Optimizing the downlink of multi-cell multiuser multiple input multiple output (MU-MIMO) networks has received substantial attention; however, the schemes in the literature consider centralized solutions requiring significant overhead in information exchange (e.g., global channel state information or CSI) and computation load (the need to solve a single large problem). This paper presents a decentralized weighted sum-rate (WSR) maximization algorithm for the multiuser downlink, accounting for beamforming, scheduling, and power allocation. We show that the signal-to-leakage-plus-noise ratio (SLNR) used in previous work suffers from significant drawbacks that limit its potential use in WSR maximization. We address this by proposing a new performance measure, the signal-to-leakage-plus-interference-plus-noise ratio (SLINR), which incorporates intra-cell interference and inter-cell leakage. The SLINR exploits the benefits of the SLNR approach, but by explicitly including interference, avoids many of its flaws. We derive an iterative and decentralized resource allocation approach under imperfect CSI, and our simulation results show that, despite BSs using only local information, the proposed algorithm comes within 3.8% of the throughput achieved by centralized schemes.

INDEX TERMS Beamforming, inter-cell interference, leakage, MIMO, SLINR.

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

With the advent of 5G wireless networks, service providers are facing increasing pressure to deliver higher data rates to more users. One of the main factors limiting the achievable rates is interference, from both intra-cell and inter-cell transmissions, which is characterized by the signal-to-interference-plus-noise ratio (SINR). Since each active link causes interference to other links, it is essential for base stations (BSs) to consider the impacts their transmissions have on the performance of the network as a whole. In this regard, a key technique proposed for multi-cell multuser multiple input multiple output (MU-MIMO) wireless networks is beamforming in the downlink [1], [2], [3], [4], which coordinates transmissions between BSs to maximize a chosen function of the SINR at scheduled users, e.g., our choice of maximizing the weighted sum-rate (WSR).

A significant additional challenge faced with beamforming is that it adds complexity to the problem of resource allocation, e.g., how should the network decide which users to serve at each point in time, and with what beamforming patterns? In [4], the authors propose an SINR-based resource allocation algorithm using fractional programming (FP) and the Hungarian algorithm to maximize the network WSR on the downlink. Nevertheless, SINR-based optimization approaches are inherently difficult to solve due to their highly-coupled nature; for each user in the network, the interference effects of each BS’s transmissions must be accounted for. This requires that cooperating BSs have global knowledge of the channel state information (CSI) between every antenna and every user in the entire network, i.e., both intra-cell and inter-cell CSI. We note that the acquisition of intra-cell CSI is a significant problem on its own, especially with the advent of massive MIMO networks [5]; requiring inter-cell CSI as well seems highly impractical.

Given the challenges with acquiring CSI, SINR-based solutions are likely unrealistic in practice, but provide a
B. LITERATURE REVIEW

1) SINR-BASED OPTIMIZATION

As mentioned, direct optimization of rates necessitates SINR-based optimization. Since the SINR is a ratio, optimizing network utility functions, such as the WSR, is non-convex and NP-hard [8]; we can, at best, find local optima. The quadratic transform, an FP technique, is proposed in [9] to decouple the ratio terms. This transforms the problem into a form amenable to iterative optimization.

In [4], a centralized SINR-based resource allocation algorithm is proposed to optimize the network WSR on the downlink using FP based on the quadratic transform. It also makes use of the fact that, for fixed beamforming vectors throughout the network, a BS can optimally schedule users locally without changing the inter-cell interference pattern. This optimal scheduling is accomplished using the Hungarian algorithm. This algorithm iterates between designing optimal beamforming vectors given the current set of scheduled users, then refining the optimal set of users given the current beamforming vectors; this procedure converges to a stationary point.

2) SLNR-BASED OPTIMIZATION

SLNR-based approaches for resource allocation have garnered some interest [10], [11], [12], [13], [14], [15]. The fundamental benefit is that optimizing an SLNR-based function requires only local CSI. SLNR-based techniques have, therefore, been the dominant approach in the literature when developing distributed solutions using only local information. It is well-known that a closed-form solution exists for the SLNR-optimal beamformer—a significant simplification from the SINR case [6].

The work in [10] addresses SLNR-based beamforming under per-antenna power constraints and devises an approach that is robust to imperfect CSI. SLNR is used in [11] to decouple a two-stage beamforming algorithm across BSs; similarly, [12] employs SLNR-based beamforming to decouple an SINR-based problem for connected vehicles into independent subproblems. In [13], an explicit expression for the SLNR-optimal beamformer is derived for the case of single-antenna receivers. It is shown that under an equal power allocation, this beamformer has performance identical to the minimum mean square error (MMSE) beamformer. The authors note that at high SNR, the SLNR converges to a scaled $F$ random variable which is independent of the SNR.

It is worth noting that power allocation is difficult to optimize under SLNR. This is because the power allocated to a beam scales both the signal and the leakage equally, thereby attenuating the effect of power control. Optimizing SLNR also complicates scheduling, which is highly coupled even within a single cell. This is because leakage captures how a beam impacts all the users being served; when this set of users changes, so must the leakage for all the beams (this is in direct contrast to the SINR-based case where the Hungarian algorithm is applicable). Thus, leakage is a user-centric measure that depends on user scheduling decisions, whereas interference is a user-centric measure that depends on beam design.

A one-pass scheduling algorithm is proposed in [14], which computes a lower bound on each user’s SLNR, then chooses the subset of users with the greatest lower bounds. This lower bound loosens when the number of users is much greater than the number of antennas, and so they also propose a successive scheduling algorithm, which incrementally schedules the best user given previous scheduling decisions. In [15], a mixed approach is proposed with one half of the users scheduled using the max-lower bound criterion and the other half using successive scheduling. The authors also propose power allocation based on the SLNR computed for each user.

3) OPTIMIZATION WITH INTERFERENCE AND LEAKAGE

In [16], the authors propose downlink beamforming for a single-cell network based on the intra-cell leakage and interference using a performance measure they call SLINR. In [17], the author considers 3 BSs and 3 paired users at the cell edge. The author develops an SLINR-based iterative beamforming scheme requiring CSI exchange across base stations, but not joint computation. Here, the SLINR is a function of the inter-cell leakage and interference, which is different than the definition used in [16]. In [18], the authors propose a downlink precoding scheme based on a performance measured called SILNR, which only requires local CSI. The SILNR incorporates intra-cell interference and the geometric mean of the inter-cell leakage terms. None of [16], [17], [18] incorporate user scheduling.

C. CONTRIBUTIONS

In this paper, we develop a valuable performance measure for guiding decentralized resource allocation. We too call this performance measure “SLINR”, but it differs significantly from the previous work in [16], [17], and [18]. Unlike in [16], our paper studies a multi-cell network, and our SLINR comprises intra-cell interference and inter-cell leakage. Unlike [17], our paper studies a general multi-cell network model with an arbitrary number of BSs and users, and our algorithm is completely decentralized and does not...
require CSI exchange. Unlike [18], our SLINR depends on the sum of the inter-cell leakage terms rather than their geometric mean.

Furthermore, we develop the motivation for our SLINR from first principles and derive an effective resource allocation algorithm which, unlike the previous work, includes scheduling and per-user weights. We optimize the WSR, and include a comprehensive performance evaluation for different choices of weights. Our results compare the proposed algorithm to other decentralized ones and the state-of-the-art in centralized resource allocation (a short version of this paper has been accepted at a conference [19]). In summary, our major contributions are:

- We motivate replacing the SINR (or SLNR) with a general, decentralized version of the SLINR, which uses inter-cell leakage and intra-cell interference. In our formulation, the SLINR can be computed without information exchange across BSs (unlike [17]).
- We develop an iterative and decentralized SLINR-based resource allocation algorithm for the multi-cell multi-user environment at hand, which obtains a stationary point. The development of the SLINR and its optimization under imperfect CSI is our main contribution.
- We show that the performance of our SLINR-based algorithm comes within 3\% of that achieved by the centralized resource allocation algorithm which, unlike the previous work, includes simulation, the SLINR can be computed without information exchange across BSs (unlike [17]).

\[ r_{kb} = \mathbf{h}_{kb,b}^H \mathbf{v}_{kb} s_{kb} \eta_{kb} + \eta_{kb}, \]

where \( \eta_{kb} \sim \mathcal{CN}(0, \sigma^2_{kb}) \) denotes additive white Gaussian noise. The achievable rate with \( W_B \) Hz of channel bandwidth is defined (in appropriate units) as

\[ R_{kb} = W_B \log(1 + \gamma_{kb}), \]

where \( \gamma_{kb} \) is the SINR, written as

\[ \gamma_{kb} = \frac{u_{kb}|\mathbf{h}_{kb,b}^H \mathbf{v}_{kb}|^2}{\sigma^2_{kb} + \sum_{b'=1}^{[B]} \sum_{k'=1}^{K_{b'}} |\mathbf{h}_{b',b'}^H \mathbf{v}_{b'}|^2}. \]

Finally, we control the relative priority of each user via the rate weighting factor, \( w_{kb} \).

