




2023

Peer-to-Peer Energy Trading in Smart Residential Environment with User Behavioral Modeling

Ashutosh Timilsina

University of Kentucky, ashutoshtimilsina@gmail.com

Author ORCID Identifier:

 <https://orcid.org/0000-0002-3674-7857>

Digital Object Identifier: <https://doi.org/10.13023/etd.2023.161>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Timilsina, Ashutosh, "Peer-to-Peer Energy Trading in Smart Residential Environment with User Behavioral Modeling" (2023). *Theses and Dissertations--Computer Science*. 133.
https://uknowledge.uky.edu/cs_etds/133

This Doctoral Dissertation is brought to you for free and open access by the Computer Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Computer Science by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@sv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Ashutosh Timilsina, Student

Dr. Simone Silvestri, Major Professor

Dr. Simone Silvestri, Director of Graduate Studies

PEER-TO-PEER ENERGY TRADING IN SMART RESIDENTIAL
ENVIRONMENT WITH USER BEHAVIORAL MODELING

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the College of Engineering
at the University of Kentucky

By

Ashutosh Timilsina

Lexington, Kentucky

Advisor: Dr. Simone Silvestri,
Associate Professor of Computer Science
Lexington, Kentucky

2023

Copyright ©Ashutosh Timilsina 2023

<https://orcid.org/0000-0002-3674-7857>

ABSTRACT OF DISSERTATION

Electric power systems are transforming from a centralized unidirectional market to a decentralized open market. With this shift, the end-users have the possibility to actively participate in local energy exchanges, with or without the involvement of the main grid. Rapidly reducing prices for Renewable Energy Technologies (RETs), supported by their ease of installation and operation, with the facilitation of Electric Vehicles (EV) and Smart Grid (SG) technologies to make bidirectional flow of energy possible, has contributed to this changing landscape in the distribution side of the traditional power grid.

Trading energy among users in a decentralized fashion has been referred to as Peer-to-Peer (P2P) Energy Trading, which has attracted significant attention from the research and industry communities in recent times. However, previous research has mostly focused on engineering aspects of P2P energy trading systems, often neglecting the central role of users in such systems. P2P trading mechanisms require active participation from users to decide factors such as selling prices, storing versus trading energy, and selection of energy sources among others. The complexity of these tasks, paired with the limited cognitive and time capabilities of human users, can result in sub-optimal decisions or even abandonment of such systems if performance is not satisfactory. Therefore, it is of paramount importance for P2P energy trading systems to incorporate user behavioral modeling that captures users' individual trading behaviors, preferences, and perceived utility in a realistic and accurate manner. Often, such user behavioral models are not known *a priori* in real-world settings, and therefore need to be learned online as the P2P system is operating.

In this thesis, we design novel algorithms for P2P energy trading. By exploiting a

variety of statistical, algorithmic, machine learning, and behavioral economics tools, we propose solutions that are able to jointly optimize the system performance while taking into account and learning realistic model of user behavior. The results in this dissertation has been published in IEEE Transactions on Green Communications and Networking 2021, Proceedings of IEEE Global Communication Conference 2022, Proceedings of IEEE Conference on Pervasive Computing and Communications 2023 and ACM Transactions on Evolutionary Learning and Optimization 2023.

KEYWORDS: P2P Energy Trading, User Behavioral Modeling, Prospect Theory, User Preference, Reinforcement Learning, Optimization

PEER-TO-PEER ENERGY TRADING IN SMART RESIDENTIAL
ENVIRONMENT WITH USER BEHAVIORAL MODELING

By

Ashutosh Timilsina

Dr. Simone Silvestri

Director of Dissertation

Dr. Simone Silvestri

Director of Graduate Studies

04/28/2023

Date

If you keep proving stuff that others have done, getting confidence,
increasing the complexities of your solutions - for the fun of it - then one
day you'll turn around and discover that nobody actually did that one!

- Richard P. Feynman

DEDICATION

To my family

for being there on every step of my life ...

ACKNOWLEDGEMENTS

I am filled with a deep sense of gratitude as I reflect on my PhD journey that led to the completion of this dissertation. It is with great pleasure and humility that I acknowledge the contributions of the individuals who supported me along the way and made this accomplishment possible.

First and foremost, I would like to thank my advisor and DGS, Dr. Simone Silvestri, for his unwavering support, guidance, and expertise. His mentorship and valuable feedback have been instrumental in shaping my research ideas and guiding me towards the successful completion of this dissertation. I would also like to extend my appreciation to my committee members, Dr. Brent Harrison, Dr. Nathan Jacobs, and Dr. Dan Ionel, for their valuable feedback during and after my qualifying examination, which significantly improved the quality of my dissertation.

I am grateful to my fellow PhD friends, Subash Khanal, Sweta Ojha, Stephen Parsons, and Enrico Casella, who have been with me throughout this journey, providing constant support, motivation, and laughter. Your unwavering friendship has made this experience a lot more manageable and enjoyable.

I also owe a debt of gratitude to my undergraduate research partners and lifelong friends, Binay, Beni, Jenesha, Saugat, Abinash, and Elina, for the foundation we built together in our earlier years. I would like to thank my undergraduate advisor, Dr. Arbind Kumar Mishra, and mentor, Dr. Nava Raj Karki, for teaching me the basics of research and providing me with opportunities to explore my interests early on in the undergraduate years.

Last but certainly not the least, I would like to express my heartfelt appreciation to my family. To my father, Hari Timilsina, and mother, Meena Baral, for their constant support, guidance, and unconditional love. You have taught me the ABC's of life, and I am grateful for everything you have done for me. To my brother, Amit Timilsina,

and my sister-in-law, Pavitra Neupane, for being my constant support system and cheering me on every step of the way.

Thank you, everyone, for being a part of my life and making this possible. Your belief in me has been the wind beneath my wings, and I am forever grateful. I owe this dissertation to all of you!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	ix
LIST OF FIGURES	x
1 INTRODUCTION	1
1.1 Background	1
1.1.1 Prosumption and Need for Alternative Energy Market	2
1.1.2 Peer-to-peer (P2P) Energy Trading	4
1.2 Objectives of This Dissertation	8
1.3 Dissertation Structure	9
2 GENERAL OVERVIEW OF PEER-TO-PEER ENERGY TRADING, USER BEHAVIORAL MODELING, REINFORCEMENT LEARNING, AND CROWDSOURCING	10
2.1 Distributed Energy Resources, Smart Grid and Their Integration . . .	10
2.2 Localized Energy Exchange and P2P Energy Trading	12
2.2.1 Components of P2P Energy Trading	14
2.2.2 Possible Market Structures for P2P Energy Trading	19
2.3 User Behavioral Modeling and Behavioral Economics	21
2.4 Reinforcement Learning	25
2.5 Spatial Crowdsourcing and Use of Electric Vehicles for Crowdsourcing- based Energy Sharing	29
3 A DETAILED LITERATURE REVIEW ON P2P ENERGY TRADING AND USER BEHAVIORAL MODELING	31
3.1 P2P Energy Trading	31
3.2 Technical Approaches Employed for P2P Energy Trading	34
3.3 Real-world Implementations of P2P Energy Trading	40

3.4	User Behavioral Modeling and Behavioral Economics in Context of Energy Trading	41
3.4.1	User Perception in Energy Sharing Systems	41
3.4.2	Prospect Theory in P2P Energy Trading	43
3.5	Reinforcement Learning in Energy Trading	45
3.6	EV-based Crowdsourcing for Energy Sharing	46
3.7	Limitations in Existing Literature and Motivation	48
4	A REINFORCEMENT LEARNING APPROACH FOR USER PREFERENCE-AWARE P2P ENERGY SHARING	51
4.1	System Model and Assumptions	52
4.2	Problem Formulation	54
4.3	A Reinforcement Learning Approach for User Preference Learning . .	57
4.4	A Constrained Maximum Weighted Matching- based Reinforcement Learning Approach	61
4.4.1	Faster Initialization Algorithm (FIA)	62
4.4.2	The BiParTite-K Algorithm	63
4.5	Experimental Results	69
4.5.1	Experimental Setup	70
4.5.2	Comparison approach	70
4.5.3	Performance Evaluation	71
4.6	Concluding Remarks	80
5	PROSPECT THEORY-INSPIRED P2P ENERGY TRADING	81
5.1	System Model and Problem Formulation	84
5.1.1	Modeling Energy Allocation	84
5.1.2	Modeling Pricing Mechanism	90
5.2	Solution Approaches and Heuristics	90
5.2.1	DEbATE	91

5.2.2	PQR	93
5.2.3	ProDQN	95
5.3	Experimental Results	99
5.3.1	Experimental Setup	99
5.3.2	Comparison Approaches	101
5.3.3	Results	102
5.4	Concluding Remarks	107
6	<i>E-UBER: A CROWDSOURCING PLATFORM FOR ELECTRIC</i> <i>VEHICLE-BASED RIDE- AND ENERGY-SHARING</i>	109
6.1	System Model	110
6.2	<i>e-Uber</i> : Problem Formulation	112
6.2.1	Preference-aware Optimal Task Recommendation Problem . .	113
6.2.2	Winning Bid Selection and Final Payment Problem	114
6.3	e-Uber Solution Approaches	116
6.3.1	CMAB-based Task Recommendation System	116
6.3.2	Winning Bid Selection using Weighted Bipartite Matching . .	121
6.4	Experiment	123
6.4.1	Experimental Setup	123
6.4.2	Results	125
6.5	Conclusion	131
7	CONCLUSION AND FURTHER RESEARCH	132
7.1	Main Contribution	133
7.2	Further Research	134
7.2.1	Modeling complex dependencies of user behavioral parameters	135
7.2.2	Fully decentralized energy allocation	135
7.2.3	Reward functions for RL frameworks	136
7.2.4	Including energy storage and electric vehicles in P2P	136

7.2.5 Blockchain technology for practical implementation	137
BIBLIOGRAPHY	138
VITA	151

LIST OF TABLES

3.1	P2P Energy Trading Approaches	35
4.1	Notation Summary	53
5.1	Statistical Analysis of experimental result in Figs. 5.9 and 5.10	104
6.1	List of Notations	110

LIST OF FIGURES

1.1	DERs and their integration with Smart Grid.	1
1.2	Visualization of Virtual Power Plant architecture. [CC BY-SA 4.0] . .	2
1.3	Incorporating User Behavioral Modeling into P2P Energy Trading . .	6
2.1	General Overview of P2P Energy Trading Network	14
2.2	Division of P2P Energy Trading Network Elements	15
2.3	P2P Energy Market Models [1]: a) Fully Decentralized Model; b) & c) Prosumer-Microgrid Model - b) Prosumer-Interconnected Microgrids, c) Prosumer-Islanded Microgrids; and d) Organized Community Model	19
2.4	Notions of User Behavioral Modeling	22
2.5	A prospect theory value function $v(x)$ with $\zeta_+ = \zeta_- = 0.88$ and $k_+ =$ $k_- = 2.25$	24
2.6	Agent-Environment Interaction in RL	25
2.7	General Schema of RL Methods. Direct approach utilizes value/policy function and indirect approach utilizes a model of the environment. .	28
4.1	P2P Energy Sharing System Overview	51
4.2	Efficiency and Satisfied Demand vs. number of consumers (keeping number of producers constant)	72
4.3	Efficiency and Satisfied Demand vs. number of consumers (keeping ratio of consumers-to-producers constant)	72
4.4	Plot for cumulative reward (energy exchange)	75
4.5	Percentage of energy losses over time.	75
4.6	Average absolute % error of preferences learned over time	75
4.7	Number of Days required for initialization vs. K	75
4.8	Efficiency of the initialization algorithms vs. number of consumers . .	78
4.9	Number of days required for initialization vs. number of consumers .	78

4.10	Energy Transferred vs. K	79
4.11	Energy Transferred vs. T	79
5.1	Incorporating User Behavioral Modeling into P2P Energy Trading . .	81
5.2	P2P Energy Trading System Overview.	82
5.3	Proposed Framework of P2P Energy Trading.	85
5.4	Overview of the ProDQN approach	96
5.5	Normalized objective value vs. number of iterations for varied population sizes.	102
5.6	Computation time vs. population sizes.	102
5.7	Normalized objective value vs. number of iterations for varied network sizes.	102
5.8	Computation time vs. network sizes.	102
5.9	Buyers' perceived values.	104
5.10	Sellers' cumulative reward.	104
5.11	Individual prices for <i>ProDQN</i>	106
5.12	Individual prices for <i>PQR</i>	106
5.13	Avg. price comparison for different network sizes between <i>PQR</i> and <i>ProDQN</i>	106
5.14	Objective values for buyer vs. network size.	107
5.15	Total rewards for sellers vs. network size.	107
6.1	e-Uber crowdsourcing platform overview	109
6.2	Working mechanism of e-Uber	111
6.3	Training Loss %	125
6.4	Bid Prediction test accuracy	125
6.5	Snapshot of obj. values & matches vs. time	126
6.6	Cumulative obj. values	127

6.7	Cumulative tasks	127
6.8	Avg. Price/task vs. Task(%)	128
6.9	Avg. price/task vs. V2G (%)	128
6.10	Mean Absolute Error vs. time	129
6.11	Cumulative reward plot	130
6.12	Obj. values/matching vs. K	130

CHAPTER 1. INTRODUCTION

1.1 Background

Over the last decade, there has been significant interest in Renewable Energy Technologies (RET) and Smart Grids (SG) fueled by the increased concern regarding climate change and carbon footprint. The current status of power system has a significant contribution to the carbon footprint given that 63% of world's total electricity production is based on fossil fuels like coal, oil, and natural gas as of 2019 [2]. In comparison, the portion of renewable sources is only 26% [2]. The figure in United States (US) alone is 60% for fossil fuels and 20% for renewable as of 2020 [3]. However, over the course of two years through 2018 and 2019, the renewable source based generation has seen significant increase globally with wind (+11.8%) and solar (+22.5%) on the rise [2]. Somewhere around 2,807 GW worth of renewable energy technologies has been installed at the end of 2020 with 10.5% growth in 2020 alone while this figure reached to 3064 GW with 9.2% growth in 2021 [4]. Main driving factors for this growth in renewable energy generation can be accounted to the ease of installation for such RETs and financial incentives that they entail.



Figure 1.1: DERs and their integration with Smart Grid.

Owing to these factors, dramatic increase has been observed in the deployment of RETs on consumers' side as well [2] with 21 GW (+28%) residential solar installation compared to 59.5 GW (+28.5%) in US [3]. In general, the energy sources with small-scale power generation capabilities installed on distribution side are referred

to as Distributed Energy Resources (DERs), which may include solar panels, wind turbines, and electric vehicles among others. The consumers equipped with DERs save the expenses in long term as they cut off the cost (in part or in whole depending on the installed generation capacity) of buying energy from the grid at costly price, followed by the fact that this allows them to have control of how and when to produce, consume and manage the energy. Recently, several researchers and government bodies have put significant efforts into the evolution of SG technologies towards the paradigm of Virtual Power Plants (VPPs) [5,6]. Unlike the traditional systems where the energy generation and distribution are centralized [1, 7], VPPs support a two-way flow of electricity and information [8]. The objective of VPPs is to aggregate DERs, (such as photovoltaics (PV), wind power, electric vehicles etc.), into the grid to provide reliable ancillary services, traditionally provided by large power plants [9]. 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 [7, 9].

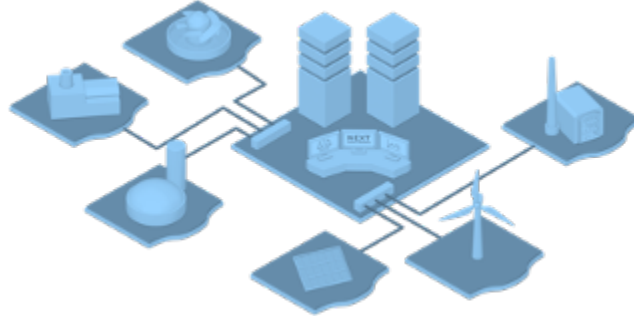


Figure 1.2: Visualization of Virtual Power Plant architecture. [CC BY-SA 4.0]

1.1.1 Prosumption and Need for Alternative Energy Market

DERs are generally intermittent and depending on the time of the day, weather condition and the energy demand on any given day, the consumers equipped with DERs can go beyond self-consumption as there will be excess energy generation which might go wasted otherwise [1, 7, 10]. There exists two extensively employed methods as of now for utilizing such excess energy generation on consumer's end [11]. The first

approach is to install an energy storage system like battery to store excess energy generation for later use. However, that will only incur additional costs in user's case not just installation cost but operational costs too. Furthermore, it has been shown in several studies that installing a reasonable size of state-of-the-art battery back-up system is either much expensive or their life-cycle and efficiency render them infeasible in long-term scenario [12, 13].

The second approach to utilize excess energy generation is to sell the energy to grid through widely used schemes like Feed-in-Tariff (FiT) mechanism [13–15]. In such mechanism, users can sell excess energy to the grid and buy again from the grid in case of deficiency [15]. Unfortunately, these mechanisms offer very marginal benefits to participating consumers [14, 15]. Mechanism like FiT generally involve net-metering for energy exchange for users meaning that users will not get paid for any excess energy they supply to the grid. Even if they do get paid, it is a nominal price that is not as motivating for users to be involved in such schemes for long term [14, 15]. As a consequence, several grids through out the world have placed capping either in terms of the amount of energy that these consumers can sell to the grid or in terms of energy price. In addition to that, there have been reports of schemes like FiT being discontinued altogether in several locations through out the world [13–15].

Instead recent efforts to create a consumer-centric energy market where consumers can participate in exchanging energy freely without any sidelining from the grid is taking pace [1, 16, 17]. A key enabler of this new market has been SG technologies [18], which has facilitated the bidirectional flow of electricity with the advanced metering infrastructures and smart circuit breakers. Furthermore, with the support of state-of-the-art Internet-of-Things (IoT) devices, SG has made it more easier for consumers to exchange, monitor and manage energy generation and consumption [19, 20]. Since consumers can monetize from excess energy that instead would have gone to waste, this acts as a primary incentive for such users to engage in the energy exchange

process itself. Moreover this has also promoted the localized energy generation and consumption which has an added benefit of reducing the energy loss incurred in transmission/distribution system in a traditional power system [10].

Additionally, the local energy exchange among consumers also helps in managing the energy at the distribution side without causing stress to the grid. Two-way benefits of such an energy market has been regarded as a promising and viable alternative to the schemes like FiT mechanism and does not require costly energy storage options [1, 15]. With the localized energy exchange, the consumers who were at the lower hierarchy in the traditional power system are now being actively involved in the energy market with the increased proliferation of DERs and SG technologies into the scene. This has, in turn, created a shift in the energy market from traditional centralized architecture to a more robust and versatile decentralized architecture where such consumers equipped with DERs can be involved in the energy exchange for any excess generation on their end [10, 13–15]. These consumers who can engage in energy exchange are called prosumers [1], as a portmanteau of *pro*-ducers and *con*-sumers. This kind of prosumer-centric market emphasizes on the decentralized structure that promotes local energy exchange through added monetary incentives for both local buyers and sellers of energy [1, 17].

1.1.2 Peer-to-peer (P2P) Energy Trading

Such a decentralized energy market is referred to as the peer-to-peer (P2P) energy trading market [10, 14, 15], as it involves the trading of energy between peers. Peers in this case would be prosumers who want to exchange energy among each other. Generally feasible in localized setting because of the amount of energy involved and the constraints of physical infrastructure [7, 15], these types of energy trading have been successfully applied in a real setting as can be seen in case of US based Brooklyn Microgrid [21], Dutch startup called Vanderbron [22] and UK-based Piclo [23]. This energy trading market modality has served as an alternative to the traditional energy

trading modality and therefore, has opened new avenues to allow prosumers to trade energy among themselves at mutually beneficial prices with added flexibility for trading energy services [1,17]. It not only benefits the prosumers in terms of finances and flexibility but also supports the overall cause of making the energy landscape green, sustainable, and efficient [7,10,14,24].

The main concern however in case of P2P energy trading is to assert sustained involvement of prosumers in the energy market to trade energy with one another [15]. This requires a continual active participation from prosumers with the system, which might not work in practical setting. It has been established that humans have limited capability to process information which reflects to their limited cognitive and time capabilities [25]. In other words, the users deviate from rational decision making when they are overwhelmed with a great deal of options to choose from or tend to settle for non-rational judgements when the problem start getting relatively complex to comprehend. This bound to human rationality is explained by the concept called *Bounded Rationality*, which has been observed in real life situations and verified by numerous studies spanning from fields like behavioral economics [26–28] to cognitive sciences and decision-making [29].

The original work on bounded rationality in [26] mentions that the users might suffer from fatigue in the face of complex problems and overwhelmed choices and hence, resort to a phenomenon called *satisficing*, i.e. resorting to bare minimum that meets their aspiration. Furthermore, the author in [30] notes that this phenomenon has become more prominent in current digital age with wealth of information and choice overload at hand. Similarly the works in [31] found that in addition to satisficing, the users might also opt to terminate the participation in the process if they are snowballed, bored or experience physical discomfort beyond certain limit. With similar rhetoric, [32] points out that in the age of information overload, the users must be subjected to only relevant information and the least amount at that. Hence, in

light of these works, it would be safe to say that a P2P energy trading market (or any such intelligent system, for that matter) should require minimal active participation from users while also ensuring their sustained involvement [15, 33]. This can be achieved through the process of automated decision making for energy exchanges among prosumers that is augmented by occasional human participation as a feedback for the system to guide it towards better decision-making.

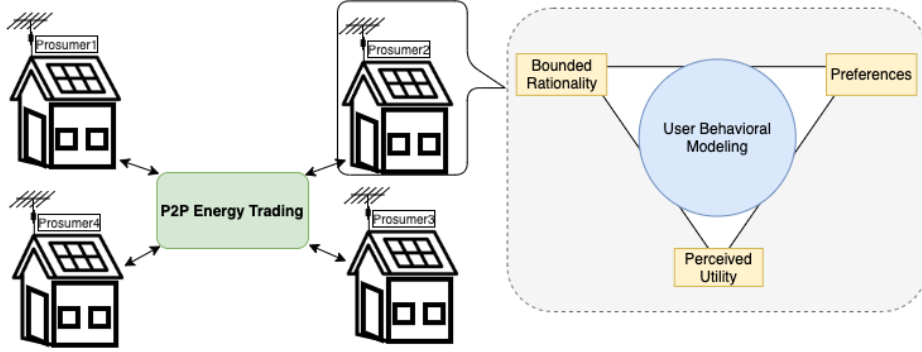


Figure 1.3: Incorporating User Behavioral Modeling into P2P Energy Trading

To do so, user behavioral modeling should be incorporated into the system for making automated decisions that closely reflects the prosumers' energy trading behaviors as shown in Fig. 1.3. In addition to that, the system also needs to maximize the overall incentives on their behalf. Therefore, the P2P energy trading needs to be prosumer-centric not just in terms of their involvement but also in terms of modeling their behavioral patterns to predict their decision-making behavior and sustain their long term involvement. The user behavior in turn needs to be learned from historical user behavioral patterns/data. But oftentimes these data are unavailable or difficult to obtain. So in absence of such data, this user behavior needs to be learned through the adaptive learning process with online feedback from the users. This is a complex problem of learning the users' behavioral patterns over time while also using those behavioral patterns for maximizing the incentives in every energy transactions.

Pertaining to the P2P energy trading, several work has been done in the domain of energy exchange between peers in [34], [35], [36], [10], [37] and [38].

