Fairness in Network-Friendly Recommendations

Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis

Fairness in Network-Friendly Recommendations

Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis

Fairness in Network-Friendly Recommendations

Theodoros Giannakas[†], Pavlos Sermpezis[‡], Anastasios Giovanidis[⋆],
Thrasyvoulos Spyropoulos[†], George Arvanitakis[‡]

[†]EURECOM, France; firstname.lastname@eurecom.fr
[‡]Aristotle University of Thessaloniki, Greece; {sermpezis, garvanitakis}@csd.auth.gr
[⋆]Sorbonne University, CNRS, LIP6, France; anastasios.giovanidis@lip6.fr

***Abstract*—As mobile traffic is dominated by content services (e.g., video), which typically use recommendation systems, the paradigm of network-friendly recommendations (NFR) has been proposed recently to boost the network performance by promoting content that can be efficiently delivered (e.g., cached at the edge). NFR increase the network performance, however, at the cost of being unfair towards certain contents when compared to the standard recommendations. This unfairness is a side effect of NFR that has not been studied in literature. Nevertheless, retaining fairness among contents is a key operational requirement for content providers. This paper is the first to study the fairness in NFR, and design fair-NFR. Specifically, we use a set of metrics that capture different notions of fairness, and study the unfairness created by existing NFR schemes. Our analysis reveals that NFR can be significantly unfair. We identify an inherent trade-off between the network gains achieved by NFR and the resulting unfairness, and derive bounds for this trade-off. We show that existing NFR schemes frequently operate far from the bounds, i.e., there is room for improvement. To this end, we formulate the design of Fair-NFR (i.e., NFR with fairness guarantees compared to the baseline recommendations) as a linear optimization problem. Our results show that the Fair-NFR can achieve high network gains (similar to non-fair-NFR) with little unfairness.**

I. Introduction

Background. The paradigm of network-friendly recommendations (NFR) has been very recently proposed as a promising solution for improving the quality and/or the cost of content delivery [1]–[21]. NFR is based on the fact that content traffic dominates the mobile traffic today [22], [23] and the majority of content services (online video, radio, social networks, etc.) employ recommendation systems (RS), which heavily affect the user choices and shape the content demand [24], [25]. The main idea behind NFR is to nudge the recommendations of the RS of the content provider towards content that can be delivered in a "network-friendly" way (e.g., cached in the mobile edge [2]–[10], or coded broadcast transmissions [15]–[17]), thus shaping the user demand in favor of this content.

The NFR paradigm involves three main parties: the network, the users, and the content provider. Existing works design NFR schemes that explicitly aim to benefit the *network*. Indeed, the envisioned network gains (lower load, congestion, resources, or costs) have been shown to be very promising [1]–[18]. Moreover, the *user* experience can improve as well due to the higher satisfaction from high quality content delivery [19]. Finally, there can be benefits for the *content provider* (e.g., higher user engagement); however, those have only been envisioned as a consequence of the higher user satisfaction, but have not been explicitly studied.

The problem: Fairness in NFR. To enable network benefits through NFR, the "cost" to be paid by the RS is that NFR (a) nudge the optimal recommendations list provided to users, which may lead to worse user satisfaction, and (b) bias the demand for different contents (by making some contents more and others less popular), which may lead to displeasure from the content owners/producers (e.g., YouTubers). The former (user perspective) has been explicitly taken into account in NFR schemes, by considering the *quality of recommendations (QoR)* in the nudged recommendations, e.g., by imposing a minimum threshold in the content similarity [4] or a window of user preferences [2]. However, the latter (content provider perspective) has been overlooked in related literature. In fact, the shaping of the content demand relates to the *fairness* of a RS towards the content producers/owners, which is a key requirement for content providers and has attracted a lot of attention recently in the design of RS [26]–[35].

On one hand, some unfairness due to NFR may be acceptable by the content providers under some conditions (during periods of network congestion, peak hours, etc.), in order to better satisfy the users or increase their engagement by avoiding serving them content in poor quality. However, previous works have not studied *how much unfairness is created by the NFR schemes*, and whether this is acceptable by the content provider. On the other hand, a content provider may need to satisfy some explicit fairness requirements for the contents (or, the content producers/owners), e.g., not allow a change in the demand larger than 5%. Up to now, this is not an option in

This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021) implemented by the State Scholarships Foundation (IKY), and the Operational Program Competitiveness, Entrepreneurship and Innovation under the call RESEARCH-CREATE-INNOVATE (project T2EDK-04937). It is also funded by the ANR (French National Agency of Research) by the "FairEngine" project under grant ANR-19-CE25-0011, by the ANR "5C-for-5G" project under grant ANR-17-CE25-0001, and the IMT F&R, "Joint Optimization of Mobile Content Caching and Recommendation" project.

the existing NFR schemes, since *fairness requirements have not been considered as a design aspect in NFR.*

Contributions. Motivated by this gap in literature, this paper is the first to study the aspect of fairness in NFR:

- ***Fairness characterization.*** We use metrics that capture different notions of fairness in RS (Section II), and then quantify the unfairness created in a wide range of representative scenarios and NFR algorithms, and investigate the role of different system parameters (Section III).
- ***The fairness vs. network gain trade-off.*** We identify an inherent trade-off between the network gains achieved by a NFR scheme and the resulting unfairness. We analytically study this trade-off and derive bounds. We show that existing NFR schemes, frequently operate far from the optimal operating point that is given by the bound (Section IV).
- ***Optimal Fair-NFR.*** We formulate the problem of designing NFR that maximize the network gain, under fairness guarantees compared to the baseline RS. Through a series of transformations, we show that the problem of optimal fair-NFR can be expressed as a linear program (Section V).
- ***The price of fairness.*** Studying the performance of the Fair-NFR scheme shows that *by allowing a little unfairness, high network gains can be achieved*, which is a promising message for the NFR paradigm. A comparison with (non-fair) NFR schemes demonstrates that *the Fair-NFR scheme achieves equal gains with much less unfairness* (Section VI).

II. Preliminaries

A. Network-friendly Recommendations

We consider a content service that has integrated in its (web/mobile) platform a recommendation system (RS). When a user is in the platform and consumes (e.g., watches, listens to, reads, buys) a content, a list of recommendations is presented by the RS suggesting to her to consume another content next. This is a typical scenario for the majority of online video/radio services (e.g., YouTube, Netflix, Spotify), news sites, e-shops and online marketplaces (e.g., Amazon), online social networks (e.g., Facebook, Instagram), etc. In the following, we describe the generic setup considered in NFR; the main notation is summarized in Table I.

Content service. Assume that the service has a content catalog $\mathcal{K}$ ($|\mathcal{K}| = K$). Users request contents in two ways: (i) directly, e.g., by following an external link or typing the content through a search bar, or (ii) by following one of the recommendations provided by the RS of the service (users typically consume several contents when visiting the service). These are the main types of demand in most content services.

We define the *demand* p_i for a content i as the fraction of all requests (i.e., direct and through recommendations) that are for this content; we denote as $\mathbf{p} = [p_1, ..., p_K]$ the vector with the distribution of total demand for all contents.

Network. We assume that a subset of the content catalog $\mathcal{C} \subset \mathcal{K}$ can be delivered with low cost for the network (and/or in high quality). For instance, in the context of mobile edge caching considered by the majority of related work in NFR [1]–[4], [7]–[9], the contents in $\mathcal{C}$ are cached in the mobile edge. In this context, and w.l.o.g., we set the cost for delivering contents in $\mathcal{C}$ to zero and the cost for the other contents to 1. Hence, the cache hit ratio, $CHR = \sum_{i\in\mathcal{C}} p_i$, captures the total benefit for the network (i.e., the decrease in the cost by using a cache).

Recommendations. We assume a "recommendation score" u_{ij} for every pair of contents $i, j \in \mathcal{K}$, which indicates how good a recommendation for content j after content i is. The score u_{ij} may correspond to the similarity between two contents, or more generally to the relevance of recommending j after i (e.g., capturing from item-item collaborative filtering [36] to black-box deep learning architectures [37]), and can be the output of any state-of-the-art RS. W.l.o.g, we assume $u_{ij} \in [0, 1]$ and higher values denote better recommendations.

Baseline RS (BS-RS) is the standard RS (i.e., non network-friendly) that generates the recommendation scores u_{ij} and is used in production by the content/service provider. After a user has consumed content i, the BS-RS recommends to the user a list R_i^{BS} that contains the N contents with the highest recommendation score values u_{ij}.