B. CHANNEL ESTIMATION

We consider a time-division duplex system, where channels are estimated on the uplink using a pilot-based scheme. We assume channel reciprocity holds. Users within a given cell are assigned pilot sequences as having unit norm, i.e., \( \Phi_{kb} \Phi_{kb}^H = 1 \). Note that the vectors \( \Phi_{kb} \) are known by all BSs; we also assume each BS has knowledge of every user’s large-scale fading coefficient and transmit power (denoted by \( p_u \)), similarly to [21].

The training signal received by the \( b \)th BS is denoted by \( \mathbf{Y}_b \in \mathbb{C}^{M_t \times N_p} \) and, assuming all pilot signals are received coherently, is written as

\[ \mathbf{Y}_b = \sum_{b'=1}^{[B]} \sum_{k'=1}^{K_{b'}} \sqrt{p_u} \mathbf{h}_{k',b'} \Phi_{k',b'} + \mathbf{Z}_b, \]

1The specifics of how this is done are beyond the scope of our work, but some techniques might include averaging channel gains over many realizations [20], or using knowledge that might be available about users’ positions (provided there is no shadowing).
where $Z_b$ is measurement noise, with entries distributed according to $CN(0, \sigma^2_z)$.

To estimate the channel from the $b^{th}$ BS to the $k^{th}$ user of the $b^{th}$ BS, $b$ first preprocesses (4) to eliminate the contributions from users with pilot sequences orthogonal to $\Phi_{k,b}$. This is done by projecting (4) onto the user’s pilot signal, yielding $\hat{y}_{k,b,b} = \frac{1}{\sqrt{p_u}} Y_b \Phi_{k,b}^H$. BS $b$ then uses the linear minimum mean square error (LMMSE) estimator to compute the estimate $\hat{h}_{k,b,b}$:

$$
\hat{h}_{k,b,b} = R_{\tilde{y}_{k,b,b},y_{k,b,b}}^{-1} R_{\tilde{y}_{k,b,b},y_{k,b,b}} \hat{y}_{k,b,b}.
$$

(5)

In (5), $R_{\tilde{y}_{k,b,b},y_{k,b,b}}$ is the cross-covariance matrix between the measurement $\tilde{y}_{k,b,b}$ and the channel, and $R_{\tilde{y}_{k,b,b},y_{k,b,b}}$ is the covariance matrix of $\hat{y}_{k,b,b}$. Assuming spatially uncorrelated channels, we obtain the cross-covariance matrix as $R_{\tilde{y}_{k,b,b},y_{k,b,b}} = D_{k,b,b} \beta_{k,b,b} I_M$, and the covariance matrix as

$$
R_{\tilde{y}_{k,b,b},y_{k,b,b}} = \sum_{(k',b') \in U_{k,b}} |\Phi_{k',b'} \Phi_{k,b}^H|^2 D_{k',b'} b_{b'} + \sigma^2_Z I_M,
$$

(6)

where $U_{k,b}$ denotes the set of users whose pilots are non-orthogonal to $\Phi_{k,b}$ (including user $b$ at BS $b$). Since the covariance matrices are diagonals, simplifying (5) yields the estimator

$$
\hat{h}_{k,b,b} = \left( \frac{1}{1 + 1/\hat{\xi}_{k,b,b}} \right) \frac{1}{\sqrt{p_u}} Y_b \Phi_{k,b}^H.
$$

(7)

In (7), $\hat{\xi}_{k,b,b}$ is an SINR-like term for the channel estimation procedure, written as

$$
\hat{\xi}_{k,b,b} = p_u (\beta_{k,b,b} / \left( p_u \sum_{(k',b') \in U_{k,b}} |\Phi_{k',b'} \Phi_{k,b}^H|^2 D_{k',b'} b_{b'} + \sigma^2_Z \right)).
$$

(8)

From the theory of LMMSE estimation, the estimate (7) is distributed as $CN(0, A_{k,b,b})$, where

$$
A_{k,b,b} = R_{\hat{h}_{k,b,b},y_{k,b,b}}^{-1} R_{\hat{h}_{k,b,b},y_{k,b,b}} = \beta^2_{k,b,b} \left[ \sum_{(k',b') \in U_{k,b}} |\Phi_{k',b'} \Phi_{k,b}^H|^2 D_{k',b'} b_{b'} + \sigma^2_Z / p_u \right]^{-1}.
$$

(9)

and the estimation error $e_{k,b,b} = h_{k,b,b} - \hat{h}_{k,b,b}$ is distributed according to $CN(0, \Theta_{k,b,b})$, where

$$
\Theta_{k,b,b} = R_{h_{k,b,b},y_{k,b,b}}^{-1} R_{h_{k,b,b},y_{k,b,b}} = \left( \frac{1}{1 + \hat{\xi}_{k,b,b}} \right) \beta^2_{k,b,b} I_M.
$$

(10)

C. PROBLEM FORMULATION

We wish to maximize resource allocation performance throughout the network, as measured by the weighted sum-rate, in a decentralized manner (i.e., without any real-time information exchange across BSs). In this section, we state the centralized optimization problem that the network is faced with, and in the following sections, we propose and analyze a decentralized approach for solving it effectively.

The network-wide optimization is over the scheduling variables and beamformers, collected in matrices $U$ and $V$, respectively. Accordingly, the weighted sum-rate maximization problem for a single time slot can be expressed in terms of the SINR (3) as shown in (11).

The first set of constraints in (11b) enforce binary scheduling decisions. The set of constraints in (11c) ensure that the number of users scheduled at a BS does not exceed the number of antennas. Finally, the set of constraints in (11d) limit the power available for transmission at each BS.

$$
\begin{align*}
&\text{maximize} & & \sum_{b=1}^{B} \sum_{k=1}^{K_b} w_{kb} \log(1 + \gamma_{kb}) & \quad \text{(11a)} \\
&\text{subject to} & & u_{kb} \in \{0, 1\}, & b = 1, \ldots, |B| & \quad \text{(11b)} \\
& & & u_{kb} \leq M_t, & b = 1, \ldots, |B| & \quad \text{(11c)} \\
& & & \sum_{k=1}^{K_b} |u_{kb}|^2 \leq P_T, & b = 1, \ldots, |B|. & \quad \text{(11d)}
\end{align*}
$$

III. DOWNLINK OPTIMIZATION BASED ON SLINR

A. PROBLEM ANALYSIS

We begin by noting that the optimization problem (11) is mixed-integer, non-convex, and NP-hard [8]. Since optimality is infeasible, our goal is to design an algorithm with desirable properties that solves the problem effectively. Next, we emphasize that we wish to find a fully decentralized solution to this problem, i.e., each BS can only know its local scheduling decisions, beamforming vectors and CSI. This requirement alters the problem structure, since it constrains our solutions to those which can be obtained by each BS performing an independent optimization using only local information. Accordingly, our first step for finding a decentralized algorithm is to state the single-cell optimization problem, over local design variables, for the $b^{th}$ BS:

$$
\begin{align*}
&\text{maximize} & & \sum_{k=1}^{K_b} w_{kb} \log(1 + \gamma_{kb}) & \quad \text{(12)} \\
&\text{subject to} & & (11b), (11c), (11d). &
\end{align*}
$$

In (12), $u_{kb} \in \{0, 1\}^{K_b \times 1}$ and $V_b \in C^{K_b \times M_t}$ denote the scheduling variables and beamforming variables designed by the $b^{th}$ BS, respectively.

An insurmountable problem arises with (12): it is impossible for a BS to evaluate $\gamma_{kb}$ via (3), since the inter-cell interference experienced by each user depends on other BSs’ beamforming decisions and CSI (neither can be known without information exchange). Thus, neither direct rate-based optimization nor explicit interference coordination are possible in this setting. Additionally, we wish to study the practical case of imperfect CSI outlined in Section II-B, so the optimization strategy should account for channel estimation errors.
As discussed in Section III-A, optimizing rates directly using pseudorates measures, which we call as such, we need to choose among alternative performance information exchange in order to calculate inter-cell interference. Possible in the decentralized setting, because BSs require information (i.e., using an SINR maximization methodology) is not possible in the decentralized setting, because BSs require information exchange in order to calculate inter-cell interference. As such, we need to choose among alternative performance measures, which we call pseudorates. Note that in the following, we do not claim that these pseudorates are achievable, nor do we use them to evaluate performance; rather, the pseudorates are only used in our optimization procedure, and we always evaluate system performance with respect to the real (SINR-based) achievable rates.

As a starting point, consider an SINR-based optimization approach that only uses the intra-cell interference (since each BS can evaluate this using solely local information). This is a reasonable strategy that can serve as a performance baseline, but we will be interested in trying to improve upon it since it is inherently selfish. By selfish, we mean that it encourages each BS to maximize its own performance with no consideration for how its transmissions might reduce the performance of other BSs. Since we are interested in maximizing the performance of the network as a whole, it would be convenient if we had a way to balance the selfish tendency of each BS with a tendency for neighbourly behaviour.