These works focus on physical aspects like minimizing the loss or costs to maximizing the exchanged energy. However, they have failed to incorporate the user utility into the problem. There has been a work in [39] that includes the grid involvement in the trading whereby grid sets the rules and regulation for amount of energy and price. In context of user behavioral modeling in energy market, authors in [40] consider dynamical price change to alter the user behavior for demand response. In [41], prosumers with similar energy behaviors are grouped together to form a virtual prosumer-community for energy sharing among such communities.

Similarly there have been few attempts at utilizing behavioral economics concept to model the user behaviors. The authors in [16,42,43] have made some effort to capture the irrationality of users under uncertain decision-making using the *prospect theory* as a model to reflect user's perceived utility. The term "utility" is referred to as the measure of total satisfaction or benefit derived from consuming a good or service¹. The work in [15] has regarded the user participation as a central aspect of the P2P energy trading but is only limited to the coalition formation in game theoretic setting and does not explicitly consider user behavioral modeling. The work in [44] also uses a prospect theory based distributed energy trading model to optimize trading decisions for prosumers in a competitive market but with active human participation in the form of bids and aims at learning the aggressiveness of the bids over time. However, an explicit model that incorporates all of these system components for P2P energy trading market together is still lacking in existing literature. Therefore there is the requirement for developing such an energy trading market that not just facilitates the automated energy exchange between peers but also maximizes their incentives while prioritizing their energy trading behavior and perceived utility.

This dissertation aims at addressing the above limitations. Specifically, this work aims at devising a prosumer-centric model for peer-to-peer energy trading

¹Note that the term *utility*, for the rest of the dissertation, does not refer to the utility companies that provide electricity or gas and strictly refers to the economical and psychological worth

that optimizes the energy exchange between peers with their behavioral patterns like preferences, price, and perceived utility into account. In order to learn these user behavior patterns over time, mathematical optimization, machine learning and reinforcement learning algorithms along with other classical statistical tools can be used to design an all encompassing user modality. Through this study, we aim to incorporate concepts of behavioral economics including bounded rationality and prospect theory in the energy exchange problem that learns and captures the user behavior as accurately as possible and utilizes the same to automate the decision-making process on behalf of the users.

1.2 Objectives of This Dissertation

Main objectives of this dissertation are:

- To provide a general outlook on P2P energy trading along with possible market modalities, user behavioral modeling and reinforcement learning.
- To conduct a detailed literature survey on existing state of P2P energy exchange and trading approaches; along with an in-depth review of user behavioral modeling, behavioral economics and learning of the user behavior.
- To devise a P2P energy exchange mechanism among users that aims at learning the user behavior over time through reinforcement learning and maximize the energy exchange between users with user preference and bounded rationality in concern.
- To develop a prosumer-centric automated P2P energy trading platform through the lens of prospect theory that maximizes prosumer's perceived utility with monetary incentives as major driving factor.
- To automate the pricing mechanism for sellers using reinforcement learning frameworks to dynamically update selling prices based on market feedback on price and total energy sold.

- To propose an alternative modality of P2P energy trading involving Electric Vehicles for joint enabling of ride- and energy-sharing services.

1.3 Dissertation Structure

The dissertation consists of seven chapters, bibliography and appendices. Chapter 1 provides the overview of the dissertation, main contributions and motivations. Chapter 2 presents a general overview of relevant topics that sets up a foundation for the work presented in this dissertation. An in-depth review of related works is provided in Chapter 3 along with highlighting the research gaps that exist in these works. In chapter 4, a P2P energy exchange model is devised that aims at learning the user preferences through a reinforcement learning based recommendation system while also maximizing the daily energy production and consumption. The work in this chapter was published in IEEE Transactions on Green Communications and Networking [45] and Proceeding of 2023 IEEE International Conference on Pervasive Computing and Communications (IEEE PerCom) [46]. Chapter 5 proposes the incorporation of the monetary exchanges and perceived utility of prosumers, automating the trading behavior of users through deep reinforcement learning. The work in this chapter is published in the proceedings of 2022 IEEE Global Communications Conference [47] and in ACM Transactions on Evolutionary Learning and Optimization [48]. In chapter 6, we also incorporate the electric vehicle into the P2P energy market model. In this chapter, we develop a more comprehensive, effective, and realistic solution to jointly enable of ride-and energy-sharing services in a crowdsourcing setting using reverse auction, reinforcement learning and efficient matching algorithms. The works in this chapter are published in Proceeding of 2023 IEEE Transportation Electrification Conference [49] and under peer-review at the time of submitting this thesis (arXiv preprint [50]). Finally, the conclusion and future directions for the work in this dissertation are discussed in detail in chapter 7.

CHAPTER 2. GENERAL OVERVIEW OF PEER-TO-PEER ENERGY TRADING, USER BEHAVIORAL MODELING, REINFORCEMENT LEARNING, AND CROWDSOURCING

In this chapter, a quick overlook on Distributed Energy Resources (DERs), peer-to-peer (P2P) energy trading market model, user behavioral modeling and reinforcement learning is presented. We start with growth of DERs, their interconnection with the grid, rise of *prosumers* and Smart Grid (SG) and then move into detailed overview of localized energy exchange and P2P model. This is complemented with a brief introduction to user behavioral modeling and reinforcement learning approaches that will find application in this dissertation.

2.1 Distributed Energy Resources, Smart Grid and Their Integration

In this section, we discuss about DERs, how the growth in smart grid technologies have bolstered the growth of such DERs in distribution side of the grid and how it is transforming the energy landscape altogether.

The urgency to mitigate the detrimental effects of energy industry on environment has boosted the research in the field of environment-friendly, efficient and sustainable source of energy generation in recent years [51]. As a result, the Renewable Energy Technologies (RETs) and more specifically green energy technologies have been at the centre of attention [9, 51]. This has been further fueled by the IoT-enabled SG that embraces the use of cutting edge technologies to make the grid smarter and an active ecosystem for energy exchanges among all the stakeholders [18]. Through the incorporation of the novel sensing, communication, computing and monitoring technologies, this smart grid focuses on transforming the electrical grid into smarter, more resilient, efficient and flexible grid [19, 20]. SG technologies include Phasor Measurement Units (PMUs) to assess the grid stability in real time, Advanced Metering Infrastructures (AMIs) that can sense bidirectional flow of electricity and

other IoT-based multi-functional supports that can help consumers track and schedule their energy consumption as per their requirement [18]. SG also includes smart relays and circuit breakers that sense and recover from faults automatically and support the remote operation [18]. This way the SG has realized a bidirectional flow of electricity and information that can easily accommodate the flexible operation of grid in real time. It has also made the energy ecosystem more convenient and accessible for installation and operation at consumer's end. Similarly the rapid rise in adoption of Electric Vehicles (EV) and the bidirectional chargers have created new opportunities for mobile and flexible energy management through Vehicle-to-Grid (V2G), Vehicle-to-Home (V2H), Vehicle-to-Vehicle (V2V) [52], as well as Battery Swapping Technology (BST) [53].

The adoption of SG as well as the convenient installation/operation and reducing price of RETs and EVs have resulted into proliferation of such energy sources at distribution side of the grid that has been referred to as Distributed Energy Resources (DERs) [54]. As [55] notes, DERs 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. 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 [6, 56]. To this aim, the integration of such DERs into power systems has been introduced with the help of SG [10, 19, 57].

Integrating renewable energy generation into Smart Grids (SGs) already exists in the form of Feed-in-Tariff [10, 13–15]. Through this mechanism, consumers can sell the excess energy generated to the grid and also buy from the grid in case of deficiency [7, 15]. However, [14, 15] mentions that this mechanism do not offer greater incentives to consumers. It generally employs net-metering concept meaning that the grid do not offer monetary incentives for any net excess energy sold to the grid and even if they do, the price offered is very low [14, 16, 39]. Therefore, the authors in [7, 13]

highlight that this approach is neither profitable nor flexible for such DER-based producers. In addition to that, [15] also points out that the increased penetration of the intermittent distributed energy resources into the grid presents critical challenges, such as energy fluctuation in the grid and successive destabilization.

The use of energy storage like batteries could be an alternative to deal with the challenges posed by increased penetration of DERS into the grid but at the existing state of these technologies, they are expensive to install and their efficiency and life-cycle is not luring enough to encourage consumers for their wide-scale adoption [12,13]. As an alternative solution to mitigate these challenges, the trade-off between use of storage for excess energy and managing energy among users with DERs was studied in [58]. The results showed that in absence of expensive and ineffective storage, the grid can notably gain from energy exchange among the users themselves meaning that the localized energy exchange between the consumers could help in managing the energy without destabilizing the grid. 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 [59]. Therefore, trading surplus energy between users in a localized setting is more viable and attractive option [24,45,60].

2.2 Localized Energy Exchange and P2P Energy Trading

In this section, we discuss in brief about P2P energy trading modality. In doing that, several components and market structure that realizes a P2P energy trading model is presented.

As established in section 2.1, supported by the SG technologies, a localized energy trading mechanism that provides a platform to exchange energy between peers can serve as an alternative to the existing market modality. Such a mechanism is referred as Peer-to-Peer (P2P) Energy Trading. It consists of consumers with energy generation capabilities or "prosumers" (portmanteau of *pro*-ducers and *con*-sumers)

who can sell their energy to other prosumers (or consumers) at a profitable price and also buy from them at the time of their need. P2P energy trading has been gaining popularity in recent years as a decentralized alternative to the traditional energy trading modality that provides flexibility for end-users to be involved in energy trading [1, 33]. While P2P market model for energy trading mitigates the limitations of existing market modality including FiT mechanism, there are several other added benefits of this model. The localized energy trading between prosumers offer economical incentive for prosumers to be engaged in such a system and make profit out of the energy which otherwise would have gone wasted.

It also offers the consumers a wide range of choices to buy energy from according to their preferences and convenience. Specifically, P2P energy trading provides a prosumer-centric platform that allows prosumers to trade energy with each other at a negotiated price. The trading may or may not involve the grid [14, 61]. Typically, the operation range of trading price would be higher than FiT price offered by grid and lower than the electricity tariff charged by utilities. Further, electric grid companies offering rates that change with the time-of-day make the P2P paradigm even more convenient, particularly when the price of electricity charged by the grid is at its peak [61]. Monetary incentives resulting from P2P energy trading, for both selling prosumers (producers) and buying prosumers (consumers), are therefore far better compared to existing mechanisms. As a result, prosumers are more incentivized to keep engaging with the trading for the long-term [14, 61].

P2P energy trading also aims at minimizing the dependency of prosumers from grid for energy [15], resulting in an increased reliability of the overall system. Additionally, a higher amount of local energy generation and consumption resulting from P2P trading leads to the minimization of the overall system energy loss, as well as an effective way to achieve demand side management [14, 24, 35, 45, 60]. Benefits extend also to the grid operator, by providing savings in investments that would

have been otherwise required to develop/maintain transmission infrastructure in a centralized power distribution architecture [1, 14]. Therefore, P2P energy trading offers a prosumer-centric approach that has potential to benefit all stakeholders involved, as highlighted extensively in recent studies [10, 15, 34, 36, 39, 44, 45].

A general P2P energy trading mechanism is shown in Fig. 2.1. In order to provide a detailed and structured overview of P2P energy trading, the components of P2P energy trading is described in the next subsection 2.2.1 and the possible market modality is explained in the following subsection 2.2.2.

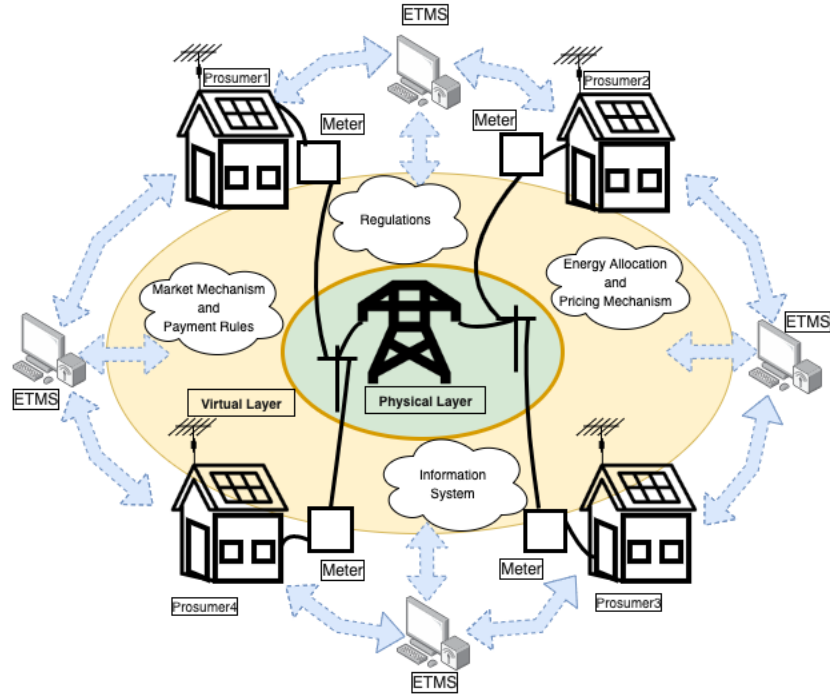


Figure 2.1: General Overview of P2P Energy Trading Network

2.2.1 Components of P2P Energy Trading

To realize a P2P network, there must be two important elements present in the system. One is the physical network that makes the actual exchange possible among peers and next is the virtual platform that provides a technical infrastructure for controlling and managing the P2P exchange. Different literature have visualized the network components of P2P energy trading in multitude ways [1, 7, 14, 33, 62]. Most

succinct way of dividing the features for such a P2P energy trading can be found in [62], which highlights seven requirements viz. (i) a localized market setup with participants and objectives, (ii) grid connection setup to transfer the energy (iii) information system for communication within the market and its monitoring, (iv) market mechanism and payment rules, (v) energy allocation and pricing mechanism, (vi) energy management trading system, and (vii) regulation. Depending on whether these features are physical requirement or virtual requirement, they can be grouped into physical layer and virtual layer. The most important requirement is the market participants without whom the market would not work. This can be referred as *Human Participation* which includes the prosumers and user behavioral modeling. So in summary, the P2P energy trading can be said to composed of these three major components as shown in the Fig. 2.2.

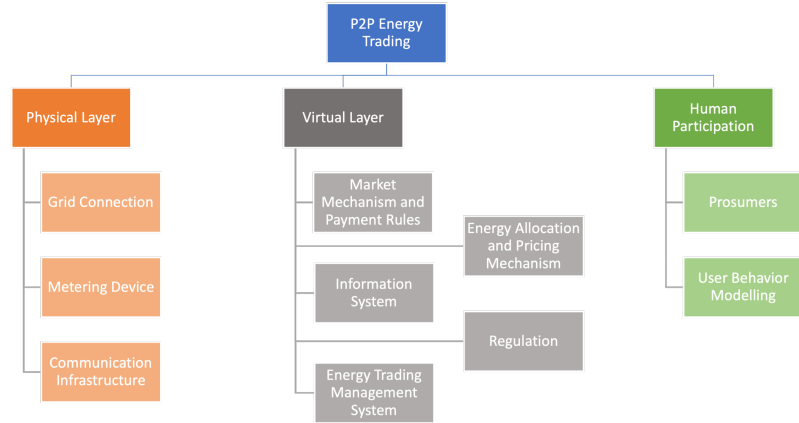


Figure 2.2: Division of P2P Energy Trading Network Elements

Based on this division, a general overview of the P2P energy trading network is shown in Fig. 2.1. The network can be visualized as the physical infrastructures that forms the core of the network to realize the actual exchange of energy through grid connection, smart meters and communication infrastructure. This physical layer is then encapsulated by and monitored through the virtual layer that overviews the energy exchanges using the market mechanism and pricing and regulations. As a P2P

network, it requires involvement from peers/prosumers in the process which either demands their active participation or could instead automate the decision making process on behalf of peers through the use of user behavioral modeling.

In the following, the three elements of the P2P energy trading and their components are explained in brief.

1. *Virtual Layer Elements*

a) **Market Mechanism and Payment Rules**

The market mechanism defines overall modality of market including payment rules and clearly defined market operation. It also enables how the trading and energy transactions will be carried out during different market-time horizon [14, 62].

b) **Energy Allocation and Pricing Mechanism**

This element follows the market operation rules to allocate the energy between buyer and seller with a specific pricing mechanism for carrying out the trading. It also needs to consider the fairness to all buyers and sellers during the allocation in order to provide an unbiased financial platform for all prosumers to be engaged in. The allocation of energy and respective pricing also depends on the overall state of energy within the market but also needs to reflect the prosumer's individual trading behavior [14, 62].

c) **Information System**

Information system forms heart of the P2P energy trading in that it enables the communication between all the market participants, provides an equal access to all prosumers for trading energy, integrates all the elements of P2P network together and ensures a secure platform for trading [14, 62]. With rise in blockchain technologies in recent times, it has served as an

efficient information system to realize a P2P energy trading network and make it a fully decentralized system altogether [14, 63, 64].

d) **Regulation**

In order to make P2P energy trading a legitimate market and ensure the participation of prosumers, there needs to be a clearly defined all-encompassing regulations and energy policies that will govern such market and determine the taxes, charges, and incentives for being involved in the market.

e) **Energy Trading Management System (ETMS)**

ETMS manages the supply of energy for market participants while carrying out the trading. It has access to demand and supply information of the prosumers and utilizes that to manage the exchange of energy. It can also involve automatic energy trading strategies on behalf of prosumers [14, 62].

2. *Physical Layer Elements*

a) **Grid Connection**

This is the fundamental element of the network through which the energy is transferred physically among prosumers and/or grid [14, 62]. Such an energy network could be connected with the main grid as a backup or it could also serve as an islanded microgrid system. In the latter case, it needs to be ensured that the energy generation is sufficient enough to meet the energy demand of the market participants for a given market-time horizon.

b) **Metering**

In order to evaluate and monitor the energy exchanges on transaction basis, the advanced metering devices need to be installed at every prosumer's connection point so as to track and record how much energy was bought

or sold during a given time period for a given prosumer. Furthermore, these metering devices also need to communicate with each other to ensure the inflow/outflow of energy as required by the virtual layer components.

c) **Communication Infrastructure**

Communication infrastructure is what makes the exchange of information and all the communication between stakeholders possible. It could employ different communication architectures to support the flow of information within the networks in real-time with specific consideration to latency, throughput, security, and reliability as mentioned in the work of [7]. The communication infrastructure includes networks and technologies that allows for distribution of information related to measurement and commands within the network [7].

3. *Human Participation*

a) **Market Participants**

There needs to be sufficient number of market participants for enabling the P2P energy trading and also a mechanism to keep them tied up with the market so as to make it sustainable over longer period of time. Furthermore, there needs to be sufficient number of sellers as well as buyers to support the trading.

b) **User behavioral modeling**

Demanding active participation from prosumers for every transactions in each time-slot can be distressing for them and might also result in potential abandonment of the system. Therefore, in order to ensure the sustained participation of users, the system should demand minimal active participation from users by automating the trading decisions on their behalf with occasional feedback from them to guide the system. This

requires implementing a proper user behavioral modeling that reflects the involved users' perception and preferences for such trading behaviors. We discuss more on user behavioral modeling in section 2.3.

2.2.2 Possible Market Structures for P2P Energy Trading

Few works have looked into what the possible market structures will look like in practical setting for P2P energy trading which includes [1] and [7]. The authors in [1] have recognized three modality to incorporate prosumers into the scene namely, i) prosumer-grid integration model, ii) fully decentralized model, and iii) prosumer-prosumer organized community group model. It further highlights that these markets have capability to maximize the energy efficiency effort along with decentralizing and democratizing the energy trading environment.

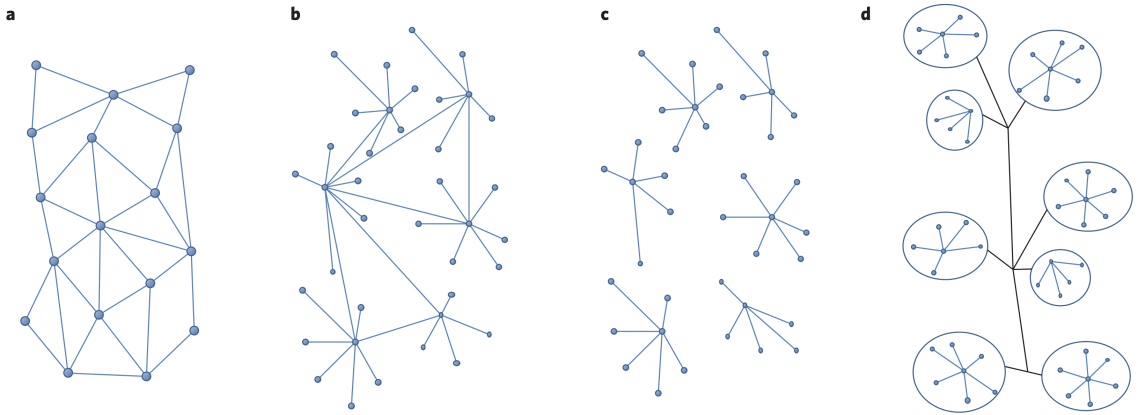


Figure 2.3: P2P Energy Market Models [1]: a) Fully Decentralized Model; b) & c) Prosumer-Microgrid Model - b) Prosumer-Interconnected Microgrids, c) Prosumer-Islanded Microgrids; and d) Organized Community Model

1. *Fully Decentralized Model*

This model involves direct energy exchanges between peers and can be considered decentralized, autonomous and flexible market modality for P2P energy trading as shown in Fig.2.3(a). It relies on bilateral contracts between individual prosumers that interconnect with each other directly without any

central influence for trading energy services. One such mechanism for bilateral contract is proposed in [65] for P2P energy trading with real-time and forward contracts between peers and utility-maximization in concern. This kind of model might reflect the sharing economic modality that has been promoted by companies like Airbnb in case of sharing accommodation spaces and Uber in case of sharing privately owned vehicles. The work in [13] has also discussed about such possibility of sharing economy modality in case of sharing electricity between prosumers. However, this could present challenges to the operation of such a market without any central influence and also regarding the ownership of infrastructures. In case of non-cooperative behavior from certain prosumers or malicious attack on the network, this could lead to the disruption of whole network and therefore, there could be potential concern for liability and accountability for ensuring safe exchange of energy between peers [1].

2. *Prosumer-Grid Model*

This model is more structured than fully decentralized model. It involves the interconnection of peers using a microgrid with (as in Fig. 2.3(b)) or without (as in Fig. 2.3(c)) the involvement of main grid [1]. In this market model, the energy can be traded among prosumers through the brokerage of the interconnected microgrid and therefore the liability for safe energy exchange is shifted toward the grid owner. If the main grid is involved, it can serve as a backup for prosumers to sell or buy excess energy generation or demand that cannot be traded within the local microgrid. In case of islanded microgrid, the energy generation and demand needs to be managed within the microgrid through load-shifting and storage options [1].

3. *Organized Community or Composite Model*

This represents a market model in which the prosumers in closer proximity

form an organized community like VPP and pool resources for use among each other [1]. These VPPs are in turn connected with each other through grid to engage in energy trading services among and within these VPPs. A general representation of such a market model is presented in Fig. 2.3(d).

The selection of the appropriate market structure depends on geographical location of such market and similarity of consumption/generation profiles of prosumers among other several factors. Another major factor could also be the role of grid in such market and how they can be integrated into the P2P energy trading scene.

2.3 User Behavioral Modeling and Behavioral Economics

User behavioral modeling in general has become a topic of attention in recent times given the rising interest towards intelligent and automated systems/algorithms. In particular the user behavioral modeling have become a beneficial tool to guide and nudge the intelligent systems and algorithms to provide personalized, customized and adaptive recommendations, feedback and choices for users [66]. User behavioral modeling focuses on building a mathematical construct that reflects the complex qualitative and behavioral patterns of users in order to determine and/or predict how the users might behave under similar circumstances for a given application [66–68].

