

Rado: A Randomized Auction Approach for Data Offloading via D2D Communication

Yifei Zhu*, Jingjie Jiang*, Bo Li*, and Baochun Li†

Department of Computer Science and Engineering, Hong Kong University of Science and Technology*

Department of Electrical and Computer Engineering, University of Toronto†

Email: yzhual@connect.ust.hk, jjiangaf@cse.ust.hk, bli@cs.ust.hk, bli@ece.toronto.edu

Abstract—Despite the growing deployment of 4G networks, the capacity of cellular networks is still insufficient to satisfy the ever-increasing bandwidth demand of mobile applications. Given the common interest of mobile users, Device-to-Device (D2D) communication has emerged as a promising solution to offload cellular traffic and enable proximity-based services. One of the main detriments for D2D communication is the lack of incentive for mobile users to share their content, since such sharing inevitably consumes limited resources and potentially jeopardizes user privacy. In this paper, we study the incentive problem in D2D communications. Specifically, we model the incentive in offloading scenario as an auction game. A trading network is constructed between an eNB and users, in which auctions are conducted to group offloading users and determine proper rewards. We further design a randomized auction mechanism to guarantee system efficiency and truthfulness. Extensive experiments verify the effectiveness of our mechanism in that it achieves a significant performance gain in comparison with baseline methods.

Index Terms—Device-to-Device communication; incentive mechanism; auction game; cellular traffic offloading

I. INTRODUCTION

Mobile data traffic has increased exponentially over the past several years. Global mobile data traffic increased by 81 percent in 2013 and is expected to surpass 15 Exabyte by 2018 [1]. Such meteoric rise in data traffic has caused unprecedented traffic overload problems in existing cellular networks. With minimal new wireless spectrum available, leading mobile operators are being forced to come up with new tactics and technologies to handle mobile data growth. Currently, Wi-Fi and femtocells are the preferred offloading technologies. To better respond to the surge in data traffic and provide better services, Device-to-Device (D2D) communication defined by the Third-Generation Partnership Project (3GPP) has emerged as a critical component of cellular networks to solve the traffic overload problem. D2D communication is defined as direct communication among user equipments (UEs) without traversing the core cellular networks. If UEs in the same proximity have the content that one user wants, this piece of content, hereafter referred to as a *message* in this paper, can be retrieved directly through D2D communication. By directly exchanging data through D2D links, a cellular network reduces the traffic directly passing through the base station, called eNB (evolved NodeB), and thus alleviates its traffic volume without incurring the high cost of building extra infrastructure.

The previous academic literature mainly focuses on how to run D2D communication efficiently as an underlay to the

cellular network [2], [3]. Quite amount of work has also confirmed the benefits of applying D2D in terms of throughput, energy saving, spectrum efficiency and so on [4], [5]. However, when diving into the D2D communication process, the nature that D2D communication requires interaction among independent individuals brings uncertainty to its bright future. To be specific, D2D communication services require users sacrifice their own indispensable resources such as battery life, computation power and privacy information to help others in close proximity [6]. But why would users sacrifice their valuable resources to help others?

Most of previous literature assumes that these independent users are altruistic and are willing to participate in D2D communication [7], [8], [9]. Chen *et al.* in [10] first propose to leverage social relationships to enhance D2D communication. However, in more common scenarios like news sharing in subscription-based services, or video sharing in special events, these potential helpers need more convincing reasons to undertake D2D communication. Lacking incentives will definitely increase the barrier for users to accept D2D, harm its wide deployment, and affect the promising benefits of D2D communication.

Thus, our problem arises as how to give independent and selfish users the incentive to get involved in D2D content sharing. In common D2D offloading scenarios, an eNB is responsible for fulfilling the message requests from users, and a group of users in the network has already archived some of these messages. A trade between the eNB and these potential helpers can then be established naturally to facilitate the exchange. In this trading market, we call the users who have the messages the sellers, and the eNB the buyer. After buying what it wants from the message holders, the eNB pays monetary rewards or virtual currency to the sellers, which serves as incentives to the sellers. Nevertheless, since the value of a message is different for each user and we cannot know such information in advance, it is hence nontrivial to set a price without triggering the overpricing or underpricing problem.

Auction is one of the most prevalent forms of trading as it allows competitive price display and efficient resource allocation even when valuation information is private. Typically, a well-designed auction ensures to elicit the true valuation of items from each bidder. Using this truthful revelation, the auctioneer selects the bidders that are consistent with the designer's goal to win the auction, and determines their clearing

prices. A good auction is expected to satisfy the following three properties: social efficiency: the sum of utilities of all players is maximised; individual rationality: the utility of each individual player is at least zero; truthfulness: the bid value of each bidder equals to its own valuation. These properties guarantee the trading result is efficient and fair to all users in the network. However, existing auction mechanisms cannot be directly applied to our scenario. The most famous auction mechanism, Vickrey-Clarke-Groves (VCG) auction, requires solving a social welfare optimization problem optimally to ensure truthfulness and system efficiency. Unfortunately, the optimization problem in our scenario is NP-hard. To make things worse, approximately solving the optimization problem harms the truthfulness of VCG mechanism [11]. As the linchpin of a good auction, truthfulness is the prerequisite of other important properties.

In this paper, we propose Rado, a Randomized Auction mechanism for Data Offloading in cellular networks. To be specific, firstly, we consider collecting message requests in aggregated time intervals, and exploit broadcast capacity in D2D communication. Secondly, we study the incentive issue in D2D communication under this new scheme, and formulate this problem into an integer programming problem. After demonstrating the difficulties in applying the existing mechanisms directly, we propose a randomized combinatorial auction mechanism, Rado, to solve it. This mechanism first simulates a fractional VCG auction, it then decomposes the fractional solution into the convex combination of feasible integral solutions. Winning players and payments are determined based on the weight of feasible solutions as the probability. Thirdly, we prove theoretically that Rado guarantees social efficiency, individual rationality, and offers truthfulness in its best effort. To complement our mechanism, we further propose a near optimal approximation algorithm. Last but not least, extensive experiments verify the excellent performance of Rado in offloading, and other desirable economical properties.