This discussion brings to mind the leakage-based approaches mentioned in Section I. The leakage philosophy dictates that a BS accounts for other cells by considering the impacts its transmissions have on their users. This is in contrast to the interference philosophy, which dictates that a BS accounts for other cells by considering the impacts their transmissions have on its own users. The leakage and interference philosophies are shown in Figs. 1a and 1b, respectively, visualizing our earlier statement that leakage is beam-centric while interference is user-centric. Unlike interference, leakage can be computed without exchanging CSI and beamformers. The power that beam \( v_{kb} \) leaks to the users in neighbouring cells, i.e., its inter-cell leakage, is [6]

\[
L_{\text{inter}}^{\text{inter}} = \sum_{b' = 1}^{|B|} \sum_{k' = 1}^{K_{b'}} u_{k'b'} |h_{k'b',b}v_{kb}|^2, \tag{13}
\]

and the power it leaks to the users in its own cell, i.e., its intra-cell leakage, is

\[
L_{\text{intra}}^{\text{inter}} = \sum_{k = 1}^{K_b} u_{k'b} |h_{k'b,b}v_{kb}|^2. \tag{14}
\]

From (13), we can see that computing the inter-cell leakage requires knowledge of other BSs’ scheduling decisions. Instead of requiring perfect knowledge of these variables (i.e., information exchange), we instead adopt a flexible approach based on probabilistic knowledge, i.e., \( \Pr(u_{k'b'} = 1) \). Our evaluation of the inter-cell leakage for beam \( v_{kb} \), then becomes an estimate

\[
L_{\text{inter}}^{\text{inter}} = \sum_{b' = 1}^{|B|} \sum_{k' = 1}^{K_{b'}} \Pr(u_{k'b'} = 1) |h_{k'b',b}v_{kb}|^2. \tag{15}
\]

Importantly, \( \Pr(u_{k'b'} = 1) \) could be set according to various policies, e.g., it may be fixed, or time-varying (since in general, not all users are equiprobable). Note that time-varying policies would not require scheduling information to be exchanged between BSs during each allocation, as long as the probabilistic knowledge of the scheduling variables in neighbouring cells changes relatively slowly. In other words,
real-time knowledge of the scheduling variables is not necessary as long as the impact this has on the inter-cell leakage computation is small. Regardless of how the probabilities are assigned, by relaxing our knowledge of the scheduling variables in neighbouring cells to probabilities, we obtain a decentralized way of evaluating a beam’s inter-cell leakage via (15); thus, BSs are fully decoupled. A user’s SLNR is defined as the ratio of its signal power to its total leakage and noise power

$$\text{SLNR}_{kb} = \frac{u_{kb}|h_{kb,b}^H v_{kb}|^2}{\sigma^2_{kb} + L_{\text{inter}}^{\text{intra}} + L_{\text{inter}}^b}. \quad (16)$$

SLNR-based optimization is attractive in certain respects: it only requires local CSI and design variables, optimal beamformers can be determined in closed form, and it is inherently “neighbourly” since it penalizes inter-cell transmissions. However, this approach has important drawbacks, as mentioned in the introduction. The first key drawback comes from the observation that scheduling decisions are highly coupled under the SLNR, since the decision of whether to schedule a user affects the leakage of all beams. The second key drawback is that, since the power of a beam appears in both the numerator and the leakage terms in the denominator of the SLNR, it struggles to guide power allocation effectively.

Neither of these drawbacks are present for the SINR. Given fixed beamforming patterns, local user scheduling can be optimally solved using the Hungarian algorithm when using the SINR (and naturally, local decision making is conducive to our goal of developing a decentralized resource allocator). Even without knowing the beamformers used by other BSs, we can still use the Hungarian algorithm to make local scheduling decisions via the intra-cell interference. Furthermore, since the SINR has different beams in the numerator and denominator, the power cancellation problem experienced by the SLNR does not occur. As such, an SINR-based measure which uses only the intra-cell interference does not suffer the drawbacks of the SLNR, but at the same time, it does not experience the benefits of the SLNR’s neighbourly intentions.

In summary, this discussion has shown that both leakage and interference bring something to the table for this decentralized environment: using the intra-cell interference allows BSs to make local scheduling decisions optimally given fixed beams, and using the inter-cell leakage allows BSs to account for their impact on other cells without exchanging CSI and design variables. However, both leakage and interference leave something to be desired when considered on their own. As such, we now consider a performance measure comprising leakage and interference, in hopes of retaining the benefits of each while diminishing the drawbacks.

We write the hybrid leakage- and interference-based measure as the SLINR, given by

$$\gamma_{kb}^{\text{SLINR}} = \frac{u_{kb}|h_{kb,b}^H v_{kb}|^2}{\sigma^2_{kb} + L_{\text{inter}}^{\text{intra}} + L_{\text{inter}}^b}. \quad (17)$$

where $I_{kb}^{\text{intra}}$ is the intra-cell interference experienced by the $k^{th}$ user of BS $b$, given as

$$I_{kb}^{\text{intra}} = \sum_{k' \neq k} u_{k'b}|h_{k'b,b}^H v_{k'b}|^2. \quad (18)$$

We emphasize that (17) can be computed without exchanging information across BSs.

To model imperfect CSI, we replace the channels with their estimates, and treat the estimation errors as additional noise like in [20] and [22]. This will persist throughout the rest of the paper. The SLINR in (17) then becomes

$$\tilde{\gamma}_{kb}^{\text{SLINR}} = \frac{u_{kb}|h_{kb,b}^H \tilde{v}_{kb}|^2}{\sigma^2_{kb} + L_{\text{inter}}^{\text{intra}} + L_{\text{inter}}^b}. \quad (19)$$

In (19), $L_{\text{inter}}^{\text{intra}}$ and $L_{\text{inter}}^b$ are identical to (15) and (18), respectively, except that the channels are replaced with their estimates. The estimation error for the inter-cell leakage, $\tilde{L}_{\text{inter}}$, is given by

$$\tilde{L}_{\text{inter}}^b = \sum_{b' = 1}^{B} \sum_{k' = 1}^{K_{b'}} \Pr \left( u_{b'b'} = 1 \right) v_{b'b'}^H \tilde{\Theta}_{b'b,b} v_{b'b}, \quad (20)$$

and the estimation error for the intra-cell interference, $\tilde{I}_{\text{intra}}^b$, is given by

$$\tilde{I}_{\text{intra}}^b = \sum_{k' \neq k}^{K_b} u_{k'b} \tilde{v}_{k'b}^H \tilde{\Theta}_{kb,b} v_{k'b}. \quad (21)$$

The signal model employed for SLINR-based processing is visualized in Fig. 2.

![FIGURE 2. Block diagram of the SLINR-based signal model.](image_url)

We now introduce the SLINR-based pseudorates as $\tilde{R}_{kb} = W_B \log \left( 1 + \tilde{\gamma}_{kb}^{\text{SLINR}} \right)$. We optimize these as a proxy for the rate maximization problem, with the understanding that system performance must ultimately be evaluated via the real (SINR-based) rates using (2). Accordingly, the weighted sum-pseudorate maximization problem can be written like the original decentralized formulation (12), except that the rates in the objective are replaced by pseudorates, i.e., for BS $b$:

$$\max_{u_k, v_k} \sum_{k=1}^{K_b} w_k \log \left( 1 + \tilde{\gamma}_{kb}^{\text{SLINR}} \right)$$

subject to (11b), (11c), (11d). \quad (22)
C. RESOURCE ALLOCATION

We begin the optimization for the $\phi^{th}$ BS by noting that (22) is a sum-of-functions-of-ratios maximization problem. To solve it, we first introduce auxiliary variables, $\phi_{kb}$, to remove the fractional SLINR terms from the logarithms in the objective, yielding an equivalent optimization problem:

$$
\text{maximize} \quad \sum_{k=1}^{K_b} w_{kb} \log (1 + \phi_{kb}) \quad (23a)
$$

subject to \quad (11b), (11c), (11d) \quad (23b)

$$
\phi_{kb} = \hat{\gamma}_L^{1/k}, \quad k = 1, \ldots, K_b. \quad (23c)
$$

Next, we introduce Lagrange multipliers, $\lambda_{kb}$, for the equality constraints, and form the Lagrangian

$$
\mathcal{L} (z_b, \lambda_b) = \sum_{k=1}^{K_b} w_{kb} \log (1 + \phi_{kb}) - \sum_{k=1}^{K_b} \lambda_{kb} (\phi_{kb} - \hat{\gamma}_L^{1/k}),
$$

where $z_b = (u_b, V_b, \phi_b)$ collects the primal variables to save space. Differentiating with respect to $\phi_{kb}$, we find the optimal Lagrange multipliers to be

$$
\lambda_{kb}^{*} = \frac{w_{kb}}{1 + \hat{\gamma}_L^{1/k}} = \frac{w_{kb} \text{den} \{\hat{\gamma}_L^{1/k}\}}{\text{den} \{\hat{\gamma}_L^{1/k}\} + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2},
$$

where $\text{den} \{\hat{\gamma}_L^{1/k}\}$ denotes the denominator of (19). Now, substituting (25) into (24) yields the revised objective function

$$
f_r (z_b) = \sum_{k=1}^{K_b} w_{kb} (\log (1 + \phi_{kb}) - \phi_{kb})
$$

$$
+ \sum_{k=1}^{K_b} w_{kb} u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2 (1 + \phi_{kb})
$$

$$
+ \sum_{k=1}^{K_b} \text{den} \{\hat{\gamma}_L^{1/k}\} + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2.
$$

