Constrained Base Station Clustering for Cooperative Post-Cellular Relay Networks

Raphael T. L. Rolny, Marc Kuhn, and Armin Wittneben
Swiss Federal Institute of Technology (ETH) Zurich,
Communication Technology Laboratory, CH-8092 Zurich, Switzerland
Email: {rolny, kuhn, wittneben}@nari.ee.ethz.ch

***Abstract*—We consider the downlink of a "post-cellular" network in which base stations (BSs) are no longer evenly spread in the entire area of service but rather placed where they can easily be installed. Accordingly, the BSs are not necessarily close to the users anymore but have to serve distant mobiles with aggressive spatial multiplexing, possibly with the help of additional low-complexity relays. In such networks, it is important to allow for flexible and dynamic cooperation between different infrastructure nodes to achieve high data rates and coverage. To this end, we develop a framework that optimizes the performance by dynamic BS clustering with joint transmission for one- and two-hop communication under practical conditions such as backhaul rate limitations and power control. The network performance is maximized by an evolutionary algorithm that optimizes the BS clustering, relay routing, and power allocation for a beamforming scheme with closed-form solution. We show that a very high performance can be achieved in an efficient way with small overhead and that high data rates can be delivered to areas without wired backhaul access.**

I. Introduction

Future cellular networks face high expectations regarding data rates, coverage, and reliability. In order to increase the performance by orders of magnitude compared to today's systems, sophisticated interference management is inevitable, as such networks are mainly interference limited. The classical approach thereby is to introduce a spatial reuse that ensures a certain separation between base stations (BSs) that use the same physical channel [1]. By exploiting the pathloss, each BS has to be closer to the users it serves than to other users it interferes with, which leads to the usual *cellular* network topology. To increase the network capacity, additional BSs can be installed in the areas where it is most needed, which reduces the cell sizes down to pico- or femto-cells [2], [3]. In practice, however, finding sufficiently many new BS sites is in many cases difficult or even impossible, e.g. due to the costs, availability of backbone access, or public acceptance.

To this end, it might be more convenient to abandon the cellular network layout but rather let backhaul access points operate in places where they can most easily be installed. With this, the typical cells of traditional networks vanish and the backhaul access points have to serve mobile stations (MSs) that are possibly far away. Such a "post-cellular" network topology, however, requires an aggressive spatial multiplexing that separates interfering users. This can be achieved by large antenna arrays at the BSs (massive MIMO) [4] or by forming large virtual antenna arrays by cooperation/coordinated multipoint (CoMP) transmission across different BS sites [5], [6]. Dedicated stationary nodes not belonging to the wired infrastructure, such as relays, can thereby assist the communication between BSs and MSs. As shown by the recently introduced relay carpet concept [7], [8], even relays of very low complexity that are spread in large numbers can substantially reduce system complexity while a favorable performance scaling is maintained. Example scenarios of such post-cellular networks could include cities where sophisticated backhaul access points/BSs are installed *around* the city where more space is available and the public acceptance is higher or the roll-out of a new technology or a developing country where a high density of BSs is available in a confined region (e.g. the capital city) but no wired infrastructure in the backcountry. The MSs in the city center or backcountry could then be served by massive relay enabled spatial multiplexing or the help of possibly available existing BSs or residential internet access points.

In this paper, we consider such a post-cellular network where sparsely located BSs should serve MSs that are possibly far away, by cooperation and the help of low-complexity relays (see Fig. 1). To this end, the BS cooperation should be flexible without fixed cooperation clusters but with dynamic cooperation sets to adapt to changing user distributions and rate requirements [9], [10]. Additionally, when different types of BSs cooperate with each other, they might have different backhaul connections and different computational capabilities which have to be taken into account by the transmission schemes. From a practical perspective, it is not feasible that all BSs cooperate with each other in a global fashion, but each MS should be served only by a subset of BSs. Such a dynamic BS clustering approach has already been studied e.g. in [10]–[12], where a sparse beamforming design is developed. These approaches, however, require complex iterative optimization procedures that are computationally very costly, accurate channel state information (CSI) of the entire network and cannot be extended to relay assisted networks. In this work, we develop a framework that allows to optimize the network performance

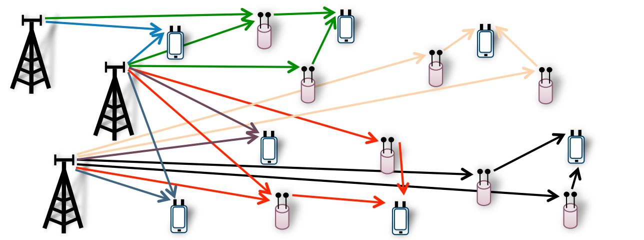

Fig. 1. A "post-cellular" network in which a cluster of BSs serves a wide area by the assistance of many distributed low-complexity relays.

by dynamic BS clustering with joint coherent transmission for one- and two-hop (relay assisted) communication under practical conditions as backhaul rate constraints and power control. To this end, we apply a leakage-based beamforming scheme [13] that can be calculated in closed-form and develop an extension thereof for two-hop communication with simple amplify-and-forward (AF) relays where the fading coefficients of the second hop channel are unknown. The BS cooperation clustering, relay routing, and power allocation are then optimized by an evolutionary algorithm [14] that is very flexible with respect to changing requirements, network conditions and/or constraints. By simulation results, we show that such a relay enabled post-cellular network can efficiently deliver high data rates to areas where no wired backhaul access is present.

