A Reinforcement Learning Approach for User Preference-aware Energy Sharing Systems

Ashutosh Timilsina [†], Atieh R. Khamesi[†], Vincenzo Agate[‡] and Simone Silvestri[†]
[†]Department of Computer Science, University of Kentucky, Lexington, KY, USA.
[‡]Department of Engineering, University of Palermo, Palermo, Italy.
Email: [†]{ashutosh.timilsina, atieh.khamesi, simone.silvestri}@uky.edu,[‡]vincenzo.agate@unipa.it

***Abstract*—Energy Sharing Systems (ESS) are envisioned to be the future of power systems. In these systems, consumers equipped with renewable energy generation capabilities are able to participate in an *energy market* to sell their energy. This paper proposes an ESS that, differently from previous works, takes into account the consumers' preference, engagement, and bounded rationality. The problem of maximizing the energy exchange while considering such user modeling is formulated and shown to be NP-Hard. To learn the user behavior, two heuristics are proposed: a Reinforcement Learning-based algorithm, which provides a bounded regret, and a more computationally efficient heuristic, named BPT-K, with guaranteed termination and correctness. A comprehensive experimental analysis is conducted against state-of-the-art solutions using realistic datasets. Results show that including user modeling and learning provides significant performance improvements compared to state-of-the-art approaches. Specifically, the proposed algorithms result in 25% higher efficiency and 27% more transferred energy. Furthermore, the learning algorithms converge to a value less than 5% of the optimal solution in less than 3 months of learning.**

***Index Terms*—Energy Sharing Systems, Virtual Power Plants, Reinforcement Learning, User Preference.**

I. Introduction

Over the past decade, there has been a growing interest in overcoming the detrimental effects of the energy industry on environment, such as the carbon footprint [1]. In fact, numerous studies have focused on energy-efficient, environmental-friendly, and yet sustainable methods for energy generation [2]. A key enabler of this revolution has been IoT-enabled energy grid, known as Smart Grid (SG) [3], which exploits novel sensing, communication, computing, and control technologies to improve the reliability, resiliency, efficiency, and flexibility of power systems [4], [5]. Recently, several researchers and government bodies have put significant efforts into the evolution of SG technologies towards the paradigm of Virtual Power Plants (VPPs), such as [6], [7]. Unlike the traditional energy systems where the energy generation and distribution are centralized [8], [9], VPPs support a two-way flow of electricity and information [10]. The objective of VPPs is to aggregate Distributed Energy Resources (DER), such as Renewable Energy Technologies (RET) (e.g., photovoltaics (PV), wind power, etc.), into the grid to provide reliable ancillary services, traditionally provided by large power plants [2]. Furthermore, the convenient and low-scale installation and operation of DER and RET will enable widespread adoption at the consumer level. As a result, VPPs represent a paradigm shift where large scale power plants will co-exist, and potentially even be partially replaced, by distributed consumer-level energy generation [2], [9].

Consumers equipped with energy generation capabilities can go beyond self-consumption as they produce surplus energy depending on the generation type and the weather condition [8], [9]. Currently in the U.S., the excess energy is either wasted or sold to the grid [11]. However, this is often not profitable to the consumers since (*i*) the grid usually have a fixed cap on the amount of energy to be purchased from each producer; and (*ii*) the price offered by the grid is often low, non-competitive, and non-negotiable [9]. An alternative is the use of consumer-level batteries for storing the excess energy. However, it has been shown that in order for this to be effective, each home should be equipped with batteries larger than 12kWh, costing more than $\$6,000$ per household [12].

Supported by the emerging paradigm of VPP, a viable and more attractive alternative is to trade the surplus energy between users through an Energy Sharing System (ESS) [13], [14]. In these systems, consumers with renewable energy generation capabilities (called *producers* for simplicity) can sell their energy at more profitable prices, and compete in an energy market with standard energy sources (e.g., nuclear, coal, etc.), and larger renewable energy power plants, to sell their energy to other consumers. ESSs are not only economically more convenient, but they can also contribute in reducing the loss incurred in energy transfer resulting from the closer proximity of users' homes with respect to the utility company [15]. Additionally, ESSs allow to buy energy from different sources, taking into account the consumers' increasing environmental concerns and awareness [16]. An example of a commercial application of ESS is the Dutch start-up Vandebron. Vandebron enables the local renewable electricity generators to sell their energy under an online peer-to-peer (P2P) marketplace platform independent of any utility or government agency [17]. Similarly, Brooklyn Microgrid (BMG) offers an energy marketplace in New York City [18]. BMG uses blockchain technology to allow solar PV owners, in both residential and commercial sectors, to sell excess energy to other NYC residents who prefer to consume the locally-generated renewable energy instead of fossil fuel-based energy.

Previous works on ESSs, such as [13], [15], [19], [20] have proposed matching and/or auction mechanisms to decide how to share energy among local producers and consumers. Authors of [15], in particular, aim at determining a proper producers and consumers matching that minimizes the transmission losses and energy waste. However, these works are mainly based on simplified models of human behavior, for example assuming that users are always available and engaged with the ESS, or will always follow the suggestions that the ESS would recommend through the matching/auction. Few works assume

more realistic models of user behavior in their formulation, for example using prospect theory to model the user's response toward energy prices [20], or modifying classical game theory to better reflect the users' perception towards perceived loss and gain [21], or capturing the user's irrational perception towards bidding results in an energy market [22]. However, these works assume that the parameters of such models are known *a priori*. This is generally an unrealistic assumption. In fact, recent researches in the social science domain have shown that users are highly heterogeneous in their preferences for energy sources [23] and in engaging with energy management systems in general [16]. As a result, previous works in ESS may fail when implemented in the real world [24]–[26].

This paper advances the state-of-the-art in ESSs by considering realistic and heterogeneous user behaviors in terms of preferences and engagement, as well as their limited time and cognitive capabilities in accordance to the principle of *bounded rationality*. A general overview of the ESS system considered in this work is presented in Fig. 1. As depicted in the overview, an energy community which implements an ESS is considered. Within this platform, users are allowed to sell and buy energy to and from other members in the community, as well as from renewable and standard power plants connected to a larger SG, and it is grounded on previous models proposed for ESS [8]. In this system, an active user participation in the energy exchange is envisaged, where users may have different preferences for different energy sources options (e.g., solar, wind, nuclear, coal, etc.), as well as a different level of engagement with the system. According to the proposed approach, ESS periodically (e.g., daily) calculates a prediction and a match of demand and production, to maximize the system performance given the users preferences and level of engagement. The matching is translated into a *personalized recommendation*, sent for example through a smartphone app. This recommendation includes a short list of energy sources and the amount that should be bought from each source to fulfill the consumer's demand. The short size of such list is of primary importance since, according to the principle of bounded rationality, consumers have limited cognitive capabilities and time to select the preferences. Hence, with too many options the user may easily get overwhelmed, leading to the potential abandonment of the system [27], [28]. If a recommendation is accepted by the user, it needs to be honored by the system. Conversely, if a user ignores a recommendation, for example because he/she is not engaged with the ESS, or because the source of energy does not match his/her preferences, the committed energy is wasted due to the limited energy storage at the producer side. In this case, the corresponding demand is supplied from the grid, likely with a higher price. As a result, to maximize the system performance, it is fundamental to take the user behavior into account while matching the produced and consumed energy.

The problem of matching the producers and consumers is formulated as a Mixed Integer Linear Programming (MILP), which aims at maximizing the amount of exchanged energy, while considering the user preference, the size of the recommendation list, as well as physical constraints imposed by the loss of energy in the transfer process. It is shown that the problem is NP-Hard and requires prior knowledge of the user behavioral model. For this reason, a Reinforcement Learning (RL) approach, called User Preference Learning (UPL), is adopted to learn the user preference while optimizing the system performance [29]. It is shown that UPL has a *bounded regret* with respect to the optimal case in which the user preference is known. However, UPL suffers from a relatively long initialization phase and requires to solve the NP-Hard MILP problem at every iteration. To address these shortcomings, Faster Initialization Algorithm (FIA) is proposed, which significantly reduces the time required to complete the initialization phase. In addition, this paper proposes a polynomial heuristic, named BiParTite-K (BPT-K). BPT-K performs multiple maximum weighted bipartite matchings to maximize the energy exchange while implementing the RL framework to learn user preferences. It is formally shown that BPT-K is *totally-correct*, i.e. it always returns a feasible solution of the MILP problem and has guaranteed termination.

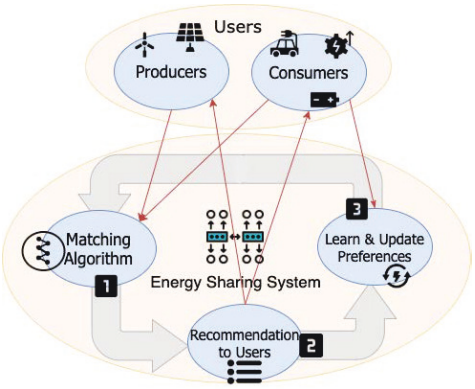

Fig. 1. Energy Sharing System Overview

This work compares the performance of the proposed algorithms with the state-of-the-art approach for ESS and the optimal solution, through simulations using the real traces for consumption and production of energy [30], [31]. Results show that the proposed approach is able to effectively learn the user preference in a much shorter time and significantly improve the performance of the system compared to the state-of-the-art. In summary, the main contributions are as follows.

- This work studies the problem of optimizing the performance of an ESS while considering realistic user behavioral model in terms of preferences, engagement and bounded rationality.
- Problem is formulated as MILP and proved NP-Hard.
- This paper proposes UPL, a heuristic based on Reinforcement Learning (RL), with bounded regret.
- A Faster Initialization Algorithm (FIA), that significantly reduces the initialization phase time of UPL, is proposed.
- A computationally-efficient heuristic, called BPT-K, is developed based on Maximum Weighted Bipartite Matching (MWBM) and RL.
- It is formally proven that BPT-K has polynomial complexity, guaranteed termination, and it is correct.
- Comprehensive experiments based on real data are conducted to evaluate the performance of the proposed algorithms compared with state-of-the-art approaches.

Rest of the paper is organized as follows. The system model and problem statement are described in Sections II and III, respectively. Then, the proposed algorithms are explained in detail in Sections IV and V. Furthermore, the experimental results are elaborated in Section VI. Then, Section VII investigates the related works. Finally, Section VIII concludes the paper and draws a future direction for this study.

II. System Model and Assumptions

The system model in this work consists of two sets of users. P defines the set of producers which includes users equipped with on-site power generators such as PV panels, larger utilities based on renewable energies (e.g., solar, wind, etc.), and traditional power plants (coal, nuclear, hydroelectric, etc.). Similarly, the set of consumers, represented as C, consists of users without power generation capabilities or with a higher consumption compared to their self-production[1].

