

Technical Report: On design challenges of an endpoint traffic engineering service in a multi-provider wireless heterogeneous network

Jianwei Liu[†], Xin Xing*, Kang Chen*, and James Martin[†]

[†]School of Computing, Clemson University, Clemson, SC 29634

Email: {jianwel, jmarty}@clemson.edu

*Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL 62901

Email: {xin.xing, kchen}@siu.edu

Abstract—Resource allocation optimization is critical to the overall performance of wireless heterogeneous networks (HetNet). Prior research on this topic focuses on optimizing the user/flow associations in a single cellular heterogeneous network. Usually it tries to reach a global optimization objective in terms of aggregated throughput and fairness under simplified assumptions. However, recent development of network stacks supports seamless handover and concurrent usage of interfaces connecting to wireless networks run by multiple operators. This forms a new type of heterogeneous network, which we call it a multi-provider HetNet. In this paper, we would like to first explore some changes of the multi-provider HetNet compared with a traditional single-provider HetNet, and the implications to the design of a resource allocation framework for it. One thing we pay specific attention is which simplified assumptions commonly made in previous literature will not hold anymore in a MP-HetNet context. After analyzing the changes in the new context and their implications, we study how to design and implement a resource allocation framework by only controlling the flow association and rates at the endpoints of communication sessions. We call it an endpoint traffic engineering service for an MP-HetNet. Then we provide a general centralized design for this service, which does not assume any control to the underlying network infrastructure. With the problem and design in mind, we then explore the performance problem when extending the existing user association schemes in a single-provider context to the new design. Though various association schemes are proposed in the prior research under different contexts, few of them provide an in-depth evaluation to several fundamental problems that are required by the problem in single/multi-provider context, i.e. 1) the distances of the association schemes to the optimal solution under various scenarios; 2) the sources and impacts of the potential throughput estimation errors. By using relatively small scale and more controlled scenarios, this paper first time provides answers to the above questions, which are valuable for guiding further studies and real system designs.

I. INTRODUCTION

Resource allocation optimization of wireless systems usually involves maximizing an overall objective related to the throughput and fairness among users from a global perspective for a subset of the whole network. One way to implement this optimization is by controlling the user association to Access Point (AP) or Base Stations (BS) at the client devices (We will use AP for both AP and BS in this paper). A number of recent publications have identified, analyzed and proposed solutions for this type of optimization problem in

a single radio access technology (RAT) and single provider wireless heterogeneous network (HetNet). For example, [1] has identified the problem for a 3G wireless network where macrocells and picocells co-exist and overlap. We first use the networks M_1 and S_{1x} in Fig. 1 to illustrate this type of heterogeneous wireless network. The networks S_{11} and S_{12} in the figure are picocells, while M_1 is a macrocell. They are operated by the same provider and share the same gateway and backhaul network (G_1) towards the outside Internet. Every mobile node (MN) or user equipment (UE) connects to corresponding nodes (CN) via Internet. Given this model of a single-provider wireless HetNet, [1] identified the user association optimization problem which tries to achieve the *overall/generalized proportional fairness* (PF) in such a mixed network, when all the APs use PF schedulers. Fundamental to the modeling is the assumption that the generalized PF is equivalent to maximizing a utility function that requires estimates of user throughput[2]. The problem is NP-hard and the authors propose offline optimal solutions and online greedy heuristics. The work in [3] provides another centralized solution based on convex optimization, and a distributed scheme to approximate it. Both designed in the scope of a single cellular system, the solutions often assume separate controllers owned by providers, as shown in Fig. 1. Though preliminary results show the optimization using the generalized PF as the objective can improve spectral efficiency and fairness from a global perspective, some of the fundamental design issues for such an endpoint traffic engineering system in even a single-provider HetNet remains untouched in prior literature, i.e.,

Problem 1: whether the throughput estimation process required by the scheduling algorithms has errors, and what are the potential impacts of them to the system performance?

Problem 2: what is the impact of system parameters, such as control frequency to the performance?

Problem 3: what is the impact of the percentage of deployment (how much percentage of users are under the control of the designed endpoint traffic engineering system) to the wireless system performance;

Problem 4: what is the impact of the backlogged traffic assumption, which is commonly made in the previous literature, to the system performance.

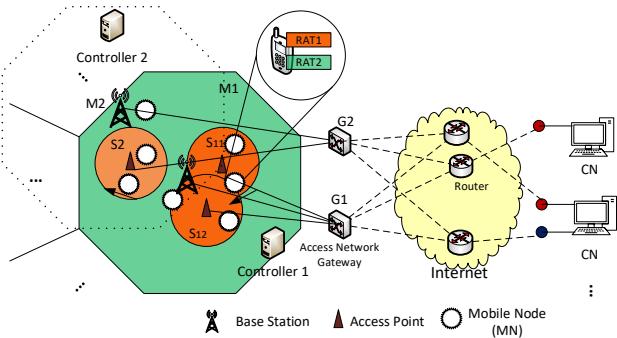


Fig. 1. System context.

Meanwhile, nowadays one location is usually covered with multiple wireless access networks, that belong to multiple providers, e.g., competitive commercial LTE networks and WiFi networks. The newly developed network stacks and applications ([4], [5]) can enable mobile applications to use one/more wireless interfaces. All the pieces exist to build large scale multi-provider heterogeneous networks (MP-HetNet) that support multi-homed mobile devices. Recent literature [6], [7] considers such systems. Fig. 1 illustrates a hypothetical MP-HetNet. Besides the networks S_{1x} and M_1 operated by one provider, we see the same location can be covered by macrocell M_2 and small cells S_2 , which are operated by another provider. Each UE owns multiple interfaces, and the applications on it can choose to use one/more RATs from a specific provider. In a single-provider HetNet, controller can reside in the backhaul network of the provider. However, in a multi-provider HetNet, the global controller in design should be operator independent, i.e. it should not reside in any operator's backhaul network, while being accessible by any mobile device. This new HetNet owns the following features, 1) the individual networks usually have different gateways and backhaul networks. For example, a Sprint LTE network and a campus WiFi will certainly have separated gateways and backhaul networks; 2) the individual APs of each network can use heterogeneous schedulers, or even sometimes unknown schedulers. For example, most of the previous literature in single-RAT HetNet assumes PF schedulers. However, WiFi networks usually use throughput fairness schedulers. Even though the four problems we identified in a single-provider context still apply in a multi-provider HetNet, because of these differences, we expect the four problems we summarized above will possess new characteristics and become more problematic.

- 1) Problem 1: 1) by aggregating multiple wireless interfaces, it is more likely devices will experience congestions at the backhaul network. Therefore, the errors in throughput estimation process could come from more sources, and become more obvious; 2) the heterogeneity of backhaul network make it impossible to assume homogeneous backhaul performance in terms of metrics such as RTT and loss rate, which are essential for end-to-end throughputs. In another words, the assumption in prior research will not hold anymore that the last hop wireless access

network is the only bottleneck. This will also add a new error source to the throughput estimation process.

- 2) Problem 2: Since the global optimization controller need to be independent of any provider, how to implement the control plane and select the right control frequency becomes more critical to system performance in a MP-HetNet.
- 3) Problem 3: Previous literature in the context of a single-provider HetNet usually assumes full deployment, because the control is usually implemented as admission control in the MAC layer. With the new MP-HetNet context and finer control granularities, it usually implies a higher layer control, which might not be fully deployed.
- 4) Problem 4: As stated in 1), aggregating multiple interfaces belonging to multiple RATs run by multiple operators increases the experienced bandwidth from application's perspective. Therefore, it is less likely that the application traffic is backlogged from the transport layer's perspective, as assumed by most of the prior research.

Also, other new problems could arise in the new context.

- 1) because of the more heterogeneous schedulers, the convex property and complexity of the optimization problem may change;
- 2) Because of the possibility for applications to use multiple interfaces concurrently, the modeling and solution of the problem may change. Further design of a flow association optimization framework in a MP-HetNet requires careful examination to these changes.

In general, we think of the traffic (under the granularity of user or flow) optimization frameworks in both types of wireless HetNets above an Over-The-Top (OTT) resource allocation network service for the corresponding HetNet environment. By OTT, we mean the design is against the previous resource allocation frameworks which assume controls/modifications to the network components, such as [8], [9], [10], [11]. For example, [8] assumes the control of the OFDMA resource allocation at the individual APs. Instead, an OTT design only assumes control to the ends of communication sessions. This paper generalizes and abstracts the problem, and provides a general and centralized solution. Then, we study some critical design issues for this type of optimization framework, such as the impact of various input errors. We are interested in the following specific scenario. **Similar to the previous literature [1], we focus on the optimization problem of downstream like video streaming from CNs to the mobile nodes, as this type of traffic takes up most of the Internet traffic.** Upstream is much more difficult when channel condition is used for throughput estimation as in [1], [3]. The scheduling scheme heterogeneity in an MP-HetNet adds more difficulty. Because some wireless networks have uncoupled uplink and downlink (like cellular networks), while the other networks may share resource across the up/down links (like 802.11n). In this case, the upstream traffic is similar to uncontrolled traffic in a traditional HetNet (Problem 3). For now, we assume downstream dominates in the scenarios we study. However, the upstream problem is not totally unsolvable under the framework we propose.

The measurement based throughput estimation method we proposed in [12] can help to alleviate these problems under our system design. For example, instead of fully relying on SINR to predict the throughput, historical average of measured throughput can reduce the scheduling algorithm's dependency to uplink SINR.

This paper focuses on the following questions that are valid in both HetNet environment, but that are likely more problematic in a multi-provider HetNet. 1) Even assuming an accurate throughput estimation model with no error, what are the distances of the previous association schemes to the optimal and status quo policy based association schemes, in terms of the PF objective function value and other global performance metrics? 2) Whether the input errors exist; and what are their impacts to the performance of the solutions?

The contributions of this paper include, 1) first time identifying the OTT resource allocation problem in a MP-HetNet, and why OTT is needed in the MP-HetNet; 2) providing a centralized OTT resource allocation framework design; 3) designing new methodology for an in-depth examination of the performance distance of previous schemes to the status quo scheme and the optimal solution; 4) testing various methods' sensitivities to the input errors and system parameter selection. The results can serve as guidance for the design and implementation of a real OTT resource allocation framework.

This paper is organized as follows. We review the previous literature in section II. Then we present a generalized mathematical model of the existing association schemes in section III. To solve a similar problem in an MP-HetNet, we present an OTT system design in section IV. In section V, we design new methodology to examine the performance distance of the previous schemes to the policy-based scheme used in real system, and the optimal solution. In section VI, we explore the causes of the input error to scheduling algorithms and prove its existence by simulation. We then test the sensitivities of various schemes to general errors and error from control frequency. Finally, we make our conclusions in section VII.

II. RELATED WORKS

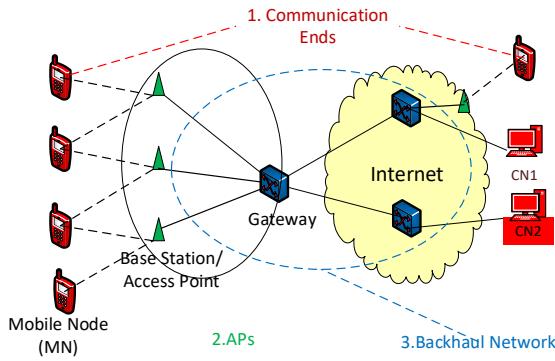


Fig. 2. Different types of resource allocation problems.

From an abstract perspective, a wireless system, as shown in Fig. 2, consists of three components,

- 1) Communication ends - at least one end is wireless.

- 2) Access Points (AP) - can be of the same or different radio access technologies (RAT) and operators. This paper will use the term AP for last hop access nodes, such as APs in WiFi and Base Stations in LTE.

- 3) Backhaul nodes - routers, etc.

The most general form of the resource allocation problem for a wireless system can control the resource allocation/usage planning all nodes belonging to the three categories mentioned above with any granularity. The controls at the communication ends can involve controlling user associations and data sending rates at the end-point client devices, as shown in [1], [6], [13]. The controls at the APs can include the innovations of the scheduling algorithms at the APs, as demonstrated in [8], [11]. The control at backhaul network usually refers to the resource allocation and flow scheduling study at the backhaul routers [14]. This paper focuses on the first route, i.e. an OTT style network service design, which only controls the communication ends. It is a natural requirement from the MP-HetNet context.

Though there are a lot of previous literature on HetNet resource allocation problem [1], [3], [15], [6], [13], [9], [8], and some of them even of OTT type [1], [3], [6], [13], there is no good modeling work that research on some of the fundamental problems in such an OTT type flow association service.

In general, these works use the generalized proportional fairness as the overall optimization objective, and focus on backlogged downlink traffic. Most of them assumes the individual schedulers use proportional fairness. Since the problem is proved to be NP-hard [6], [1], greedy heuristic and approximation dominate the prior research. However, they fail to provide a detailed comparison to the current policy based scheme used on smart phones, and the optimal solution. Also, there is no existing report of the comparison among the previous schemes. This paper tries to provide a more in-depth view of the problem itself, the comparison of different methods, and the impact of various system dynamics and parameter selections.

III. GENERALIZED ABSTRACT MODEL OF THE RESOURCE ALLOCATION PROBLEM IN THE PREVIOUS LITERATURE

The study in this paper is based on the following mathematical model, which is similar to the ones in [1], [3], [13], [6]. The models in the previous literature in general assumes, 1) generalized proportional fairness as the overall optimization goal, as PF is considered a simple and fair trade-off between the aggregated throughput and fairness [1], [2], [16]; 2) backlogged traffic or infinite application layer demand; 3) rough control granularity of association control only.

$$\begin{aligned}
 & \text{Maximize} \quad \sum_{j=1 \dots M} \sum_{i=1 \dots N} \log(T_{ij}) * x_{ij} \\
 & \text{subject to} \\
 & \quad \sum_j x_{ij} = 1, \\
 & \quad T_{ij} = \mathcal{U}(\hat{x}, \dots), \\
 & \quad x_{ij} \in \{0, 1\}
 \end{aligned} \tag{1}$$

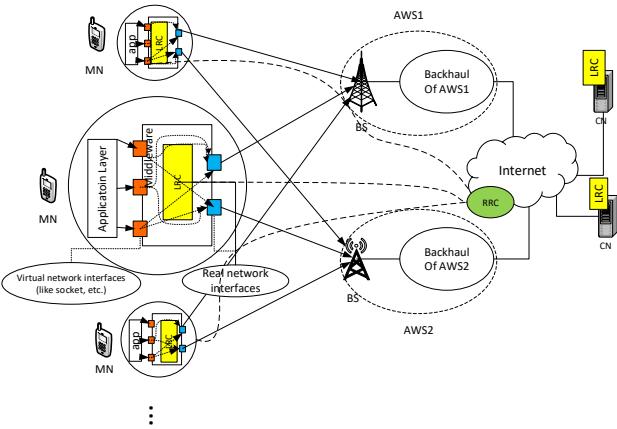


Fig. 3. A generalized OTT architecture for the flow association in an MP-HetNet.