II. System Model

Consider the downlink transmission from B BSs to M MSs which is assisted by K relays. For the sake of notational simplicity, we assume that all BSs are equipped with N_B antennas, all MSs with N_M, and the relays with N_R antennas. With BS cooperation, each MS $j \in \{1, \ldots, M\}$ can be served jointly by a subset of BSs, either directly or via relays. In order to represent which BSs serve which MSs, we define the *routing* or *BS clustering* matrix $\mathbf{C} \in \{0,1\}^{M\times B}$ as the matrix whose element $\mathbf{C}_{(j,b)}$ in the j-th row and b-th column is 1 if MS j is served by BS b and 0 otherwise. Each MS can thus be served by an arbitrary number of BSs and the BS cooperation clusters for different MSs can overlap or contain different nodes. Furthermore, we use $\mathcal{I}_j$ to denote the set of indices of the non-zero elements of the j-th row $\mathbf{C}_{(j,:)}$ (the BSs that serve MS j) and $\mathcal{J}_b$ to denote the index set of the non-zero elements of the b-th column $\mathbf{C}_{(:,b)}$ (the MSs that are served by BS b).

For a practical implementation, we limit ourselves to linear precoding and describe the transmit signal of BS b by $\mathbf{x}_b = \sum_{j\in\mathcal{J}_b} \mathbf{Q}_{j,b} \cdot \mathbf{s}_j$, where $\mathbf{s}_j \in \mathbb{C}^{d_s}$ is the data symbol vector with d_s elements i.i.d. $\mathcal{CN}(0,1)$ intended for MS j and $\mathbf{Q}_{j,b} \in \mathbb{C}^{N_\mathrm{B}\times d_s}$ the corresponding precoding matrix from BS b. The transmit signals are constrained to a per-BS sum transmit power constraint

$$\mathrm{Tr}\Big\{\sum_{j\in\mathcal{J}_b} \mathbf{Q}_{j,b}\cdot\mathbf{Q}_{j,b}^\mathsf{H}\Big\} \le P_\mathrm{B}. \tag{1}$$

In the following, we assume that the MSs are either served directly by their associated BSs or they are served via relays in a two-hop fashion where the direct channel from the BSs is not used[1]. For the former case, we use $\mathbf{H}_{j,b} \in \mathbb{C}^{N_\mathrm{M}\times N_\mathrm{B}}$ to describe the block fading channel from BS b to MS j and $\mathbf{H}_{j,\mathcal{I}_j} = \left[\{\mathbf{H}_{j,b}\}_{b\in\mathcal{I}_j}\right]$ for the concatenated channel matrix from all BSs $b \in \mathcal{I}_j$ to MS j. With $\mathbf{w}_j$ being the additive noise with elements i.i.d. $\mathcal{CN}(0,\sigma_w^2)$, the receive signal at MS j, split into desired signal, interference, and noise, is given by

$$\mathbf{y}_j = \mathbf{H}_{j,\mathcal{I}_j}\cdot\mathbf{Q}_j\cdot\mathbf{s}_j + \sum_{\substack{i=1\\i\ne j}}^{M} \mathbf{H}_{j,\mathcal{I}_i}\cdot\mathbf{Q}_i\cdot\mathbf{s}_i + \mathbf{w}_j, \tag{2}$$

[1] A combination of direct and two-hop communication can be included into the framework, but is not presented here.

where $\mathbf{Q}_j = \left[\{\mathbf{Q}_{j,b}^\mathsf{T}\}_{b\in\mathcal{I}_j}\right]^\mathsf{T}$ is the concatenated precoding matrix for the signal to MS j from all its serving BSs.

In the two-hop case, we use $\mathbf{H}^{(1)}_{k,\mathcal{I}_j}$ to denote the first hop channel from all BSs in the set $\mathcal{I}_j$ to relay k. The receive signal of relay k is then

$$\mathbf{r}_k = \sum_{j=1}^{M} \mathbf{H}^{(1)}_{k,\mathcal{I}_j}\cdot\mathbf{Q}_j\cdot\mathbf{s}_j + \mathbf{n}_k, \tag{3}$$

with $\mathbf{n}_k$ being the relay noise with elements i.i.d. $\mathcal{CN}(0,\sigma_n^2)$. In order to use relay nodes of very low complexity and low cost, such that they can be implemented in massive numbers, we consider simple AF relays as in [8] that perform a frequency conversion and amplify the received signal with a gain matrix $\mathbf{G}_k = \alpha_k\cdot\mathbf{I}_{N_\mathrm{R}}$, where α_k is the amplification factor. Such relays can be implemented in a very inexpensive way (analog) and introduce no additional delays. With $\mathbf{H}^{(2)}_{j,k}$ describing the second hop channel between relay k and MS j, the receive signal at MS j follows as

$$\mathbf{y}_j = \sum_{k=1}^{K}\sum_{i=1}^{M} \mathbf{H}^{(2)}_{j,k}\mathbf{G}_k\mathbf{H}^{(1)}_{k,\mathcal{I}_i}\mathbf{Q}_i\mathbf{s}_i + \sum_{k=1}^{K}\mathbf{H}^{(2)}_{j,k}\mathbf{G}_k\mathbf{n}_k + \mathbf{w}_j. \tag{4}$$