In the envisioned system, energy exchanges are performed daily, for example during the evening for the next day. For each producer $i \in P$ the ESS estimates the production capacity r_i, and for each consumer $j \in C$ the energy demand w_j, which are expected for the next day. It has been shown that these can be accurately predicted with time-series analysis techniques, such as exponential moving average [15]. This paper considers an ESS in which users, and specifically consumers, have an active role in the exchange process. Specifically, the ESS sends a daily *personalized recommendation* to each consumer through a smartphone app. This recommendation consists of a list of producers, the amount of energy to be bought from each of them, and the cost. The cost may differ for each producer, but it is assumed that such cost does not change over time. Different from previous works in this area, e.g., [15], [32], which consider the users to always be compliant and engaged with the system, the current work considers a realistic user behavioral model in which users may accept, reject, or ignore each of the recommendations in the list. This behavior is dictated by the level of engagement of consumers with the ESS, by their preferences for the source of produced energy (e.g., coal, renewable, nuclear, etc.), and by the price at which energy is sold by a producer. This preference is modeled as a Bernoulli random variable with success probability $p_{ij} \in [0, 1]$, representing the likelihood that consumer j would buy energy from producer i. The probability is initially unknown, and a *Reinforcement Learning* (RL) approach is adopted to learn it. It is assumed that this probability does not change over time. However, several statistical tests, such as the χ^2 test [33] and the Student t-test [34], could be used to detect changes in the user behavior, and restart the learning.

Several studies in the domain of behavioral economics have shown that humans' decisions and actions follow the principle of *bounded rationality* [27]. Specifically, humans possess limited information, time, and cognitive capabilities which prevent them to act optimally. These aspects of human behavior are modeled in this work by limiting the size of the recommendation list to a maximum length K. This reduced size of recommendation list prevents overwhelming the users by reducing the time, information, and effort to select the energy sources to buy energy from. It is also considered that when producer i sells energy to consumer j, there is an *energy loss* during the energy transfer [15]. This loss depends on the physical distance between i and j and it is directly proportional to the amount of energy exchanged. The loss is modeled as a fraction $L_{ij} \in [0, 1]$ of the energy exchanged. It is also assumed that there is a maximum loss threshold L_{max} that the ESS allows and therefore considers only those recommendations that are within this threshold. Moreover, the energy exchanged between two users should be greater than a minimum value α, since it is not convenient to exchange infinitesimal amounts of energy. Note that, if a recommendation is accepted, the ESS will fulfill this exchange. Conversely, if a user ignores or rejects a recommendation, the grid would serve as a backup producer to satisfy the user's demand. Therefore, a recommendation is a *commitment* of energy resources. Consequently, if a recommendation is rejected or ignored, it will result in an energy waste (or in energy sold to the utility company for a much lower price). As a result, recommendations need to be carefully designed to maximize energy exchange and overall performance of system.

[1]In practice, some users may behave as both producers and consumers, also known as *prosumers*, depending on the relative amount of produced and consumed energy. This paper assumes that such roles do not change over time, although the proposed approach could be easily extended to this case.

TABLE I
Notation Summary

Notation	Description
P	Set of producers
r_i	Production capacity of i^{th} producer
C	Set of consumers
w_j	Energy demand of j^{th} consumer
d	Time index corresponding to day
P_{ij}	Random variable corresponding to preference of consumer j buying from producer i
p_{ij}	Mean of P_{ij}
$\widehat{p}_{ij}$	Estimation of p_{ij}
m_{ij}	Number of times producer i has been recommended to consumer j
L_{ij}	Transmission loss between producer i and consumer j
$\mathbf{A}$	RL action matrix
T	Size of the exchangeable unit of energy
K	Max. length of recommendations list

III. Problem Formulation

The goal of the ESS optimization problem is to find the recommendations to be sent to the consumers so that the expected energy exchanged is maximized. This results in minimizing the amount of wasted energy for local producers. Problem formulation is presented in Eq. (1). Table I summarizes the notations used throughout the paper. The decision variables of the problem are $x_{ij} \in [0, 1]$. Given the energy demand w_j of consumer j, x_{ij} represents the fraction of w_j that consumer j is being recommended to buy from producer i. The goal is to maximize the expected amount of exchanged energy, considering the probability p_{ij} with which consumer j will accept the recommendation. The binary decision variable $z_{ij} \in \{0, 1\}$ is equal to 1 if $x_{ij} > 0$, i.e., if producer i is included in the recommendation of consumer j.

$$\text{maximize} \quad \sum_{i \in P} \sum_{j \in C} w_j p_{ij} x_{ij} \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in C} (1 + L_{ij}) w_j x_{ij} \leq r_i, \qquad \forall i \tag{1a}$$

$$\sum_{i \in P} x_{ij} \leq 1, \qquad \forall j \tag{1b}$$

$$\sum_{i \in P} z_{ij} \leq K, \qquad \forall j \tag{1c}$$

$$\alpha z_{ij} \leq w_j x_{ij} \leq w_j z_{ij}, \qquad \forall i, j \tag{1d}$$

$$z_{ij} \geq x_{ij}, \qquad \forall i, j \tag{1e}$$

$$x_{ij} \in [0, 1], \; z_{ij} \in \{0, 1\}, \qquad \forall i, j \tag{1f}$$

The constraint in Eq. (1a) guarantees that the production capacity of producer i is not exceeded, considering the loss that is incurred in the transmission. Similarly, constraint (1b) ensures that the demand of consumer j is not exceeded. The variables z_{ij} are used in the constraint (1c) to make sure that the recommendation list is of maximum length K. Finally, Eq. (1d) certifies that an exchange is larger than the minimum exchangeable allowed amount α, and Eqs. (1e)-(1f) define the domain of the decision variables. Note that, the problem allows exchanges between all pairs of producers and consumers, given the problem constraints. Nevertheless, an additional constraint can be added to prevent losses above the maximum allowed fraction L_{max} by setting $x_{ij} = 0$ if $L_{ij} > L_{max}$.

The following theorem shows that the problem is NP-Hard.

Theorem 1. *The optimization problem in Eq.* (1) *is NP-Hard.*

Proof. In a general instance of GAP [35], there are n tasks and m processors. A task can be assigned to a single process, and the goal is to find the assignment that provides the maximum profit given the resources of the processors. Processor i has r_i resources. By assigning task j to processor i, a profit f_{ij} and resource consumption of g_{ij} is observed. From this general GAP formulation, an instance of the problem can be created through reduction. A consumer for each task and a producer for each processor are created. K is set to 1, so that the recommendation for a consumer can contain at most a single producer. Furthermore, there is $(1 + L_{ij})w_j = g_{ij}$ and the energy production of producer i is set to r_i. It also sets $L_{max} = \infty$ so that all exchanges are possible. At this point, the only difference between the reduced problem and the GAP problem is that the decision variables x_{ij} are continuous, while the decision variable under GAP are discrete. However, infinitesimal exchanges are not allowed in the proposed system, as they need to be greater than or equal to α. By setting $\alpha = w_j$, the constraint in Eq. (1d) forces the decision variable x_{ij} to coincide with the discrete variable z_{ij}. As a result, the solution of the reduced problem provides the assignment that maximizes the profit within the constrained processors' resources. Therefore, the proposed problem is at least as hard as GAP, and thus it is NP-Hard. □

Note that, in addition to the NP-Hardness, the solution of such optimization problem requires the knowledge of the expected user preferences (p_{ij}), the expected production capacity (r_i), and the expected demand (w_j). As mentioned, the latter two can be predicted using time series analysis [15]. Conversely, learning the user behavior is challenging, as users may significantly differ in their preferences and engagement with the ESS [16], [23]. For these reasons, a Reinforcement Learning (RL) approach, called *User Preference Learning* (UPL) is proposed in this paper to learn user preferences, inspired by [29]. UPL consists of the initialization phase that aims at probing the user preferences at least once and optimization phase that requires the optimal solution of a similar version of the optimization problem in Eq. (1) to guarantee the bounded regret. Both phases of UPL are further extended. Specifically, a Faster Initialization Algorithm (FIA) to speed up the initialization phase, and a computationally-efficient heuristic called *BiParTite-K* (BPT-K), based on graph matching theory, for the optimization phase are proposed.

IV. A Reinforcement Learning Approach for User Preference Learning

The optimization problem in Eq. (1) requires the knowledge of the user preferences, expressed in terms of the probabilities p_{ij}. A possible way of predicting the expected user preference is to directly ask users when the ESS is installed in their homes. However, social behavioral studies show that such information does not always reflect the actual preferences. These situations typically occur when users make choices that are not always motivated by a well-defined logic, such as in the case considered in [36]. Given this lack of initial knowledge, it is necessary to learn the users' preferences at run time, by sending recommendations to them while at the same time optimize the system performance. The assumption on independence of the preference probabilities, and the linear nature of the objective function in Eq. (1), allow to formulate this problem through the framework of *combinatorial multi-armed bandit* [29]. Specifically, it is possible to select a subset of the available matches (arms), observe their realization (accept/reject), and gain the linear sum of the outcomes (exchanged energy). This learning process is guided by a balance between *exploration* of the unknown user preference, and *exploitation* of what is already learned.

Reinforcement Learning (RL) is an effective way to solve the multi-armed bandit problem. A naive approach to tackle this problem is to utilize the standard UCB1 algorithm which regards each arm as an independent action [37]. However, this approach ignores the inherent dependencies among the arms, and therefore, ends up learning the information about the observed actions independently [29]. Therefore, a more efficient learning approach is to learn from the observations of the correlated actions and select better decisions based on these correlations. For this reason, this work extends the approach proposed in [29], to the problem of finding the best matching between consumers and producers while simultaneously learning the users preferences. The unknown environment in the problem formulation consists of the players, i.e., consumers; and the available arms, i.e., producers. Besides, the action in this case is the matching between the consumers and producers. Therefore, the reward corresponds to the total energy exchanged among all the consumers and

Algorithm 1: User Preference Learning (UPL)

/* **Initialization Phase** */
for *each* $i \in P$ *and each* $j \in C$ **do**
  Select any $\mathbf{A} \in \mathcal{F}$ s.t. $a_{ij} > 0$;
  Update $[\widehat{p}_{ij}]_{|P|\times|C|}$ and $[m_{ij}]_{|P|\times|C|}$;
end
/* **Optimization Phase** */
while *True* **do**
  $d = d + 1$;
  Select an action $\mathbf{A}$ s.t.

$$\mathbf{A}(d) = \arg\max_{\mathbf{A}\in\mathcal{F}} \sum_{i\in P}\sum_{j\in C} w_j a_{ij}\left(\widehat{p}_{ij} + \sqrt{\frac{(Q+1)\ln d}{m_{ij}}}\right);$$

  Update $[\widehat{p}_{ij}]_{|P|\times|C|}$ and $[m_{ij}]_{|P|\times|C|}$;
end

the producers. The action played during a day d is modeled by the *action matrix* $\mathbf{A}(d)$. The matrix has dimension $|P|\times|C|$ and an element a_{ij} ranges in the interval $[0, 1]$. The value of a_{ij} represents the fraction of demand that producer i is selling to consumer j, similar to the x_{ij} variables of the optimization problem. If $a_{ij} = 0$, there is no exchange between these two actors. Conversely, if $a_{ij} > 0$, a recommendation is sent to consumer j to buy from i. The consumer decision is observed, and the corresponding probability is updated.

