Fairness in Network-Friendly Recommendations
Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis

To cite this version:

Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis. Fairness in Network-Friendly Recommendations. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Jun 2021, Pisa (virtual), Italy. pp.71-80, 10.1109/WoWMoM51794.2021.00020. hal-03188830

HAL Id: hal-03188830
https://hal.archives-ouvertes.fr/hal-03188830
Submitted on 2 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Fairness in Network-Friendly Recommendations
Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis

To cite this version:
Theodoros Giannakas, Pavlos Sermpezis, Anastasios Giovanidis, Thrasyvoulos Spyropoulos, George Arvanitakis. Fairness in Network-Friendly Recommendations. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, Jun 2021, Pisa, Italy. hal-03188830

HAL Id: hal-03188830
https://hal.archives-ouvertes.fr/hal-03188830
Submitted on 2 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 a 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
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 \( p = [p_1, ..., p_K] \) 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 \( R_i^{BS} = [a] \). Let only \( b, c \in \mathcal{C} \) be cached; then the NF-RS would recommend \( R_i^{NF} = [b] \), because this would bring network gains, and would have QoR = \( \frac{u_b}{u_b} = 0.8 \), which is higher than if \( c \) was recommended instead of \( b \) (QoR=0.5).

Finally, the resulting demand \( p \) 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., \( p^{BS} \) for the BS-RS and \( p^{NF} \) for a NF-RS. The differences between the vectors \( p^{BS} \) and \( p^{NF} \) 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].
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, involving the 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 \( p \) created by the NF-RS. Thus, a generic measure \( F \) can be used:

\[
F = f(p_{BS}^{NF}, p_{NF})
\]

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:\

- **F-max** \( F_{max} = \max_{i \in K} |p_{NF}^{i} - p_{BS}^{i}| \) 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** \( F_{tv} = \frac{1}{2} \cdot \sum_{i \in K} |p_{NF}^{i} - p_{BS}^{i}| \) 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** \( F_{kl} = \sum_{i \in K} p_{BS}^{i} \cdot \log \left( \frac{p_{BS}^{i}}{p_{NF}^{i}} \right) \) 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 \( \Delta p \) in the demand \( p_{BS}^{i} \) leads to a higher increase in \( F_{kl} \) when \( p_{BS}^{i} \) is small [33].

**Remark:** Note that \( F_{max}, F_{tv} \in [0, 1] \), whereas \( F_{kl} \in [0, \infty] \) \((F_{kl} \to \infty \text{ when } p_{NF}^{i} = 0 \text{ and } p_{BS}^{i} \neq 0)\). 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 \( p_{NF}^{i} \to (1 - w) \cdot p_{NF}^{i} + w \cdot p_{BS}^{i} \), with \( w = 0.01 \) and normalize with its upper bound \( \frac{1}{w} \), i.e.,

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

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 \( b, c \in C \), 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 \( U = \{u_{ij}\} \) 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 \( U \). As the resulting \( U \) 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
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 ($p_{BS}^d$).

**Content demand.** Similarly to previous works [4], [7], [8], [12], [15], [17], we assume that a user follows a recommendation with probability $\alpha$, 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 ($\alpha=0.5$) [24] and Netflix ($\alpha=0.8$) [25] and an extreme value where users follow almost always recommendations ($\alpha=0.99$), 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 $p^{(d)}$.

**NF-RS algorithms.** Several NRF variants have been proposed. To avoid restricting our study to a single algorithm or setup, we consider three representative NRF 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 $\alpha$. 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 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., $\sum_{j \in R_i^{BS}} u_{ij} \geq q \cdot \sum_{j \in R_i^{BS}} u_{ij}$ [4], [12]. In CABaRet, the QoR is implicitly determined by the width $W_{BS}$ and the depth $D_{BS}$ parameters of the BFS [5]. In our simulations, we consider values $q \in \{0.5, 0.8, 0.9\}$, and $W_{BS} \in \{N, 2N\}$ and $D_{BS} = \{1, 2\}$.

