sensors

MDPI

Article

Energy-Efficient Nonuniform Content Edge Pre-Caching to Improve Quality of Service in Fog Radio Access Networks

Yi Cen [1], Yigang Cen [2], Ke Wang [3,*] and Jingcong Li [1]

[1] School of Information Engineering, Minzu University of China, Beijing 100081, China; yi_cen@126.com (Y.C.); jcli@pku.org.cn (J.L.)
[2] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China; ygcen@bjtu.edu.cn
[3] School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing 100876, China
* Correspondence: wangke@bupt.edu.cn

Received: 9 February 2019; Accepted: 15 March 2019; Published: 22 March 2019

Abstract: The fog radio access network (F-RAN) equipped with enhanced remote radio heads (eRRHs), which can pre-store some requested files in the edge cache and support mobile edge computing (MEC). To guarantee the quality-of-service (QoS) and energy efficiency of F-RAN, a proper content caching strategy is necessary to avoid coarse content storing locally in the cache or frequent fetching from a centralized baseband signal processing unit (BBU) pool via backhauls. In this paper we investigate the relationships among eRRH/terminal activities and content requesting in F-RANs, and propose an edge content caching strategy for eRRHs by mining out mobile network behavior information. Especially, to attain the inference for appropriate content caching, we establish a pre-mapping containing content preference information and geographical influence by an efficient non-uniformed accelerated matrix completion algorithm. The energy consumption analysis is given in order to discuss the energy saving properties of the proposed edge content caching strategy. Simulation results demonstrate our theoretical analysis on the inference validity of the pre-mapping construction method in static and dynamic cases, and show the energy efficiency achieved by the proposed edge content pre-caching strategy.

Keywords: fog radio access network; non-uniform mobile edge caching; preference inference; group partition; non-convex matrix/tensor completion

1. Introduction

Recently, the fog radio access network (F-RAN) has been proposed as an emerging network architecture of a cloud radio access network (C-RAN) for fifth generation wireless systems (5G), which aims to address the limitations of previous cellular standards and be a prospective key enabler for future Internet-of-Things (IoT) [1]. In a typical C-RAN, a centralized baseband signal processing unit (BBU) pool is equipped for baseband processing of a remote radio heads (RRHs) set, connected to the BBUs by fronthaul links, to save on operational expenditures and reduce energy consumption [2–5]. Although some efficient signal compression methods have been proposed for C-RANs [6], it is insufficient to satisfy the dramatically increasing requirements from mobile users for real-time services with high quality-of-service (QoS) guarantees. To address this challenge, the enhanced architecture of F-RAN allowing the RRHs with the capability of storage and signal processing functionalities, was proposed [7–9]. With the enhanced RRHs (eRRHs), edge caching can be performed to pre-fetch some requested files to the eRRHs local caches (as illustrated in Figure 1). Subsequently, the burden on

backhaul is relieved and higher spectral efficiency or lower delivery latency can be obtained for users' requesting cached files from incorporating caching units. In general, the goal of F-RAN architecture is to optimize the system performance in terms of delivery rate by leveraging both BBU and edge caching, which is different from that in C-RAN.

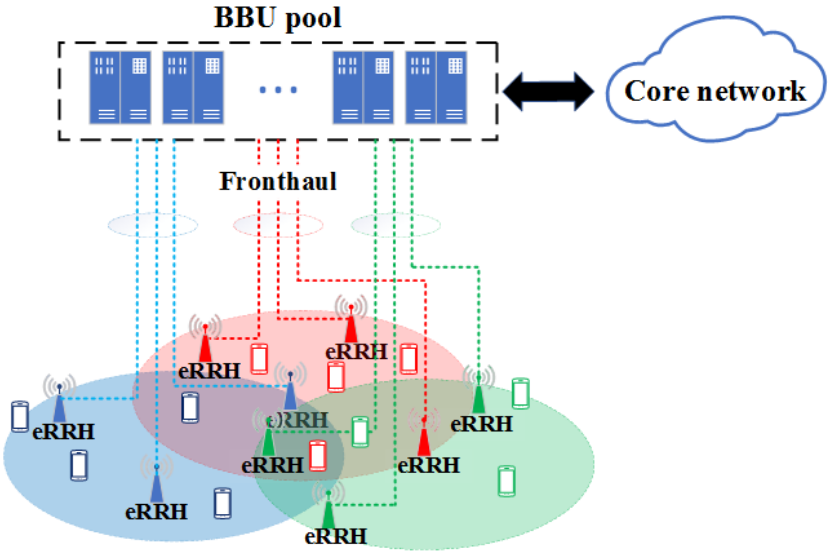

Figure 1. An architecture of F-RAN scheduling framework. The UEs with the same preference are denoted by the same color and the dotted loops of different colors denote the areas where the corresponding preferences are dominant.

As a cache-aided system, F-RAN operates in the pre-fetching and delivery phases [10–12]. The pre-fetching normally stores the content by constant popularity ranking in the large time scale corresponding to multiple transmission intervals. Instead, the delivery phase operates separately on each transmission interval based on the cached file messages. Several recent studies involved the scheme and the performance of F-RAN in this context. In [7,10] the fronthaul-aware design via the pre-fetching policy was studied to minimize the average delivery latency with the cache memory constraints, and in [13] the optimal caching and delivery strategies that minimize the delivery latency are characterized for system designing. In [12,14], trade-off between the total power consumed and the total backhaul capacity needed in the downlink of the cache-aided RAN was studied. It has shown that the energy consumption is decreased due to the increase of spatial diversity by cooperative communication, meanwhile the backhaul burden is increased due to the delivery of uncached files to more eRRHs. The aforementioned research indicates that the content edge-storage strategy design plays a critical role in improving the performances of F-RAN.

Designing content edge caching strategies to more closely meet the needs can greatly alleviate the burden on backhauls, and reduce the delay and energy consumption for a large number of users. Notice that caching popular files can make the design of the delivery phase more meaningful and sensible to meet the needs of most users; almost all the content edge-storage strategies via pre-fetching were directly proposed due to this most-popular-caching premise. To the best of our knowledge, the content edge-storage strategies in the literature are considered to be based on an assumption that all files are available at the BBU and the popularity of the file is modeled by Zipf distribution [15]. More specifically, for Zipf distribution, let the files be labeled in the order of popularity from the most to the least popular ones, such that the most popular file has index $f = 1$ and the least popular file has index $f = F$. Then the probability of a requested file $f \in \{1, 2, \ldots, F\}$ is $P(f) = cf^{-s}$ satisfying $\sum_{f=1}^{F} P(f) = 1$. Therefore, consider that user equipments (UEs) in an F-RAN are served by multiple eRRHs that are connected to a BBU pool through digital fronthaul links, UE_k selects file $f_k \in \{1, 2, \ldots, F\}$ with the probability $P(f_k)$, and the requested files f_k are independent across

the index k. The increasing of the exponent s makes the probability of selecting a small group of files larger. When the parameter s is adjusted appropriately, only a few popular files are frequently requested by UEs.

Although the above content selecting criterion can simplify the discussion, the real situation of the content obtaining in F-RANs is much more complex due to the users' social and activity limitation, which has noticeable impact on the performances of F-RAN. For instance, there are usually some certain activity regions for different users, which endows the content request of the individual with regional features. Also, the preferences of the users in different districts, such as the area around the school, the airport or the mall, are often relatively different, which can make the content distribution non-uniformed for different groups. Moreover, for some multimedia users, the obtained information is inevitably lagging behind once the cached contents are not desired by them. It means that judging which content to be cached from past requesting actions is reasonable and necessary in order to guarantee the QoS. It is shown that such contradictions are particularly prominent under the condition of limited edge cache in eRRH, and a straightforward way to utilize Zipf distribution for content caching is inappropriate to realistic needs. However, notice that the massive data on the activity of request is recorded and the powerful computational capacity is provided by the computing resources of the cloud, it is possible to design more efficient content-obtaining strategies relying on these foundations.

With this consideration, in this paper, we propose a caching content selection strategy by digging out users' network behavior information and improving the distribution on content allocation. We analyze the relationship between scope of eRRH/UEs' activities and content requests in an F-RAN, and then establish pre-mapping inferred by an efficient matrix completion algorithm for an appropriate content edge pre-caching. Especially, the proposed matrix completion algorithm gets better at the accuracy in the case of non-uniform data sampling and computational efficiency. Furthermore, we analyze the energy consumption of the content edge pre-caching strategy based on the proposed pre-mapping. Numerical results are provided to prove the effectiveness of the given inference method for the pre-mapping construction in static and dynamic cases, and illustrate the energy saving characteristics of the content edge pre-caching.

This paper is structured as follows. In Section 2 we introduce the system model considered and state the caching file selection strategy. The data structure corresponding to the UEs' requests for the entire researched F-RAN is established mathematically as well. in Section 3, the optimization problem of the content edge caching pre-mapping construction is presented and the non-uniform non-convex matrix completion algorithm is proposed for solving the problem. The corresponding energy consumption analysis for the caching content selection strategy is provided in Section 4. Simulation results of the algorithm and the energy consumption performances are shown in Section 5, followed by the conclusion in Section 6.

2. System Model and Problem Statement

We consider the downlink transmission of an F-RAN as illustrated in Figure 1. N eRRHs are deployed in the network and cooperatively serve all users. Each eRRH, equipped with L antennas and a cache of the same size, can connect to the BBU pool via individual backhaul links with finite capacity. The cluster-scale joint management, such as scheduling and resource allocation, can be implemented. On the other hand, UEs are uniformly and independently distributed within the network. In each scheduling interval, K users will be scheduled, and send their content requests according to some preferences. Assume that each user can request at most one content at its scheduled time and the content cache in BBU pool stores the set of all content objects required by the users, which is denoted as a set $S_c = \{F_1, \ldots, F_C\}$, with $F_1, \ldots, F_C$ all of the same size and belonging to I types. Then in their own activity regions, UEs requesting the contents that belong to the same subset $S_m \subset S_c$ (with the same color icon shown in Figure 1) can form a multicast UE cluster G_m. The m-th cluster G_m, limited by the preferences and activity regions, is served cooperatively by a cluster of eRRHs, which is denoted as R_m.

It should be noted that, although the eRRH clusters serving UE clusters can overlap with each other, the overlap of the eRRH clusters is relatively small and irregular since the service area of an eRRH is limited and the eRRH equipments are deployed in the sufficiently large area. With this consideration, we ignore this characteristic here to simplify the analysis and assume that the users in the overlap are served by only one eRRH cluster with high probability, i.e., $G_m \cap G_{m'} = \emptyset$ and $R_m \cap R_{m'} = \emptyset$, $m \neq m'$.

Content caching and transmission: Without loss of generality, we set $K > N$ and focus on a typical eRRH/UE cluster, i.e., G_m and R_m (represent by the icons with the same color in Figure 1). While the content cache of BBU pool is deployed to fully exploit the potential of content caching in the F-RAN, the content cache of each eRRH in the same cluster only contains the contents that are most likely to be requested by the users in the same area. The content requests from served users are aggregated at the edge eRRH cluster in the F-RAN, and can be treated as follows: First, the eRRH cluster checks its cluster content caches and the requests can be served immediately if the desired content is available at the caches (illustrated as the procedure (1) in Figure 2). Otherwise, the requests will be forwarded to the BBU pool content cache, and then the corresponding content can be provided through the fronthaul link from the BBU pool. Then the requests can be handled similarly to the case in which the content resides in the content caches in the cluster (illustrated as the procedure (2) in Figure 2).