Given the action matrix, including acceptance or rejection of a recommendation by the consumer, the preference of consumer j, with respect to accepting a recommendation for buying energy from producer i, is modeled as a random variable P_{ij}. The realization of such variable at day d is referred to as $P_{ij}(d) \in \{0, 1\}$. The mean value of P_{ij} is denoted as p_{ij} and it is initially unknown. It is also assumed that P_{ij} evolves as an i.i.d. process over time. Given the energy consumption/production predictions for day d, the ESS decides which recommendations should be sent to the consumers based on the action matrix for day d, $\mathbf{A}(d) = [a_{ij}(d)]_{|P|\times|C|}$The total number of unknown variables is $Q = |P| \times |C|$. Moreover, the solution space $\mathcal{F}$ includes all feasible action matrices that would satisfy all the constraints of the optimization problem.

Similar to the optimization problem, the amount of exchanged energy is to be maximized in this case too. At each iteration of the optimization phase d, the ESS chooses the action matrix $\mathbf{A}(d)$ that maximizes the optimization function given the current knowledge. This knowledge is represented by the estimated expected $\widehat{p}_{ij}(d)$ for each random variable P_{ij}. For an action matrix $\mathbf{A}(d)$, the *reward* is defined as

$$\mathbf{R}_{\mathbf{A}(d)}(d) = \sum_{i,j} w_j a_{ij}(d) P_{ij}(d). \tag{2}$$

Since the distribution of variables P_{ij} are initially unknown, the goal is to find a policy, denoted by series of action matrices in $\mathcal{F}$, that minimizes the *regret* up to the current time d. This is calculated as the difference between the expected reward having perfect knowledge of the variables realizations and that obtained by the policy. Formally, the regret is expressed as

$$\mathcal{R}(d) = d\mathbf{R}^*_{\mathbf{A}(d)}(d) - \mathbb{E}[\sum_{t'=1}^{d} \mathbf{R}_{\mathbf{A}(t')}(t')], \tag{3}$$

where $\mathbf{R}^*_{\mathbf{A}(d)}(d)$ is the reward obtained with perfect knowledge of users' preferences. Minimizing the regret is a hard problem, given the initially unknown variable distribution. However, an efficient algorithm based on RL is adopted that ensures a *bounded regret* with respect to the optimal [29]. Bounded regret is a desirable property, as it ensures that the algorithm picks a non-optimal action only a limited number of times; which in this case translates into ensuring that in a finite time the optimal set of matches are identified and the best recommendation are sent. This way, the system performance are eventually maximized although the user preferences are initially unknown. The pseudo-code of the algorithm is shown in Alg. 1, namely User Preference Learning (UPL). It is composed of two consecutive phases: *initialization* and *optimization*. During the initialization phase, Q actions are played randomly in order to observe all the Q random variables at least once. Then, in the optimization phase, the system plays an action that maximizes the function defined in line 8 of Alg. 1, over the solution space $\mathcal{F}$. This can be accomplished by solving an optimization problem with the same constraint as in Eqs. (1a)-(1f), and the following objective function:

$$\mathbf{A}(d) = \arg\max_{\mathbf{A}\in\mathcal{F}} \sum_{i\in P}\sum_{j\in C} w_j a_{ij}\left(\widehat{p}_{ij} + \sqrt{\frac{(Q+1)ln\ d}{m_{ij}}}\right), \tag{4}$$

The optimization problem solved at day d is based on the estimation of the expected values p_{ij} at day $(d-1)$, denoted as $\widehat{p}_{ij}(d-1)$. If the selected action at time d includes an energy transaction between consumer j and producer i, i.e., $a_{ij}(d) \neq 0$, a new realization $P_{ij}(d)$ of the random variable P_{ij} is observed. This information is used to update the current knowledge estimation of $\widehat{p}_{ij}(d)$, as well as the total number $m_{ij}(d)$ of observations of the variable P_{ij}, as follows:

$$\widehat{p}_{ij}(d) = \begin{cases} \frac{\widehat{p}_{ij}(d-1)m_{ij}(d-1)+P_{ij}(d)}{m_{ij}(d-1)+1} & \text{if } a_{ij}(d) \neq 0, \\ \widehat{p}_{ij}(d-1) & \text{otherwise.} \end{cases} \tag{5}$$

$$m_{ij}(d) = \begin{cases} m_{ij}(d-1) + 1 & \text{if } a_{ij}(d) \neq 0, \\ m_{ij}(d-1) & \text{otherwise.} \end{cases} \tag{6}$$

Theorem 2. *Let w_j be homogeneous across users for sufficient amount of time, UPL provides bounded regret given by:*

$$\mathcal{R}(d) \leq \left[\frac{4a^2_{max}Q^3(Q+1)\ln(d)}{(\Delta_{min})^2} + \frac{\pi^2}{3}Q^2 + Q\right]\Delta_{max}, \tag{7}$$

where, a_{max} is defined as $\max_{\mathbf{A}\in\mathcal{F}}\max_{i,j} a_{ij}$. Besides, $\Delta_{min} = \min_{\mathbf{R_A}<\mathbf{R}^*}(\mathbf{R}^* - \mathbf{R_A})$ and $\Delta_{max} = \max_{\mathbf{R_A}<\mathbf{R}^*}(\mathbf{R}^* - \mathbf{R_A})$ are the minimum and maximum difference to the reward obtained with perfect knowledge of the users' preferences, respectively.

Proof. The proof is obtained following Theorem 2 of [29]. □

V. A Constrained Maximum Weighted Matching-based Reinforcement Learning Approach

This section first describes the Faster Initialization Algorithm (FIA), to improve the initialization phase of UPL. Subsequently, it discusses the heuristic BPT-K for the optimization

phase. The initialization phase of UPL, similar to the one originally presented in [29], has the purpose of observing each of the Q variables at least once, by selecting random action matrices, before starting the optimization phase. However, Q grows with the number of producers and consumers. Since it takes 24 hours to play an action and observe a realization of the random variables P_{ij}, it would be very inefficient to wait Q days before starting the optimization phase, which serves as the motivation to design FIA. Additionally, given the NP-hardness of the optimization problem in Eq. (1), the optimization phase of UPL is also NP-hard. Therefore, a computationally efficient heuristic algorithm, named BPT-K, is proposed for the optimization phase of UPL to maximize the energy exchange while exploiting RL to simultaneously learn the user preferences. Finally, it is formally proved that the heuristic terminates and it is correct, i.e., it always returns a solution that does not violate the problem constraints. It is also shown that BPT-K has a polynomial complexity.

A. Faster Initialization Algorithm (FIA)

The pseudo-code of the Faster Initialization Algorithm (FIA) is shown in Alg. 2. The primary objective of FIA is to minimize the number of days required to play all variables at least once, in order to meet the requirement of the initialization phase of UPL. Secondarily, the algorithm tries to maximize the amount of satisfied demand of the users corresponding to the played actions. To achieve these objectives, the algorithm keeps track of the already played variables in a binary matrix $\mathcal{B}$, whereby element b_{ij} is equal to 1 if the variable P_{ij} has been played, and zero otherwise. For a given consumer j, each day the algorithm selects at most K previously unassigned producers (i.e., producers such that $b_{ij} = 0$), in order to maximize the number of played actions. Additionally, FIA evenly spreads the demand w_j across such producers (i.e., assigns up to $\frac{w_j}{K}$ to each producer) in order to satisfy the consumer demand. It also excludes variables that cannot be played because they violate the loss threshold L_{max} (line 2).

Algorithm 2: Faster Initialization Algorithm (FIA)

Input : Sets of Producers (P) and Consumers (C), Producer's Capacity($[r_i]_{|P|}$), Consumer's Demand ($[w_j]_{|C|}$), $[m_{ij}]_{|P|\times|C|}$, $[\hat{p}_{ij}]_{|P|\times|C|}$, α
Output: Updated $[m_{ij}]_{|P|\times|C|}$ and $[\hat{p}_{ij}]_{|P|\times|C|}$

1 $\mathcal{B} = [b_{ij}]_{|P|\times|C|} = 0$; // Binary Matrix $\mathcal{B}$ to keep record of actions played
2 $\forall i \in P,\ j \in C$, if $L_{ij} > L_{max}$, then set $b_{ij} = 1$;
/* **Run until all actions are played; J: all-ones matrix** */
3 **while** $\mathcal{B} \neq J$ **do**
4 $\mathcal{A} = [a_{ij}]_{|P|\times|C|} = 0$;
5 **for** $j \in C$ **do**
6 $e = \max\{\frac{w_j}{K}, \alpha\}$;
7 **while** $\big(\sum_{i\in P} a_{ij} < 1\big)$ *and* $\big(\exists i \mid (b_{ij} = 0$ *and* $(r_i \geq e))\big)$ **do**
8 $i \leftarrow$ Select a producer at random from P s.t. $b_{ij} = 0$ and $r_i \geq e$;
9 $a_{ij} = \frac{e}{w_j}$;
10 $r_i = r_i - e$;
11 $b_{ij} = 1$; // Update element $b_{ij} \in \mathcal{B}$
12 **end**
13 **end**
14 Select $\mathcal{A}$ as actions and update $[\hat{p}_{ij}]_{|P|\times|C|}$ and $[m_{ij}]_{|P|\times|C|}$;
15 **end**

The *while* loop (lines 3 – 15) is run until all the elements of $\mathcal{B}$ are equal to 1 (i.e., $\mathcal{B} = J_{|P|,|C|}$). An iteration of the *while* loop identifies the variables to play and the energy exchanges to take place in that day. The matrix $\mathcal{A} = [a_{ij}]_{|P|\times|C|}$ keeps track of the fraction of demand satisfied for that day between consumer j and producer i. An action is played if $a_{ij} > 0$. At each iteration of the *while* loop, the inner *for* loop iterates over the set of consumers C. For each consumer $j \in C$, a random producer i is selected such that the variable P_{ij} was not previously observed (i.e., $b_{ij} = 0$) and also producer i has capacity greater than $e = max\{\frac{w_j}{K}, \alpha\}$ (line 7). The amount of a_{ij}, capacity of the producer r_i, and the elements b_{ij} are updated accordingly (lines 9 – 11). At the end of each iteration of the *while* loop, the actions in $\mathcal{A}$ are played and the observed realizations are updated according to Eqs. (5) and (6). The *while* loop terminates as soon as all variables are observed, and then the optimization phase begins.

B. The BiParTite-K Algorithm

1) Overview: The problem introduced in Eq. (1) is an extension of the generalized matching problem (see Theorem 1), with the additional constraint that consumer-nodes' degrees cannot exceed K (see Eq. (1c)). Recall that such K-constraint is a practical requirement for bounded rationality to prevent overwhelming users with a large list of recommendations [27].