The user behavior models in earlier days relied on hand-crafted knowledge that in turn are accumulated from inferences made through observation about users [68]. But with the wealth of data that is generated through internet and internet-of-things, statistical models have stood up as a favorable alternative to traditional knowledge-based method for user modeling [68]. There are several techniques under predictive statistical models that have been reinforced with the emergence of Artificial Intelligence. These statistical models have been used to learn user behavior with the help of tools like decision trees, machine learning, neural networks, Bayesian networks and so on [68]. These techniques are data-driven which learns the underlying pattern accordingly from the provided datasets and therefore has served as a beneficial way to

learn the user behaviors. This mechanism for behavioral modeling involves learning of the model from historical data.

Another mechanism to learn the user behavioral pattern include an online learning process where the parameters of the model are learned over the time with human-in-the-loop. It can also be content-based meaning that the behavior of a user is predicted from their past behaviors or it can be collaborative meaning that behavior of a user is predicted from behavior of other users with certain similarities [68]. Either way, as pointed out by [66, 68], the accurate representation of user behavior is hugely significant in the applicability of an intelligent system/algorithm that focuses on automating the tasks on behalf of users.

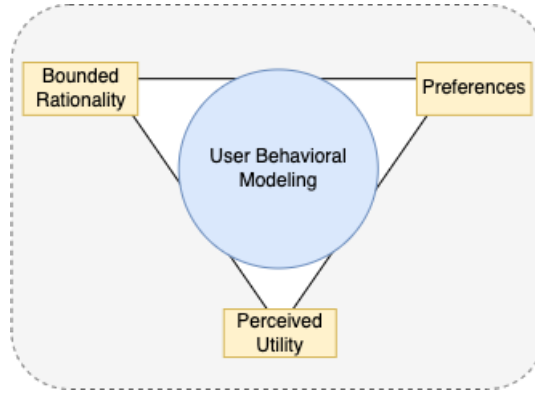


Figure 2.4: Notions of User Behavioral Modeling

It is very difficult to model human psychology and more so to duplicate human decision making under risk [25, 27, 69, 70], which is where the behavioral economics comes into play. Behavioral economics extends the traditional economic model with realistic psychological foundations that emphasizes on limited cognitive capabilities of humans and their perception of loss and gain as deviation from a reference point [25, 69, 70]. This helps to model the user behavior more accurately and particularly their non-rational decision making under different circumstances [25, 69, 70]. *Bounded Rationality* is one aspect of behavioral economics that captures the limited cognitive capabilities of humans and provides explanation to their non-rational decision making

in the face of overwhelmed choices or intractability of the decision problem [25, 70].

Another major concept in behavioral economics is *Prospect Theory* (PT) [69] which can be used to model the non-rational user behavior in the face of uncertain decision-making. It is often regarded as fairly accurate mathematical representation of human behavior [17, 42, 69]. PT specifically models the risk-seeking and loss-averting behaviors of humans into a mathematical construct. It extends the concept of *financial value* of loss and gain and transforms it into the *perceived value* of loss and gain. This perceived value is determined with the help of the diminishing marginal sensitivities for both loss and gain with respect to deviation from some reference point and adding probability weighting to the prospects before making decision on which prospect to choose [69, 70]. Eq. (2.1) represents the prospect theory value function that accounts for perceived utility as a deviation from a reference point, r_0 .

$$v(x) = \begin{cases} k_+(x - r_0)^{\zeta_+}, & x \geq r_0 \\ -k_-(r_0 - x)^{\zeta_-}, & x < r_0 \end{cases} \quad (2.1)$$

As shown in figure 2.5 and recreated from [69], the value function is evidently non-linear with respect to x , concave in gain domain, i.e. ($x \geq r_0$), convex in loss domain, i.e. ($x < r_0$) and steeper for loss than gains emphasizing the loss-averse tendency of humans [69].

This notion of reference point, as mentioned in [69], is compatible with basic human perception and judgement apparatus. Human response to qualitative and subjective attributes are always perceived in relation to something they are familiar with or in relation to their current preference/status in regards to such attribute [69]. Some examples would be loudness and temperature, which people prefer what they are familiar with or have been subjected to either in the past or in the present [69, 70]. However, in addition to that, the reference point could also be an expected value that the agent aims to achieve regardless of their current status. As an illustration,

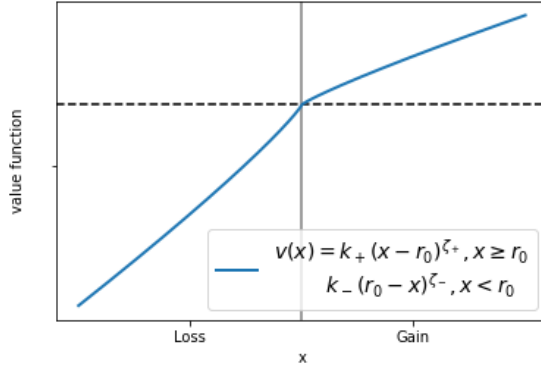


Figure 2.5: A prospect theory value function $v(x)$ with $\zeta_+ = \zeta_- = 0.88$ and $k_+ = k_- = 2.25$

the relevant reference point for determining perceived loss and gain could either be the current status of wealth/welfare or some expected value that they strive to achieve [70]. The perceived value is complemented with the integration of risk-seeking and loss-averting behaviors in prospect theory to model the decision-making under uncertainty and risk [27, 69, 70].

As highlighted in [70], user behavioral modeling requires learning of the behavioral patterns of user over time and making refinement to the existing model to capture the realistic non-rational behavioral patterns. Human psychology is a complex and dynamic thing [67, 69, 70] and the parameters of a user behavioral model is not known *a priori* as such in realistic setting. In order to learn any parameters associated with a model, one could either resort to statistical approaches that learn from the historical data and then apply the knowledge obtained from those data to predict in same or some another setting. However, this approach requires the availability of data which might not always be possible. Another approach to learn the parameters is associated with online adaptive learning process. In such process, the system is fed with initial estimate and then based on the outcome the estimate is updated accordingly over each iteration which is expected to converge after some iterations. One of such adaptive learning methods is *Reinforcement Learning* (RL) which utilizes a dynamic learning

process to learn an optimal policy [71]. In the next section, we explain RL in brief.

2.4 Reinforcement Learning

Reinforcement Learning is a sequential decision-making and behavior learning model that learns to take the best actions (in terms of maximizing the reward or minimizing the regret) through trial and error of interacting with the environment. In RL, the actions are taken over time by balancing exploration of the unknown states and exploitation of the known states based on existing knowledge. The reward from the environment for taking that particular action is then observed as a reinforcement or feedback to update the existing knowledge [71, 72]. This adaptive learning process of reinforcement learning renders it practical in settings where the complete knowledge of the environment is missing or complete control over environment is not required. Therefore, in recent times there has been a huge surge in interest over reinforcement learning based algorithms which has found application in diverse fields ranging from games to self-driving vehicles to robots to smart grids [72].

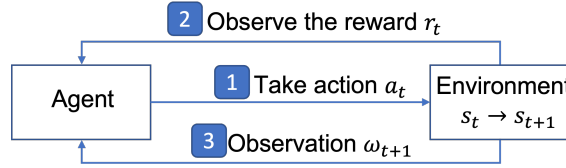


Figure 2.6: Agent-Environment Interaction in RL

The general framework for an RL problem is governed by Markov Decision Process (MDP). An MDP consists of 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$ where:

- \mathcal{S} is the state space of the environment
- \mathcal{A} is the action space that an agent can choose to take an action from
- \mathcal{T} is the transition probability of the agent transitioning from one state to the next

- $\gamma \in [0, 1)$ is the discount factor that puts weight on how much the immediate rewards value over the long term rewards
- \mathcal{R} is the reward function that provides a quantitative feedback from the environment for the actions taken by the agent.

RL in its basic form can be regarded as an interaction between an agent and environment as shown in Fig. 2.6. At each time step t , the agent takes an action ($a_t \in \mathcal{A}$) according to exploration-exploitation strategy which takes the agent to a new state ($s_{t+1} \in \mathcal{S}$) from state ($s_t \in \mathcal{S}$). The agent obtains a reward ($r_t \in \mathcal{R}$) from the environment as a feedback on how good the action taken by the agent was and the agent obtains an observation ($\omega_{t+1} \in \Omega$). The objective of an RL is to learn an optimum *policy* and the policy in case of RL is how an agent selects action(s). Policies can be either deterministic or indeterministic:

- Deterministic case: the policy is defined by states alone as $\pi(s) : \mathcal{S} \rightarrow \mathcal{A}$
- Stochastic case: the policy is defined by state-action pair as $\pi(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ and $\pi(s, a)$ is the probability that action a may be chosen in state s

The RL agent is seeking to find a policy $\pi(s, a) \in \Pi$ that maximizes an overall expected return $V^\pi(s) : \mathcal{S} \rightarrow \mathbb{R}$ over infinite horizon. This expected return is called V-value function which is defined as

$$V^\pi(s) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, \pi \right] \quad (2.2)$$

The optimal expected return or V-value function, therefore would be the return that is received if the policy chosen is maximum one i.e.

$$V^*(s) = \max_{\pi \in \Pi} V^\pi(s) \quad (2.3)$$

In addition to V-value function, another widely used metric in RL is the Q-value function $Q^\pi(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow R$ defined as

$$Q^\pi(s, a) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a, \pi \right] \quad (2.4)$$

and this equation in turn can be rewritten recursively for an MDP using Bellman's equation as

$$Q^\pi(s, a) = \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') \left(R(s, a, s') + \gamma Q^\pi(s', a = \pi(s')) \right) \quad (2.5)$$

The optimal Q-value function, therefore would be similar to Eq. 2.3:

$$Q^*(s, a) = \max_{\pi \in \Pi} Q^\pi(s, a) \quad (2.6)$$

and the optimal policy can be found through optimal Q-value function

$$\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q^*(s, a) \quad (2.7)$$

Eq. 2.7 is the reason the Q-value function is more preferred than V-value function most of the time as it allows to determine optimal policy directly from optimal Q-value function.

From the above discussion, it can be stated that there are three different approaches to learn a policy which are:

- Value function representation that provides prediction of how good each state or each state/action pair is
- a direct representation of policy $\pi(s)$ or $\pi(s, a)$
- a model of the environment that is represented by the estimated transition function and estimated reward function

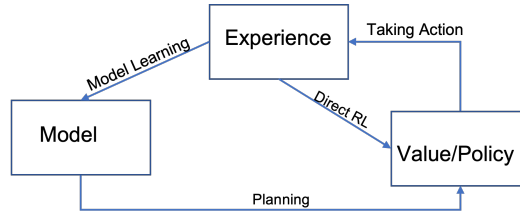


Figure 2.7: General Schema of RL Methods. Direct approach utilizes value/policy function and indirect approach utilizes a model of the environment.

The first two approaches belong to a broader category of *model-free* RL and the third approach is *model-based* RL which is clarified through pictorial representation in Fig. 2.7. The direct approach involves learning value or policy function while the indirect approach involves learning of the model of the environment itself.

Based on the three approaches to learning policies, different RL algorithms can also be classified under three headings as follows:

1. Value-based Methods

- Q-learning
- Fitted Q-learning
- Deep Q-Networks (DQN)

2. Policy-based Methods

- Stochastic Policy Gradient
- Deterministic Policy Gradient
- Actor-Critic Method

3. Model-based Method

- Lookahead Search - Monte Carlo Tree Search (MCTS)
- Trajectory Optimization - PILCO, Guided Policy Search

2.5 Spatial Crowdsourcing and Use of Electric Vehicles for Crowdsourcing-based Energy Sharing

Spatial crowdsourcing is a recently emerging concept that leverages the power of the crowd to complete spatial tasks such as mapping, monitoring, and surveillance [73,74]. It involves the use of mobile devices equipped with various sensors and IoT devices such as GPS, cameras, and accelerometers, which allow users to completing the spatial tasks. Spatial crowdsourcing has a wide range of applications, including disaster response, environmental monitoring, urban planning, and transportation [75].

One emerging area of research in spatial crowdsourcing is the use of electric vehicles for crowdsourcing-based energy sharing through vehicle-to-grid (V2G) and vehicle-to-home (V2H) technologies [50]. V2G and V2H allow electric vehicles to be used as energy storage systems that can help balance the electricity grid and provide backup power to homes during blackouts. This is particularly important for renewable energy sources such as solar and wind, which can be intermittent and unpredictable. V2G and V2H require a large number of electric vehicles to be connected to the grid or home energy system, and this is where spatial crowdsourcing comes in. By incentivizing electric vehicle owners to participate in crowdsourcing activities, such as data collection or task completion, a large pool of electric vehicles can be created that can be used for V2G and V2H. This can help to reduce the cost of implementing these technologies and make them more accessible to a wider range of people.

Overall, spatial crowdsourcing and the use of electric vehicles for energy sharing through V2G and V2H are promising areas of research that have the potential to revolutionize the energy sector. By harnessing the power of the crowd and electric vehicles, researchers can create a more efficient, sustainable, and resilient energy system that is beneficial to all the stakeholders. The benefits can be extended by also incorporating the ride-sharing along with energy-sharing which allows for diverse range of options for the EV drivers.

The works put forth in this dissertation derives extensively from the concepts discussed in this chapter. Specifically, we utilize user behavioral notions like user preferences, bounded rationality and perceived utility along with the reinforcement learning frameworks like Combinatorial Multi-Armed Bandit (CMAB), Q-learning and Deep Q-network (DQN) to devise an automated and prosumer-centric P2P energy trading modality through optimization techniques and crowdsourcing mechanism. The details of the study carried for this dissertation are discussed in detail in the following chapters.

CHAPTER 3. A DETAILED LITERATURE REVIEW ON P2P ENERGY TRADING AND USER BEHAVIORAL MODELING

In this chapter, existing works carried out on the fields of P2P energy trading, user behavioral modeling and reinforcement learning is presented in detail. In doing so, the major limitations of the existing approaches and research in the field is highlighted that serves as the major motivation for the proposed prosumer-centric peer-to-peer energy trading modality.

3.1 P2P Energy Trading

The energy exchange traditionally has been confined between large power producers and grid wherein the grid purchases from these large producers at a pre-agreed price and sells the energy to the consumers through its interconnected network of transmission and distribution lines. In recent times, researchers in this field have focused in diversifying the energy market modality adopted by the grids to accommodate changing energy landscape including topics like coordinated operation of large power plants; incorporating intra-day and real-time energy trading modality into the traditional market and so on. As an example, a privacy-preserving framework is proposed in recent work [76] 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 [77] incorporating the intra-day trading mechanism with traditional day-ahead trading models to increase the profitability and minimize the risks. These are few representative studies among plethora of works being done in the field.

With the growth of microgrids, however, energy exchange between microgrids themselves and also between microgrids and the grid has also been studied in literature and also implemented in real-world setting. There exist several works focusing on the

possible energy exchanges between smaller microgrids with or without the involvement of main grid. Solution to coordinated distributed generation and demand response management problem has been presented in [78] using multi-agent based approach while similar work is done in [59], which adopts game theoretic and hierarchical optimization approaches to minimize the power mismatch in and among microgrids in a multiagent-based energy market. Operational management of a multi-microgrid system is modeled using a joint constraint in a cooperative manner in [79] using stochastic predictive control mechanism. Local energy trading has also been explored among the interconnected microgrids in [80] and [81] in consideration with uncertain parameters in the system.

Of late, the increasing involvement of prosumers into energy market and rise in SG and VPP have made it possible for the prosumers to be engaged with energy market. There have been several mechanisms in place to trade energy between consumers and grid with most widely adopted mechanism being FiT mechanism. However as an alternative to selling energy to the grid, the localized energy exchange among prosumers in a P2P fashion is on the rise. [1] offers perspectives on prosumer-centric energy market and the challenges that entails with them. It highlights that P2P energy market has capability to maximize the energy efficiency effort along with decentralizing and democratizing the energy trading environment. However, it also mentions that the biggest caveat in doing so is the fact that the mechanism required for such prosumer-centric market are either still being developed or not accessible to all prosumers due to lack of support from utility or lack of regulations to govern such mechanisms. Furthermore, in devising such market structures, there needs to be proper consideration for grid reliability, privacy preservation and user behavioral aspect; absence of which is bound to drive prosumers away from such market modality rather than attracting them [1]. P2P energy trading, also referred to as *sharing economy model for energy*, has been regarded as a highly prospective

electricity market model [13, 82]. Its ability to accommodate the need for a rising number of prosumers and to provide benefits to the grid and other stakeholders in today's changing electricity market, has resulted in notable attention from the research community in recent years. Authors in [14] note that in P2P energy trading, the prosumers themselves are in control of setting the terms of the transactions and therefore the individual gains that they receive from participating in such trading is significant in determining the energy trading model. It further mentions that based on current emergence of SG and blockchain technologies, deployment of P2P energy trading can provide a highly efficient and cost-effective energy management technique in a decentralized way.

There are several real-life implementation of such energy exchanging modality and other numerous literary works that consider energy trading among small-scale local prosumers and consumers. The energy exchange in a localized setting is studied in [35] which utilizes a DC power sharing among nearby homes. It addresses the problem of mismatching between energy harvesting and consumption in a microgrid and proposes a greedy approach that maximizes the energy exchanges among users while minimizing loss and energy waste. Furthermore, self-consumption of locally generated energy within a microgrid has been studied in [24], which presents a 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. The authors in [10] have developed a basic P2P model to incentivize the prosumers and also address the power loss reduction and line congestion in physical electricity lines. Similarly, the power loss in electric lines is also considered in [36] along with privacy of the prosumers in a P2P energy trading market and effectiveness of the decentralized energy market is established through simulation results.

The work in [15] has regarded the user participation as a central aspect of the P2P energy trading but is only limited to the coalition formation in game theoretic setting

and does not explicitly consider user behavioral modeling. The work in [44] also uses a prospect theory based distributed energy trading model to optimize trading decisions for prosumers in a competitive market but with active human participation in the form of bids and aims at learning the aggressiveness of the bids over time. A mid-market price based P2P energy trading market is presented in [34] that allocates energy from the cheapest sellers first to the buyers in order of their registration. In [83], 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 with uncertainty of energy demand and energy production in consideration.

These are some of the works that consider some form of energy exchanges between peers. There exist several studies that consider several aspects of P2P energy trading and in the next section we review these literature based on the approaches they employ in their work.

3.2 Technical Approaches Employed for P2P Energy Trading

There have been several works focusing on different aspects of P2P energy trading that use several mathematical, economic and computer science-based constructs to realize it. Based on these works, we can basically divide the existing P2P energy trading mechanisms into four major approaches : 1) Game Theory, 2) Auction Mechanism, 3) Constrained Optimization, and 4) Blockchain. Table 3.1 presents a tabulated overview of general approach and methods employed for P2P energy trading in these literature.

1. *Game Theory*

Game theory is a mathematical tool that helps in decision-making among rational agents through their strategic interactions [14,33,39,84]. This strategic interaction can be non-cooperative or cooperative. In non-cooperative game, the agents with partially or completely conflicting interests take their decisions

Table 3.1: P2P Energy Trading Approaches

Technical Approach	Methods Employed	Literature
<i>Game Theory</i>	Stackelberg Game, Non-Cooperative Nash Game, Canonical Coalition Game	[39], [84], [63], [85], [15], [33]
<i>Auction Mechanism</i>	Double Auction, Reverse Auction	[86], [64], [87], [88]
<i>Constrained Optimization</i>	Linear Programming (LP), Multiple Integer Linear Programming (MILP), Alternating Direction Method of Multipliers (ADMM), Non-Linear Programming	[45], [60], [89], [90], [91], [37]
<i>Blockchain</i>	Smart Contract, Consortium Blockchain, Elecbay, IBM Hyperleger Fabric, Ethereum	[63], [64], [92], [93], [34]

without communicating with each other (eg. Nash Equilibrium) while in cooperative game, the rational agents form a coalition with other agents to improve their position in the game (eg. characteristic form) [33]. It is to be noted that game theory considers the agents to be rational entities that are always capable of making optimal decisions [14, 33]. Furthermore, it requires constant active participation of these agents in the decision-making process which can be static or dynamic depending on whether the agents can take action only once or can continually update their actions depending on how other agents act.

When it comes to P2P energy trading, huge amount of works tend to resort to game theory for carrying out the trading between rational peers. A cooperative Stackelberg game is formulated with dynamic energy pricing for P2P energy trading in [39] with grid as the leader that sets the prices and the prosumers as the followers. The prosumers seek to get incentivized for managing the energy demand locally through coalition and helping to reduce the energy demand for grid for a given time period. Similarly, the work by [84] also uses non-cooperative Nash game theoretic formulation to propose a hybrid approach for load scheduling and energy sharing. They conclude that the approach serves significant cost benefits to the participating users without sacrificing the fairness. However, the work depends on finding a Nash bargaining solution

which is significantly challenging computation-wise and expected to perform worse with increase in scale. In [85], a distribution side energy trading market is proposed as a non-cooperative, multiplayer game that finds a Nash equilibrium solution through an extremum seeking algorithm. The problem consists of determining market clearing prices and quantity for each players based on Nash equilibrium solution. [63] uses the Stackelberg game to design a pricing scheme for P2P energy trading that uses consortium blockchain technology as a platform for secure transactions. In [15], a canonical coalition game is devised to share energy among groups of peers through social coalition for ensuring user's sustainable participation in the market.

2. *Auction Mechanism*

Auction mechanisms are the market institution that determine the resource allocation and price on the basis of bids and ask price submitted by market participants. Mostly employed auction mechanisms in case of energy markets are the double auction and reverse auction. Double auction involves buyers submitting their bids to an auctioneer while sellers, meanwhile, submit their ask prices. The sellers are arranged in increasing order of their ask prices and the buyers are arranged in decreasing order of their reservation bids. An intersection point is determined from aggregated supply and demand curve that eventually determines the breakeven auction price and list of sellers/buyers that will engage in trading process. This list consists of sellers with ask price below the breakeven price and buyers with bid above the breakeven price. On the other hand, reverse auction is the market mechanism wherein sellers submit bids and compete to sell their product to the buyers. For auction mechanism to work efficiently, the reservation prices and/or bids reported by market participants need to be truthful [14], which therefore requires the auction mechanism to have individual-rationality and incentive compatibility properties.

There are few works that employ auction mechanism as the tool to emulate energy trading mechanism among prosumers. The authors in [86] aim at designing the decentralized P2P architecture that explicitly deals with the physical network constraints while enabling the energy trading between buyers and sellers in a step-by-step continuous double auction approach. Similarly, the study in [87] deals with learning the interaction between prosumers' bidding actions and market response from historical transaction data and uses that to achieve optimal operation and maximizing profits in a P2P electricity market in a double auction setting. A consortium blockchain-enabled double auction mechanism to determine energy and price traded between prosumers is presented in [64]. Likewise, [88] proposes a blockchain-enabled reverse auction mechanism for dynamic pricing between charging and discharging EVs in a peer-to-peer fashion using vehicle-to-grid network.

