional rewards for active user

requests.

Specifically, at time slot t , if user i is activated for content j , then there is an additional reward growing with time, controlled by weight factor ρ . The longer such requests are not delivered, the larger the (unbounded) additional reward.

$$f_{ij}^{(t)} = \gamma_{ij}^{(t)} \cdot \left(1 - \phi_{ij}^{(t)}\right) \cdot [f_{ij} + \rho (t - \lceil \mathcal{T}_{ij} \rceil)] \quad (4.34)$$

Of course, the system reduces to a ‘push’-only system when $\rho = 0$.

In the simulations, we choose the following parameters:

1. number of users $M = 300$, number of contents $N = 600$;
2. number of disseminating contents $N_d = 0.05N = 30$;
3. scheduling horizon $\mathcal{T}_H = 120s$, slot length $T = 1s$;
4. bandwidth $B = 20$ MHz.

We use MovieLens data [14] as the user reward values of the non-disseminating contents.

We map the nodes V_j of social dissemination graph G_j for content j randomly to the users.

All the curves in 4.5.2 (without predictions from social networks) are plotted in dash lines, compared with those with predictions (4.5.3) in solid lines.

The performance for disseminating contents is evaluated using the following metrics:

1. normalized activated and delivered users are plotted in Fig.4.8 (ideal situation yields $u(t - 1)$);

2. delay in Fig.4.9 (ideal situation yields $u(t)$);
3. number of total transmissions per content in Fig.4.10;
4. number of users served per transmission in Fig.4.11.

The overall reward for the system is shown in Fig.4.12.

Without any prioritization ($\rho = 0$), the disseminating contents suffer from significant performance degradation.

4.5.3 Look-Ahead Scheduling

As described in Section 4.4, we employ Monte Carlo Estimation to predict the activation states $\{\hat{\gamma}_{ij}^{(t)}\}$ based on the social dissemination graph without delivery delay. We use threshold value of $\xi = 0.3$ for estimation. Therefore, for scheduling purposes, we have:

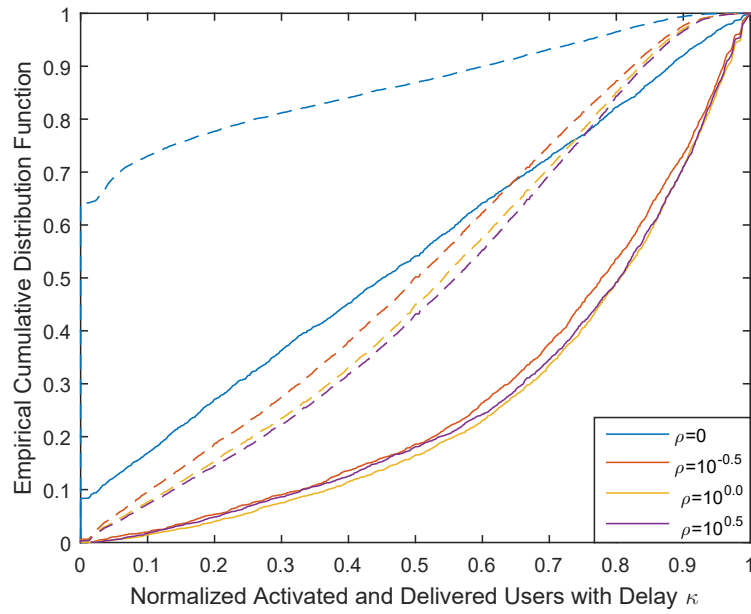
$$\hat{f}_{ij}^{(t)} = \hat{\gamma}_{ij}^{(t)} \cdot \left(1 - \phi_{ij}^{(t)}\right) \cdot \left[f_{ij} + \rho \cdot \gamma_{ij}^{(t)} \cdot (t - \lceil \mathcal{T}_{ij} \rceil)\right] \quad (4.35)$$

It is easy to verify that (4.25) and (4.35) are equivalent when $\rho = 0$. The additional rewards significantly improve the system performance.