For all tuples in the constraint set, the numerators of the fractional terms are non-negative and the denominators are strictly positive. Therefore, we can use the quadratic transform [9] to decouple the numerator and denominator of each fraction, yielding a new objective function

$$
f_q (z_b, y_b) = \sum_{k=1}^{K_b} \left[ w_{kb} (\log (1 + \phi_{kb}) - \phi_{kb})
$$

$$
+ 2 \text{Re} \left( y_{kb}^{*} \sqrt{w_{kb} u_{kb} (1 + \phi_{kb})} h_{kb,b}^{H} h_{kb,b}^{H} v_{kb}\right) - |y_{kb}|^2 \left( \text{den} \{\hat{\gamma}_L^{1/k}\} + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2 \right) \right],
$$

and a new optimization problem:

$$
\text{maximize} \quad f_q (z_b, y_b)
$$

subject to \quad $y_{kb} \in \mathbb{C}$, (11b), (11c), (11d) \quad (28)

The problem in (28) is equivalent to (23), in the sense that the optimal values of the primal variables and objective functions are identical; this follows directly from properties C2 and C3 of the quadratic transform, detailed in Section II-B of [9].

We now show how (28) can be solved using cyclic maximization with closed-form updates. Note that cyclic maximization falls under the broader class of minorization-maximization algorithms, which are guaranteed to converge to a stationary point under natural assumptions [23]. Keeping $u_b$, $V_b$, and $y_b$ fixed, we set $\frac{\partial f_q(z_b, y_b)}{\partial \phi_{kb}} = 0$ to find the optimal value of $\phi_{kb}$ as

$$
\phi_{kb}^{*} = \frac{\gamma_L}{\hat{\gamma}_L^{1/k}}.
$$

Similarly, we set $\frac{\partial f_q(z_b, y_b)}{\partial y_{kb}} = 0$ to find the optimal value of $y_{kb}$ as

$$
y_{kb}^{*} = \sqrt{\frac{w_{kb} u_{kb} (1 + \phi_{kb}) v_{kb}^{H} h_{kb,b}^{H} h_{kb,b}^{H} v_{kb}}{\text{den} \{\hat{\gamma}_L^{1/k}\} + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2}}.
$$

Beamformer and Power Constraint: To find the optimal $v_b$, we first introduce a dual variable $\mu_b \geq 0$ to capture the per-base-station power constraint, and form the Lagrangian

$$
\mathcal{L} (z_b, y_b, \mu_b) = f_q (z_b, y_b) - \mu_b \left( \sum_{k=1}^{K_b} \|v_{kb}\|^2 - P_T \right).
$$

By setting the $k^{th}$ element of the gradient $\nabla v_b \mathcal{L} (z_b, y_b, \mu_b)$ equal to zero and solving for $v_{kb}$, it can be shown (see Appendix A) that the optimal beamforming vector is

$$
v_{kb}^{*} = \sqrt{\frac{w_{kb} u_{kb} (1 + \phi_{kb}) X_{kb}^{-1}(\mu_b) h_{kb,b}^{H}}{\text{den} \{\hat{\gamma}_L^{1/k}\} + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2}}.
$$

where the intermediate matrix $X_{kb}(\mu_b)$ is given as

$$
X_{kb}(\mu_b) = \mu_b M_b
$$

$$
+ u_{kb} \sum_{k=1}^{K_b} |y_{kb}|^2 \left( h_{k,b,b}^{H} h_{k,b,b}^{H} + \Theta_{k,b,b} \right)
$$

$$
+ \sum_{b'=1}^{B} \sum_{k'=1}^{K_{b'}} \text{Pr} (u_{kb'} = 1)
$$

$$
\times \left( h_{k',b'}^{H} h_{k',b'}^{H} + \Theta_{k',b'} \right).
$$

In (33), $\mu_b$ can be chosen using a bisection search in order to satisfy the sum power constraint. Due to the constraint on the maximum number of users that can be scheduled as given by (11c), the beam design step yields $N_0 \leq M_b$ nonzero beams, denoted for the $b^{th}$ BS as $v_{kb}$, $n = 1, \ldots, N_b$.

Scheduling Variables: The last step in the iteration is to update the local schedule, $u_b$. Crucially, the inter-cell leakage in (15) is independent of local scheduling decisions for a given beam, and for a fixed set of beams, the intra-cell interference pattern is independent of local scheduling decisions too. Thus, when the $k^{th}$ user of the $b^{th}$ BS is served on the $b^{th}$ channel, we can write the intra-cell interference it experiences as the total interference minus the new signal term

$$
|\hat{\gamma}_L^{1/k}|^2 + u_{kb} \|h_{kb,b}^{H} v_{kb}\|^2 - |h_{kb,b}^{H} v_{kb}|^2.
$$

(34)
The new estimation error follows similarly, written as
\[ (\tilde{I}_{kb}^{\text{intra}} + u_{kb}v_{kb}^H\Theta_{kb}v_{kb}) - w_{nb}^H\hat{\Theta}_{kb,b}v_{nb} + \tilde{L}_{kb}^{\text{inter}}(v_{nb}), \]  
(35)
and we denote the new inter-cell leakage as \( \tilde{L}_{kb}^{\text{inter}}(v_{nb}) \). Here, \( \tilde{L}_{kb}^{\text{inter}}(v_{nb}) \) and \( \tilde{L}_{kb}^{\text{inter}}(v_{nb}) \) are identical to (15) and (20), respectively, except that \( v_{nb} \) is used in the calculation instead of \( v_{kb} \). With the previous equations in mind, we can see that the SLINR for this user can be written as
\[ \gamma_{kb,n}^{(1)}(v_{nb}) = \frac{|\hat{h}_{kb,b}^Hv_{nb}|^2}{c_{kb} + \hat{\xi}_{kb}(v_{nb})}, \]  
(36)
where the fixed quantity \( c_{kb} = \hat{r}_{kb}^H + \tilde{I}_{kb}^{\text{intra}} + \tilde{I}_{kb}^{\text{intra}} + u_{kb}(\hat{h}_{kb,b}^Hv_{kb})^2 + v_{kb}^H\hat{\Theta}_{kb,b}v_{kb} \) is independent of the beam choice, and \( \hat{\xi}_{kb}(v_{nb}) = L_{kb}^{\text{inter}}(v_{nb}) + L_{kb}^{\text{inter}}(v_{nb}) - |\hat{h}_{kb,b}^Hv_{nb}|^2 - \hat{\Theta}_{kb,b}v_{nb} \) is dependent on the beam choice.

The weighted pseudorate when this user is served on the \( n \)th non-zero beam is then given by (37).
\[ \tilde{R}_{kb,n} = w_{kb} \log \left(1 + \frac{|\hat{h}_{kb,b}^Hv_{nb}|^2}{c_{kb} + \hat{\xi}_{kb}(v_{nb})}\right). \]  
(37)
Updating the local scheduling variables, \( u_{kb} \), is then a matter of determining the optimal user-to-beam matchings, given the beams designed via (32). For the \( b \)th BS, this can be formulated as the linear sum assignment problem in (38).
\[
\max_{\text{X}} \sum_{k=1}^{K_b} \sum_{n=1}^{N_b} R_{kb,n} x_{kb,n} \quad \text{(38a)}
\]
subject to
\[
\sum_{k=1}^{K_b} x_{kb,n} = 1, \quad n = 1, \ldots, N_b \quad \text{(38b)}
\]
\[
\sum_{k=1}^{K_b} x_{kb,n} = 1, \quad k = 1, \ldots, K_b \quad \text{(38c)}
\]
\[
x_{kb,n} \in \{0, 1\}, \quad n = 1, \ldots, N_b, \quad k = 1, \ldots, K_b. \quad \text{(38d)}
\]
In (38), the binary variables \( x_{kb,n} \) indicate whether the \( \ell \)th user of the \( b \)th BS has been scheduled on the \( n \)th beam (in such a case, we set \( u_{kb} = 1 \) and \( v_{kb} = v_{nb} \) for the next iteration). Importantly, (38) can be solved optimally in \( O(K_bN_b^2) \) steps using well-established techniques such as the Hungarian algorithm [24].