For given clustering, precoding, and relay gain matrices, the achievable rate for MS j can be computed by

$$R_j = \log_2\det\left\{\mathbf{I} + \left(\mathbf{K}_j^{(\mathrm{i+n})}\right)^{-1}\cdot\mathbf{K}_j^{(\mathrm{sig})}\right\}, \tag{5}$$

with the covariance matrices of the desired signal and interference plus noise for the direct transmission

$$\mathbf{K}_j^{(\mathrm{sig})} = \mathbf{H}_{j,\mathcal{I}_j}\mathbf{Q}_j\mathbf{Q}_j^\mathsf{H}\mathbf{H}_{j,\mathcal{I}_j}^\mathsf{H}$$
$$\mathbf{K}_j^{(\mathrm{i+n})} = \sum_{\substack{i=1\\i\ne j}}^{M}\mathbf{H}_{j,\mathcal{I}_i}\mathbf{Q}_i\mathbf{Q}_i^\mathsf{H}\mathbf{H}_{j,\mathcal{I}_i}^\mathsf{H} + \sigma_w^2\cdot\mathbf{I}_{N_\mathrm{M}}.$$

In the case of the two-hop transmission, a prelog factor $\frac{1}{2}$ applies to the achievable rate due to the half-duplex relays. The covariance matrices are in this case

$$\mathbf{K}_j^{(\mathrm{sig})} = \sum_{k=1}^{K}\sum_{l=1}^{K}\mathbf{H}^{(2)}_{j,k}\mathbf{G}_k\mathbf{H}^{(1)}_{k,\mathcal{I}_j}\mathbf{Q}_j\mathbf{Q}_j^\mathsf{H}\mathbf{H}^{(1)\mathsf{H}}_{l,\mathcal{I}_j}\mathbf{G}_l^\mathsf{H}\mathbf{H}^{(2)\mathsf{H}}_{j,l}$$
$$\mathbf{K}_j^{(\mathrm{i+n})} = \mathsf{E}\left[\mathbf{y}_j\mathbf{y}_j^\mathsf{H}\right] - \mathbf{K}_j^{(\mathrm{sig})}.$$

In the following, we develop how the precoding and the cooperation clustering can be designed to maximize the network performance. To this end, we first consider the case of direct transmission without relays and extend the framework afterwards to the two-hop case.

III. Cooperative Precoding

In order to achieve a good network performance, the cooperation sets described by $\mathbf{C}$ need to be chosen appropriately and the precoding matrices $\mathbf{Q}_j$ need to be optimized. In practical systems, it is thereby hardly possible that all BSs of the entire network can cooperate with each other. Not only would the computational complexity be prohibitive but also the required data traffic between BSs would exceed any feasible backhaul capacity. It is therefore desirable that each MS is only served by a small or moderate number of BSs, i.e. that the clustering matrix $\mathbf{C}$ is *sparse*, and that practical conditions as limited backhaul rates are considered in the optimization. In this way, the complexity of sharing data between different BSs and joint signal processing can be reduced.

This aspect has already been studied in [11], where the authors attempt to maximize the system performance under per-BS backhaul rate constraints. The resulting optimization problem can be solved by approximating these backhaul constraints as ℓ_1-norm constraints and iterating over a number of equivalent convex optimization problems that minimize the weighted minimum mean square errors (WMMSE). The backhaul constraints are thereby an important ingredient to enforce sparse clusters, as the more stringent the backhaul constraints are, the less MSs can be served by a BS. While this approach converges to a locally optimal precoding, it requires global CSI and high computational power, as the algorithm consists of an iterative procedure in which a convex optimization problem has to be solved in each iteration. In the following, we attempt to maximize the network performance in a suboptimal but more efficient way that can be distributed and extended to two-hop networks which is not directly possible with the WMMSE algorithm.

A. Leakage Based Precoding

In order to get a low-complexity algorithm, we apply a leakage-based beamforming scheme as in [13]. Although it is suboptimal in terms of achievable rate, it can be calculated in closed form, which is efficient, and is a well-proven and suitable choice for flexible BS cooperation, as it decouples the transmissions to different users.

For a given BS clustering matrix $\mathbf{C}$, let $\mathbf{H}_{j,\mathrm{L}} = \left[\{\mathbf{H}_{i,b}\}_{i\neq j, b\in\mathcal{I}_j}\right]$ be the matrix that contains all channels from the BSs that are involved in the transmission to MS j to all other MSs $i \neq j$, i.e. the channel that *leaks* interference to other users. Furthermore, we induce a per-MS power budget $\mathrm{Tr}\left\{\mathbf{Q}_j\mathbf{Q}_j^{\mathsf{H}}\right\} \leq P_{\mathrm{M}}$ instead of the per-BS transmit power constraint (1). This allows a scaling of the precoding matrices

$$\tilde{\mathbf{Q}}_j = \sqrt{P_{\mathrm{M}}/\mathrm{Tr}\{\mathbf{Q}_j\mathbf{Q}_j^{\mathsf{H}}\}} \cdot \mathbf{Q}_j \tag{6}$$

across the different BSs that are involved. The per-BS transmit power constraint is handled later. The signal-to-leakage-and-noise ratio (SLNR) follows as [13]

$$\mathrm{SLNR}_j = \frac{\mathrm{Tr}\left\{\mathbf{H}_{j,\mathcal{I}_j}\tilde{\mathbf{Q}}_j\tilde{\mathbf{Q}}_j^{\mathsf{H}}\mathbf{H}_{j,\mathcal{I}_j}^{\mathsf{H}}\right\}}{\mathrm{Tr}\left\{\mathbf{H}_{j,\mathrm{L}}\tilde{\mathbf{Q}}_j\tilde{\mathbf{Q}}_j^{\mathsf{H}}\mathbf{H}_{j,\mathrm{L}}^{\mathsf{H}} + \sigma_w^2\cdot\mathbf{I}_{N_{\mathrm{M}}}\right\}} \tag{7}$$