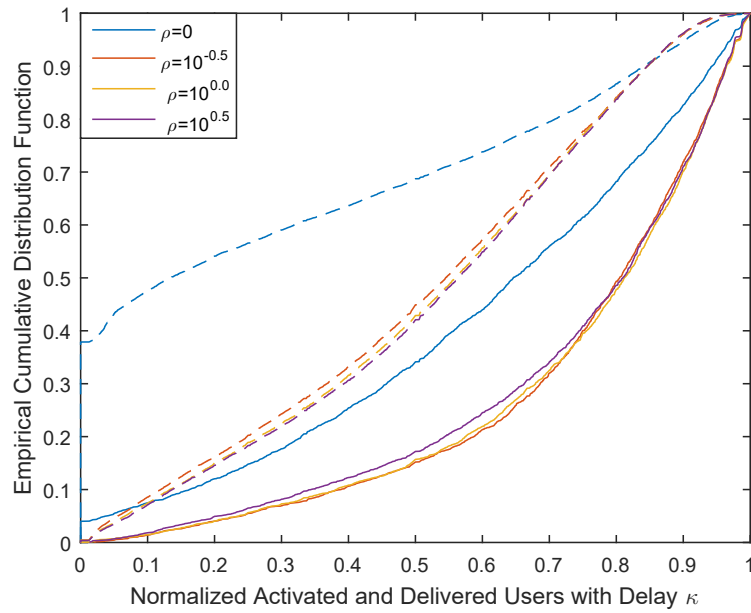
All the curves in 4.5.3 (with predictions from social networks) are plotted in solid lines, compared with those without predictions (4.5.2) in dash lines.

Compared with on-demand/no-lookahead scheduling (illustrated in dash lines), we note that with predictions (illustrated in solid lines):

1. The overall system reward is significantly improved (Fig.4.12).
2. The performance for disseminating contents is greatly improved (Fig.4.8).