We have now derived updates for all variables. Accordingly, the SLINR-based downlink resource allocation algorithm for the \( b \)th BS is stated in Algorithm 1. We emphasize that this algorithm is executed at each BS independently, i.e., the overall resource allocation is distributed. We also emphasize that it is impossible for any of steps 4 to 7 to decrease the value of the objective function (28), since they are each locally optimal with respect to the function. Accordingly, the objective function must be monotonically non-decreasing throughout the optimization procedure. This, plus the fact that the objective is bounded above (since the weighted sum-rate must be finite), ensures that Algorithm 1 converges to a stationary point.

**Algorithm 1 Proposed Resource Allocation Algorithm**

1. Initialize \( u_b, V_b, N_{\text{iterations}} \)
2. Set \( i = 1 \)
3. repeat
4. Update \( \phi_b \) using (29)
5. Update \( y_b \) using (30)
6. Update \( V_b \) using (32)
7. Update \( u_b \) and \( V_b \) jointly by solving (38)
8. Increment \( i \)
9. until convergence or \( i = N_{\text{iterations}} \)

We begin the analysis with Step 4, where \( \phi_{kb}^{(i)} \) is updated using (29). This is equivalent to evaluating (19), which is dominated by the matrix-vector multiplication in \( \tilde{L}_{kb}^{\text{inter}}(v_{nb}) \) and \( \tilde{I}_{kb}^{\text{intra}} \), yielding an overall complexity for this step of \( O(|B|K_M^2 + K_bM_b^2) \). In Step 5, the update for \( y_b^{(i)} \) in (30) can reuse the denominator \( \gamma_{kb,n}^{(i)} \) term computed as part of Step 4; with this caching, the overall complexity of Step 5 just comes from the two inner products in (30), and is therefore \( O(M_b^2) \). Step 6 consists of several sub-steps, the first of which is computing the outer products and matrix additions comprising \( X_{kb}(\mu_b) \), which is \( O(|B|K_M^2 + K_bM_b^2) \). Next, \( X_{kb}(\mu_b) \) needs to be inverted, which is \( O(M_b^3) \), but we treat this search as having constant complexity since it scales with search parameters that are independent of the scenario under study. These sub-steps need to be performed for up to \( M_b \) non-zero beams, yielding an overall complexity for Step 6 of \( O(|B|K_M^2 + K_bM_b^2 + M_b^3) \). Step 7 also consists of sub-steps, the first of which is setting up the \( \tilde{R}_{kb,n} \) matrix and the second of which is solving (38). It can be shown that setting up \( \tilde{R}_{kb,n} \) is dominated by \( K_bM_b \) computations of \( \tilde{L}_{kb}^{\text{inter}}(v_{nb}) \), which is \( O(|B|K_bM_b^2) \). As mentioned previously, (38) can be solved in \( O(K_bM_b^2) \), so the overall complexity for Step 7 is \( O(|B|K_bM_b^2 + K_bM_b^2) \).

We have now derived the computational complexity of each step in Algorithm 1. Putting them together and keeping the highest-order terms yields a per-iteration complexity of \( O(|B|KK_bM_b^2 + M_b^3) \) for BS \( b \), which is dominated by setting up \( \tilde{R}_{kb,n} \) in Step 7 and inverting \( X_{kb}(\mu_b) \) in Step 6.
In summary, in this section we motivated and developed resource allocation based on a hybrid leakage- and interference-based performance measure, the SLINR, which serves as a proxy for the SINR when optimizing rates. The SLINR allows for a completely distributed solution like the well-known SLNR, but does not suffer from its drawbacks (discussed in Section III-B). Using SLINR helps deal with interference while also allowing for effective power control and scheduling. Our next section illustrates the efficacy of the proposed approach.

### IV. NUMERICAL RESULTS

In this section, the performance of the proposed algorithm (Algorithm 1) is compared to the state-of-the-art centralized SINR-based algorithm from [4], and some decentralized algorithms. Evaluations are made for max sum-rate (equal weights) and proportionally fair weighting on the basis of throughput and fairness. Unless otherwise noted, we assume perfect CSI is available.

We simulated a 7-cell environment (|B| = 7) with wraparound. For convenience, we implemented hexagonal cells with a BS at the center of each cell. Throughout this paper, a topology refers to a particular placement of users in the cells; these user placements were randomly generated according to a spatially uniform distribution. We model Rayleigh fading channels by choosing $g_{kj} \sim \mathcal{CN}(0, I_M)$ and set the large-scale path loss as $p_{kj} = (1 + d_{kj}/d_0)^{-\alpha}$, where $d_0$ and $\alpha$ are given, along with the other main system parameters, in Table 1.

#### TABLE 1. System parameters used for simulation.

| Parameter                  | Value                  |
|----------------------------|------------------------|
| Number of cells            | |B| = 7                  |
| Distance between adjacent BSs | 1000 m                |
| Number of antennas per BS  | $M_t = 8$              |
| Channel bandwidth          | $W_B = 20$ MHz         |
| Base station maximum transmit power | $P_T = 43$ dBm    |
| Noise power spectral density | $N_0 = -174$ dBm/Hz   |
| Noise figure               | $N_P = 9$ dBm          |
| Reference distance         | $d_0 = 0.3920$ m       |
| Path loss exponent         | $\alpha = 3.76$       |
| User transmit power        | $P_u = 17$ dBm         |
| Forgetting factor          | $\alpha_f = 0.05$ [4]  |

Each time a resource allocator was run, the (SINR-based) rates achieved by each user were recorded. For simulations using proportionally fair weights, the weights inputted to the algorithm for the $(n+1)^{th}$ time slot were computed according to $w_{k_ib}(n+1) = 1/\hat{R}_{k_ib}^{(n)}$, where $\hat{R}_{k_ib}^{(n)}$ denotes the user’s long-term average data rate over an exponentially-decaying window, calculated via the update equation $\hat{R}_{k_ib}^{(n)} = (1 - \alpha_f)\hat{R}_{k_ib}^{(n-1)} + \alpha_f\hat{R}_{k_ib}^{(n)}$, with $\alpha_f$ denoting the forgetting factor. Throughout this section, the proposed algorithm used $Pr(u_{k_i} = 1) = M_t/K_{b}^{(n)}$, i.e., we assumed all users in neighboring cells were equally likely to be scheduled (see (15)).

Unless otherwise specified, we initialized both the proposed algorithm and the SINR-based algorithm using the best known initialization: the maximum interference-free weighted sum-rate (MIFWSR). We computed the term $w_{k.ib} \log \left(1 + \frac{|h_{k.ib}|^2 P_T}{\sigma_{b}^{K_{b}}}ight)$ for each user, and chose the $\min (M_t, K_{b})$ users with the largest values. For both the proposed algorithm and the SINR-based algorithm, the initial guess for the beamformers (i.e., at the start of each iterative optimization process) was chosen to be matched filtering weights with equal power.

#### FIGURE 3. Sum-rate convergence of SLINR-based and SINR-based algorithms ($K_{b} = 80$).

Importantly, we plot both the true sum-rate convergence and the sum-pseudorate convergence for the SLINR-based algorithm. The local pseudorates are the information that the SLINR-based algorithm has access to, whereas the true rates are what the users experience. We see that the SLINR-based algorithm converges slower than the SINR-based algorithm, albeit only by 1 to 2 iterations. The SLINR-based algorithm also converges in a monotonically nondecreasing fashion, both in terms of the true rates and the pseudorates. The sum-pseudorate is quite close to the true sum-rate, suggesting that the SLINR is a reasonable proxy for the SINR. Finally, we note that the true sum-rate converges to a slightly higher value than the sum-pseudorate.

#### B. LEAKAGE WEIGHTING FACTOR

In this section, we consider a variation of (17):

$$\gamma_{k_ib}^{LI} = \frac{u_{k_ib} |h_{k_ib}|^2}{\sigma_{b} + \epsilon_{k_ib} \text{inter} + \epsilon_{k_ib} \text{intra}}.$$  \hspace{1cm} (39)

In (39), $\epsilon \geq 0$ is the leakage weighting factor; by varying it, we can study the role of the inter-cell leakage in detail.
When $\epsilon = 0$, the proposed algorithm is still decentralized but leakage no longer plays a role; as such, we refer to the algorithm as "SLINR (no leakage)" in this case. For the results in this section, we consider a distance of 500 m between adjacent BSs and $K_b = 40$. Results are averaged over 1000 channel realizations on a single topology.

First, we consider the max sum-rate case ($w_{kb} = 1$), which causes the algorithm to focus on the users with the largest instantaneous rates (usually near the cell center). From Fig. 4a, we can see that the throughput of the proposed algorithm is concave in $\epsilon$, with a maximum being achieved near the interval $[0.4, 0.6]$. Performance improves by about 100 Mbps when increasing $\epsilon$ from 0 to 1; although, in percentage terms, this improvement is not large, it nevertheless demonstrates that penalizing leakage can yield gains.