To solve this problem efficiently, inspired by bipartite matching theory, an iterative algorithm, named BiParTite-K (BPT-K) is proposed. In order to perform the assignment, BPT-K uses Maximum Weighted Bipartite Matching (MWBM) as a sub-routine, which can be solved polynomially, for example with the Hopcroft-Karp algorithm [38] or Edmond's Algorithm [39], [40]. Since MWBM provides a one-to-one matching, this would result in significant waste of energy. Therefore, BPT-K enforces a discretization of energy production capacity and consumption demand into *units of exchangeable energy* of size T. BPT-K implements two views of a bipartite graph of producers and consumers, referred to as *aggregated* and *disaggregated* graphs. The vertices of the aggregated graph are the set of producers P and consumers C. In this graph, there exists an edge between a producer and a consumer if they can *potentially* exchange energy, i.e., the loss is less than the threshold L_{max}. Conversely, the disaggregated graph provides a finer grained view based on the notion of unit of exchangeable energy. Specifically, in this graph each consumer demand and producer capacity is expanded into a proportionate number of nodes of equivalent size T. Similar to the aggregated graph, in the disaggregated graph there is an edge between a demand unit of a consumer and a capacity unit of a producer, if the loss between them is within L_{max}. By applying iteratively MWBM on the disaggregated graph, BPT-K allows producers to sell to multiple consumers, and consumers to buy from multiple producers (at most K). This also speeds up the learning rate of user preferences by allowing to probe more variables each day.

The algorithm fulfills two major tasks, namely (i) matching demand and consumption considering the user preference, and (ii) learning such preferences by observing the user

responses to recommendation. As a result, BPT-K combines matching with reinforcement learning to achieve both tasks. The algorithm takes as input the set of producers P and consumers C, with respective capacities and demands, and builds a disaggregated graph G. It returns a matching graph Φ_{out}, with nodes $P \cup C$ and initially no edges. Subsequently, BPT-K runs the MWBM on G resulting in the disaggregated bipartite matching graph Φ_G. Then, Φ_G is used to update Φ_{out} without violating the K-constraint (more details are given in the algorithm description).

Since the proposed algorithm is iterative, this process is repeated until Φ_{out} keeps changing, i.e., the algorithm updates the set of producers and consumers based on residual capacities and demands and repeats the matching iteratively. Once the output graph Φ_{out} is left unchanged, it means that either the producers' capacity and/or the consumers' demand have already exhausted; or there are no possible matching among producers and consumers without violating the K-constraint. Eventually the algorithm breaks out of the loop and terminates by sending the recommendations to the consumers according to the matching expressed by Φ_{out}. At the end of the algorithm, the users' preferences are learned accordingly based on the observed responses. To this aim, the same approach of UPL is adopted, where the preferences and total number of observations are updated according to Eqs. (5) and (6). Note that, the parameter T can be set as a trade-off between complexity and efficiency of the energy exchange. A smaller value of T increases the granularity of the algorithm, thus increasing the amount of exchanged energy. However, such improvement in performance is at the expense of an increased complexity. In Section VI, a sensitivity analysis with respect to the size T is provided. Obviously, T must be set greater than minimum exchangeable allowed energy α (see Eq. (1e)).

2) Algorithm Description: The pseudocode for the BiParTite-K algorithm (BPT-K) is presented in Alg. 3. The output is the graph Φ_{out}, initialized in line 1. The algorithm initializes a temporary graph Φ_{temp} in line 2 used to verify if Φ_{out} has changed. BPT-K is an iterative algorithm so it utilizes a $do - while$ loop (lines $3 - 24$) to run Maximum Weighted Bipartite Matching (MWBM) in an iterative fashion. As explained in the previous subsection, inside the $do-while$ loop, the algorithm starts with the aggregated bipartite graph in order to generate the disaggregated graph, G, based on exchangeable units of energy of size T. To this aim, it first updates the set of producers (P) and consumers (C) to keep only those which have energy capacity and demand greater than or equal to T (lines $4-5$). The algorithm then discretizes the production and demand into the units of size T to obtain the sets P_d and C_d (lines $6-8$), and it builds the disaggregated bipartite graph G using P_d and C_d (line 9). In line 11 weighted edges are added between pairs of nodes in P_d and C_d considering the maximum tolerable loss L_{max} and the K-constraint (lines $10 - 14$). To keep track of the K-constraint, following two conditions are verified. First, for each pair (i, j), corresponding to producer i and consumer j, an edge is added if either j has degree less than K, or secondly it has degree exactly K and has already been assigned to producer i in Φ_{out}. In the pseudocode, degree of node j in Φ_{out} is denoted by $|(., j)|_{E_{\Phi_{out}}}$.

Subsequently, the algorithm computes the Maximum Weighted Bipartite Matching (MWBM) on graph G (line 15) resulting in the graph Φ_G. It then sets $\Phi_{temp} = \Phi_{out}$ and updates Φ_{out} given Φ_G (lines $18 - 23$). For this purpose, the algorithm first sorts the edges in E_{Φ_G} by decreasing weight. Then, for each edge $(u, v) \in E_{\Phi_G}$ it updates the edge in Φ_{out}, between the corresponding producer i and consumer j, only if it does not violate the K-constraint. Then the edge is removed from E_{Φ_G}. The $while$ loop in lines $18-23$ terminates as soon as E_{Φ_G} is empty. If Φ_{out} has changed as a consequence of these updates (line 24), BPT-K performs the next iteration of the $do-while$ loop. Otherwise, it sends the recommendations based on the output graph Φ_{out}, observes the performed exchanges and updates the estimated preferences $\hat{p}_{ij}$ and number of times each preferences has been observed m_{ij}. The algorithm then terminates for the corresponding day and is repeated again for the subsequent day with the new demands and productions based on the latest estimated preferences.

Lemma 1. *BPT-K algorithm returns a feasible solution of the optimization problem in Eq.* (1)*.*

Proof. To prove the Lemma, it is sufficient to show that the solution provided by BPT-K does not violate the constraints of the optimization problem Eq. (1). Since the maximum weighted matching is always performed considering the residual capacity and unsatisfied demand, BPT-K trivially never violates the capacity and demand constraints in Eqs. (1b) and (1c). Moreover, by setting the size of the unit of exchangeable energy $T > \alpha$, constraint (1e) is also satisfied. Finally, the K-constraint in Eq. (1d), requires each consumers to be provided with no more than K recommendations. To this purpose, BPT-K either updates the weights of the existing edges of Φ_{out} (line 20) or adds new edges to Φ_{out} (line 21). A weight update clearly does not violate the constraint. Similarly, an edge is added only if a consumer node j has degree less than K in Φ_{out}, thus preventing to violate the K-constraint. □

Lemma 2. *BPT-K algorithm has a guaranteed termination.*

Proof. BPT-K algorithm consists of a $do - while$ loop (lines $3 - 24$) and other non-iterative instructions. Since the latter are certain of terminating, the rest focuses on the termination of the $do - while$ loop. At the end of each iteration of the $do - while$ loop, the $while$ loop in lines $18 - 23$ updates the weight of existing edges in Φ_{out} (line 20) or adds new edges that do not violate the K-constraint in Φ_{out} (line 21). Each edge update increases the weight of an amount of energy equal to, or larger than, T. Since producers' capacities and consumers' demands are bounded, this update can occur only a finite amount of times. Similarly, an edge (i, j) is added to Φ_{out} only if it does not violate the K-constraint, i.e. if $|(., j)|_{E_{\Phi_{out}}} + 1 \leq K$. Clearly, at most $K \times |C|$ edges can be added. As a result, output graph Φ_{out} can be updated only a finite times, after which the $do - while$ loop terminates.

The proof is concluded by noting that the $while$ loop in lines $18-23$ considers at each iteration an edge $(u, v) \in E_{\Phi_G}$, corresponding to a unit of exchangeable energy assigned

Algorithm 3: BiParTite-K (BPT-K)

Input : Sets of Producers (P) and Consumers (C), Producer's Capacity($[r_i]_{1\times|P|}$), Consumer's Demand ($[w_j]_{1\times|C|}$), Unit of Exchangeable Energy (of size T), Recommendation Size (K), $[m_{ij}]_{|P|\times|C|}$, $[\hat{p}_{ij}]_{|P|\times|C|}$, day (d)
Output: K-Recommendations Graph (Φ_{out}), Updated $[m_{ij}]_{|P|\times|C|}$ and $[\hat{p}_{ij}]_{|P|\times|C|}$

1 $\Phi_{out} = \{P \cup C, E_{\Phi_{out}} = \emptyset\}$;
2 $\Phi_{temp} = \{P \cup C, E_{\Phi_{temp}} = \emptyset\}$;
/* **Iterative matching loop** */
3 **do**
4 Remove from P producers with residual capacity less than T;
5 Remove from C consumers with unsatisfied demand less than T;
/* **Generate disaggregated bipartite graph** G */
6 $\forall i \in P$, let P_i be the set of units of exchangeable energy for producer i;
7 $\forall j \in C$, let C_j be the set of units of exchangeable energy for consumer j;
8 Let $P_d = \bigcup_{i\in P} P_i$ and $C_d = \bigcup_{j\in C} C_j$;
9 Build Bipartite Graph $G = \{P_d \cup C_d, E_G = \emptyset\}$;
10 **for** *each node* $u \in P_i, v \in C_j$ **do**
11 **if** $L_{ij} \le L_{max}$ *and* $\left((|(.,j)|_{E_{\Phi_{out}}} < K)$ *or* $(|(.,j)|_{E_{\Phi_{out}}} = K$ *and* $(i,j) \in E_{\Phi_{out}})\right)$ **then**
12 Add edge (u,v) to E_G with weight, $\mathcal{W}_G(u,v) = \left(T * \left(\hat{p}_{ij} + \sqrt{\frac{(Q+1)\ln d}{m_{ij}}}\right)\right)$;
13 **end**
14 **end**
15 Perform Maximum Weighted Bipartite Matching on G and output graph $\Phi_G = \{P_d \cup C_d, E_{\Phi_G}\}$, where $E_{\Phi_G} \subseteq E_G$;
16 $\Phi_{temp} = \Phi_{out}$;
/* **Add/update the edge in** Φ_{out} **from** Φ_G **without violating the** K**-constraint** */
17 Sort edges in E_{Φ_G} by decreasing weight;
18 **while** $E_{\Phi_G} \neq \emptyset$ **do**
19 Consider next edge $((u,v) \in E_{\Phi_G}$ s.t. $u \in P_i, v \in C_j)$;
20 **if** $(i,j) \in E_{\Phi_{out}}$ **then** update the edge weight, $\mathcal{W}_{\Phi_{out}}(i,j) = \mathcal{W}_{\Phi_{out}}(i,j) + \sum_{\substack{u\in P_i \\ v\in C_j}} \mathcal{W}_G(u,v)$;
21 **else if** $(|(.,j)|_{E_{\Phi_{out}}} + 1 \le K)$ **then** add edge (i,j) to $E_{\Phi_{out}}$ with weight, $\mathcal{W}_{\Phi_{out}}(i,j) = \sum_{\substack{u\in P_i \\ v\in C_j}} \mathcal{W}_G(u,v)$;
22 Remove (u,v) from E_{Φ_G};
23 **end**
24 **while** $\Phi_{out} \neq \Phi_{temp}$;
25 Produce a recommendation list from Φ_{out} and send them to respective consumers;
26 Observe the performed exchanges and update $[\hat{p}_{ij}]_{|P|\times|C|}$ and $[m_{ij}]_{|P|\times|C|}$;

between producer i and consumer j. The loop continues until $E_{\Phi_G} \neq \emptyset$. Since at the end of each iteration the edge (u,v) is removed from E_{Φ_G} (line 22), the $while$ loop also terminates. □

By definition, an algorithm is referred to as *totally-correct*, if it returns a feasible solution and also terminates. Following theorem proves the correctness of BPT-K on the basis of Lemmas 1 and 2.