$$= \frac{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\mathbf{H}_{j,\mathcal{I}_j}^{\mathsf{H}}\mathbf{H}_{j,\mathcal{I}_j}\mathbf{Q}_j\right\}}{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\left(\mathbf{H}_{j,\mathrm{L}}^{\mathsf{H}}\mathbf{H}_{j,\mathrm{L}} + \frac{\sigma_w^2 N_{\mathrm{M}}}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}\right)\mathbf{Q}_j\right\}}, \tag{8}$$

where the second equality follows from (6) and the cyclic shift property of the trace. The SLNR is maximized by choosing $\mathbf{Q}_j$ as the (scaled according to (6)) generalized eigenvector (GEV) that corresponds to the largest generalized eigenvalue of the matrix pair $\mathbf{H}_{j,\mathcal{I}_j}^{\mathsf{H}}\mathbf{H}_{j,\mathcal{I}_j}$ and $\mathbf{H}_{j,\mathrm{L}}^{\mathsf{H}}\mathbf{H}_{j,\mathrm{L}} + \frac{\sigma_n^2 N_{\mathrm{M}}}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}$, in short

$$\mathbf{Q}_j^{\star} = \max\mathrm{GEV}\left\{\mathbf{H}_{j,\mathcal{I}_j}^{\mathsf{H}}\mathbf{H}_{j,\mathcal{I}_j}, \mathbf{H}_{j,\mathrm{L}}^{\mathsf{H}}\mathbf{H}_{j,\mathrm{L}} + \frac{\sigma_w^2 N_{\mathrm{M}}}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}\right\}. \tag{9}$$

In order to support multiple spatial streams to each MS, we extend this solution and choose for $\mathbf{Q}_j$ the d_{s} largest generalized eigenvectors. Note that this form of precoding does not impose any conditions on the number of BSs, MSs, or their antennas and each MS is served by its own subset of BSs that can overlap with other subsets associated with other users.

B. Precoding for Two-Hop Networks

When the MSs are far away from the BSs, the performance can be enhanced by relays that are spread in large numbers. To this end, the relays should be of low cost and low complexity and can therefore not fulfill complicated tasks. We thus assume that no CSI is available from the second hop, i.e., the relays are not able to estimate any channels. The BSs, however, have *local* CSI of their first hop channels. This CSI is simple to acquire when the relays are fixed. In this case, the BSs only have to track quasi-static channels which simplifies channel estimation [8]. Additionally, the BSs might have *statistical* CSI of the channels between relays and MSs. This can e.g. be obtained when the positions of the relays and MSs are known, possibly by reporting GPS information. For this scenario, we develop two strategies to adapt the SLNR beamforming.

1) Ignoring the 2nd hop: A simple and straightforward way to calculate the precoding is to treat selected relays as the receivers and to ignore the second hop. To this end, we use the relay clustering matrix $\mathbf{D} \in \{0,1\}^{M\times K}$ which is defined, like the BS clustering matrix $\mathbf{C}$, as the matrix whose element $\mathbf{D}_{(j,k)}$ is 1 if MS j is served via relay k and 0 otherwise and $\mathcal{I}_j^{(\mathrm{R})}$ for the corresponding index set. The (assumed) channel for the desired signal is then $\mathbf{H}_j^{(1)} = \left[\{\mathbf{H}_{k,b}^{(1)}\}_{k\in\mathcal{I}_j^{(\mathrm{R})}, b\in\mathcal{I}_j}\right]$ and the covariance matrix for the SLNR calculation is $\mathbf{R}_j = \mathbf{H}_j^{(1)\mathsf{H}}\cdot\mathbf{H}_j^{(1)}$. Similarly, the leakage covariance matrix is $\mathbf{R}_{j,\mathrm{L}} = \mathbf{H}_{j,\mathrm{L}}^{(1)\mathsf{H}}\cdot\mathbf{H}_{j,\mathrm{L}}^{(1)}$, where $\mathbf{H}_{j,\mathrm{L}}^{(1)}$ is the concatenated matrix that contains all channels from the selected BSs in $\mathcal{I}_j$ to all relays that are not selected for transmission to MS j. With this, the precoding for MS j can, as in (9), be found by

$$\mathbf{Q}_j^{\star} = \max\mathrm{GEV}\left\{\mathbf{R}_j, \mathbf{R}_{j,\mathrm{L}} + \frac{\sigma_n^2 N_{\mathrm{R}}|\mathcal{I}_j^{(\mathrm{R})}|}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}\right\}, \tag{10}$$

which maximizes the SLNR at the selected relays

$$\mathrm{SLNR}_j = \frac{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\mathbf{R}_j\mathbf{Q}_j\right\}}{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\left(\mathbf{R}_{j,\mathrm{L}} + \frac{\sigma_n^2 N_{\mathrm{R}}|\mathcal{I}_j^{(\mathrm{R})}|}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}\right)\mathbf{Q}_j\right\}}. \tag{11}$$

The signals transmitted by the BSs are then received by the relays, amplified and forwarded to the MSs. For appropriately chosen relay gains and clustering matrices $\mathbf{C}$ and $\mathbf{D}$, the interference that these relays generate should be kept low.