Network-friendly RS (NF-RS) is a RS that takes into account the network conditions (e.g., delivery cost [4], [19], cached contents [3], [8], wireless channel [16], [17]) and provides a list of recommendations R_i^{NF} to the user. In general, the lists R_i^{NF} can be the same as those of the BS-RS R_i^{BS}, partially overlap with them, or be totally disjoint sets. Typically, the recommendations of NF-RS tend to (i) include more recommendations to contents that can be delivered in a network-friendly way (e.g., cached contents), while (ii) trying to maintain the quality of recommendations (QoR) by recommending contents with relatively high scores u_{ij}. In a simple example, with one user, three contents a, b, c with scores $u_a = 1$, $u_b = 0.8$, $u_c = 0.5$, and a BS-RS recommending only one content $R^{BS} = [a]$. Let only $b, c \in \mathcal{C}$ be cached; then the NF-RS would recommend $R^{NF} = [b]$, because this would bring network gains, and would have QoR$= \frac{u_b}{u_a} = 0.8$, which is higher than if c was recommended instead of b (QoR=0.5).

Finally, the resulting demand $\mathbf{p}$ depends on the underlying RS: a RS that selects more frequently a content i in the recommendation lists, will lead to an increase in the demand p_i. In the remainder, we denote with a superscript the RS that corresponds to the content demand, e.g., $\mathbf{p^{BS}}$ for the BS-RS and $\mathbf{p^{NF}}$ for a NF-RS. The differences between the vectors $\mathbf{p^{BS}}$ and $\mathbf{p^{NF}}$ capture the notion of fairness, which we formally define below.

B. Fairness Definition

Content providers aim to satisfy two parties, their users (consumers) and the content owners (producers) [26]–[28], [32], [38], while at the same time maximizing their own utility (e.g., revenue) [26]. In general, the goal of a *fair RS* is to strike a balance between *utility* and *satisfaction* of the involved parties [32], [38].

TABLE I: Important notation.

$\mathcal{K}$	content catalog ($\lvert\mathcal{K}\rvert = K$)
p_i	demand for content i; $\mathbf{p} = [p_1, ..., p_K]$ and $\sum_{i\in\mathcal{K}} p_i$=1
$\mathcal{C}$	set of cached contents ($\lvert\mathcal{C}\rvert = C$)
CHR	cache hit rate, $CHR = \sum_{i\in\mathcal{C}} p_i$
u_{ij}	recommendation score, $u_{ij} \in [0, 1]$
R_i	list of recommendations after content i
N	number of recommendations.
α	probability a user to follow a recommendation
$p_i^{(d)}$	probability that a "direct request" is for content $i \in \mathcal{K}$; $\mathbf{p^{(d)}} = \left[p_1^{(d)}, ..., p_K^{(d)}\right]$ and $\sum_{i\in\mathcal{K}} p_i^{(d)} = 1$
G	network gain, $G = CHR^{NF} - CHR^{BS}$

In the context of NFR, user satisfaction is taken into account with the concept of quality of recommendations (QoR). However, the content owner/producer satisfaction, which is identified as a key component in the design of fair RS, especially in multistakeholder settings [26], [27], has been neglected in previous works in NFR. Hence, we focus on the need of the content provider to satisfy content owners/producers, by providing recommendations that are fair with respect to them (which in literature is referred to also as *p-fairness* [26], [27]).

Fairness in RS can be defined in several ways [26]–[34], depending on the system, the involved parties, the needs of the content provider, etc. The fairness of a RS can be measured with respect to the recommendations of *a fair RS*. In our setting, where the goal is to quantify the (un)fairness of NFR, this fair RS is by convention the BS-RS (i.e., any standard RS) and the fairness captures the deviation in the total demand $\mathbf{p}$ created by the NF-RS. Thus, a generic measure F can be used:

$$F = f(\mathbf{p^{BS}}, \mathbf{p^{NF}}) \tag{1}$$

In general, the function f can be defined at will according to the use case or requirements of the content provider. For example, [32] suggests that f can be any probability divergence measure. Different measures f can capture different notions of fairness. In this paper, we consider the three fairness measures that are most commonly used in literature and practice[1]:

F-max $F_{max} = \max_{i\in\mathcal{K}} |p_i^{NF} - p_i^{BS}|$ relates to the *individual fairness* [31] and accounts for the "worst case", i.e., no content will have a demand difference larger than F_{max}.

F-tv $F_{tv} = \frac{1}{2} \cdot \sum_{i\in\mathcal{K}} |p_i^{NF} - p_i^{BS}|$ is the *total variation distance* between the two distributions, i.e., the average (absolute) change in the content demand [30]. It allows more flexibility than F_{max} in shaping the demand, since it does not impose a constraint for every single content; e.g., a large difference in a content demand can be compensated by small demand differences in other contents.

F-kl $F_{kl} = \sum_{i\in\mathcal{K}} p_i^{BS} \cdot \log\left(\frac{p_i^{BS}}{p_i^{NF}}\right)$ is the Kullback–Leibler (KL) divergence, a widely used measure for the difference between distributions, and commonly used to quantify fairness in RSs [32]–[34]. F_{kl} is more sensitive to changes in contents with lower demand, e.g., an increase Δp in the demand p_i^{BS} leads to a higher increase in F_{kl} when p_i^{BS} is small [33].

[1]Note that, since we aim to capture the fairness in recommendations, we use metrics from the RS field. Other fairness measures from other fields, e.g., resource allocation [39], would be less relevant.

Remark: Note that $F_{max}, F_{tv} \in [0, 1]$, whereas $F_{kl} \in [0, \infty]$ ($F_{kl} \to \infty$ when $p_i^{NF} = 0$ and $p_i^{BS} \neq 0$). For the sake of presentation, in the results we normalize the values of F_{kl} so that it takes values in $[0, 1]$ and is comparable with the other metrics. In particular, we use the smoothed version of [32], [33], where we substitute $p_i^{NF} \to (1 - w) \cdot p_i^{NF} + w \cdot p_i^{BS}$, with $w = 0.01$ and normalize with its upper bound $\log \frac{1}{w}$; i.e.,

$$F_{kl} = \frac{1}{\log \frac{1}{w}} \cdot \sum_{i\in\mathcal{K}} p_i^{BS} \cdot \log\left(\frac{p_i^{BS}}{(1-w)\cdot p_i^{NF} + w\cdot p_i^{BS}}\right)$$

The above metrics reflect different notions of fairness and requirements of the content provider. In general, it is not possible to satisfy all notions of fairness at the same time [31]. In this paper, we consider all these metrics, and study their characteristics in relation to NFR (Sections III and IV) and take them into account in the design of fair NF-RS (Section V).

QoR vs. fairness. As a remark, we stress that the notions of QoR (considered in previous works) and fairness (not considered before) describe *different* quantities in NFR; the former relates to the satisfaction of the users/consumers, and the latter to the satisfaction of the content owners/producers. The following example demonstrates this distinction: Let two users and three contents a, b, c with scores $u_a = 1$, $u_b = 0.8$, $u_c = 0.8$ (same for both users), and $b, c \in \mathcal{C}$, i.e., are cached. The BS-RS recommends content a, with the highest score u, to both users. Let's assume two NF-RS that nudge the BS-RS recommendations towards cached contents: A NF-RS recommends b to both users, and another NF-RS recommends b to the first user and c to the second user. Since, $u_b = u_c$ the QoR in both NF-RS is the same. However, the former NF-RS is less fair, e.g., in terms of *F-max*, since it increases twice the demand for content b compared to the latter NF-RS.

III. Characterization of Unfairness in NFR

In this section, we aim to understand the (un)fairness F in NFR. To this end, we employ an empirical approach where we (i) consider a wide range of scenarios, (ii) apply the BS-RS and different NF-RS algorithms that have been proposed in previous works, and (iii) calculate the resulting content demand and its unfairness (Section III-A). We analyze the results to investigate whether existing NFR schemes create unfairness, and what are the key factors that cause it (Section III-B).

A. Simulation Setup

Content catalogs. We consider content catalogs and matrices $\mathbf{U} = \{u_{ij}\}$ extracted from two datasets of real services:

Last.fm. We use a dataset from the Last.fm database [40], where we applied the "getSimilar" method to the content IDs' and populate the matrix $\mathbf{U}$. As the resulting $\mathbf{U}$ matrix is quite sparse, for the purpose of demonstration, we keep the largest component of the underlying graph, and round to $u_{ij} = 1$ the values above a threshold of 0.1.