Theorem 3. *BPT-K, proposed in Alg. 3, is totally-correct.*

Proof. Following the statement made in Lemma 1, BPT-K returns a correct solution. In addition, Lemma 2 guarantees the termination. Therefore, by definition, the BPT-K algorithm is provably totally-correct. □

Theorem 4. *Complexity of the BPT-K algorithm is* $O\left(\min\left\{|P_d|, |C_d|\right\} \times \left(|P_d| + |C_d|\right)^3\right)$.

Proof. The complexity of the algorithm is dominated by the $do - while$ loop (lines 3 – 24). Let $|P_d| = \left\lfloor \frac{\sum_{i\in P} r_i}{T} \right\rfloor$ and $|C_d| = \left\lfloor \frac{\sum_{j\in C} w_j}{T} \right\rfloor$. At each iteration of the $do - while$ loop, an edge weight is updated or an edge is added to Φ_{out}. Lemma 2 shows that the number of such operations is limited. Specifically, the number of edge updates is bounded by $O\left(\min\left\{|P_d|, |C_d|\right\}\right)$ and the number of edges that can be added is bounded by $O\left(K|C|\right)$. Inside the $do - while$ loop there are four main operations: the for loop (lines 10 – 14), the maximum weighted matching (line 15), sorting of the edges in E_{Φ_G}, and the $while$ loop (lines 18 – 23). The for loop has complexity equal to $O\left(|P_d||C_d|\right)$. The maximum weighted matching can be solved with the Edmond's algorithm with complexity $O\left(\left(|P_d| + |C_d|\right)^3\right)$ [39], [40]. The cardinality of E_{Φ_G} is upper bounded by $O\left(|P_d||C_d|\right)$, therefore sorting the edges has complexity $O\left(|P_d||C_d| \log\left(|P_d||C_d|\right)\right)$, and the while loop has a number of iterations upper bounded by $O\left(|P_d||C_d|\right)$. Since the maximum weighted matching algorithm dominates the operations within the $do-while$ loop, and $K|C|$ is generally less than $\min\left\{|P_d|, |C_d|\right\}$, the overall complexity of BPT-K is $O\left(\min\left\{|P_d|, |C_d|\right\} \times \left(|P_d| + |C_d|\right)^3\right)$. □

VI. Experimental Results

In this section, performance of the proposed approaches is evaluated versus a state-of-the-art approach, named Zhu, proposed in [15]. First, the experimental setup is presented, then the Zhu algorithm is described followed by the discussion on the comparison results. Furthermore, the performance of the

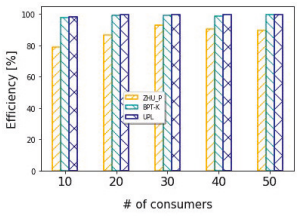

Fig. 2. Efficiency versus number of consumers with constant number of producers.

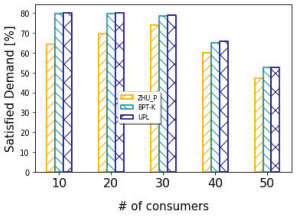

Fig. 3. Satisfied Demand versus number of consumers with constant number of producers.

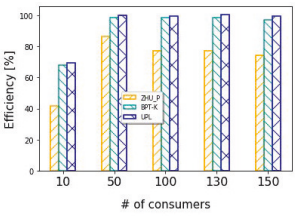

Fig. 4. Efficiency versus number of consumers with constant ratio of consumer-to-producer.

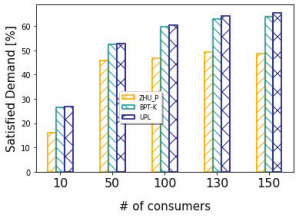

Fig. 5. Satisfied Demand versus number of consumers with constant ratio of consumer-to-producer.

FIA algorithm is investigated and also a sensitivity analysis to relevant parameters of BPT-K is provided.

A. Experimental Setup

Realistic datasets for energy production and consumption are used for experiment. Real consumption dataset is obtained from [30] that contains daily aggregated energy consumption data of 53 residential buildings of different types and sizes over the course of 2014. 16 solar energy producers located in Lexington, Kentucky, USA are considered. These producers are equipped with Photovoltaic (PV) generation capabilities. Half producers are equipped with a 8kW power plant, while the other half with a 4kW power plant. Furthermore, the NREL's PVWatts Calculator of the U.S. Department of Energy [31] is used to generate the energy production over time given the solar irradiance in Lexington and the size of the PV plants. It is assumed that the amount of demand and production for the next day is predicted using an Exponentially Weighted Moving-Average (EWMA) with parameter 0.5. This prediction has been shown to be particularly accurate in [14], [41]. Preference probabilities are selected uniformly at random from the set $\{0.1, 0.5, 1\}$. Additionally, unless otherwise stated, T is set to 1kWh and K is equal to 5. A sensitivity analysis of these parameters is also provided. Finally, losses are assigned uniformly at random from the set $\{1\%, 2\%, 3\%, 4\%\}$, the maximum tolerable loss is $L_{max} = 2.5\%$ and $\alpha = 50$Wh. UPL and BPT-K implement Gurobi optimizer [42] and NetworkX python library respectively.

B. Comparison approach

The proposed algorithms, UPL and BPT-K, are compared to the "Zhu" algorithm presented in [15]. Zhu matches producers and consumers in order to minimize the transmission loss. In this method, consumers are sorted in descending order based on the amount of energy demand. Then, the algorithm follows such order and matches the consumers' demand with the available producers by giving precedence to those that provide the minimum loss. The interested reader is referred to [15] for more details. To the best of our knowledge, [15] is the closest work in context of the proposed system of this paper which aims at finding an optimal matching among the producers and consumers in a localized energy sharing system.

It is to be noted that the Zhu algorithm uses minimization of loss as heuristic for the best match and does not take into account the consumers' preferences nor the maximum size K of the recommendation list as explored in this paper. To provide a fair comparison, a modified version of Zhu algorithm is adopted. This modified version replaces the matching criteria based on loss with the consumers' preferences to maximize the likelihood that the recommendation is accepted. Specifically, it follows sorted order of consumers and matches each consumer j with the producer i that has the highest preference p_{ij} and satisfy the loss threshold L_{max}. Additionally, it stops the matching for consumer j as soon as the number of producers assigned to it reaches K. The modified approach is denoted as "Zhu$_P$". Note that Zhu$_P$ only addresses the matching problem but not the challenge of learning the user preferences. For fairness, it is assumed that Zhu$_P$ has perfect knowledge of such preferences. The experiments compare UPL and BPT-K to Zhu$_P$. Experimental results of the original Zhu algorithm are provided in the conference version of this paper [14].

C. Performance Evaluation

Four experimental scenarios are considered. The first scenario compares performance of the proposed algorithms by scaling the network size. The second scenario focuses on the cumulative reward of RL over time, that is the cumulative energy transfer. In the third scenario, the advantages of the Fast Initialization Algorithm (FIA) is compared to the original initialization of UPL. Finally, the fourth scenario provides sensitivity analysis to study the impact of K and T on the performance of proposed as well as comparison approaches.

Experimental Scenario 1. In this scenario UPL, BPT-K, and Zhu$_P$ are compared with respect to *efficiency* and percentage of *satisfied demand*. Efficiency is defined as the ratio of exchanged energy over the optimum value obtained by solving the optimization problem in Eq. (1) optimally, given the perfect knowledge of the user behavior. The efficiency of each algorithm is calculated every day, and it averages the value over a period of a year. The percentage of satisfied demand refers to the amount of consumers' demand that has been satisfied through exchanged energy, i.e., the recommendations sent by the algorithms and accepted by the consumers.

The purpose of this experiment is to determine how these metrics are affected by the scale of the network. Two possibilities for scaling the network are identified: (i) increasing the number of consumers while keeping the producers constant, and (ii) increasing the number of consumers and proportionally increasing the number of producers. These scenarios are similar, but they present different challenges for the proposed algorithms. In the former, the amount of available

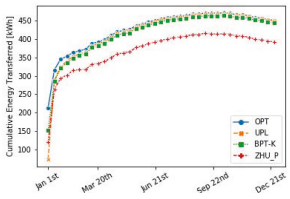

Fig. 6. Cumulative reward (transferred energy) divided by time.

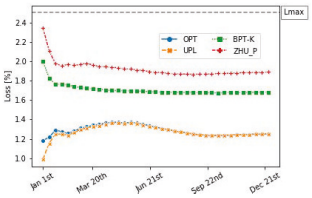

Fig. 7. Percentage of energy losses over time.

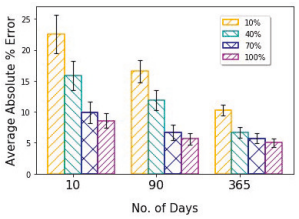

Fig. 8. Average absolute percentage error of preferences learned over time

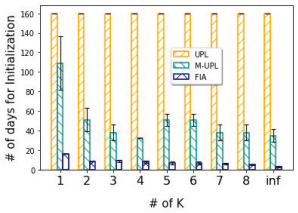

Fig. 9. Number of Days required for initialization vs. K

energy shrinks with respect to the demand, but the number of matching options only increases linearly. Conversely, in the latter, the amount of available energy increases with the network size, but the number of matching options increases quadratically. As a result, the first scenario is more challenging for the matching algorithm, while the second for the learning algorithm, as there are more preferences to learn.

Accordingly, the efficiency achieved by the considered approaches in the first scenario is shown in Fig. 2. Zhu$_P$ shows the worst performance among the considered algorithms, even though it has perfect knowledge of the consumers' preferences. This is due to the greedy nature of this algorithm, which may lead to poor performance when some inadequate greedy decisions are taken. Specifically, to satisfy the demand of a given consumer, Zhu$_P$ assigns all possible energy from one producer before considering the next one. This may prevent to find better solutions where the demand of a user can be satisfied by multiple producers. On the contrary, UPL achieves the best performance both in terms of efficiency and satisfied demand. By exploiting RL and solving the optimization problem on the basis of the current knowledge, UPL is able to achieve performance close to the optimum (i.e., 100% efficiency) in all scenarios at the expense of a higher computational complexity. On the other hand, BPT-K, by means of the iterative matching and RL, is also able to provide a solution close to the optimum while benefiting from a lower complexity.

Fig. 3 shows the percentage of satisfied demand. Since in this case the number of producers are constant (i.e., constant amount of produced energy), the satisfied demand decreases by increasing the number of consumers for all approaches. Nevertheless, UPL and BPT-K significantly outperform Zhu$_P$ even though they need to learn the user preferences through RL. Note that, the satisfied demand under UPL and BPT-K is around 80% until the number of consumers is less than or equal to 30. This is due to two reasons: (i) not enough energy is available for all consumers on some days over the year depending on weather conditions; and (ii) consumers may reject some recommendations, which prevents reaching 100% of satisfied demand although enough energy is available.