The remainder of this paper is organized as follows. In Section. II, we first provide necessary preliminaries for our problem. We then formally formulate our incentive problem in the offloading scenario in Section. III. In Section. IV, we start from the overview of Rado and provide its detailed design after that. To backup our mechanism, we present an approximation algorithm in Section. V. Extensive experiments are provided in Section. VI, followed by the related work part in Section. VII. Finally we conclude our paper in Section. VIII.

II. BACKGROUND AND PRELIMINARIES

A. D2D communication

The D2D communication we discuss in this paper specifically refers to LTE D2D techniques in 3GPP, which involves both an eNB and UEs in a hybrid network, and the eNB exerts light control over D2D communication. In LTE, the spectrum is divided into resource blocks, each consisting of 12 adjacent sub-carriers. To avoid interference between different D2D links, we consider that the eNB allocates a resource block to each device to be used in D2D communication.

The interference mitigation is accomplished at the beginning of the communication to avoid interference between devices [12]. The detailed interference management system is beyond the scope of our work. We study the incentive issue in the offloading scenario on top of these underlay mechanisms.

The application of D2D is generally categorized into two fields: public safety services and commercial use. Of the two, the public safety service is currently regarded as a priority to develop. Unlike the unicast scheme adopted by existing D2D communication techniques (*e.g.*, LTE-direct), the broadcast scheme is a required feature for public safety applications. What is more, owing to the broadcasting nature of wireless signal, one-to-many broadcasting also surpasses one-to-one unicasting in terms of device discovery, energy savings and traffic offloading. Researchers have also investigated the possibility of applying broadcast D2D transmission to some application scenarios [13], [9]. Given the advantages presented above, we also adopt broadcast D2D communication in this paper, and study its effect on offloading performance.

Researchers in previous literature either assume D2D links (communication pairs) are predetermined or simply make a helper respond to the first request it receives as long as the QoS requirement is satisfied. Such a First-Come, First-Served mechanism, though easy to implement, misses the opportunity to further improve offloading performance. This phenomenon is especially obvious in broadcast D2D communication. Simply satisfying the first request of a single user immediately, while abandoning the chance of broadcasting a message that could benefit multiple users would definitely degrade the offloading performance. Therefore, without incurring much transmission delay, we consider dividing the communication period into small time intervals; each helper collects requests at every interval and responds to a bunch of requests through broadcast.

B. Auction game

Typically, there are two roles in an auction, bidder and auctioneer. Bidders could be sellers or buyers. An auctioneer hosts the auction, and collects bids; bidders send asks or bids to the auctioneer in order to sell or buy items. After receiving all bids, the auctioneer decides the allocation of items and the clearing price. Behind all of these, it is the designer's job to design an auction that is both fair and efficient. This process is called mechanism design. Unlike conventional forward auctions where multiple buyers bid for items, reverse auctions refer to the situations that several sellers compete to win the items from the auctioneer. Correspondingly, double auctions are applied in multiple-buyer, multiple-seller situations. Reverse auctions are extensively used in the transport industry. In these reverse auctions, bidders send bids about how much they must be offered at least for taking delivery services, and the auctioneer selects bidders to complete the delivery services with the lowest hiring cost.

A critical point for D2D communication to be widely accepted is that it should be transparent to users. Users just need to send their requests and then they receive what they

want without knowing the source. Besides that, after paying their monthly fees to service operators, obviously there is no reason to ask users join the auction paying extra money to get what they deserve. Therefore, we construct reverse auctions between the eNB and helpers to model our scenario rather than choosing double auctions or forward auctions. In our auction, the eNB acts as the auctioneer, users with cached messages act as sellers. The auctioneer could also be the cloud platform like other works have done [14]. Auctions are periodically taken in each auction round. Compared with the long term leasing such as spectrum auction in FCC, we take single round auction instead of iterative auction to reduce communication time as well as overheads. Players are assumed to be independent of each other. We do not consider collusion situations in this paper and leave designing a collusion-free auction mechanism for D2D communications to future work.

III. PROBLEM FORMULATION

We consider incentive issues in the D2D offloading scenario. Auctions are constructed to facilitate exchanges between an eNB and helpers, where the eNB buys messages from the helpers, and helpers get payments for their sharing messages. We have k users that request m distinct messages from n potential helpers in the network. Let M denote the total message set that is requested by the users, D denote the set of the corresponding users that requests these messages, S denote the set of potential helpers, a.k.a sellers in the auction. Each helper caches a subset of M for sharing; Each seller s_i names its bid b_i^n for message M_n based on an individual private cost function. Let the bid vector of seller s_i be $B_i = (b_i^1, b_i^2, b_i^3 \dots b_i^m)$. Sellers further send this information to the auctioneer. Based on the messages cached in seller s_i , s_i is capable of satisfying different subsets of users corresponding to different messages. We index the resultant subsets of users in the network for convenience and have q as the index of each subset. Index q can also represent its corresponding subset interchangeably.

We first use the following example in Fig. 1 to illustrate the problem we have, and introduce the formal formulation after that. In our example, four users request three messages. We have a seller set $S = \{s_1, s_2\}$ and their corresponding user subset group $\{1.\{D_1, D_2\}, 2.\{D_3\}, 3.\{D_2\}, 4.\{D_4\}\}$. Index of each user subset is also listed in front of each subset. Among the helpers, s_1 caches M_1 and M_2 , which is requested by the user set $\{D_1, D_2\}$ and $\{D_3\}$ in proximity respectively, and s_2 caches M_1 and M_3 , which is requested by $\{D_2\}$ and $\{D_4\}$ in proximity respectively. The valuation to transmit the corresponding message is also listed out after each message. Since for s_1 and s_2 , they can only choose one message to send at each time, a decision on which message to send has to be made. We hope that all the requests can be satisfied through the D2D links and the expenditure for hiring these sellers can be minimised. The optimal solution in this example is to satisfy $\{D_1, D_2\}$ and $\{D_4\}$ by allowing s_1 send M_1 , s_2 send M_3 with total cost 3. The request from D_3 is left to the eNB. While for larger instances, this determination problem is not

that easy to solve. Notice that for the eNB, buying the same kind of message from different helpers may generate different results to it. For instance, having seller 1 share M_1 and having seller 2 share M_1 produce different results to the eNB. This can happen when D_1 is out of the communication range of seller 2. Thus our problem cannot be directly modelled as a single item auction of selling distinct messages or multi-unit auction with copies of each message. Instead, we have a combinatorial auction situation here.