The index i is for UE, while j is for AP. N is the total number of UEs, while M is that for APs. x_{ij} is the association variable that represents whether the traffic of user i should be received from AP j . It is either 0 or 1, which is called *integral solution*. The solution to the above problem is a matrix of x_{ij} (\hat{x}), which is the flow association guideline provided to UEs by the framework. The first constraint means every UE can only connect to exactly one interface. T_{ij} is the end-to-end throughput of UE i if it is connected to AP j . We purposely do not give the detailed definition of T_{ij} here, because it is dependent to the assumptions on the scheduling schemes used on the individual APs. We will details the models in V. The second constraint shows that the end-to-end throughput is a function that involves the association plan \hat{x} , and other factors. Those can include the signal-to-noise-ratio of a link and the other internal node properties, like loss rate of a router in the path. Note that this function \mathcal{U} abstracts the individual scheduling scheme used by every AP. This model is similar to the model in Eq. (1) - (4) in [1], but with a more generalized form.

IV. OTT SYSTEM DESIGN

To meet the OTT requirement of the above MP-HetNet, we assume the following general conceptual system architecture and components in the solutions we discuss below. As shown in Fig. 3, a Local Resource Controller (LRC) is located at every end device, i.e. a User Equipment (UE) or a remote server. It collects network connection status information from the device and relays it to the Regional Resource Controller (RRC). The RRC then uses the aggregated information to model the throughput of every client device under various planning choices. Centralized algorithm running at the RRC then produces a solution for how every flow should be associated so that system-wise performance objective can be optimized. Every LRC receives the plan periodically and enforces it locally. Though we present the RRC as a centralized service, its storage and computation can be distributed. It is only conceptually a centralized service.

The abstracted data flow of the system is shown in Fig. 4. Clients first send flow information measured from the

two ends of communication sessions to the RRC. The RRC then conducts resource allocation algorithm which invokes the throughput estimation. It is because the scheduling algorithms require estimated throughputs to judge which association policy is better. The figure shows the relation of different modules, and why throughput estimation is important.

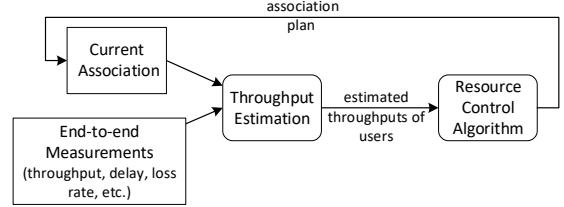


Fig. 4. Work flow of the system.

V. BASELINE STUDY: PERFORMANCE OF THE EXISTING USER ASSOCIATION SCHEMES

Given the above optimization problem in Eq. 1 and the system design in section IV, the first question is, how well will the scheduling algorithms in the previous work behave when compared with the status quo association scheme and the optimal solution. Specifically, we are interested in the scenarios where all the scheduling algorithms have a perfect input, i.e. 100% accurate throughput estimation (which is generally assumed in the previous literature). In this section, we try to provide these information using a representative set of previous methods and a new methodology.

First, we set up a scenario that has N UEs and M APs. We test with $N = \{5, 10\}$ and $M = 3$. The scales of M and N are purposely kept small so that the optimal solution is tractable. To make the percentage of UEs with good connection to each AP more controllable, we did not use random mobility as in previous literature [1], [6]. Instead, we control the percentage of users connecting to each AP. One AP with large coverage (like a LTE macrocell) and several APs with smaller coverage (like a WiFi AP) coexist. We control the percentage of UEs under the small cells (P_s). The P_s values we tested are $\{0.8, 0.6, 0.4, 0.2\}$. For each time slot, we generate the distance to APs randomly following the P_s . We first show the result of $P_s = 0.8$ below. For results of more P_s values, please refer to Section VIII-C.

Also, we test with two cases concerning the type of APs:

Case 1. All proportional fair APs. (as in [1], [3], shorten for 'pf only' in the figures).

This assumption actually has significant implication to the modeling. With the former, the throughput T_{ij} in Eq. 1 is usually modeled as follows,

$$T_{ij}(\hat{x}) = \frac{r_{ij} * G}{\sum x_{ij}}$$

where x_{ij} is still the association variable of the traffic of user i to AP j . The matrix \hat{x} represents an association plan, with element x_{ij} the association variable as in Eq. 1 (For now, we consider intergral association, i.e. $x_{ij} = \{0, 1\}$). Note T_{ij} is a function of \hat{x} . r_{ij} is the maximum channel rate of user i

given its connection status. G is a constant that counts for the multi-user gain and overhead from various layers.

We can see the function of $T_{ij}(x)$ is strictly concave (the detailed proof is in Appendix VIII-B), and the whole objective function is strictly convex after applying the logarithm function. Therefore, we can use convex solvers (e.g. the interior point method solver in MATLAB) to solve the problem easily.

Case 2. A mix of proportional fair and throughput fair APs. (as in [13], [15], shorten for 'both' in the figures).

The throughput fairness in [13], [15] are defined like the following. It is based on the observations in [17], [18], i.e. every station get similar throughput in the saturated case. They model the 802.11 (b) station throughput like the following,

$$T_{ij} = \frac{G * p_{ij}}{T_{cycle}}$$

where p_{ij} is the priority of a flow, which in our case should always be the same for each flow. T_{cycle} is the the average time between two consecutive transmissions of the station, and G is a constact factor counting for priority of flows and overhead in various layers, and

$$T_{cycle} = \sum \frac{p_{ij}}{r_{ij}}$$

In Case 2, with the added complexity of throughput fairness scheduling APs, the T_{ij} is not concave anymore, rendering the convex solvers unusable. We will show its impact to the convex solver based methods (the first centralized method in [3]) in the results below. In the simulation, we use the distance to throughput mapping we measured from NS3 [19] for the peak rates before resource divisions. The detailed mapping is provided in the appendix VIII-A.

We compare the following five methods.

- 1) policy-based: This represents the interface selection scheme on most of smart phones, i.e. when WiFi is available, connect to the WiFi with the best signal; otherwise, connect to the cellular network.
- 2) optimal-brute-force: An optimal solution using a brute-force method which iterates through all the possible user association configurations, and returns the one with the largest objective function value.
- 3) ATOM: Greedy algorithm generalized from the Algorithm 1 in [13], which is shown in the Algorithm 1 below. We generalize the algorithm in [13] to m LTE BSs instead of a single BS, and select k WiFi APs to offload instead of one. The basic idea is that every macrocell forms a set containing UEs that can connect to it. For each set s , we first initialize all the UEs in it to the macrocell, and then try to offload them to k WiFi APs. From the n^k cases, we select the one that results in the maximum objective function value, and finalize the UEs to the corresponding small cells. The algorithm repeats this process until all the sets are checked.
- 4) random: Returning the best solution in 5 randomly generated user association configurations.

- 5) round-off-interior-point: Solving the problem using a non-linear solver (like *fmincon* in MATLAB), and return the round-off integral solution.

Algorithm 1: Generalized greedy heuristic.

```

1 for  $u$  in  $\mathbb{U}$  do
2   | if  $R_{ij} > 0$  then
3   |   |  $u$  into set  $S_m$ ;
4   | end
5 end
6 for  $s \in S$  do
7   | select  $k$  small cells and try to offload UEs in  $s$  to the small cells ( $n^k$ 
     | scenarios in total);
8   | select the scenario above that results in the maximum object value, and
9 end

```

We compare the following three metrics,

- 1) PF value: The objective function value in Eq. 1.
- 2) Spectral efficiency: Sum of all the user throughputs divided by the overall bandwidth.
- 3) Jain's fairness metric: $J(T_1, T_2, \dots, T_n) = \frac{(\sum_{i=1 \dots n} \frac{T_i}{Tb_i})^2}{n * \sum_{i=1 \dots n} (\frac{T_i}{Tb_i})^2}$, where T_i is the throughput of user i , and Tb_i is the throughput of user i of the optimal global proportional fairness solution (e.g. the one generated by the brute-force method below).

Fig. 5 shows the results for the Case 1, i.e. only PF scheduling APs, while Fig. 6 is the results for the Case 2, i.e. both PF scheduling APs and throughput fairness APs.

We have the following observations from these figures,

- (1) Concerning PF value, optimal-brute-force always produces the best PF value for both cases, which serves as the baseline and a sanity check. In Case 1, round-off-int has close-to-optimal performance under Case 1 most of time. However, in some rare cases the round-off can make way worse configuration that results in only about half the optimal PF value and aggregated throughput, which are even worse than those of the policy based and random. This is perhaps a result of local optimality.
- (2) In both cases, ATOM is the second-closest to the optimal, and with lower standard deviation compared with random and policy-based. This means, though not optimal, the heuristics like ATOM work pretty well. However, ATOM has larger standard deviations in terms of both aggregated throughput and Jain's fairness index, compared with the optimal, even though having a similar PF value.
- (3) The methods have similar Jain's fairness index performance under the Case 2. But in Case 1, the layers among different methods are relatively clear. We can see basically optimal \approx round-off-int \leq policy \leq ATOM \leq random. The optimal also has larger standard deviation in terms of Jain's fairness index under Case 2 compared with Case 1.
- (4) The random method has the worst performance and the largest deviation most of time. The policy based method is only slightly better than the random generated configurations.

The result in this section shows, even with perfect information, the current algorithms will have some distance to the optimal global PF objective. This has very important

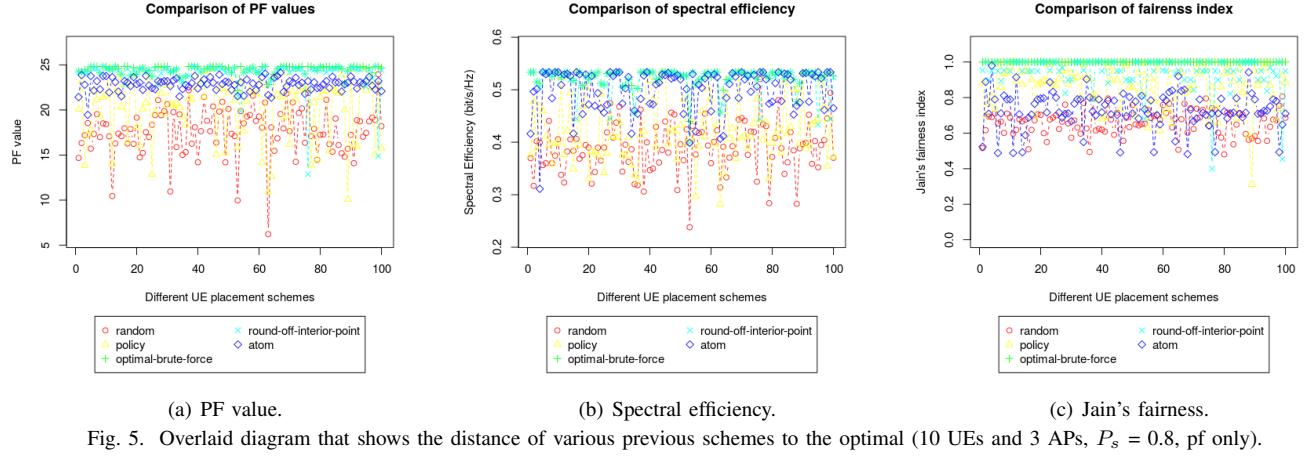


Fig. 5. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.8$, pf only).

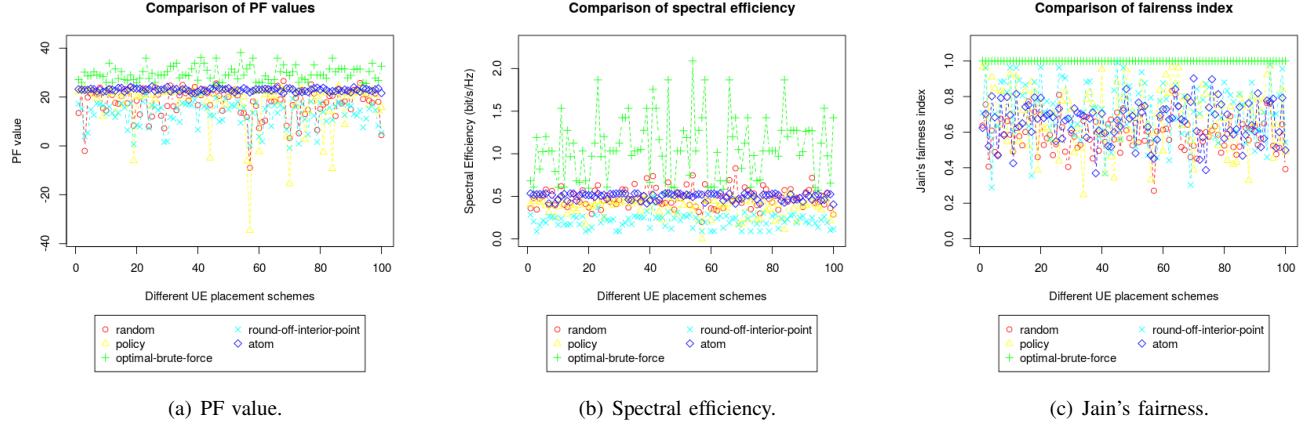


Fig. 6. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.8$, both).

implication for the study below about the impact of the input error to the algorithms, i.e. when estimating the impact of input error in the section VI, we need to start with the optimal-brute-force solution. Because only the results of the optimal solution have no influence from the error introduced by algorithms.

VI. SOURCES AND IMPACTS OF INPUT ERRORS

In the last section, we use a similar assumption as those in the previous literature, i.e. there is an explicit formula that can estimate user throughput accurately. For example, in [3], they use the Shannon equation. However, this assumption is invalid in real systems. In general, the throughput estimation errors can come from,

- 1) **Type I:** throughput model, which includes,
 - a) capacity and individual scheduler modeling error;
 - b) failing to consider application demands in the model;
 - c) failing to include impacts from the backhaul networks.
- 2) **Type II:** the change of inputs (e.g. connection status and association information) between sampling when the plan is enforced at the LRC. It consists of two phases, i.e. 1) from sampling to RRC scheduling 2) from RRC scheduling to local policy enforcement.

A. existence of errors

1) Type I (a) error: We first show why the rough throughput estimation model based on Shannon capacity equation (like in [3]) will introduce Type I error. We use the following simulation in NS3 to demonstrate the concept. We the error rates of the following methods when estimating one UE under a LTE Base Station (BS), and then compared with the TCP and UDP goodput measured by *iperf*. During the simulation, we position the UE at various distances to the BS. We tested with two cases, i.e. 1) no fading; 2) with a more realistic fading model based on the Annex B.2 of 3GPP TS36.104 [20]. We used pedestrian model with a speed of 3kmph.