In order to further study the scalability of the system, efficiency and satisfied demand are investigated by varying the number of consumers from 10 to 150 and proportionally increasing the number of producers from 3 to 45. Note that, since the datasets used in this paper only provide data for 53 consumers and 16 producers, an augmented dataset was created by selecting producers and consumers at random. This results in an average produced energy which is around 60% of the demand. The results are shown in Figs. 4 and 5. When the number of consumers/producers is low, some recommendations may include less preferred producers, for lack of better alternatives. This results in a lower efficiency. Nevertheless, the efficiency rapidly increases as the scale of the network increases. The efficiency of both UPL and BPT-K reaches values close to 100% around 50 consumers, and remains almost constant after that point. This shows the impressive scalability of the proposed solutions. Conversely, Zhu$_P$ is not able to perform well due to its greedy matching strategy and saturates around 75% only. As a result, our approaches provide a consistent 25% improvement in efficiency compared to the state-of-the-art solution. The percentage of satisfied demand is compared in Fig. 5. The satisfied demand increases as the number of producers and consumers increases. In fact, as the size of the network is increased, there are more producers from whom a given consumer is willing to buy with high probability (i.e., preference). The maximum satisfied demand approaches 65% (i.e., the production to demand ratio) with 150 consumers and 45 producers, as most consumers receive highly preferred recommendations. Also in this case, the RL-based algorithms, UPL and BPT-K, significantly outperform Zhu$_P$.

Experimental Scenario 2. This scenario studies the widely adopted measure of RL algorithms, that is the cumulative reward over time. In this case, it corresponds to the cumulative energy exchanged. To this purpose, for each day d, the cumulative energy exchanged up to that day is calculated, then it is divided by d. Note that, to better focus on the reward, the results are shown after the initialization phase of UPL and BPT-K has completed. As a result, day $d = 0$ corresponds for UPL and BPT-K to the first day after the end of their respective initialization, which may have a different length for each algorithm. The length of the initialization phase is explicitly studied in experimental scenario 3.

As the results presented in Fig. 6 show, UPL outperforms all approaches, demonstrating outstanding performance with negligible gap with respect to the optimal solution which assumes perfect knowledge of the user preferences. This results from the ability of UPL to quickly learn the user preferences and by solving the optimization problem optimally at each iteration. Once the preferences are sufficiently learned, UPL and OPT provide the same solution. BPT-K shows a reward that closely matches the performance of UPL, without

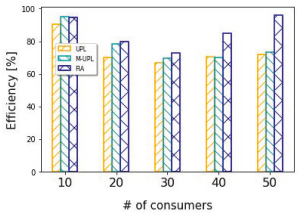

Fig. 10. Efficiency of the initialization algorithms vs. No. of consumers with constant number of producers.

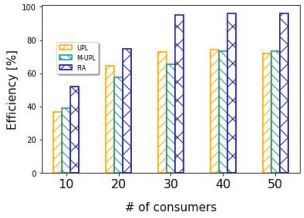

Fig. 11. Efficiency of initialization algorithms vs. No. of consumers with constant ratio of consumer-to-producer.

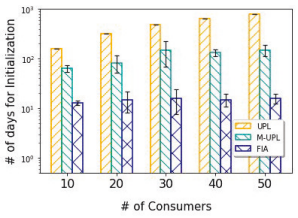

Fig. 12. Number of days required for initialization vs. number of consumers with constant number of producers.

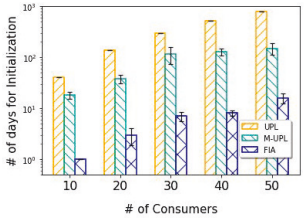

Fig. 13. No. of days required for initialization vs. No. of consumers with constant ratio of consumer-to-producer.

requiring the solution of a NP-Hard problem at each time step. On the contrary, Zhu$_P$ clearly exhibits its inability to provide satisfactory performance due to its greedy matching approach. Overall, both UPL and BPT-K are within 5% of the optimal solution in less than three months of learning. Additionally, they provide an average 27% gain in energy transferred compared to Zhu$_P$. It is worth noting that since realistic energy production data is obtained from [31], there is a seasonal effect causing the non-monotonic trend of all approaches in Fig. 6. In fact, during the Fall/Winter months there is a decrease in energy production of solar panels, which implies a decrease in the exchanged energy.

For completeness, Fig. 7 illustrates the percentage of energy loss. In these experiments, there is $L_{max} = 2.5\%$. None of the algorithms is specifically targeting loss as an optimization metric, as long as no more than L_{max} energy is lost in each transaction. As a result, all approaches incur a loss lower than L_{max}. Finally, the rate of learning user preferences under various ratios of produced energy versus demand is studied. To this purpose, Fig. 8 shows the average percentage absolute error in learning the consumers' preferences, i.e., the probabilities p_{ij}, under UPL. Results for BPT-K were omitted because similar trends were observed. In these experiment, the produced energy is given as a percentage (10%, 40%, 70%, and 100%) of the demand, and corresponding learning error after 10 days, 3 months, and 1 year is observed. Intuitively, when less energy is produced, less exchanges are possible which results in a slower learning phase. As expected, the error decreases as the amount of energy increases, as well as with time. It is worth noting that, even under just 10 days, the error is below 10% if at least 70% of the required energy is available. Interestingly, the error never reaches zero, and it tends to stabilize around 5%. This is due to the nature of reinforcement learning, which prefers exploitation over exploration, once sufficient knowledge is acquired. In fact, once the best matches (i.e., those with higher chances of acceptance) are identified, these are selected more often, in order to maximize the system performance. As a result, other consumers' preferences are not learned exactly but this does not prevent the system from achieving high efficiency.

Experimental Scenario 3. In the third scenario, the performance of the Faster Initialization Algorithm (FIA) is studied. Both the primary objective, i.e., minimizing the number of days required to complete the initialization phase, as well as the secondary objective, which is improving the amount of exchanged energy during the initialization, are considered. In this scenario, FIA is compared with the initialization phase of UPL originally proposed in [29]. The goal of the original initialization is to probe all the variables (here preferences) at least once. However, the UPL initialization has a fixed duration of $|P| \times |C|$ days, due to the for loop in Alg. 1 line 1. For a fair comparison, a modification of this approach is adopted, called "M-UPL", wherein the algorithm breaks out of the for loop as soon as the goal of probing all the variables at least once is met. First the number of days required to complete the initialization phase is studied by varying the value of K from 1 to 8. It is also considered the case of $K = \infty$, corresponding to no limit on the size of the recommendation list. The number of consumers and producers are constant and equal to 10 and 16, respectively. As shown in Fig. 9, FIA is able to significantly shorten the initialization time by maximizing the number of probed variables at each iteration, without violating the K-constraint. Conversely, the original UPL initialization has a constant initialization time of $|P| \times |C| = 160$ days. Modified version M-UPL improves the performance of UPL, but it still achieves a termination time which is 7 times higher than FIA on average. This is due to the fact that M-UPL probes single variable at every iteration, while selecting other variables randomly until all variables are probed at least once.

Next, the impact of network size with respect to the length of initialization time is studied. Similar to experimental scenario 1, number of consumers is increased by keeping number of producers constant and also by increasing the producers proportionally. These experiments set $K = 5$. Figs. 12 and 13 present the results. In both cases, FIA significantly outperforms UPL and M-UPL (note the log-scale on the y-axis). Note that the initialization time increases more significantly for all approaches when number of producers increases with the number of consumers. This is due to the number of variables that increases linearly when producers are kept constant, and quadratically when they scale with the number of consumers.

Finally, the experiment focuses on the secondary objective of FIA, which is the amount of exchanged energy. To this purpose, the efficiency of FIA, UPL, and M-UPL are compared during their respective initialization phases. The efficiency is calculated as the total amount of exchanged energy by the algorithms, during the initialization phases, divided by energy exchanged by the optimal solution, with perfect knowledge of preferences, during the same period. K is fixed at 5 and number of consumers and producers is increased as before. Results are presented in Figs. 10 and 11. As the figures show

FIA, even not targeting energy exchange as primary objective, significantly outperforms the original UPL and M-UPL.

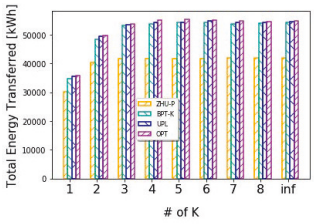

Fig. 14. Energy Transferred vs. K.

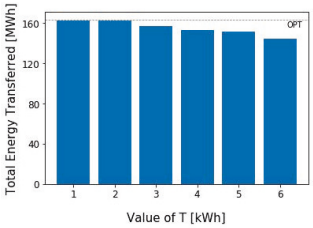

Fig. 15. Energy Transferred vs. T.

Experimental Scenario 4. The final experimental scenario performs a sensitivity analysis to investigate the impact of the values K and T on the performance of the proposed methods and the comparison approaches. First, it focuses on the value of K. In this experiment, K varies from 1 to 8 and it also includes $K = \infty$. The number of consumers and producers are equal to 10 and 16, respectively. Fig. 14 illustrates the total energy transferred as a function of K. As observed, UPL and BPT-K outperform Zhu$_P$ for each value of K and perform close to the optimum. The technique of discretization into unit of exchangeable energy of size T, results in slightly worse performance of BPT-K compared to UPL. Similar to previous experiments, Zhu$_P$ performs worse than others even with perfect knowledge of users' preferences. Numerically, Zhu$_P$ saturates at 75% of the optimum value, while UPL and BPT-K reach 99% and 98%, respectively. This experiment reveals that the proposed system aligns well with social-science and behavioral economic theories of bounded rationality [27]. In fact, there is no noticeable performance improvement for values of K larger than 5. Therefore, sending a shorter list of recommendations to consumers is convenient for them to interact with the system, without sacrificing the system performance. Finally, sensitivity analysis of BPT-K to size of the unit of exchangeable energy T is presented in Fig. 15. The trend of the total energy transferred over a year, by varying T from 1kWh to 6kWh. It also sets $K = 5$ and considers 50 consumers and 16 producers. For $T = 1$kWh, the total energy exchange is close to the optimal value. Increasing T reduces the algorithm computational complexity (see Theorem 4), at the expense of a small decrease in performance.

VII. Related Works

Energy industry, and specifically electric power sector, is responsible for a major emissions of carbon worldwide [43]. Today, coal-fired generators supply 41% of total electricity requirement while they account for 73% of global greenhouse gas (GHG) emissions [44]. Accordingly, green energy industry is at the core attention. To this aim, the integration of Distributed Energy Resources (DERs) such as Renewable Energy Technologies (RET) in power systems has been introduced [4], [45]. DERs, defined as the power generations at the customer side of the distribution network [46], can be aggregated to optimize generation, storage, as well as demand-side resources for maximizing the utility of both the end-users and the grid operator [47]. The idea of aggregating DERs has resulted into the paradigm of Virtual Power Plants (VPPs), which has attracted significant interest from both the academic and industry community [7], [48]. Integrating renewable energy generation into Smart Grids (SGs) already exists in the form of net-metering but presents critical challenges, such as energy fluctuation in the grid. To reduce such fluctuation, [49] studied the trade-off between use of storage and DERs. The results showed that in absence of large storage, the grid can notably gain from exchange energy among the users. Besides, a recent work shows that the energy mismatch within and between microgrids pose a significant problem which needs to be optimized through an energy market to reduce the dependency on the grid [32]. This work adopts game theoretic and hierarchical optimization approaches to minimize the power mismatch in and among microgrids in a multiagent-based energy market. Unlike the mentioned works which aim at integrating the produced energy into the grid, this work proposes an ESS to trade energy among the users locally in order to avoid energy fluctuation in the grid altogether. Further, it has been shown that trading of this excess energy with grid is neither profitable nor flexible for producers [9], [50].