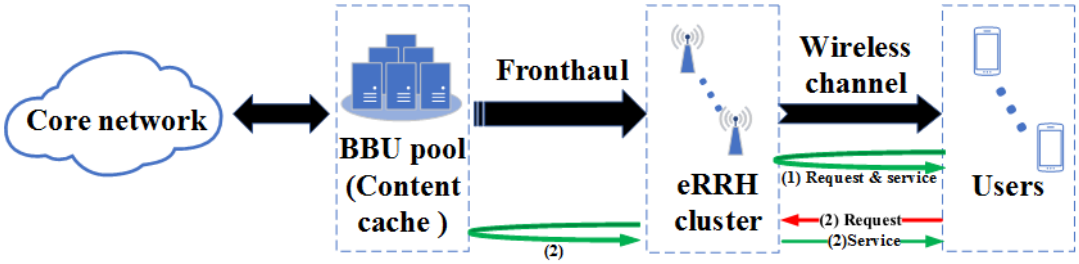

Figure 2. Queueing model of content object transmission with one eRRH-cluster content caching in F-RANs.

To increasing the operability, the preferences of the UEs' requests and the corresponding contents will be classified into different types. For each type, the contents will be sorted according to the timeliness and the popularity, and pre-cached in the eRRHs according to the UEs' preferences inferred by the requests. Because the caching via the preferences of the individual can reduce the probability of the procedure (2) execution, which can lead to improved QoS guarantees with low power consumption in practical F-RANs, it is reasonable to propose an eRRH content per-caching strategy depending on the UEs' behaviors.

Content preference data model: Note that the restricted mobility and the randomness of request for UEs in line with the assumptions at the beginning of this section, we establish the content preference data model associated with the UEs' requests based on the following setting: (i) the users usually stay in some limited areas depending on the occupation and habit, etc. This results in that only a limited part of the eRRHs (compared with the overall scale in the researched F-RAN) can serve the certain UEs; (ii) Due to the restricted mobility of each UE, the eRRH in the service cluster to receive the UE's request can be deemed to be selected randomly. This means that some eRRHs in the serving cluster S_m may not obtain the request of the UE in the G_m; (iii) It is also obvious that the number of content types interested in is far less than the scale of the users. Hence, with these prerequisites, we first let j and k denote the indices of eRRHs and UEs respectively, the basic dataset reflecting the content types that the UEs' requests and can be structured as the preference matrix $\mathbf{A} \in \{a_1, a_2, \ldots, a_I, 0\}^{N \times K}$ with the entry as follows:

$$A_{jk} = \begin{cases} \varphi_{m,i} & \text{if } k \text{ is served by the } j\text{-th eRRH belonging to } \mathrm{R_m} \text{ and has} \\ & \text{sent the request to } j \text{ for the content in the } i\text{-th type,} \\ 0 & \text{if the relationship between } k \text{ and } j \text{ does not exist,} \end{cases} \quad (1)$$

where $\varphi_{m,i}$ is a positive integer and $\varphi_{m,i} = a_i \in \{a_1, a_2, \ldots, a_I\}$. The value of a_i can be set as an arbitrary integer for simplicity provided that there is no popular rank. However, if the popular rank exists, the value of a_i will be set according to the way used for Zipf distribution setting and refer to the request history of the corresponding user.

We now show that the preference matrix **A** with the complete request dataset is low rank. For UE k, the entries A_{jk} corresponding to the m-th eRRH cluster R_m are assigned as the same a_i since all eRRHs in the cluster should serve UE k. Thus, with a suitable realignment of users and eRRHs, **A** can be represented as a block-diagonal matrix where the entries within the diagonal blocks are positive and the others within the off-diagonal blocks are all 0's. Furthermore, for each "positive" block, all entries of the diagonal sub-blocks are $\varphi_{m,i}$ (shown in Figure 3). The following theorem provides the upper bound for the rank of the complete preference matrix.

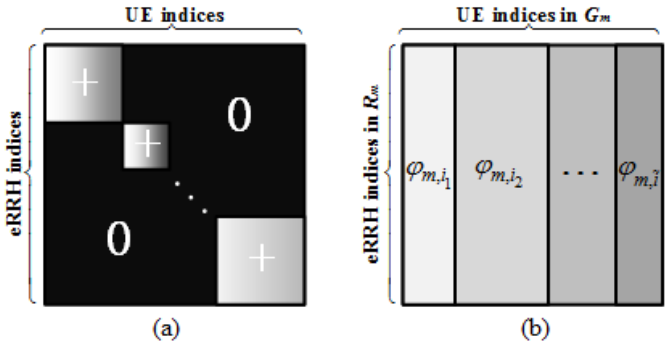

Figure 3. The structure of matrix **A** corresponding to the complete request data for the entire F-RAN. (**a**) presents the structure of the entire matrix; (**b**) is used to illustrate each positive sub-block with different shades of gray, the rows and the columns of the sub-block correspond to the users and the eRRHs in the same cluster.

Theorem 1. (Low rank structure of request preference data model): *The complete preference matrix* $\mathbf{A} \in \{a_1, a_2, \ldots, a_I, 0\}^{N \times K}$ *has rank* $r = N_{e-ug} I$ *at most, where* N_{e-ug} *and* I *correspond to the number of multicast eRRH/UE clusters and that of content types respectively.*

Proof. Suppose that all clusters of eRRHs are in service, then the eRRHs set can be divided into N_{e-ug} clusters, i.e., R_m, $m = 1, \ldots, N_{e-ug}$. Further, for each eRRH/UE cluster, the UEs of G_m can be classified into several common preference sub-groups, denoted by $G_{m(i)}$ and the number of the sub-group $G_{m(i)}$ in the entire network is no more than $N_{e-ug} I$. After suitable reordering indices of nodes (UEs) in $G_{m(i)}$, the corresponding row vector of **A** are all identical to the following form

$$\mathbf{a} = (0, \ldots, 0, \underbrace{\varphi_{m,i_1}, \ldots, \varphi_{m,i_1}, \ldots, \varphi_{m,i_{d_m}}, \ldots, \varphi_{m,i_{d_m}}}_{\text{the } m\text{th cluster}}, 0, \ldots, 0), \quad (2)$$

where d_m is the cardinality of the requested content type subset $\{i_1, \ldots, i_{d_m}\}$ in G_m and $d_m \leq I$. Then let us take all UE subgroups into account, we operate column vectors of the $m(i)$ sub-group by the column elementary transformation, which is equivalent to subtracting the first column vector of the i-th sub-group from the other ones, etc. Do the similar operations for row vectors of **A** and rearrange the rows and columns of **A**, we obtain

$$\mathbf{A} \to \begin{pmatrix} \mathbf{\Phi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \text{ and } \mathbf{\Phi} = \begin{pmatrix} \tilde{\varphi} & 0 & \cdots & 0 \\ 0 & \tilde{\varphi} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\varphi} \end{pmatrix}, \tag{3}$$

where $\mathbf{\Phi} \in R^{r \times r}$ and $rank(\mathbf{A}) = rank(\mathbf{\Phi})$. $\tilde{\varphi}$ denotes arbitrary a_i for convenience. It is easy to deduce that $rank(\mathbf{\Phi}) \leq N_{e-ug}I$, therefore $rank(\mathbf{A}) \leq N_{e-ug}I$. □

Despite of a clear structure retained by complete preference matrix **A**, the restricted mobility and the randomness of request for UEs (i.e., the Prerequisite (i) and (ii)) lead to a serious observation missing on the complete request preference data (the preference matrix **A**). The more likely situation is that a very limited number of requests are recorded and can be indicated by $P_\Omega(\mathbf{A})$, where Ω is the index set of the recorded entries, P_Ω denotes the orthogonal projection operator onto the span of matrices vanishing outside of Ω, so that the (*j*,*k*)-th component of $P_\Omega(\mathbf{A})$ is equal to the (*j*,*k*)-th component of **A** when $(j,k) \in \Omega$ and zero otherwise. Especially, for each row (column) of **A**, we find that the number of non-zero entries in the row indicates the activity of the corresponding UE. Therewith, we denote the number of non-zero entries in the *k*th column as N_k, and then define the activity of the *k*th UE as $\tilde{q}_k = N_k/N_\Omega$, where $N_\Omega = |\Omega|$ represents the number of observed non-zero entries in **A**. The activity vector of all UEs can be further defined as $\mathbf{q} = (\tilde{q}_1, \ldots, \tilde{q}_K)^T$. Similarly, the activity of eRRH means the measurement that the eRRH caches the proper contents and serves UEs. The activity vector of all eRRHs can also be defined as $\mathbf{p} = (\tilde{p}_1, \ldots, \tilde{p}_N)^T$ where $\tilde{p}_j = N_j/N_\Omega$ (N_j is the number of non-zero entries in the *j*th row). It is observed that activity brings out the non-uniform observations/samples.

With these known conditions, we define the content edge-caching pre-mapping construction as preference data matrix inferring. More specifically, the core mission is inferring the unknown potential relationships between the UE and eRRH in the F-RAN, and further partitioning the cluster according to the criterion associated with preference and randomness of requests in active area. On this basis, **our eRRH content per-caching strategy** is designed as follows: Guided by the entries of the inferred preference matrix (i.e., the pre-mapping), BBU pool selects the most popular content of each type according to the inferred results, and then pre-sends contents to several alternative eRRHs for the corresponding UE cluster service. Besides, if the eRRH cache exists in free space, the most popular contents except the pre-sent ones will be transmitted to the eRRH until no space left in the edge-cache. Due to the prediction, this strategy based on pre-mapping seems to satisfy the users' needs with high probability and meaningful in the scenario such as the super-resolution videos pre-caching for high throughput transmission QoS and low energy consumption.