2) Expectation over 2nd hop: As an alternative, the BSs can make use of statistical CSI of the second hop. Under the assumption of Rayleigh fading, the second hop channels can be written as $\mathbf{H}_{j,k}^{(2)} = \sqrt{L_{j,k}} \cdot \mathbf{F}_{j,k}$, where $\mathbf{F}_{j,k} \sim \mathcal{CN}(\mathbf{O}, \mathbf{I})$ is the unknown small scale fading and $L_{j,k}$ is the known path gain (variance). With this, we can calculate the expectation with respect to $\mathbf{F}_{j,k}$ of the desired signal power at MS j

$$\begin{aligned} P_j^{(\mathrm{sig})} &= \mathsf{E}_{\mathbf{F}}\left[\mathsf{E}_{\mathbf{s}}\left[\left\|\sum_{k=1}^{K}\mathbf{H}_{j,k}^{(2)}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\mathbf{Q}_j\mathbf{s}_j\right\|_2^2\right]\right] \\ &= \mathsf{E}_{\mathbf{F}}\left[\mathrm{Tr}\left\{\sum_{k,l}\sqrt{L_{j,k}L_{j,l}}\mathbf{F}_{j,k}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\mathbf{Q}_j\mathbf{Q}_j^{\mathsf{H}}\mathbf{H}_{l,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_l^{\mathsf{H}}\mathbf{F}_{j,l}^{\mathsf{H}}\right\}\right] \\ &= \sum_{k,l}\sqrt{L_{j,k}L_{j,l}}\mathrm{Tr}\left\{\mathsf{E}_{\mathbf{F}}\left[\mathbf{F}_{j,l}^{\mathsf{H}}\mathbf{F}_{j,k}\right]\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\mathbf{Q}_j\mathbf{Q}_j^{\mathsf{H}}\mathbf{H}_{l,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_l^{\mathsf{H}}\right\} \\ &= \mathrm{Tr}\left\{\sum_{k=1}^{K}L_{j,k}N_{\mathrm{M}}\cdot\mathbf{Q}_j^{\mathsf{H}}\mathbf{H}_{k,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_k^{\mathsf{H}}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\mathbf{Q}_j\right\} \\ &= \mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\left(N_{\mathrm{M}}\cdot\sum_{k=1}^{K}L_{j,k}\mathbf{H}_{k,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_k^{\mathsf{H}}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\right)\mathbf{Q}_j\right\}, \end{aligned}$$

which follows from the cyclic shift property of the trace and

$$\mathsf{E}_{\mathbf{F}}\left[\mathbf{F}_{j,l}^{\mathsf{H}}\mathbf{F}_{j,k}\right] = \begin{cases} N_{\mathrm{M}}\cdot\mathbf{I}_{N_{\mathrm{R}}}, & \text{if } k=l \\ \mathbf{O}, & \text{if } k\neq l \end{cases}$$

as the small scale fading is assumed to be independent for channels between different nodes. Likewise, the expected leakage power can be calculated by

$$\begin{aligned} P_j^{(\mathrm{leak})} &= \mathsf{E}_{\mathbf{F}}\left[\mathsf{E}_{\mathbf{s}}\left[\left\|\sum_{i\neq j}\sum_{k=1}^{K}\mathbf{H}_{i,k}^{(2)}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\mathbf{Q}_j\mathbf{s}_j\right\|_2^2\right]\right] \\ &= \mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\left(N_{\mathrm{M}}\cdot\sum_{i\neq j}\sum_{k=1}^{K}L_{i,k}\mathbf{H}_{k,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_k^{\mathsf{H}}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)}\right)\mathbf{Q}_j\right\} \end{aligned}$$

and the expected noise power follows the same lines as

$$P_j^{(\mathrm{noise})} = N_{\mathrm{M}}\cdot\mathrm{Tr}\left\{\sum_{k=1}^{K}\sigma_n^2L_{j,k}\mathbf{G}_k\mathbf{G}_k^{\mathsf{H}}\right\} + N_{\mathrm{M}}\cdot\sigma_w^2.$$

The SLNR, in which N_{M} cancels out, is then given by

$$\mathrm{SLNR}_j = \frac{P_j^{(\mathrm{sig})}}{P_j^{(\mathrm{leak})}+P_j^{(\mathrm{noise})}} = \frac{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\cdot\mathbf{R}_j\cdot\mathbf{Q}_j\right\}}{\mathrm{Tr}\left\{\mathbf{Q}_j^{\mathsf{H}}\cdot\mathbf{R}_{j,\mathrm{L+N}}\cdot\mathbf{Q}_j\right\}} \quad (12)$$

with

$$\begin{aligned} \mathbf{R}_j &= \sum_{k=1}^{K}L_{j,k}\mathbf{H}_{k,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_k^{\mathsf{H}}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)} \\ \mathbf{R}_{j,\mathrm{L+N}} &= \sum_{i\neq j}\sum_{k=1}^{K}L_{i,k}\mathbf{H}_{k,\mathcal{I}_j}^{(1)\mathsf{H}}\mathbf{G}_k^{\mathsf{H}}\mathbf{G}_k\mathbf{H}_{k,\mathcal{I}_j}^{(1)} + \\ &\quad + \frac{\mathrm{Tr}\left\{\sum_{k=1}^{K}\sigma_n^2L_{j,k}\mathbf{G}_k\mathbf{G}_k^{\mathsf{H}}\right\}+\sigma_w^2}{P_{\mathrm{M}}}\cdot\mathbf{I}_{N_{\mathrm{B}}}, \end{aligned}$$

which is again maximized by

$$\mathbf{Q}_j^{\star} = \max\mathrm{GEV}\left\{\mathbf{R}_j, \mathbf{R}_{j,\mathrm{L+N}}\right\}. \quad (13)$$

IV. Constrained Cluster Optimization

After having calculated the precoding for fixed and given clustering matrices, we now attempt to find optimal clustering matrices $\mathbf{C}$ (and $\mathbf{D}$ if required). To this end, we apply an evolutionary algorithm [14] with which we also incorporate power control to further improve the performance and to satisfy the per-BS transmit power constraint (1). Finding the optimal clustering matrix is a combinatorial problem that is generally hard to solve. Evolutionary algorithms are heuristic search techniques that simulate natural selection and evolution with which such problems can be solved efficiently.