- 1) Shannon Equation: $T = B * \log(1 + SNR)$, where B is the bandwidth of the spectrum in use, and T is the estimated user throughput. $SNR = \frac{SP}{NP}$, where SP is the power of signal and NP is the power of noise (in watt).
- 2) Modulation and Coding Scheme (MCS). For example, if 64QAM is used, every symbol has 6bits. If c symbols can be supported for a 10MHz channel, the prediction result will be $6c$ bps.
- 3) Nyquist Equation: $T = 2B * \log_2(Nbits)$, Where $Nbits$ is the number of bits used for the coding scheme.
- 4) Transport Block (TB) size. A MAC layer rate limiter in LTE, which sets an upper bound for the maximum bytes one UE can transmit given a specific MCS and number

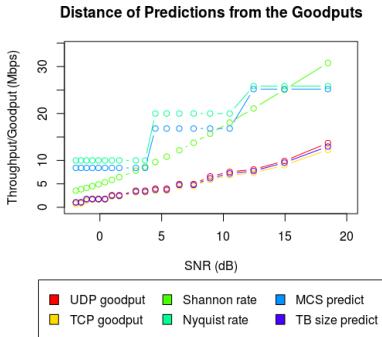


Fig. 7. Overlaid diagram that shows the distance of estimation methods to the application layer goodputs.

of resource blocks. The TB size is based on the table in 3GPP standard TS36.101, Annex A.2.1.2 [21].

Fig. 7 shows the distance of the raw results using the above throughput estimation methods to the goodputs. Fig. 8(a) shows the best curve fitting of SNR based estimation curves to the TCP goodput when no fading added; while Fig. 8(b) shows that when the fading model added.

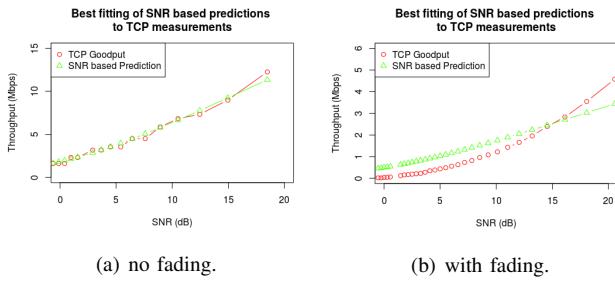


Fig. 8. Best curve fitting of SNR based estimation to the TCP goodput.

From the results, we can see that the throughput estimation, even with the best fitting parameters, can deviate from the application layer goodput, especially with the case of realistic fading. More importantly, the directions of this deviation is random. Therefore, in real system, the errors can not be easily offset by a constant factor. Therefore, this at least proves that the Type I throughput estimation error exists using the previous SNR based throughput estimation methods.

2) *Type I (b) & (c)*: In real world, applications are not always backlogged. When we browse websites, the traffic will not always satuate the wireless capacity. Even the video streaming session usually have an upper bound due to the resolution of the video, which can be less than some of the newer wireless connection capacity like LTE-A and the newer 802.11 wireless networks. For example, all the major wireless operators have introduced gigabit LTE. The current LTE-A network in use in US is reported to have a peak downlink rate of 200Mbps in 2017.[22] However, even a 4K video usually only require a bandwidth of 50Mbps. We can expect much higher mobile wireless bandwidth in the near future. [23]

For Type I(c), it is clear that the delay and loss in the backhaul network will have significant impact on at least TCP sessions.

3) *Type II*: This is clearly an issue, as the mobility and fading of wireless signal will always make the sampling inaccurate to some distance.

B. impact of of errors

Meanwhile, as shown in Fig. 4, the throughput estimation model provides input to the resource control (a generalization of the controls to association and rate) algorithm. However, the rough throughput estimation models used in the previous literature inevitably introduce errors to this input. None of the prior work studies the sensitivity of their algorithms to the input errors, i.e., assuming the throughput model has a 100% accuracy in simulations.

We first use the following simple example to illustrate the impact of throughput estimation error to the system performance. Like shown in Fig. 9, there are two APs with overlapping area. Both APs use proportional fairness schedulers. The white circle represents UEs, while the number in it represents the total of UEs. As we can see, the current association plan has 10 users connecting to the AP1 (Macro), while 40 users to the AP2 (small). Assume each AP has a total capacity of 20Mbps, and all the UEs has the same connection status to the AP it should connect to based on the association plan. If the 10 UEs connecting to the AP1 were estimated to have a throughput of 2Mbps per user. But actually each UE has a throughput of 0.1Mbps instead due to any kind of errors or their combinations. The neighboring AP2 has 40 users with an estimated throughput of 0.5M, but a real demand of 6Mbps per user. Each UE on AP2 can only reach 0.5Mbps. The scheduler thought the overall throughput is 20M + 20M = 40M, but actually it is only 1M + 20M = 21M. Moving some of the 6Mbps users to the first AP can definitely better balance the system and achieve better proportional fairness utility, but scheduler would fail to do so because the input throughput estimation errors. From the above results, at least intuitively the throughput estimation accuracy matters for the outcome of overall throughput. In this section, we will try to inspect the impact of it to various scheduling algorithms in details.

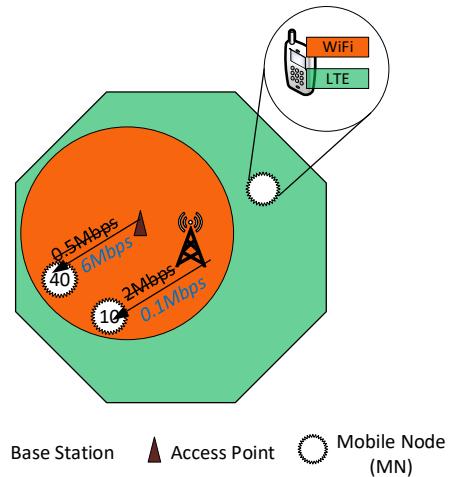


Fig. 9. Simple example showing why the throughput estimation accuracy matters.

Therefore, in this section, we study the sensitivities of some representative user association schemes to various levels of input errors. We would select the same set of algorithms as in section V. However, the problem is invalid for the policy based and the random methods, since they do not use throughput estimation. Thus, we only compare three schemes in this section, i.e. 1) optimal-brute-force; 2) ATOM; 3) round-off-interior-point. The subsection VI-A1 deals with the Type I error, while the subsection VI-C builds simplified model to test the impact of the Type II errors.

In this section, we test the sensitivities of the previous user flow association schemes to the input errors. We set up a similar scenario like in section V. However, this time we purposely insert errors to the inputs of the algorithms, i.e. the estimated user throughput (T). The new throughput after error insertion $T' = (1 \pm e) * T$, where e is the error rate. We use $e = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ in our evaluation. Every experiment lasts for 5000 time slots. Fig. 10 shows the results for the method optimal-brute-force. The figures plot the Cumulative Density Function (CDF) of the performance metric change rate. It is defined as $\frac{S_e - S_o}{S_o} \times 100\%$, where S_o is the original performance metric without errors, and S_e the one with errors. The x axis is the performance metric change rate, while the y axis is the CDF value. We first show the impact to the optimal-brute-force methods. Because all the other methods' results contain the errors from the method itself, as demonstrated in the Figs. 5 and 6. We provide the results of the other methods and more AP types (Case 1 and 2) in the Section VIII-D, and more P_s values in ?? From Fig. 10, we observe that,

- (1) For the optimal solution, the input error will always lead to degraded PF objective function value (shown in Fig. 21(a)). Meanwhile, the percentage of performance degradation is nearly proportional to the error rate inserted. With an error rate of 0.5, the PF value can degrade up to 50%.
- (2) The system performance in terms of both aggregated throughput can get better or worse. Actually, for methods except the optimal-brute-force method(in ??), the Jain's fairness index performance also has the same behavior when errors are inserted. This is because the proportional fairness is only a tradeoff between the two optimization objectives. There is still space that either or both metrics can be improved [16]. This is the artifact of the proportional fairness itself. However, as a widely used and simple resource allocation optimization objective, this paper is more interested in how it behaves, instead of whether a better objective can be used.
- (3) From Fig. 21(b), we observe that, for the aggregated throughput, the CDF curves are biased to the left of x=0. This means, when errors inserted, there is high possibility to get a performance degradation in terms of aggregated throughput. With an error rate of 0.5, the aggregated throughput can degrade up to 80%.
- (4) Fig. 21(c) demonstrates the performance of Jain's fairness will always degrade when error inserted. This is mainly because the definition of it always targets to the optimal solution. With an error rate of 0.5, the Jain's fairness index

can degrade up to 60%.

To observe the detailed impact of the input errors to the aggregated throughput and fairness respectively, we overlay the changes of both throughput and Jain's fairness metric onto the same figure. Making the figure readable, we sample the first 50 timeslots from the 5000 timeslots run. Fig. 11 shows the results of optimal-brute-force under various input error rates. For the results of the other methods, please refer to the Section VIII-E. In the figures, the x axis is the time sequence; while the y axis means the percentage of change of the two different metrics. If the aggregated throughput was positive, it is drawn in blue box; while in gray box if negative. If Jain's fairness metric was positive, it is drawn in red box; while white box if negative.

C. Type II errors

We simulate and verify the impact of the Type II & III errors in the following way. Assuming an atomic time unit (t_a) of SNR change, we vary the control time interval (t_c) as the multiples of this time unit. The multiples α we tried are $\{1, 3, 7, 15, 31\}$. We continue to use $N=5$, $M=3$, and the WiFi coverage rate $P_s=0.8$. For the results with more P_s values, please refer to the appendix in the full version of the paper [24]. We test with SNR change rate (C_r)= $\{0.1, 0.3, 0.5\}$. $C_r = \frac{SNR_t}{SNR_{t-1}}$ where t represents the current time slot. Figs. 39 shows the results for two methods under the Case 1 APs. From the results, we have the following observations,

- 1) When the control time interval is larger than 10 times of t_a ($\alpha = \{15, 31\}$), the optimization objective value starts to degrade. For the scenarios with a smaller t_c , the performance degradation is negligible. This implies, in real system design and implementation, we most likely want to use control frequency that is less than 10 times of average SNR change frequency, and be cautious about the opposite case.
- 2) The impact of control frequency is also related to the SNR change rate. For example, for SNR change rate of 10%, larger control interval will have little impact. However, when the SNR change rate increased to 0.3 and 0.5, the degradations are quite obvious.

VII. CONCLUSION

From the analysis and simulation results, we can see the input throughput error exists, which is more problematic in a MP-HetNet. We also observe that, directly applying the existing user association schemes in SP-HetNets to the new context can result in performance degradation and large deviations. We also tested the sensitivities of the existing schemes to the errors introduced by system control frequency. From the results, we learn the range of usable control frequencies without drastic performance degradation. These insights can serve as the guidance for future system design and implementation for a real OTT optimization framework. For example, knowing the existence and impacts of input errors, we can use measurement based method to reduce the throughput estimation error as we proposed in [12]. In the future, we would like to investigate the scenarios when multiple interfaces can be simultaneously

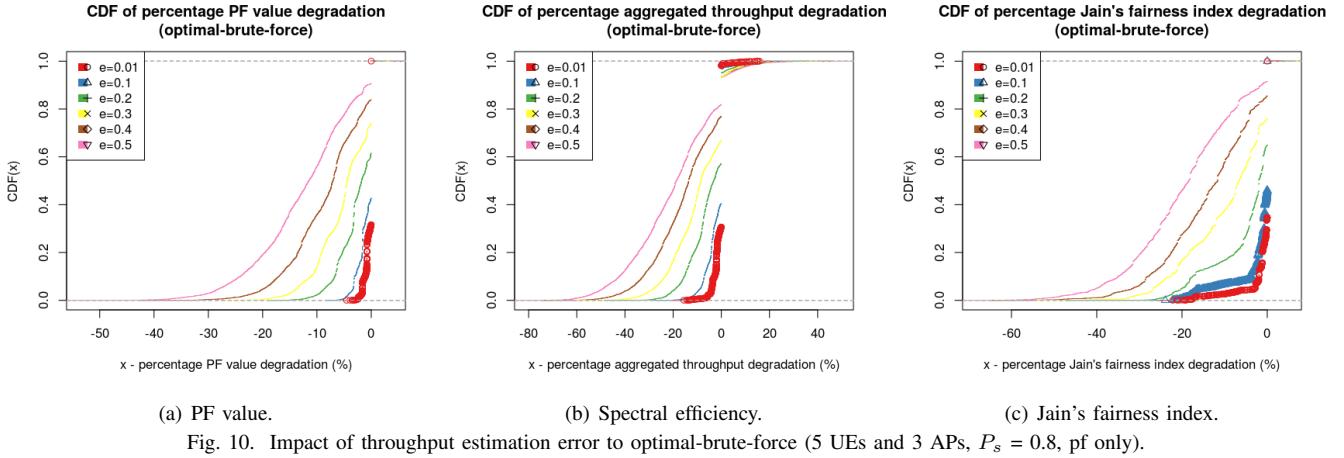


Fig. 10. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.8$, pf only).