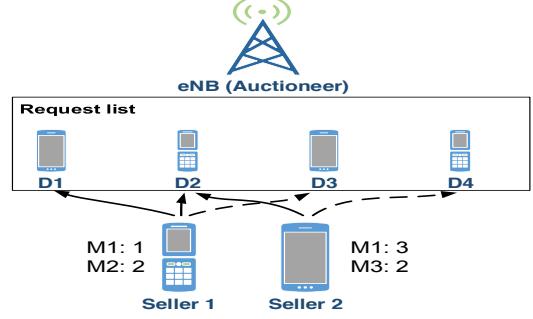


Fig. 1. Sellers sell messages to the eNB, and the eNB wants to reduce cost and cover all requests

Therefore, in the general framework, the system-wide social welfare maximisation problem in our reverse auction translates into the objective of minimising the aggregated cost, while covering all the message requests. Our winner determination problem then can be formulated as follows:

We have a set of 0-1 variable $x_{i,q}$ for each seller s_i . Variable $x_{i,q}$ equals one if the seller i is chosen to satisfy requests from user subset q . This subset corresponds to a message in s_i . We have another variable $x_{eNB,q}$ equals one if q is satisfied through cellular links with cost $C_{eNB,q}$. To better represent constraints in the matrix form, we introduce one more variable: $a_{i,q}^j$ that equals one if device D_j is in the subset q , and seller s_i also bids for this subset.

$$\min \sum_i \sum_q x_{i,q} b_{i,q} + \sum_q C_{eNB,q} x_{eNB,q} \quad (1)$$

$$\text{s.t. } \sum_{q \in D} x_{i,q} \leq 1, \forall s_i \in S \quad (2)$$

$$\sum_i \sum_q x_{i,q} a_{i,q}^j + \sum_{\forall D_j \in q} x_{eNB,q} \geq 1, \forall D_j \in D \quad (3)$$

$$x_{i,q}, x_{eNB,q} \in \{0, 1\} \quad (4)$$

Constraints (2) represent that for each seller, it can only sell one message, namely, each helpers can only choose one message to send at each time. Constraints (3) guarantee that each message request can be satisfied by one helper or the eNB. As a result, this winner determination problem (WDP) is NP-hard (proved in Theorem. 1), making it non-trivial to design an auction mechanism that simultaneously satisfies truthfulness and social efficiency. We summarise the notation

TABLE I
TABLE OF NOTATIONS

S	total seller set	$x_{eNB,q}$	indication of whether eNB has to satisfy user subset q
D	total user set	$b_{i,q}$	the bid of seller s_i for user subset q
s_i	seller i	$C_{eNB,q}$	the cost of eNB to cover user subset q
D_j	user j that requests a message	$d_{i,q}^j$	indication of whether user D_j is in subset q of seller s_i
q	q^{th} user subset	λ_j^r	Lagrangian multiplier of user j in iteration r
$ q $	the number of users in subset q	h_r	step size at iteration r
$x_{i,q}$	whether seller s_i is chosen for user subset q		

in Table I for convenience.

Theorem 1. *WDP in (1) is NP-hard.*

Proof. We present a polynomial-time reduction from the set covering problem, a NP-hard problem [15]. The set covering problem is defined as:

$$\begin{aligned} & \min \sum_S w(S) \cdot x_S \\ & \text{s.t } \sum_S a_{e,S} \cdot x_S \geq 1, \forall e \in U, x_S \in \{0, 1\}. \end{aligned}$$

Let us consider a special case of the WDP. Suppose each seller only transmits one bid, and all requests can be satisfied by the D2D links. The WDP degrades into:

$$\begin{aligned} & \min \sum_{i=1}^n x_{i,q} b_{i,q} \\ & \text{s.t } \sum_{q:D_j \in q} x_{i,q} \geq 1, \forall D_j \in D, x_{i,q} \in \{0, 1\}, \end{aligned}$$

which is exactly a set covering problem. Obviously, the reduction can be done within polynomial time. Therefore, WDP in (1) is NP-hard. \square

IV. RANDOMIZED AUCTION MECHANISM DESIGN

We propose a randomised auction mechanism, Rado, to solve our problem ensuring system efficiency and truthfulness simultaneously. The general framework of our randomised auction can be divided into three major steps, as illustrated in Algorithm 1.

Algorithm 1 A randomised auction mechanism

- 1: Simulating a fractional VCG auction
 - 2: Decomposing the fractional solution into the convex combination of feasible integral solutions
 - 3: Randomised winner selection and payment determination
-

We review these three key steps one by one in details in the following:

A. The fractional VCG auction

If we allow fractional results, we are capable of relaxing the original integer linear program to get an optimal fractional solution and simulate a fractional VCG auction based on that. To be specific, after relaxing the integer constraints in the WDP with new constraints $x_{i,q}, x_{eNB,q} \geq 0$, we have a linear programming relaxation (LPR) of our original problem, which can be solved optimally, OPT_{LPR} . We can further determine payments using the VCG mechanism. Denote the optimal fractional allocation and the payment as $(x_{i,q}^*, x_{eNB,q}^*, p_i^*)$.

The payment $p_{\tilde{i}}^*$ for $s_{\tilde{i}}$ is defined as

$$\begin{aligned} p_{\tilde{i}}^* = & (\sum_q \sum_i \tilde{x}_{i,q} b_{i,q} + \sum_q C_q \tilde{x}_{eNB,q}) \\ & - (\sum_q \sum_{i \neq \tilde{i}} x_{i,q}^* b_{i,q} + \sum_q C_q x_{eNB,q}^*) \end{aligned} \quad (5)$$

based on the conventional VCG mechanism [16]. In the first term, $\tilde{x}_i, \tilde{x}_{eNB}$ is an optimal fractional solution without considering seller \tilde{i} ; the second term is the social welfare of others according to the determined outcome. The intuition behind VCG is to calculate the opportunity cost of letting a seller win, namely how much this allocation decision hurts others. According to the VCG mechanism, the resulting allocation solution is optimal, consequently the corresponding payment guarantees truthfulness and individual rationality. However, this result is not applicable in real cases because the allocation result is fractional.