3. *Constrained Optimization*

Constrained optimization is a widely used mathematical tool for optimization. It involves optimizing a certain objective function under the influence of constraints that limit the region of operation of variables. These objectives can be linear or non-linear and therefore, are called Linear Programming (LP) and Non-Linear Programming (NLP) respectively. Similarly, it could also involve types of variable that are employed and one such special case is Mixed-Integer Linear Programming (MILP) that involves mixed set of integer and non-integer variables. A canonical form of constrained optimization can be expressed as

$$\text{Maximize } f(\mathbf{x}) \tag{3.1}$$

subject to some constraints for variables like $\mathbf{Ax} \leq \mathbf{c}$ and $\mathbf{x} \geq 0$ where \mathbf{x} is the vector of variables that is to be determined, c is the coefficient

vector while \mathbf{A} is the coefficient matrix. Eq. (3.1) is the objective function that is to be optimized while determining the variable vector that satisfy the constraints of the problem. Depending on whether Eq. (3.1) has linear or non-linear function, the optimization problem becomes linear (LP) or non-linear optimization problem (NLP).

There are several approaches devised to solve complex optimization problem, one of such methods is Alternating Direction Method of Multipliers (ADMM). Basically a modified version of Lagrangian scheme, ADMM divides the complex optimization problem into dual problem with two sets of dual variables. It then solves the dual problem using partial update between these dual variables. The mathematical form of the ADMM is expressed as

$$\text{Maximize}_{\mathbf{x}, \mathbf{z}} f(\mathbf{x}) + g(\mathbf{z}) \quad (3.2)$$

subject to constraints like $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$ and $\mathbf{x} \geq 0$ where \mathbf{z} is the dual variable vector of \mathbf{x} , and $f(\mathbf{x})$ and $g(\mathbf{z})$ are the dual problem and hence can be separated into two objectives as well.

In [89], the authors propose a two-stage control mechanism to realize P2P energy sharing in microgrids. In first stage, they formulate a constrained non-linear optimization problem that minimizes the overall energy cost of the microgrid and in second stage, they conduct a rule-based controlling of actual set-points based on real-time measurement with optimal scheduling obtained from stage one as a constraint for updating the said set-points in physical application. Another work employing constrained optimization include [90] which utilizes MILP to optimize the operating decision for P2P electricity trading scenario considering practical constraints surrounding DG resource like rooftop PV and battery storage. ADMM is used instead in [91] to minimize the overall

individual cost functions of all participants in a P2P market in a decentralized way that reflects the negotiation between the buyers and sellers. ADMM is employed in [65] as well to optimize the utility of all participants of P2P energy trading in the form of their individual preferences. Similarly the reactive power optimization in P2P energy trading is studied in [37] using optimization framework and differential evolution-based algorithm is developed to solve this problem with higher dimensionality.

4. *Blockchain*

First introduced in [94], the Blockchain Technology (BCT) is a distributed data structure that provides a decentralized alternative to secured trading platform. Based on replication of blockchains and consensus mechanism within members of network, it allows the electronic transactions to be carried out without the need for a trusted intermediary. Given the decentralized structure of P2P energy trading, BCT can only complement the P2P energy trading with its secured decentralized financial platform. There has been some efforts in integrating BCT with P2P energy trading modality.

The works in [63] and [64] use blockchain-enabled P2P energy trading with Stackelberg game and double auction mechanism respectively. Another literature that takes advantage of state of the art blockchain technology IBM hyperledger fabric architecture is [92]. It realizes a crowdsourced energy systems that supports the P2P energy trading between prosumers and/or the grid. With the inspiration from original BCT, an efficient and secure method to carry out P2P energy trading in microgrids called *Elecbay* has been proposed in [93] that utilizes game theory and Nash equilibrium to determine the energy traded among peers during the bidding process. A simple P2P energy trading mechanism is conceived in [34] that considers mid-market price between buyer

and seller and assigns energy from the cheapest seller to the first registered buyer in the blockchain registration order before moving onto the next buyer and then to next seller accordingly.

3.3 Real-world Implementations of P2P Energy Trading

In addition to the works in literature, there has been some real-life attempts at realizing the P2P energy trading modality albeit in a preliminary form. A real life commercial implementation of P2P energy market is offered by Brooklyn Microgrid (BMG) [21], Vanderbron [22] and Piclo [23]. BMG offers a localized energy marketplace in New York City. It 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 [21]. Similarly, the Dutch start-up Vanderbron enables the local renewable electricity generators to sell their energy under an online P2P marketplace platform independent of any grid or government agency with a small flat subscription charge for both buyers and sellers [22]. Piclo [23] is another start-up based on United Kingdom that aims at building and promoting decentralized energy landscape. It conducted a pilot program for its P2P energy trading initiative [95] that allows renewable producers to set the price and sell energy to local consumers. It notes that this modality offers significant empowerment of distribution side costumers and removes the hurdle for individual participation in electricity market traditionally monopolized by large, and often times fossil fuel based, producers.

Noting the participation rate from public and positive reception of BMG, the authors in [62] made a study with BMG [21] as a test case. They acknowledge the market potential for localized energy trading between energy buyers and sellers, which will only be further amplified in future through the facilitation of blockchain technologies and improving smart grid technologies [62].

3.4 User Behavioral Modeling and Behavioral Economics in Context of Energy Trading

As pointed out by [66, 68], the accurate representation of user behavior is hugely significant in the applicability of an intelligent system/algorithm that focuses on automating the tasks on behalf of users. Furthermore, the mathematical constructs that these user behavioral models require are oftentimes complex and touch on multifaceted domains. One of these fields is human psychology [96] which is driven by numerous known factors and many more unknown factors. Similarly, the user behavioral modeling also requires socio-economic and demographic information which are often times sensitive and private information that are hard to obtain and also lead towards prejudiced outcomes, if proper consideration is not given to fairness in devising the model [97, 98]. [99] further notes that just understanding how the mind works is not enough to build such a system and must be complemented by analysis of data on user behavior, learning the pattern and then making continual refinements on the model over time. It also emphasizes on user-friendliness and simplistic design for such system to encourage sustained user participation.

Note that none of the papers mentioned in this chapter so far consider the complex aspects of user behavior, thus assuming users to be either extraneous to the system or fully compliant with the system decision. They usually assume the users to be rational entity and objective decision-maker. Therefore, for these reasons the lack of realistic modeling may cause failure when implementing P2P energy trading in the real world [96, 100, 101].

3.4.1 User Perception in Energy Sharing Systems

In developing a sharing economy business model for energy, due consideration must be given to behavioral aspects of prosumers along with easing of regulatory barriers for such market modality [82]. Similarly, [14] and [1] have given emphasis on accommodating the user behavioral modeling in the P2P energy trading problem

so as to ensure the sustained participation from prosumers while incentivizing their contribution in such trading. P2P energy trading mechanism needs to manage the energy exchange between buyers and sellers on the basis of their individual preferences, behaviors, perception of loss and gains that will accurately model the user behavior in this setting. These behavioral patterns are then used to maximize the performance of overall system while simultaneously automating the task on behalf of users. So the need for modeling user behavior as accurately as possible is essential for ensuring sustainable participation of prosumers in the P2P energy market. The works [14, 17, 43, 44] have highlighted on the importance of modeling the non-rational behavior of users in the face of uncertainty and risk. [17] notes that the user-behavior can deviate from rational principles of conventional approaches when it comes to making choices and averting losses. Hence, assuming that these users are some rational entities, who always make optimal and correct decisions, is essentially flawed [17, 42, 100] and therefore, in devising a P2P energy trading market, due consideration must be given to the user behavioral modeling with concepts from behavioral economics as a foundation for the mathematical construct.

In general, there have been some researches that focus on user behavioral modeling in case of energy market. Modeling user behavior in SGs has been considered in the context of Demand Response (DR) [40, 102] 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. In [41], several internal and external factors that affect the prosumer's energy sharing behavior are identified that can be used to segment prosumers with similar energy behaviors together in the form of virtual prosumer-community for energy sharing with another such community. [103] and [100] devise several social-behavioral models to define a user utility on perceived importance of appliances which is sought to maximize through a constrained optimization problem. They have found that including social-behavioral

models, that capture complex human psychology into the problem, significantly improves the energy consumption efficiency in a smart residential environment.

It is to be highlighted that P2P trading needs to manage the energy exchange between buyers and sellers on the basis of their individual preferences, behaviors, and perception of loss and gains [1, 45]. These behavioral patterns can then be used to maximize the performance of the overall system. Further, as established in [1, 15, 45, 104], accommodating the user behavioral modeling and motivational psychology in P2P energy trading is essential to ensure sustainable participation of prosumers in the P2P energy market while incentivizing their contribution. In line with that rationale, the study in [45] tries to model bounded rationality and user preferences in a P2P energy trading scene but it requires continuous human participation and assumes a simplistic linear model for user perception. Conversely, the studies in [15, 105] focus a game-theoretic approach for energy trading between prosumers through social coalition formation with motivational psychology models as framework including economic benefit and positive reinforcement models. Regardless, they do not explicitly consider user behavioral modeling in the system design and focus on how coalition game theoretic setting motivates the users to act in cumulative benefit.

3.4.2 Prospect Theory in P2P Energy Trading

Recently, there has been a few efforts in integrating Prospect Theory (PT) in energy related applications as well to capture the irrationality of users under uncertain decision-making [16, 42–44, 106]. These papers notice that the classical game-theoretic approaches consider users to be rational decision-makers which does not reflect the actual behaviors 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 relation to peer-to-peer energy trading, the authors in [44] has proposed a prospect theory-based distributed energy trading model to

optimize trading decisions for prosumers in a competitive market with active human participation as well as an adaptive learning process for learning the aggressiveness of the bids over time. In [16], authors present a framework for energy storage management to allow users to store or sell the energy while modeling the user's subjective perceptions of probable outcomes using prospect theory.

Similarly, a non-cooperative game between prosumers is formulated to meet their energy demands by incorporating prospect theory to model the prosumer's perception of the probable profits surrounding stochastic wind generation and decision-making pertaining to either selling or storing energy at a given time [17]. The authors in [42] 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. [107] studied the energy exchange between microgrids with capability of energy generation as a prospect theory based energy exchange game. There has also been an effort on modeling the user's perception towards bidding result in a power market using prospect theory [43], that uses genetic algorithm for solving the optimal power market bidding problem. Additionally, the authors in [108] use prospect theory to model the user response to energy prices, and focus on the impact of such realistic behaviors on the system. The work in [109], on the other hand, uses prospect theory for demand-side management to figure out for grid customers whether or not to participate in load shifting based on their subjective perception towards such decisions. Similarly, the study in [106] considers prospect theory in the optimization framework for P2P energy trading to incorporate the risk-aversion attitude of end-users. The authors in [110] also try to capture the complexity of prosumer's behavior in P2P trading between microgrids.

Although these papers model the user behavior in some ways, they require active participation from users and also assume that such behavior (e.g., the parameters of the PT) is known *a priori* and homogenous for all the users. Social science studies,

such as the one conducted in Italy to investigate the social acceptance of nuclear energy using an online survey [111], 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 [112]. Neuroscience studies have also stressed the heterogeneity of humans in reference to PT parameters [113]. Not capturing such heterogeneity provides little benefits in terms of user behavioral modeling. Therefore such individual preferences and perceptions need to be captured in user behavior models through an adaptive learning process that is tailored for individual prosumers. Furthermore, as [99] has highlighted, the behavior of users might change over time and hence, the user behavioral modeling also needs occasional human refinement/feedback to reflect the changed user behavior accordingly.

3.5 Reinforcement Learning in Energy Trading

RL has gained widespread interest in recent years, more so in applications where an agent need to learn an optimal policy over time based on the feedback it receives from the environment. It has also found its way into energy related applications for example energy management in residential setting, power exchanges in a microgrid and also in large scale grid settings. Pertaining to peer-to-peer energy trading, in [87], energy trading between local prosumers is proposed that exploits Deep Reinforcement Learning (DRL) to learn optimal energy trading strategy for each prosumers with energy storage at each time-step. [114] proposes a distributed energy marketplace for grid to leverage prosumers' storage capacity with monetary incentives and uses reinforcement learning to learn an optimal decision-making that maximizes the economic benefit of all agents i.e. grid and other prosumers. Similarly, [115] uses Coordinated Q-learning to optimally schedule the battery for PV-based prosumers to reduce power consumption from grid in a distributed and cooperative way.

The authors in [116] devised a reinforcement learning framework, specifically Temporal Difference(0) (TD(0)) and Q-learning with risk-sensitivity of agents included. This approach termed *Risk-Sensitive Reinforcement Learning*, the authors model risk-sensitivity through a transformation function with desired risk sensitivity parameter to transform the temporal differences that are encountered in learning process. Similarly, [117] extends the work of [116] by utilizing prospect theory as a risk-sensitive metric to transform temporal difference error while learning. This approach integrates the behavioral modeling through prospect theory valuation function into the reinforcement learning framework to make informed decision-making that reflects risk-sensitivity (i.e. risk-seeking and/or averting behavior) of agents.

3.6 EV-based Crowdsourcing for Energy Sharing

Crowdsourcing services has received increasing attention in recent years because of their flexibility and convenience in facilitating the completion of tasks by a set of workers [73]. There exists a plethora of research works that focus on different aspects of crowdsourcing from optimal task allocation [52] to preference-aware decision-making [118] to privacy-preserving [119, 120]. Some other focus on designing an effective and informed incentive mechanism that motivates workers for their sustained engagement in the system [120].

Reverse auction mechanism has been widely utilized for designing incentive mechanism including bidding and winner selection in crowdsourcing works [121–124]. In [122], a secure reverse auction protocol is devised for task assignment for spatial crowdsourcing along with an approximation algorithm. Similarly, [123] proposes a truthful reverse auction mechanism for location-aware crowdsensing while authors in [121] focus on generalized second-price auction for stable task assignment. The work in [124] also uses a truthful reverse auction mechanism to devise incentives for workers in urban parcel delivery.

In context of electric vehicles (EV), the work in [52] employs crowdsourcing for

solving charging problems of EVs. A V2V energy-sharing framework has been proposed that crowdsources the charging request from EV owners and allocates the energy considering energy trading prices, EV parameters and privacy. Some other crowdsourcing literature focus on different problems like route optimization of EVs [125] and parcel delivery using EVs [126]. Closer to our problem setting, some literature have explored the use of crowdsourcing for integrating energy-sharing services with EVs. For instance, authors in [75] proposed a V2H-based omni-sharing modality system in a microgrid community, where energy is crowdsourced from EVs to reduce the overall cost of the community and decrease the need for energy storage. Another study [127] suggested an autonomous EV-based energy crowdsourcing approach, which enables EVs to participate in energy-sharing tasks for cloud-based energy consumers. However, this approach is challenging to implement and doesn't consider workers' preferences or the impact of sub-optimal decision-making in the context of crowdsourcing-based energy-sharing.

In fact, most of these crowdsourcing works ignore the user behavioral modeling in task assignment. The spatial crowdsourcing work in [118] tried to solve the task assignment problem by considering worker preferences, but this solution is better suited for group tasks and doesn't account for other behavioral aspects of user behavioral modeling like *bounded rationality* [27] and irrational decision-making that drastically affects the system performances. Additionally, the existing works neglect the task recommendation problem and other realistic budget constraints, such as the energy budgets required by the grid companies or microgrids for any time period. Furthermore, these works are limited to homogeneous tasks like energy-sharing or delivery services only, which can result in significant idle hours for EVs during off-peak periods as such tasks have similar pattern.

3.7 Limitations in Existing Literature and Motivation

As discussed in this chapter, there has been a number of researches and works in context of localized energy exchange and some form of user behavioral modeling along with reinforcement learning. With large scale integration of DERs into the grid following the advent in the field of smart grid technologies, the energy grid has been transforming towards decentralized and distributed energy architecture as seen with the proliferation of Virtual Power Plant (VPP) architecture and rise of *prosumers* from *consumers*. Given the limitations of storage options and trading with grid for such prosumers, the exchange of energy between prosumers themselves is beneficial to manage the energy consumption and demand response for grid at a local level. It also provides a financial platform for prosumers to trade energy among themselves at a profitable price compared to the grid. Hence the P2P energy trading modality can be deemed as the future of energy market.

It has been established in this chapter that failure to accommodate complex human psychology and capture their decision-making behavior through user behavioral modeling in an intelligent system will only lead to failure of such system. Therefore, for the efficient management of a P2P energy trading market and for ensuring sustained participation of users in such market, there needs to be proper behavioral modeling that aims at automating the task on behalf of users with due consideration to their preferences, convenience and perceived utility. This kind of algorithmic trading will demand less interaction and lesser active participation from humans which is in line with the findings of several literature reviewed in this chapter. Also having flexibility for the prosumers for engaging in energy exchange will bolster the sustainability of such P2P energy market. Therefore incorporating prosumers with EVs through V2G, V2H and V2V paradigms in addition to renewable energy sources will be helpful to advance the P2P energy trading market a step further.

Even though there has been some work in localized energy exchange setting and

some work in user behavioral modeling, an explicit work that aims at developing a peer-to-peer energy trading modality among prosumers that captures their preference and perceived utility is lacking in the literature. Pertaining to EVs in energy sharing modalities, existing literature in crowdsourcing mechanisms have contributed to task assignment, incentive design, privacy and energy-sharing services, there is still room for improvement in terms of behavioral aspect like preference-aware task recommendation and online learning of these preferences; task assignment with overall cost minimization and energy budgets; and heterogeneity in crowdsourcing tasks. Through this dissertation, we aim to bridge the gaps in existing literature by devising the relevant frameworks necessary to realize the P2P energy trading in smart residential environment through the use of several notions of behavioral economics as constructs to learn the user behavior model. Our proposed work focuses on addressing these limitations and developing more comprehensive, effective, and realistic solution to enabling the P2P energy market including the joint enabling of ride-and energy-sharing services through use of different concepts from power systems, computer science, mathematics, psychology and economics.

However, in our work, we do not consider the physical layer related studies except for loss consideration and energy available or required for prosumers, as there exists plethora of research that deal with different physical layer aspects in P2P energy trading setting. Instead, for physical layer related study, we refer the readers to existing works including [10, 36–38] that focus on reactive power optimization, voltage regulation, line congestion among others. Also it needs to be highlighted that although we primarily use energy as the commodity for trading, the power injection and specifically reactive power consideration can affect the such energy exchanges in real-world implementation. We request readers to go over the works in [37, 38] regarding such considerations.

As such behavioral economics offer some basic construct that can be applied into

the problem. We aim to use these concepts from behavioral economics as a basis to model the user behavior and learn the parameters of the model accordingly using reinforcement learning to optimize the energy exchanges between prosumers in a localized setting.

CHAPTER 4. A REINFORCEMENT LEARNING APPROACH FOR USER PREFERENCE-AWARE P2P ENERGY SHARING

In this chapter, a P2P energy exchange system is proposed considering realistic and heterogeneous user behaviors in terms of preferences and engagement. In addition to that, limited time and cognitive capabilities of prosumers is also incorporated into the model in accordance to the principle of bounded rationality. The work done in this chapter was published in IEEE Transactions on Green Communications and Networking [45] and Proceedings of IEEE International Conference on Pervasive Computing and Communications (IEEE PerCom) 2023 [46].

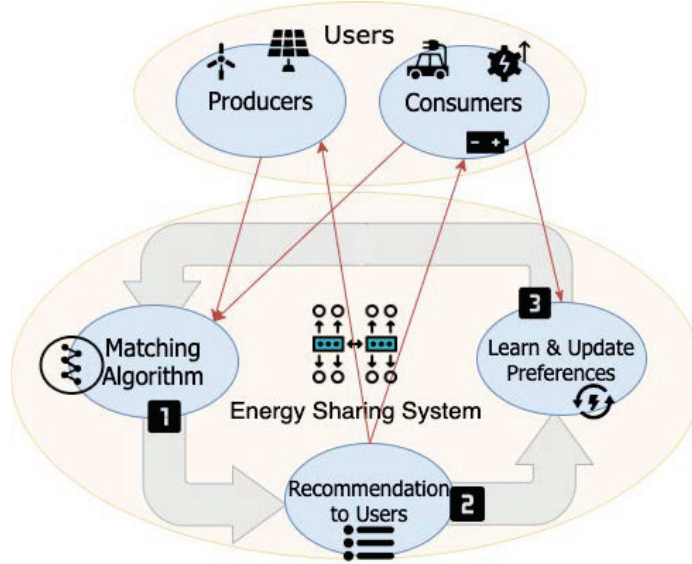


Figure 4.1: P2P Energy Sharing System Overview

A general overview of the P2P system considered in this work is presented in Fig. 4.1. As depicted in the overview, a localized energy sharing community is considered. Within this platform, prosumers are allowed to sell and buy energy to and from other prosumers in the community, as well as from renewable and standard power plants connected to a larger SG. This model is grounded on previous models proposed for such energy exchange modality [1]. In this system, an active user participation in

the day-ahead energy exchange modality is envisaged, where users may have different preferences for different energy sources (e.g., solar, wind, nuclear, coal, etc.), as well as a different level of engagement with the system.

The problem of matching producers and consumers is formulated as a Mixed Integer Linear Programming (MILP), which aims at maximizing the amount of energy exchanged among prosumers considering their individual preferences, the bounded rationality, as well as other physical constraints that affect the energy exchange. Rest of the chapter is organized as follows. The system model and problem statement are described in Sections 4.1 and 4.2, respectively. Then, the proposed algorithms are explained in detail in Sections 4.3 and 4.4. Furthermore, the experimental results are elaborated in Section 4.5 along with the discussion on the observed results.

4.1 System Model and Assumptions

Proposed P2P energy exchange model consists of two sets of users. P defines the set of producers which includes prosumers equipped with DERs 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 consumers without the power generation capabilities or prosumers with a higher consumption compared to their self-production.

Day-ahead energy market modality is adopted with energy exchanges performed on a daily basis. For each producer $i \in P$, the system estimates 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 the daily production and demand capacity can be accurately predicted with time-series analysis techniques, such as exponential moving average and other machine learning techniques like Long Short Term Memory (LSTM) [35, 128]. The users, and specifically consumers, are considered to have an active role in the exchange process. Specifically, the system is envisioned to send a daily personalized recommendation to each consumer through a smartphone app.

Table 4.1: 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}
\hat{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

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., [35, 59], 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 system, 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 [129] and the Student t -test [130], 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 [28]. Specifically, humans possess limited information, time, and cognitive capabilities which prevent them to act optimally. These aspects of human behavior are modeled in this problem by limiting the size of the recommendation list to a maximum length K . This reduced size of recommendation list prevents the users from being overwhelmed 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 in electric lines during the transfer [35]. 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 system 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 system 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.

4.2 Problem Formulation

The goal of the proposed system is to find the recommendations to be sent to the consumers so that the expected energy exchanged is maximized. Therefore, this requirement is moulded into the optimization problem as presented in Eqs. (4.1)-(4.1f). Table 4.1 summarizes the notations used throughout this chapter. 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 objective function in Eq. (4.1) aims at maximizing 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 .

$$\text{maximize} \quad \sum_{i \in P} \sum_{j \in C} w_j p_{ij} x_{ij} \quad (4.1)$$

$$\text{s.t.} \quad \sum_{j \in C} (1 + L_{ij}) w_j x_{ij} \leq r_i, \quad \forall i \quad (4.1a)$$

$$\sum_{i \in P} x_{ij} \leq 1, \quad \forall j \quad (4.1b)$$

$$\sum_{i \in P} z_{ij} \leq K, \quad \forall j \quad (4.1c)$$

$$\alpha z_{ij} \leq w_j x_{ij} \leq w_j z_{ij}, \quad \forall i, j \quad (4.1d)$$

$$z_{ij} \geq x_{ij}, \quad \forall i, j \quad (4.1e)$$

$$x_{ij} \in [0, 1], \quad z_{ij} \in \{0, 1\}, \quad \forall i, j \quad (4.1f)$$

The constraint in Eq. (4.1a) guarantees that the production capacity of producer i is not exceeded, considering the loss that is incurred in the transmission. Similarly, constraint (4.1b) ensures that the demand of consumer j is not exceeded. The variables z_{ij} are used in the constraint (4.1c) to make sure that the recommendation list is of maximum length K . Finally, Eq. (4.1d) certifies that an exchange is larger than the minimum exchangeable allowed amount α , and Eqs. (4.1e)-(4.1f) 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. (4.1) is NP-Hard.*

Proof. In general instance of GAP [131], 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 contains 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, unlike discrete in GAP. 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. (4.1d) 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. \square

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 [35]. Conversely, learning the user behavior is challenging, as users may significantly differ in their preferences and engagement with the system [111, 112]. For these reasons, a Reinforcement Learning (RL) approach, called User Preference Learning (UPL) is proposed in section 4.3 to learn user preferences, inspired by [132]. 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. (4.1) to guarantee the bounded regret. However, as a workaround for NP-Hard UPL algorithm, two polynomial time algorithms are proposed for both phases of UPL in section 4.4. 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.