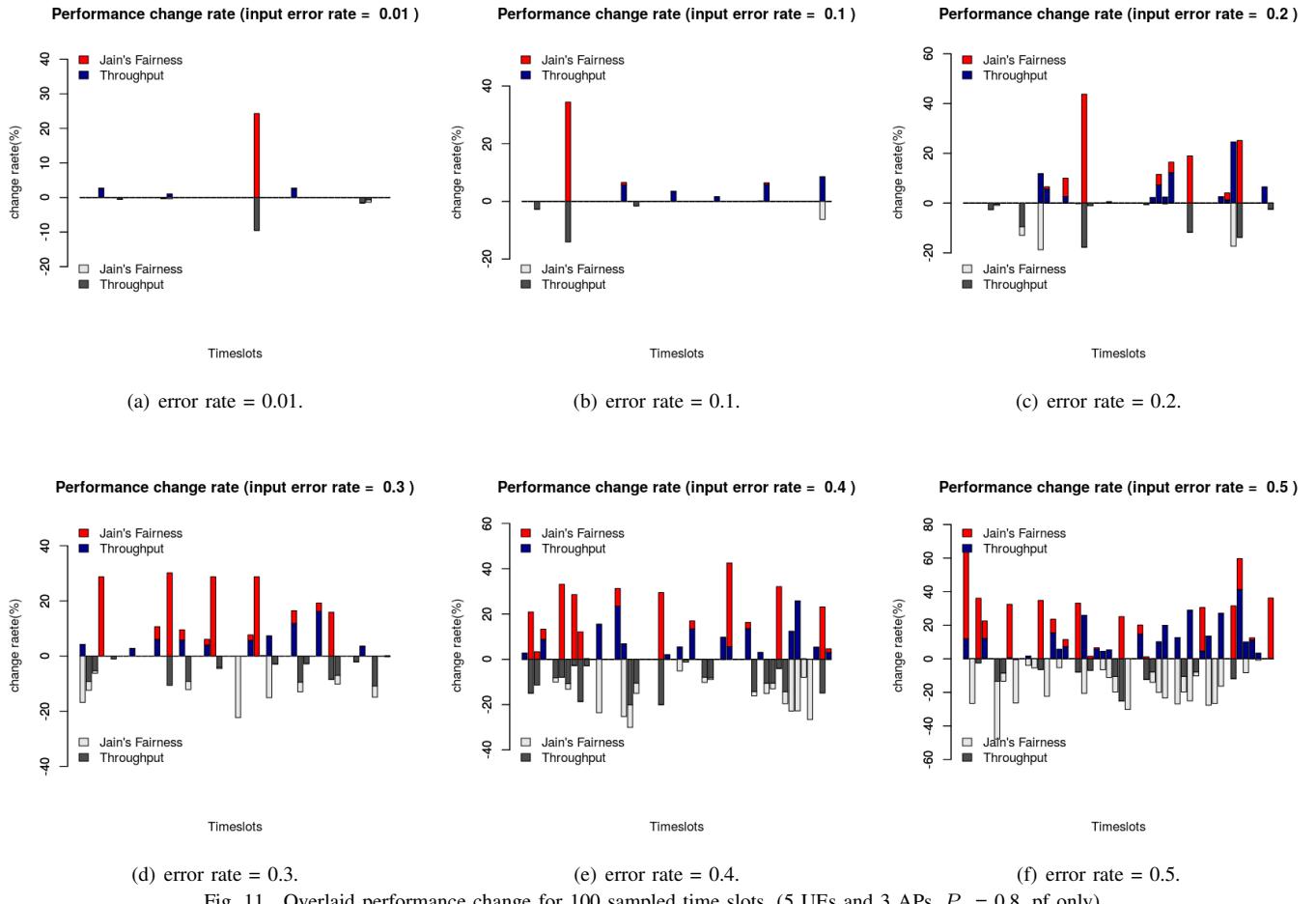
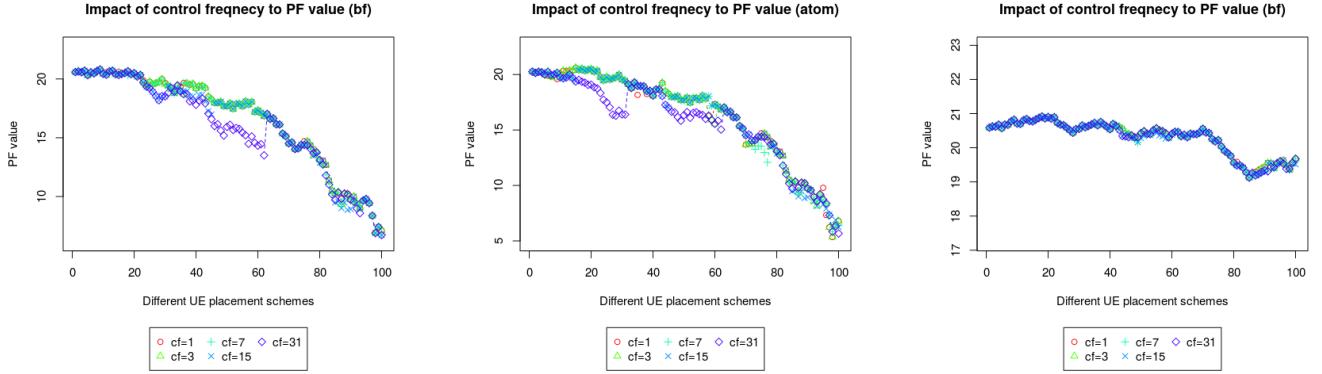


Fig. 11. Overlaid performance change for 100 sampled time slots. (5 UEs and 3 APs, $P_s = 0.8$, pf only).

(a) $c = 30\%$ optimal-brute-force(b) $c = 30\%$ atom(c) $c = 10\%$ optimal-brute-forceFig. 12. Impact of control frequency to the PF value of various methods (5 UEs and 3 APs, $P_s = 0.8$, pf only).

utilized by UEs, and the scenarios with various control granularities.

VIII. APPENDIX

A. Rate mapping used in simulation

For the macrocell, we use the LTE rate tested in NS3 using one UE and one macrocell. The following table gives the rate we tested and used.

<i>Distance(m)</i>	<i>UDP Goodput(mbps)</i>
2000	13.718178
3000	9.884191
4000	8.075719125
5000	7.559518
6000	6.603277375
7000	4.968599375
8000	4.968599375
9000	3.9995205
10000	3.9995205
11000	3.554007375
12000	3.554007375
14000	2.5937585
15000	2.5937585
16000	1.814277875
17000	1.814277875
18000	1.814277875
19000	1.761634
20000	1.0840515
21000	1.03905825

For the small cell, we use the rate tested using the WiFi module in NS3.

<i>Distance(m)</i>	<i>UDP Throughput(mbps)</i>
5	9.99916
10	10.0008
15	9.99916
20	10.0008
25	9.99916
30	9.99916
35	9.99916
40	10.0008
45	9.99916
50	10.0008
55	9.99916
60	10.0008
65	9.99916
70	10.0008
75	8.74086
80	8.86047
85	8.63764
90	5.57384
95	4.88899
100	4.92995
105	4.90701
110	4.42368
115	1.85631
120	0.01638
125	0

B. Proof of convex property of traditional throughput model

The tradition throughput is like the following (the constant is removed),

$$f(x) = \frac{1}{x_1 + x_2 + \dots + x_k}$$

we want to get the relation between $af(x) + bf(y)$ and $f(ax + by)$, where $0 \leq a \leq 1$ and $0 \leq b \leq 1$. If we make $X = x_1 + x_2 + \dots + x_k$, and $Y = y_1 + y_2 + \dots + y_k$, we need to get relation of $\frac{a}{X} + \frac{b}{Y}$ and $\frac{1}{aX + bY}$. If we assume the function is convex, we only need to prove,

$$\frac{a}{X} + \frac{b}{Y} \geq \frac{1}{aX + bY}$$

$$\frac{aY + bX}{XY} \geq \frac{1}{aX + bY} \quad (2)$$

$$(aY + bX)(aX + bY) \geq XY \quad (3)$$

$$a^2XY + abY^2 + abX^2 + b^2XY \geq XY \quad (4)$$

$$(a + b)^2Y - 2abXY + ab(X^2 + Y^2) \geq XY \quad (5)$$

$$(6)$$

since $(a+b) = 1$, we only need to prove,

$$ab(X^2 + Y^2 - 2XY) \geq 0 \quad (7)$$

$$ab(X - Y)^2 \geq 0 \quad (8)$$

Therefore, it is proved.

C. More P_s values for the baseline distance study in Section V

Fig. 13 - Fig. 18 show the results with $P_s = \{0.6, 0.4, 0.2\}$.

D. Results for more methods and AP types of the study in Section VI-B

1) *Case 1: pf only:* Fig. 19-Fig. 21 show the impact of errors to the optimal-brute-force method. Fig. 22-Fig. 24 show that of the atom method. Fig. 25-Fig. 27 show that to the round-off-int method.

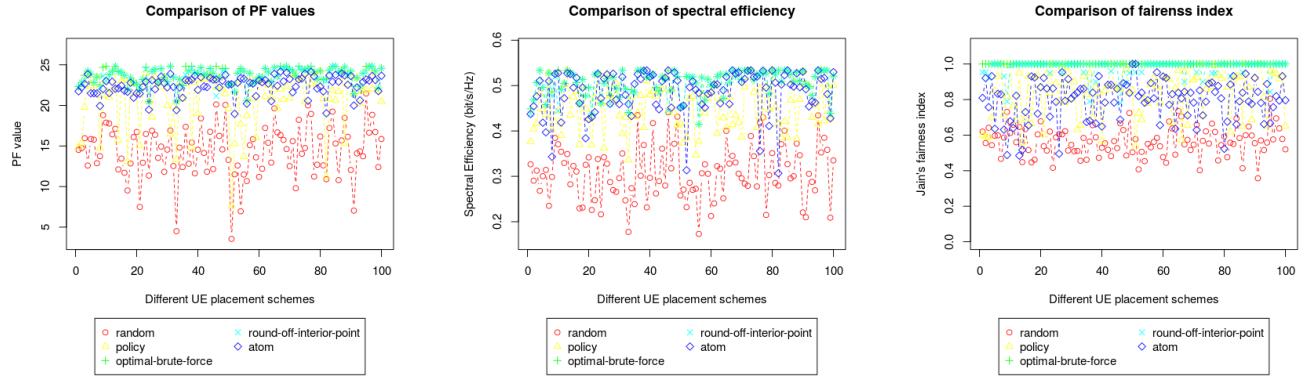


Fig. 13. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.6$, pf only).

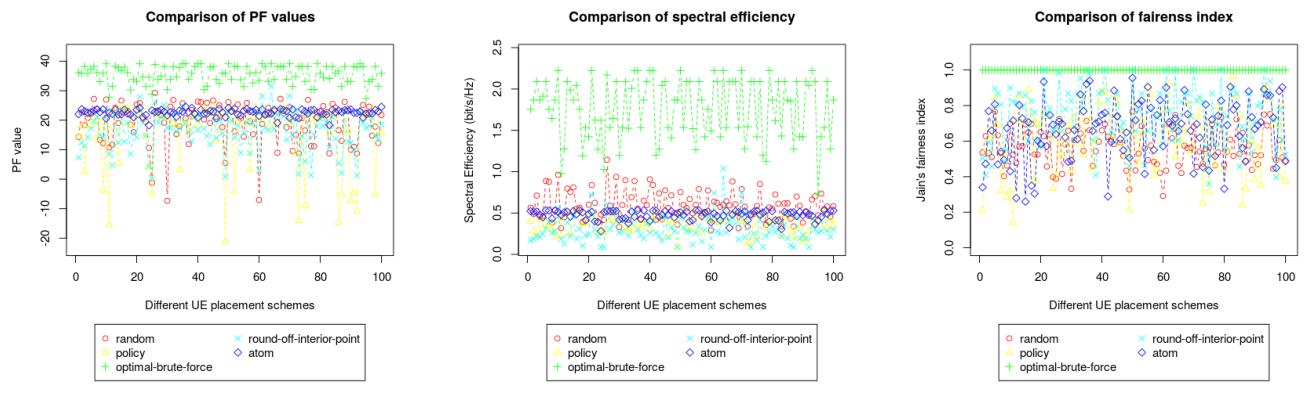


Fig. 14. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.6$, both).

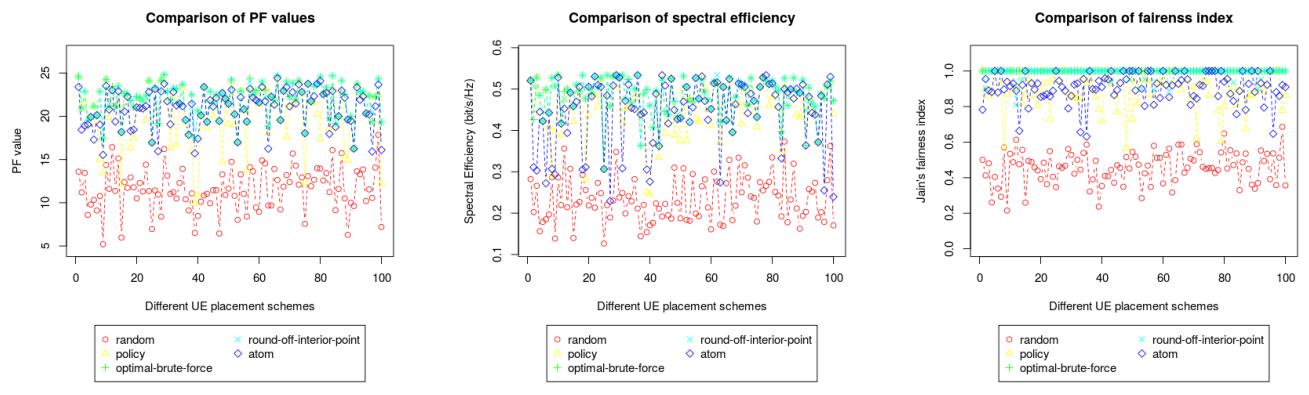


Fig. 15. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.4$, pf only).

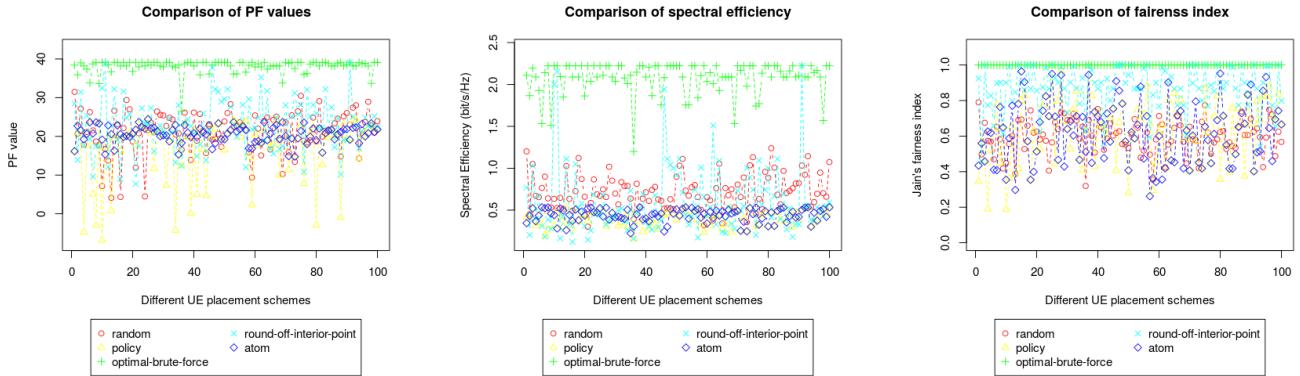


Fig. 16. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.4$, both).

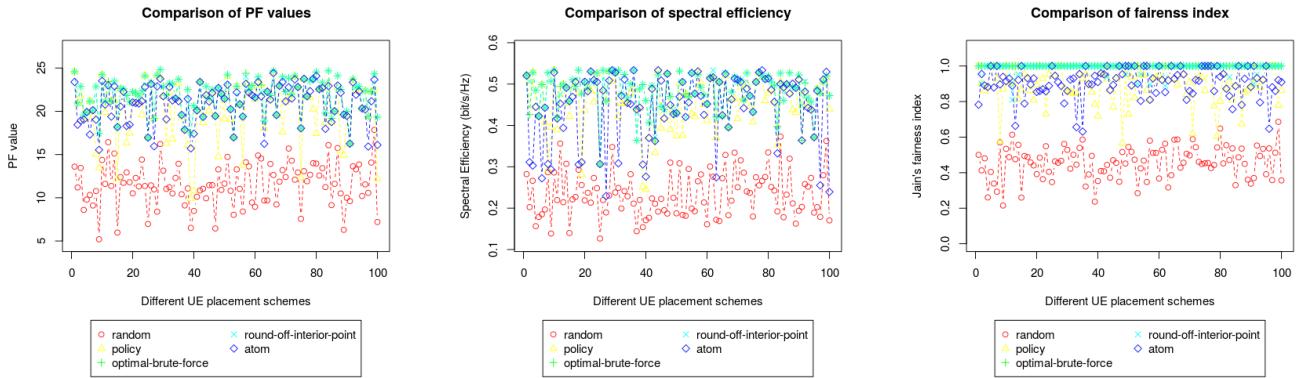


Fig. 17. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.2$, pf only).