MovieLens. We use the Movielens movies-rating dataset [41], containing 69162 ratings (0 to 5 stars) of 671 users for 9066

TABLE II: Parameters of the simulation scenarios (in total, all their combinations give 1296 scenarios)

U: last.fm ($K=757$), Movielens ($K=1060$)	$N \in \{2,5,10\}$
$\mathbf{p^{(d)}} \sim \{Zipf(s=1), uniform\}$	$C \in \{5,10,20\}$
$\alpha \in \{0.5, 0.8, 0.99\}$	$q \in \{0.5, 0.8, 0.9\}$
$W_{BFS} \in \{N, 2N\}$	$D_{BFS} \in \{1,2\}$

movies. We apply an item-to-item collaborative filtering (using 10 most similar items) to extract the missing user ratings, and then use the cosine distance to calculate the similarity for each pair of contents. We set $u_{ij} = 1$ for contents with cosine distance larger than 0.6, and 0 otherwise.

Caching. We consider cache sizes $C \in \{5,10,20\}$, with a popularity-based caching policy, i.e., the cache contains the C contents with the highest demand under the BS-RS ($\mathbf{p^{BS}}$).

Content demand. Similarly to previous works [4], [7], [8], [12], [15], [17], we assume that a user follows a recommendation with probability α, or directly requests a content with probability $1-\alpha$. We set $\alpha \in \{0.5, 0.8, 0.99\}$, to capture the behavior reported for YouTube (α=0.5) [24] and Netflix (α=0.8) [25] and an extreme value where users follow almost always recommendations (α=0.99), e.g. as in YouTube autoplay or online radio services like Last.fm, Jango, etc..

We assume that direct requests for different contents follow a Zipf distribution with exponent s, where we used a typical scenario with $s=1$ [24] and an extreme scenario with $s=0$ (i.e., uniform distribution). We denote the distribution of direct requests as $\mathbf{p^{(d)}}$.

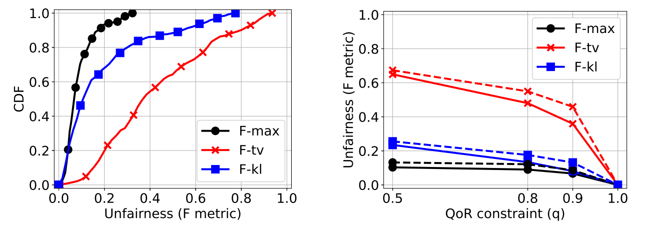
NF-RS algorithms. Several NFR variants have been proposed. To avoid restricting our study to a single algorithm or setup, we consider three representative NF-RS algorithms.

Greedy NF-RS includes in each recommendation list R_i as many cached contents as possible, without violating a minimum QoR threshold q. It aims to maximize the CHR by considering every request independently (without taking into account the long term performance). It can be seen as a simplified version of only the recommendation part of the CawR algorithm [8] (with the cache assumed already filled)[2], or the "Myopic" version of CARS [4].

Multi-step NF-RS [12] is an algorithm that includes in each recommendation list R_i a set of contents that satisfy a QoR constraint (similarly to the Greedy NF-RS) and maximizes the network gains in the long term, i.e., by taking into account requests made directly and through recommendations, and the probability α. It returns the optimal solution in our model setup under *no fairness* requirements.

CABaRet [5] follows a different approach, by leveraging the BS-RS and assuming no explicit knowledge on the scores u_{ij}. For each content i, it does a breadth-first search (BFS) starting

[2]We note that CawR [8] optimizes at the same time the caching and recommendation policies. Since the scope of this paper is on the fairness of the recommendations in NF-RS, we focus on the resulting recommendations of NF-RS algorithms, given a pre-filled cache; we discuss implications of joint NF-RS and caching policy optimization algorithms in Section VIII.



(a) CDF of unfairness (b) Unfairness vs. QoR

Fig. 1: (a) CDF of unfairness created by the NF-RS algorithms in all scenarios (Table II). (b) Unfairness vs. QoR, in Movielens scenarios with α=0.8, N=5, C=10, uniform $\mathbf{p^{(d)}}$, Greedy (continuous lines) and Multi-step (dashed lines) NF-RS.

from the list R_i^{BS} (depth 1), and then to the lists R_j^{BS}, $\forall j \in R_i^{BS}$, (depth 2), and so on. It returns a recommendation list that contains the cached contents found in the BFS and, if needed, fills the list with the initial recommendations R_i^{BS}.

In all cases recommendation lists are of size $N \in \{2,5,10\}$.

Quality of Recommendations (QoR). In the Greedy and Multi-step NF-RS, the QoR constraint is explicitly defined as a fraction of the recommendation quality of the BS-RS by a parameter $q \in [0,1]$, i.e., $\sum_{j \in R_i^{NF}} u_{ij} \geq q \cdot \sum_{j \in R_i^{BS}} u_{ij}$ [4], [12]. In CABaRet, the QoR is implicitly determined by the width W_{BFS} and depth D_{BFS} parameters of the BFS [5]. In our simulations, we consider values $q \in \{0.5, 0.8, 0.9\}$, and $W_{BFS} \in \{N, 2N\}$ and $D_{BFS} = \{1,2\}$.

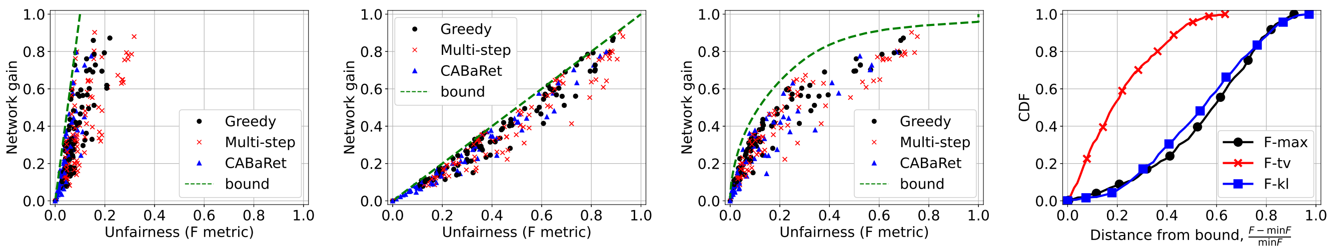
Table II summarizes the parameters of the considered scenarios. In total, we simulated 1296 scenarios, accounting *for all the combinations* of the parameters.

B. Unfairness in NFR

In the scenarios we simulate, we calculate the content demand under the BS-RS ($\mathbf{p^{BS}}$) and the different NF-RS algorithms ($\mathbf{p^{NF}}$), and then the resulting unfairness captured with the metrics $F(\mathbf{p^{NF}}, \mathbf{p^{BS}})$ defined in Section II-B.

Unfairness in existing NF-RS algorithms. We first quantify the unfairness created by existing NF-RS algorithms. Figure 1(a) presents the CDF of the values of the fairness metrics F among all the scenarios we tested; large values of the fairness metric F denote more unfair systems (e.g., the system is fair for F=0 and very unfair for F=1). We see that *the NF-RS algorithms create unfairness, which is very high in several cases* (we remind that the presented F metrics take values in $[0,1]$). Moreover, comparing the curves of the different metrics (or, notions) of fairness, we can see that F_{max} that captures the individual fairness takes lower values, whereas F_{tv} that is averaged over all contents takes the highest values (even up to 1). The CDF of F_{kl}, which considers all contents while also giving emphasis on individual contents whose demand deviates a lot from $\mathbf{p^{BS}}$, lies between the other two metrics.

The role of the QoR. Figure 1(b) presents the resulting unfairness (y-axis) by applying the Greedy NF-RS (continuous lines) and the Multi-step NF-RS (dashed lines) in scenarios



(a) F_{max} (b) F_{tv} (c) F_{kl} (d) CDF of distance from bounds

Fig. 2: (a),(b),(c): Network gain G (y-axis) vs. fairness metric F (x-axis) in all scenarios of Table II with $C = 10$, and under the *Greedy* (circles), *Multi-step* (crosses), *CABaRet* (triangles) NF-RS. The bounds are denoted with dashed lines. (d): CDF of the relative distance $\frac{F-\min F}{\min F}$ of the operating points (F, G) of the NF-RS algorithms from the corresponding bounds $(\min F, G)$.

with different QoR constraints q (x-axis). It is clearly seen that as the QoR constraint becomes looser (lower values in x-axis) the unfairness of the system increases (higher values in y-axis). This is due to the fact that by relaxing the required QoR, the NF-RS has more flexibility in changing the recommendation lists, and consequently this nudges the content demand $\mathbf{p^{NF}}$ farther from $\mathbf{p^{BS}}$. However, it is interesting to note the change in unfairness is not linear to the QoR, but is rather a concave function that decreases more steeply for higher values of QoR. In fact, a key observation in Fig. 1(b) is that even a small decrease in QoR from the maximum value that corresponds to the BS-RS ($q = 1$), can already lead to significant unfairness (e.g., see the F values for $q = 0.9$). This finding (similar behavior holds in all the scenarios we tested) highlights the following insight, which is a main motivation for this paper:

"The QoR constraint commonly used in NF-RS to satisfy the users, may not suffice to (implicitly) impose fairness for the content provider as well. To account for fairness, one needs to explicitly take it into account when designing a NF-RS."