**FIGURE 4.** (a) Average sum-rate as $\epsilon$ is varied. (b) Dependence of rate on distance, as $\epsilon$ is varied.

Fig. 4b helps explain the trend in the sum-rate. The initial increase appears to be coming from the improvement in the average rates of the cell-center users; however, we can see that increasing $\epsilon$ worsens the average rates of the mid-cell users, and causes a negligible difference for the cell-edge users (who are rarely scheduled). As such, the improvement in the cell-center users’ rates is eventually insufficient to make up for the loss in the rates of the mid-cell users. With a uniform user distribution, there are fewer cell-center users than mid-cell users, hence the eventual decrease in sum-rate as $\epsilon$ increases.

As for why performance shifts from the mid-cell users to those near the cell center, this can be explained by the fact that as $\epsilon$ increases, each BS starts accounting for the leakage it causes to other cells. This decreases the estimated performance of the users who are far from the cell center, but has a negligible impact on those near it (since their SLINRs are large, and therefore their pseudorates are relatively insensitive to the introduction of the leakage penalty). For the max sum-rate case we are considering, this effectively de-prioritizes users who are further away from their BS. As we will see next, the story changes when a different weighting scheme is adopted.

**FIGURE 5.** (a) Average weighted sum-rate achieved as $\epsilon$ is varied. (b) 10th percentile user rates.

Next, we consider the case of proportionally fair weights, which causes the algorithm to prioritize each user in proportion to the inverse of its average rate. With this choice of weights, throughput is not our primary concern; instead we focus on fairness-related measures, such as the WSR shown in Fig. 5a. The decrease in WSR for $0 < \epsilon < 0.4$ shows that penalizing small amounts of inter-cell leakage harms the overall performance of the network. One explanation is that in this regime, we are paying for penalizing the best users through our attempt to be leakage-aware; at the same time, we are not sufficiently leakage-aware to reap the rewards. As $\epsilon$ increases from 0.4 to 1, we recover the initial performance of the network (with a slight gain).

Importantly, we note that Fig. 5b shows a 10% improvement in the 10th percentile rate within this interval, meaning that although the transition from $\epsilon = 0$ to $\epsilon = 1$ has not improved the WSR drastically, it has improved the rates of the cell-edge users to a noticeable extent. This indicates
the suppression of inter-cell leakage is paying off. Finally, we note that as $\epsilon$ increases beyond 1, the WSR increases to a maximum of 8.5% higher than at $\epsilon = 0$, and the 10th percentile rate improves by up to 17%, thus demonstrating the efficacy of SLINR.

### C. PERFORMANCE COMPARISONS ACROSS ALGORITHMS

Throughout the rest of our results, we return to the setup described in Table 1.

1) **MAX SUM-RATE PERFORMANCE**

We begin with a sum-rate performance comparison between our proposed algorithm, the state-of-the-art centralized SINR-based algorithm from [4], and the SLNR-based algorithms from [14] and [15]. The SLNR-based algorithms work in a one-shot fashion: they first determine beam directions for each user (note: unit power), then compute an SLNR-based priority for each user, and finally determine a schedule greedily. For a fair comparison with our proposed algorithm, we make the SLNR-based algorithms perform two additional steps: (1) repeating the beam calculations after local scheduling, and (2) performing an equal power allocation. These steps are performed locally at each BS without any information exchange.

Fig. 6 shows a network sum-rate comparison of the algorithms based on data gathered over 10 topologies, each with 100 channel realizations. We emphasize that for all cases, the throughput is obtained by summing the true SINR-based rates in (2) after the algorithms are executed. While the (centralized) SINR-based algorithm with the MIFWSR initialization performs the best, the proposed decentralized SLINR-based algorithm is not far behind. Furthermore, our algorithm outperforms the SLNR-based ones, and the SINR-based one with a random initial schedule. The SINR, SLINR and SLNR Max. LB algorithms achieve similar performance; this is as expected, since setting equal weights encourages scheduling the cell-center users (with high signal power).

![Figure 6](image)

**FIGURE 6.** Average sum-rate achieved as a function of the number of users per cell.

Table 2 summarizes fairness and throughput for the best-performing algorithms in the case where $K_b = 80$. The throughput of our proposed fully decentralized SLINR-based algorithm comes within 3.8% of the SINR-based algorithm’s throughput; even though this is a performance loss, we believe this slight drawback is outweighed by the important benefit of *not needing any information exchange*. Furthermore, our proposed algorithm also provides for more fairness across users relative to the SINR case (MIFWSR scheduling), with 9% more users in the network achieving at least 1 Mbps. When comparing our proposed algorithm to the “SLINR (no leakage)” case (recall from Section IV-B that this refers to the SLINR approach with $\epsilon = 0$ in (39)), the results demonstrate that incorporating inter-cell leakage improves both fairness and throughput. Finally, we observe that the SLNR Max. LB algorithm results in the worst fairness.

2) **PROPORTIONALLY FAIR PERFORMANCE**

We now investigate the performance of the SINR, SLINR and SLNR Max. LB algorithms when proportionally fair weights are used.

Fig. 7 presents the cumulative distribution functions (CDFs) for the average user data rates over 1000 channel realizations on a single topology, and Table 3 provides accompanying numerical data (edge rates refer to 10th percentile rates). All performance measures decrease when moving from a centralized scheme to a decentralized one; this is unavoidable since proportionally fair weights require cell-edge users to be scheduled, and explicit interference coordination for these users is not possible in a decentralized setting (thus, their rates suffer). However, we once again

![Figure 7](image)

**FIGURE 7.** CDFs of long-term average data rates when proportionally fair weights are used.

| Algorithm                           | Users Achieving ≥ 1 Mbps (%) | Throughput (Mbps) |
|-------------------------------------|------------------------------|-------------------|
| SINR (MIFWSR init.)                | 24                           | 7877              |
| SINR (random init.)                | 60                           | 4647              |
| SLINR (no leakage)                 | 21                           | 7410              |
| SLINR (proposed)                   | 33                           | 7576              |
| SLNR Max. LB                       | 18                           | 7295              |

Table 2. SINR- and SLINR-based allocation for $K_b = 80$. Equal weights.
Table 3. SINR-, SLINR- and SLNR-based allocation for $K_b = 80$.
Proportionally fair weights.

| Algorithm                  | Edge User Rate (Mbps) | Throughput (Mbps) | WSR ($\times 10^3$) |
|----------------------------|-----------------------|-------------------|---------------------|
| SINR (MIFWSR init.)       | 3.39                  | 2745              | 1423                |
| SINR (random init.)       | 3.07                  | 2764              | 952                 |
| SLINR (no leakage)        | 1.67                  | 1629              | 926                 |
| SLINR (proposed)          | 1.96                  | 2008              | 1021                |
| SLNR Max. LB              | 0.62                  | 5784              | 168                 |

Notice that by incorporating inter-cell leakage, our proposed algorithm outperforms “SLINR (no leakage)” across all performance measures; in particular, Fig. 7 shows a more desirable rate CDF. As before, we notice the lack of fairness characterizing the SLNR Max. LB algorithm.

D. CHANNEL ESTIMATION

In this section, we study the max sum-rate performance of the proposed SLINR-based algorithm under imperfect CSI. Pilots were orthogonal within each cell, and reused across cells. Results were obtained by averaging over 30 topologies, each with 400 channel realizations.

Fig. 8 shows that with imperfect CSI (and no overhead considered), we suffer a throughput decrease of 11–16%, with the relative loss increasing with $K_b$. A larger performance decrease is observed once the training overhead is factored in; this is because with overhead, the achievable rate must be multiplied by the fraction of time available to transmit the data, i.e., $(\tau_d - \tau_p)/\tau_d$, where $\tau_d = 200$ is the length of the downlink transmission phase, and $\tau_p = K_b$ is the length of the training phase [19]. The sum-rate begins decreasing in $K_b$ after $K_b = 50$, indicating diminishing returns when too many resources are dedicated to the training phase.

![Figure 8. Average sum-rate achieved under imperfect CSI.](image)

Notice how (40) is identical to the inter-cell leakage value (15) computed under the “Statistical CSI” method. In summary, these two points show how, as the inter-cell channel estimation quality worsens under the “Standard” method, the estimation error for the inter-cell leakage (20) approaches

$$L_{\text{inter}}^k = \sum_{b' \in B \setminus b} \sum_{k' = 1}^{K_b} \Pr (u_{k'b'} = 1) |v_{kb'}|^2 \beta_{k'b',k}. \ (40)$$

Notice how (40) is identical to the inter-cell leakage value (15) computed under the “Statistical CSI” method. In summary, these two points show how, as the inter-cell channel estimation quality worsens under the “Standard” method, the inter-cell leakage value reduces and the estimation error for this quantity approaches a weighted sum of the large-scale fading coefficients. This is the fundamental reason why similar performance is obtained when we replace our inter-cell channel estimates with long-term information (such as the large-scale fading coefficients or traffic distribution) which considerably improve the algorithm’s practicality and scalability.