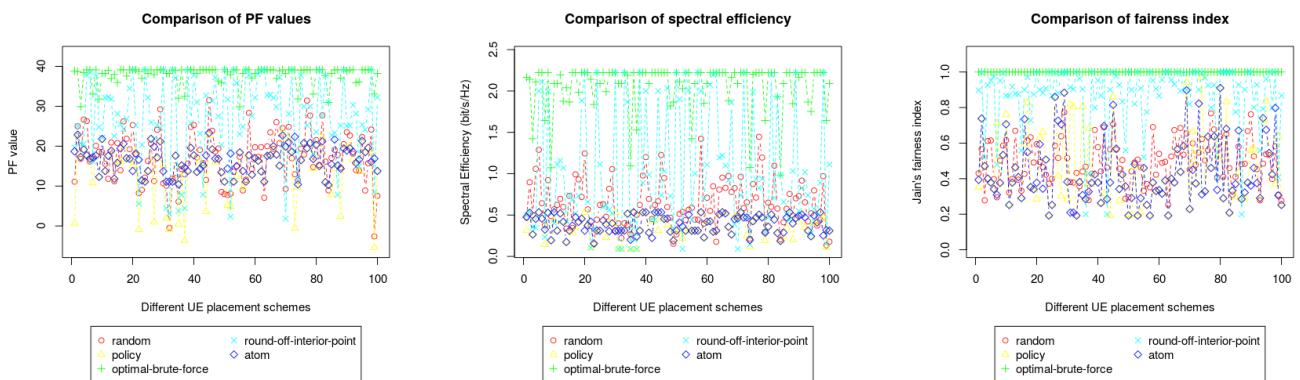


Fig. 18. Overlaid diagram that shows the distance of various previous schemes to the optimal (10 UEs and 3 APs, $P_s = 0.2$, both).

2) *Case 2: both*: Fig. 28-Fig. 30 show the impact of errors to the optimal-brute-force method. Fig. 31-Fig. 33 show that of the atom method. Fig. 34-Fig. 36 show that to the round-off-int method.

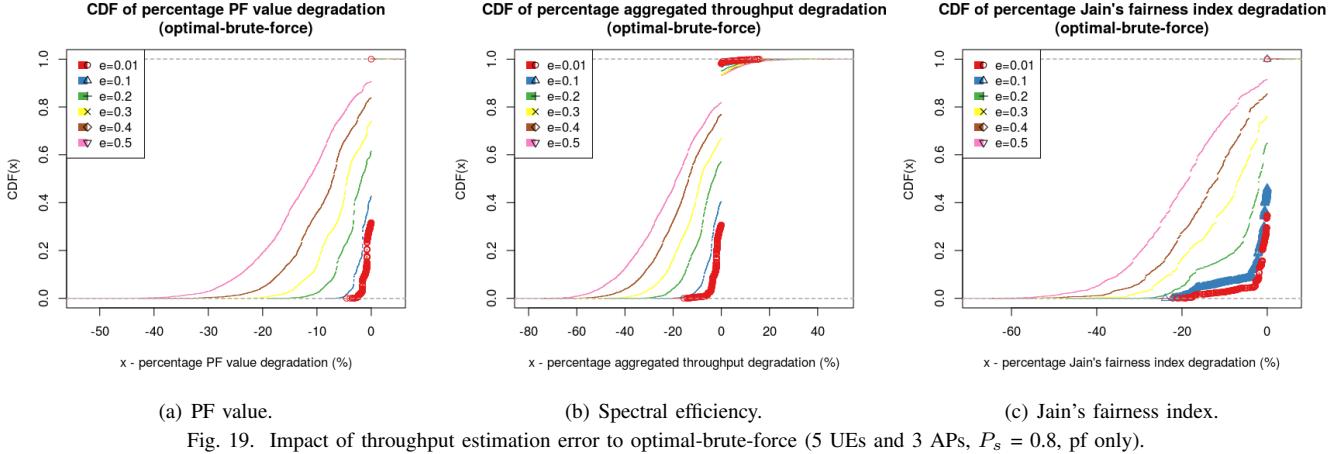


Fig. 19. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.8$, pf only).

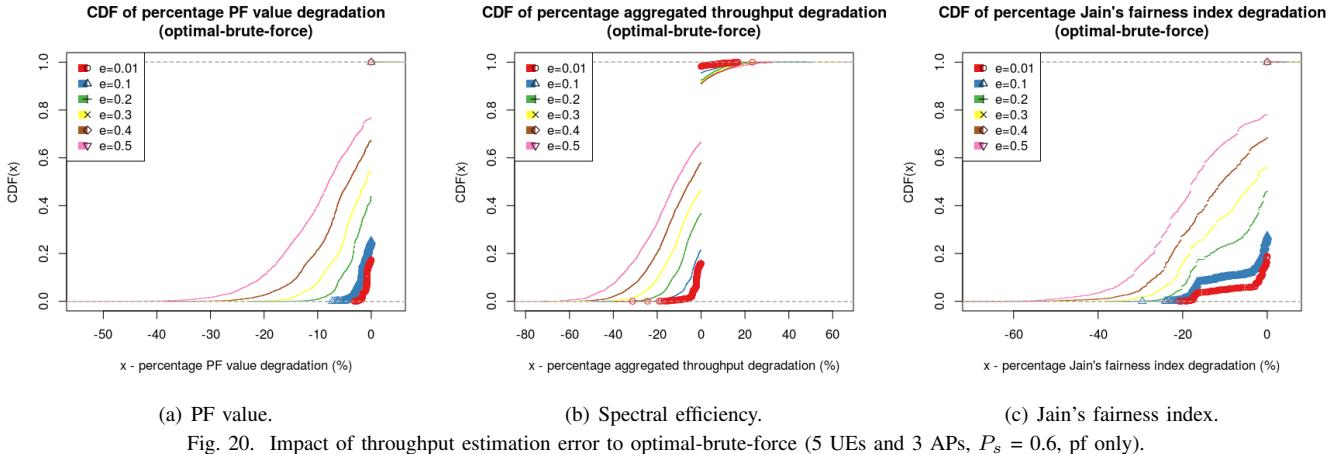


Fig. 20. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.6$, pf only).

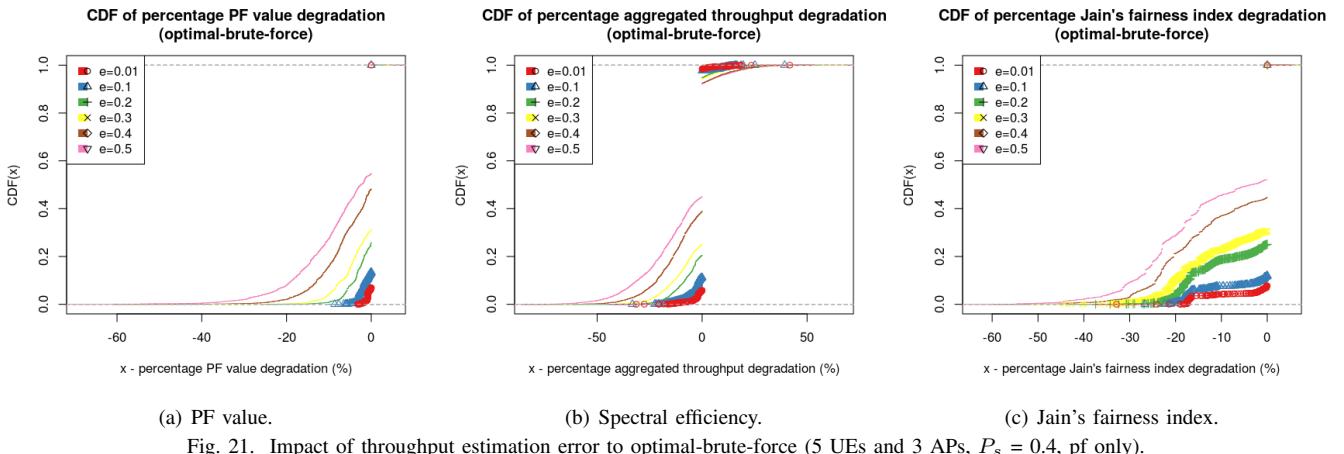


Fig. 21. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.4$, pf only).

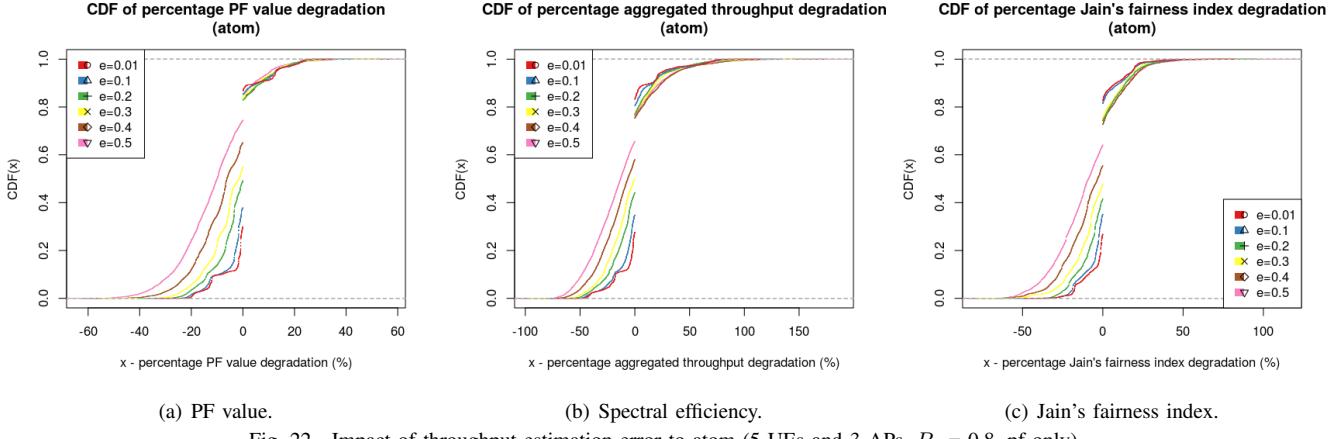


Fig. 22. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.8$, pf only).

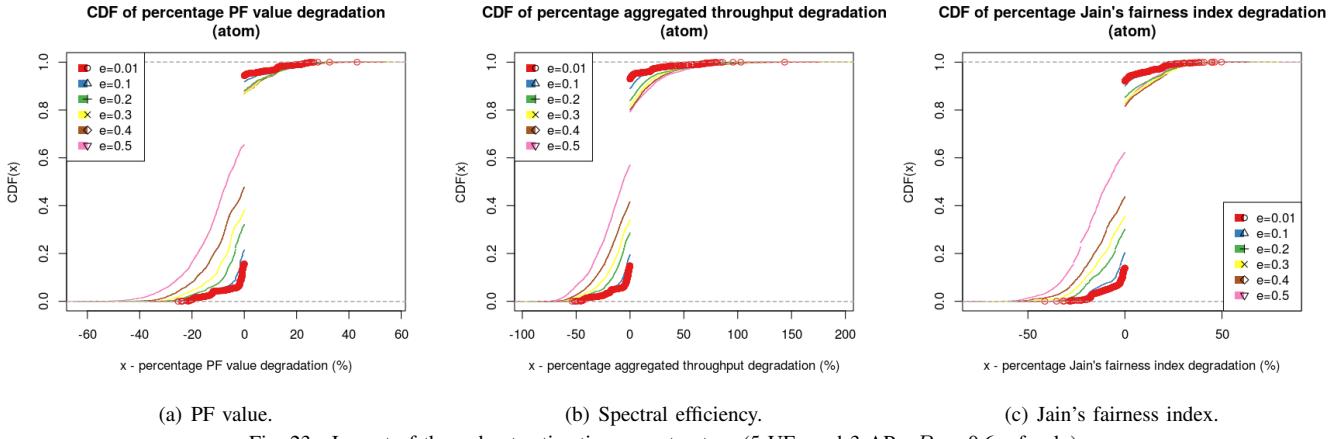


Fig. 23. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.6$, pf only).

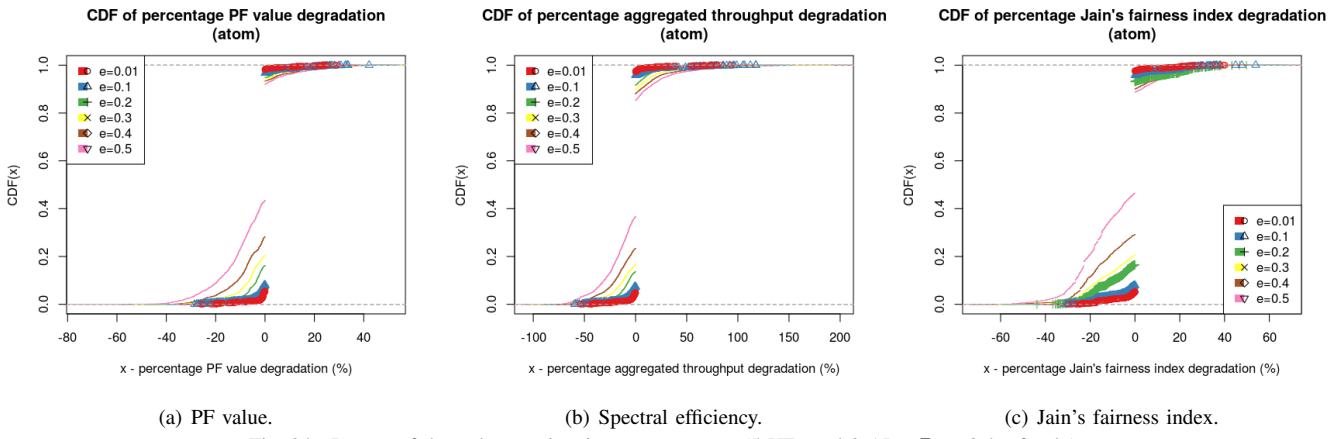


Fig. 24. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.4$, pf only).

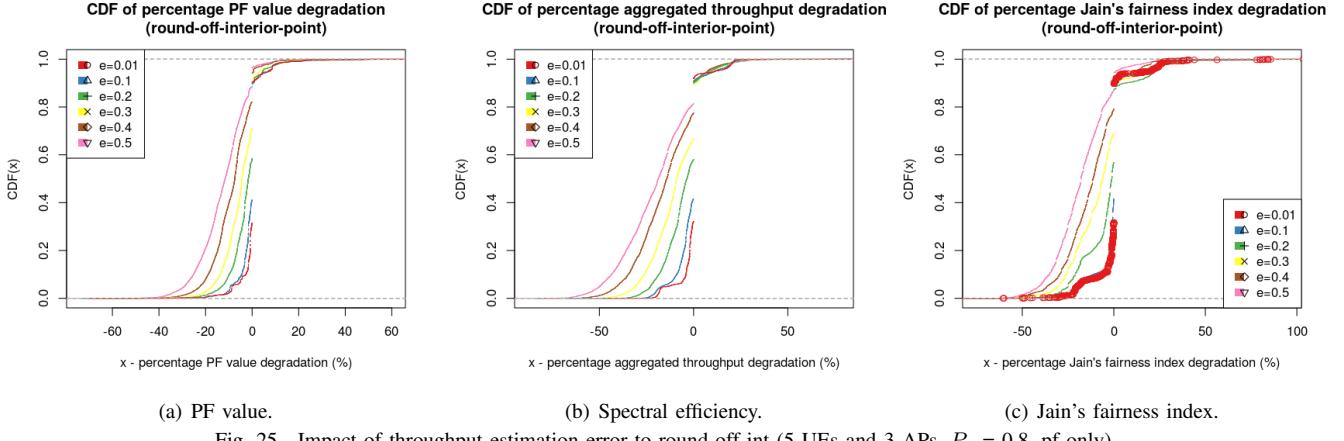


Fig. 25. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.8$, pf only).