3. Nonuniform Pre-Mapping Construction via Nonconvex Optimization

In this section, on the basis of Theorem 1 and properties aforementioned, we establish a non-uniform pre-mapping to infer the unknown potential preference data in the F-RAN and further achieve eRRH/UE cluster partition according to the structure of the preference matrix. Since the complete preference matrix **A** obeys the low-rank structure, it is reasonable to utilize the low-rank matrix completion [16–18] to achieve pre-mapping construction. However, main methods based upon low-rank matrix completion algorithms are assumed that the data are sampled under uniform distribution. This does not hold for our scenario and system model, owing to the close relationship between the preference data distribution and the distinct activity levels of participants. To overcome the obstacle of non-uniform observation, we propose the non-convex matrix completion with iteratively re-weighted modified trace norm regularization for clustering:

Given a distribution P_{jk} ($j = 1, \ldots, N$ and $k = 1, \ldots, K$) that reflects the UEs' activity of requesting, for the objective matrix $\mathbf{X} \in \{a_1, a_2, \ldots, a_I, 0\}^{N \times K}$, the modified trace norm of $\mathbf{X} \in R^{N \times K}$ is defined as $\|\mathbf{X}\|_{*(\mathbf{p},\mathbf{q})} = \left\| diag(\mathbf{p})^{1/2} \mathbf{X} diag(\mathbf{q})^{1/2} \right\|_*$. Here, $\|\bullet\|_*$ denotes the trace

norm, $\sigma_{\tilde{i}}\left(diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2}\right)$ denotes the $\tilde{i}$-th singular value of the matrix and redesignated as $\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})}$ for simplicity. $diag(\mathbf{p})^{1/2}$ is the diagonal matrix with the activity vector of all participants: $\mathbf{p} = \left(p_1,\ldots,p_j,\ldots,p_N\right)^T$, where p_j represents the row marginal, i.e., $p_j = \sum_{k=1}^{K} P_{jk}$. Similarly $diag(\mathbf{q})^{1/2}$ corresponds to the columns of $\mathbf{X}$ and possesses similar relations (i.e., $q_k = \sum_{j=1}^{N} P_{jk}$) (see Section 2). With this definition, matrix completion with modified trace norm regularization involves the following optimization

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|_{*(\mathbf{p},\mathbf{q})} \; s.t. \; \mathcal{P}_{\Omega}(\mathbf{X}) = \mathcal{P}_{\Omega}(\mathbf{A}). \tag{4}$$

Note that compared to l_1 norm, the l_ϑ quasi-norm , $0 < \vartheta < 1$, makes a closer approximation to the counting norm l_0, which is the number of nonzero entries of x, the variants of non-convex l_s for $0 < \vartheta < 1$ have been used to develop algorithms for recovering low-rank matrices in [19,20]. With this consideration, we let $\|\mathbf{X}\|^{\vartheta}_{\vartheta(\mathbf{p},\mathbf{q})} = \sum_i \sigma^{\vartheta}_{i,(\mathbf{p},\mathbf{q})}$ and the substitution of the optimization (4) can be given as follows:

$$\arg\min_{\mathbf{X}} \|\mathbf{X}\|^{\vartheta}_{\vartheta(\mathbf{p},\mathbf{q})} \; s.t. \; \mathcal{P}_{\Omega}(\mathbf{X}) = \mathcal{P}_{\Omega}(\mathbf{A}), \tag{5}$$

where $0 < \vartheta < 1$. We then propose an accelerated non-convex non-uniformed matrix completion algorithm (ANNMC) via variant quasi-norm optimization to adapt to the preference inference and eRRH/UE cluster partition for pre-caching in this section. The iteratively re-weighted framework and the accelerated framework are utilized for the algorithm design.

3.1. Nonconvex Nonuniformed Matrix Completion Agorithm

Inspired by the iterately re-weighted framework via l_1 norm in compressed sensing, an iterative procedure for solving the minimization problem (5) is as follows:

1. Set the iteration count $t = 1$ and $w_{\tilde{i}}^{(0)} = 1$, $\tilde{i} = 1,\ldots,N$.
2. Solve the weighted modified trace norm minimization problem

$$\mathbf{X}^{(t)} = \arg\min_{\mathbf{X}} \sum_{\tilde{i}} w_{\tilde{i}}^{(t-1)} \sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} \; s.t. \; \mathcal{P}_{\Omega}(\mathbf{X}) = \mathcal{P}_{\Omega}(\mathbf{A}), \tag{6}$$

3. Update the weights: for each $\tilde{i} = 1,\ldots,N$,

$$w_{\tilde{i}}^{(t)} = \frac{1}{\left(\sigma^{(t)}_{\tilde{i},(\mathbf{p},\mathbf{q})} + \varepsilon_{\tilde{i}}\right)^{1-\vartheta}}, \tag{7}$$

   where $\varepsilon_{\tilde{i}} > 0$ in order to provide stability and to ensure that a zero-valued $\sigma_{\tilde{i}}^{(t)}\left(diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2}\right)$ does not strictly prohibit a nozero estimate at the next step.
4. Terminate on convergence or when $t = t_{max}$. Otherwise, increment t and go to step 2.

In this method, using an iterative framework to construct the $w_i^{(t)}$ tends to allow for successively better estimation of the nonzero coefficients. Even though the early iterations may find inaccurate signal estimates, the largest signal coefficients are most likely to be identified as nonzero. Once these locations are identified, their influence is downweighted in order to allow more sensitivity for identifying the remaining small but nonzero signal coefficients.

3.2. Accelerated Agorithm for Convex Subproblem

Subsequently, we develop an accelerated variant for non-convex non-uniformed low-rank matrix completion algorithm. By referring optimization in [21], the subproblem (6) in iteration t can come down to

$$\arg\min_{\mathbf{X}} \tau \sum_{\tilde{i}} w_{\tilde{i}}^{(t-1)} \sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2} \|\mathbf{X}\|_F^2 \ s.t. \ \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{A}). \tag{8}$$

Its Lagrangian function is defined as

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = \tau \sum_{\tilde{i}} w_{\tilde{i}}^{(t-1)} \sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2} \|\mathbf{X}\|_F^2 + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{A}) \rangle, \tag{9}$$

and its dual function is

$$f(\mathbf{Y}) = \inf_{\mathbf{X}} \mathcal{L}(\mathbf{X}, \mathbf{Y}). \tag{10}$$

We then intend to utilize the dual function (10) to solve the subproblem (8). To this end, We first deduce the properties of the dual function $f(\mathbf{Y})$ and then illustrate how to achieve the optimal solution of the subproblem (8) from its dual optimum directly. Now the following results should be given, which are essential to obtain the properties of $f(\mathbf{Y})$. We omit the proofs of the properties here and present them in Appendix A.

Theorem 2. *For $\tau \geq 0$, $\mathbf{Y} \in R^{N \times K}$ and $\mathbf{w} = \{w_{\tilde{i}}\}_{\tilde{i} \in N^+}$, $0 \leq w_1 \leq \cdots \leq w_N$, the solution of the optimal problem $\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \sum_{\tilde{i}} w_{\tilde{i}} \sigma_{i,(\mathbf{p},\mathbf{q})}$ obeys*

$$\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{Y}) = diag(\mathbf{p})^{1/2} \mathbf{U}_{(\mathbf{p},\mathbf{q})} \mathbf{\Sigma}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})} \mathbf{V}_{(\mathbf{p},\mathbf{q})}^T diag(\mathbf{q})^{1/2}, \tag{11}$$

where $\mathbf{\Sigma}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})} = diag\left(\sigma_{\tilde{i}}\left(diag(\mathbf{p})^{-1/2} \mathbf{Y} diag(\mathbf{q})^{-1/2}\right) - \tau w_{\tilde{i}}\right)_+$, and $diag(\mathbf{p})^{-1/2} \mathbf{Y} diag(\mathbf{q})^{-1/2} = \mathbf{U}_{(\mathbf{p},\mathbf{q})} \mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})} \mathbf{V}_{(\mathbf{p},\mathbf{q})}^T$.

Theorem 2 plays the crucial role in formulating the optimal of the subproblem (8). Additionally, $\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\bullet)$ is equal to $\mathcal{D}_\tau(\bullet)$ when vector $\mathbf{w} = \mathbf{p} = \mathbf{q} = \mathbf{1} = (\mathbf{1}, \ldots, \mathbf{1})^T$, which is the crucial value for the traditional trace norm minimization solving. Based on the properties of Moreau-Yosida regularization and Theorem 2, we obtain the following result.

Theorem 3. *For any $\mathbf{X}, \mathbf{Y} \in R^{N \times K}$, we have:*

$$\left\| \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{X}) - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{Y}) \right\|_F^2 \leq \left\langle \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{X}) - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \right\rangle, \tag{12}$$

which indicates that $\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{Y})$ is globally Lipschitz continuous with modulus 1.

With Theorems 2 and 3, we obtain the following property of the dual function $f(\mathbf{Y})$.

Theorem 4. *For any $\tau \geq 0$, the dual function $f(\mathbf{Y})$ is continuously differentiable with Lipschitz continuous gradient at most 1, and the primal optimal $\hat{\mathbf{X}}$ of the subproblem (8) is given by $\hat{\mathbf{X}} = \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}\left(\mathcal{P}_\Omega\left(\hat{\mathbf{Y}}\right)\right)$, when the dual optimal $\hat{\mathbf{Y}}$ of the subproblem (8) is obtained.*

With the above properties, let $q(\mathbf{Y}) = -f(\mathbf{Y})$. Since $f(\mathbf{Y})$ is the dual function of (8), $f(\mathbf{Y})$ is concave and subsequently $q(\mathbf{Y})$ is convex. Thus, for any $\mathbf{Y}_1, \mathbf{Y}_2 \in R^{N \times K}$, $\langle q(\mathbf{Y}_1) - q(\mathbf{Y}_2), \mathbf{Y}_1 - \mathbf{Y}_2 \rangle \geq 0$. From Equations (A8) and (A10) in Appendix A, it is also easy to show that $q(\mathbf{Y})$ satisfies $\nabla q(\mathbf{Y}) = \mathcal{P}_\Omega\left(\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}((\mathbf{Y})) - \mathbf{A}\right)$ and belongs to $\mathcal{S}_{0,1}^{1,1}(R^{N \times K})$, which is the class of convex functions with Lipschitz gradient (i.e., for some $0 \leq \mu \leq 1$ and any $\mathbf{Y}_1, \mathbf{Y}_2 \in R^{N \times K}$, $q(\mathbf{Y})$ satisfies

$\|\nabla q(\mathbf{Y}_1) - \nabla q(\mathbf{Y}_2)\|_F \leq \|\mathbf{Y}_1 - \mathbf{Y}_1\|_F$ and $\langle \nabla f(\mathbf{Y}_1) - \nabla f(\mathbf{Y}_2), \mathbf{Y}_1 - \mathbf{Y}_2 \rangle \geq \mu \|\mathbf{Y}_1 - \mathbf{Y}_2\|_F^2$). Therefore, optimization (8) can be solved by minimizing the objective function $q(\mathbf{Y})$, i.e.,

$$\min_{\mathbf{Y}} q(\mathbf{Y}). \tag{13}$$

After acquiring this equivalent optimization $\min_{\mathbf{Y}} q(\mathbf{Y})$, in the following we propose to solve this smooth convex optimization problem by using the Nesterov's method, a very powerful optimization technique for class $\mathcal{S}_{\mu,L}^{1,1}(R^{N\times K})$, $\mu \geq 0$, $L < +\infty$ [22]. For $q(\mathbf{Y})$ belonging to $\mathcal{S}_{0,1}^{1,1}(R^{N\times K})$, The Nesterov's method for this problem utilizes two sequences: $\{\mathbf{Y}_l\}$ and $\{\mathbf{Z}_l\}$, $\mathbf{Y}_l, \mathbf{Z}_l \in R^{N\times K}$,

$$\mathbf{Z}_l = \mathbf{Y}_l + \beta_l(\mathbf{Y}_l - \mathbf{Y}_{l-1}),\ \mathbf{Y}_{l+1} = \mathbf{Z}_l - \frac{1}{L_l}\nabla q(\mathbf{Z}_l).$$

where β_l is a tuning parameter, and $1/L_l$ is the step size. By utilizing the Nemirovski's line search scheme [23], which developed from the Nesterov's method, for L_l and β_l, the update scheme is that $L_{l+1} = 2L_l$ and β_l is independent on L_l. Starting from an initial point $\mathbf{Y}_0$, $\mathbf{Z}_l$ and $\mathbf{Y}_{l+1}$ can be computed recursively, and arrive at the optimal solution $\hat{\mathbf{Y}}$. We get Algorithm 1 to achieve the optimal solution of the subproblem (8) in the t-th iteration from its dual optimum directly. By using the Nesterov's and Nemirovski's scheme framework, the Algorithm 1 for the subproblem can achieve the convergence rate of $O\left(1/t_{\max}^2\right)$.