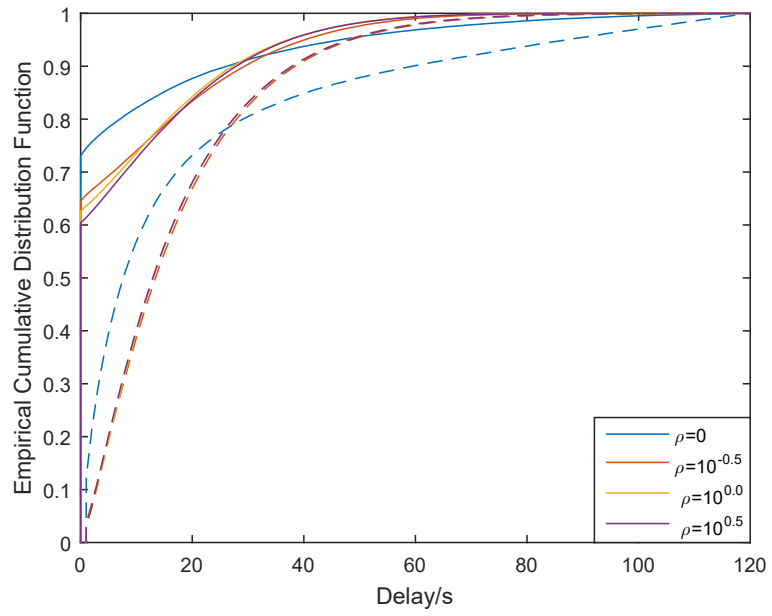


(a) ML-1M

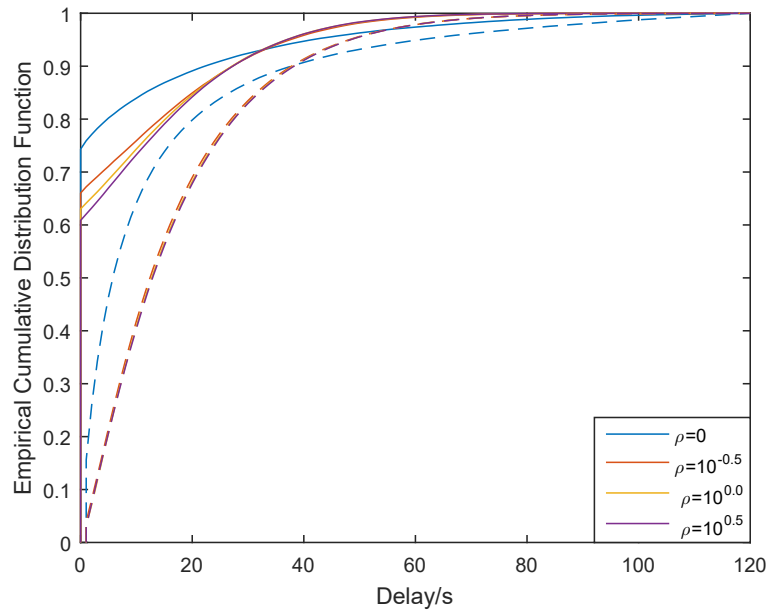


(b) Yahoo-Music

Figure 4.8: Social Performance for Disseminating Contents in Hybrid System

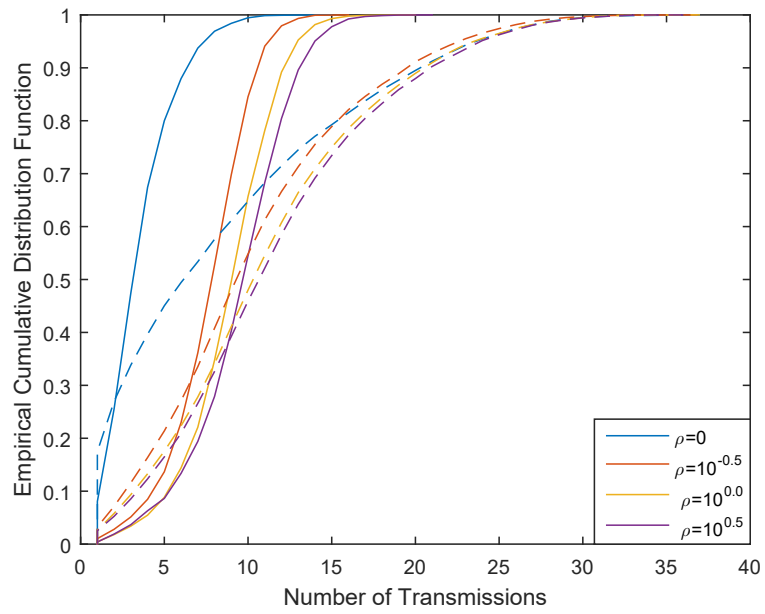


(a) ML-1M

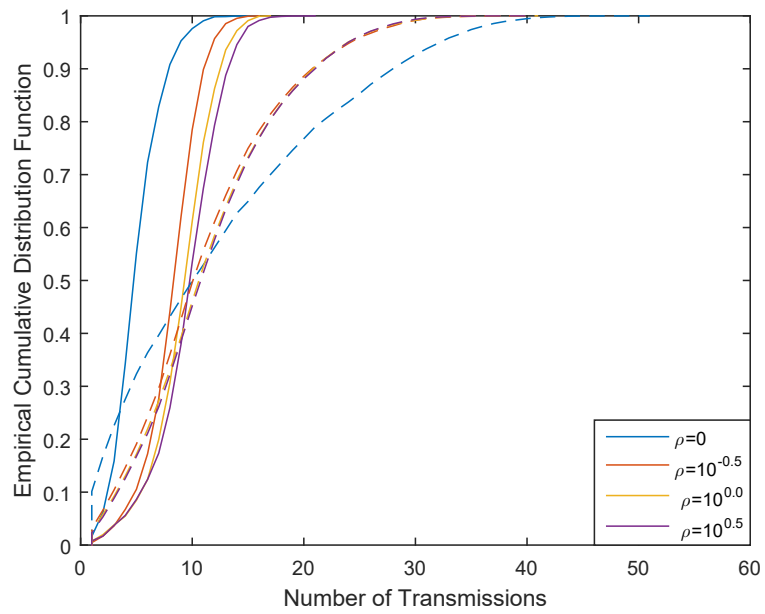