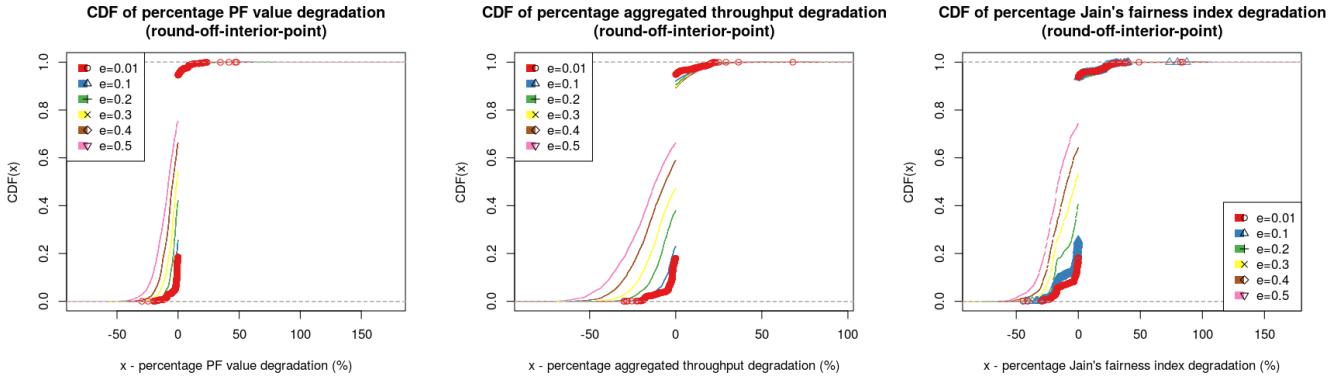


Fig. 26. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.6$, pf only).

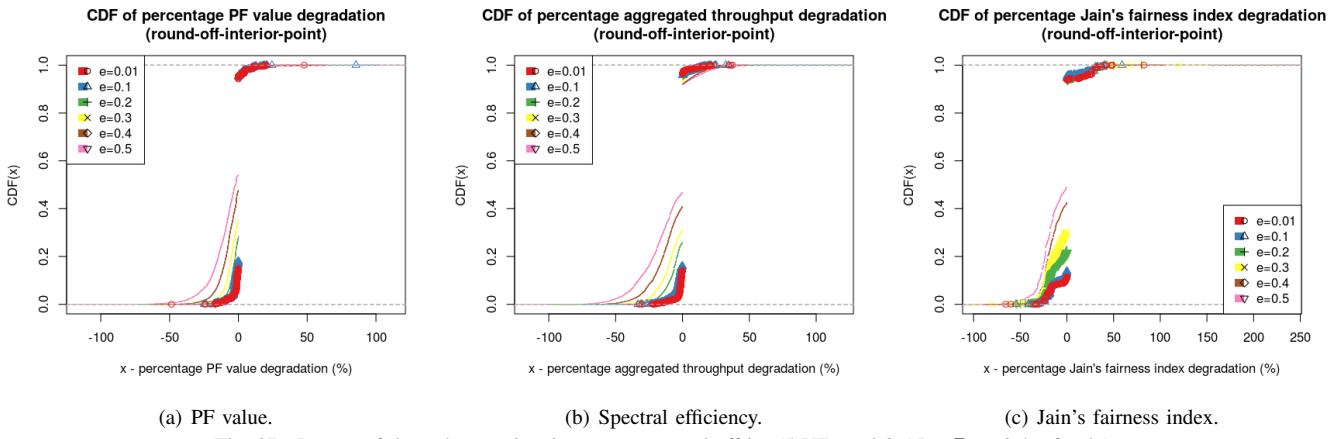


Fig. 27. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.4$, pf only).

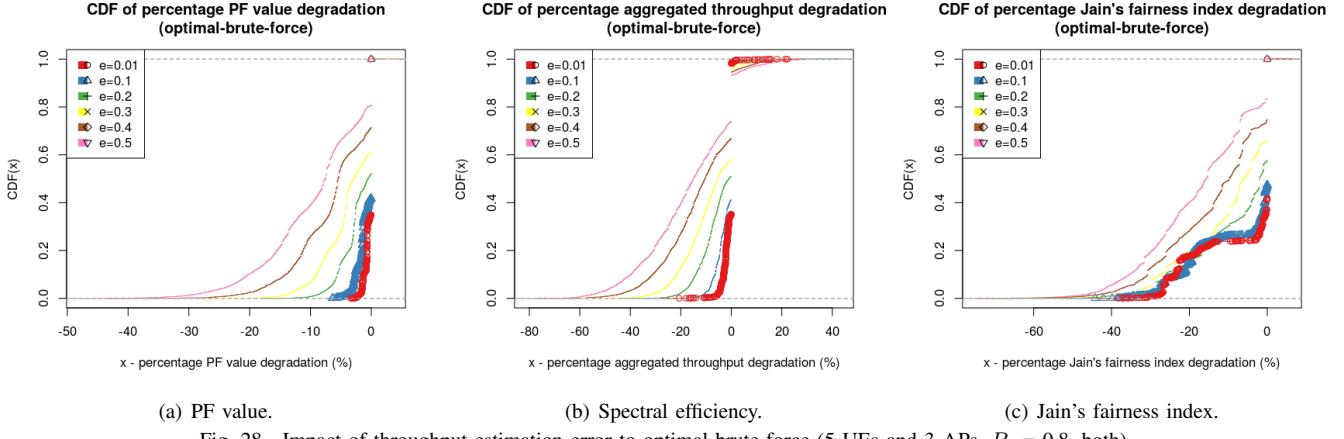


Fig. 28. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.8$, both).

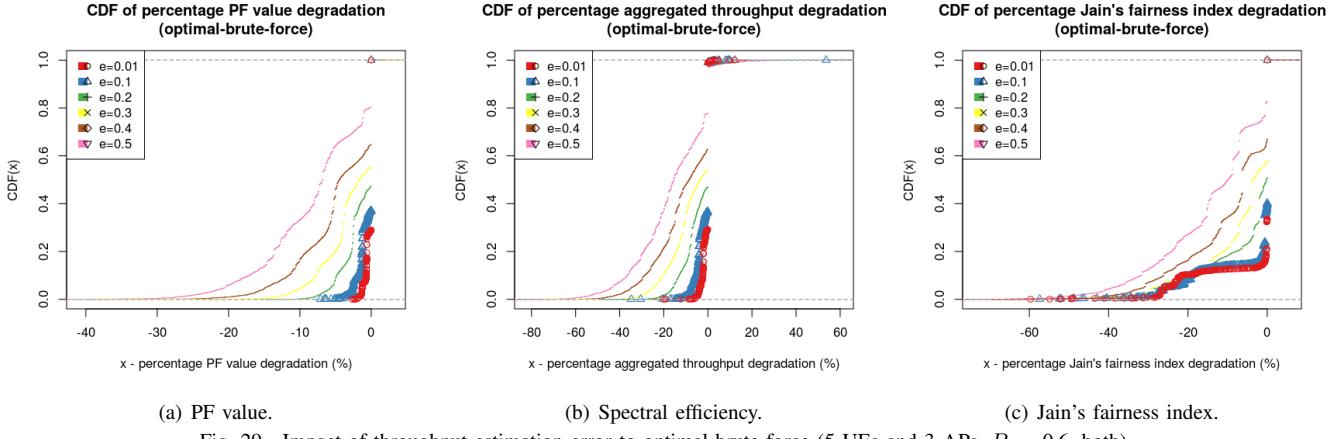


Fig. 29. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.6$, both).

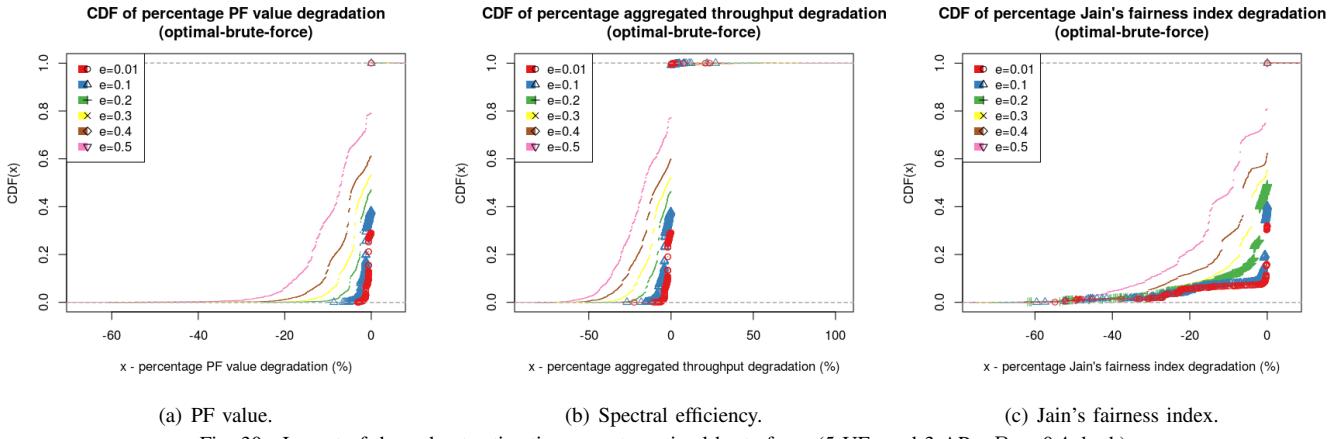


Fig. 30. Impact of throughput estimation error to optimal-brute-force (5 UEs and 3 APs, $P_s = 0.4$, both).

E. Sampling the first 100 raw error rates in Section VI-B

Figs. 37 and 38 shows the results of atom and round-off-int under various input error rates.

REFERENCES

- [1] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in *Proc. of INFOCOM*, April 2006, pp. 1–12.
- [2] D. Bertsimas, V. F. Farias, and N. Trichakis, "The price of fairness," *Operations research*, vol. 59, no. 1, pp. 17–31, 2011.
- [3] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 6, pp. 2706–2716, June 2013.
- [4] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "Rfc6824: Tcp extensions for multipath operation with multiple addresses," 2013.
- [5] A. L. Ramaboli, O. E. Falowo, and A. H. Chan, "Bandwidth aggregation in heterogeneous wireless networks: A survey of current approaches and issues," *Journal of Network and Computer Applications*, vol. 35, no. 6, pp. 1674–1690, 2012.
- [6] S. Deb, K. Nagaraj, and V. Srinivasan, "Mota: Engineering an operator agnostic mobile service," in *Proc. of MobiCom*. New York, NY, USA: ACM, 2011, pp. 133–144.
- [7] A. Sridharan, R. Sinha, R. Jana, B. Han, K. Ramakrishnan, N. Shankaranarayanan, and I. Broustis, "Multi-path tcp: Boosting fairness in cellular networks," in *Proc. of ICNP*, Oct 2014, pp. 275–280.
- [8] R. Amin, J. Martin, J. Deaton, L. DaSilva, A. Hussien, and A. Eltawil, "Balancing spectral efficiency, energy consumption, and fairness in future heterogeneous wireless systems with reconfigurable devices," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 969–980, May 2013.
- [9] H. Zhang, F. Bai, and X. Ju, "Heterogeneous vehicular wireless networking: A theoretical perspective," in *Proc. of WCNC*. IEEE, 2015, pp. 1936–1941.
- [10] A. Anand and G. de Veciana, "Invited paper: Context-aware schedulers: Realizing quality of service/experience trade-offs for heterogeneous traffic mixes," *WiOpt*, 2016.
- [11] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1066–1075, March 2012.
- [12] J. Liu, A. Rayamajhi, and J. Martin, "Using mptcp subflow association control for heterogeneous wireless network optimization," in *Pro. of WiOpt*, May 2016, pp. 1–8.
- [13] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated lte-wifi networks," in *Proc. of MobiCom*. ACM, 2014, pp. 189–200.
- [14] M. Shreedhar and G. Varghese, "Efficient fair queuing using deficit round-robin," *IEEE/ACM Transactions on networking*, vol. 4, no. 3, pp. 375–385, 1996.
- [15] W. Wang, X. Liu, J. Vicente, and P. Mohapatra, "Integration gain of heterogeneous wifi/wimax networks," *Mobile Computing, IEEE Transactions on*, vol. 10, no. 8, pp. 1131–1143, Aug 2011.
- [16] A. Sediq, R. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 7, pp. 3496–3509, July 2013.
- [17] E. Garcia, D. Viamonte, R. Vidal, and J. Paradells, "Achievable bandwidth estimation for stations in multi-rate ieee 802.11 wlan cells," in *2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, June 2007, pp. 1–8.
- [18] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*, vol. 2, March 2003, pp. 836–843 vol.2.
- [19] "NS3," <https://www.nsnam.org/>.
- [20] "ETSI", "Base station (bs) radio transmission and reception, 3gpp ts 36.104 release 14," 2017.
- [21] ETSI, "User equipment (ue) radio transmission and reception, 3gpp ts 36.101 release 14," 2017.
- [22] P. Magazine, <https://www.pcmag.com/Fastest-Mobile-Networks>.
- [23] "Ntt docomo sets 10gbps mobile network speed record," <https://www.extremetech.com/computing/149541-ntt-docomo-sets-10gbps-mobile-network-speed-record>.
- [24] "Technical report on conceptual study of over-the-top resource allocation in heterogenous wireless networks," <https://jianwei-liu.github.io/icc-full.pdf>.

F. More results on the impact of control frequencies

Fig. 39 to Fig. 44 show the impact of control frequencies to the three methods (optimal-brute-force, atom, round-off-int) in terms of PF value, aggregated throughput and Jain's fairness index.

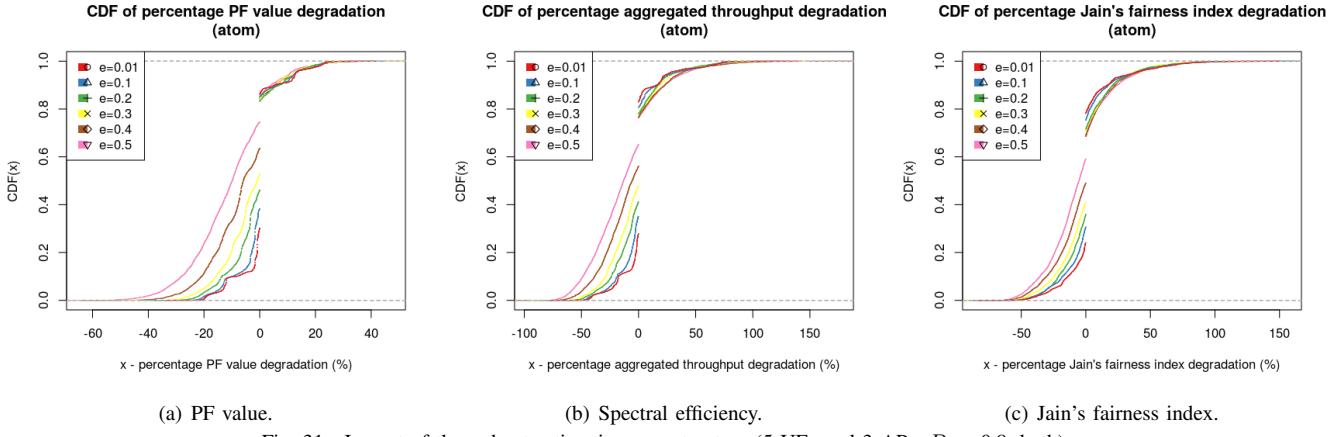


Fig. 31. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.8$, both).