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 NRF**

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

**Unfairness in existing NRF algorithms.** We first quantify the unfairness created by existing NRF algorithms. Figure 1(a) presents the CDF of the unfairness $F$ for 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 NRF 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 $p_{BS}^d$, 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

<table>
<thead>
<tr>
<th>U: last.fm ($K = 757$), Movielens ($K = 1000$)</th>
<th>N \in {2, 5, 10}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^{(d)} \sim \text{Zipf}(s = 1)$, uniform</td>
<td>$C \in {5, 10, 20}$</td>
</tr>
<tr>
<td>$\alpha \in {0.5, 0.8, 0.99}$</td>
<td>$q \in {0.5, 0.8, 0.9}$</td>
</tr>
<tr>
<td>$W_{BS} \in {N, 2N}$</td>
<td>$D_{BS} \in {1, 2}$</td>
</tr>
</tbody>
</table>

---

$^2$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.
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 ($\rho$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$F_{max}$</th>
<th>$F_{tv}$</th>
<th>$F_{kl}$</th>
<th>$G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>$\rho=0.40$</td>
<td>$\rho=0.38$</td>
<td>$\rho=0.40$</td>
<td>$\rho=0.42$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\rho=0.46$</td>
<td>$\rho=0.81$</td>
<td>$\rho=0.75$</td>
<td>$\rho=0.69$</td>
</tr>
<tr>
<td>$N$</td>
<td>$\rho=0.47$</td>
<td>$\rho=0.13$</td>
<td>$\rho=0.16$</td>
<td>$\rho=0.20$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\rho=0.13$</td>
<td>$\rho=0.06$</td>
<td>$\rho=0.03$</td>
<td>$\rho=0.18$</td>
</tr>
</tbody>
</table>

We observed the same relation between fairness and network gains, when varying the other system parameters as well; see Table III. Specifically, increasing the $\alpha$ 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 NF-RS scheme.

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

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

where $C \subset K$ 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

\[
F_{\text{max}} \geq \frac{1}{C} \cdot G \tag{3}
\]

\[
F_{\text{tv}} \geq G \tag{4}
\]

\[
F_{\text{kl}} \geq -H \cdot \log\left(1 + \frac{G}{H}\right) - (1 - H) \cdot \log\left(1 - \frac{G}{1 - H}\right) \tag{5}
\]

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

**Proof.** The proof is given in the Appendix. \qed

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 $F_{\text{tv}}$ value, and (ii) increases with the cache size $C$ in the $F_{\text{max}}$ case. This indicates that larger caches can allow network gains without compensating in individual fairness ($F_{\text{max}}$), while this is not the case in aggregate fairness ($F_{\text{tv}}$). In the case of $F_{\text{kl}}$, 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 $H = \text{CHR}^{BS}$).

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 $F_{\text{max}}$ and $F_{\text{kl}}$ and 20% larger for $F_{\text{tv}}$. 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) ($F_{ij}$ case) correspond to different values $\text{CHR}^{BS}$ (e.g., due to different $p^{BS}$ 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.

\section{Optimal Fair NF-RS}

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., $C_{i} = \sum_{i \in C} p_{i}^{NF}$.

**Decision variables.** An NF-RS algorithm selects which content provider according to its operational requirements. 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).

**Constraints.** First, we require that the decision variables are bounded, i.e., $r_{ij} \in [0, 1]$ for the QoR constraint similarly to previous works, i.e., $\sum_{j \in K} r_{ij} = 1$ [12], [18]. Second, we use a threshold $q \in [0, 1]$ for the QoR constraint similarly to previous works, i.e., $\sum_{j \in K} r_{ij} \cdot u_{ij} \geq q \cdot q^{BS}$, where $q^{BS}$ is 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., $F(p^{BS}, p^{NF}) \leq c_{f}$, where $F$ is any of the metrics $F_{\text{max}}$, $F_{\text{tv}}$, or $F_{\text{kl}}$. 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:

\[
\begin{align*}
\text{maximize} & \quad \sum_{i \in C} \left( p_{i}^{NF} \cdot \sum_{j \in K} w_{ij} \cdot p^{NF}(j) \right), \quad \forall j \in K \tag{6a} \\
\text{subject to} & \quad p_{i}^{NF} - \frac{\alpha}{N} \cdot \sum_{j \in K} w_{ij} = p^{NF}(j), \quad \forall j \in K \tag{6b} \\
& \quad \sum_{j \in K} u_{ij} - p_{i}^{NF} \cdot q^{BS} \geq 0, \quad \forall i \in K, \quad \forall j \in K, \tag{6c} \\
& \quad \sum_{j \in K} w_{ij} - N \cdot p_{i}^{NF} = 0, \quad w_{ii} = 0, \quad \forall i \in K \tag{6d} \\
& \quad w_{ij} - p_{i}^{NF} \leq 0, \quad w_{ij} \geq 0, \quad \forall i, j \in K \tag{6e} \\
& \quad S(z, p^{NF}) \geq 0 \tag{6f}
\end{align*}
\]

where $z \in \mathbb{R}^{K}$, $W \in \mathbb{R}^{K \times K}$, and $S(z, p^{NF})$ a set of linear constraints given in Table IV for each fairness metric.
Proof. The objective (and the fairness constraint) involves terms of the content demand \( p_{NF} \), which depends on the recommendations \( R \). 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 \( R \), the direct requests \( p(d) \) and the probability \( \alpha \). 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 \( p_{NF} \) is given by

\[
p_{NF} = (1 - \alpha) \cdot p(d) \cdot \left( I - \frac{\alpha}{N} \cdot R \right)^{-1}
\]  

(7)

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

Corollary 1. \( p_{NF} > 0 \), \( \forall i \in K \), for \( \alpha \in (0,1) \) and \( p(d) > 0 \).

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

\[
\text{Eq. (6b)} \iff p_{i}^{NF} - \frac{\alpha}{N} \cdot \sum_{i \in K} p_{i}^{NF} \cdot r_{ij} = (1 - \alpha) \cdot p_{i}^{d} \quad (9a)
\]

\[
\text{Eq. (6c)} \iff \sum_{j \in K} r_{ij} \cdot u_{ij} \geq q \cdot q_{i}^{BS} \quad (9b)
\]

\[
\text{Eq. (6d)} \iff \sum_{j \in K} r_{ij} = N, \quad r_{ii} = 0 \quad (9c)
\]

\[
\text{Eq. (6e)} \iff r_{ij} \leq 1, \quad r_{ii} \geq 0 \quad (9d)
\]

where Eq. (9a) is equivalent to Eq. (8) (and guarantees that \( p_{NF} \) is a stationary distribution for \( R \)), 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 \( F(p_{BS}, p_{NF}) \leq c_{f} \). In the following, we transform the fairness constraint in a set of linear constraints \( S(z, p_{NF}) \) for each fairness metric of Section II-B. **F-max.** In the case of \( F_{max} \) the fairness constraint is

\[
F_{max}(p_{NF}, p_{BS}) = \max_{i \in K} \left| p_{i}^{NF} - p_{i}^{BS} \right| \leq c_{f}
\]

(10)

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

\[
p_{i}^{BS} - p_{i}^{NF} \leq c_{f} \quad \forall i \in K
\]

(11)

where we first set \( |p_{i}^{NF} - p_{i}^{BS}| \leq c_{f} \) \( \forall i \in K \) 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

\[
F_{tv}(p_{NF}, p_{BS}) = \frac{1}{2} \cdot \sum_{i \in K} \left| p_{i}^{NF} - p_{i}^{BS} \right| \leq c_{f}
\]

(12)

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

\[
\sum_{i \in K} z_{i} \leq c_{f}
\]

\[
|p_{i}^{BS} - p_{i}^{NF}| \leq z_{i}, \quad \forall i \in K
\]

(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

\[
F_{kl} = \sum_{i \in K} p_{i}^{BS} \cdot \left( \log(p_{i}^{BS}) - \log(p_{i}^{NF}) \right) \leq c_{f}
\]

(14)

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

\[
\sum_{i \in K} p_{i}^{BS} \log(p_{i}^{NF}) \geq - \left( c_{f} - \sum_{i \in K} p_{i}^{BS} \log(p_{i}^{BS}) \right)
\]