The role of the NF-RS algorithm and the system parameters. Comparing the curves of the two NF-RS algorithms in Fig. 1(b), we see that the unfairness introduced by the Multi-step NF-RS is higher than the Greedy NF-RS, under any fairness metric F. This finding holds in all (for F_{tv}, F_{kl}) and in 95% (for F_{max}) of the scenarios we tested, and is due to the fact that the Multi-step NF-RS, by accounting the long term behavior, can shape in a larger degree than the Greedy NF-RS (or other heuristics) the content demand under the same QoR constraint. Hence, on the one hand the Multi-step NF-RS achieves *higher* network gains, but on the other hand it leads to *less* fairness (e.g., 10% higher CHR and 45% higher F_{kl} than Greedy among all scenarios).

TABLE III: Relation between system parameters, fairness F and network gain G: monotonicity and correlation (ρ).

	F_{max}	F_{tv}	F_{kl}	G
$q \nearrow$	$\searrow$ (ρ=-0.40)	$\searrow$ (ρ=-0.38)	$\searrow$ (ρ=-0.30)	$\searrow$ (ρ=-0.42)
$\alpha \nearrow$	$\nearrow$ (ρ=0.46)	$\nearrow$ (ρ=0.81)	$\nearrow$ (ρ=0.75)	$\nearrow$ (ρ=0.69)
$N \nearrow$	$\searrow$ (ρ=-0.47)	$\searrow$ (ρ=-0.13)	$\searrow$ (ρ=-0.16)	$\searrow$ (ρ=-0.20)
$C \nearrow$	$\searrow$ (ρ=-0.13)	– (ρ=0.06)	– (ρ=0.03)	$\nearrow$ (ρ=0.18)

We observed the same relation between fairness and network gains, when varying the other system parameters as well; see Table III. Specifically, increasing the α means that the users choices are affected more by the RS, and the same happens for small N since there are less choices (recommendations); this makes the shaping of the demand caused by a NF-RS more intense, and leads to higher network gains, and as we present in Table III, less fairness as well. The cache size C does not significantly affect the fairness, but it also had a small effect on the network gains in the scenarios we tested.

The above observations raise the following question, on which we focus in the next section:

"Does great network gain come with great unfairness?"

IV. The Fairness vs. Network Gains Trade-off

In this section we proceed to study the trade-off between the network gains that can be achieved by a NF-RS algorithm and the unfairness it creates. We first analyze the simulation results to verify that such a trade-off exists, and then study it analytically and derive analytic bounds (closed form expressions) for the minimum possible unfairness as a function of the network gains under any NFR scheme.

Let us first formally define the *network gain* G as the increase in the cache hit rate (CHR) achieved by a NF-RS:

$$G = CHR^{NF} - CHR^{BS} = \textstyle\sum_{i\in\mathcal{C}}(p_i^{NF} - p_i^{BS}) \qquad (2)$$

where $\mathcal{C} \subset \mathcal{K}$ is the set of cached contents. In other words, the network gain is the extra content demand that can be served by the cache when applying a NF-RS.

In Fig. 2 we present scatter plots, where each marker corresponds to a simulation scenario and its (x, y)-coordinates correspond to the resulting fairness metric F and network gain G values, respectively. The results verify our previous observations: as the achieved network gain increases, the unfairness of the system increases as well. This positive correlation holds for all fairness metrics. However, the exact behavior differs among the different F metrics (note that all subplots of Fig. 2 present the same simulation scenarios, i.e., with the same network gains); for instance, F_{max} sees a lower increase and with values up to 0.3, while F_{tv} has a larger increase with values up to 1.

In the following theorem we analytically study the observed behavior, and derive theoretical bounds for the trade-off.

Theorem 1. *Under any NF-RS and any system parameters, for the fairness F vs. network gain G trade-off it holds that*

$$F_{max} \geq \frac{1}{C} \cdot G \tag{3}$$

$$F_{tv} \geq G \tag{4}$$

$$F_{kl} \geq -H \cdot \log(1 + \frac{G}{H}) - (1 - H) \cdot \log(1 - \frac{G}{1-H}) \tag{5}$$

where $C = |\mathcal{C}|$ is the number of cached contents, and $H = CHR^{BS} = \sum_{i \in \mathcal{C}} p_i^{BS}$.

Proof. The proof is given in the Appendix. □

The expressions in Theorem 1 state that the maximum network gain that can be achieved by any NF-RS (i) cannot be larger than the desired F_{tv} value, and (ii) increases with the cache size C in the F_{max} case. This indicates that larger caches can allow network gains without compensating in *individual fairness* (F_{max}), while this is not the case in *aggregate fairness* (F_{tv}). In the case of F_{kl}, the bound is given by a non-linear function (convex in G), which depends on the cache size and the distribution of the demand under the BS-RS (captured by the parameter $H = CHR^{BS}$).

Comparing the results in Fig. 2 with the bounds, we can see that in some scenarios the achieved network gains are close to (or, coincide with) the bound, i.e., *the bounds are tight*[3].

However, in the majority of scenarios, *the operating point of the considered NF-RS algorithms is far from the bound*. For instance, in Fig 2(d) that gives the CDF of the distance (along the x-axis) between the operating points and the bound, we can see that in half of the cases (i.e., for 0.5 in the y-axis) the resulting unfairness (x-axis) is at least 50% larger than the value of the bound for F_{max} and F_{kl} and 20% larger for F_{tv}. The fact that NF-RS algorithms do not operate on the bound can be due to (i) the system parameters (e.g., the QoR constraint) that restrict an algorithm from shaping arbitrarily the content demand, and in this case the bound may not be achievable, or (ii) the NF-RS algorithms themselves, which were designed to optimize the network gain without taking the fairness into account. Thus, a question that follows naturally is:

"Can a NF-RS be designed to operate closer to the bound, and achieve the optimal fairness vs. network gain trade-off?"

In Section V we address the above question and design an optimal NF-RS algorithm that achieves the maximum network gain under a fairness constraint, and in Section VI we study how the introduced constraint affects the network gains.

[3]Note that some of the scenarios/markers in Fig. 2(c) (F_{kl} case) correspond to different values CHR^{BS} (e.g., due to different $\mathbf{p^{BS}}$ distributions), and thus different bounds. In favor of readability, we avoid depicting several bounds and show only the worst-case bound among those scenarios; i.e., for some scenarios/markers the bound is tighter than the depicted bound.

V. OPTIMAL FAIR NFR

In this section we formulate the problem of designing the optimal NF-RS that takes fairness into account. We first model and describe the problem, and then prove that it can be expressed as a linear program (LP) whose solution is the *optimal fair NF-RS*.

Objective. The objective in NFR is to maximize the network gains (or, equivalently minimize the network cost), which in our framework is captured by the CHR, i.e., $\sum_{i \in \mathcal{C}} p_i^{NF}$.

Decision variables. An NF-RS algorithm selects which contents to recommend, i.e., the recommendation lists R_i[4]. We model the recommendation decisions with a set of variables r_{ij}, which denote the probability (or frequency) that a content j appears in the recommendation list of i, i.e., $r_{ij} = \text{Prob}\{j \in R_i\}$. We denote as $\mathbf{R}$ the $K \times K$ matrix that contains all the variables r_{ij} $\forall i, j \in \mathcal{K}$. We follow a probabilistic approach, i.e., $r_{ij} \in [0, 1]$ (instead of the deterministic $r_{ij} \in \{0, 1\}$), to capture variations of recommendations among different users, or even for the same user (e.g., to not show always the same recommendations for a given content).

Constraints. First, we require that the decision variables are probabilities ($r_{ij} \in [0, 1]$) and the recommendation lists contain N recommendations ($\sum_{j \in \mathcal{K}} r_{ij}{=}N$) [12], [18]. Second, we use a threshold $q \in [0, 1]$ for the QoR constraint similarly to previous works, i.e., $\sum_{j \in \mathcal{K}} r_{ij} \cdot u_{ij} \geq q \cdot q_i^{BS}$, where q_i^{BS} the maximum QoR achieved by the BS-RS. Finally, we introduce the fairness constraint, by using a threshold c_f for the maximum allowed unfairness, i.e., $F(\mathbf{p^{BS}}, \mathbf{p^{NF}}) \leq c_f$, where F is any of the metrics F_{max}, F_{tv}, or F_{kl}. Both thresholds q and c_f, and the fairness metric F, can be selected by the content provider according to its operational requirements.