Algorithm 1 Accelerated Algorithm for Subproblem (6)

Input: $\tilde{\mu}, \alpha_{-1} = 0.5, \mathbf{Y}_{-1}^{(t)} = \mathbf{Y}_0^{(t)} = \mathbf{Y}_n^{(t-1)}, L_{-1} = L_0, \gamma_0 \geq \tilde{\mu}, \lambda_0 = 1, \vartheta, \varepsilon_{\tilde{i}} > 0;$
Output: $\mathbf{Y}_n^{(t)}, rank\left(\mathbf{Y}_n^{(t)}\right);$
1: **for** $l = 0, 1, 2, \ldots, n$ **do**
2: **while** true **do**
3: compute $\alpha_l \in (0,1)$ as the root of $L_l\alpha_l^2 - (1-\alpha_l)\gamma_l - \alpha_l\tilde{\mu} = 0$, $\gamma_{l+1} = (1-\alpha_l)\gamma_l + \alpha_l\tilde{\mu}$, $\beta_l = \frac{(1-\alpha_{l-1})\gamma_l}{(\gamma_l + L_l\alpha_l)\alpha_{l-1}}$;
4: compute $\mathbf{Z}_l^{(t)} = \mathbf{Y}_l^{(t)} + \beta_l\left(\mathbf{Y}_l^{(t)} - \mathbf{Y}_{l-1}^{(t)}\right)$; $\mathbf{Y}_{l+1}^{(t)} = \mathbf{Z}_l^{(t)} - \frac{1}{L_l}\nabla q\left(\mathbf{Z}_l^{(t)}\right)$; $L_l = 2L_l$;
5: **end while**
6: $\lambda_{l+1} = (1-\alpha_l)\lambda_l$;
7: **end for**
8: **return** $\mathbf{Y}_n^{(t)}, rank\left(\mathbf{Y}_n^{(t)}\right);$

We then summarize the proposed ANNMC algorithm and represent the execution steps. For the t-th iteration, we solve the subproblem (8) by Algorithm 1 to get a coarse result firstly. Since the entries of the objective preference matrix belong to the set $\{a_1, a_2, \ldots, a_I, 0\}$, we add the quantification steps to ensure the result matrix in the feasible domain. To fit the observations better, the quantification thresholds are amended by the rate of value a_i in the observations. Thus let the rate $p_{a_i} = N_\Omega(a_i)/N_\Omega$ where $N_\Omega(a_i)$ is the number of a_i in the observation and the quantification rule is shown in Algorithm 2. It aims to reduce the number of iterations as much as possible that the quantification is integrated in the t-th iteration but not after solving problem (5). In addition, the convergence of the ANNMC algorithm can be guaranteed due to the convergence of the iteratively re-weighted minimization for l_s quasi-norm [19]. Specifically, because the ANNMC algorithm is based on the iteratively re-weighted framework and the $\left\{\mathbf{Y}^{(t)}\right\}$ is generated by solving subproblem (8), there exists $\hat{\mathbf{Y}}$, an accumulation point of $\left\{\mathbf{Y}^{(t)}\right\}$, which is the first-order stationary point of problem (5). The method of the convergence proof is analogous to the method in [19,24], we omit it for conciseness.

There exists an additional remark that in practice the pre-mapping constructed by the proposed algorithms can only predict the preference-caching relations between UEs and eRRHs accurately

with high probability due to the posteriori estimation. However, based on sufficient sampled data in a certain period, the preference/caching manner can be mining via the pre-mapping. The edge pre-caching strategy mentioned in Section 2 can facilitate the service with high QoS for users based on recommendation.

At the end of this section, we provide a concise example from an application to illustrate our algorithm at work. Assume that there exist C files prepared to be sent to eRRHs for pre-caching. Then, we execute the following steps to accomplish the task: (i) By recording the fragmentary request information of the UEs in the designated region, we first use the ANNMC algorithm to get the pre-mapping which concludes the UEs' preference inferences and the corresponding eRRHs accessing information associated with the geographical influence. (ii) Since there are many intersection among the preference sets of different UEs in the same group, we merge these preference sets and regard the derived union as the criterion of the edge pre-caching. (iii) Classify the C files as I types according to the UEs' preferences, and rank the file types and the files of each type respectively via the popularity in society and history of the UEs' requests. (iv) Determine which eRRH cluster serves the corresponding UE group by using the pre-mapping. Then divide files into messages of the same size and deliver to the eRRHs based on the criterion of the edge pre-caching, until the cache capacity is exhausted. (V) If the pre-cached messages match the requests of the served UEs, the messages will be sent to the UEs and get the subsequent messages of the same files from BBU pool. Otherwise the fragmentary request information record of the UEs will be updated and then the pre-mapping is modified based on the ANNMC algorithm, and so on.

Algorithm 2 ANNMC Algorithm

Input: $\tilde{\mu}, \alpha_{-1} = 0.5, \mathbf{Y}_{-1} = \mathbf{Y}_0, L_{-1} = L_0, \gamma_0 \geq \tilde{\mu}, \lambda_0 = 1, \vartheta, \varepsilon_{\tilde{i}} > 0$;
Output: $\mathbf{Y}_{t_{max}}, rank\left(\mathbf{Y}_{t_{max}}\right)$;
1: **for** $t = 1, 2, \ldots, t_{max}$ **do**
2: solve subproblem (8) by Algorithm 1;
3: **end for**
4: **if** the (k,j)-th component of $\mathbf{Y}_{t_{max}}$ satisfies
$$\left(p_{a_{i-1}} a_{i-1} + p_{a_i} a_i\right) / \left(p_{a_{i-1}} + p_{a_i}\right) \leq \left(\mathbf{Y}_{t_{max}}\right)_{j,k} < \left(p_{a_i} a_i + p_{a_{i+1}} a_{i+1}\right) / \left(p_{a_i} + p_{a_{i+1}}\right),$$
where $p_{a_i} = N_\Omega\left(a_i\right) / N_\Omega$ and $i = 2, \ldots, I$;
5: **then**
6: $\left(\mathbf{Y}_{t_{max}}\right)_{j,k} = a_i$;
7: **else**
8: $\left(\mathbf{Y}_{t_{max}}\right)_{j,k} = 0$;
9: **end if**
10: **return** $\mathbf{Y}_{t_{max}}, rank\left(\mathbf{Y}_{t_{max}}\right)$;

4. Energy Consumption Analysis

In light of mass data generated by billions of devices, we always prefer less energy consumption in data transmission, storing and processing. For this reason, edge caching aims to reduce repeated data transmission from original servers, which means that unnecessary energy consumption on packet delivery between the edge tier and the server tier can be saved. Moreover, caching itself also consumes extra energy while keeping RAM or disk memory running. To determine and sum up the overall cost of the entire simulation on the F-RAN network architecture, in summary, we may consider three parts of the energy consumption, the energy for device maintaining, content caching and transmission. Based on the content service strategy mentioned in Section 2 (also shown in Figure 2), we present the calculation of total energy consumption for each eRRH as follows:

$$E_{total} = E_{device} + E_{cache} + E_{trans}. \tag{14}$$

Note that once the devices are deployed and work, the part to maintain all devices in the F-RAN can not be expressed in the total energy consumption since it is a fixed cost and can only be reduced by shutting down some devices [25]. Therefore, consider a fixed number of devices (eRRHs and users) in the F-RAN obeying the content per-caching strategy in Section 2 and all eRRHs are received and caches contents with the same size and number as the initial state. Then the total energy consumption analysis for each eRRH can be equivalent to the discussion on its total energy consumption change, which is the total energy cost for the content caching and the transmission for content update:

$$E_{total\Delta} = E_{cache} + E_{trans\Delta}. \tag{15}$$

Especially, the caching energy cost

$$E_{cache} = \eta E_{content}, \tag{16}$$

where $E_{content}$ is the power consumption of keeping a content object in an eRRH and η is the is the size of the eRRH cache (the number of the content object). Meanwhile, since the content transmission energy cost of each eRRH contains the content sending and receiving energy cost by BBU pool and eRRH respectively, the transmission energy cost for the single content update $E_{content\Delta} = E_{send} + E_{recv}$. Then similar to the energy consumption linear model in [26,27],

$$E_{trans\Delta} = \eta \left((1 - \theta) E_{content\Delta} + \delta \right), \tag{17}$$

where δ represents fixed costs and θ represents the rate of the required contents for the cluster users in the cache (i.e., $1 - \theta$ represents the content update rate).

With this energy consumption analysis model, we discuss $E_{total\Delta}$ under the condition upon the implement of the pre-mapping constructed by the proposed preference inference algorithm, i.e., ANNMC. For E_{cache}, one part of the $E_{total\Delta}$, because the inference algorithm with appropriate non-uniform sampling rate may accurately estimate the preference data of the cluster with high probability, the number of the caching contents η would be less than that of the blind caching. In other words, the accurate preference estimation may prompt the eRRH cluster cache only several certain kinds of contents but not as many kinds as possible to ensure the QoS of F-RAN. On the other hand, for $E_{trans\Delta}$, θ will be determined by the inference error of the preference data. Thus, the $E_{trans\Delta}$ directly associates with the accuracy of the proposed inference algorithm. Let η_{min} denote the minimum number of cached content objects and η_{max} denote the specific maximum number of contents allowed for caching, then the range of the total energy consumption change for each eRRH with the inference guidance is

$$\eta_{\min} E_{content} \leq E_{total\Delta}|_{infer} \leq \eta_{\max} \left(E_{content} + E_{content\Delta} + \delta \right). \tag{18}$$

Especially, with the low-rank assumption of the preference data and the proper low non-uniform sampling rate, the proposed ANNMC for inference can ensure the total energy consumption to approximate the lower bound with high probability, due to the foreseeable algorithm accuracy.

5. Performance Evaluations

The performance evaluations are illustrated in this section by using MATLAB. We first perform experiments on the synthetic networks, including static and dynamic cases, and show that our system model and the proposed pre-mapping construction method is feasible on the task of content preference distribution inference for the pre-caching in the F-RAN. Additionally, experimental simulations about the energy consumption performance of the pre-caching strategy are provided as well. To ensure that our results are reliable, we conduct all experiments 200 times, and average the results from all of the trials.

5.1. Performance of the Preference Inference

To verify the validity, we first consider a pre-mapping (complete preference matrix $\mathbf{A}$) constructed by the synthetic F-RAN link data. The observation matrix $P_\Omega(\mathbf{A})$ is formed by sampling some entries from $\mathbf{A}$. To be specific, for the F-RAN, we set several eRRH clusters. The clusters of UEs are formed on the basic pattern of the $\mathbf{A}$ and located randomly. Each UE cluster is served by one of eRRH cluster. Accordingly, $\mathbf{A}$ possesses complete diagonal-block structure which is mentioned and analysed in Section 2. The size of each UE cluster is larger than 20 and the sum of the sizes is $K = 1000$ and the size of each eRRH cluster is larger than 5 and the sum of the sizes is $N = 200$. Meanwhile, in the F-RAN the content types is sorted via popularity and its number $I = 15$. For convenience, the preferred type subset of each UE cluster is sampled from the type set by Zipf distribution obeying the sorting. We further assume that only a part of requests are recorded by non-uniform sampling and the file types are valued by different integers (the max number of the content types contained by each eRRH cluster is set to be 5 as an example). The probability distribution of sampling ensures that each row (column) of the original matrix is sampled with different $p_s \in (0, 1)$, and the practical sampling-rate of the original matrix $\mathbf{A}$ is defined as $\|P_\Omega(\mathbf{A})\|_{0,1} \Big/ \|\mathbf{A}\|_{0,1}$, where $\|\bullet\|_{0,1}$ is equal to the number of non-zero entries in the matrix. Then we use our proposed ANNMC algorithm to estimate the complete matrix $\mathbf{A}$ and compare the performance of our approach to traditional Alternating Least Square (tALS) [28], the weighted trace norm regularization (WTNR) [29] and Accelerated Singular Value Thresholding (ASVT) [30] for the content caching pre-mapping construction problem. The details of the ANNMC parameters are shown in Table 1, and the experimental settings of the compared methods are the same with the corresponding literatures.