(b) Yahoo-Music

Figure 4.9: Delay for Disseminating Contents in Hybrid System

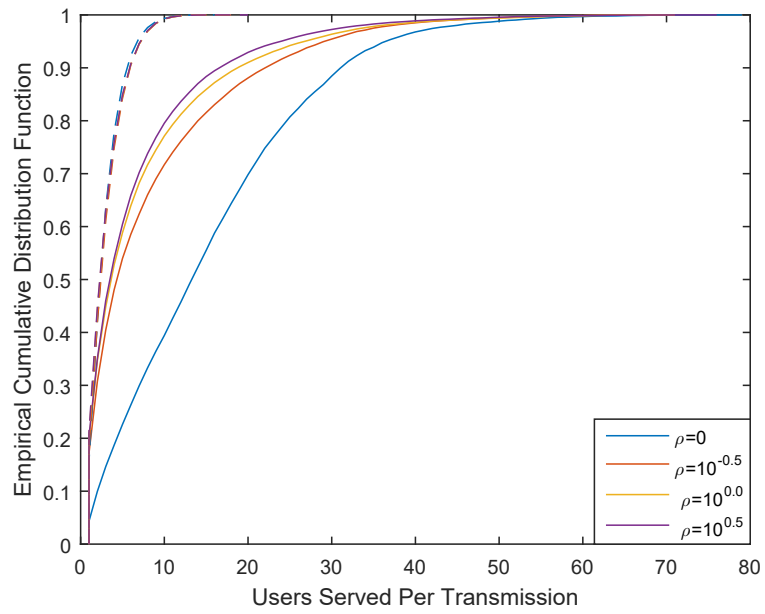


(a) ML-1M

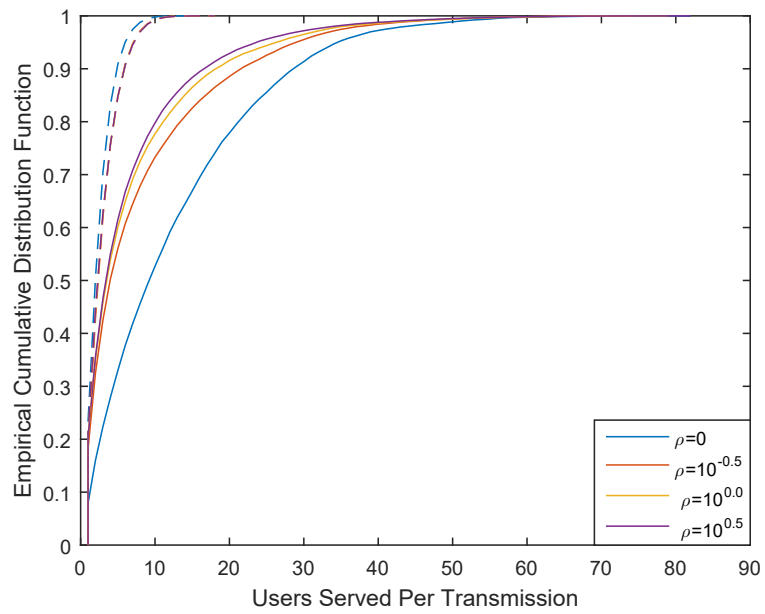


(b) Yahoo-Music

Figure 4.10: Total Transmissions per Content for Disseminating Contents in Hybrid System