In the following theorem we express the above problem as a LP. To do this, we need to introduce a set of auxiliary variables and transform the non-linear expressions in the objective and constraints; the remainder of this section gives the proof, which includes all the needed details.

Theorem 2. *The optimal fair NF-RS is given by the solution of the following linear optimization problem:*

$$\underset{\mathbf{z}, \mathbf{p^{NF}}, \mathbf{W}}{\text{maximize}} \quad \sum_{i \in \mathcal{C}} p_i^{NF} \tag{6a}$$

$$\text{subject to} \quad p_j^{NF} - \frac{\alpha}{N} \cdot \sum_{i \in \mathcal{K}} w_{ij} = p^d(j), \quad \forall j \in \mathcal{K} \tag{6b}$$

$$\sum_{j \in \mathcal{K}} w_{ij} \cdot u_{ij} - p_i^{NF} \cdot q \cdot q_i^{BS} \geq 0, \quad \forall\ i\ \in \mathcal{K}, \tag{6c}$$

$$\sum_{j \in \mathcal{K}} w_{ij} - N \cdot p_i^{NF} = 0,\ w_{ii} = 0, \quad \forall\ i \in \mathcal{K} \tag{6d}$$

$$w_{ij} - p_i^{NF} \leq 0,\ w_{ij} \geq 0, \quad \forall\ i, j \in \mathcal{K} \tag{6e}$$

$$\mathbf{S}(\mathbf{z}, \mathbf{p^{NF}}) \tag{6f}$$

where $\mathbf{z} \in \mathbb{R}^K$, $\mathbf{W} \in \mathbb{R}^{K \times K}$, and $\mathbf{S}(\mathbf{z}, \mathbf{p^{NF}})$ a set of linear constraints given in Table IV for each fairness metric.

[4]There are NF-RS algorithms that select also the network policy, e.g., caching [3], [7]–[9]. While our framework can be generalized in this direction, this is out of the scope of this paper (see also discussion in Section VIII).

TABLE IV: Set of linear fairness constraints $\mathbf{S}(\mathbf{z}, \mathbf{p^{NF}})$.

F_{max}:	$p_i^{BS} - p_i^{NF} \leq c_f$ $p_i^{NF} - p_i^{BS} \leq c_f \quad \forall\, i \in \mathcal{K}$
F_{tv}:	$\sum_{i\in\mathcal{K}} z_i \leq c_f$ $p_i^{BS} - p_i^{NF} \leq z_i \quad \forall\, i \in \mathcal{K}$ $p_i^{NF} - p_i^{BS} \leq z_i \quad \forall\, i \in \mathcal{K}$
F_{kl}:	$\sum_{i\in\mathcal{K}} p_i^{BS} \cdot z_i \geq -\left(c_f - \sum_{i\in\mathcal{K}} p_i^{BS} \log(p_i^{BS})\right)$ $z_i \leq e^{(m-1)\cdot s} \cdot p_i^{NF} - (m-1)s - 1, \quad \forall i \in \mathcal{K}, m \in \{1, ..., M\}$

Proof. The objective (and the fairness constraint) involves terms of the content demand $\mathbf{p^{NF}}$, which depends on the recommendations $\mathbf{R}$. In the considered framework, and similarly to previous works (e.g., [4], [7], [8], [12], [17]), the content demand can be modeled with a Markov Chain, with transition probabilities that depend on the recommendations $\mathbf{R}$, the direct requests $\mathbf{p^{(d)}}$ and the probability α. Hence, using the result of [42] (the detailed proof is omitted due to space limitations), we prove the following lemma.

Lemma 1. *The content demand $\mathbf{p^{NF}}$ is given by*

$$\mathbf{p^{NF}} = (1-\alpha)\cdot \mathbf{p^{(d)}} \cdot \left(\mathbf{I} - \frac{\alpha}{N}\cdot \mathbf{R}\right)^{-1} \tag{7}$$

for $\alpha \in (0,1)$ and $\mathbf{p^{(d)}} > 0$; $\mathbf{I}$ is the $K \times K$ identity matrix.

Lemma 1 gives $\mathbf{p}^{NF}$ as a function of an inverse matrix of $\mathbf{R}$, which in general is non-convex on the variable $\mathbf{R}$. To overcome this non-linearity, we explicitly introduce $\mathbf{p^{NF}}$ as an auxiliary optimization variable. The only constraint we need for the variable $\mathbf{p^{NF}}$ is Eq. (7). To express this constraint as a linear equation, we first multiply both sides with the term $\left(\mathbf{I} - \frac{\alpha}{N}\cdot \mathbf{R}\right)$ and write:

$$\mathbf{p^{NF}} - \frac{\alpha}{N}\cdot \mathbf{p^{NF}} \cdot \mathbf{R} = (1-\alpha)\cdot \mathbf{p^{(d)}} \tag{8}$$

Since Eq. (8) involves products of the variables $\mathbf{p^{NF}} \cdot \mathbf{R}$ (i.e., a quadratic term), we substitute the optimization variables r_{ij} with the new auxiliary variables w_{ij}, where $r_{ij} = \frac{w_{ij}}{p_i^{NF}}$. This substitution is possible because $p_i^{NF} > 0$ for the cases of interest, as stated in the following corollary (whose proof follows by observing Eq. (8)) .

Corollary 1. $p_i^{NF} > 0$, $\forall i \in \mathcal{K}$, *for* $\alpha \in (0,1)$ *and* $\mathbf{p^{(d)}} > 0$.

Having introduced the new auxiliary variables, it is easy to show how the constraints of Eq. (6) are derived, by substituting $r_{ij} = \frac{w_{ij}}{p_i^{NF}}$, as follows:

$$\text{Eq. (6b)} \Leftrightarrow p_j^{NF} - \frac{\alpha}{N}\cdot \sum_{i\in\mathcal{K}} p_i^{NF}\cdot r_{ij} = (1-\alpha)\cdot p_i^d \tag{9a}$$

$$\text{Eq. (6c)} \Leftrightarrow \sum_{j\in\mathcal{K}} r_{ij}\cdot u_{ij} \geq q\cdot q_i^{BS} \tag{9b}$$

$$\text{Eq. (6d)} \Leftrightarrow \sum_{j\in\mathcal{K}} r_{ij} = N,\ r_{ii} = 0 \tag{9c}$$

$$\text{Eq. (6e)} \Leftrightarrow r_{ij} \leq 1,\ r_{ij} \geq 0 \tag{9d}$$

where Eq. (9a) is equivalent to Eq. (8) (and guarantees that $\mathbf{p^{NF}}$ is a stationary distribution for $\mathbf{R}$), Eq. (9b) is the QoR constraint, and Eq. (9c) and Eq. (9d) are constraints on the recommendation variables.

Up to now, we have transformed all the constraints, apart from the fairness constraint $F(\mathbf{p^{BS}}, \mathbf{p^{NF}}) \leq c_f$. In the following, we transform the fairness constraint in a set of linear constraints $\mathbf{S}(\mathbf{z}, \mathbf{p^{NF}})$ for each fairness metric of Section II-B.

F-max. In the case of F_{max} the fairness constraint is

$$F_{max}(\mathbf{p^{NF}}, \mathbf{p^{BS}}) = \max_{i\in\mathcal{K}}\{|p_i^{NF} - p_i^{BS}|\} \leq c_f \tag{10}$$

Eq. (10) is not a linear inequality. However, it can be expressed as the intersection of the following $2\cdot K$ linear inequalities

$$\begin{array}{l} p_i^{BS} - p_i^{NF} \leq c_f \\ p_i^{NF} - p_i^{BS} \leq c_f \end{array} \quad \forall\, i \in \mathcal{K} \tag{11}$$

where we first set $|p_i^{NF} - p_i^{BS}| \leq c_f$ $\forall i \in \mathcal{K}$ as equivalent to constraining the $\max$, and then substituted each absolute term $|x| \leq c_f$ with two constraints $x \leq c_f$ and $-x \leq c_f$.