Table 1. The parameter settings of the ANNMC.

Parameter	τ	ϑ	ε_i	$\tilde{\mu}$	α_{-1}	L_0	λ_0
Setting	$2\sqrt{NK}$	0.5	10^{-7}	0.1	0.5	$p_s/1.1$	1

We evaluate the performances by the similarity between the inferred matrix $\hat{\mathbf{A}}$ and the original matrix $\mathbf{A}$ to indicate the accuracy of estimation, the definition of which is $S_{\mathbf{A},\hat{\mathbf{A}}} = \left|\langle \mathbf{A}, \hat{\mathbf{A}} \rangle\right| / \|\mathbf{A}\|_F \left\|\hat{\mathbf{A}}\right\|_F$. The range of the practical sampling-rate is from 0.1 to 0.9 and plot the inference accuracy in Figure 4. Apparently, the proposed non-uniform algorithm outperforms others due to higher accuracy. Moreover, we choose matrices (200×1000) with ranks $r = 6, 8, \ldots, 24, 26$ (i.e., the number of the clusters) and non-uniform sampling rate $p_s = 0.25$ for matrix completion. For each algorithm, we complete the structure with different N_{e-ug} and compute recovery relative accuracy by similarity. The result is shown in Figure 5a. It is observed that ANNMC possesses the better robustness of non-uniform matrix completion in the certain range of eRRH/UE cluster number. We further compare ANNMC and WTNR due to their better estimation performance than the others in the simulation. As the results shown in Figure 5b, ANNMC outperforms WTNR with non-uniform sampling on low sampling-rate based estimation, even though its convergence rate is only slightly faster than WTNR. An visualized example of the pre-mapping constructed by ANNMC algorithm is also shown in Figure 6 ($N_{e-ug} = 15$ and $I = 15$ with sampling rates $p_s = 0.2$ and $p_s = 0.4$), which consists with our results.

5.2. Performance of the Dynamic Preference Inference

In general, the pre-mapping $\mathbf{A}$ varies during a time period long enough. Accordingly, the observations of this dynamic preference matrix over time essentially introduce time dimension to the problem of mining the potential eRRH/content-UE relationships. Therefore, a more realistic scenario is inferring the $\mathbf{A}$ by utilizing the historical records of the eRRH content caching and UEs' requests in global and temporal evolvement perspectives. In particular, assume that we are given an incomplete non-uniformed observation tensor (or 3-dimensional array), which consists of the preference matrix pattern corresponding to the snapshots of the underlying dynamic relationships at

time $T = T_0+1, T_0+2, \ldots, T_0+\hat{T}$. Then the preference inference task is to estimate the possible pattern of the dynamic preference matrix at time $T_0+\hat{T}$ based on the given the 3-dimensional observation tensor. It is notice that, despite maintaining the dynamic property, the underlying eRRH/content-UE relationships in reality always display some "redundancy" attributed to the gradual periodic variation and the relatively stability [31]. With this property, the tensor consisting of the preference matrix patterns at T can be considered to be low-rank.

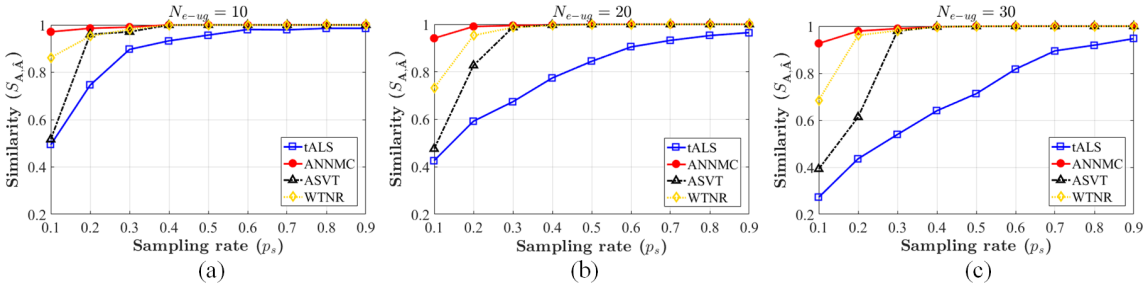

Figure 4. Comparison of similarity by different preference inferring algorithms (tALS, ASVT, WTNR and ANNMC) on the synthetic dataset. The number of blocks $N_{e-ug} = 10$ corresponding to (**a**), 20 corresponding to (**b**) and 30 corresponding to (**c**). The range of the non-uniform sampling-rate p_s is varied from 0.1 to 0.9 and the content types $I = 15$.

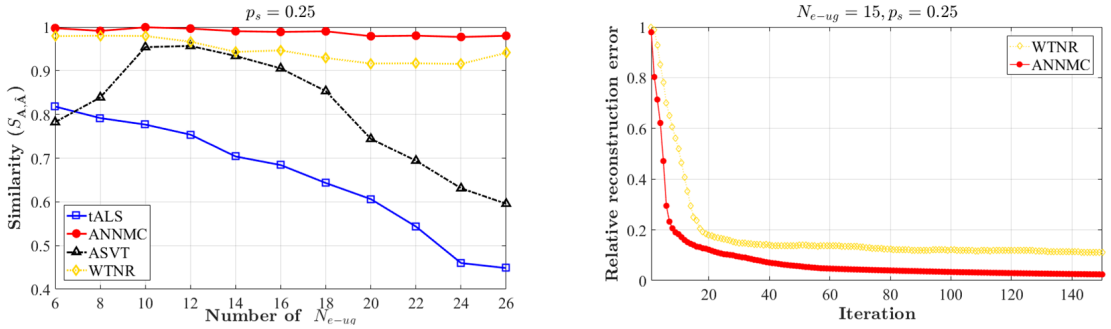

Figure 5. (**a**) shows the estimation similarity of the structure estimation for the preference matrix **A** with the number of the content types $I = 15$, the number of groups $N_{e-ug} = 6, 8, \ldots, 26$ and the non-uniform sampling-rate $p_s = 0.25$; (**b**) is Convergence rate of ANNMC and WTNR on the synthetic data **A** with the number of the content types $I = 15$, $N_{e-ug} = 15$ and the non-uniform sampling-rate $p_s = 0.25$.

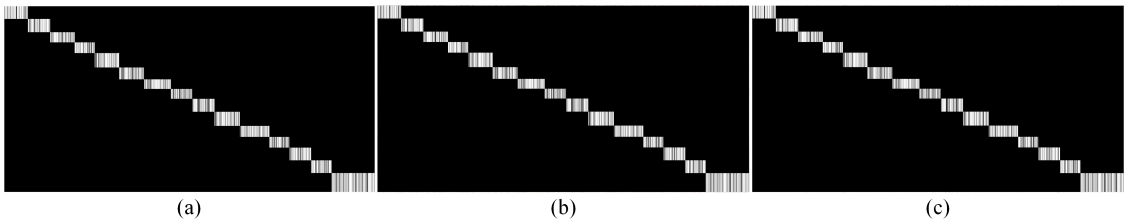

Figure 6. An visualized example of the pre-mapping construction for the preference matrix **A** ($N_{e-ug} = 15$, $I = 15$). (**a**) is the original preference matrix; (**b**) is the estimation result with 20% non-uniform samples (the similarity is 0.9942); (**c**) is the estimation result with 40% non-uniform samples (the similarity is 0.9991).

To construct synthetic networks for simulation, we first consider a complete dynamic eRRH/content-UE relationship network whose preference tensor is $\mathcal{A} \in R^{I_1 \times I_2 \times I_3}$, $I_1 = N, I_2 = K, I_3 = \hat{T}$. The slide of $\mathcal{A}$ at time T is an preference matrix pattern $\mathbf{A}^{(T)}$ in the form of (1). In addition, for the gradual periodic variation, only a few kinds of $\mathbf{A}^{(T)}$ (eRRH-UE group partition styles) exist in $\mathcal{A}$ and the patterns similar to each others are usually close in time. The observation tensor $P_\Omega(\mathcal{A})$ is formed by sampling some entries from $\mathcal{A}$ non-uniformly. Concretely, we let the F-RAN with the same settings in Section 5.1. Besides, for $\mathcal{A}$, let the eRRH-UE cluster number be varied from 10 to 30 with step size 5, and for each cluster, 4 eRRH-UE group styles are generated randomly with Gaussian distribution as the basic styles. Then each basic style generates the other derived styles by perturbation and the total number of group style set is 40. We further assume that $I_3 = \hat{T} = 100$ and each group style appears randomly but more than once. The non-uniform sampling-rate of the original tensor $\mathcal{A}$ from 0.1 to 0.9 (it is realized by non-uniform sampling each slide with the given sampling-rate). Then with the given observed $P_\Omega(\mathcal{A})$, the task of the preference inference at time $T_0 + \hat{T}$ is achieved by solving the following problem:

$$\arg\min_{\mathcal{X} \in R^{I_1 \times I_2 \times I_3}} \sum_{\hat{i}=1}^{3} \left\| \mathbf{X}_{(\hat{i})} \right\|_{\vartheta(\mathbf{p},\mathbf{q})}^{\vartheta} \; s.t. \; \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{A}), \tag{19}$$

where $\mathbf{X}_{(\hat{i})} \in R^{I_{\hat{i}} \times I_1 \cdots I_{\hat{i}-1} I_{\hat{i}+1} \cdots I_3}$ is the kth mode on $\mathcal{X}$, and extracting the slide of $\mathcal{X}$ at the time $T_0 + \hat{T}$ as the inferred pattern $\hat{\mathbf{A}}^{(T_0+\hat{T})}$.

Similar with the approach in Section 3, we use ten-ANNMC, the variation of the algorithm (2) for tensor completion problem (19) to estimate $\hat{\mathbf{A}}^{(T_0+\hat{T})}$, and compare the performance of our approach to the variations of the same algorithms in Section 5.1 for tensor completion (ten-tALS, ten-ASVT and ten-WTNR). $S_{\mathbf{A}^{(T_0+\hat{T})}, \hat{\mathbf{A}}^{(T_0+\hat{T})}}$, the similarity between the inferred matrix $\hat{\mathbf{A}}^{(T_0+\hat{T})}$ and the original setting $\mathbf{A}^{(T_0+\hat{T})}$, is utilized to indicate the accuracy of estimation as well. We implement the algorithms to infer one of the group patterns generated before with the eRRH-UE cluster number $N_{e-ug}^{(\hat{T})}$ = 10, 20, 30 respectively and plot the inference accuracy in Figure 7. The inferred results is shown that our inference approach can almost be accurately estimated the objective group style and apparently outperforms the others. Meanwhile, the results for tensor completion case is worse than the matrix completion in Section 5.1 when the sampling-rate is low ($p_s = 0.1$ and 0.2). This phenomenon is due to the fact that the inference based on tensor completion blends the the observation records of the different patterns with the $\hat{\mathbf{A}}^{(T_0+\hat{T})}$. However, this strategy is benefit for the content pre-caching since the inferred matrix pattern can partly integrate the features of the observation $\mathbf{A}^{(T)}$ ($T = T_0 + 1, \ldots, T_0 + \hat{T} - 1$) and be more likely to meet the requests.