B. Decomposing the fractional solution

Based on the convex decomposition techniques in [17], we intend to decompose the fractional results into the convex combination of integral feasible solutions. We need an β -approximation algorithm to WDP serving as a separation oracle to help us accomplish this decomposition, namely we need the solution found by our approximation algorithm to satisfy

$$\sum_q \sum_i x_{i,q}^p b_i(q) + \sum_q C_{eNB,q} x_{eNB,q}^p \leq \beta OPT_{LPR}. \quad (6)$$

Our goal of decomposition is to find out a set of nonnegative coefficient ω^p , such that $\sum_p \omega^p = 1$, $\sum_{p \in P} \omega^p x_{i,q}^p \leq \beta x_{i,q}^*$, $\sum_{p \in P} \omega^p x_{eNB,q}^p \leq \beta x_{eNB,q}^*$, where P is the set of extreme points in the integer polyhedron of WDP. We also say the vector $\beta x_{i,q}^*$ dominates the convex combination of integral solutions. This is accomplished by solving the subsequent well designed primal-dual linear program:

Primal:

$$\max \quad \sum_{p \in P} \omega^p \quad (7)$$

$$\text{s.t.} \quad \sum_{p \in P} \omega^p x_{eNB,q}^p \leq \beta x_{eNB,q}^* \quad (8)$$

$$\sum_{p \in P} \omega^p x_{i,q}^p \leq \beta x_{i,q}^*, \quad (9)$$

$$\sum_{p \in P} \omega^p \leq 1, \quad \omega^p \geq 0 \quad (10)$$

Dual:

$$\min \quad \beta \sum_q (\sum_i \mu_{i,q} x_{i,q}^* + \nu_q x_{eNB,q}^*) + \eta \quad (11)$$

$$\text{s.t.} \quad \sum_q (\sum_i \mu_{i,q} x_{i,q}^p + \nu_q x_{eNB,q}^p) + \eta \geq 1 \quad (12)$$

$$\mu, \nu, \eta \geq 0 \quad (13)$$

The solutions of the primal problem explicitly present us desired convex decompositions if its objective value is one. Since the primal problem (7) has exponential number of variables, it may take exponential time to solve it directly. Therefore we intend to solve the dual problem (11) using the ellipsoid method first, though it has exponential number of constraints [18]. We first prove the existence of such decomposition in Lemma. 1, then prove that we can solve this decomposition in polynomial time through solving its dual problem first in Theorem. 2.

Lemma 1. *The primal problem (7) and the dual problem (11) have optimal objective value of 1.*

Proof. We examine the dual problem first. It is easy to see that $\mu^* = 0, \nu^* = 0, \eta^* = 1$ is a feasible solution to the dual problem (11) with value 1. Therefore, the optimal solution of the dual problem is at most 1. We will use contradiction to prove that this dual problem (11) is at least 1. Suppose we have $\mu^*, \nu^*, \eta^* \geq 0$ such that the objective function of the dual problem $\beta(\sum_q \sum_i x_{i,q}^* \mu_{i,q} + \sum_q \nu_q x_{eNB,q}^*) + \eta < 1$. Leveraging an β -approximation algorithm, we know there exists an integral solution $x_{i,q}^p, x_{eNB,q}^p$ satisfying $\sum_q \sum_i x_{i,q}^p \mu_{i,q} + \sum_q \nu_q x_{eNB,q}^p \leq \beta OPT_{LPR}$, where the objective function is $\mu_{i,q}, \nu_q$.

Since $OPT_{LPR} \leq \sum_q \sum_i x_{i,q}^* \mu_{i,q} + \sum_q \nu_q x_{eNB,q}^*$, we then have $\sum_q \sum_i x_{i,q}^p \mu_{i,q} + \sum_q \nu_q x_{eNB,q}^p \leq \beta(\sum_q \sum_i x_{i,q}^* \mu_{i,q} + \sum_q \nu_q x_{eNB,q}^*) < 1 - \eta$, which violates the constraints (12) in the dual problem. The objective value of the dual problem thus is at least 1. This completes the proof that the optimal objective value of the dual problem is 1. According to the strong LP duality, we then have the optimal objective value of the primal problem is 1 as well. The proof is now complete. \square

Theorem 2. *The primal problem (7) can be solved in polynomial time.*

Proof. We use the ellipsoid algorithm to solve the dual problem in polynomial time. In doing so, we use polynomial number of hyperplanes to cut the ellipsoid to get the final optimal solution. Each hyperplane corresponds to a feasible integral solution that is generated by our approximation algorithm of a certain objective function. Each feasible integral solution further corresponds to a variable ω^p in the primal problem. We then have a small-sized primal problem with polynomial

number of variables which enable us to solve it in polynomial time. The proof is complete. \square

C. The randomised winner selection

After solving the primal problem, we regard the weight of each integral solution in the second step as the probability distribution of various deterministic auction results. We select each solution with its weight as the probability. Then the payment for seller s_i is defined as $p_i^* \frac{\sum_q x_{i,q}^p b_{i,q}}{\sum_q x_{i,q}^* b_{i,q}}$. We prove next that our mechanism guarantees system efficiency in expectation, individual rationality and offers truthfulness in its best effort.

We first show that Rado guarantees individually rationality, namely, the utility of the seller is at least zero. We have the expected utility of a seller s_i as follows:

$$E(p_i) - E(cost) = (p_i^* - \sum_q x_{i,q}^* b_{i,q}) \frac{\sum_q \omega^p x_{i,q}^p b_{i,q}}{\sum_q x_{i,q}^* b_{i,q}}.$$

Since the fractional VCG satisfies individual rationality, we have $(p_i^* - \sum_q x_{i,q}^* b_{i,q}) \geq 0$. Therefore, the expected utility of each seller is also greater and equal to zero, meaning that at least sellers can get the price of what they bid for. The expected social welfare is

$$\begin{aligned} & \sum_{i=1}^n \sum_q \omega^p x_{i,q}^p b_i(q) + \sum_q \omega^p C_{eNB,q} x_{eNB,q}^p \\ & \leq \sum_{i=1}^n \sum_q \beta x_{i,q}^* b_i(q) + \sum_q \beta C_{eNB,q} x_{eNB,q}^* \\ & \leq \beta OPT_{LPR}. \end{aligned} \quad (14)$$

Therefore, the expected social welfare of randomised solutions is bounded by β times the social welfare of fractional solutions. As for truthfulness, from the bidder's point of view, it is unclear whether taking other untruthful strategies will generate higher utility due to the randomness nature of our mechanism. Thus, our mechanism offers truthfulness at its best effort.