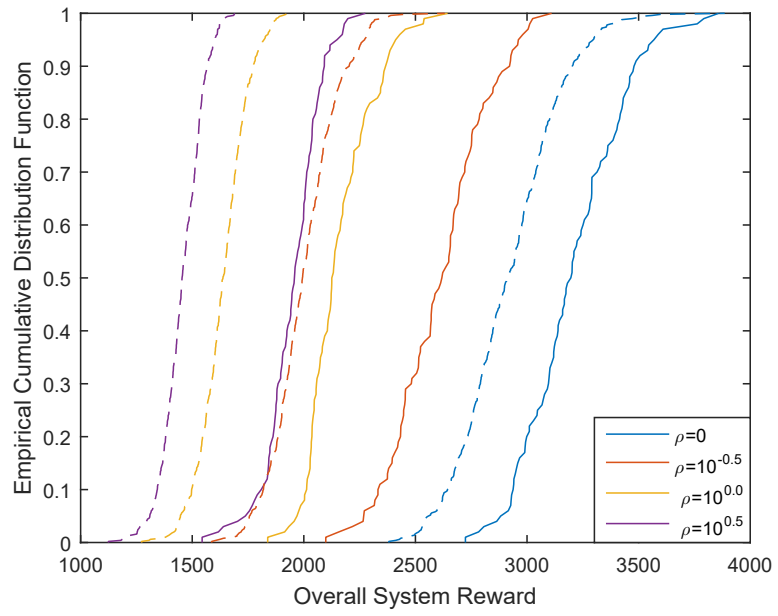


(a) ML-1M

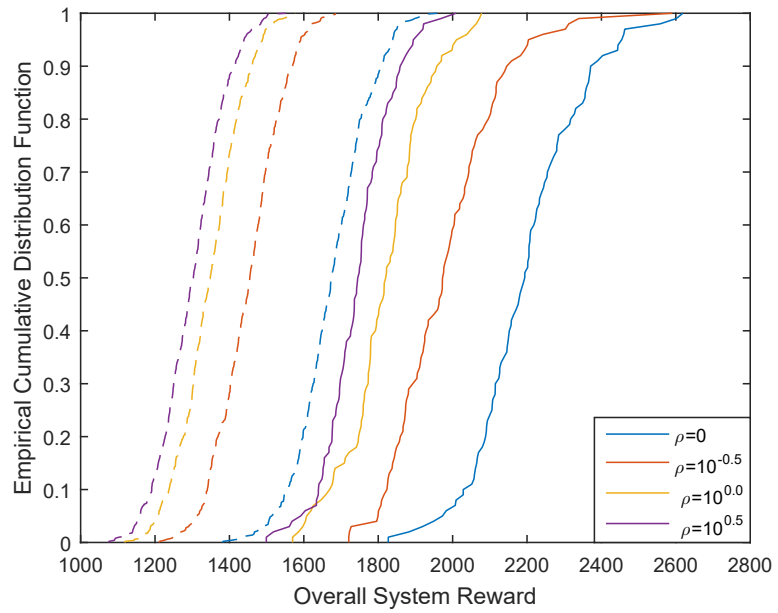


(b) Yahoo-Music

Figure 4.11: Total Transmission per Content for Disseminating Contents in Hybrid System



(a) ML-1M



(b) Yahoo-Music

Figure 4.12: Overall System Reward for Hybrid System

3. Delay for the disseminating contents is greatly improved and could reach 0 now due to the precaching(Fig.4.9).
4. Redundant transmissions are greatly reduced (Fig.4.10) and the users served per transmission are increased (Fig.4.11). This is great news for the system operator since better resource utilization usually brings about higher profits.

We also plot the ratio of missed deliveries (in which users are activated but do not receive the contents) versus ratio of unnecessary transmissions (users are not activated but are delivered the contents) in Fig.4.13, from 100 randomly selected disseminating contents up until the scheduling horizon \mathcal{T}_H .

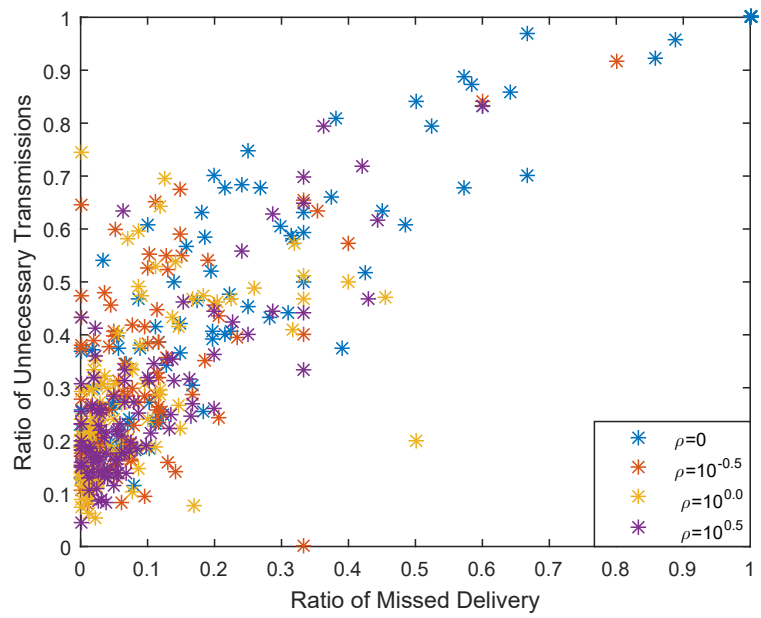
$$P'_F = \frac{|\{i : \gamma_i = 0 \wedge \phi_i = 1\}|}{|\{i : \phi_i = 1\}|} \quad (4.36)$$

$$P'_M = \frac{|\{i : \gamma_i = 1 \wedge \phi_i = 0\}|}{|\{i : \gamma_i = 1\}|} \quad (4.37)$$