For the optimization, we consider one of the following utility or fitness functions of the clustering $\mathbf{C}$ (and thus the achievable rates): (i) the sum rate $f_{\Sigma}(\mathbf{C}) = \sum_{j=1}^{M}R_k$, (ii) the minimum rate $f_{\min}(\mathbf{C}) = \min_j\{R_1,\ldots,R_M\}$, or (iii) the outage probability $f_{\mathrm{out}}(\mathbf{C}) = \Pr\{R_j < R^{\star}\}$ for some target rate $R^{\star}$. In the optimization process, a population of N_{ind} individual clustering matrices $\mathbf{C}_n^{(t)}$, for $n=1,\ldots,N_{\mathrm{ind}}$ and iteration index $t=0,1,2,\ldots$, is generated, their evaluated fitness functions are compared, and the clusters are updated by mutation until a sufficiently good solution is found.

As initialization, N_{ind} clustering matrices $\mathbf{C}_n^{(0)}$ are randomly generated. For each of them, the SLNR precoding as well as the resulting achievable rates are calculated as described before. All individuals of the population are then sorted according to the evaluated fitness function and the N_{sur} best (surviving) clustering matrices are selected. In each iteration step t, the surviving clustering matrices are reused and N_{child} children are generated from each of them. To this end, each bit of the actual clustering matrix $\mathbf{C}_n^{(t)}$ is flipped with a probability p_t that can change during the iteration steps for an accelerated convergence, i.e.

$$\mathbf{C}_n^{(t+1)} = \left(\mathbf{C}_n^{(t)} + \mathbf{C}_{\mathrm{rand}}(p_t)\right) \mod 2. \quad (14)$$

For a good tradeoff between performance and convergence speed, we update the bit flip probabilities with $p_{t+1} = \tau\cdot p_t$, for some update factor $\tau\in(0,1)$. Additional to the survivors and the children, we also generate $N_{\mathrm{new}} = N_{\mathrm{ind}} - N_{\mathrm{sur}} - N_{\mathrm{sur}}\cdot N_{\mathrm{child}}$ new individuals randomly in each iteration step. In the following, the algorithm is extended to handle power control as well as additional constraints.

1) Power Control: The SLNR precoding matrices are designed based on a per-MS power budget. Due to the cooperation between multiple BSs, a per-BS transmit power constraint is not straightforward to apply, because the entire precoding matrix $\mathbf{Q}_j$ needs to be scaled with one scaling factor, otherwise the SLNR optimality would be destroyed. To this end, the per-MS power budgets $P_{\mathrm{M},j}$ are applied which can be chosen such that the per-BS power constraint is met during the iterations of the optimization. To achieve this, the clustering matrix $\mathbf{C}$ is extended to $\tilde{\mathbf{C}} = [\mathbf{C}, \mathbf{P}]$, where row j of $\mathbf{P}$ is a binary representation with N_{bits} bits that describes the fraction of the maximal power $P_{\mathrm{M},\max}$ that is allocated to MS j. The extended matrix $\tilde{\mathbf{C}}$ is then optimized in the evolutionary algorithm as before, attempting not only to find the optimal BS-MS association, but also an optimal power scaling for each link.

The per-BS power constraint can be incorporated by setting

$$\tilde{R}_j = \begin{cases} R_j, & \text{as in (5) if power constraint is met} \\ 0, & \text{if power constraint is violated} \end{cases}, \quad (15)$$

i.e., a rate resulting from a precoding that violates the constraint is rejected and the optimizer tries to find a valid solution.

2) Additional Constraints: In a similar way, we can incorporate additional constraints as e.g. a per-BS backhaul rate constraint

$$R_{\Sigma,b} = \sum_{j \in \mathcal{J}_b} R_j \leq C_b, \quad \forall b, \quad (16)$$

where the total delivered data rate of BS b must not exceed the capacity C_b of its backhaul connection.[2] To this end, $R_{\Sigma,b}$ is calculated in each iteration. If this value fulfills constraint (16), the current achievable rates are used, otherwise, they are set to zero as in (15). In this way, the algorithm can also deal with additional, possibly difficult, constraints.

3) Relay Networks: The performance of the relay network can also be optimized in the same way. If the SLNR scheme that ignores the second hop is applied, the relay selection matrix $\mathbf{D}$ has to be included into the optimization. It is thereby treated in the same way as the BS clustering matrix $\mathbf{C}$. For the case of the second SLNR scheme in which the statistical CSI of the second hop is considered, no such relay selection matrix is required as the relays are selected implicitly by the BS beamforming. In order to optimize the relay gain factors α_k, we can proceed as with BS power control: the applied relay gains are encoded as binary bit strings, where an optimal fraction of $\alpha_{\max}$ is applied to each relay.

With this algorithm, we can solve the difficult problem of BS clustering, relay routing, and power allocation under practical constraints. To this end, the achievable rates have to be evaluated in each iteration. This can either be realized in a distributed way by rate feedbacks from the MSs or centrally by generating random virtual fading coefficients of the unknown second hop channels and by maximizing the sample means $\overline{R}_j = \frac{1}{t}\sum_{i=0}^{t} R_j^{(i)}$. As a result, the algorithm can track changes in the network on the fly without starting from the beginning but by updating existing solutions, e.g. when some fading coefficients change or when MSs drop out of the network or if new nodes come in. In this way, the algorithm can efficiently adapt to new requirements or network conditions.