V. THE APPROXIMATION ALGORITHM

To use the ellipsoid method, we need a separation oracle to find out a violated constraint in each iteration. This constraint gives us an integral solution that can further help us solve the primal problem. We, therefore, propose a Lagrangian based approximation algorithm to act as the separation oracle. Notice that constraint (3) is the only set of constraints that couples the selection variables of different helpers. We naturally use Lagrangian relaxation to decouple these constraints, and transform the complex optimisation problem involving multiple helpers into easier subproblems for each helper.

We introduce non-negative Lagrangian multipliers $\lambda = \{\lambda_j, j = 1, 2, \dots, k\}$ for each D_i in constraints (3) in WDP. To remove constraint (3), we add our penalty term λ into the

original objective function, which then becomes:

$$\begin{aligned} \min & \sum_i \sum_q x_{i,q} (b_{i,q} - \sum_{j=1}^k \lambda_j a_{i,q}^j) + \sum_q x_{eNB,q} (\\ & C_{eNB,q} - \sum_{j=1}^k \lambda_j d_{j,q}) + \sum_{j=1}^k \lambda_j. \end{aligned} \quad (15)$$

It is easy to prove that our newly constructed objective function provides a lower bound for WDP. Leveraging the primal-dual scheme, the Lagrangian dual problem of our WDP is:

$$\max \quad Z(\lambda) \quad (16)$$

$$\text{s.t.} \quad \sum_{q \in D} x_{i,q} \leq 1, \forall s_i \in S \quad (17)$$

$$x_{i,q}, x_{eNB,q} \in \{0, 1\} \quad (18)$$

$$\lambda_j \geq 0, \forall \lambda_j \in \lambda \quad (19)$$

$$\text{where } Z(\lambda) = \min \sum_i \sum_q x_{i,q} (b_{i,q} - \sum_j \lambda_j a_{i,q}^j) + \sum_q x_{eNB,q} (C_{eNB,q} - \sum_j \lambda_j d_{j,q}) + \sum_j \lambda_j.$$

Observing this dual problem, we have the subproblem $Z_i(\lambda)$ for each helper, with $Z_i(\lambda)$ equals to

$$\min \quad \sum_q x_{i,q} (b_{i,q} - \sum_j \lambda_j a_{i,q}^j) \quad (20)$$

$$\text{s.t.} \quad x_{i,q} \leq 1, x_{i,q} \in \{0, 1\}, \forall s_i \in S. \quad (21)$$

We propose an efficient but near optimal algorithm to solve our primal problem as well as its Lagrangian dual problem in Algorithm 2.

Our algorithm involves three steps: the first two steps are based on a subgradient method and work in an iterative fashion; the last one takes a heuristic approach to generate the final solution. We first greedily define λ_j^0 as $\min \frac{b_{i,q}}{|q|}$, $\forall D_j \in D$ where $|q|$ is the number of devices covered by subset q . In each iteration r , in the first step, given λ^r , each subproblem is solved independently. It is easy to see that for each helper, they just need to select one message that has the smallest objective value. Sending this chosen message can in turn satisfy a certain group of users, denoted as q_o . Therefore, $q_o = \arg \min_j (b_{i,q} - \sum_j \lambda_j^r a_{i,q}^j)$, $\forall s_i \in S$. We further determine $x_{i,q}$ as follows: $x_{i,q}$ equals one only if it satisfies $q = q_o$ and $b_{i,q} - \sum_j \lambda_j^r a_{i,q}^j < 0$; $x_{i,q}$ equals zero in other cases. Since λ_j can be interpreted as the marginal benefit of covering an additional user D_j , the benefit of choosing a message should be greater than its cost. In the iteration, we actually set $x_{eNB,q} = 0$ for all subset q on purpose, because we expect all requests to be handled by D2D links.

We next solve our dual problem with a subgradient method, and further update λ in the second step. Given the allocation result in step 1, we update λ^{r+1} as $\lambda_j^{r+1} = \lambda_j^r + h_r(1 - \sum_i \sum_q x_{i,q} a_{i,q}^j)$ if $\lambda_j^r + h_r \lambda_j^r \geq 0$. Correspondingly, this condition can be interpreted as we want the resulting marginal

Algorithm 2 A Lagrangian-based approximation algorithm

- 1: **-A subgradient method**
 - 2: Initiate $\lambda_j^0 = \min \frac{b_{i,q}}{|q|}$, $\forall D_j \in D$
 - 3: **repeat**
 - 4: -Primal subproblem:
 - 5: Given λ , $q_o = \arg \min_q (b_{i,q} - \sum_j \lambda_j^r a_{i,q}^j)$
 - 6: $x_{i,q} = 1$ when $q = q_o$ and $b_{i,q} - \sum_j \lambda_j^r a_{i,q}^j < 0$
 - 7: -Dual multiplier update: Given $x_{i,q}$, $\lambda_j^{r+1} = \lambda_j^r + h_r(1 - \sum_i \sum_q x_{i,q} a_{i,q}^j)$
 - 8: **until** $h_r < \varepsilon$
 - 9: When converge, output $x_{i,q}$
 - 10: **-A heuristic approach to further modify the solution**
 - 11: **repeat**
 - 12: Select the helper with the smallest $b_{i,q}$ from unallocated helpers
 - 13: **if** it could help neighbours **then**
 - 14: $x_{i,q} = 1$, update unallocated helpers and users
 - 15: **else**
 - 16: Update unallocated helpers
 - 17: **end if**
 - 18: **until** All users are satisfied
 - 19: Output the final solution $x_{i,q}, \forall i, \forall q$
-

benefit of covering a request greater than zero. The subgradient of $Z(\lambda)$ is $1 - \sum_i \sum_q x_{i,q} a_{i,q}^j$. h_r is the step length. Step length can be defined into different forms following different step length rules. We set $h_r = \frac{3}{2r+1}$ in this paper. Another commonly used step length is $h_r = \frac{c(Z^* - Z)}{\|1 - \sum_i \sum_q x_{i,q} a_{i,q}^j\|^2}$ where c is a scalar between 0 and 2, and Z^* is an upperbound on Z . The updated λ is then sent back to each subproblem, and new subproblems are solved in the first step again. This iteration ends until step size is smaller than the threshold. After that we have our temporary solution $x_{i,q}$. Since not all users can be satisfied by the D2D links, the solution we have in the previous iteration may not satisfy the constraint that all users must be covered. Therefore, we use a heuristic approach to generate the final solution and satisfy all constraints. We first check for any users that are not satisfied in the second step. We then select the helper with the smallest bid in the unallocated helper set to cover these users, and set its corresponding variable $x_{i,q}$ to one. We repeat the selection until all constraints are satisfied. If after reviewing all unallocated helpers, there are still some unsatisfied users. We turn to the eNB for help, and set the corresponding variable $x_{eNB,q}$ to one. In the end, we have the final allocation solution that satisfies all constraints in WDP.