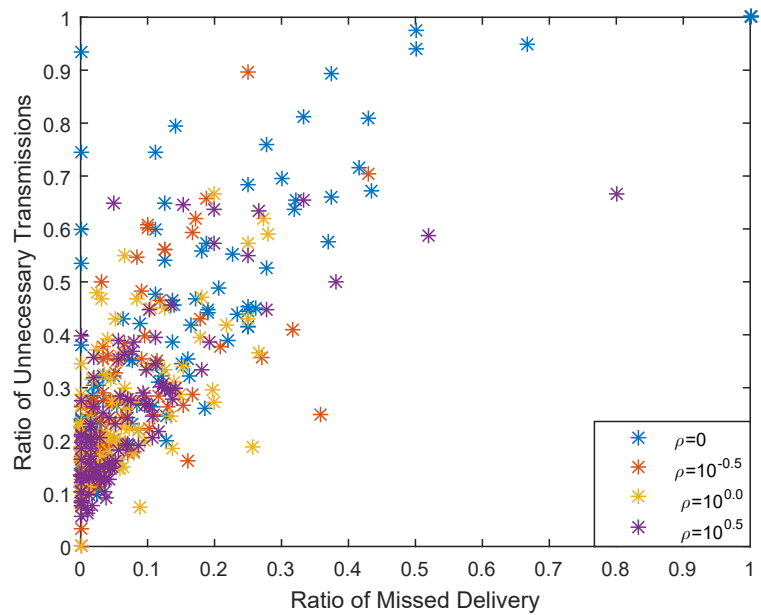
In fact, the definitions (4.36)(4.37) are somewhat identical to the concept in social networks (4.22)(4.23), as plotted in Fig.4.4. The empirical distributions of the ratio of unnecessary transmissions are plotted in Fig.4.14. We observe that the hybrid scheduling framework greatly improves the prediction precision. This is most probably due to the fact that the hybrid system drives towards actual user requests, increasing the number of activations in general and triggering more re-evaluations.

4.6 Summary

In this chapter, we investigate how to facilitate content dissemination in the presence of wireless capacity constraints. We leverage the predictions of social dy-