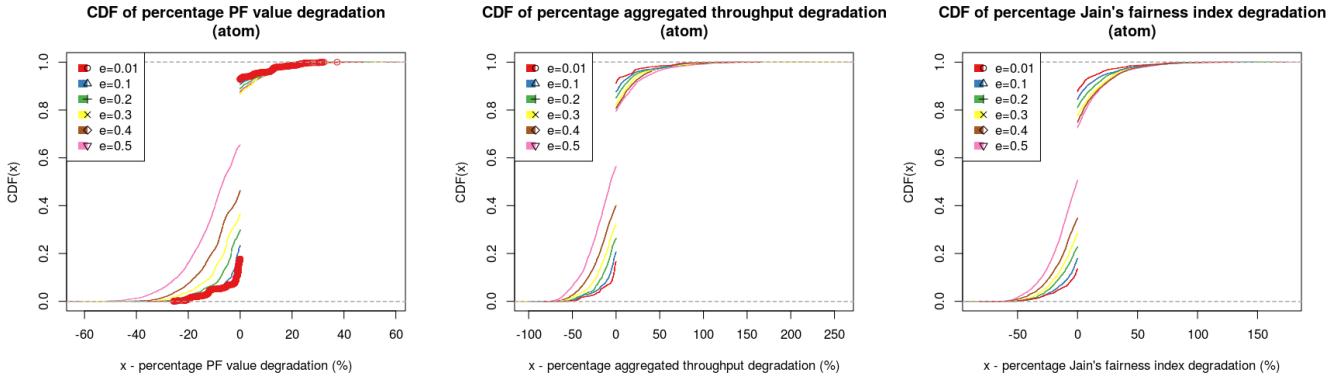


Fig. 32. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.6$, both).

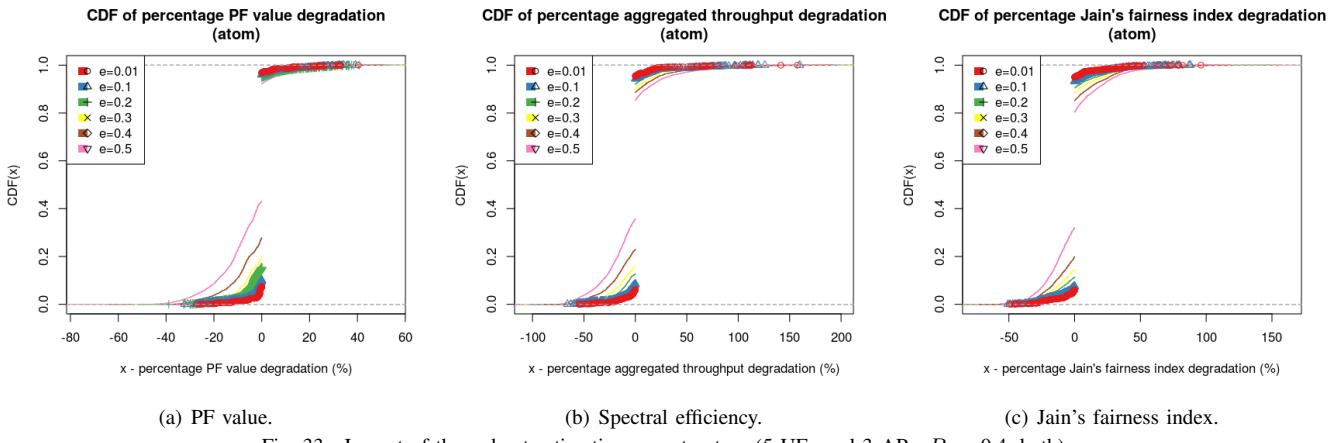


Fig. 33. Impact of throughput estimation error to atom (5 UEs and 3 APs, $P_s = 0.4$, both).

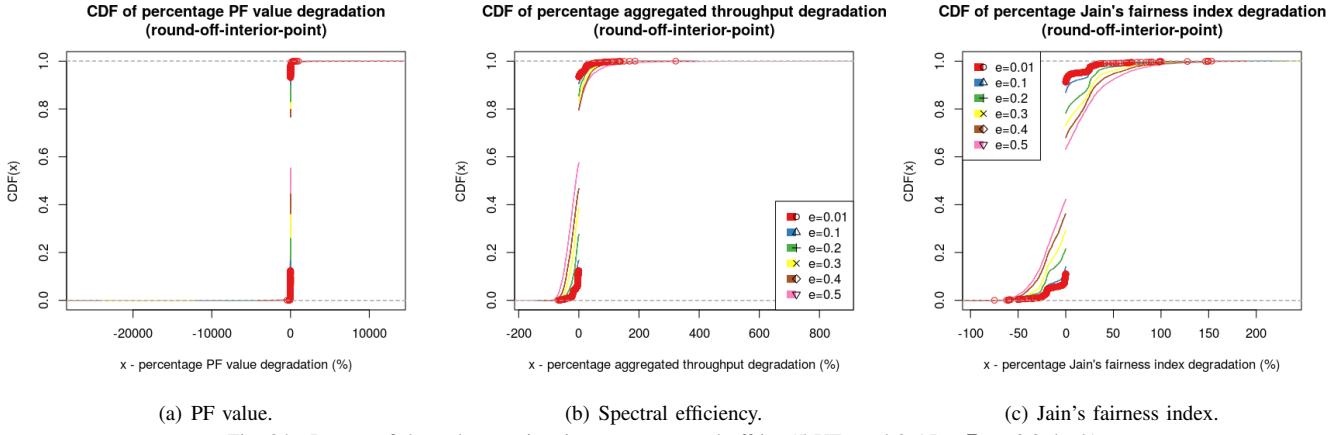


Fig. 34. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.8$, both).

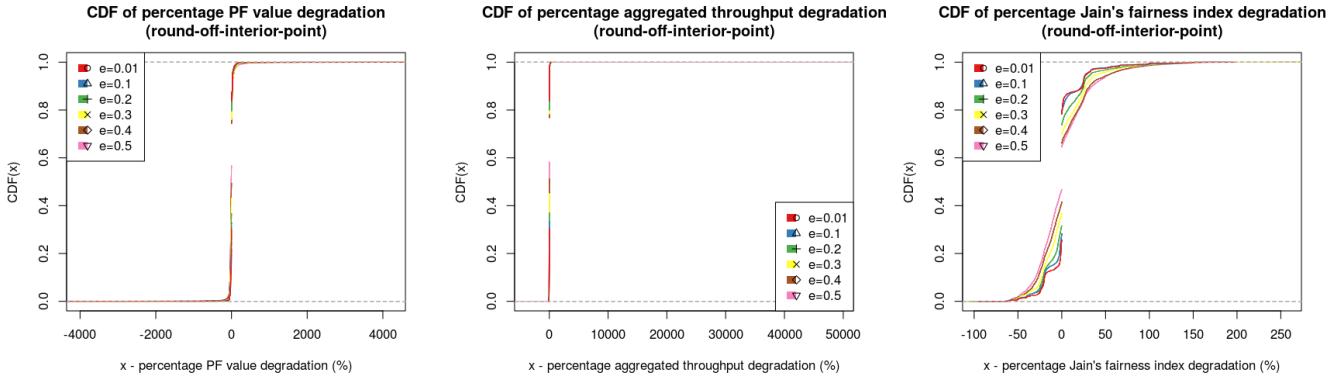


Fig. 35. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.6$, both).

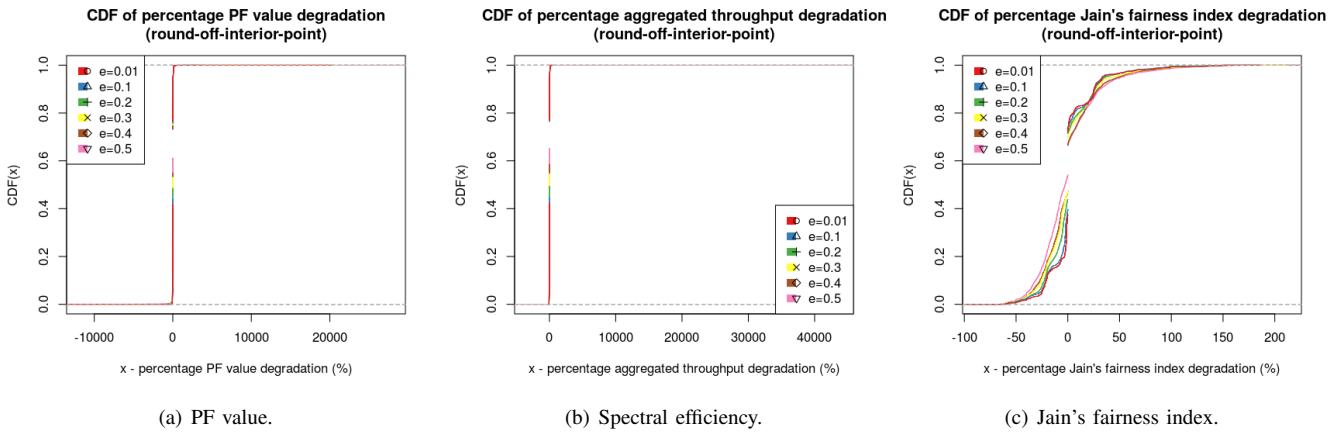


Fig. 36. Impact of throughput estimation error to round-off-int (5 UEs and 3 APs, $P_s = 0.4$, both).

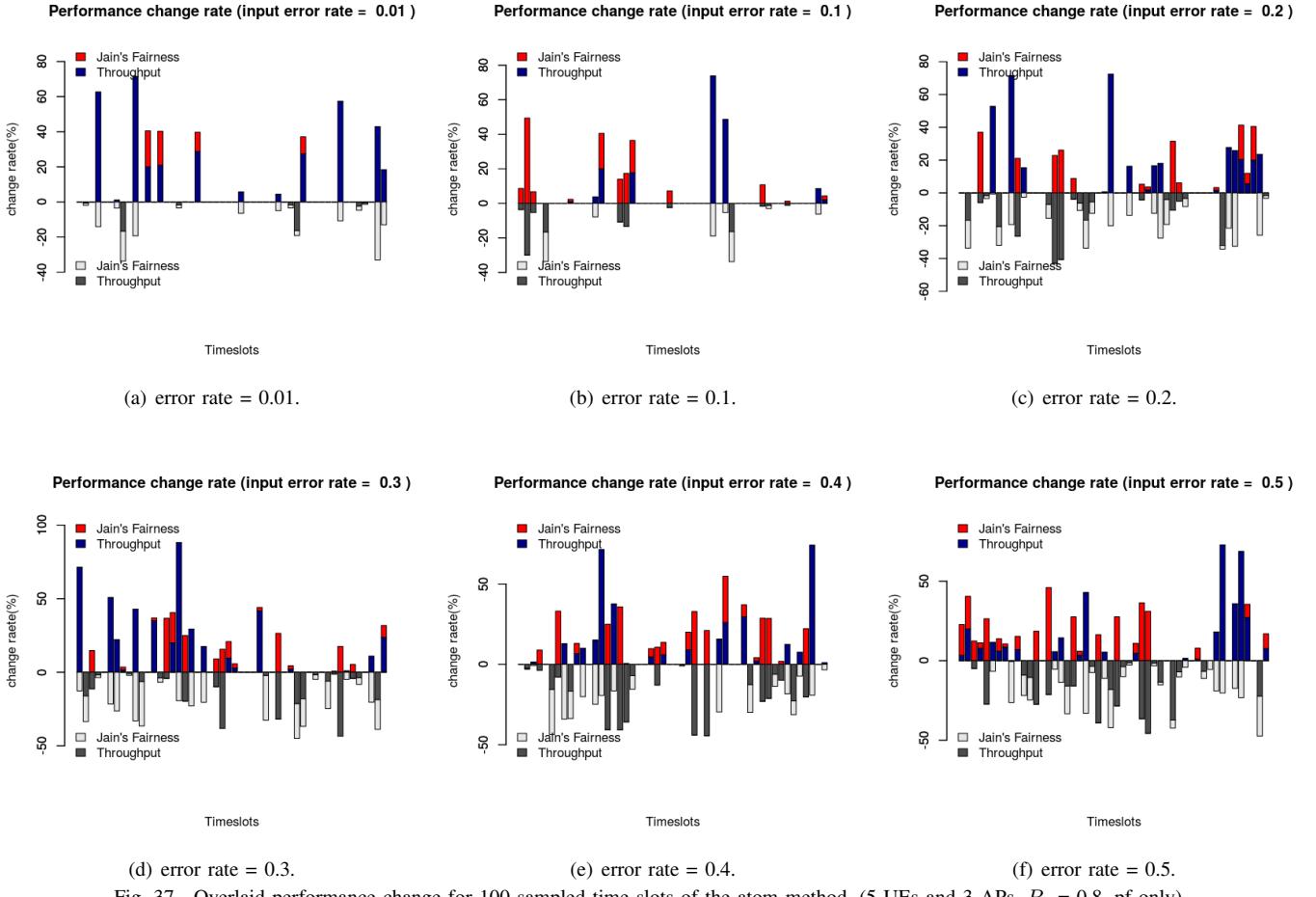


Fig. 37. Overlaid performance change for 100 sampled time slots of the atom method. (5 UEs and 3 APs, $P_s = 0.8$, pf only).