V. Simulation Results & Discussion

In the following, we asses the performance of the constrained evolutionary optimization by means of computer simulations. To this end, we first consider a network without relays where $M = 15$ MSs are served directly by $B = 5$ BSs, all randomly located in a square area of $1000\,\text{m} \times 1000\,\text{m}$ according to a poisson point process (ppp) [15]. The BSs and the MSs are equipped with $N_\text{B} = 4$ and $N_\text{M} = 2$ antennas, respectively, and we apply a channel model with Rayleigh fading, pathloss, and shadowing according to the WINNER II model as in [7]. Assuming a system bandwidth of $100\,\text{MHz}$ and a noise figure of $5\,\text{dB}$, the noise variance at the MSs is $\sigma_w^2 = 5 \cdot 10^{-12}\,\text{W}$.

[2]Only the actual user data is included in the backhaul constraints. Additional traffic (e.g. control signals) can easily be included by an additional constant.

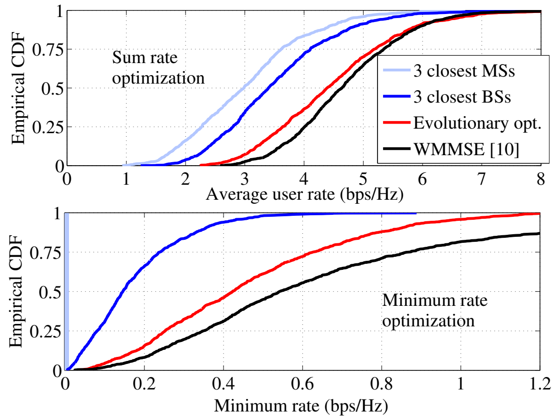

Fig. 2. Empirical CDFs of average user rates and minimum rates.

The maximal allowed transmit power at the BSs is set to $P_\text{B} = 40\,\text{W}$ which can be reduced by the optimization with a resolution of $N_\text{bits} = 32$.

Fig. 2 shows the empirical cumulative distribution functions (CDFs) of average and minimum user rates for different schemes. The evolutionary optimization is compared with the WMMSE approach from [11] and 2 static BS clustering schemes with SLNR beamforming (each MS chooses the three closest BSs and each BS serves the three closest MSs). It can be seen that the static schemes are clearly outperformed by the optimized ones, especially for the minimum rates. In this case, some MSs are not served at all ($f_\text{min} = 0$) when each BS only serves its three closest MSs. Even though the evolutionary algorithm is suboptimal, it is close to the WMMSE algorithm which is proven to converge to a local optimum. The evolutionary scheme, however, can be extended to much more general setups as shown in the following.

The outage performance for a target rate $R^\star = 1\,\text{bps/Hz}$ of the evolutionary scheme with fitness functions f_out and f_min is plotted in Fig. 3 versus different backhaul rate constraints C_b. It can be seen that the optimization is very robust with respect to limited backhaul rates, as the optimization can adapt the backhaul traffic by properly selecting the BSs for each MS. The performance saturates around $\frac{1}{3}$ because the chosen BS configuration can serve only 10 out of 15 MSs. For comparison, the performance of the reference schemes is also shown. Again, the evolutionary scheme outperforms the static ones and is close to the WMMSE algorithm. The latter, however, is not directly comparable, as we used f_min as utility because f_out is not applicable in this approach.

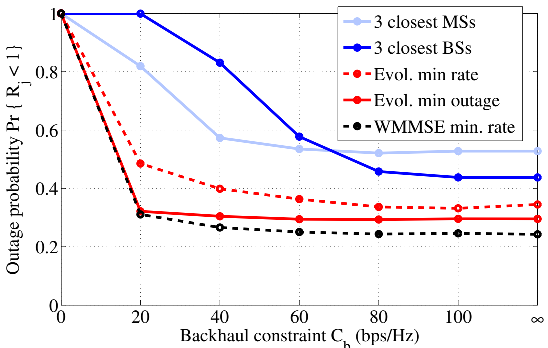

Fig. 3. The probability that a mobile user is in outage for different backhaul rate constraints.

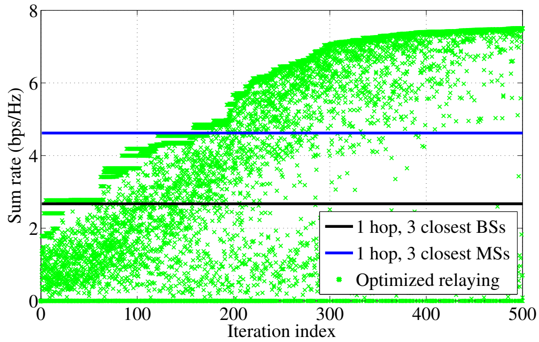

Fig. 4. Convergence behavior of the sum rate maximization for a typical network realization with relays.

In the following, we consider a relay assisted post-cellular network as depicted in Fig. 1. In this network, $B = 5$ BSs are randomly located in the first $500\,\mathrm{m}$ and serve $M = 15$ MSs in the rest of a $2000\,\mathrm{m}$ long network area (the width is $1000\,\mathrm{m}$ as before). $K = 30$ relays with $N_\mathrm{R} = 2$ antennas, also distributed in the same range as the MSs, assist the communication. MSs that are located in the same area as the BSs do not profit from two-hop communication due to the prelog factor $\frac{1}{2}$ that applies in the case of relaying and are better served by the BSs in a direct transmission, i.e. as in the previously discussed case. For such a two-hop network, we show the convergence behavior of the evolutionary algorithm in Fig. 4. The relay noise is $\sigma_n^2 = \sigma_w^2 = 5 \cdot 10^{-12}\,\mathrm{W}$ and the maximal relay amplification gain $\alpha_{k,\max}$ is chosen such that the average transmit power of each relay is $P_\mathrm{B}/10 = 4\,\mathrm{W}$. For the optimization, we use $N_\mathrm{ind} = 8$, $N_\mathrm{sur} = N_\mathrm{child} = 2$ and a bit flip probability update with $\tau = 0.95$. It can be seen that the algorithm converges after about 500 iterations, which is quite efficient for a network size of $5 \times 15 \times 30$ nodes.