VI. EVALUATION

In this section, we evaluate the performance of our approximation algorithm through simulations. The basic setting is 1000 devices uniformly distributed in a square area of $1000 \times 1000 m^2$. Each device has a limited D2D communication range of 37 meters. Considering the significant impact

of cache to the content sharing, we enable the cache ability on each device and apply the Least-Frequently Used caching policy. For simplicity, the cache capacities are assumed to be the same. We assume all messages are of unit size. The valuation of messages for each helper in turn varies following the random distribution. Each device independently sends a request with probability 0.5 in each time slot. Message request distribution follows the well-known Zipf-like distribution. The Zipf distribution describes the frequency distributions of ranked data, and key parameter α determines the distribution pattern. The larger the α , the higher the possibility that requests focus on few messages. This distribution has already been proved to be efficient in describing the distribution of users' requests on the Internet. Due to the limited range of D2D communication, D2D is suggested to be designed for relative stationary links [3]. In our current experiment, we also assume that the interaction time among devices is sufficient for their D2D communications.

There are two metrics we care about, offloading ratio and social welfare. Offloading ratio describes how D2D communication can help the eNB offload its traffic. Social welfare denotes the aggregated cost in the whole network. We first inspect the overall performance of our approximation algorithm.

A. The performance of approximation algorithm

We review the overall system performance over 60 time slots in Fig. 2(a). For simplicity, we denote the number of devices as $|D|$, the number of messages as $|M|$. The system parameter settings of Fig. 2(a) are: $\alpha = 2$, $|D| = 1000$, and $|M| = 100$. Since the cache is empty for each device at the beginning, all message requests have to be handled by the eNB initially. The aggregated cost thus is the largest at this moment, correspondingly, the offloading ratio is 0. As time goes by, more and more requests can be satisfied locally, and the aggregated cost decreases. Due to limited cache capacity, finite communication range, and other factors, it is hard to satisfy all requests locally in real cases. The average offloading ratio reaches 69% in Fig. 2(a), and the cost also reduces by 69.2% on average compared to the cost at the beginning.

As the performance of our approximation algorithm is essential to the separation oracle, we use duality gap to evaluate the effectiveness of our approximation algorithm. It is defined as the ratio of difference between the primal and dual objective values divided by the primal objective value here. For a linear programming problem, its duality gap is 0 even under Lagrangian relaxation, namely strong duality holds. Things become non-trivial when our problem is NP-hard. We present the duality gap of our algorithm under different instance sizes in Fig. 2(b). In our experiment, small instances, like $\alpha = 4$, $|M| = 50$, $|D| = 1000$, tend to have better duality gap, and the duality gap of our algorithm is within 5% in most cases. This also means our algorithm generates near optimal solutions.

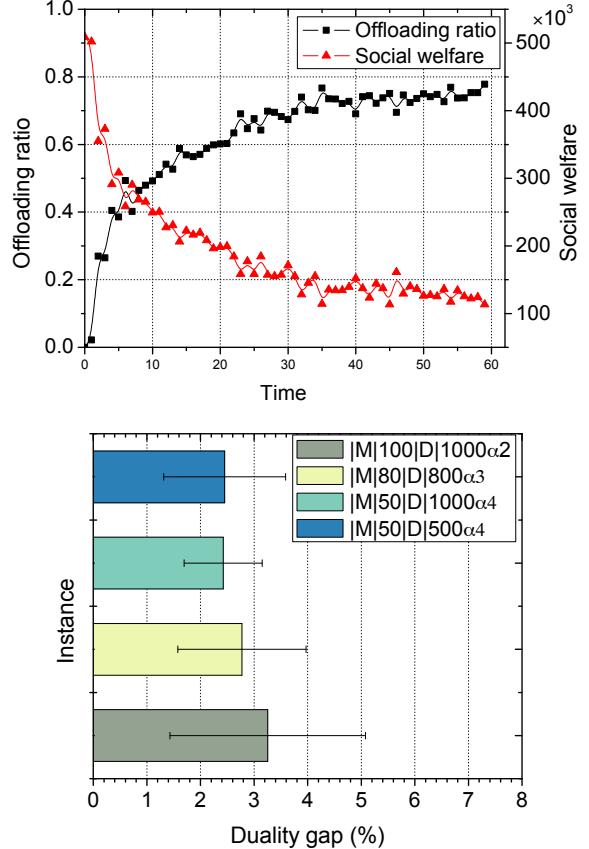


Fig. 2. The system performance and duality gap of Rado.

B. Approximation algorithm under different system parameters

We next examine the system performance of our algorithm under different system parameters in Fig. 3. We start with the influence of α on our algorithm in terms of social welfare in Fig. 3(a). The leftmost point denotes the value of social welfare when all the requests are handled by the eNB. Aggregated social welfare decreases as the value of α grows. The larger the α means the more overlapping interest among users, the broadcast message also serves more users. When $\alpha = 4$, about 70% of the aggregated cost can be saved.

We then test the influence of other system parameters in Fig. 3(b) and Fig. 3(c). We omit the influence of α on the offloading performance, and use the offloading ratios to demonstrate the performance of our algorithm under other parameters in the remainder of evaluation. In Fig. 3(b), average offloading ratio increases as the number of devices increases. Only 27% of requests can be satisfied by D2D communication when we have 100 devices in the network. This number increases to 69% when there are 1000 devices in the network. This is because when we have more devices in the network, the device density increases. Given the limited communication range, each user has more neighbours, and thus has bigger chances to find what it wants in the neighbourhood. Therefore, more requests can be satisfied by the D2D communication.