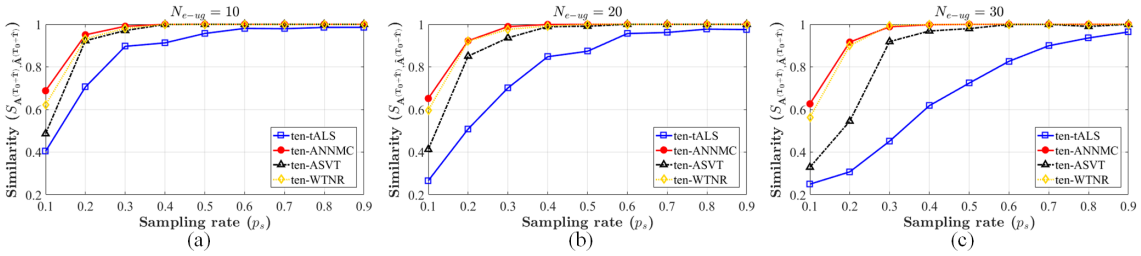

Figure 7. Comparison of similarity by different preference inferring algorithms for tensor completion (ten-tALS, ten-ASVT, ten-WTNR and ten-ANNMC) on the synthetic dataset. The number of blocks $N_{e-ug}^{(\hat{T})} = 10$ corresponding to (**a**), 20 corresponding to (**b**) and 30 corresponding to (**c**). The range of the non-uniform sampling-rate p_s is varied from 0.1 to 0.9 and the content types $I = 15$.

5.3. Performance of the Energy Consumption

In this section, simulation results are provided to evaluate the energy consumption of the proposed cluster content caching strategy in the F-RAN. The terms for energy consumption simulation are consistent with the description in Section 4. Referring to the linear energy consumption model and the real energy measurements in [26], the power consumption of the single content caching for the eRRH and backhaul transmissions are set as $E_{content} = 0.05$ W·sec and $E_{content\Delta} = 0.24$ W·sec, respectively. Assume that the content type number I and N_{cont}, the number of the best popular contents corresponding to the different types, are given and large enough. Also, all of the eRRHs in one cluster cache the same contents to ensure the service quality. We then construct the complete preference matrix $\mathbf{A}$ according to the method mentioned in Section 5.1 at the beginning as reference. The contents cached in each eRRH cluster for UEs' requests are determined by using pre-caching strategy in Section 2 (which can be presented by the value set of the corresponding block in $\mathbf{A}$). Since the caching contents in this scenario are exactly matched the UEs' requests without retransmission, the corresponding average total energy cost for each eRRH can be considered to be the minimum, i.e, $E_{min-total} = Aver\{0.05\eta_{\min}\}$ W·sec.

With the aforementioned assumptions, two pre-caching strategies are applied and compared for energy consumption analysis. The first one is the proposed edge caching/transmission strategy based on the content caching pre-mapping and description in Section 2. Due to the fact that η is the is the size of the eRRH cache (see Equation (16)) and $\eta \geq \eta_{\min}$, to improve the QoS for users, we will use the residual memory space to non-repetitively cache the contents by Zipf distribution, which obeys the popularity sorting for all contents. Subsequently, all cache in eRRH will be used and the average caching anergy cost is 0.05η W·sec. Furthermore, note that the probability of retransmission is closely associated with the accuracy of the $\mathbf{A}$ inferred by ANNMC, we set θ is equal to the similarity between the $\hat{\mathbf{A}}$ and the $\mathbf{A}$, i.e., $S_{\mathbf{A},\hat{\mathbf{A}}}$. Therefore, with (16) and (17), the average total energy cost for each eRRH based on the proposed strategy is presented as $Aver\{E_{total\Delta}|_{infer}\} = (0.05 + 0.24(1-\theta))\eta$ W·sec (δ in (17) is omitted since it is extremely smaller than the other terms). To compare performances, the traditional uniform per-caching strategy is given for simulation. For the traditional one, the active eRRHs in the F-RAN utilize all caches to retain the same content set S_η selected by Zipf distribution obeying the popularity sorting for all contents. Hence, while the E_{cache} is equal to that of the proposed strategy, the retransmission energy cost is $Aver\{0.24\eta_{retrans}\}$ W·sec where $\eta_{retrans} = |S_m - S_\eta|$. Then $Aver\{E_{total\Delta}|_{trad}\} = (0.05\eta + 0.24Aver\{|S_m - S_\eta|\})$ W·sec. Both of the strategies are compared with the minimum consumption $E_{min-total}$, which represents the lower bound.

The energy consumption performances of the per-caching strategies are evaluated as follows. For one thing, given the numbers of the UEs and eRRHs respectively ($K = 1000$, $N = 200$) and fixed the UE/eRRH cluster number ($N_{e-ug} = 15$), the number of the all alternative contents N_{cont} and the Zipf exponent s are changed respectively ($N_{cont} = 15, 25$; $s = 0, 0.5, 1, 2$). The non-uniform sampling rate for pre-mapping construction by ANNMC is set to be $p_s = 0.2$ and the corresponding inference similarity $S_{\mathbf{A},\hat{\mathbf{A}}} = 0.9942$. The average energy consumptions generated by the different number of caching contents in each eRRH are illustrated in Figure 8. By given $N_{e-ug} = 15$ and $s = 0.5$, the trends of the energy consumption via the changes of the caching content number and the alternative content number are shown in Figure 9 as well. As shown in the figures, the proposed caching strategy via the pre-mapping inference emerges better performance of the energy efficiency than the traditional one.

Meanwhile, the energy efficiency performance of the traditional strategy is approximating to the caching strategy via inference as s increasing, since the user interests converge to fewer popular content objects stored in the BBU pool. For the other, we fix N_{cont} and the number of caching contents in each eRRH (μ) and show the energy consumption associated with the number of the UE/eRRH clusters in Figure 10. While the better performance of the proposed strategy and the similar properties related to the Zipf exponent s are illustrated in this figure, the average energy consumption of the proposed strategy slightly increases as the UE/eRRH cluster number is enlarged (i.e., the size of the cluster is decreased), since the rank of the matrix (i.e., N_{e-ug}) effects the inferring accuracy by using ANNMC

($S_{\mathbf{A},\hat{\mathbf{A}}} = 0.9954, 0.9908, 0.9866, 0.9799, 0.9749, 0.9712$ when N_{e-ug} changes from 15 to 40 with 5 interval). Also, due to the user interests converging to fewer popular content objects when s is increasing as well, the average energy consumption increasing trend is weaken. Furthermore, by given $\mu = 10$, $s = 0.5$ and varied N_{e-ug} form 40 to 200, the trends of the energy consumption via the changes of the caching content number and the UE/eRRH cluster number are shown in Figure 11. It illustrates that the energy efficiency performance is weaken along with the increasing of the caching content number and the UE/eRRH cluster number. Especially, fixing the caching content number, the worst energy efficiency performance of our proposed strategy is almost reached when the UE/eRRH cluster number meets the maximum (i.e., $N_{e-ug} = 200$).

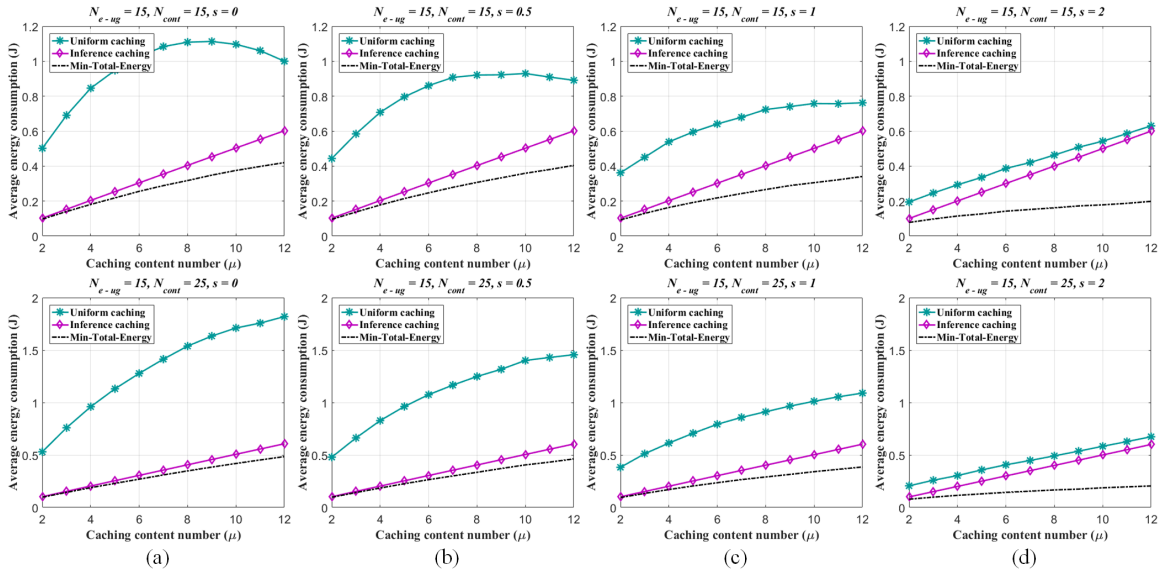

Figure 8. Average energy consumption of each eRRH in the F-RAN on the synthetic dataset (measured as J due to J = W·sec). Two different edge caching strategies are considered: (i) uniform caching strategy via content Zipf distribution, (ii) inferring caching strategy via pre-mapping (inferred by ANNMC with non-uniform sampling rate $p_s = 0.2$) and Zipf distribution. The average energy consumptions of both strategies are compared with the minimum consumption $E_{min-total}$. The range of the caching content number of each eRRH is varied from 2 to 12. The Zipf exponent $s = 0, 0.5, 1, 2$, and (**a**–**d**) correspond to different s with $N_{e-ug} = 15$, $N_{cont} = 15$ and 25.

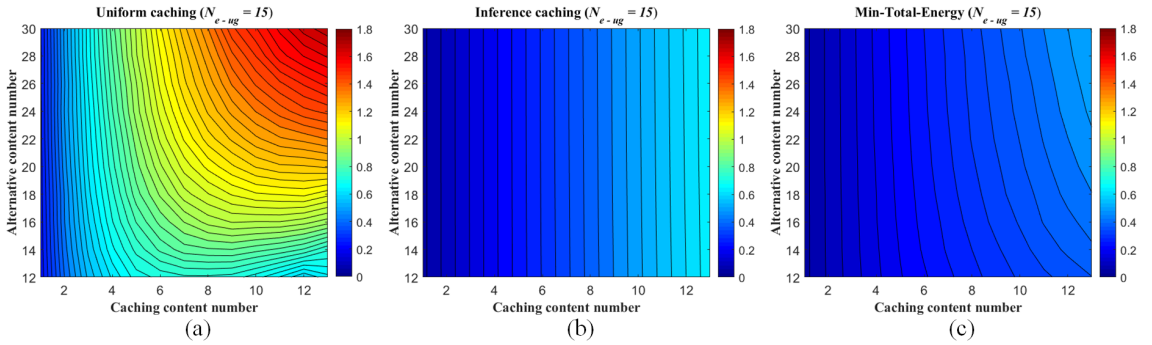

Figure 9. Average energy consumption range of each eRRH in the F-RAN on the synthetic dataset (measured as J) with the setting that $N_{e-ug} = 15$ and Zipf exponent $s = 0.5$. (**a**) is the uniform caching strategy via content Zipf distribution; (**b**) is the proposed inferring caching strategy via pre-mapping (inferring by ANNMC with non-uniform sampling rate $p_s = 0.2$) and Zipf distribution; (**c**) is the minimum consumption $E_{min-total}$.