F-tv. A similar approach could be applied for the constraint

$$F_{tv}(\mathbf{p^{NF}}, \mathbf{p^{BS}}) = \tfrac{1}{2}\cdot \textstyle\sum_{i\in\mathcal{K}} |p_i^{NF} - p_i^{BS}| \leq c_f \tag{12}$$

However, it would lead to 2^K linear inequalities, which is impractical for large catalogs. Hence, we introduce an auxiliary set of variables $\mathbf{z} \in \mathbb{R}^K$ (a K-sized vector) and substitute Eq. (12) with the following constraints

$$\begin{array}{l} \sum_{i\in\mathcal{K}} z_i \leq c_f \\ |p_i^{BS} - p_i^{NF}| \leq z_i, \quad \forall\, i \in \mathcal{K} \end{array} \tag{13}$$

The first constraint is a linear inequality, and the remaining K inequalities of Eq. (13) can be substituted with $2\cdot K$ linear inequalities similarly to Eq. (11).

F-kl. In the F_{kl} case, the constraint can be written as

$$F_{kl} = \textstyle\sum_{i\in\mathcal{K}} p_i^{BS}\cdot \left(\log(p_i^{BS}) - \log(p_i^{NF})\right) \leq c_f \tag{14}$$

Eq. (14) involves a logarithmic function, thus, we cannot proceed as in F_{max} or F_{tv}. We first rewrite Eq. (14) as

$$\textstyle\sum_{i\in\mathcal{K}} p_i^{BS} \log(p_i^{NF}) \geq -\left(c_f - \sum_{i\in\mathcal{K}} p_i^{BS}\log(p_i^{BS})\right) \tag{15}$$

where we remind that p_i^{NF} are optimization variables and p_i^{BS} are given constants. Then, we introduce an auxiliary set of K variables $\mathbf{z} \in \mathbb{R}^K$, and we demand the following $K+1$ inequalities which are equivalent to Eq. (15)

$$\begin{array}{l} \sum_{i\in\mathcal{K}} p_i^{BS}\cdot z_i \geq -\left(c_f - \sum_{i\in\mathcal{K}} p_i^{BS}\log(p_i^{BS})\right) \\ \log(p_i^{NF}) \geq z_i, \quad \forall\, i \in \mathcal{K} \end{array} \tag{16}$$

The first inequality of Eq. (16) is linear. The remaining K inequalities are nonlinear due to the presence of the logarithm. To transform them to linear constraints, we approximate the logarithm with a general family of linear cuts. Specifically, we define M lines for every i, as $f(p_i^{NF}) = a_{m,i}\cdot p_i^{NF} + b_{m,i}$, that are tangent to the $\log(p_i^{NF})$ function in the interval $p_i^{NF} \in [0,1]$. Essentially we sample the logarithm at the points

$$\{e^{-(m-1)\cdot s}\ ,\ \log e^{-(m-1)\cdot s}\}$$

where $m = 1,\ldots,M$ and $s < 1$. The M slopes $a_{m,i}$ and the corresponding constants $b_{m,i}$, which are the same for every dimension $i \in \mathcal{K}$, of these tangent lines can be straightforwardly calculated. Thus, instead of using the K non-linear inequalities of Eq. (16), we use the following $M \cdot K$ inequalities that are linear on the variables z_i and p_i^{NF}

$$z_i \leq e^{(m-1)\cdot s} \cdot p_i^{NF} - (m-1)s - 1, \quad \forall i \in \mathcal{K}, m \in \{1, ..., M\}$$

Remark: The sampling step s and the number of linear cuts M, play a major role in the optimization process. s that is small enough for dense sampling, and an M according to the size of the catalog. In our scenarios, we found that $s = 0.05$ and $M = 160$ suffices for a catalog of $K \approx 1000$ contents. □

VI. The Price of Fairness

In this section, we employ the Fair NF-RS of Section V to the simulation setup of Section III-A. We consider different values for the fairness constraints c_f, and in Fig. 3 we present the performance, i.e., the achieved CHR (y-axis) of the Fair NF-RS (continuous lines) vs. the resulting unfairness (x-axis). We present two indicative scenarios for the LastFM/Movielens datasets (red/blue color). Also, we present the bound for each scenario (dash lines), the operating points {*unfairness, CHR*} of the BS-RS (star markers) and the NF-RS schemes that do not consider fairness (star, cross, and hexagonal markers for the Greedy, Multi-step, and CABaRet NF-RS, respectively).

Below we discuss the main findings stemming from Fig. 3, which provide useful insights for the effect of imposing fairness in NFR and the price we have to pay for this.

Key finding 1: ***"The Fair NF-RS always achieves a better performance trade-off than other NF-RS algorithms"***

The first observation that verifies the correctness of the proposed approach is that the Fair NF-RS performs better than other NF-RS (both in fairness and network gains), i.e., the curve of the Fair NF-RS is above (higher CHR) and/or on the left (less unfairness) of the markers that indicate the operation points of the other NF-RS algorithms. Extending the Fair NF-RS curve towards (i) small values of F (x-axis) leads to the operating point of the BS-RS ($F = 0$), and (ii) large values of F leads to the operating point of the Multi-step NF-RS, which is equivalent to the Fair NF-RS *without* fairness constraint.

Key finding 2: ***"By allowing a little unfairness, high network gains can be achieved"***

Comparing the Fair NF-RS performance with this of the BS-RS, we see that the increase in the network gains is steep for a small relaxation in the unfairness (i.e., for small values in the x-axis). In fact, we can see that the curve of the Fair NF-RS coincides with the bound, which means that the optimal fairness-gains trade-off is achievable by the Fair NF-RS for small values of fairness constraints. *This is a promising message for the practical feasibility of the NFR framework*: significant network gains are possible even when a level of fairness is required by the content provider.

Key finding 3: ***"The price (wrt. the network gain) of imposing fairness is small"***

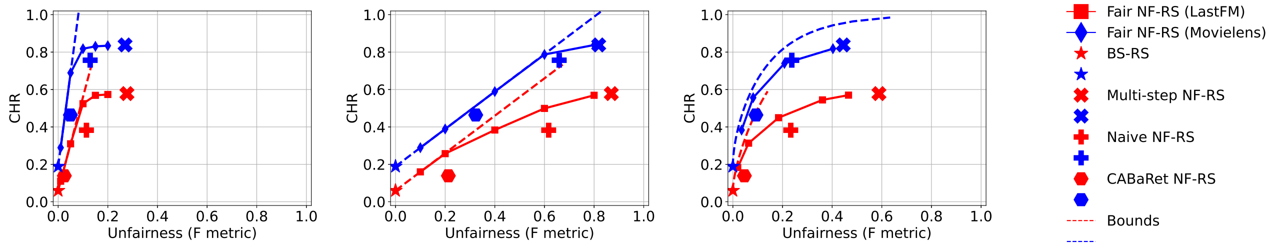
Moving our attention on the other side of the Fair NF-RS curve (i.e., for higher values of F), two interesting observations can be made: (i) the curve of the Fair NF-RS is concave, and (ii) the gains in CHR diminish for large values of F. These findings show that *similar network gains* to the Multi-step NF-RS (which achieves the best performance under no fairness constraints) can be achieved *with much less unfairness*. In particular in the case of F_{max} that captures the individual fairness, this behavior is clearer: a CHR very close to the highest possible can be achieved by the Fair NF-RS even with 3 times lower F_{max} compared to the Multi-step NF-RS .

Moreover, we can see that while the bounds are linear in the case of F_{max} and F_{tv}, the curve of the Fair NF-RS is concave: this is a positive finding indicating that the Fair NF-RS stays close to the bound and deviates for it only when large values of unfairness are allowed. Even in the case of F_{kl} where the bound is also concave, the Fair NF-RS curve approaches the highest possible CHR with a faster rate.

VII. Related Work

NFR. The paradigm of network-friendly (or, network-aware) recommendations has been recently proposed and studied under different network setups and content services [1]–[5], [7]–[17], [19]–[21]. The proposed NFR schemes aim to increase the network gains (and/or improve the quality of content delivery) by selecting recommendations [4], [5], [12], [20] or by jointly designing the recommendation and network policy (e.g., caching) [2], [3], [7]–[11], [13]–[17]. The majority of related works considers *cache-friendly* recommendations in mobile networks [1]–[4], [7]–[9]. However, the same principles apply to generic network setups [12], such as coded caching [7], broadcast communications [15]–[17], user association to base stations [13], or swarming systems [21]. While some of the proposed schemes take into account the user perspective by accounting the QoR, none of them has considered the fairness in recommendations from the perspective of the content provider. In this context, our work studies the dimension of fairness, thus providing a more complete view of the NFR paradigm. The proposed Fair NF-RS retains the efficiency of previous NFR schemes for achieving high network gains, while reduces the unfairness.