The two different beamforming schemes for the two-hop communication (SLNR ignoring the second hop and SLNR with averaged second hop) are compared in Fig. 5 for different numbers of MSs ($M = 5, 10, 15$) and a varying number of relays. Thereby, the statistical CSI of the second hop leads to a considerable performance gain and the MSs can clearly benefit when more relays are present. The influence of power control (at BSs and relays) on the performance versus backhaul rate constraints at the $B = 5$ BSs is show in Fig. 6 for again $M = 15$ MSs and $K = 30$ relays. It can be observed that the first one or two bits have the strongest impact. Beyond 16 bits,

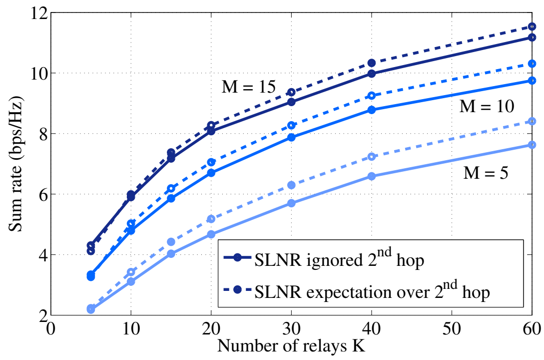

Fig. 5. Achievable sum rates for different network configurations with relays.

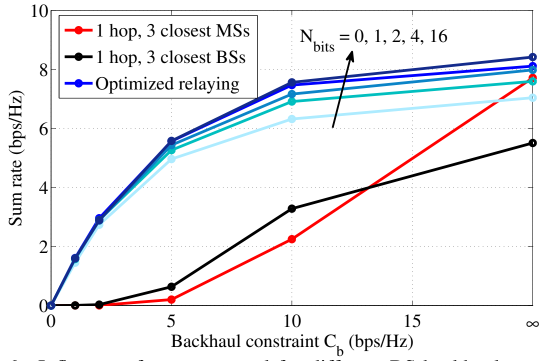

Fig. 6. Influence of power control for different BS backhaul constraints.

the performance gain is negligible.

The simulation results indicate that the evolutionary algorithm is an efficient tool to optimize the BS clustering and power allocation for both classical cellular one-hop networks as well as relay assisted post-cellular networks where the joint BS selection and relay routing is a difficult problem. The proposed scheme with extended two-hop SLNR beamforming is also very flexible with respect to practical considerations such as backhaul rate constraints and can adapt to changes in the network setup. The post-cellular network architecture with the help of many low-complexity relays can thus offer a dynamic and flexible solution for mobile communication networks of future generations.

References

[1] S.-E. Elayoubi, O. Ben Haddada, and B. Fouresti, "Performance evaluation of frequency planning schemes in OFDMA-based networks," *IEEE Trans. Wireless Comm.*, vol. 7, no. 5, pp. 1623–1633, May 2008.

[2] T. Nakamura *et al.*, "Trends in small cell enhancements in LTE-Advanced," *IEEE Comm. Mag.*, vol. 51, no. 2, pp. 98–105, Feb. 2013.

[3] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Comm. Mag.*, vol. 50, no. 6, pp. 54 –64, June 2012.

[4] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Comm.*, vol. 9, no. 11, pp. 3590–3600, 2010.

[5] D. Gesbert *et al.*, "Multi-cell MIMO cooperative networks: a new look at interference," *IEEE JSAC*, vol. 28, No. 9, Dec. 2010.

[6] J. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-Advanced systems," *IEEE Comm. Mag.*, vol. 50, no. 11, pp. 44 –50, Nov. 2012.

[7] R. Rolny, T. Rüegg, M. Kuhn, and A. Wittneben, "The cellular relay carpet: distributed cooperation with ubiquitous relaying," *Springer International Journal of Wireless Information Networks*, Jun. 2014.

[8] R. Rolny, C. Dünner, and A. Wittneben, "Power control for cellular networks with large antenna arrays and ubiquitous relaying," in *SPAWC*, Jun. 2014.

[9] I. Garcia, N. Kusashima, K. Sakaguchi, and K. Araki, "Dynamic cooperation set clustering on base station cooperation cellular networks," in *PIMRC*, Sep. 2010, pp. 2127 –2132.

[10] M. Hong *et al.*, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE JSAC*, vol. 31, no. 2, February 2013.

[11] B. Dai and W. Yu, "Sparse beamforming design for network MIMO system with per-base-station backhaul constraints," in *SPAWC*, Jun. 2014.

[12] M. Sanjabi, M. Hong, M. Razaviyayn, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission using statistical channel state information," in *SPAWC*, Jun. 2014.

[13] M. Sadek and S. Aissa, "Leakage based precoding for multi-user MIMO-OFDM systems," *IEEE Trans. Wireless Comm.*, vol. 10, no. 8, pp. 2428–2433, August 2011.

[14] K. Y. Lee and M. A. El-Sharkawi, *Modern Heuristic Optimization Techniques*, 1st ed. J. Wiley and Sons, Inc., 2008.

[15] R. Heath and M. Kountouris, "Modeling heterogeneous network interference," in *ITA*, Feb 2012, pp. 17–22.