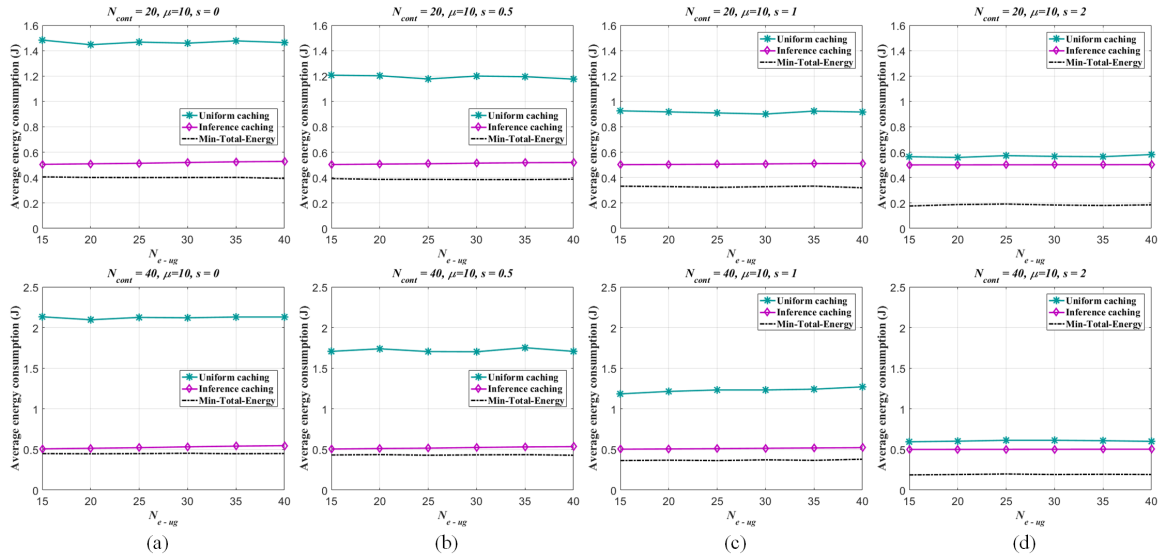

Figure 10. Average energy consumption of each eRRH in the F-RAN on the synthetic dataset (J = W·sec). Two different edge caching strategies are considered: (i) uniform caching strategy via content Zipf distribution, (ii) inferring caching strategy via pre-mapping (inferred by ANNMC with non-uniform sampling rate $p_s = 0.2$) and Zipf distribution. The average energy consumptions of both strategies are compared with the minimum consumption $E_{min-total}$. The range of the eRRH/UE cluster number is varied from 15 to 40. The Zipf exponent $s = 0, 0.5, 1, 2$, and (**a**–**d**) correspond to different s with $\mu = 10$, $N_{cont} = 20$ and 40.

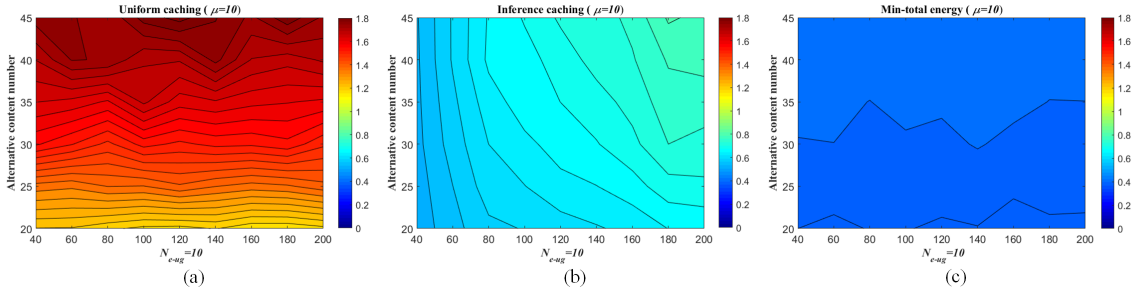

Figure 11. Average energy consumption range of each eRRH in the F-RAN on the synthetic dataset (measured as J) with the setting that $\mu = 10$ and Zipf exponent $s = 0.5$. (**a**) is the uniform caching strategy via content Zipf distribution; (**b**) is the proposed inferring caching strategy via pre-mapping (inferring by ANNMC with non-uniform sampling rate $p_s = 0.2$) and Zipf distribution; (**c**) is the minimum consumption $E_{min-total}$.

6. Conclusions

In this paper, based on digging out the non-uniform observed users' network behavior information, we propose an edge content pre-caching strategy in F-RANs to improve the QoS for the users and energy efficiency. Especially, we analyze the relationships among the users' activity, the content requesting and the state of the eRRH caching, and established pre-mapping for users' content preference inferring by an accelerated non-convex matrix completion algorithm with the non-uniform observations. The dynamic scenario is also discussed and the developed variation based on tensor completion possesses good performance for the inferring task. In addition, the energy consumption analysis is given to reveal the properties of the pre-mapping based caching strategy. The simulation results show that the inferring algorithm for pre-mapping construction (preference inferring) is effective. Meanwhile, the inferring-based caching strategy possesses the advantages in energy saving compared to traditional uniform caching via Zipf distribution.

Author Contributions: Y.C. (Yi Cen) contributes mostly to this paper on introduction, system model and solutions. Y.C. (Yi Cen) and K.W. main contribution is to the algorithm design. Y.C. (Yi Cen), Y.C. (Yigang Cen) and J.L. main contribution is to the simulation and analysis part.

Funding: This research was funded by the National Natural Science Foundation of China (61602538, 61872034, 61572067), Project D010109, the Double-First Class Fund and the Scientific Research Fund for Young Teachers of Minzu University of China.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Proof of Theorem 2. Let $\mathbf{\Sigma} = diag\left(\sigma_{\tilde{i},\mathbf{X}}\right)$ and $\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})} = diag\left(\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})}\right)$, which implies the nonzero entries of the diagonal matrices are in non-increasing order. Since the penalty term only depends on the singular values of $diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2}$, problem $\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 + \tau\sum_{\tilde{i}} w_{\tilde{i}}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})}$ can be equivalently written as

$$\min_{\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}} \left\{ \min_{\mathbf{X}} \left(\frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 + \tau\sum_{\tilde{i}} w_{\tilde{i}}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} \right) \right\}. \tag{A1}$$

For the inner minimization, due to von Neumann's trace inequality, we have

$$\begin{aligned}
\frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 &= tr\left(\mathbf{Y}\mathbf{Y}^T\right) - 2tr\left(\mathbf{Y}\mathbf{X}^T\right) + tr\left(\mathbf{X}\mathbf{X}^T\right) \\
&= tr\left(\mathbf{\Sigma}_{\mathbf{Y}}^2\right) - 2tr\left(diag(\mathbf{p})^{-1/2}\mathbf{Y}diag(\mathbf{q})^{-1/2}\left(diag(\mathbf{p})^{-1/2}\mathbf{X}diag(\mathbf{q})^{-1/2}\right)^T\right) + tr\left(\mathbf{\Sigma}^2\right) \\
&\geq tr\left(\mathbf{\Sigma}_{\mathbf{Y}}^2\right) - 2tr\left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\right) + tr\left(\mathbf{\Sigma}^2\right).
\end{aligned}$$

Especially the equality holds when $diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2}$ admits the singular value decomposition $diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2} = \mathbf{U}_{(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\mathbf{V}_{(\mathbf{p},\mathbf{q})}^T$, where $\mathbf{U}_{(\mathbf{p},\mathbf{q})}$ and $\mathbf{V}_{(\mathbf{p},\mathbf{q})}$ are defined as the left and right singular matrices of $\mathbf{p}^{-1/2}\mathbf{Y}\mathbf{q}^{-1/2}$. Meanwhile $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are defined as the left and right singular matrices of $\mathbf{Y}$. Then the optimization reduces to

$$\min_{\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}} \left\{ \frac{1}{2}tr\left(\mathbf{\Sigma}^2\right) - tr\left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\right) + \tau tr\left(diag\left(w_{\tilde{i}}\right)\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\right) + \frac{1}{2}tr\left(\mathbf{\Sigma}_{\mathbf{Y}}^2\right) \right\}. \tag{A2}$$

Note that $diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2} = diag(\mathbf{p})^{1/2}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T diag(\mathbf{q})^{1/2}$, i.e.,

$$\mathbf{\Sigma} = \mathbf{U}^T diag(\mathbf{p})^{1/2}\mathbf{U}_{(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\mathbf{V}_{(\mathbf{p},\mathbf{q})}^T diag(\mathbf{q})^{1/2}\mathbf{V}, \tag{A3}$$

the objective function is completely separable and is minimized only when

$$\nabla_{\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}} \left(\frac{1}{2}tr\left(\mathbf{\Sigma}\mathbf{\Sigma}^T\right) - tr\left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\right) + tr\left(\tau diag\left(w_{\tilde{i}}\right)\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\right) \right) = \mathbf{0}, \tag{A4}$$

where $\nabla_{\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}}(\bullet)$ denotes the gradient for $\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}$.

Utilizing $\nabla_{\mathbf{X}} tr\left(\mathbf{A}^T\mathbf{X}\mathbf{B}\right) = \mathbf{A}\mathbf{B}^T$ and $\nabla_{\mathbf{X}} tr\left(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^T\right) = \mathbf{A}^T\mathbf{X}\mathbf{B}^T + \mathbf{A}\mathbf{X}\mathbf{B}$, we have

$$\begin{aligned}
&\mathbf{U}_{(\mathbf{p},\mathbf{q})}^T diag(\mathbf{p})^{1/2}\mathbf{U}\left(\mathbf{U}^T diag(\mathbf{p})^{1/2}\mathbf{U}_{(\mathbf{p},\mathbf{q})}\mathbf{\Sigma}_{(\mathbf{p},\mathbf{q})}\mathbf{V}_{(\mathbf{p},\mathbf{q})}^T diag(\mathbf{q})^{1/2}\mathbf{V}\right)\mathbf{V}^T diag(\mathbf{q})^{1/2}\mathbf{V}_{(\mathbf{p},\mathbf{q})} \\
&= \left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})} - diag\left(\tau w_{\tilde{i}}\right)\right)_+,
\end{aligned}$$

i.e.,

$$\mathbf{U}_{(\mathbf{p},\mathbf{q})}^T diag(\mathbf{p})^{1/2}\mathbf{X}diag(\mathbf{q})^{1/2}\mathbf{V}_{(\mathbf{p},\mathbf{q})} = \left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})} - diag\left(\tau w_{\tilde{i}}\right)\right)_+ \tag{A5}$$