(15)

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

\[
\sum_{i \in K} p_{i}^{BS} \cdot z_{i} \geq - \left( c_{f} - \sum_{i \in K} p_{i}^{BS} \log(p_{i}^{BS}) \right)
\]

(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 \( f(p_{i}^{NF}) = a_{m,i} p_{i}^{NF} + b_{m,i} \), 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

\[
\left\{ e^{-(m-1) \cdot s} \log e^{-(m-1) \cdot s} \right\}
\]

where \( m = 1, \ldots, M \) and \( s \leq 1 \). The \( M \) slopes \( a_{m,i} \) and the corresponding constants \( b_{m,i} \), which are the same

|\begin{array}{|c|c|}
|\hline
| F_{max}: & p_{i}^{BS} - p_{i}^{NF} \leq c_{f} \\
| & \forall i \in K \\
| F_{tv}: & p_{i}^{BS} - p_{i}^{NF} \leq z_{i} \\
| & \forall i \in K \\
| F_{kl}: & \sum_{i \in K} p_{i}^{BS} \cdot \log(p_{i}^{NF}) \leq c_{f} \\
| & \forall i \in K \\
| F_{tv}: & p_{i}^{BS} - p_{i}^{NF} \leq c_{f} \\
| & \forall i \in K \\
|\end{array}|
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 \( p_i^{NF} \):

\[
z_i \leq e^{(m-1)i} \cdot p_i^{NF} - (m-1)\forall i \in \mathcal{K}, m \in \{1, \ldots, M\}
\]

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 \( M \) 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 \( (\text{unfairness}, \text{CHR}) \) 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
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 $\alpha=0.99$, $N=2$, $\rho=0.9$, $C=5$ (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

for $i \in K \setminus C$ 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

$$G \leq 2 \cdot F_{tv} - G \implies G \leq F_{tv}$$

**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:

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

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

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

Substituting Eq. (22) (see $F_{kl}$ definition; Sec. II-B). Setting $\frac{\partial \mathcal{L}}{\partial p_{i}^{NF}} = 0$ for the optimal solution, gives:

$$p_{i}^{NF} = \begin{cases} \frac{\lambda \cdot k^{-1}}{\mu} \cdot p_{i}^{BS}, & i \in C \\ 1, & i \in K \setminus C \end{cases}$$

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

$$G = \sum_{i \in C} \mu \cdot p_{i}^{BS} - p_{i}^{NF} \Rightarrow \frac{\lambda}{\mu} = 1 = \sum_{i \in C} G_{i}^{BS} = 1 + \frac{G}{H}$$

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

$$\sum_{i \in C} \left( 1 + \frac{G}{H} \right) \cdot p_{i}^{BS} + \sum_{i \in C} \frac{\lambda}{\mu} \cdot p_{i}^{BS} = 1 \implies \left( 1 + \frac{G}{H} \right) \cdot H + \frac{\lambda}{\mu} \cdot (1 - H) = 1 \implies \frac{\lambda}{\mu} = 1 - \frac{G}{1 - H}$$

where we used $\sum_{i \in K \setminus C} p_{i}^{BS} = 1 - \sum_{i \in C} p_{i}^{BS} = 1 - H$. Now, substituting from Eq. (21), Eq. (22) and Eq. (23), in the expression for the $F_{kl}$, gives

$$F_{kl} = \sum_{i \in C} p_{i}^{BS} \cdot \log \left( \frac{p_{i}^{BS}}{1 - G} \right) + \sum_{i \in K \setminus C} p_{i}^{BS} \cdot \log \left( \frac{1 - G}{1 - H} \right) = - \sum_{i \in C} p_{i}^{BS} \cdot \log \left( 1 + \frac{G}{H} \right) - \sum_{i \in K \setminus C} p_{i}^{BS} \cdot \log \left( 1 - \frac{G}{1 - H} \right)$$

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

The problem Eq. (17) involves also the constraints $0 \leq p_{i}^{NF} < 1$, $\forall i \in K$, which need to be accounted in the Lagrangian. However, if $p_{i}^{NF} = 0$ 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).