4.3 A Reinforcement Learning Approach for User Preference Learning

The optimization problem in Eq. (4.1) 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 system 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 [27]. 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. (4.1), allow to formulate this problem through the framework of combinatorial multi-armed bandit [132]. 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 [133]. However, this approach ignores the inherent dependencies among the arms, and therefore, ends up

learning the information about the observed actions independently [132]. 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, we extend the approach proposed in [132], to the problem of finding the best matching between consumers and producers while simultaneously learning the user 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 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 the corresponding arm i and player j . 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, 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 system 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.

At each iteration of the optimization phase d , the system chooses the action matrix

Algorithm 1: User Preference Learning (UPL)

```

/* Initialization Phase */
1 for each  $i \in P$  and each  $j \in C$  do
2   | Select any  $\mathbf{A} \in \mathcal{F}$  s.t.  $a_{ij} > 0$ ;
3   | Update  $[\hat{p}_{ij}]_{|P| \times |C|}$  and  $[m_{ij}]_{|P| \times |C|}$  according to Eqs. (4.5) and (4.6);
4 end
/* Optimization Phase */
5 while True do
6   |  $d = d + 1$ ;
7   | Select an action  $\mathbf{A}$  s.t.  $\mathbf{A}(d) = \arg \max_{\mathbf{A} \in \mathcal{F}} \sum_{i \in P} \sum_{j \in C} w_j a_{ij} \left( \hat{p}_{ij} + \sqrt{\frac{(Q+1) \ln d}{m_{ij}}} \right)$ ;
8   | Update  $[\hat{p}_{ij}]_{|P| \times |C|}$  and  $[m_{ij}]_{|P| \times |C|}$  according to Eqs. (4.5) and (4.6);
9 end

```

$\mathbf{A}(d)$ that maximizes the objective function given the current knowledge. This knowledge is represented by the estimated expected $\hat{p}_{ij}(d)$ for each random variable P_{ij} . For an action matrix $\mathbf{A}(d)$, the reward is defined as

$$\mathbf{R}_{\mathbf{A}(d)}(d) = \sum_{i,j} w_j a_{ij}(d) P_{ij}(d). \quad (4.2)$$

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

$$\mathcal{R}(d) = d\mathbf{R}_{\mathbf{A}(d)}^*(d) - \mathbb{E}\left[\sum_{t'=1}^d \mathbf{R}_{\mathbf{A}(t')}(t')\right], \quad (4.3)$$

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 [132]. 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. (4.1a)-(4.1f), and the following objective function:

$$\mathbf{A}(d) = \arg \max_{\mathbf{A} \in \mathcal{F}} \sum_{i \in P} \sum_{j \in C} w_j a_{ij} \left(\hat{p}_{ij} + \sqrt{\frac{(Q+1) \ln d}{m_{ij}}} \right), \quad (4.4)$$

The optimization problem solved at day d is based on the estimation of the expected values p_{ij} at day $(d-1)$, denoted as $\hat{p}_{ij}(d-1)$. 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 $\hat{p}_{ij}(d)$, as well as the total number $m_{ij}(d)$ of observations of the variable P_{ij} , as follows:

$$\hat{p}_{ij}(d) = \begin{cases} \frac{\hat{p}_{ij}(d-1)m_{ij}(d-1) + P_{ij}(d)}{m_{ij}(d-1) + 1} & \text{if } a_{ij}(d) \neq 0, \\ \hat{p}_{ij}(d-1) & \text{otherwise.} \end{cases} \quad (4.5)$$

$$m_{ij}(d) = \begin{cases} m_{ij}(d-1) + 1 & \text{if } a_{ij}(d) \neq 0, \\ m_{ij}(d-1) & \text{otherwise.} \end{cases} \quad (4.6)$$

Theorem 2. *Let w_j be homogeneous across users for sufficient amount of time, UPL*

provides bounded regret given by:

$$\mathcal{R}(d) \leq \left[\frac{4a_{max}^2 Q^3 (Q+1) \ln(d)}{(\Delta_{min})^2} + \frac{\pi^2}{3} Q^2 + Q \right] \Delta_{max}, \quad (4.7)$$

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 [132]. \square

4.4 A Constrained Maximum Weighted Matching- based Reinforcement Learning Approach

Faster Initialization Algorithm (FIA) is presented in this section as an improvement to the initialization phase of UPL. Subsequently, a heuristic algorithm BPT- K is proposed for the optimization phase. The initialization phase of UPL, similar to the one originally presented in [132], 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. (4.1), 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.

4.4.1 Faster Initialization Algorithm (FIA)

The Faster Initialization Algorithm (FIA) is presented in Alg. 2. 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. Secondly, 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  $(\sum_{i \in P} a_{ij} < 1)$  and  $(\exists i \mid (b_{ij} = 0 \text{ and } (r_i \geq e)))$  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. (4.5) and (4.6). The *while* loop terminates as soon as all variables are observed, and then the optimization phase begins.

4.4.2 The BiParTite-K Algorithm

Overview

The problem introduced in Eq. (4.1) is an extension of the generalized matching problem (see Theorem 1), with the additional constraint that consumer-nodes' degrees cannot exceed K (see Eq. (4.1c)). Recall that such K -constraint is a practical requirement for bounded rationality to prevent overwhelming users with a large list of recommendations [28].

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 [134] or Edmond's Algorithm [135, 136]. 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, viz. (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. (4.5) and (4.6). 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 4.5, 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. (4.1e)).