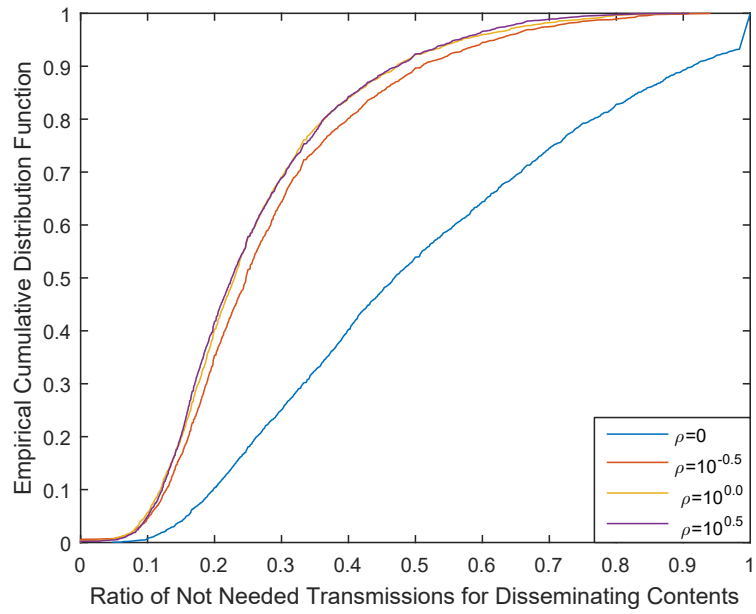


(a) ML-1M

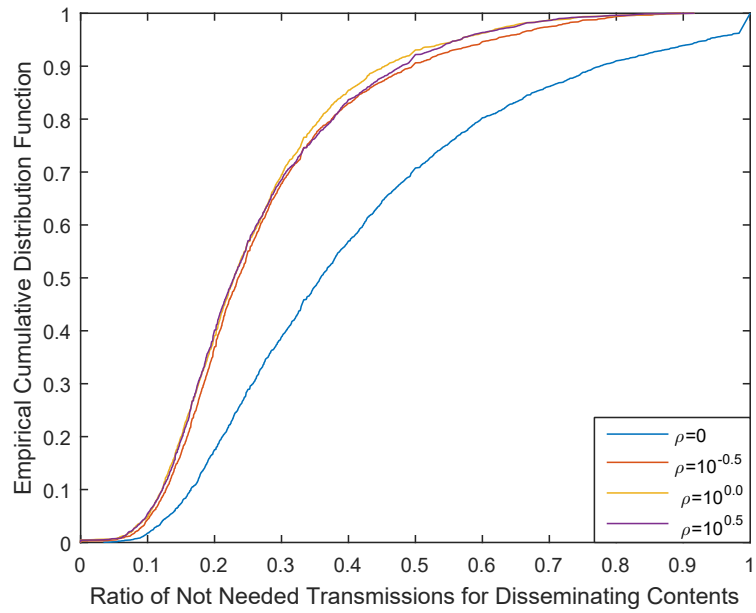


(b) Yahoo-Music

Figure 4.13: Prediction Precision for Hybrid System



(a) ML-1M



(b) Yahoo-Music

Figure 4.14: Ratio of Unnecessary Transmissions for Hybrid System

namics to transmit the contents before the users actually request them. Results indicate that with the help of predictions, we could mitigate the performance degradation for disseminating contents and improve the overall system performance. Results also indicate that the hybrid systems proposed in Chapter 2 greatly improve the performance of content dissemination, because they introduce proper prioritization for active user requests and thus prioritize social content dissemination.

Future work includes extending the analysis with more accurate predictions and extending the design to scenarios with multiple base stations.

Table 4.1: Summary of Variables

Notations	Definition
M	Number of users.
N	Number of contents.
f_{ij}	Reward for delivering content j to user i .
$\gamma_{ij}^{(t)}$	Binary activation state for user i to consume content j at time slot t .
$\phi_{ij}^{(t)}$	Binary delivery state whether content j has been transmitted to user i before time slot t .
I_i^{\leftarrow}	Influencer set for user i .
I_i^{\rightarrow}	Influencee set for user i .
τ_{vu}	Diffusion parameter of (exponential) influence time for user v to activate user u .
π_u	Wireless delivery delay for user u .
$\alpha_{ij}^{(t)}$	Binary decision variable whether to transmit content j to user i at time slot t .
$B^{(t)}$	Total available bandwidth at time slot t .
$s_j^{(t)}$	Wireless resource allocated for content j at time slot t .
$\text{SINR}_i^{(t)}$	Signal-to-Interference-Noise ratio of user i at time slot t
SINR_k^{th}	SINR threshold for transmission mode k
W_j	Size of content j in bits

Table 4.2: Monte Carlo Instance Runtime (in microseconds)

Min	P90	P95	P99	Max	Mean
2.8	58.8	77.4	137.6	900.9	32.6

Chapter 5: Conclusions and Future Work

In this dissertation, we propose a novel cross-layer joint optimization framework to schedule the delivery of social contents more efficiently, in the presence of wireless capacity constraints and changing wireless channels. Our proposed scheduling framework:

1. utilizes the social nature of contents and broadcast nature of wireless communication by multicasting the contents to groups of users;
2. takes advantage of reward predictions obtained from social networks using sophisticated machine learning algorithms and ‘big’ data of users;
3. requires minor changes of the existing system architecture and minimum information exchange between social networks and wireless networks;

Simulations indicate that our design would greatly improve user experience and increase the efficiency of spectrum resource, compared to the existing layered solutions. The major functions of social and wireless layers are still separated: the delivery scheduling is strictly implemented at the base stations, rather than at the central servers; the reward predictions are completed regardless of the volatility of user’s network conditions. The connection between the two layers is conveniently

made by assigning content-based reward, which drives the wireless networks to deliver one complete content package rather than just multiple packets.

We repeatedly take advantage of the limited number of multicast transmission modes to reduce the scale of our problems. In fact, in actual wireless communication systems, it is only possible to implement limited modes. Therefore, for each mode, we can conveniently aggregate users and reduce the optimization complexity. We further reduce the open hard problem involving multiple base stations to a set of simple and parallel feasibility tests. This pragmatic approach makes real-time online scheduling possible in practice.

Finally, we seamlessly incorporate the predictions of social dynamics into the scheduling system. Not only does our scheduling framework work well when the content reward remains unchanged during the scheduling horizon, it also works well when we have time-variant reward due to social dynamics, with the leverage of the hybrid systems.