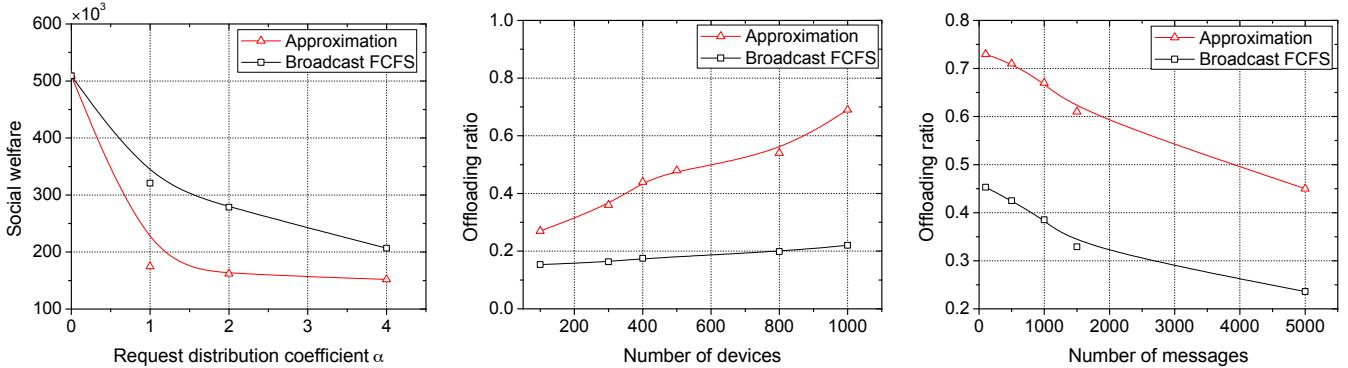


Fig. 3. Our algorithm has better performance when users have more neighbours and more common interests

In Fig. 3(c), the offloading ratio decreases with the increases in number of messages. Recalling that, α and number of messages decide the frequency distribution of message requests. Given the α , the larger the message pool, the smaller common interest among users. All messages become equally important with respect to helpers. When no popular message appears, the offloading ratio decreases drastically. Offloading ratio drops from 73% when we have 100 messages to 45% when we have 5000 messages. Therefore, scattered interests undermine the performance of our algorithm.

C. Randomized auction

We implement our randomised auction mechanism using the proposed approximation algorithm and the ellipsoid method. Given the randomised nature of this auction, we simulate randomised auction 20 times in each time interval, and calculate the expected social welfare of that time period. The performance is compared with that of the fractional VCG mechanism. Noticing that the fractional VCG auction, though has better performance than Rado, is not realistic in real cases. The results are shown in Fig. 4(a) and Fig. 4(b). Figures presented also validate that the average social welfare of the randomized auction is at most β times of the fractional VCG social welfare as proved in the previous section. We further compared the average performance of our randomised auction approach with other two baseline methods, namely, broadcast FCFS and unicast FCFS in Fig. 4(c). Rado has 25% performance gain on average compared with broadcast FCFS, and 30% performance gain when compared with unicast FCFS.

VII. RELATED WORK

Most of previous academic literature in D2D communication assumes devices are altruistic, and they mainly discuss leveraging D2D to improve system performance like spectrum efficiency, power efficiency, and so on. Lin *et al.* in [19] model the multicast D2D scenario, and study the influence of network assistance on multicast D2D performance. It further studies the optimal transmission rate to maximise D2D performance. Pyattaev *et al.* in [20] discuss the possibilities of applying network-assisted WiFi Direct to offload traffic. It demonstrates the improvement of network throughput by randomly choosing

D2D pairs. In our paper, we select trading partners to maximise social welfare rather than just randomly pick one. Besides, without a reward policy, these mechanisms cannot effectively stimulate mobile users to accept D2D communication in the first place.

Incentive issues have also been widely discussed in other types of wireless networks, such as delay tolerant networks (DTNs) [21], [22] and multi-hop cellular networks [23]. Most incentive solutions in DTNs rely on various kinds of probability functions to make decisions due to the opportunistic nature of DTNs. Ning *et al.* in [21] use a specific value function based on the probability of reaching the destination to make their bids. In our paper, we claim that value functions are determined by different user profiles, which may contain unknown number of factors to consider for different users. We thus have no prior knowledge about them, instead we design a mechanism to elicit the true valuations from players. Zhuo *et al.* in [22] construct a reverse auction to motivate users wait longer for future communication opportunities in DTNs in order to offload cellular traffic. While they neglect the incentive for the sender side, and they only accomplish truthfulness without achieving social efficiency after approximately solving the NP-hard Knapsack problem. Salem *et al.* in [23] study the incentive problem to encourage users to relay packets on the route in a multi-hop cellular network. While they focus on designing protocols to support such communication model without exploring how to set the payment properly and how to improve the overall social welfare. In addition, since relaying nodes only help out-of-coverage mobile users to reach the base station [24], all the traffic still has to traverse the cellular network. In contrast, the eNB in our offloading scenario strives to motivate users to handle their requests through neighbours.

The randomised auction framework is based on the decomposition techniques developed in the theoretical computer science field [25], [17]. It successfully translates inapplicable fractional solutions into the convex combination of integral solutions with some desired properties preserved. This framework has just started being used to solve cloud computing [26] and other networking problems [27] recently. Facing various requirements and constraints in different application

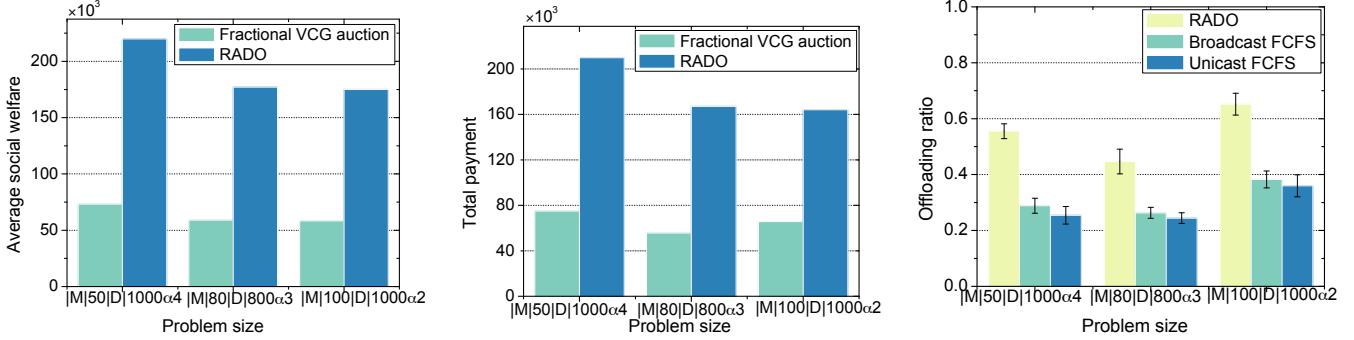


Fig. 4. Average system performance compared with the fractional VCG auction and other baseline methods

scenarios, appropriate auction models and formulations with specific objective goals are proposed. We are also the first to modify this framework to solve D2D problems.