There have been several works focusing on the possible energy exchanges among large producers, the grid, microgrids, and also among small-scale local producers and consumers. In [15], a DC power sharing among nearby homes has been introduced to address the problem of mismatching between energy harvesting and consumption in microgrids. It presents a greedy approach that maximizes the energy exchanges while minimizing loss and energy waste. Furthermore, self-consumption of locally generated energy in a microgrid scenario has been studied in [13], which presents a peer-to-peer (P2P) energy sharing model with price-based demand response (PBDR) program. The efficacy of the method is verified in terms of cost-savings and improved energy-exchange. A privacy-preserving framework is proposed in [51] to facilitate the coordinated operation of large-scale operators like renewable power system operators and private industrial energy hub operators while minimizing overall operation costs. Similarly, an offering and bidding mechanism for a hybrid power producer is proposed by [52] incorporating the intra-day trading mechanism with traditional day-ahead trading models to increase the profitability and minimize the risks. Solution to coordinated distributed generation and demand response management problem has been presented in [53] using multi-agent based approach. However, it does not consider the autonomous decision making among the concerned agents to determine the solution. Operational management of multi-microgrid system is modelled using a joint constraint in a cooperative manner in [54] using stochastic predictive control mechanism. Local energy trading has also been explored among the interconnected microgrids in [55] and [56] in consideration with uncertain parameters in the system. Energy exchange and management between microgrids is achieved in [55] with focus on utilising the chance-constrained programming to model uncertain parameters of the system; while [56] presents a hybrid approach of information-gap decision theory and stochastic programming to capture uncertainties in energy trading among microgrids and proportionate cost-saving among them based on their size.

These papers, however, do not consider the case of local energy trading among the users themselves.

In regards to energy trading among the users in a peer-to-peer (P2P) approach, authors in [57] discuss the existing P2P electricity trading technologies and the challenges associated with them. As per their findings, the existing techniques on P2P energy trade is based on one of the following: game theory, auction mechanisms, constrained optimization and blockchain. It further notes that based on current emergence of SG and blockchain technologies, the deployment of P2P energy trading can provide a highly efficient and cost-effective energy management technique in a decentralized way. In [58], the authors design a decentralized algorithm for an energy trading market with renewable energy generators and price-responsive load aggregators. The goal is to propose a receding horizon energy trading algorithm for the load aggregators and generators. Although they address uncertainty of energy demand and energy production, they do not consider real user involvement. Note that, none of the above mentioned papers considers the complex aspects of user behavior, thus assuming users to be either extraneous to the system or perfectly compliant with the system decision. Therefore, the lack of realistic modeling may cause failure when implementing ESS in the real world [25], [59], [60].

Modeling user behavior in SGs has been considered in the context of Demand Response (DR) [61], [62] that is concerned with preventing the occurrence of demand peaks. For instance in price-based DR, the price of electricity is changed dynamically to alter the user behavior. The authors in [20] use a reverse approach, in which *prospect theory* is used to model the user response to energy prices, and focus on the impact of such realistic behaviors on the system. Despite relatively easy implementation through the use of advanced metering infrastructure (AMI) [63], [64], DR adoption rates are low [65], and its effectiveness is not clear as it can even lead to an increase of energy consumption [66].

In addition to [20], recently there has been some efforts in integrating prospect theory in energy related application to capture the irrationality of users under uncertain decision making [21], [22], [67]. These papers notice that the classical game theoretic approaches consider users to be rational decision makers which does not reflect the actual scenario exhibited by the users, specially under uncertain situation where the users may deviate from rational decision making to avert the perceived loss or magnify the perceived utility. In [21], authors present a framework for energy storage management to allow users to store or sell the energy while modelling the user's subjective perceptions of probable outcomes using Prospect Theory. Similarly, [67] uses Stackelberg Game Theory to optimize energy trading between prosumers and grid where the players make decisions in the face of uncertain future energy price using framing effect in prospect theoretic framework that accounts for deviation of utility from a certain reference point. There has also been an effort on modelling the user's perception towards bidding result in a power market [22], that uses genetic algorithm for solving the optimal power market bidding problem. Although these papers model the user behavior more realistically, they assume that such behavior (e.g., the parameters of the Prospect Theory curve) is known a priori. Social science studies, such as the one conducted in Italy to investigate the social acceptance of nuclear energy using an online survey [23], show that users exhibit significant heterogeneity in their preferences for the sources of energy. In fact, it is found that the preferences of users are affected by not only the environmental aspects but also the financial aspects resulting from the installation of DERs and also the engaging with energy management systems in general [16]. Not capturing such heterogeneity provides little benefits in terms of user modeling. For this reason, this paper focuses on learning the users' preferences in the optimization of the ESS using a Reinforcement Learning (RL) approach based on exploration and exploitation trade-off while simultaneously maximizing the system performance. Recent work in [68] is the only approach that applies Multi-agent Q-learning algorithm to tackle the problem of energy consumption scheduling for home energy management by minimizing both electricity price and DR induced dissatisfaction. However, similar to previous works in this context, this paper also relies on a simplified user modeling of dissatisfaction, i.e., a quadratic function of energy consumption difference, and thus lacks realistic psychological user behavior model. Furthermore, their problem setting is limited to a house level and particularly not suited for the proposed model where a community level engagement between producers and consumers is envisioned.

To the best of the authors' knowledge, this paper is the first effort that combines optimization, reinforcement learning, and user behavioral modeling in the context of ESS under a Virtual Power Plant (VPP) framework.

VIII. Conclusion

This work studies the problem of exchanging locally-generated energy through an Energy Sharing System (ESS) enabled by the paradigm of Virtual Power Plants (VPPs). This problem was formulated as a Mixed-Integer Linear Programming (MILP) and proved its NP-Hardness. Unlike the existing works that mostly overlook or oversimplify the role of human behavior in ESSs, a realistic user behavioral model in terms of the consumer preference, engagement, and bounded rationality is incorporated. To learn the user preferences, a Reinforcement Learning (RL)-based algorithm called User Preference Learning (UPL) is proposed. An efficient RL heuristic is also proposed, called BPT-K, which is based on Maximum Weighted Bipartite Matching (MWBM). Extensive experimental evaluation, with real energy consumption and production data, show that the proposed approaches perform close to the optimal and substantially outperform the comparison method.

However, there are different potential open problems that could lead to further research in this domain. Specifically, it is assumed that the user behavior can be modeled as independent probabilities and it is stationary over time. There could also be effects of previous actions on future actions, and the user behavior may mutate over time, which could be further explored. Additionally, the impact of price is also not explicitly modeled in this work. In practice, lower prices may drive a user towards less preferred sources of energy.

Furthermore, the proposed approaches requires humans to be actively engaged with the system, which might be challenging to sustain over long period of time; making a semi-automated system more practical. Finally, an additional future direction for this research can include integrated energy storage systems and considering a more complex model user preferences that reflects the irrational and variable nature of human decision would be another interesting research scope.

ACKNOWLEDGMENT

This work is supported by the National Institute for Food and Agriculture (NIFA) under the grant 2017-67008-26145, the NSF grant EPCN 1936131, and the NSF CAREER grant CPS-1943035.

REFERENCES

[1] M. Moretti, S. N. Djomo, H. Azadi, K. May, K. De Vos, S. Van Passel, and N. Witters, "A systematic review of environmental and economic impacts of smart grids," *Renewable and Sustainable Energy Reviews*, vol. 68, pp. 888–898, 2017.

[2] W. Strielkowski, *Social Impacts of Smart Grids: The Future of Smart Grids and Energy Market Design*. Elsevier, 2019.

[3] U.S. Department of Energy, "The smart grid: An introduction," U.S. Department of Energy, Tech. Rep., November 2008.

[4] M. H. Rehmani, M. Reisslein, A. Rachedi, M. Erol-Kantarci, and M. Radenkovic, "Integrating renewable energy resources into the smart grid: Recent developments in information and communication technologies," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 2814–2825, 2018.

[5] S. Ciavarella, J.-Y. Joo, and S. Silvestri, "Managing contingencies in smart grids via the internet of things," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 2134–2141, 2016.

[6] J. Johnson, J. Flicker, A. Castillo, C. Hansen, M. El-Khatib, D. Schoenwald, M. Smith, R. Graves, J. Henry *et al.*, "Design and implementation of a secure virtual power plant," *Sandia Technical Report*, 2017.

[7] O. Palizban, K. Kauhaniemi, and J. M. Guerrero, "Microgrids in active network management—part i: Hierarchical control, energy storage, virtual power plants, and market participation," *Renewable and Sustainable Energy Reviews*, vol. 36, pp. 428–439, 2014.

[8] Y. Parag and B. Sovacool, "Electricity market design for the prosumer era," *Nature Energy*, vol. 1, p. 16032, March 2016.

[9] O. Jogunola, A. Ikpehai, K. Anoh, B. Adebisi, M. Hammoudeh, S.-Y. Son, and G. Harris, "State-of-the-art and prospects for peer-to-peer transaction-based energy system," *Energies*, vol. 10, no. 12, 2017.

[10] K. Wang, X. Hu, H. Li, P. Li, D. Zeng, and S. Guo, "A survey on energy internet communications for sustainability," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 3, pp. 231–254, July 2017.

[11] (2009) Freeing the grid: Best and worst practices in state netmetering policies and interconnection procedure. [Online]. Available: http://www.newenergychoices.org/uploads/FreeingTheGrid2009.pdf

[12] T. Zhu, A. Mishra, D. Irwin, N. Sharma, P. Shenoy, and D. Towsley, "The case for efficient renewable energy management in smart homes," in *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 2011, pp. 67–72.

[13] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, and J. Lei, "Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3569–3583, Sep. 2017.

[14] V. Agate, A. R. Khamesi, S. Silvestri, and S. Gaglio, "Enabling peer-to-peer user-preference-aware energy sharing through reinforcement learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–7.

[15] T. Zhu, Z. Huang, A. Sharma, J. Su, D. Irwin, A. Mishra, D. Menasche, and P. Shenoy, "Sharing renewable energy in smart microgrids," in *2013 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, April 2013, pp. 219–228.

[16] E. Ropuszyńska-Surma and M. Węglarz, "Profiling end user of renewable energy sources among residential consumers in poland," *Sustainability*, vol. 10, no. 12, p. 4452, 2018.

[17] Vandebron. [Online]. Available: https://vandebron.nl/

[18] Brooklyn micorgrid. [Online]. Available: https://www.brooklyn.energy/

[19] S. Althaher, P. Mancarella, and J. Mutale, "Automated demand response from home energy management system under dynamic pricing and power and comfort constraints," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1874–1883, July 2015.