5.1 Future Work

Future work based on this dissertation includes:

1. incorporating the uncertainty of reward prediction in the system: in this dissertation, we did not consider how the prediction error affects system performance;
2. allocating the resource dynamically:
 - (a) between multicast and unicast: certain requests, especially security and

privacy related ones, are inherently unicast;

(b) between different base stations: apparently, the allocation should be adaptive to the change of traffic load;

(c) incorporate frequency-selective and fast fading;

3. improving performance with better client-side design.

Bibliography

- [1] Tim Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and high Technology law*, 2:141, 2003.
- [2] William Stallings. *Handbook of Computer-communications Standards; Vol. 1: The Open Systems Interconnection (OSI) Model and OSI-related Standards*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1987.
- [3] Mung Chiang, Steven H Low, A Robert Calderbank, and John C Doyle. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 95(1):255–312, 2007.
- [4] Konstantinos Stathatos, Nick Roussopoulos, and John S Baras. Adaptive data broadcast in hybrid networks. Technical report, DTIC Document, 1997.
- [5] Yanan Bao, Xiaolei Wang, Sheng Zhou, and Zhisheng 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*, pages 566–571. IEEE, 2012.
- [6] O Shoukry, M Abd El-Mohsen, John Tadrous, Hesham El Gamal, Tamer El-Batt, N Wanas, Y Elnakieb, and M Khairy. Proactive scheduling for content pre-fetching in mobile networks. In *Communications (ICC), 2014 IEEE International Conference on*, pages 2848–2854. IEEE, 2014.
- [7] Omar K Shoukry and Magda B Fayek. Evolutionary scheduler for content pre-fetching in mobile networks. In *2013 AAAI Fall Symposium Series*, 2013.
- [8] Faisal Alotaibi, Sameh Hosny, Hesham El Gamal, and Atilla Eryilmaz. A game theoretic approach to content trading in proactive wireless networks. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2216–2220. IEEE, 2015.
- [9] K.C.J. Lin, C.W. Chen, and C.F. Chou. Preference-aware content dissemination in opportunistic mobile social networks. In *INFOCOM, 2012 Proceedings IEEE*, pages 1960–1968, 2012.

- [10] Xiangnan Weng and John 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.
- [11] Xiangnan Weng and John Baras. Joint Optimization for Social Content Delivery in Heterogeneous Wireless Networks. In *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2016 14th International Symposium on*, May 2016.
- [12] 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.
- [13] Raj Jain, Arjan Duresi, and Gojko Babic. Throughput fairness index: An explanation. Technical report, The Ohio State University, 1999.
- [14] MovieLens dataset. <http://www.grouplens.org/data/>, 2003.
- [15] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. TS 36.213, 3rd Generation Partnership Project (3GPP), 12 2015.
- [16] Ralf Irmer, Heinz Droste, Patrick Marsch, Michael Grieger, Gerhard Fettweis, Stefan Brueck, Hans-Peter Mayer, Lars Thiele, and Volker Jungnickel. Coordinated multipoint: Concepts, performance, and field trial results. *Communications Magazine, IEEE*, 49(2):102–111, 2011.
- [17] Yahoo! webscope dataset, ydata-ymusic-rating-study-v1_0-train. http://research.yahoo.com/Academic_Relations, 2009.
- [18] Rashmi R Sinha and Kirsten Swearingen. Comparing recommendations made by online systems and friends. In *DELOS workshop: personalisation and recommender systems in digital libraries*, volume 1, 2001.
- [19] Sergey Edunov, Carlos Diuk, Ismail Onur Filiz, Smriti Bhagat, and Moira Burke. Three and a half degrees of separation. <https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/>, February 2016.
- [20] Justin Cheng, Lada Adamic, P. Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, pages 925–936, 2014.
- [21] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.
- [22] Mohamed Akkouchi. On the convolution of exponential distributions. *J. Chungcheong Math. Soc.*, 21(4):501–510, 2008.

- [23] Jure Leskovec and Christos Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *Proceedings of the 24th international conference on Machine learning*, pages 497–504. ACM, 2007.
- [24] Jure Leskovec and Rok Sosič. SNAP: A general purpose network analysis and graph mining library in C++. <http://snap.stanford.edu/snap>, June 2014.