VIII. CONCLUSION

D2D communication emerges as a promising approach for tackling the traffic overload problem in cellular networks. Mobile users, however, need to sacrifice their limited resources to share content with others using D2D. Noticing such barriers, we design Rado, a randomized auction mechanism, to provide incentives for users to participate in D2D content sharing. The randomized decision is realized by solving a well-designed primal-dual linear program to decompose the optimal fractional solution into weighted integral solutions. We further design a Lagrangian-based approximation algorithm to back up our mechanism. We prove that Rado achieves guaranteed social welfare and offers truthfulness in its best effort. Extensive experiments have also demonstrated the significant performance gain of our mechanism compared with other baseline methods.

REFERENCES

- [1] “Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018,” http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.
- [2] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Túriányi, “Design aspects of network assisted device-to-device communications,” *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, 2012.
- [3] K. Doppler, M. P. Rinne, P. Janis, C. Ribeiro, and K. Hugl, “Device-to-device communications; functional prospects for lte-advanced networks,” in *Proc. IEEE ICC*, 2009, pp. 1–6.
- [4] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, “Power optimization of device-to-device communication underlaying cellular communication,” in *Proc. IEEE ICC*, 2009, pp. 1–5.
- [5] K. Doppler, C.-H. Yu, C. B. Ribeiro, and P. Janis, “Mode selection for device-to-device communication underlaying an lte-advanced network,” in *Proc. IEEE WCNC*, 2010, pp. 1–6.
- [6] A. Carroll and G. Heiser, “An analysis of power consumption in a smartphone,” in *Proc. ACM USENIX*, 2010, pp. 21–21.
- [7] S. Jung, U. Lee, A. Chang, D.-K. Cho, and M. Gerla, “Bluetorrent: Cooperative content sharing for bluetooth users,” *Pervasive and Mobile Computing*, vol. 3, no. 6, pp. 609–634, 2007.
- [8] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, “Resource sharing optimization for device-to-device communication underlaying cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, 2011.
- [9] N. Abedini, S. Sampath, R. Bhattacharyya, S. Paul, and S. Shakkottai, “Realtime streaming with guaranteed qos over wireless d2d networks,” in *Proc. ACM MobiHoc*, 2013, pp. 197–206.
- [10] X. Chen, B. Proulx, X. Gong, and J. Zhang, “Social trust and social reciprocity based cooperative d2d communications,” in *Proc. ACM MobiHoc*, 2013, pp. 187–196.
- [11] P. Maillé and B. Tuffin, “Why vcg auctions can hardly be applied to the pricing of inter-domain and ad hoc networks,” in *Proc. IEEE NGI*, 2007, pp. 36–39.
- [12] E. Yaacoub and O. Kubbar, “Energy-efficient device-to-device communications in lte public safety networks,” in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, 2012, pp. 391–395.
- [13] D. Tsolkas, E. Lioutou, N. Passas, and L. Merakos, “Enabling d2d communications in lte networks,” in *Proc. IEEE PIMRC*, 2013, pp. 2846–2850.
- [14] D. Yang, G. Xue, X. Fang, and J. Tang, “Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing,” in *Proc. ACM MobiHoc*, 2012, pp. 173–184.
- [15] E. Balas and M. W. Padberg, “Set partitioning: A survey,” *SIAM review*, vol. 18, no. 4, pp. 710–760, 1976.
- [16] T. Groves, “Incentives in teams,” *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.
- [17] R. Lavi and C. Swamy, “Truthful and near-optimal mechanism design via linear programming,” *JACM*, vol. 58, no. 6, p. 25, 2011.
- [18] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997, vol. 6.
- [19] X. Lin, R. Ratasuk, A. Ghosh, and J. G. Andrews, “Modeling, analysis, and optimization of multicast device-to-device transmissions,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4346–4359, 2014.
- [20] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, “3gpp lte traffic offloading onto wifi direct,” in *Proc. IEEE WCNCW*, 2013, pp. 135–140.
- [21] T. Ning, Z. Yang, H. Wu, and Z. Han, “Self-interest-driven incentives for ad dissemination in autonomous mobile social networks,” in *Proc. IEEE INFOCOM*, 2013, pp. 2310–2318.
- [22] X. Zhuo, W. Gao, G. Cao, and Y. Dai, “Win-coupon: An incentive framework for 3g traffic offloading,” in *Proc. IEEE ICNP*, 2011, pp. 206–215.
- [23] N. B. Salem, L. Buttyán, J.-P. Hubaux, and M. Jakobsson, “A charging and rewarding scheme for packet forwarding in multi-hop cellular networks,” in *Proc. ACM MobiHoc*, 2003, pp. 13–24.
- [24] C. Zhang, X. Zhu, Y. Song, and Y. Fang, “C4: A new paradigm for providing incentives in multi-hop wireless networks,” in *Proc. IEEE INFOCOM*, 2011, pp. 918–926.
- [25] R. Carr and S. Vempala, “Randomized metarounding,” in *Proc. ACM STOC*, 2000, pp. 58–62.
- [26] L. Zhang, Z. Li, and C. Wu, “Dynamic resource provisioning in cloud computing: A randomized auction approach,” in *Proc. IEEE INFOCOM*, 2014, pp. 433–441.
- [27] ———, “Randomized auction design for electricity markets between grids and microgrids,” in *Proc. ACM SIGMETRICS*, 2014, pp. 99–110.