![Figure 9. Average sum-rate achieved with different leakage computation methods, both as a function of the number of users per cell.](image)
E. HEAT MAPS

This section presents heat maps to visualize the spatial distribution of rates and scheduling decisions. Data was collected using 80 users per cell \(K_b = 80\). White spots on the heat maps correspond to locations where no user was ever placed. The x-axis and y-axis denote position in metres. A logarithmic color scale is used to reveal features that are difficult to see on a linear scale (e.g., a color mapping to the value 2 on the color bar should be thought of as \(10^2\)).

Fig. 10 helps visualize the spatial distribution of rates throughout the network. Comparing the max sum-rate and proportionally fair cases, we see that the highest rates are tightly concentrated around the cell center in the max sum-rate case (as expected), whereas they are more “spread out” in the proportionally fair case. In particular, we see that the SLINR-based algorithm is the most spread out for the max sum-rate case, which visualizes the observation made in Section IV-C1 that the SLINR-based algorithm provides more fairness across users for this case.

For the proportionally fair heat maps, we see that all algorithms have “hot spots” at the cell center, but only “cool off” slightly as we approach the cell-edge. The SLINR-based algorithm has a slightly “cooler” edge than the SINR-based algorithm, which, as discussed in Section IV-C2, is due to its lack of interference coordination.

Fig. 11 visualizes the spatial distribution of scheduling decisions. Comparing the left side to the right side, we see that much like the rates, the scheduled users are highly concentrated around their BS for the max sum-rate case, whereas they are “spread out” in the proportionally fair case.

For the proportionally fair heat maps, we see that the SLINR-based algorithm schedules cell edge users most frequently, and schedules users less frequently the closer they are to the center of the cell. Since the users at the cell edge tend to have smaller rates than those at the cell center, this seems like a reasonable strategy—what may be surprising about it is that the edge rates for the SLINR-based algorithm are much lower than those of the SINR-based algorithm, which schedules users at the cell edge less frequently. The explanation is that since the SLINR-based algorithm is decentralized, when a BS schedules edge users, it causes significant amounts of interference to edge users scheduled in neighboring cells (since the BSs do not coordinate). Even incorporating inter-cell leakage can only combat this to a certain extent.

Furthermore, we see faint rings in each cell for the case of the SINR-based algorithm. This indicates that it focuses on neither the cell-edge users nor cell-center users, but rather, the users in between these two extremes. Interestingly, the heat map of rates presented in Fig. 10 does not show such a ring. The explanation is that by centrally coordinating resource allocation, the SINR-based algorithm will only schedule cell-edge users when it knows they can achieve high rates; this way, they do not have to be scheduled frequently. Similarly, the users in the cell center are scheduled relatively infrequently since they tend to have excellent channels, and therefore experience very high rates each time they are scheduled. However, the rest of the users between these two extremes are far enough away from their BS that they do not experience good channels all the time, and are far enough away from the cell edge that interference coordination does not protect them; hence, they need to be scheduled more often to achieve good average rates.

In summary, both algorithms have a spatial region corresponding to users who are frequently scheduled, but achieve rates that do not reflect this. This region is the cell edge for the SLINR-based algorithm, and it is a ring near the middle of the cell for the SINR-based algorithm. In conclusion, these regions are where interference suppression is the least effective.
V. CONCLUSION

In this paper, we studied downlink resource allocation for a multi-cell MU-MIMO network. Although resource allocation significantly improves network performance, the schemes available in the literature are largely centralized and accrue a significant overhead, especially in terms of CSI estimation and exchange. The focus here was on developing an effective decentralized resource allocation algorithm, building on the well-accepted concept of leakage. In this regard, we motivated the development of a hybrid performance measure, the SLINR, which retains attractive qualities of both SLNR and SINR; specifically, it accounts for the impact on other cells through the inter-cell leakage, and is amenable to optimal local scheduling through the intra-cell interference. Furthermore, it can be computed without any information exchange across BSs.

Using the SLINR to design an effective resource allocation algorithm required the derivation of a new approach. The results presented in Section IV compared the performance of the proposed SLINR-based algorithm to other decentralized algorithms, and the centralized SINR-based algorithm from [4]. We showed that the proposed algorithm outperforms the decentralized ones, and performs within 3.8% of the centralized one. Our algorithm converges slightly slower than the centralized one, but does so in a monotonically non-decreasing fashion. It also performs well under imperfect CSI, provided the fraction of time spent on training is limited, and interestingly, its performance is robust under limited CSI when computing the inter-cell leakage. Our results concluded with heat maps showing the spatial distributions of rates and scheduling decisions throughout the network, and a discussion relating these to the previous performance analyses.

There are various directions to consider for future work. For example, incorporating quality-of-service constraints into the resource allocator would be a timely topic, given the heterogeneous communication requirements of users in 5G networks. Evaluating performance under a shadowing model, to mimic outdoor scenarios with realistic power variations, would also likely be of practical interest. Similarly, alternative network scenarios could be investigated, such as heterogeneous cell shapes or BS transmit power constraints. Finally, since each user’s achieved rate (and therefore SINR) is available to its BS as a measured feedback signal, it may be worth investigating how this information could be used to improve the resource allocation decisions. For example, if it could be exploited to improve interference suppression at the cell edge, then performance could be improved while maintaining our requirement of full decentralization, which would be a valuable contribution.

APPENDIX A

DETAILED DERIVATION OF OPTIMAL BEAMFORMERS

In this section, we present the detailed derivation for the optimal beamforming vectors used in Section III-C, i.e., $\mathbf{v}_b^*$.

Starting from the Lagrangian in (31), consider the $k$th element of the gradient $\nabla^*_{k} L (\mathbf{z}_b, \mathbf{y}_b, \mu_b)$:

$$\frac{\partial}{\partial \mathbf{v}_{kb}} L (\mathbf{z}_b, \mathbf{y}_b, \mu_b) = \frac{\partial}{\partial \mathbf{v}_{kb}} f_q (\mathbf{z}_b, \mathbf{y}_b) - \mu_b \mathbf{v}_{kb}. \tag{42}$$

Because $f_q (\mathbf{z}_b, \mathbf{y}_b)$ sums over all users, indexed by $m$, we consider the cases $m = k$ and $m \neq k$ separately when computing $\frac{\partial}{\partial \mathbf{v}_{kb}} f_q (\mathbf{z}_b, \mathbf{y}_b, \mu_b)$. For $m \neq k$, we have

$$\frac{\partial}{\partial \mathbf{v}_{kb}} \log (1 + \phi_{mb}) - \phi_{mb}$$

$$+ \frac{2 \Re (\mathbf{v}_{mb}^H \mathbf{w}_{mb} \mathbf{u}_{mb} (1 + \phi_{mb}) \mathbf{v}_{mb}^H \mathbf{h}_{mb,b})}{\mathbf{v}_{kb}^H \mathbf{v}_{kb}}$$

$$- \frac{\partial}{\partial \mathbf{v}_{kb}} |y_{mb}|^2 \left( \text{den} [\gamma_{mb}^{11}] + u_{mb} \mathbf{h}_{mb,b}^H \mathbf{v}_{mb}^2 \right). \tag{43}$$

Simplifying (43) yields a nonzero contribution from the intra-cell interference terms inside $\text{den} [\gamma_{mb}^{11}]$:

$$-|y_{mb}|^2 \sum_{k' = 1}^{K_b} \frac{\partial}{\partial \mathbf{v}_{kb}} u_{k', b} \left( \mathbf{h}_{mb,b}^H \mathbf{v}_{k', b} + \Theta_{k', b,b} \right) \mathbf{v}_{kb}, \tag{44}$$

Changing the summation index from $m$ to $k'$ in (44), we can write the contribution to $\frac{\partial}{\partial \mathbf{v}_{kb}} f_q (\mathbf{u}_b, \mathbf{V}_b, \phi_{mb}, \mathbf{y}_b)$ for all $k' \neq k$ as

$$-u_{kb} \sum_{k' = 1}^{K_b} \sum_{k' = 1}^{K_b} |y_{k', b}|^2 \left( \mathbf{h}_{k', b}^H \mathbf{h}_{k', b} + \Theta_{k', b,b} \right) \mathbf{v}_{kb}. \tag{45}$$

For the case where $m = k$, we have

$$\frac{\partial}{\partial \mathbf{v}_{kb}} \log (1 + \phi_{kb}) - \phi_{kb}$$

$$+ \frac{2 \Re (\mathbf{v}_{kb}^H \mathbf{w}_{kb} \mathbf{u}_{kb} (1 + \phi_{kb}) \mathbf{v}_{kb}^H \mathbf{h}_{kb,b})}{\mathbf{v}_{kb}^H \mathbf{v}_{kb}}$$

$$- \frac{\partial}{\partial \mathbf{v}_{kb}} |y_{kb}|^2 \left( \text{den} [\gamma_{kb}^{11}] + u_{kb} \mathbf{h}_{kb,b}^H \mathbf{v}_{kb}^2 \right). \tag{46}$$