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

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  $L_{ij} \leq L_{max}$  and  $\left( (|(\cdot, j)|_{E_{\Phi_{out}}} < K) \text{ or } (|(\cdot, j)|_{E_{\Phi_{out}}} = K \text{ and } (i, j) \in E_{\Phi_{out}}) \right)$ 
12      then
13        Add edge  $(u, v)$  to  $E_G$  with weight,  $\mathcal{W}_G(u, v) = \left( T * \left( \hat{p}_{ij} + \sqrt{\frac{(Q+1) \ln d}{m_{ij}}} \right) \right)$ ;
14      end
15  end
16  Perform Maximum Weighted Bipartite Matching on  $G$  and output graph
   $\Phi_G = \{P_d \cup C_d, E_{\Phi_G}\}$ , where  $E_{\Phi_G} \subseteq E_G$ ;
17   $\Phi_{temp} = \Phi_{out}$ ;
  /* Add/update the edge in  $\Phi_{out}$  from  $\Phi_G$  without violating the
   $K$ -constraint */
18  Sort edges in  $E_{\Phi_G}$  by decreasing weight;
19  while  $E_{\Phi_G} \neq \emptyset$  do
20    Consider next edge  $((u, v) \in E_{\Phi_G} \text{ s.t. } u \in P_i, v \in C_j)$ ;
21    if  $(i, j) \in E_{\Phi_{out}}$  then update the edge weight,
       $\mathcal{W}_{\Phi_{out}}(i, j) = \mathcal{W}_{\Phi_{out}}(i, j) + \sum_{\substack{u \in P_i \\ v \in C_j}} \mathcal{W}_G(u, v)$ ;
22    else if  $(|(\cdot, j)|_{E_{\Phi_{out}}} + 1 \leq K)$  then add edge  $(i, j)$  to  $E_{\Phi_{out}}$  with weight,
       $\mathcal{W}_{\Phi_{out}}(i, j) = \sum_{\substack{u \in P_i \\ v \in C_j}} \mathcal{W}_G(u, v)$ ;
23    Remove  $(u, v)$  from  $E_{\Phi_G}$ ;
24  end
25  while  $\Phi_{out} \neq \Phi_{temp}$ ;
26  Produce a recommendation list from  $\Phi_{out}$  and send them to respective consumers;
27  Observe the performed exchanges and update  $[\hat{p}_{ij}]_{|P| \times |C|}$  and  $[m_{ij}]_{|P| \times |C|}$ ;

```

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 $|(\cdot, j)|_{E_{\Phi_{out}}}$.

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. (4.1).*

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. (4.1). 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. (4.1b) and (4.1c). Moreover, by setting the size of the unit of exchangeable energy $T > \alpha$, constraint (4.1e) is also satisfied. Finally, the K -constraint in Eq. (4.1d), 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. \square

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 $|(\cdot, j)|_{E_{\Phi_{out}}} + 1 \leq K$. 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 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. \square

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. \square

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

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 $O(\min\{|P_d|, |C_d|\})$ and the number of edges that can be added is bounded by $O(K|C|)$. 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 $O(|P_d||C_d|)$. The maximum weighted matching can be solved with the Edmond’s algorithm with complexity $O((|P_d| + |C_d|)^3)$ [135, 136]. The cardinality of E_{Φ_G} is upper bounded by $O(|P_d||C_d|)$, therefore sorting the edges has complexity $O(|P_d||C_d| \log(|P_d||C_d|))$, and the *while* loop has a number of iterations upper bounded by $O(|P_d||C_d|)$. Since the maximum weighted matching algorithm dominates the operations within the *do – while* loop, and $K|C|$ is generally less than $\min\{|P_d|, |C_d|\}$, the overall complexity of BPT- K is $O(\min\{|P_d|, |C_d|\} \times (|P_d| + |C_d|)^3)$. \square

4.5 Experimental Results

In this section, performance of the proposed approaches is evaluated versus a state-of-the-art approach, named Zhu, proposed in [35]. 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 FIA algorithm is investigated and also a sensitivity analysis to relevant parameters of BPT- K is provided.

4.5.1 Experimental Setup

Realistic datasets for energy production and consumption are used for experiment. Real consumption dataset is obtained from [137] 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 [138] 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 [60, 139]. 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 $\alpha = 50\text{Wh}$. UPL and BPT- K implement Gurobi optimizer [140] and NetworkX python library respectively.

4.5.2 Comparison approach

The proposed algorithms, UPL and BPT- K , are compared to the “Zhu” algorithm presented in [35]. 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 [35] for more details. To the best of our knowledge, [35] is the closest work in context of the

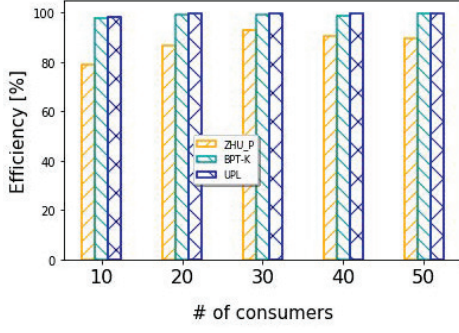
proposed system which aims at finding an optimal matching among the producers and consumers in a localized energy sharing system.

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. 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.

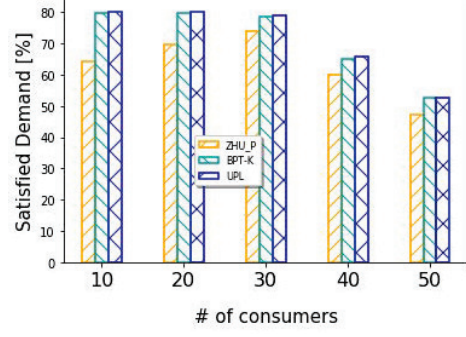
4.5.3 Performance Evaluation

Four experimental scenarios are considered for performance evaluation of the proposed P2P energy sharing system. 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

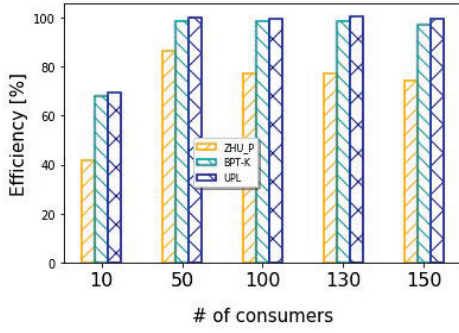


(a) Efficiency

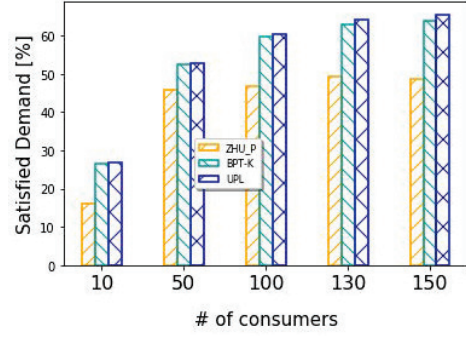


(b) Satisfied Demand

Figure 4.2: Efficiency and Satisfied Demand vs. number of consumers (keeping number of producers constant)



(a) Efficiency



(b) Satisfied Demand

Figure 4.3: Efficiency and Satisfied Demand vs. number of consumers (keeping ratio of consumers-to-producers constant)

optimization problem in Eq. (4.1) 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 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. 4.2a. 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. 4.2b 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 dataset used consists of only 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. 4.3a and 4.3b. 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. 4.3b. 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.

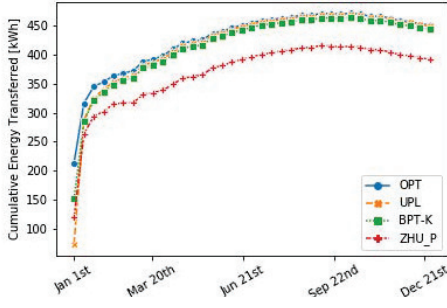


Figure 4.4: Plot for cumulative reward (energy exchange)

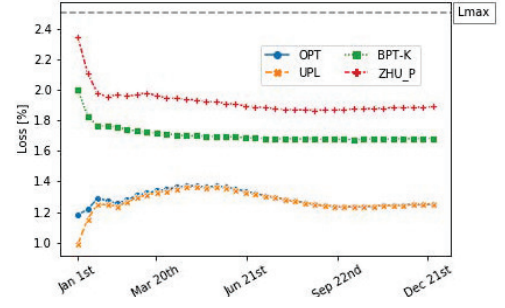


Figure 4.5: Percentage of energy losses over time.

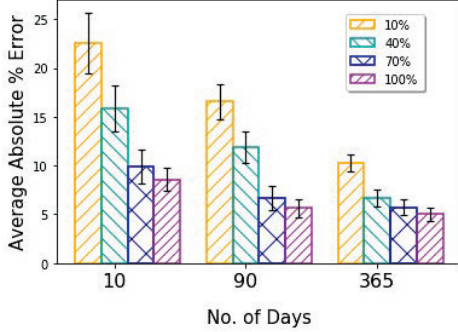


Figure 4.6: Average absolute % error of preferences learned over time

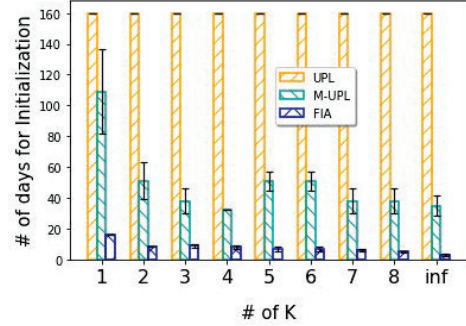


Figure 4.7: Number of Days required for initialization vs. K

As the results presented in Fig. 4.4 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

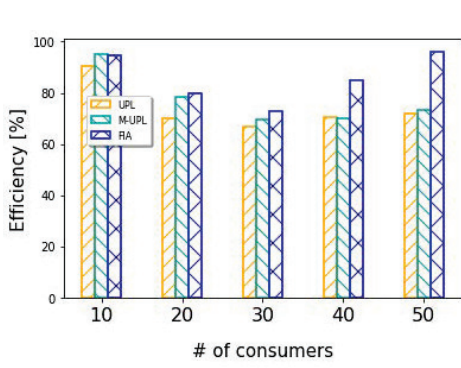
optimization problem optimally at each iteration. Once the preferences are sufficiently learned, UPL and OPT provide the same solution. BPT- K shows a reward that closely matches the performance of UPL, without requiring the solution of a NP-Hard problem at each time step. On the contrary, Zhu_P clearly 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 [138], there is a seasonal effect causing the non-monotonic trend of all approaches in Fig. 4.4. 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. 4.5 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. 4.6 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

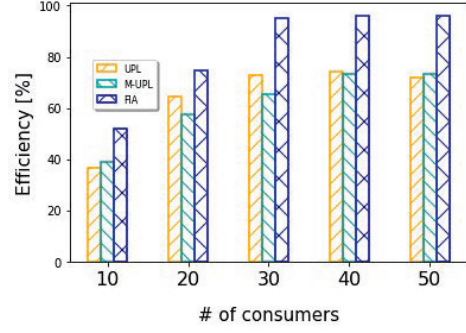
knowledge is acquired. In fact, once the best matches (i.e., those with higher chances of acceptance) are identified, these are selected more often, in order to maximize the system performance. As a result, other consumers' preferences are not learned exactly but this does not prevent the system from 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 [132]. 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. 4.7, 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



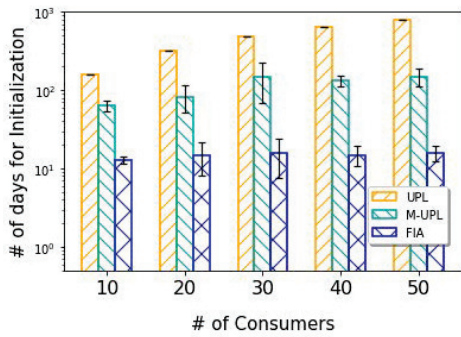
(a) Constant number of producers



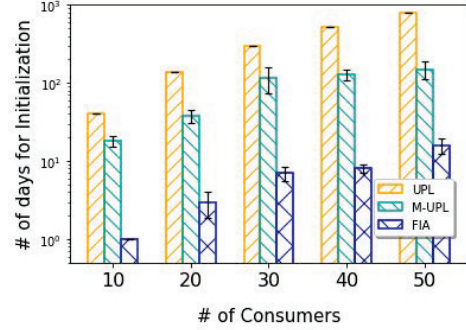
(b) Constant ratio of no. of consumers-to-producers

Figure 4.8: Efficiency of the initialization algorithms vs. number of consumers

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. 4.9a and 4.9b 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.



(a) Constant number of producers



(b) Constant ratio of no. of consumer-to-producer.

Figure 4.9: Number of days required for initialization vs. 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. 4.8a and 4.8b. As the figures show FIA, even not targeting energy exchange as primary objective, significantly outperforms the original UPL and M-UPL.

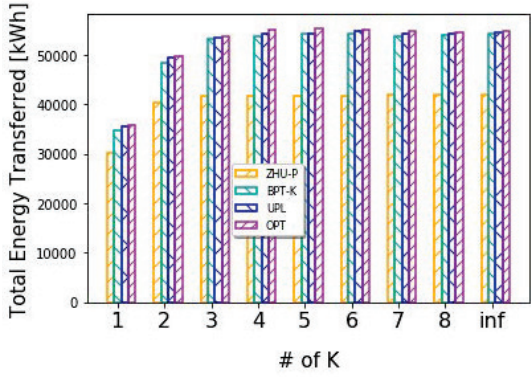


Figure 4.10: Energy Transferred vs. K .

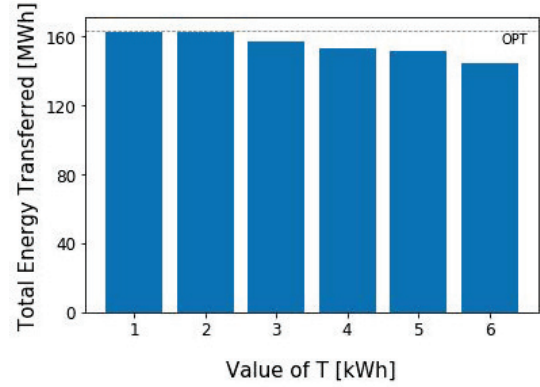


Figure 4.11: Energy Transferred vs. T .

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. 4.10 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 [28]. 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. 4.11. 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 $T = 1\text{kWh}$, 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.

4.6 Concluding Remarks

In this chapter, the problem of exchanging locally-generated energy through a P2P energy sharing mechanism was studied. Formulated as a Mixed-Integer Linear Programming (MILP), the problem was proved to be NP-Hard. Unlike the existing works that mostly overlook or oversimplify the role of human behavior in P2P energy exchange, 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. In addition to that, 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.

CHAPTER 5. PROSPECT THEORY-INSPIRED P2P ENERGY TRADING

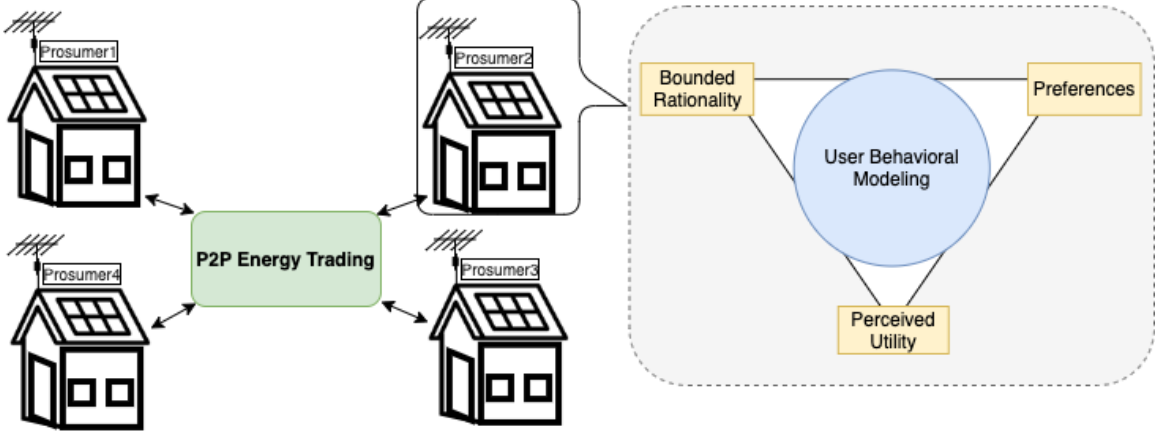


Figure 5.1: Incorporating User Behavioral Modeling into P2P Energy Trading

Like in any trading market, the monetary incentives form major motivating factor for ensuring sustainable participation of prosumers in the P2P energy trading market. However, it requires the prosumers to be actively involved in the trading process which might cause prosumers to be overwhelmed with the decision-making problem owing to the *bounded rationality*. As established in previous chapters, requiring constant active participation from prosumers might even lead to their possible abandonment and therefore, failure of implementation of such systems in real setting. To this end, the prosumers' decision-making behavior and perceived utility must be effectively incorporated in the P2P energy trading system to create a prosumer-centric trading environment that requires little active involvement from their side. This can be achieved through user behavioral modeling that integrates bounded rationality, individual preferences and perceived utility of users together as shown in the Fig. 5.1 to duplicate their decision-making as accurately as possible. In this chapter, we devise mechanisms to automate the process of P2P energy trading that incorporates notion of Prospect Theory to maximize the user's perceived utility for trading energy as well as utilize reinforcement learning frameworks to dynamically update the seller's

selling price based on the previous-day sell. The work conducted in this chapter was published in Proceedings of IEEE Global Communication Conference (IEEE GLOBECOM) 2022 [47] and in ACM Transactions on Evolutionary Learning and Optimization 2022 [48].

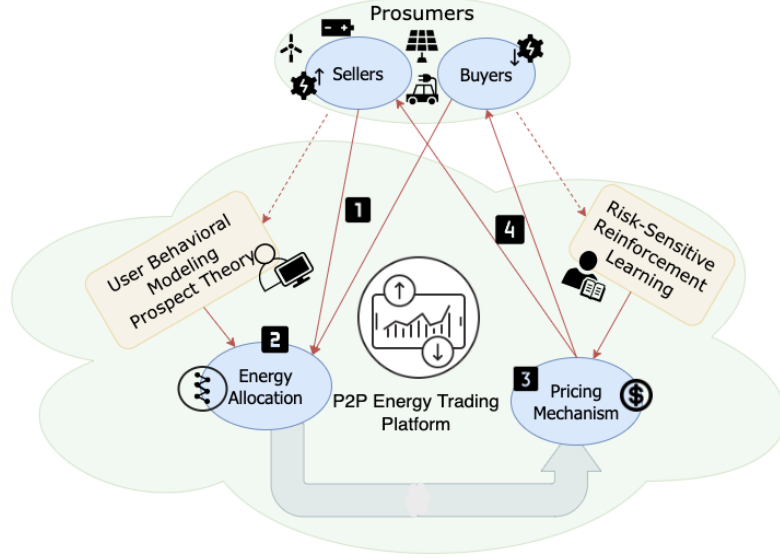


Figure 5.2: P2P Energy Trading System Overview.

Therefore, to model the user behavioral patterns and perceptions in our P2P energy trading system, we turn to behavioral economics to formulate a *prospect theory*-inspired optimization problem that maximizes the perceived value of energy trading transactions for individual prosumers in automated environment. This automated energy trading environment needs to be designed carefully to integrate prosumer’s energy trading behavior while also maximize their financial incentives for being involved with the market. It will not only save the user’s involvement time but will also help in maximizing their perceived benefits in terms of loss/gains through the use of user behavioral modeling.

To this purpose, we utilize the widely accepted notion in behavioral decision-making called *Prospect Theory* (PT) [69]. PT captures the non-rational decision making of humans in the face of uncertainty, and it provides a mathematical tool to

quantify the subjective perception loss and gain. Specifically, we propose a PT-based optimization framework for prosumer-centric P2P energy trading that incorporates perceived utility into the trading and automates the price updating for sellers using reinforcement learning. This is presented clearly in Fig. 5.2. The proposed framework aims at matching energy demand and production between buyers and sellers (step 1 in Fig. 5.2). The objective is to maximize the perceived utility of individual buyers, by taking into account their intrinsic perception and heterogeneity. We formalize this as a non-linear and non-convex optimization problem, and prove that it is NP-hard. Given the non-linear and non-convex nature of the problem on top of being NP-hard, we further propose a Differential Evolution-based Algorithm for Trading Energy (*DEbATE*) to find a solution to the problem in polynomial time (*energy allocation*, step 2).

In order to require minimal participation of prosumers, we employ a Reinforcement Learning (RL) framework, called Pricing mechanism with Q-learning and Risk-sensitivity (*PQR*), which is executed in tandem with *DEbATE*, to automate the pricing mechanism for sellers (*pricing mechanism*, step 3). Sellers are not aware of their competitiveness in the market. Therefore, *PQR* adjusts the price dynamically based on the market demand as well as seller's competitiveness and perceived utility. *PQR* learns the optimal selling price for each sellers using PT-based risk-sensitive RL approach [117]. However, *PQR* inherits the typical scalability and stability limitations of a standard tabular approach for the Q-learning function. To avoid such limitations, we further improve *PQR* by proposing a Deep Reinforcement Learning based alternative heuristic, called *ProDQN*, that uses a PT-based loss function to accommodate the seller's perceived utility. The output of the RL algorithms are then published to the prosumers for the next matching of demand and production. Finally, the output of the matching is then implemented by the P2P energy system to execute the physical energy transactions (step 4).

The proposed techniques address the limitations of previous works by modeling individual prosumers' behavior, incorporating perceived utility, and automating the price updating process for sellers. Employing a differential evolution-based heuristic, paired with reinforcement learning based pricing mechanisms, allows to efficiently find a solution to the non-linear and non-convex problem, which is especially critical for large systems with many prosumers. Additionally, by incorporating PT-based approaches, the individual subjective perception loss and gain can be quantified, which is an essential aspect of prosumer-centric P2P energy trading. Finally, through the use of reinforcement learning, the system can learn from the prosumers' behavior and adapt to changes in market conditions, leading to a more efficient and effective P2P energy trading system.

Major works proposed in this section are:

- Develop a prospect theory inspired optimization framework for P2P energy trading between prosumers.
- Devise an energy allocation mechanism based on the optimization framework that maximizes the perceived value of loss and gain for all buyers
- Take seller's individualized risk attitudes and sensitivity in account to design a dynamic pricing mechanism through risk-sensitive reinforcement learning.

5.1 System Model and Problem Formulation

The components of the proposed framework for P2P energy trading system are shown in Fig. 5.3. The systems consists of three distinct components, namely (i) Prosumers, (ii) Energy allocation, and (iii) Pricing mechanism. We describe the modeling of the three components in detail in following subsections.

5.1.1 Modeling Energy Allocation

In this subsection, we present the energy allocation mechanism which determines how to match the buyers' demand with the sellers' production, while maximizing

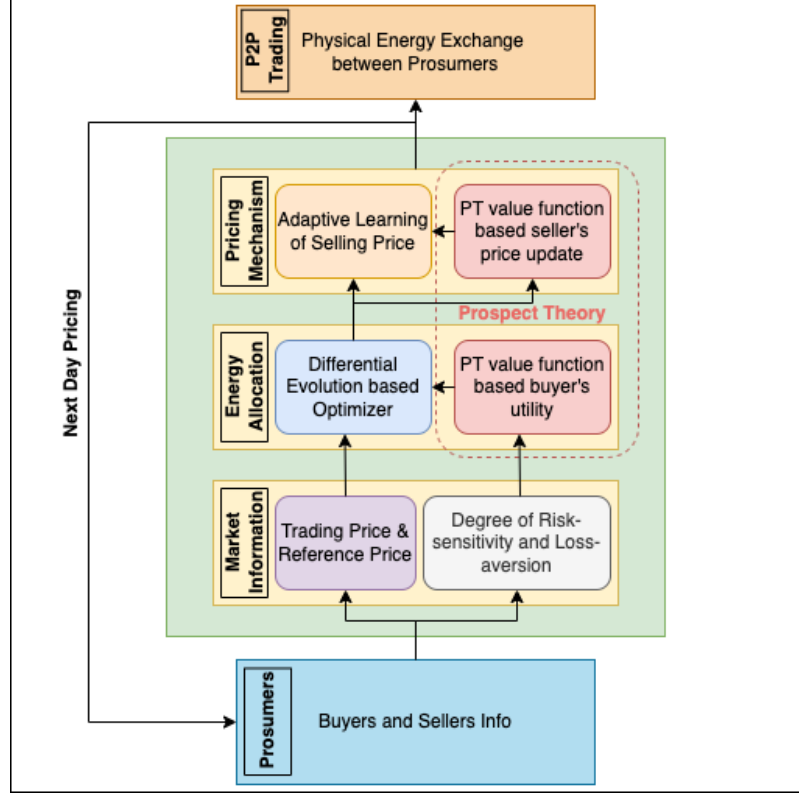


Figure 5.3: Proposed Framework of P2P Energy Trading.

their individual perception. We model the perceived loss and gain of prosumers using the *prospect theory* (PT) value function to capture user perception of gains and losses as shown in Fig. 5.3. Specifically, consider the excess energy generation of seller $i \in S$ be r_i and demand of buyer $j \in B$ be w_j . Then, let $x_{ij} \in [0, 1]$ be a variable representing the fraction of w_j that buyer j buys from seller i . Also, let ρ_{gs}, ρ_{gb} be the energy selling and purchasing prices from the grid, respectively, ρ_i be the selling price of seller i , and ρ_j is the reference price of buyer j . Prices are expressed as cost per kWh .

We adopt a modified PT value function to model realistic user perception in an energy market [69]. The function quantifies the human perceived utility towards gain and loss based on the degree of deviation from a reference point. In our problem, it captures the difference between the total actual buying cost for the buyer j i.e. y_j and their desired reference cost $\rho_j w_j$ for buying w_j amount of energy at their reference

price. The utility function is then formulated as

$$v(y_j) = \begin{cases} k_{+,j}(\rho_j w_j - y_j)^{\zeta_{+,j}}, & y_j < \rho_j w_j \\ -k_{-,j}(y_j - \rho_j w_j)^{\zeta_{-,j}}, & y_j \geq \rho_j w_j \end{cases} \quad (5.1)$$

where $k_{+,j}$, $k_{-,j}$, $\zeta_{+,j}$, $\zeta_{-,j}$ are the parameters that control the degree of loss-aversion and risk-sensitivity. These parameters can be obtained through either adaptive learning using models presented in chapter 4. Similarly to [113, 141, 142], we assume that these parameters can be obtained by surveys completed by the prosumers when the energy trading system is installed in their home, and potentially updated later on with sporadic feedback to the energy management system. Recent studies have shown that these parameters are highly heterogeneous and vary from person to person based on factors like gender and age group [141, 142]. In the equation above, y_j is the total actual cost of buying energy for buyer j that incorporates the total cost of buying energy from P2P setting as well as the grid – in case the demands are not met locally. Therefore the term y_j is given by

$$y_j = \sum_{i \in S} \rho_i x_{ij} w_j + \rho_{gs} (1 - \sum_{i \in S} x_{ij}) w_j$$

Note that, similar to the PT value function in [69], the utility function in Eq. (5.1) is concave in the gain domain (i.e., $y_j < \rho_j w_j$) while convex in loss domain (i.e., $y_j \geq \rho_j w_j$).

The problem of matching demand and production of heterogeneous prosumers is

formalized as follows.

$$\text{maximize} \quad f(y) : \sum_{j \in B} v(y_j) \quad (5.2)$$

$$\text{s.t.} \quad \sum_{j \in B} (1 + l_{ij}) x_{ij} w_j \leq r_i, \quad \forall i \quad (5.2a)$$

$$\sum_{i \in S} x_{ij} \leq 1, \quad \forall j \quad (5.2b)$$

$$x_{ij} = 0, \text{ if } l_{ij} \geq l_{max}, \quad \forall i, j \quad (5.2c)$$

$$\rho_{gb} \leq \rho_i, \rho_j \leq \rho_{gs}, \quad \forall i \quad (5.2d)$$

$$\mu_j z_{ij} \leq x_{ij} w_j \leq w_j z_{ij}, \quad \forall i, j \quad (5.2e)$$

$$z_{ij} \geq x_{ij}, \quad \forall i, j \quad (5.2f)$$

$$x_{ij} \in [0, 1], \quad z_{ij} \in \{0, 1\}, \quad \forall i, j \quad (5.2g)$$

The problem maximizes the sum of perceived utility for buyers in Eq. (5.2). There is an energy loss during the physical energy transfer through wires [35], which depends on the wire-length between i and j and it is directly proportional to the amount of energy exchanged. We model such loss as a fraction $l_{ij} \in [0, 1]$ of the energy exchanged. As a result, the constraint in Eq. (5.2a) prevents the problem from exceeding the amount of energy being sold by each sellers while incorporating the losses in electric lines. The constraint in Eq. (5.2b) ensures that the energy demand for each buyer is not exceeded, while constraint (5.2c) limits the loss between sellers and buyers to be within the loss threshold l_{max} . On the other hand, constraint (5.2d) limits the upper and lower bound for energy price to the selling and buying price of the grid. Constraint (5.2e) sets the minimum amount μ_j of an energy transaction for buyer j , using a binary decision variable z_{ij} , that is equal to 1 if $x_{ij} > 0$, and to 0 otherwise (constraint (5.2f)). The value of μ_j is generally a system parameter to prevent impractical solutions containing infinitesimal amounts [45]. Finally, the

range of operation of the decision variables are defined in (5.2g).

Theorem 5. *The optimization problem in Eq. (5.2) is NP-hard.*

Proof. We present a reduction from the Generalized Assignment Problem (GAP) [131] as a proof of NP-hardness of our optimization problem. In a general instance of GAP, there are n tasks and m processors. Each processor i has a resource budget given by r_i . By assigning task j to processor i , we obtain a profit p_{ij} while consuming g_{ij} amount of resources. A task can only be assigned to a single process, and therefore, the goal is to find the assignment that provides maximum profit given the resource budget of the processors. The GAP can be formulated as an integer linear programming problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^m \sum_{j=1}^n p_{ij} x_{ij} \end{aligned} \tag{5.3}$$

$$\text{s.t.} \quad \sum_{j=1}^n g_{ij} x_{ij} \leq r_i, \quad \forall i \tag{5.3a}$$

$$\sum_{i=1}^m x_{ij} = 1, \quad \forall j \tag{5.3b}$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j \tag{5.3c}$$

From a general GAP instance, we can create a reduced instance of our problem as follows. We create a buyer for each task and a seller for each processor. We set $(1+L_{ij})w_j = g_{ij}$ and set the energy production of a seller i to r_i . We also set $l_{max} = \infty$ so that all exchanges are possible (i.e., a task can be assigned to any processor). An important difference between our reduced problem and the GAP is that the decision variables x_{ij} are continuous instead of discrete. However, infinitesimal exchanges are not allowed in our system, as they need to be greater than or equal to μ_j . By setting $\mu_j = w_j$, the constraint in Eq. (5.2e) forces the decision variable x_{ij} to be either 0 or 1, same as binary decision variable z_{ij} . Additionally, it also forces the system to assign a buyer (i.e., a task in the GAP problem) to a single seller.

We set all the loss-aversion parameters $(k_{+,}, k_{-,})$ to 1 and the risk-sensitive parameters $(\zeta_{+,}, \zeta_{-,})$ to 1. We also set $\rho_j = 0$ for all sellers in S . In summary, the objective function becomes linear, i.e.,

$$\sum_{j \in B} \sum_{i \in S} (\rho_{gs} - \rho_i) w_j x_{ij} - \sum_{j \in B} \sum_{i \in S} \rho_{gs} w_j$$

The term $-\sum_{j \in B} \sum_{i \in S} \rho_{gs} w_j$ is just a constant, and can thus be ignored for the purpose of the maximization problem. Now, by setting $(\rho_{gs} - \rho_i) w_j = p_{ij}$, we have successfully reduced the objective function to

$$\sum_{j \in B} \sum_{i \in S} p_{ij} x_{ij}$$

As a result, the solution of our reduced problem provides the assignment that maximizes the profit within the constrained processors' resources. Therefore, our problem is at least as hard as GAP, and thus it is NP-Hard. \square

It is to be noted that, in addition to the NP-Hardness, the problem in Eq. (5.2) is also non-linear and non-convex. There is not any general procedure to solve such optimization problems dealing with continuous solution sets [143]. Hence, in order to solve this combinatorial problem of matching demand and supply of energy, we propose a heuristic based on Differential Evolution [144] which finds a feasible solution through iterative recombination and improvement of the candidate solutions along with constraint handling. Adopting this heuristic approach is particularly beneficial in large systems, where the complexity of the problem would make it impossible to find the optimal solution in reasonable time. In the next section, we motivate the need for a dynamic pricing mechanism, before presenting the differential evolution heuristic in Section 5.2.

5.1.2 Modeling Pricing Mechanism

In the proposed P2P energy trading model, the selling price is considered to be a fixed amount for a given trading period, and it is used as the trading price for a transaction. However, the reference price ρ_i of seller i is a personal value which may under- or over-estimate the competitiveness of market. In order to improve the sellers' competitiveness, we implement a dynamic pricing model for sellers as exhibited in the Fig. 5.3. Note that, to expect sellers to manually adjust the price based on the performance of the energy trading system (e.g., their revenue) is impractical. Demanding such active participation could easily be overwhelming, and severely affect their performance and level of engagement with the system. To avoid such active participation, we model the adjustment of the price as a Markov Decision Process, and exploit reinforcement learning to update the selling price at each trading period.

To maximize the sellers' perceived objectives through prospect theory, we resort to a risk-sensitive reinforcement learning approaches that forms the basis of the automated pricing mechanism within our P2P energy trading framework. In the following section 5.2, we present two algorithms based on reinforcement learning that incorporate the seller's perceptions on loss and gains to update the prices while automating the process in order to ensure sustained prosumer participation. First, in subsection subsection 5.2.2, we employ risk-sensitive Q-learning algorithm [117] and then given its efficiency limitation due to the tabular representation of the Q-function, we also present a Deep Q-network (DQN) [145] based algorithm in subsection 5.2.3 that proposes a novel loss function founded on prospect theory value function.

5.2 Solution Approaches and Heuristics

In this section, we describe the *Differential Evolution-based Algorithm for Trading Energy (DEbATE)* heuristic (Alg. 4), designed for matching demand and production according to the problem presented in Section 5.1, the *Pricing mechanism with Q-learning and Risk-sensitivity (PQR)*, and the *Prospect theory-based Deep Q-Network*

(*ProDQN*), designed to dynamically adjust the sellers' prices.

5.2.1 DEbATE

Algorithm 4: DEbATE

Input : set of buyers B , sellers S , fitness function $f(\cdot)$, max iterations G_{max} , population size NP , crossover probability CR , differential weight F

Output: best identified feasible solution \mathbf{x}^*

- 1 Update set of buyers B and sellers S , $count = 0$;
- 2 Generate initial population $\mathcal{X} = \{\mathbf{x}_k | k = 1, \dots, NP\}$;
- 3 **while** $count < G_{max}$ **do**
- 4 **for** each $\mathbf{x}_k \in \mathcal{X}$ **do**
- 5 Choose 3 different vectors $\{\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c\} \in \mathcal{X}$ at random and $R \sim U(1, |S| \times |B|)$;
- 6 Create mutated solution $\bar{\mathbf{x}}_k = \mathbf{x}_k$;
- 7 /* Mutation and Crossover */
- 8 **for** each $i \in |S|, j \in |B|$ **do**
- 9 Select $u \sim U(0, 1)$;
- 10 **if** $u < CR || (i \times j) == R$ **then**
- 11 $\bar{x}_{ij}^{(k)} = x_{ij}^{(a)} + F \times (x_{ij}^{(b)} - x_{ij}^{(c)})$;
- 12 $\bar{x}_{ij}^{(k)} = \min(1, \max(0, \bar{x}_{ij}^{(k)}))$;
- 13 **end**
- 14 /* Check Constraints */
- 15 $\forall i, j$, **if** $l_{ij} \geq l_{max}$ **then** $\bar{x}_{ij} = 0$;
- 16 $\forall i$, **if** $\sum_j (1 + l_{ij}) \bar{x}_{ij} w_j > r_i$ **then** $\bar{x}_{ij} = \frac{\bar{x}_{ij} r_i}{\sum_j (1 + l_{ij}) \bar{x}_{ij} w_j}$;
- 17 $\forall j$, **if** $\sum_i \bar{x}_{ij} > 1$ **then** $\bar{x}_{ij} = \frac{\bar{x}_{ij}}{\sum_i \bar{x}_{ij}}$;
- 18 /* Compare fitness */
- 19 **if** $f(\bar{\mathbf{x}}_k) > f(\mathbf{x}_k)$ **then** $\mathcal{X} = (\mathcal{X} \setminus \{\mathbf{x}_k\}) \cup \{\bar{\mathbf{x}}_k\}$;
- 20 **end**
- 21 $count = count++$;
- 22 **end**
- 23 /* Find the best solution and execute trading */
- 24 Let $\mathbf{x}^* = \arg \max_{\mathbf{x}_k \in \mathcal{X}} f(\mathbf{x}_k)$;
- 25 Execute transactions for each prosumers to \mathbf{x}^* ;

DEbATE is executed at each trading period (e.g., 12 hours) to solve the non-linear optimization problem presented in Section 5.1. It uses differential evolution to determine an optimal amount of energy to be traded between prosumers that maximizes the perceived utility of buyers. *DEbATE* initially updates the list of

buyers (B) and sellers (S) based on the expected production and consumption for the current trading period. These can be predicted accurately with recent approaches [128, 146]. The differential evolution-based optimization begins on line 2 where an initial population \mathcal{X} is generated with population size of NP . An element $\mathbf{x}_k \in \mathcal{X}$, with $k = 1, 2, \dots, NP$ is a candidate solution vector of variables x_{ij} representing the amount of energy to be traded between the i^{th} seller and j^{th} buyer. These variables correspond to the decision variables of our optimization problem.

The *while*-loop (line 3–19) is the differential evolution loop that aims at finding a solution to the non-linear optimization problem with Eq. (5.2) as the fitness function. The loop is executed for G_{max} iterations. At each iteration, for each candidate solution $\mathbf{x}_k \in \mathcal{X}$, the algorithm creates a mutated solution $\bar{\mathbf{x}}_k$. Initially, $\bar{\mathbf{x}}_k = \mathbf{x}_k$. The mutated solution is subsequently updated through mutation and crossover with 3 random candidates $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c \in \mathcal{X}$ (line 5). A value $R \in [1, |S| \times |B|]$ is selected at random. R will be used in the following *for*-loop to ensure a minimum mutation. The for loop in line 7 iterates over the components (dimensions in evolutionary terms) of $\bar{\mathbf{x}}_k$. During each iteration, a value $u \in [0, 1]$ is sampled at random as mutation probability (line 8). Subsequently, a mutation occurs for the component ij of $\bar{\mathbf{x}}_k$ with crossover probability CR (line 9). The mutation occurs irrespective of the probability if $(i \times j) = R$ (to ensure at least one minimum mutation). A mutation is executed by combining the corresponding component of \mathbf{x}_a , \mathbf{x}_b , and \mathbf{x}_c with the differential weight parameter $F \in [0, 2]$ as in line 10. The mutated component $\bar{\mathbf{x}}_{ij}^{(k)}$ is clipped to ensure that it falls within $[0, 1]$ as minimum and maximum threshold to satisfy constraint Eq. (5.2g) in line 11 of the algorithm.

After the mutated solution is finalized, it is checked, and adjusted if needed, to meet the constraints in Eqs. (5.2a)-(5.2c) of the optimization problem. Specifically, line 13 ensures that no exchange occurs (i.e., $\bar{\mathbf{x}}_{ij}^{(k)} = 0$) between users having a loss higher than l_{max} . Lines 14–15 ensure that the production of a seller and the demand

of each buyer are not exceeded, respectively. Finally, in line 16, the fitness function $f(\cdot)$ of the mutated solution $\bar{\mathbf{x}}_{\mathbf{k}}$ is compared against the original candidate solution $\mathbf{x}_{\mathbf{k}}$. If $f(\bar{\mathbf{x}}_{\mathbf{k}}) > f(\mathbf{x}_{\mathbf{k}})$, then $\bar{\mathbf{x}}_{\mathbf{k}}$ replaces $\mathbf{x}_{\mathbf{k}}$ in the set of candidate solutions \mathcal{X} . At the end of the while loop, *DEbATE* selects the best solution \mathbf{x}^* in \mathcal{X} (line 20) and executes the transactions accordingly (line 21). In the following Lemma 3, we show that *DEbATE* has polynomial time complexity, and hence it is computationally efficient. The theorem focuses on the asymptotic complexity, a typical mathematical formulation to characterize the upper-bound of the running-time for sufficiently large inputs [147].

Lemma 3. *The time complexity of the DEbATE algorithm is $O(G_{max} \times NP \times |S||B|)$.*

Proof. The complexity is dominated by the *while* loop (lines 3–19), which is executed G_{max} times. Within this loop, the *for*–loop (lines 4 – 17) does $|\mathcal{X}| = NP$ total iterations. In each iteration, the inner *for*–loop (lines 7 – 12) iterates over the sets S and B , and only contains constant operations. Similarly, checking the constraints (lines 13 – 15) requires to iterate over the same sets. Finally, calculating the function $f(\cdot)$ (line 16) has cost $|B|$. Overall, the complexity is $O(G_{max} \times NP \times (|S||B| + 3|S||B| + |B|)) = O(G_{max} \times NP \times |S||B|)$ \square

5.2.2 PQR

Algorithm 5: PQR