Therefore $\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathbf{Y}) = \mathbf{p}^{1/2}\mathbf{U}_{(\mathbf{p},\mathbf{q})}\left(\mathbf{\Sigma}_{\mathbf{Y},(\mathbf{p},\mathbf{q})} - diag\left(\tau w_{\tilde{i}}\right)\right)_{+}\mathbf{V}^{T}_{(\mathbf{p},\mathbf{q})}\mathbf{q}^{1/2}$ is a global optimal solution to the optimization problem. The uniqueness of the solution follows by the equality condition for the von Neumann's trace inequality when has distinct nonzero singular values, and the uniqueness of the strictly convex optimization (A2). This concludes the proof. □

Proof of Theorem 4. Since $\mathcal{L}(\mathbf{X},\mathbf{Y}) = \tau\sum_{\tilde{i}} w_{\tilde{i}}^{(t)}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2}\|\mathbf{X}\|_F^2 + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{A})\rangle$,

$$\begin{aligned} f(\mathbf{Y}) &= \inf_{\mathbf{X}} \mathcal{L}(\mathbf{X},\mathbf{Y}) = \inf_{\mathbf{X}}\left(\tau\sum_{\tilde{i}} w_{\tilde{i}}^{(t)}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2}\|\mathbf{X}\|_F^2 + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{A})\rangle\right) \\ &= \inf_{\mathbf{X}}\left(\tau\sum_{\tilde{i}} w_{\tilde{i}}^{(t)}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2}\|\mathbf{X}\|_F^2 + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{X})\rangle - \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{A})\rangle\right) \\ &= \inf_{\mathbf{X}}\left(\tau\sum_{\tilde{i}} w_{\tilde{i}}^{(t)}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2}\|\mathbf{X} - \mathcal{P}_\Omega(\mathbf{Y})\|_F^2\right) + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{A})\rangle - \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y})\|_F^2 \\ &= g(\mathbf{Y}) + \langle \mathbf{Y}, \mathcal{P}_\Omega(\mathbf{A})\rangle - \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y})\|_F^2 \end{aligned} \tag{A6}$$

For $g(\mathbf{Y})$, using the well-known properties of Moreau-Yosida Regularization [32], we get the results that $g(\mathbf{Y})$ is a globally continuously differentiable convex function.

Moreover, $\nabla g(\mathbf{Y}) = \mathcal{P}_\Omega\left(\mathbf{Y} - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}))\right)$ and $\nabla g(\mathbf{Y})$ is continuously differentiable with Lipschitz continuous gradient 1, i.e., for any $\mathbf{Y}_1, \mathbf{Y}_2 \in R^{N\times K}$,

$$\|\nabla g(\mathbf{Y}_1) - \nabla g(\mathbf{Y}_2)\|_F \le \|\mathcal{P}_\Omega(\mathbf{Y}_1) - \mathcal{P}_\Omega(\mathbf{Y}_2)\|_F \le \|\mathbf{Y}_1 - \mathbf{Y}_2\|_F. \tag{A7}$$

Then the gradient of $f(\mathbf{Y})$ can be obtained as follows:

$$\begin{aligned} \nabla f(\mathbf{Y}) &= \nabla g(\mathbf{Y}) + \mathcal{P}_\Omega(\mathbf{A}) - \mathcal{P}_\Omega(\mathbf{Y}) \\ &= \mathcal{P}_\Omega\left(\mathbf{Y} - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}))\right) + \mathcal{P}_\Omega(\mathbf{A}) - \mathcal{P}_\Omega(\mathbf{Y}) \\ &= \mathcal{P}_\Omega\left(\mathbf{A} - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}))\right). \end{aligned} \tag{A8}$$

It follows that for any $\mathbf{Y}_1, \mathbf{Y}_2 \in R^{N\times K}$,

$$\begin{aligned} \|\nabla f(\mathbf{Y}_1) - \nabla f(\mathbf{Y}_2)\|_F &= \left\|\mathcal{P}_\Omega\left(\mathbf{A} - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_1))\right) - \mathcal{P}_\Omega\left(\mathbf{A} - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_2))\right)\right\|_F \\ &= \left\|\mathcal{P}_\Omega\left(\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_1)) - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_2))\right)\right\|_F \\ &\le \left\|\mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_1)) - \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\mathbf{Y}_2))\right\|_F. \end{aligned} \tag{A9}$$

Following from Theorem 3 and Cauchy-Schwarz inequality, we have

$$\|\nabla f(\mathbf{Y}_1) - \nabla f(\mathbf{Y}_2)\|_F \le \|\mathcal{P}_\Omega(\mathbf{Y}_1) - \mathcal{P}_\Omega(\mathbf{Y}_2)\|_F \le \|\mathbf{Y}_1 - \mathbf{Y}_2\|_F. \tag{A10}$$

When the dual optimal $\hat{\mathbf{Y}}$ is obtained, by using the result of Equation (A6), we can get

$$\begin{aligned} \hat{\mathbf{X}} &= \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X},\mathbf{Y}) = \arg\min_{\mathbf{X}}\left(\tau\sum_{\tilde{i}} w_{\tilde{i}}^{(t)}\sigma_{\tilde{i},(\mathbf{p},\mathbf{q})} + \frac{1}{2}\|\mathbf{X} - \mathcal{P}_\Omega(\hat{\mathbf{Y}})\|_F^2\right) \\ &= \mathcal{D}_{\tau,\mathbf{w},(\mathbf{p},\mathbf{q})}(\mathcal{P}_\Omega(\hat{\mathbf{Y}})). \end{aligned} \tag{A11}$$

This concludes the proof. □

References

1. Zikria, Y.; Kim, S.; Afzal, M.; Wang, H.; Rehmani, M. 5G Mobile services and scenarios: Challenges and solutions. *Sustainability* **2018**, *10*, 3626. [CrossRef]
2. Checko, A.; Christiansen, H.L.; Yan, Y.; Scolari, L.; Kardaras, G.; Berger, M.S.; Dittmann, L. Cloud RAN for mobile networks—A technology overview. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 405–426. [CrossRef]
3. Peng, M.; Wang, C.; Lau, V.; Poor, H.V. Fronthaul-constrained cloud radio access networks: Insights and challenges. *IEEE Wirel. Commun.* **2015**, *22*, 152–160. [CrossRef]
4. Wang, K.; Li, X.; Ji, H.; Du, X. Modeling and optimizing the LTE discontinuous reception mechanism under self-similar traffic. *IEEE Trans. Veh. Technol.* **2016**, *65*, 5595–5610. [CrossRef]
5. Wang, K.; Yu, X.; Lin, W.; Deng, Z.; Liu, X. Computing aware scheduling in mobile edge computing system. *Wirel. Netw.* **2019**. [CrossRef]
6. Park, S.H.; Simeone, O.; Sahin, O.; Shitz, S.S. Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory. *IEEE Signal Process. Mag.* **2014**, *31*, 69–79. [CrossRef]
7. Peng, M.; Yan, S.; Zhang, K.; Wang, C. Fog-computing-based radio access networks: Issues and challenges. *IEEE Netw.* **2015**, *30*, 46–53. [CrossRef]
8. Bi, S.; Zhang, R.; Ding, Z.; Cui, S. Wireless communications in the era of big data. *Commun. Mag. IEEE* **2015**, *53*, 190–199. [CrossRef]
9. Park, S.H.; Simeone, O.; Shitz, S.S. Joint optimization of cloud and edge processing for fog radio access networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 7621–7632. [CrossRef]
10. Peng, X.; Shen, J.C.; Zhang, J.; Letaief, K.B. Backhaul-aware caching placement for wireless networks. In Proceedings of the 2015 IEEE Global Communications Conference, San Diego, CA, USA, 6–10 December 2015; pp. 1–6.
11. Tandon, R.; Simeone, O. Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 2029–2033.
12. Tao, M.; Chen, E.; Zhou, H.; Yu, W. Content-centric sparse multicast beamforming for cache-enabled cloud RAN. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 6118–6131. [CrossRef]
13. Sengupta, A.; Tandon, R.; Simeone, O. Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs. *IEEE Trans. Inf. Theory* **2017**, *63*, 6650–6678. [CrossRef]
14. Zhao, Z.; Peng, M.; Ding, Z.; Wang, W.; Poor, H.V. Cluster content caching: An energy-efficient approach to improve quality-of-service in cloud radio access networks. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 1207–1221. [CrossRef]
15. Shanmugam, K.; Golrezaei, N.; Dimakis, A.G.; Molisch, A.F.; Caire, G. FemtoCaching: Wireless content delivery through distributed caching helpers. *IEEE Trans. Inf. Theory* **2013**, *59*, 8402–8413. [CrossRef]
16. Candes, E.J.; Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.* **2009**, *9*, 717. [CrossRef]
17. Keshavan, R.H.; Montanari, A.; Oh, S. Matrix completion from noisy entries. *J. Mach. Learn. Res.* **2009**, *11*, 2057–2078.
18. Boumal, N.; Absil, P.A. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra Appl.* **2015**, *475*, 200–239. [CrossRef]
19. Lai, M.J.; Xu, Y.; Yin, W. Improved iteratively re-weighted least squares for unconstrained smoothed \ell_q minimization. *SIAM J. Numer. Anal.* **2013**, *51*, 927–957. [CrossRef]
20. Marjanovic, G.; Solo, V. On l_q optimization and matrix completion. *IEEE Trans. Signal Process.* **2012**, *60*, 5714–5724. [CrossRef]
21. Cai, J.F.; Candès, E.J.; Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **2010**, *20*, 1956–1982. [CrossRef]
22. Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*; Springer Science and Business Media: Berlin, Germany, 2013; Volume 87.
23. Nemirovski, A. *Efficient Methods in Convex Programming*; Technion: Haifa, Israel, 2005.

24. Cen, Y.; Cen, Y.; Wang, K.; Li, J.; Chen, S.; Zhang, L.; Tao, D. Low-rank tensor estimation via generalized norm/quasi-norm difference regularization. In Proceedings of the 2018 4th International Conference on Big Data Computing and Communications (BIGCOM), Chicago, IL, USA, 7–9 August 2018; pp. 144–149.
25. Gabry, F.; Bioglio, V.; Land, I. On energy-efficient edge caching in heterogeneous networks. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 3288–3298. [CrossRef]
26. Feeney, L.M.; Nilsson, M. Investigating the energy consumption of a wireless network interface in an ad hoc networking environment. In Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2001), Anchorage, AK, USA, 22–26 April 2001; Volume 3, pp. 1548–1557.
27. Xu, J.; Ota, K.; Dong, M. Saving energy on the edge: In-memory caching for multi-tier heterogeneous networks. *IEEE Commun. Mag.* **2018**, *56*, 102–107. [CrossRef]
28. Hsieh, C.J.; Chiang, K.Y.; Dhillon, I.S. Low rank modeling of signed networks. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 507–515.
29. Salakhutdinov, R.; Srebro, N. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2, Vancouver, BC, Canada, 6–9 December 2010; pp. 2056–2064.
30. Hu, Y.; Zhang, D.; Liu, J.; Ye, J.; He, X. Accelerated singular value thresholding for matrix completion. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 298–306.
31. Dunlavy, D.M.; Kolda, T.G.; Acar, E. Temporal link prediction using matrix and tensor factorizations. *ACM Trans. Knowl. Discov. Data (TKDD)* **2011**, *5*, 10. [CrossRef]
32. Hiriart-Urruty, J.B.; Lemaréchal, C. *Convex Analysis and Minimization Algorithms I: Fundamentals*; Springer Science and Business Media: Berlin, Germany, 2013; Volume 305.