Simplifying (46) yields (41), shown at the bottom of the page, and adding (45) to (41) gives $\frac{\partial}{\partial \mathbf{v}_{kb}} f_q (\mathbf{u}_b, \mathbf{V}_b, \phi_{mb}, \mathbf{y}_b)$. Substituting this result into (42) yields $\frac{\partial}{\partial \mathbf{v}_{kb}} L (\mathbf{z}_b, \mathbf{y}_b, \mu_b) = 0$, we finally obtain (32).

$$y_{kb}^* \sqrt{w_{kb} u_{kb}} (1 + \phi_{kb}) \mathbf{h}_{kb,b} - |y_{kb}|^2 \sum_{k' = 1}^{K_b} \sum_{b' = 1}^{B_b} \Pr (u_{k'b' = 1}) \left( \mathbf{h}_{k'b', b}^H \mathbf{h}_{k'b', b} + \Theta_{k'b', b,b} \right) + u_{kb} \left( \mathbf{h}_{kb,b}^H \mathbf{h}_{kb,b} + \Theta_{kb,b} \right) \mathbf{v}_{kb}. \tag{41}$$
REFERENCES

[1] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” IEEE Trans. Signal Process., vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[2] K. Hosseini, C. Zhu, A. A. Khan, R. S. Adve, and W. Yu, “Optimizing the MIMO cellular downlink: Multiplexing, diversity, or interference nulling?” IEEE Trans. Commun., vol. 66, no. 12, pp. 6068–6080, Dec. 2018.

[3] A. A. Khan, R. Adve, and W. Yu, “Optimizing multicast scheduling and beamforming via fractional programming and Hungarian algorithm,” in Proc. IEEE Globecom Workshops (GC Wksps), Dec. 2018, pp. 1–6.

[4] A. A. Khan, R. Adve, and W. Yu, “Optimizing downlink resource allocation in multiuser MIMO networks via fractional programming and the Hungarian algorithm,” IEEE Trans. Wireless Commun., vol. 19, no. 8, pp. 5162–5175, Aug. 2020.

[5] T. L. Marzetta, “Massive MIMO: An introduction,” Bell Labs Tech. J., vol. 20, pp. 11–22, 2015.

[6] M. Sadek, A. Tarighat, and A. H. Sayed, “Active antenna selection in multiuser MIMO communications,” IEEE Trans. Signal Process., vol. 55, no. 4, pp. 1498–1510, Apr. 2007.

[7] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, “Distributed resource allocation for user-centric cell-free MIMO networks,” IEEE Trans. Wireless Commun., vol. 21, no. 5, pp. 3099–3115, May 2022.

[8] Z.-Q. Luo and S. Zhang, “Dynamic spectrum management: Complexity and duality,” IEEE J. Sel. Topics Signal Process., vol. 2, no. 1, pp. 57–73, Feb. 2008.

[9] K. Shen and W. Yu, “Fractional programming for communication systems—Part I: Power control and beamforming,” IEEE Trans. Signal Process., vol. 66, no. 10, pp. 2616–2630, May 2018.

[10] H. Shen, W. Xu, A. L. Swindlehurst, and C. Zhao, “Transmitter optimization for per-antenna power constrained multi-antenna downlinks: An SLNR maximization methodology,” IEEE Trans. Signal Process., vol. 64, no. 10, pp. 2712–2725, May 2016.

[11] B. Bojadzieva and A. Klein, “Distributed two-stage beamforming with power allocation in multicell massive MIMO,” in Proc. IEEE Int. Conf. Commun. (ICC), May 2019, pp. 1–7.

[12] J. Zhao, H. Dai, X. Wu, X. Xu, and W.-P. Zhu, “Joint beamforming design for energy efficient wireless communications in heterogeneous intelligent connected vehicles networks,” IEEE Access, vol. 7, pp. 170134–170143, 2019.

[13] P. Patcharanarumol, S. M. D. Armour, and A. Doufexi, “On the equivalence between SLNR and MMSE precoding schemes with single-antenna receivers,” IEEE Commun. Lett., vol. 16, no. 7, pp. 1034–1037, Jul. 2012.

[14] X. Xia, G. Wu, J. Liu, and S. Li, “Leakage-based user scheduling in MU-MIMO broadcast channel,” Sci. China Ser. F, Inf. Sci., vol. 52, no. 12, pp. 2259–2268, Dec. 2009.

[15] Y. Liu, J. Wang, P. Wang, and R. Yu, “Joint scheduling and precoding based on SLNR criteria in MU-MIMO systems,” in Mechanical Components and Control Engineering III (Applied Mechanics and Materials), vol. 668. Stafa-Zurich, Switzerland: Trans Tech Publications Ltd, Nov. 2014, pp. 1273–1277.

[16] W. Luo, A. Guo, W. Tan, and Y. Ji, “An improved beamforming method based on SLNR for downlink multi-user multi-stream MIMO system,” Wireless Pers. Commun., vol. 72, no. 4, pp. 2673–2683, Oct. 2013.

[17] M. Wu, “Inter-cell interference mitigation in multi-cell coordinating system based on SLNR method,” J. Netw., vol. 9, no. 1, pp. 153–160, Jan. 2014.

[18] D. Han and N. Lee, “Distributed precoding using local CSIT for MU-MIMO heterogeneous cellular networks,” IEEE Trans. Commun., vol. 69, no. 3, pp. 1666–1678, Mar. 2021.

[19] Z. Li, T. Gamvrelis, H. A. Ammar, and R. Adve, “Decentralized user scheduling and beamforming in multi-cell MIMO networks,” in Proc. IEEE Int. Conf. Commun., May 2022, pp. 1980–1985.

[20] C. Pan, H. Mehrpouyan, Y. Liu, M. Elkashlan, and N. Arumugam, “Joint pilot allocation and robust transmission design for ultra-dense user-centric TDD C-RAN with imperfect CSI,” IEEE Trans. Wireless Commun., vol. 17, no. 3, pp. 2038–2053, Mar. 2018.

[21] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-free massive MIMO versus small cells,” IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 1834–1850, Jan. 2017.

[22] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan, and L. Hanzo, “Weighted sum-rate maximization for the ultra-dense user-centric TDD C-RAN downlink relying on imperfect CSI,” IEEE Trans. Wireless Commun., vol. 18, no. 2, pp. 1182–1198, Feb. 2019.

[23] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, “Optimization of MIMO device-to-device networks via matrix fractional programming: A minorization-maximization approach,” IEEE/ACM Trans. Netw., vol. 27, no. 5, pp. 2164–2177, Oct. 2019.

[24] S. Bougleux and L. Brun, “Linear sum assignment with edition,” 2016, arXiv:1603.04380.

TYLER GAMVRELIS (Student Member, IEEE) received the B.A.Sc. degree in engineering science with a major in electrical and computer engineering and the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2020 and 2022, respectively. His research interests include network slicing, market design, and signal processing algorithms for multi-antenna systems. He was a recipient of the Edward S. Rogers Sr. Graduate Scholarship, the Ontario Graduate Scholarship, and the William Taylor Scholarship in Radiocommunications.

ZEHUA LI (Student Member, IEEE) is currently pursuing the B.A.Sc. degree in engineering science with a major in electrical and computer engineering and minor in robotics and mechatronics with the University of Toronto, Toronto, ON, Canada. His research interests include wireless network optimization, signal processing, and robotics. He was a recipient of the William V. Hull Scholarship and the Faculty of Applied Science & Engineering Admission Scholarship.

AHMAD ALI KHAN (Student Member, IEEE) received the B.E.E. degree from the King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 2015, and the M.A.Sc. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include optimization theory, resource allocation in communications networks, signal processing, and machine learning for wireless communications. He was a recipient of the Edward S. Rogers Sr. Scholarship, the Queen Elizabeth II Graduate Scholarship in Science and Technology, the Ontario Graduate Scholarship, and the Wallberg Memorial Fellowship.

RAVIRAJ S. ADVE (Fellow, IEEE) was born in Mumbai, India. He received the B.Tech. degree in electrical engineering from IT Bombay, in 1990, and the Ph.D. degree from Syracuse University, in 1996, From 1997 to 2000, he was a Research Associate with Defense Conversion Inc., on contract with the Air Force Research Laboratory, Rome, NY, USA. He joined as a Faculty Member of the University of Toronto, in 2000, where he is currently a Professor. His research interests include analysis and design techniques for cooperative and heterogeneous networks, energy harvesting networks, and signal processing techniques for radar and sonar systems. He received the 2009 Fred Nathanson Young Radar Engineer of the Year Award. His thesis received the Syracuse University Outstanding Dissertation Award.