Fairness in RS. A variety of fairness metrics are used by the RS community [26]–[35] to capture different notions and needs of the content providers. Moreover, the fairness in RS can be defined with respect to the consumer (c-fairness) or the provider (p-fairness) [26], [27]. The former is typically used to design recommendation algorithms whose output is independent of sensitive user traits, e.g., race or gender [27], [29]. In other words, c-fairness aims to capture discrimination between users. Hence, it is orthogonal to our study, e.g., it could be considered as a part of the BS-RS and depends on the recommendation scores u_{ij} for which we consider a generic definition. The notion of p-fairness, which we use in this paper, aims to capture potential discrimination of the content provider towards different content producers/owners (or, individual contents). The proposed Fair NF-RS provides recommendations that achieve a balance between the user satisfaction (QoR), the content provider (fairness), and the



(a) F_{max} (b) F_{tv} (c) F_{kl}

Fig. 3: The price of fairness: Comparison of the performance (fairness at x-axis vs. CHR at y-axis) of the Fair NF-RS (continuous lines) with other RS (markers) and the bounds (dashed lines). Red colors correspond to the LastFM dataset scenario and blue colors to the Movielens dataset scenario, with parameters α=0.99, N=2, q=0.9, C=5 (LastFM) and C=10 (Movielens).

network gains. A similar issue is addressed in [35], from a multi-stakeholders perspective.

VIII. Conclusion

Previous works have shown that NFR can bring significant gains for the network, however, without considering the fairness, which is a key factor for content providers. This work is the first to study the dimension of fairness in NFR, and explore the trade-offs between controlling fairness and increasing network gains. Our results show that fairness *need* and *can* be taken into account in NFR, while the price (wrt. network cost) that one has to pay to impose fairness is small.

We believe that the findings of this paper can motivate further research on fairness in NFR. For example, under NFR schemes that *jointly* select the recommendation and network policies, we expect a more aggressive shaping of the demand. Hence, it is of interest to investigate if, and how the introduced unfairness and the trade-offs change under such schemes. In terms of fairness notions, an extension can be towards group fairness [32], [43], where the contents belong to classes (e.g., of the same genre or producer) [35], [38], and the fairness is defined among the aggregate demand of content classes. This more relaxed fairness metric, probably allows more flexibility in the decisions of the NF-RS and, thus, higher network gains.

References

[1] T. Spyropoulos and P. Sermpezis, "Soft cache hits and the impact of alternative content recommendations on mobile edge caching," in *Proc. ACM Workshop on Challenged Networks (CHANTS)*, 2016.

[2] L.-E. Chatzieleftheriou, M. Karaliopoulos, and I. Koutsopoulos, "Caching-aware recommendations: Nudging user preferences towards better caching performance," in *Proc. IEEE INFOCOM*, 2017.

[3] P. Sermpezis, T. Giannakas, T. Spyropoulos, and L. Vigneri, "Soft cache hits: Improving performance through recommendation and delivery of related content," *IEEE Journal Selected Areas in Communications*, 2018.

[4] T. Giannakas, P. Sermpezis, and T. Spyropoulos, "Show me the cache: Optimizing cache-friendly recommendations for sequential content access," in *Proc. IEEE WoWMoM*, 2018.

[5] S. Kastanakis, P. Sermpezis, V. Kotronis, and X. Dimitropoulos, "Cabaret: Leveraging recommendation systems for mobile edge caching," in *Proc. ACM SIGCOMM workshops*, 2018.

[6] S. Kastanakis, P. Sermpezis, V. Kotronis, D. S. Menasche, and T. Spyropoulos, "Network-aware recommendations in the wild: Methodology, realistic evaluations, experiments," *IEEE Trans. on Mobile Comp.*, 2020.

[7] B. Zhu and W. Chen, "Coded caching with joint content recommendation and user grouping," in *Proc. IEEE GLOBECOM*, 2018.

[8] L. E. Chatzieleftheriou, M. Karaliopoulos, and I. Koutsopoulos, "Jointly optimizing content caching and recommendations in small cell networks," *IEEE Trans. on Mobile Computing*, vol. 18, no. 1, 2019.

[9] M. Costantini, T. Spyropoulos, T. Giannakas, and P. Sermpezis, "Approximation guarantees for the joint optimization of caching and recommendation," in *Proc. IEEE ICC*, 2020.

[10] M. Garetto, E. Leonardi, and G. Neglia, "Similarity caching: Theory and algorithms," in *Proc. IEEE INFOCOM*, 2020.

[11] K. Qi, B. Chen, C. Yang, and S. Han, "Optimizing caching and recommendation towards user satisfaction," in *IEEE WCSP*, 2018.

[12] T. Giannakas, T. Spyropoulos, and P. Sermpezis, "The order of things: Position-aware network-friendly recommendations in long viewing sessions," in *Proc. WiOpt*, 2019.

[13] L. Chatzieleftheriou, G. Darzanos, M. Karaliopoulos, and I. Koutsopoulos, "Joint user association, content caching and recommendations in wireless edge networks," *PER*, vol. 46, no. 3, pp. 12–17, 2019.

[14] S. Gupta and S. Moharir, "Effect of recommendations on serving content with unknown demand," *ACM TOMPECS*, vol. 4, no. 1, p. 4, 2019.

[15] Z. Lin and W. Chen, "Joint pushing and recommendation for susceptible users with time-varying connectivity," in *IEEE GLOBECOM*, 2018.

[16] L. Song and C. Fragouli, "Making recommendations bandwidth aware," *IEEE Trans. Information Theory*, vol. 64, no. 11, 2018.

[17] Z. Lin and W. Chen, "Content pushing over multiuser miso downlinks with multicast beamforming and recommendation: A cross-layer approach," *IEEE Trans. on Communications*, vol. 67, no. 10, 2019.

[18] T. Giannakas, A. Giovanidis, and T. Spyropoulos, "Soba: Session optimal mdp-based network friendly recommendations," in *Proc. IEEE INFOCOM*, 2021.

[19] P. Sermpezis, S. Kastanakis, J. I. Pinheiro, F. Assis, D. Menasché, and T. Spyropoulos, "Towards qos-aware recommendations," in *ACM RecSys workshops (CARS workshop)*, 2020.

[20] D. Krishnappa, M. Zink, C. Griwodz, and P. Halvorsen, "Cache-centric video recommendation: an approach to improve the efficiency of youtube caches," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 4, p. 48, 2015.

[21] D. Munaro, C. Delgado, and D. S. Menasché, "Content recommendation and service costs in swarming systems," in *Proc. IEEE ICC*, 2015.

[22] Cisco, "Visual networking index: Forecast and trends, 2017-2022," 2018.

[23] Ericsson, "Ericsson mobility report," 2018, white paper.

[24] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proc. ACM IMC*, 2010.

[25] C. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, p. 13, 2016.

[26] H. Abdollahpouri and R. Burke, "Multi-stakeholder recommendation and its connection to multi-sided fairness," in *Proc. RMSE workshop at ACM RecSys*, 2019.

[27] R. Burke, "Multisided fairness for recommendation," in *Workshop on Fairness, Accountability, Transparency in Machine Learning*, 2017.

[28] R. Burke, N. Sonboli, and A. Ordonez-Gauger, "Balanced neighborhoods for multi-sided fairness in recommendation," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 202–214.

[29] B. Edizel, F. Bonchi, S. Hajian, A. Panisson, and T. Tassa, "Fairecsys: Mitigating algorithmic bias in recommender systems," *International Journal of Data Science and Analytics*, vol. 9, no. 2, pp. 197–213, 2020.
[30] G. K. Patro, A. Chakraborty, N. Ganguly, and K. Gummadi, "Incremental fairness in two-sided market platforms: On smoothly updating recommendations," in *Proc. AAAI conf. on Artificial Intelligence*, 2020.
[31] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.
[32] D. Sacharidis, K. Mouratidis, and D. Kleftogiannis, "A common approach for consumer and provider fairness in recommendations," in *Proc. ACM RecSys (Late-breaking Results,)*, 2019.
[33] H. Steck, "Calibrated recommendations," in *Proc. ACM RecSys*, 2018.
[34] K. Yang and J. Stoyanovich, "Measuring fairness in ranked outputs," in *Proc. SSDBM*, 2017.
[35] W. Liu and R. Burke, "Personalizing fairness-aware re-ranking," in *Proc. FATREC workshop at ACM RecSys*, 2018.
[36] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. ACM WWW*, 2001.
[37] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. ACM RecSys*, 2016.
[38] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems," in *Proc. ACM CIKM*, 2018.
[39] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, "An axiomatic theory of fairness in network resource allocation," in *IEEE INFOCOM*, 2010.
[40] "Last.fm dataset," https://labrosa.ee.columbia.edu/millionsong/lastfm.
[41] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. on Interactive Intelligent Systems (TiiS)*, 2016.
[42] A. N. Langville and C. D. Meyer, "Deeper inside pagerank," *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.
[43] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, and M. Shaoping, "Fairness-aware group recommendation with pareto-efficiency," in *Proc. ACM RecSys*, 2017.