[20] K. Jhala, B. Natarajan, and A. Pahwa, "Prospect theory based active consumer behavior under variable electricity pricing," *IEEE Transactions on Smart Grid*, pp. 1–1, 2018.

[21] W. Saad, A. L. Glass, N. B. Mandayam, and H. V. Poor, "Toward a consumer-centric grid: A behavioral perspective," *Proceedings of the IEEE*, vol. 104, no. 4, pp. 865–882, 2016.

[22] Y. Wang, L. Zhang, Q. Ding, and K. Zhang, "Prospect theory-based optimal bidding model of a prosumer in the power market," *IEEE Access*, vol. 8, pp. 137 063–137 073, 2020.

[23] D. Contu, E. Strazzera, and S. Mourato, "Modeling individual preferences for energy sources: The case of iv generation nuclear energy in italy," *Ecological Economics*, vol. 127, pp. 37–58, 2016.

[24] V. Dolce, C. Jackson, S. Silvestri, D. Baker, and A. DePaola, "Social-behavioral aware optimization of energy consumption in smart homes," in *2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 2018, pp. 163–172.

[25] S. Silvestri, D. A. Baker, and V. Dolce, "Integration of social behavioral modeling for energy optimization in smart environments," in *ACM Social Sense*, 2017, pp. 97–97.

[26] A. R. Khamesi, E. Shin, and S. Silvestri, "Machine learning in the wild: The case of user-centered learning in cyber physical systems," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2020, pp. 275–281.

[27] G. Gigerenzer and R. Selten, *Bounded rationality: The adaptive toolbox*. MIT press, 2002.

[28] A. Szollosi and B. R. Newell, "People as intuitive scientists: Reconsidering statistical explanations of decision making," *Trends in Cognitive Sciences*, 2020.

[29] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, Oct 2012.

[30] Pecan street inc. [Online]. Available: www.pecanstreet.org

[31] Solar Resource Data. [Online]. Available: pvwatts.nrel.gov/pvwatts.php

[32] M. M. Esfahani, A. Hariri, and O. A. Mohammed, "A multiagent-based game-theoretic and optimization approach for market operation of multimicrogrid systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 280–292, Jan 2019.

[33] P. E. Greenwood and M. S. Nikulin, *A guide to chi-squared testing*. John Wiley & Sons, 1996, vol. 280.

[34] D. Hull, "Using statistical testing in the evaluation of retrieval experiments," in *Proceedings of 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993, pp. 329–338.

[35] L. Fleischer, M. X. Goemans, V. S. Mirrokni, and M. Sviridenko, "Tight approximation algorithms for maximum general assignment problems," in *Proceedings of 17th ACM-SIAM symposium on Discrete algorithm*. Society for Industrial and Applied Mathematics, 2006, pp. 611–620.

[36] D. Kahneman, "Maps of bounded rationality: Psychology for behavioral economics," *American Econ. Rev.*, vol. 93, no. 5, pp. 1449–1475, 2003.

[37] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[38] J. E. Hopcroft and R. M. Karp, "An n^5/2 algorithm for maximum matchings in bipartite graphs," *SIAM Journal on Computing*, vol. 2, no. 4, pp. 225–231, 1973.

[39] J. Edmonds, "Maximum matching and a polyhedron with 0, 1-vertices," *Journal of Research of the National Bureau of Standards B*, vol. 69, no. 125-130, pp. 55–56, 1965.

[40] Z. Galil, "Efficient algorithms for finding maximum matching in graphs," *ACM Computing Surveys (CSUR)*, vol. 18, no. 1, pp. 23–38, 1986.

[41] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 6, no. 4, p. 32, 2007.

[42] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2020. [Online]. Available: http://www.gurobi.com

[43] C. Kang, T. Zhou, Q. Chen, J. Wang, Y. Sun, Q. Xia, and H. Yan, "Carbon emission flow from generation to demand: A network-based model," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, p. 2386, 2015.

[44] M. Pourakbari-Kasmaei, M. Lehtonen, J. Contreras, and J. R. S. Mantovani, "Carbon footprint management: a pathway toward smart emission abatement," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 935–948, 2019.

[45] S. Hadayeghparast, A. S. Farsangi, and H. A. Shayanfar, "Day-ahead stochastic multi-objective economic/emission operational scheduling of a large scale virtual power plant," *Energy*, vol. 172, pp. 630–646, 2019.

[46] T. Ackermann, G. Andersson, and L. Söder, "Distributed generation: a definition," *Electric Power Systems Research*, vol. 57, no. 3, pp. 195 – 204, 2001.

[47] P. Asmus, "Microgrids, virtual power plants and our distributed energy future," *The Electricity Journal*, vol. 23, no. 10, pp. 72–82, 2010.

[48] M. Vasirani, R. Kota, R. L. Cavalcante, S. Ossowski, and N. R. Jennings, "An agent-based approach to virtual power plants of wind power generators and electric vehicles," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1314–1322, 2013.

[49] S. Lakshminarayana, T. Q. S. Quek, and H. V. Poor, "Cooperation and storage tradeoffs in power grids with renewable energy resources," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 7, pp. 1386–1397, July 2014.

[50] D. Kalathil, C. Wu, K. Poolla, and P. Varaiya, "The sharing economy for the electricity storage," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 556–567, 2017.

[51] M. Z. Oskouei, B. Mohammadi-Ivatloo, M. Abapour, M. Shafiee, and A. Anvari-Moghaddam, "Privacy-preserving mechanism for collaborative operation of high-renewable power systems and industrial energy hubs," *Applied Energy*, vol. 283, p. 116338, 2021.

[52] H. Khaloie, M. Mollahassani-pour, and A. Anvari-Moghaddam, "Optimal behavior of a hybrid power producer in day-ahead and intraday markets: A bi-objective cvar-based approach," *IEEE Transactions on Sustainable Energy*, 2020.

[53] A. Anvari-Moghaddam, A. Rahimi-Kian, M. S. Mirian, and J. M. Guerrero, "A multi-agent based energy management solution for integrated buildings and microgrid system," *Applied Energy*, vol. 203, p. 41, 2017.

[54] N. Bazmohammadi, A. Tahsiri, A. Anvari-Moghaddam, and J. M. Guerrero, "Stochastic predictive control of multi-microgrid systems," *IEEE Transactions on Industry Applications*, vol. 55, no. 5, pp. 5311–5319, 2019.

[55] M. Daneshvar, B. Mohammadi-Ivatloo, S. Asadi, A. Anvari-Moghaddam, M. Rasouli, M. Abapour, and G. B. Gharehpetian, "Chance-constrained models for transactive energy management of interconnected microgrid clusters," *Journal of Cleaner Production*, 2020.

[56] M. Daneshvar, B. Mohammadi-Ivatloo, K. Zare, S. Asadi, and A. Anvari-Moghaddam, "A novel operational model for interconnected microgrids participation in transactive energy market: A hybrid igdt/stochastic approach," *IEEE Transactions on Indust. Inform.*, 2020.

[57] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-peer trading in electricity networks: an overview," *IEEE Transactions on Smart Grid*, 2020.

[58] S. Bahrami, M. H. Amini, M. Shafie-Khah, and J. P. Catalao, "A decentralized renewable generation management and demand response in power distribution networks," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 4, pp. 1783–1797, 2018.

[59] A. R. Khamesi, S. Silvestri, D. A. Baker, and A. D. Paola, "Perceived-value-driven optimization of energy consumption in smart homes," *ACM Transactions on Internet of Things*, vol. 1, no. 2, pp. 1–26, 2020.

[60] E. Shin, A. R. Khamesi, Z. Bahr, S. Silvestri, and D. A. Baker, "A user-centered active learning approach for appliance recognition," in *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2020, pp. 208–213.

[61] T. F. E. R. Commission, "Reports on demand response & advanced metering," The Federal Energy Regulatory Commission, Tech. Rep., December 2015.

[62] A. R. Khamesi and S. Silvestri, "Reverse auction-based demand response program: A truthful mutually beneficial mechanism," in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2020, pp. 427–436.

[63] S. Bhattacharjee, A. Thakur, S. Silvestri, and S. K. Das, "Statistical security incident forensics against data falsification in smart grid advanced metering infrastructure," in *ACM CODASPY*, 2017, pp. 35–45.

[64] S. Bhattacharjee, V. P. K. Madhavarapu, S. Silvestri, and S. K. Das, "Attack context embedded data driven trust diagnostics in smart metering infrastructure," *ACM Transactions on Privacy and Security (TOPS)*, vol. 24, no. 2, pp. 1–36, 2021.

[65] F. Sioshansi, "The sorry state of demand response in the us," *Energypost*, 2018. [Online]. Available: https://energypost.eu/the-sorry-state-of-demand-response-in-the-u-s/

[66] R. Earle and A. Faruqui, "Toward a new paradigm for valuing demand response," *The Electricity Journal*, vol. 19, no. 4, pp. 21–31, 2006.

[67] G. El Rahi, S. R. Etesami, W. Saad, N. B. Mandayam, and H. V. Poor, "Managing price uncertainty in prosumer-centric energy trading: A prospect-theoretic stackelberg game approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 702–713, 2017.

[68] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, 2020.

Ashutosh Timilsina received his Bachelor's Degree in Electrical Engineering from Tribhuvan University, Institute of Engineering, Pulchowk Campus, Nepal. He is currently pursuing PhD in Computer Science in Cyber-Physical-Human Systems Lab, University of Kentucky since 2019. His research interests include Energy Optimization in Smart Grid, P2P Energy Trading and application of machine learning in these fields.

Atieh R. Khamesi [S'15-M'18] is currently a post-doctoral scholar at the Department of Computer Science, University of Kentucky, USA. She received her Ph.D. Information Engineering in 2018 from the University of Padova, Italy. She received her bachelor's and master's degrees in electrical engineering and communication systems from the Ferdowsi University of Mashhad, Iran, and the University of Tehran, Iran, in 2011 and 2014, respectively. In 2017, she was a visiting scholar with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, USA. Recently, she has conducted research on Cyber-Physical-Human Systems (CPHSs) where she investigates the interaction of humans with smart environments and IoT systems. Other areas of her interests include machine learning, optimization and algorithm design.

Vincenzo Agate received his Bachelor Degree and Master Degree in Computer Engineering from University of Palermo, Italy, in 2012 and 2016, respectively, and his Ph.D. in Computer Engineering from University of Palermo, Italy, in 2020. His current research is focused on distributed systems, reputation management systems and privacy-preserving computation.

Simone Silvestri is currently an Associate Professor in the Department of Computer Science of the University of Kentucky. He received his Ph.D. in Computer Science in 2010 from the Department of Computer Science of the Sapienza University of Rome, Italy. Dr. Silvestri's research interest are in the area of Cyber-Physical-Human Systems, Internet of Things, and Wireless Networks. The research is funded by several national and international agencies such as NIFA, NATO and NSF. He received the NSF CAREER award in 2020. He has published more than 70 papers in top-tier international journals and conferences.