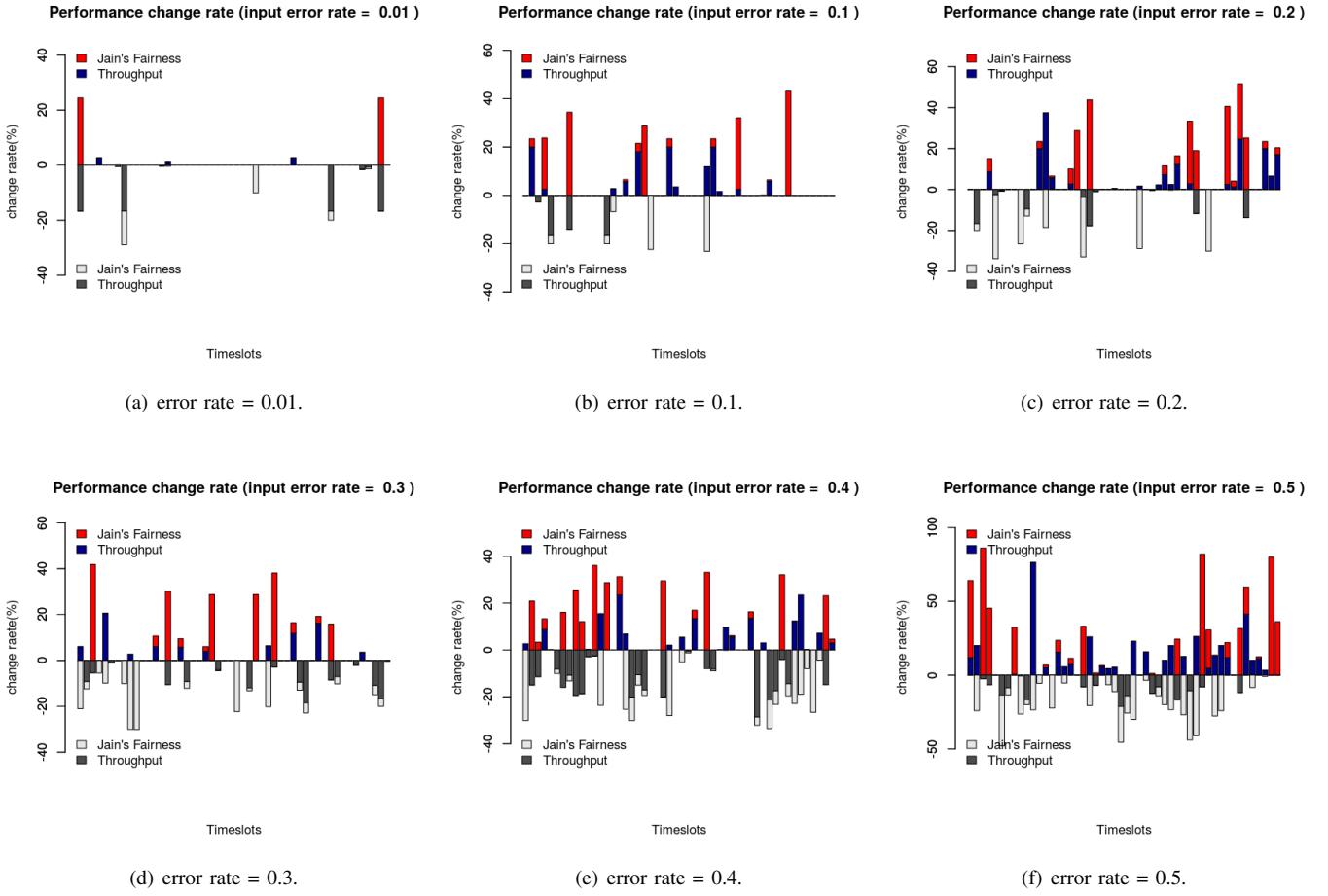


Fig. 38. Overlaid performance change for 100 sampled time slots of the round-off-int method. (5 UEs and 3 APs, $P_s = 0.8$, pf only).

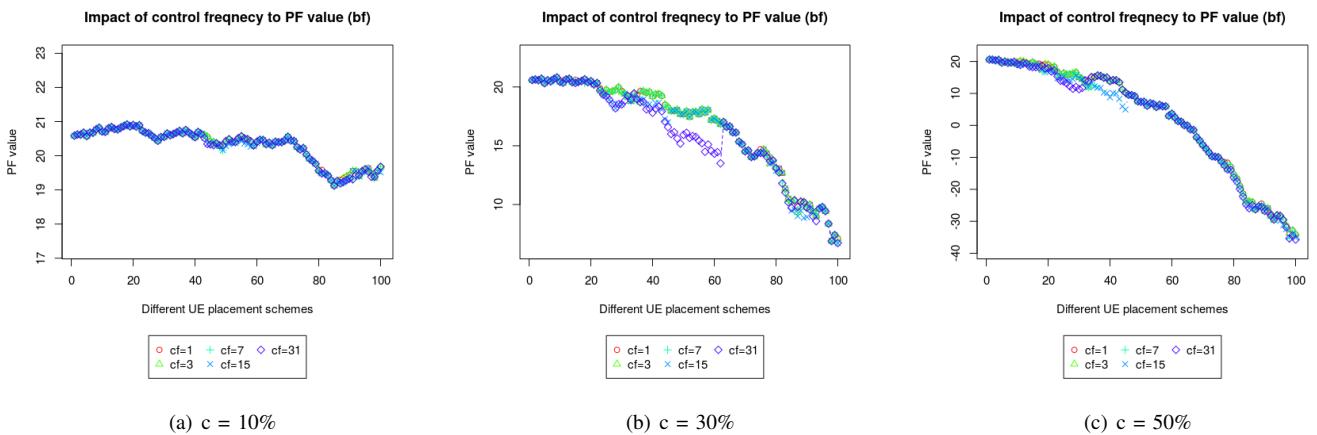


Fig. 39. Impact of control frequency to the PF value of optimal-brute-force (5 UEs and 3 APs, $P_s = 0.8$, pf only).

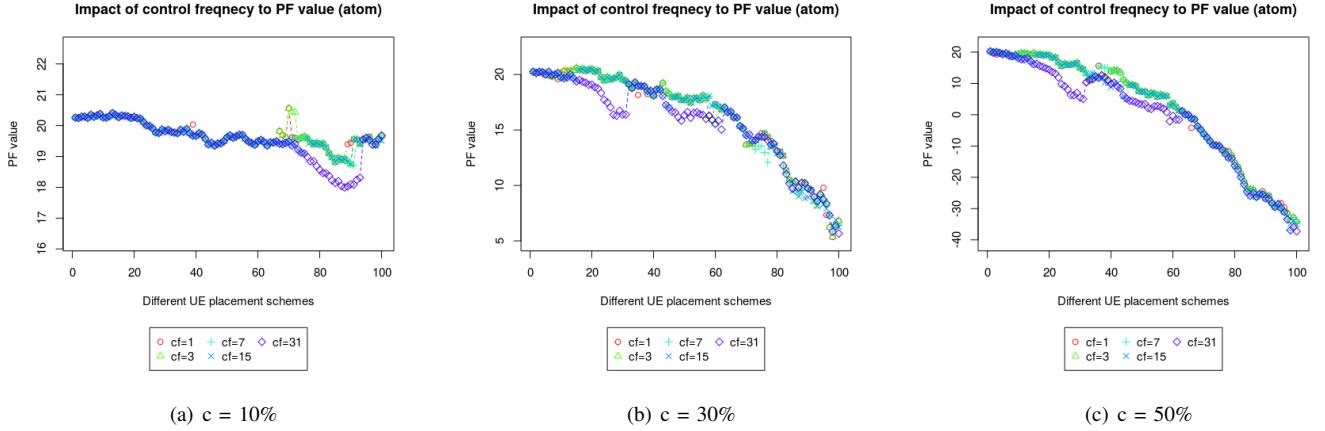


Fig. 40. Impact of control frequency to the PF value of ATOM (5 UEs and 3 APs, $P_s = 0.8$, pf only).

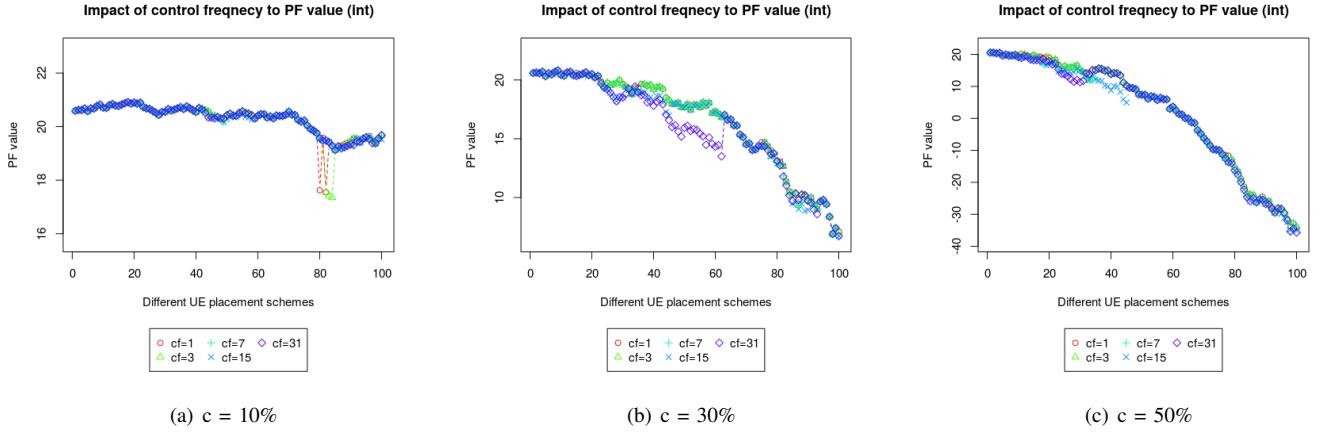


Fig. 41. Impact of control frequency to the PF value of round-off-interior-point (5 UEs and 3 APs, $P_s = 0.8$, pf only).

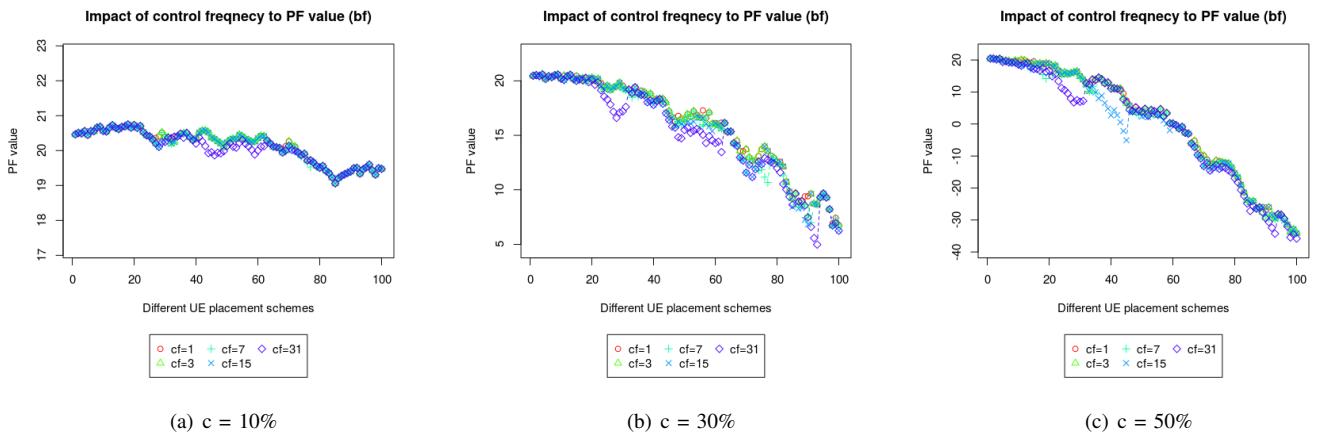


Fig. 42. Impact of control frequency to the PF value of optimal-brute-force (5 UEs and 3 APs, $P_s = 0.8$, both).

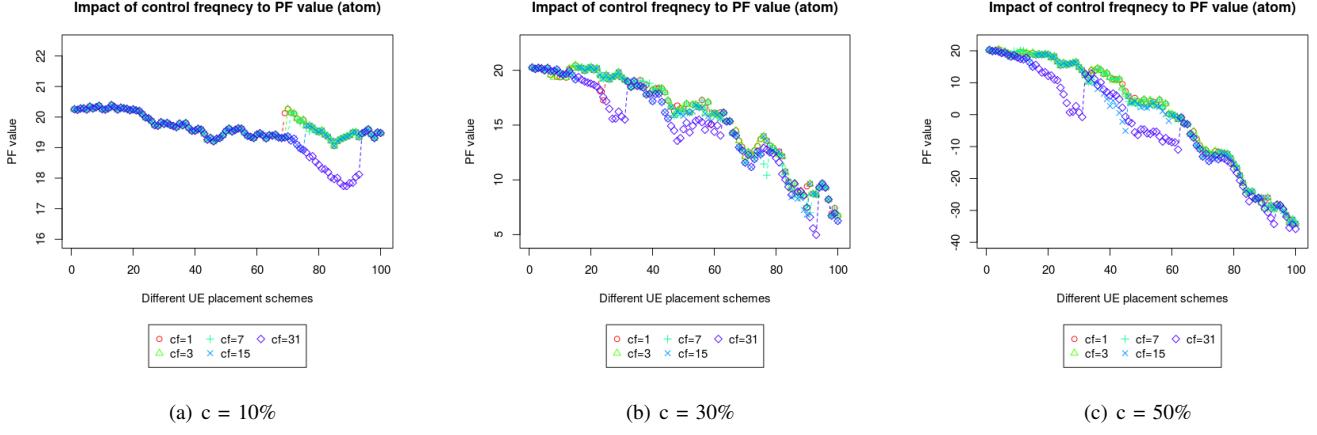


Fig. 43. Impact of control frequency to the PF value of ATOM (5 UEs and 3 APs, $P_s = 0.8$, both).

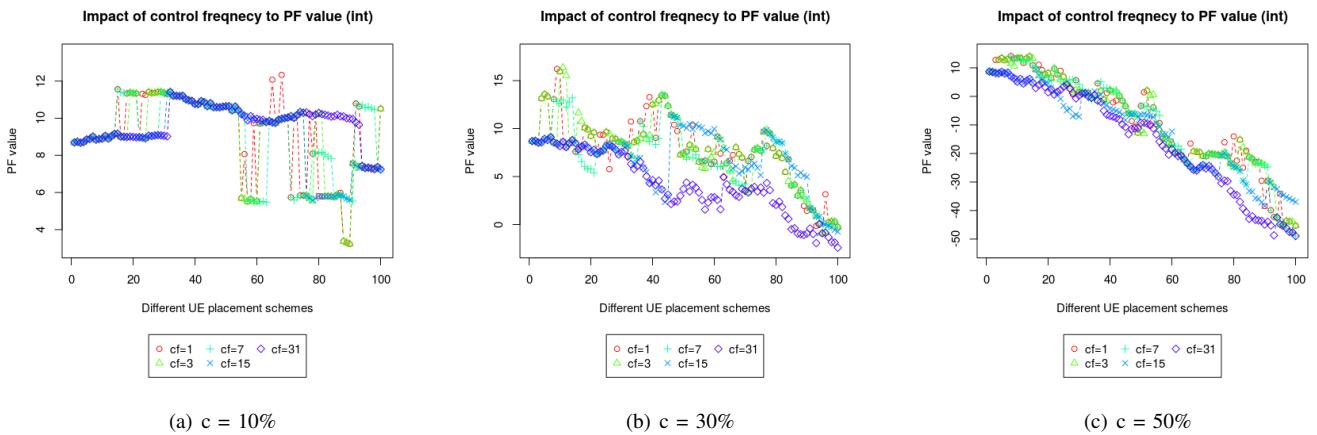


Fig. 44. Impact of control frequency to the PF value of round-off-interior-point (5 UEs and 3 APs, $P_s = 0.8$, both).