Appendix: Proof of Theorem 1

The bound for the network gain G vs. fairness F trade-off is given by the solution of the optimization problem

$$\max_{\mathbf{p^{NF}}} \quad G \qquad s.t. \;\; F \leq c_F \tag{17}$$

where G is defined in Eq. (2), F is the fairness metrics (such as the F_{max}, F_{tv}, F_{kl} of Section II-B), c_F a constant, and $\mathbf{p^{NF}}$ has to be a probability distribution.

F-max. For the network gain it holds that

$$G \leq \sum_{i\in\mathcal{C}} |p_i^{NF} - p_i^{BS}| \leq C \cdot \max_{j\in\mathcal{K}} |p_j^{NF} - p_j^{BS}| = C \cdot F_{max}$$

where the first inequality follows by applying the property $x \leq |x|$ to the terms in the expression of G (Eq. (2)), the second inequality holds because $|p_i^{NF} - p_i^{BS}| \leq \max_{j\in\mathcal{K}} |p_j^{NF} - p_j^{BS}|$, $\forall i \in \mathcal{C}$, since $\mathcal{C} \subset \mathcal{K}$, and in the last equality we simply substituted from the definition of F_{max} (Section II-B).

F-tv. Starting similarly to the F_{max} case, we get

$$G \leq \sum_{i\in\mathcal{C}} |p_i^{NF} - p_i^{BS}| = 2 \cdot F_{tv} - \sum_{i\in\mathcal{K}\backslash\mathcal{C}} |p_i^{NF} - p_i^{BS}| \tag{18}$$

where the equality follows from the definition of F_{tv}. The right hand side of Eq. (18) increases when $\sum_{i\in\mathcal{K}\backslash\mathcal{C}} |p_i^{NF} - p_i^{BS}|$ decreases. The min value of this term can be calculated as:

$$\sum_{i\in\mathcal{K}} p_i^{BS} = \sum_{i\in\mathcal{K}} p_i^{NF} \quad \Rightarrow \quad \sum_{i\in\mathcal{K}\backslash\mathcal{C}} (p_i^{BS} - p_i^{NF}) = \sum_{i\in\mathcal{C}} (p_i^{NF} - p_i^{BS})$$
$$\Rightarrow \sum_{i\in\mathcal{K}\backslash\mathcal{C}} |p_i^{BS} - p_i^{NF}| \geq G \tag{19}$$

where in the first equation both sums equal to 1 (probability distributions), the second equation follows by moving all terms for $i \in \mathcal{K}\backslash\mathcal{C}$ to the left hand side, and in the third equation the left hand side follows from the property $x \leq |x|$ and the right hand side directly from the definition of G (Eq. (2)).

Substituting Eq. (19) in Eq. (18) gives

$$G \leq 2 \cdot F_{tv} - G \qquad \Rightarrow \qquad G \leq F_{tv}$$

F-kl. Due to the logarithm involved in the expression of F_{kl}, we cannot proceed similarly to the cases of F_{max} or F_{tv}, and we calculate the bound by solving the optimization problem of Eq. (17) with the method of Lagrangian multipliers. We first formulate the Lagrangian function $\mathcal{L}$ as follows[5]

$$\mathcal{L} = \textstyle\sum_{i\in\mathcal{C}} (p_i^{NF} - p_i^{BS}) - \lambda \cdot (F_{kl} - c_f) - \mu \cdot \left(\sum_{i\in\mathcal{K}} p_i^{NF} - 1\right)$$

The derivative of $\mathcal{L}$ with respect to p_i^{NF} is

$$\frac{\partial\mathcal{L}}{\partial p_i^{NF}} = \begin{cases} 1 + \lambda \cdot \frac{p_i^{BS}}{p_i^{NF}} - \mu & , i \in \mathcal{C} \\ \lambda \cdot \frac{p_i^{BS}}{p_i^{NF}} - \mu & , i \in \mathcal{K}\backslash\mathcal{C} \end{cases} \tag{20}$$

where we calculate $\frac{\partial F_{kl}}{\partial p_i^{NF}} = -\frac{p_i^{BS}}{p_i^{NF}}$ (see F_{kl} definition; Sec. II-B). Setting $\frac{\partial\mathcal{L}}{\partial p_i^{NF}} = 0$ for the optimal solution, gives:

$$p_i^{NF} = \begin{cases} \frac{\lambda}{\mu-1} \cdot p_i^{BS} & , i \in \mathcal{C} \\ \frac{\lambda}{\mu} \cdot p_i^{BS} & , i \in \mathcal{K}\backslash\mathcal{C} \end{cases} \tag{21}$$

To calculate the Lagrange multipliers, we use the definition of G (Eq. (2)), substitute from Eq. (21), and get

$$G = \sum_{i\in\mathcal{C}} \frac{\lambda}{\mu-1} p_i^{BS} - p_i^{BS} \Rightarrow \frac{\lambda}{\mu-1} = 1 + \frac{G}{\sum_{i\in\mathcal{C}} p_i^{BS}} = 1 + \frac{G}{H} \tag{22}$$

where for brevity we denoted $H = CHR^{BS} = \sum_{i\in\mathcal{C}} p_i^{BS}$. Then we consider the constraint $\sum_{i\in\mathcal{K}} p_i^{NF} = 1$ and substituting from the expressions in Eq. (21) and Eq. (22) we get

$$\sum_{i\in\mathcal{C}} \left(1 + \frac{G}{H}\right) \cdot p_i^{BS} + \sum_{i\in\mathcal{K}\backslash\mathcal{C}} \frac{\lambda}{\mu} \cdot p_i^{BS} = 1 \Rightarrow$$
$$\left(1 + \frac{G}{H}\right) \cdot \sum_{i\in\mathcal{C}} p_i^{BS} + \frac{\lambda}{\mu} \cdot \sum_{i\in\mathcal{K}\backslash\mathcal{C}} p_i^{BS} = 1 \Rightarrow$$
$$\left(1 + \frac{G}{H}\right) \cdot H + \frac{\lambda}{\mu} \cdot (1 - H) = 1 \quad \Rightarrow \quad \frac{\lambda}{\mu} = 1 - \frac{G}{1-H} \tag{23}$$

where we used $\sum_{i\in\mathcal{K}\backslash\mathcal{C}} p_i^{BS} = 1 - \sum_{i\in\mathcal{C}} p_i^{BS} = 1 - H$.

Now, substituting from Eq. (21), Eq. (22) and Eq. (23), in the expression for the F_{kl}, gives

$$F_{kl} =$$
$$\sum_{i\in\mathcal{C}} p_i^{BS} \cdot \log\left(\frac{p_i^{BS}}{\left(1 + \frac{G}{H}\right) p_i^{BS}}\right) + \sum_{i\in\mathcal{K}\backslash\mathcal{C}} p_i^{BS} \cdot \log\left(\frac{p_i^{BS}}{\left(1 - \frac{G}{1-H}\right) p_i^{BS}}\right)$$
$$= -\sum_{i\in\mathcal{C}} p_i^{BS} \cdot \log\left(1 + \frac{G}{H}\right) - \sum_{i\in\mathcal{K}\backslash\mathcal{C}} p_i^{BS} \cdot \log\left(1 - \frac{G}{1-H}\right)$$
$$= -H \cdot \log\left(1 + \frac{G}{H}\right) - (1 - H) \cdot \log\left(1 - \frac{G}{1-H}\right) \tag{24}$$

The above equality holds for the optimal $\mathbf{p^{NF}}$, i.e., the maximum network gain G; for any other $\mathbf{p^{NF}}$ the gains will be lower, which makes the above the inequality of Theorem 1.

[5]The problem Eq. (17) involves also the constraints $0 \leq p_i^{NF} \leq 1, \forall i \in \mathcal{K}$, which need to be accounted in the Lagrangian. However, if p_i^{NF} is 0 or 1, the F_{kl} diverges and thus the constraint in Eq. (17) is not satisfied. Hence, for any feasible solution it will hold that $0 < p_i^{NF} < 1$ and the corresponding Lagrange multipliers will be equal to zero (Karush–Kuhn–Tucker conditions).