```

/* Pricing with Risk-sensitive Q-learning */
1 Collect transaction information for each prosumers from DEbATE (Alg. 4)
  for current timestep  $t$ ;
2 for each  $i \in S$  do
3   Select an action,  $a \in \{+\delta, -\delta, 0\}$  based on the  $\epsilon$ -greedy strategy ;
4    $s = \rho_i$ ;  $s_{new} = s + a$ ;  $R_i = s_{new} \sum_{j \in B} x_{ij} w_j$ ;
5    $\rho_i = s_{new}$ ;
6   Update  $Q(s, a)$  as in Eq. (5.4) and (5.5);
7   Send information on updated price  $\rho_i$  to seller  $i$ ;
8 end

```

After determining the solution to the energy allocation problem in *DEbATE*, the selling price for sellers is then updated through the Pricing mechanism with Q-learning and Risk-sensitivity (*PQR*) algorithm as presented in Alg. 5. In order to learn the optimal selling price dynamically over time, we model the sellers as independent learning agents. Note that, to preserve the privacy and avoid the conflict between prosumers, these agents do not have access to information about other sellers or buyers. The state space (s) of the Markov Decision Process, in the Q-learning formulation, consists of the prices between the grid buying/selling ρ_{gb} and ρ_{gs} , discretized by a step size, δ , i.e., $\rho_i \in \{\rho_{gb}, \rho_{gb}+\delta, \rho_{gb}+2\delta, \dots, \rho_{gb}+(\frac{\rho_{gs}-\rho_{gb}}{\delta}-1)\delta, \rho_{gs}\}$.

The action space consists of a price increasing action, price decreasing action, and no change action, i.e. $a \in \{+\delta, -\delta, 0\}$ where δ is the amount by which price is increased or decreased. The agent goes to a new state after taking action a which is referred as s_{new} . Seller i reward function is the total revenue generated at the current trading period i.e. $R_i = (\rho_i + a) \sum_{j \in B} x_{ij} w_j$. For updating Q-values, we modify the approach proposed in [117] by considering the following Q-learning update rule that includes the PT-based perceived utility of sellers.

$$Q(s, a) \leftarrow Q(s, a) + \alpha v(y_i) \quad (5.4)$$

$$v(y_i) = \begin{cases} k_{+,i}(y_i)^{\zeta_{+,i}}, & y_i > 0 \\ -k_{-,i}(-y_i)^{\zeta_{-,i}}, & y_i \leq 0 \end{cases} \quad (5.5)$$

where, $y_i = R_i + \gamma \max_a Q(s_{new}, a) - Q(s, a)$ is the Temporal Difference (TD) error of i^{th} seller for current iteration, and $v(y_i)$ is transformation of TD error to capture each seller's personalized perceived utility on loss. α refers to the learning rate for updating Q-values in Eq. (5.4).

PQR, as summarized in Alg. 5, initially collects the current trading information

from *DEbATE* in line 1. The subsequent *for*-loop (lines 2 – 8) updates the selling price for each sellers. At each iteration, a seller $i \in S$ is considered. For that seller, the action (whether to increase/decrease the price, or no change) is selected based on a ϵ -greedy exploration-exploitation strategy [71] (line 3). Specifically, ϵ refers to the probability of exploration and it is initially set to 1. It is then decreased over time using an ϵ -decay factor, that is $\epsilon = \text{decay factor} \times \epsilon$. This way, exploitation gains more importance as the system learns the optimal policy. The algorithm returns an action a , that is used to update the current state s into the new state s_{new} , and to update the reward R_i (line 4). Additionally, the Q-value is updated accordingly (line 6) . The updated selling price is then sent to the respective seller i for the next trading period in line 7.

As discussed in the experimental section, *PQR* is able to correctly learn the optimal policy (or sellers’ prices, in our case). However, it needs to be noted that in a P2P energy trading model, like in most realistic scenarios, the state spaces can be very large and multidimensional. In fact, since the Q-learning must maintain Q-values for each state-action pair, even with just three actions, a finely discretized state space could lead to a huge number of state-action pairs that needs to be stored and updated continually. This is worsened by increasing the number of agents (sellers). As a result, *PQR* may suffer from severe scalability issues, due to its tabular approach of determining Q-values, as the system grows. Therefore, in the next subsection, we utilize a widely employed deep neural network-based *function approximator* that can be used to predict the Q-values using a learned function given state-action pairs.

5.2.3 ProDQN

In this subsection, we adopt a reinforcement learning approach based on Deep Q-Network (*DQN*) [145], for learning the seller’s optimal price, in order to overcome the scalability limitations of *PQR*. DQN is a reinforcement learning paradigm that exploits a deep neural network, called *Q-network*, as a non-linear function

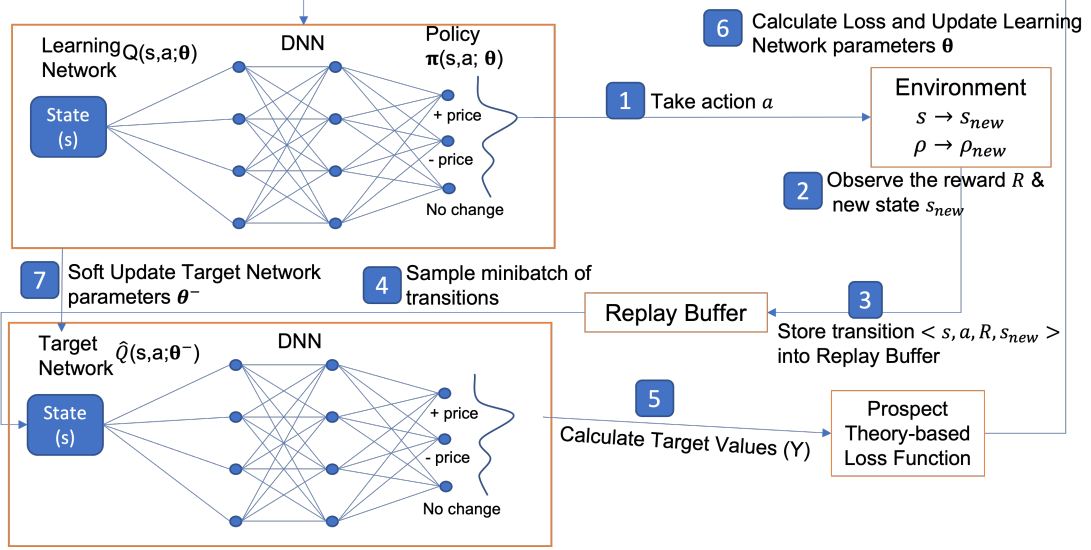


Figure 5.4: Overview of the ProDQN approach

approximator. Its parameters (or *weights*) are denoted by θ , thus the Q-value function $Q(s, a)$ becomes $Q(s, a; \theta)$. Note that, approximating the function through a neural network allows to not only represent the Q-values in a compressed form compared to the tabular Q-learning algorithm, but also to generalize over similar states.

We extend *DQN* to incorporate Prospect Theory elements, in order to devise a perceived utility-based pricing mechanism. We refer to our heuristic as Prospect theory-based DQN (*ProDQN*). An overview of the heuristic is presented in Fig. 5.4. It is important to note that, using a single Q-network for reinforcement learning may result in instability. This is due to the need of training the neural network itself while using it as a Q-function approximator. This is known as the issue of *moving target*, where the target, i.e. the expected optimal price in our application, is varying after each training period. *DQN* solves this problem by utilizing two different networks. One network is used for learning, while the other one to determine the target respectively. The first network is the *learning network*, denoted by $Q(s, a; \theta)$, which is used to take the best action given the current state. Secondly, we have a *target network*, denoted by $\hat{Q}(s, a; \theta^-)$, which is used to determine how close the output of the learning network is. The main difference between these network is that

the learning network is updated after every training period while the target network is updated less frequently. Thanks to these less frequent updates, the target network is kept relatively stable, and thus the overall learning also becomes stable.

An overview of *ProDQN* is shown in Fig. 5.4. *ProDQN* also employs two Q-networks – learning and target networks. Similar to *PQR*, each seller is represented by an individual *ProDQN* agent, and these agents do not share any information with each other to avoid conflicts and preserve privacy. Additionally the state spaces and action spaces are also the same as considered in *PQR*. As shown in Fig. 5.4, the learning network Q is used with parameters θ to predict the current action a , given the current state $s = \rho_i$ as input. The action is either to increase $(+\delta)$, decrease $(-\delta)$, or no change (item 1). Following this, the action is executed and the consequence of action taken is observed (i.e. new state s_{new} and new price ρ_{new} , and the resulting reward (R) is observed (item 2). The transition tuple $\langle s, a, R, s_{new} \rangle$ is stored in a Replay Buffer D . A minibatch of transitions $[z]$ of size z is randomly sampled from D (item 4) and passed to the target network \hat{Q} . The reason behind random sampling is to avoid bias due to high correlation in subsequent tuples. The target network returns a target value for each tuple (item 5), which is used to determine the error (or loss) in learning (item 6) and finally update the learning network parameters θ .

Specifically, for each sample $m = \{s^{(m)}, a^{(m)}, R^{(m)}, s_{new}^{(m)}\}$ in the minibatch $[z]$, the target value $Y^{(m)}$ is given by

$$Y^{(m)} = R^{(m)} + \gamma \max_{a'} \hat{Q}(s_{new}^{(m)}, a'; \theta^-) \quad (5.6)$$

The computation of term $\max_{a'} \hat{Q}(s_{new}^{(m)}, a'; \theta^-)$ is obtained from a single forward pass in the target network \hat{Q} for a given state $s_{new}^{(m)}$. According to the original version of *DQN* [145], given the target values as in Eq. (5.6), the parameters θ of the learning network $Q(s, a; \theta)$ are updated through stochastic gradient descent by minimizing a

standard loss function, usually the square loss. Conversely, in our work we propose a novel loss function in Eq. (5.7) based on PT-value function similar to the one proposed in Eqs. (5.1) and (5.5). Specifically, given the target value $Y^{(m)}$ of tuple m , the PT-based loss function $\mathcal{L}^{(m)}$ is defined as:

$$\mathcal{L}^{(m)} = \begin{cases} k_{+,.}(Y^{(m)} - Q(s, a; \boldsymbol{\theta}))^{\zeta_{+,.}}, & Y^{(m)} > Q(s, a; \boldsymbol{\theta}) \\ -k_{-,.}(Q(s, a; \boldsymbol{\theta}) - Y^{(m)})^{\zeta_{-,.}}, & Y^{(m)} \leq Q(s, a; \boldsymbol{\theta}) \end{cases} \quad (5.7)$$

Recall that $k_{+,.}, k_{-,.}, \zeta_{+,.}, \zeta_{-,.}$ are the PT parameters that quantify the perceived utility. After calculating the loss for each sample, the mean loss is determined by averaging the loss for all samples in the minibatch i.e. $\mathcal{L} = \frac{1}{z} \sum_m \mathcal{L}^{(m)}$. The

Algorithm 6: ProDQN

```

/* Pricing with Risk-sensitive Deep Q-Network */
1 Collect transaction information for each prosumers from DEbATE (Alg. 4)
   for current timestep  $t$ ;
2 for each  $i \in S$  do
3   Select an action,  $a \in \{+\delta, -\delta, 0\}$  based on exploration and exploitation ;
4   Observe the new state and reward:
      $s = \rho_i; s_{new} = s + a; R_i = s_{new} \sum_{j \in B} x_{ij} w_j$ ;
5   Store transition  $(s, a, R, s_{new})$  in Replay Buffer  $D$ ;
6   Sample random minibatch of transitions  $[z]$  from Replay Buffer  $D$ ;
7   For each  $k \in [z]$ , set the target values  $Y_i$  as in Eqs. (5.6);
8   Update  $\boldsymbol{\theta}$  as in Eq. (5.8) by performing gradient descent on PT-based
     loss function  $\mathcal{L}$ ;
9   Soft update  $\hat{Q}$  as in Eq. (5.9);
10   $\rho_i = s_{new}$ ;
11  Send information on updated price  $\rho_i$  to seller  $i$ ;
12 end

```

learning network's parameters are then updated by performing gradient descent step on network parameters $\boldsymbol{\theta}$, using the newly calculated loss \mathcal{L} as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \cdot \frac{1}{z} \sum_{m \in [z]} [(Y^{(m)} - Q(s, a; \boldsymbol{\theta}))] \nabla_{\boldsymbol{\theta}} Q(s, a; \boldsymbol{\theta}) \quad (5.8)$$

Finally, the parameters $\boldsymbol{\theta}^-$ of the target network are updated through soft updates. Specifically, an exponential moving average with parameter τ is used as follows:

$$\boldsymbol{\theta}_i^- \leftarrow \tau * \boldsymbol{\theta}_i + (1 - \tau) * \boldsymbol{\theta}_i^- \quad (5.9)$$

This process is then repeated for all the sellers to adjust their selling price in an automated manner similar to *PQR* algorithm (Alg. 5).

To the best of our knowledge, this is the first work using a PT value function-based loss calculation to update the Q-network parameters. This loss function is specially suited in our application scenario as it provides a way to capture the perceived utility of sellers based on the deviation from target values. It is to be noted that, with this PT-based loss function, the prediction of the Q-network $Q(s, a; \boldsymbol{\theta})$ tends to the target value (Y) (and therefore, to optimal Q-function i.e. $Q^*(s, a)$), transformed by the perceived utility of sellers as we update the parameters $\boldsymbol{\theta}$.

The system runs the algorithms *DEbATE* and *PQR/ProDQN* sequentially at every trading period. The input of *DEbATE* is updated based on the prices calculated by *PQR* or *ProDQN*, while *PQR/ProDQN* take as input the reward resulting from the energy transactions executed by *DEbATE*.

5.3 Experimental Results

In this section, we discuss the experimental setup, the comparison approaches, and then provide a detailed discussion on performance of both of the proposed solutions versus the comparisons. In the following discussion, we refer to *DEbATE* paired with *PQR* as *DEbATE – PQR*, and similarly to *DEbATE* paired with *ProDQN* as *DEbATE – DQN* in the results.

5.3.1 Experimental Setup

The experimental setup consists of a system with 40 prosumers, split evenly as buyers and sellers. This is considered a representative number of prosumers in a

microgrid or set of houses supplied by a single distribution transformer. We use a realistic dataset for buyer’s energy consumption obtained from [148]. Similarly, we consider sellers equipped with $4kW$ rooftop solar located in Lexington, Kentucky, USA. The energy generated is estimated using NREL’s PVWatts Calculator [149] given the solar irradiance in Lexington and size of solar panels. Losses are assigned uniformly at random (UAR) from set $\{1\%, 2\%, 3\%, 4\%\}$ and maximum loss threshold $L_{max} = 2.5\%$, while the minimum amount of energy to be exchanged (μ_j) is set to 50Wh for each buyers.

We assume that prosumers complete a survey before joining the system to estimate their individual prospect theory parameters, similar to [113,141,142]. For the purpose of experimentation, we use realistic prospect theory parameters from [113,141,142]. Specifically, the risk-averting parameter for gains (ζ_+) $\in [0.60, 0.88]$, the risk-seeking parameter for losses (ζ_-) $\in [0.52, 1.0]$, the loss-aversion parameters for gain and loss (k_+), (k_-) $\in [2.10, 2.61]$ for each individual prosumers. The grid or utility company generally sells energy at a higher price compared to the price it purchases energy. Therefore, we set the price at which the grid buys energy to $\rho_{gb} = \$0.06$ and the price at which it sells energy to $\rho_{gs} = \$0.12$, based on Kentucky’s average net-metering rate and energy selling price. With the P2P energy trading paradigm, sellers and buyers can exploit this gap to sell/buy energy among each other at a more convenient price than the grid. Therefore, we set grid’s energy selling price as upper bound for reference price for energy sellers and grid’s buying price as lower bound for reference price for energy buyers, respectively. Specifically, the reference price for each sellers is initially randomly sampled from range $[0.09, 0.12]$. It is then updated using either *PQR* or *ProDQN* at each iteration. The reference price for each buyer is selected in the range $[0.06, 0.10]$ and considered static for the duration of experiments, which is 365 days. The parameters for *PQR* algorithm are set as follows: learning rate $\alpha = 10^{-4}$, step size for discretizing state space $\delta = \$0.001$, and ϵ -decay is set 0.965.

ProDQN uses two Q-networks, learning and target network. Each of these consists of an input and output layer connected by two hidden layers with 64 nodes each. The input layer consists of a single node for state and output layer consists of three nodes for three actions. Other hyperparameters are set as follows: learning rate $\alpha = 0.0075$, replay buffer size $|D| = 1000$, minibatch size $z = 4$, discount factor $\gamma = 0.8$, soft update rate for target network $\tau = 0.01$. These hyperparameters were chosen manually for best results. Hyperparameter optimization techniques could also be adopted such as grid search, Bayesian search, and population-based evolutionary search for further fine-tuning. We developed the P2P energy trading simulation environment and implemented the algorithms in Python using SciPy and PyTorch libraries.¹

5.3.2 Comparison Approaches

In order to highlight the efficacy of our proposed approaches *DEbATE – PQR* and *DEbATE – DQN*, we compare their performance against two recently proposed state-of-the-art approaches. The first approach, referred to as *Rule*, is proposed in [34]. *Rule* allocates energy using a greedy heuristic that assigns cheapest sellers to buyers. Buyers are selected based on their registration order to the system. A mid-market pricing strategy is employed, i.e., the final price of a transaction is the mid value of seller’s and buyer’s asking price.

The second approach has been proposed in [35], to which we refer as *Zhu* from the name of the first author. This approach has been proposed to minimize the loss of the energy transactions. It employs a greedy algorithm to assign the energy among buyers and sellers. The algorithm considers buyers in decreasing order of demand. At each iteration, a buyer is selected and assigned to the sellers with the smallest loss for that buyers, until the demand of the buyer is satisfied. In this approach, the transaction price is given by the seller’s asking price.

¹Scripts for the simulation can be found at this Github link: https://github.com/ashutoshtmlsna/P2P_energy_trading

It is worth nothing that, both approaches do not consider the perceived utility of the buyers and they do not dynamically adjust the price of sellers. As discussed in the following, our approach matches demand and production by generating a market in which both the needs and perceptions of buyers and sellers are taken into account.

5.3.3 Results

We consider several experimental scenarios and performance metrics, as discussed in the following.

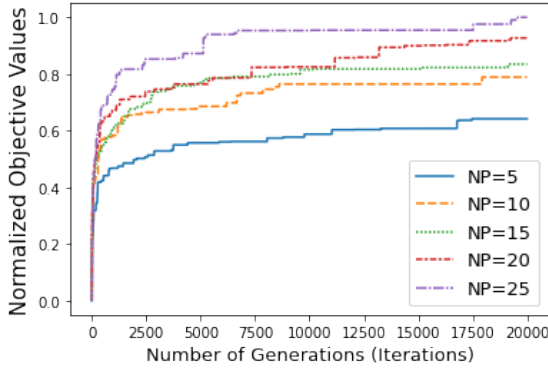


Figure 5.5: Normalized objective value vs. number of iterations for varied population sizes.

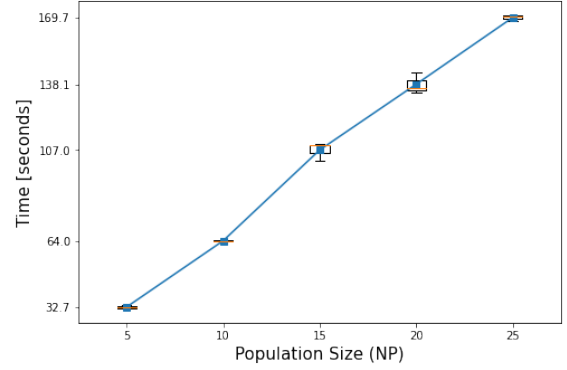


Figure 5.6: Computation time vs. population sizes.

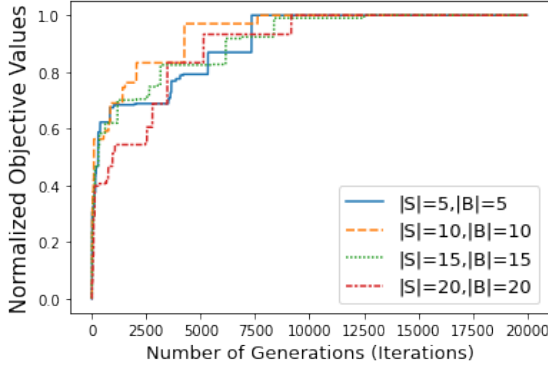


Figure 5.7: Normalized objective value vs. number of iterations for varied network sizes.

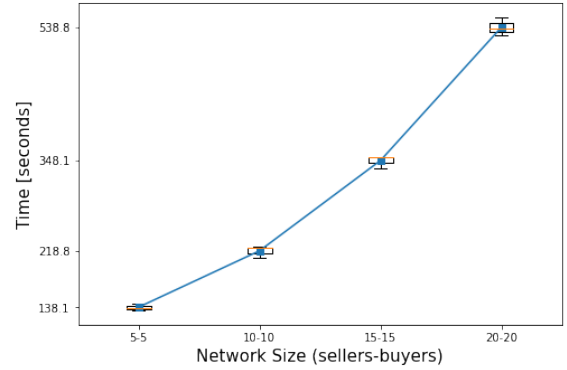


Figure 5.8: Computation time vs. network sizes.

Experimental Scenario 1: We first run two experiments to study the evolutionary aspects and convergence of *DEbATE*. Specifically, we want to know

the impact of the number of generations (G_{max}) and population size (NP) on the quality of the solution, i.e., on the value of the objective function. We first study the impact of the population size NP on the value of the objective function of the optimization problem in Eq. 5.2 and on the computational time. Specifically, we vary NP from 5 to 25. In this experiments, we consider a system with 5 sellers and 5 buyers. Fig. 5.5 shows the respective plot averaged over 10 runs. It can be seen that, as the population size is increased, *DEbATE* is able to find a better solution. Additionally, with all the considered population sizes, *DEbATE* is able to quickly converge towards a good solution with few iterations. We show in Fig. 5.6 the computational time versus the population size along with error bars. The figure clearly shows that the computation time grows linearly with the population size. This is in accordance with the complexity derived in Lemma 3.

We now consider the impact of the number of generations (iterations) G_{max} on the quality of the solution and execution time. We consider different network sizes by linearly scaling the number of sellers and buyers from $|S| = |B| = 5$ to $|S| = |B| = 20$, and setting $NP = 20$. Fig. 5.7 shows the normalized objective value as a function of G_{max} . The results show that by setting $G_{max} = 10,000$ iterations is sufficient for the algorithm to converge in the considered settings. We plot in Fig. 5.8 the computation time of *DEbATE* by increasing the system size along with the error bars. According to Lemma 3, the computation complexity is proportional to $|S| \times |B|$. As a result, by increasing both buyers and sellers linearly, we incur in a quadratic increase in the execution time.

Given the results of the aforementioned experiments, in the following, we select a trade-off between computation time and quality of the solution. For the remaining of the experiments, we therefore set the population size $NP = 20$, since it yields a solution with similar objective value while requiring 22% less execution time, and set $G_{max} = 10,000$. This helps to ensure that the algorithm will generate a quality

solution.

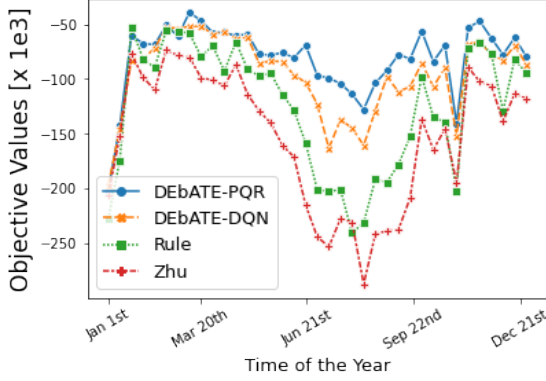


Figure 5.9: Buyers' perceived values.

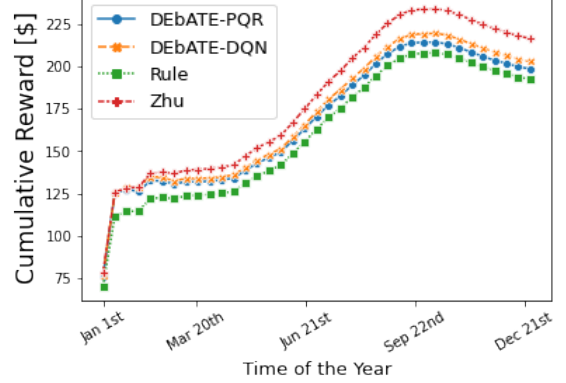


Figure 5.10: Sellers' cumulative reward.

Table 5.1: Statistical Analysis of experimental result in Figs. 5.9 and 5.10

Metric	Approach	DEbATE-PQR	DEbATE-DQN	Rule	Zhu
<i>Obj. Val</i>	Mean	-83.182	-97.818	-128.599	-155.312
	Std. Deviation	34.738	37.800	60.288	61.983
<i>Reward</i>	Mean	197.9	202.3	191.8	215.8
	Std. Deviation	79.778	82.963	79.916	90.198

Experimental Scenario 2: In the second experimental scenario, we study the performance of the considered approaches over time. Two performance metrics are considered, namely the buyers' objective value and the sellers' cumulative reward. These are represented in Figs. 5.9 and 5.10, respectively, with a moving average of 10 days. In these experiments, we consider 15 buyers and 15 sellers. Note that, the buyers' objective values are negative because they are paying higher prices than their reference purchase price. Therefore, transactions are seen as loss from a prospect theory perspective. Clearly, *DEbATE-PQR* and *DEbATE-DQN* perform better than *Rule* in both metrics. The greedy nature of *Rule* penalizes the quality of the resulting matching, significantly reducing the buyers' perceived value while both our approaches optimize the energy assignments to maximize the buyers perceived utility. Additionally, our approaches are able to generate higher rewards than *Rule* by dynamically learning the prices for sellers. *Zhu* however performs the worst in

terms of the buyers' objective values, but performs the best in terms of cumulative reward. This is because the energy assignment is driven by loss minimization, not taking into consideration the buyers' reference price. This, paired with the trading price set as the sellers' asking price, results in a market heavily biased towards sellers, achieving a very low perceived utility for buyers. We present a statistical analysis of experimental results in Figs. 5.9 and 5.10 with respect to both mean and standard deviation in table 5.1. As the table shows, the mean objective value is significantly higher for $DEbATE - PQR$ and $DEbATE - DQN$, with respect to the comparison approaches. This is also paired with a lower standard deviation, which implies more stable system performance. $DEbATE - DQN$ produces a slightly higher mean reward and a higher standard deviation. This is due to higher randomness engraved in deep learning frameworks. In line with our observation from Fig. 5.10, *Zhu* highly favors sellers with respect to buyers.

It is worth noting that, the benefits of $DEbATE - PQR$ and $DEbATE - DQN$ over *Rule* and *Zhu* are more prominent from April through October, when the energy demand and production are higher. Note that, the energy consumption is higher during summer months due to the higher use of air conditioning equipment. Similarly, the energy production is higher due to the increased solar radiation in these months. Comparing $DEbATE - PQR$ and $DEbATE - DQN$ we notice that *ProDQN* slightly penalizes buyers (lower utility) in favor of sellers (higher rewards). This slight imbalance is however compensated by the better scalability of *ProDQN*. In general, the sellers' reward decreases after mid-September for all four approaches, due to the reduced energy production during winter.

Experimental Scenario 3: We further study the performance over time by considering the evolution of individual sellers' prices. We consider a smaller system of 5 sellers and 5 buyers for ease of representation of the results. Fig. 5.11 shows the individual prices set by *ProDQN* algorithm while Fig. 5.12 shows the individual

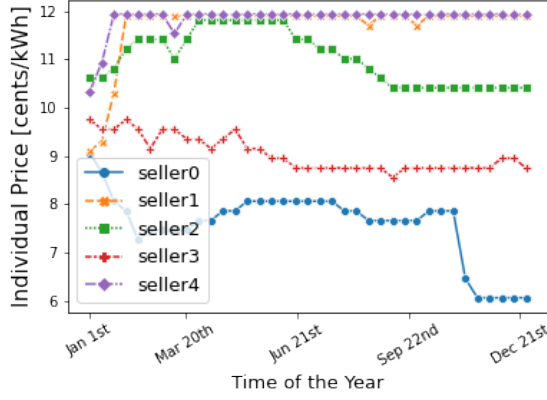


Figure 5.11: Individual prices for *ProDQN*.

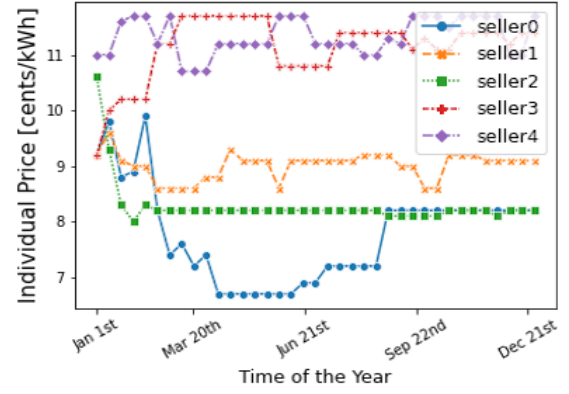


Figure 5.12: Individual prices for *PQR*.

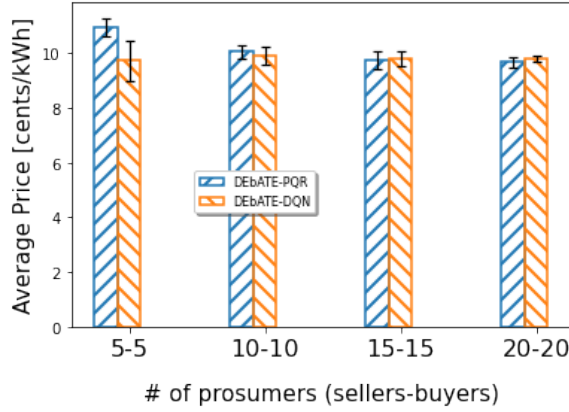


Figure 5.13: Avg. price comparison for different network sizes between *PQR* and *ProDQN*.

prices by *PQR*. Both approaches proposed in our system are able to learn and adjust the price over time to improve the buyers' perceived value while considering the sellers' competitiveness. The competitiveness of a seller is a function of buyers' reference prices, the seller production, and their location in the system (e.g., loss w.r.t. buyers). Note that, although both algorithms adjust prices based on the output of the transactions, which indirectly reflects the sellers' competitiveness, the evolution of prices under *ProDQN* and *PQR* may differ. When taken collectively, both algorithms are able to find a balance between buyers perceived utility and sellers reward. To support this statement, we show in Fig. 5.13 the average sellers' prices, after a year

of execution of the algorithms, with different system sizes. Both approaches converge towards similar prices, with negligible differences as the system grows.

Experimental Scenario 4: In this scenario we test the scalability of the proposed approaches with respect to the system size through a year-long aggregated analysis. Specifically, we increase the system proportionately from $|S| = 5$ sellers and $|B| = 5$ buyers to $|S| = 20$ sellers and $|B| = 20$ buyers. Figs. (5.14)-(5.15) show the buyers' total perceived value and the sellers' reward, respectively, over a year. By considering the loss-averse and risk-seeking PT-value functions, *DEbATE - PQR* and *DEbATE - DQN* achieve an increasing advantage as the system size increases compared to *Rule*, for both sellers and buyers. *Zhu*, as previously discussed, creates a heavily biased market that penalizes buyers and favors sellers. As a numerical example, *DEbATE - PQR* achieves as much as 26% increase in buyers' perceived value while ensuring 7% profit improvement for sellers compared to *Rule*. Similarly, *DEbATE - DQN* achieves 8% more profit for sellers with almost 23% more in buyer's perceived utility.

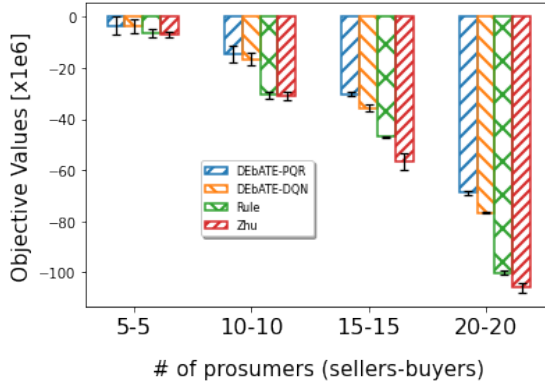


Figure 5.14: Objective values for buyer vs. network size.

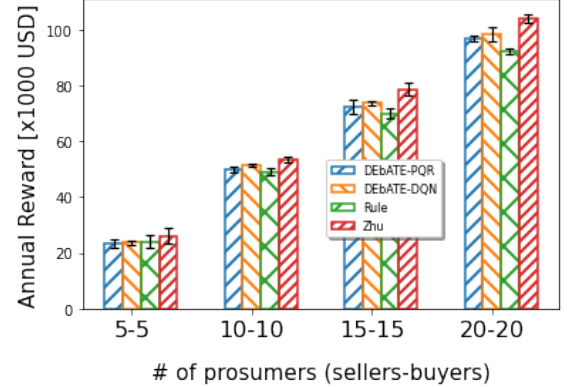


Figure 5.15: Total rewards for sellers vs. network size.

5.4 Concluding Remarks

In this chapter, we brought together the concept of perceived utility from behavioral economics and reinforcement learning into P2P energy trading. Unlike existing

literature, we propose an automated and dynamic P2P energy trading problem that maximizes the perceived value for buyers while simultaneously learning the optimal selling price for sellers. Given the non-linear and non-convex nature of the problem, we propose a novel differential evolution-based metaheuristic algorithm, called *DEbATE*. *DEbATE* is paired with a prospect theory enhanced Q-learning algorithm, called *PQR*, to adjust the selling price over time. Given the limitations of the tabular *Q-learning* approach of *PQR*, we propose a Deep Q-Network-based algorithm called *Pro-DQN* that proposes a novel loss function based on PT value function to model the seller’s perceived utility. Results show the advantages of the proposed approaches with respect to a state of the art solution using real energy consumption and production data.

The work in this chapter shows that integrating concepts from the behavioral economics and reinforcement learning can lead to more efficient and effective energy exchange in peer-to-peer (P2P) energy trading systems. It is also supported by the results showing how the proposed algorithms, i.e., *DEbATE*, *PQR*, and *Pro-DQN*, outperform existing solutions in maximizing perceived value for buyers as well as learning the optimal selling prices for sellers.

CHAPTER 6. *E-UBER*: A CROWDSOURCING PLATFORM FOR ELECTRIC VEHICLE-BASED RIDE- AND ENERGY-SHARING

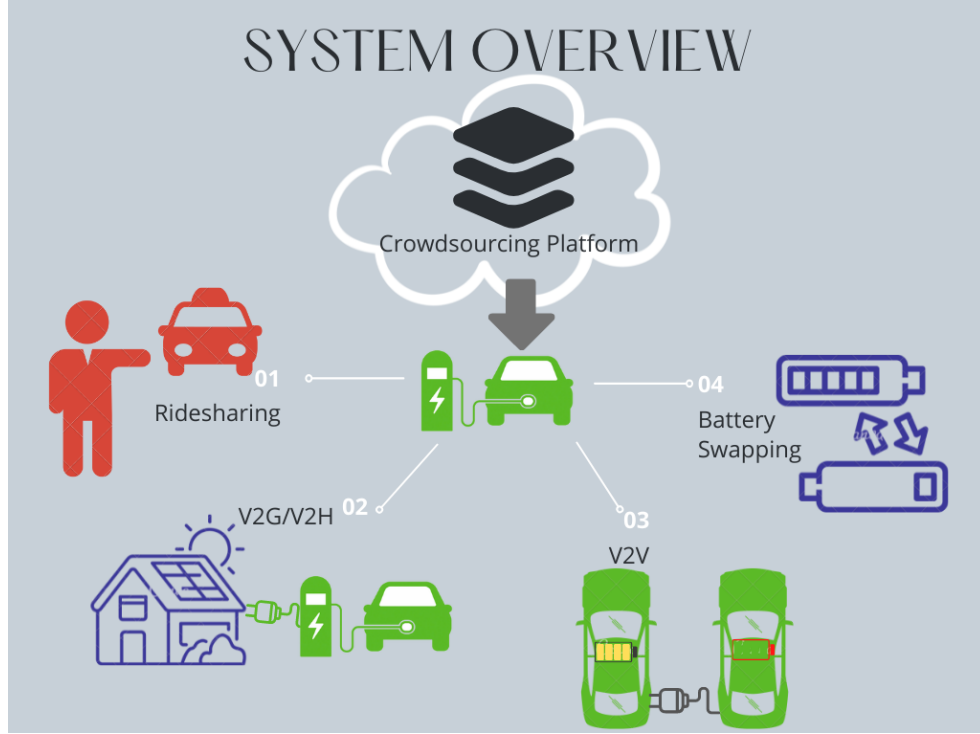


Figure 6.1: e-Uber crowdsourcing platform overview

The sharing-economy-based business model has recently seen success in the transportation and accommodation sectors with companies like Uber and Airbnb. There is growing interest in applying this model to energy systems, with modalities like peer-to-peer (P2P) Energy Trading, Electric Vehicles (EV)-based Vehicle-to-Grid (V2G), Vehicle-to-Home (V2H), Vehicle-to-Vehicle (V2V), and Battery Swapping Technology (BST). In this chapter, we exploit the increasing diffusion of EVs to realize a crowdsourcing platform called *e-Uber* that jointly enables ride-sharing and energy-sharing through V2G and BST. *e-Uber* exploits *spatial crowdsourcing*, reinforcement learning, and *reverse auction* theory. Specifically, the platform uses reinforcement learning to understand the drivers' preferences towards different ride-sharing and energy-sharing tasks. Based on these preferences, a personalized list is recommended

to each driver through CMAB-based Algorithm for task Recommendation System (CARS). Drivers bid on their preferred tasks in their list in a reverse auction fashion. Then e-Uber solves the task assignment optimization problem that minimizes cost and guarantees V2G energy requirement. We prove that this problem is NP-hard and introduce a bipartite matching-inspired heuristic, Bipartite Matching-based Winner selection (BMW), that has polynomial time complexity. Results from experiments using real data from NYC taxi trips and energy consumption show that e-Uber performs close to the optimum and finds better solutions compared to a state-of-the-art approach.

6.1 System Model

Table 6.1: List of Notations

\mathcal{S}_t	List of all tasks at timeslot t
$s_j = \langle z_j, c_j, d_j \rangle$	j^{th} task represented by type of task (z_j), start position (c_j), destination (d_j)
\mathcal{W}_t	List of all workers available at timeslot t
$w_i = \langle c_i, e_i, r_i, r_i^{min} \rangle$	i^{th} worker represented by current location (c_i), the energy per unit range (e_i), remaining range of the EV (r_i) and minimum range threshold (r_i^{min})
\mathcal{B}_t	List of all the bids received at timeslot t
b_{ij}	Bid submitted by worker i for task j
α_{iz_j}	Acceptance probability of worker i to the j^{th} task type z_j
K	Maximum number of tasks to be recommended
λ	Proximity distance
\mathcal{E}_t	Amount of energy that must be satisfied through V2G/V2H
\mathbf{q}^*	Optimal solution to $f(\cdot)$ / Winning bids

We assume time to be divided in time slots. At each time slot t , the set of tasks is referred to as \mathcal{S}_t , which are crowdsourced to the workers. We refer to \mathcal{W}_t as the set of workers at time t . Each task in \mathcal{S}_t is denoted by a tuple $s_j \stackrel{\text{def}}{=} \langle z_j, c_j, d_j \rangle$ where z_j is the type of task (0—ride-share, 1—battery swapping, and 2—V2G), c_{s_j} is the start position and d_j is the destination of task. For energy-sharing tasks, although spatial in nature, start position c_{s_j} is same as destination d_j . We assume the utility company submits energy tasks as a result of an energy requirement \mathcal{E} . This is a

typical assumption for demand response solutions [20, 103, 150]. As a result, the total amount of energy provided by workers through V2G must be at least \mathcal{E} . Each worker in \mathcal{W}_t is denoted by a tuple $w_i \stackrel{\text{def}}{=} \langle c_{w_i}, e_i, r_i, r_i^{\min} \rangle$, where c_i is the current position of the EV worker w_i which can be different to spatial task location c_{s_j} , e_i is the energy per unit range value in (kWh/km) that gives information about how much energy the EV consumes to drive a unit distance, r_i is the available range of electrical vehicle in km given by the remaining energy level in their batteries, and r_i^{\min} is the minimum energy not to be exceeded after completing the task to ensure sufficient energy for traveling to a charging location. The energy required to perform task s_j by worker w_i is denoted by l_{ij} . e-Uber provides that a list of tasks, called recommendation list, is sent out to each worker. Workers then submit bids to these tasks. The bid $b_{ij} \in \mathcal{B}$ represents the cost asked by worker w_i to perform task s_j , where \mathcal{B} is the set of all the bids submitted by workers.

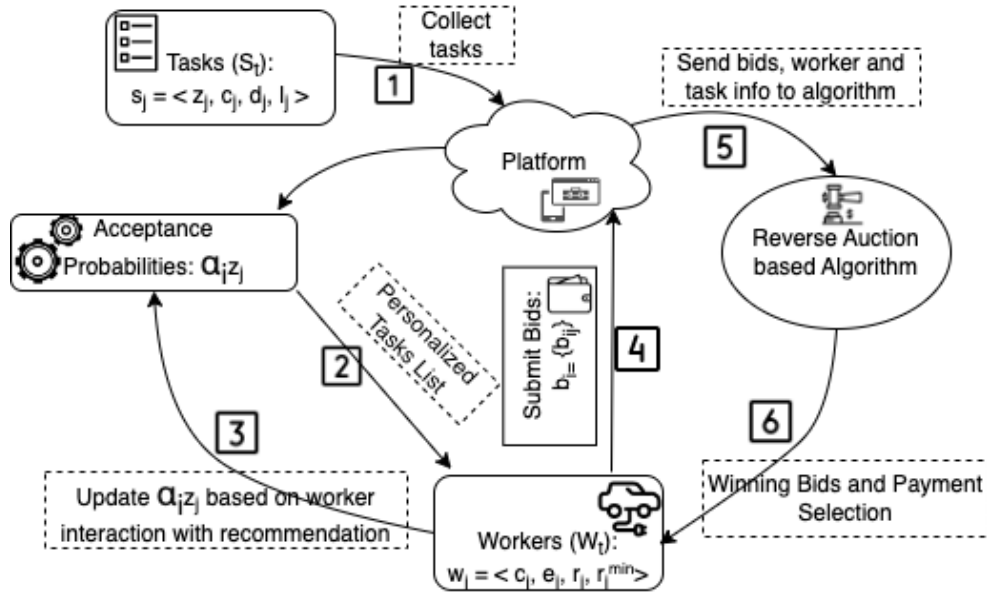


Figure 6.2: Working mechanism of e-Uber

Previous works in crowdsourcing and energy-sharing using EVs has generally assumed that workers would have complete access to the list of available tasks and

would pick the best task for them or, conversely, the crowdsourcing platform would assign tasks to workers regardless of their preference. These assumptions are both undesirable. On the one hand workers have limited time and ability to go over potentially a very long list of tasks [45], and on the other hand workers may have different preferences on the tasks to complete. In this work, we recommend a limited list of relevant tasks to each worker based on their preferences. We model the preferences as follows. We denote by $\alpha_{iz_j} \in [0, 1]$ the probability that worker w_i bids for a task of type z_j . These are called *bidding probabilities*. We assume that these probabilities are unknown and thus need to be learned over time by observing the workers' behavior.

6.2 e-Uber: Problem Formulation

Fig. 6.2 summarizes the steps involved in the e-Uber platform. e-Uber collects a list of tasks \mathcal{S}_t at time t as requested by task-requesters which need to be crowdsourced to the EV-based workers in \mathcal{W}_t (step 1). The platform sends a personalized list of tasks to the workers based on their preferences (step 2) to which they respond by submitting bids to the platform for the tasks (step 3). Based on the received bids \mathcal{B}_t (step 4), the platform uses reverse auction based algorithm to determine the winning bids \mathbf{q}^* along with final payment \mathbf{P} for winners (step 5). Finally, the worker preferences are updated based on their feedback for the next time step (step 6). Given the nature of the considered tasks, worker-task assignment is performed one-to-one.

As described above, the system involves solving two different problems. One is to recommend a set of tasks which maximizes the likelihood of generating the maximum number of bids, and thus improving the overall system performance. Another problem is to select the winning bids for task assignment and determine the final payment to crowdsource the tasks to the workers. These two problems are discussed below.

6.2.1 Preference-aware Optimal Task Recommendation Problem

Our objective is to recommend a limited subset of tasks to each workers which maximizes the likelihood of bidding for these tasks, while avoiding to overwhelm workers with a list above their cognitive capabilities. We formalize this through the Preference-aware Optimal Task Recommendation (POTR) problem as follows. In short, the problem aims at maximizing the overall task bidding probabilities (hereafter referred interchangeably as preferences) while limiting the size of the recommended list to K as well as ensuring that each task is recommended to at least ψ workers.

$$\text{maximize} \quad \sum_{w_i \in \mathcal{W}} \sum_{s_j \in \mathcal{S}} \alpha_{iz_j} x_{ij} \quad (6.1)$$

$$\text{s.t.} \quad \sum_{s_j \in \mathcal{S}} x_{ij} \leq K, \quad \forall w_i \quad (6.1a)$$

$$\sum_{w_i \in \mathcal{W}} x_{ij} \geq \psi, \quad \forall s_j \quad (6.1b)$$

$$\sum_{s_j \in \mathcal{S}} g(z_j) x_{ij} \geq \frac{|V2G|}{|\mathcal{S}|} K, \quad \forall w_i \quad (6.1c)$$

$$l_{ij} x_{ij} \leq (r_i - r_i^{\min}) e_i, \quad \forall w_i, s_j \quad (6.1d)$$

$$x_{ij} = 0, \text{ if } |c_{s_j} - c_{w_i}| > \lambda, \quad \forall w_i, s_j \quad (6.1e)$$

$$x_{ij} \in \{0, 1\}, \quad \forall w_i, s_j \quad (6.1f)$$

$$g(z_j) = \begin{cases} 1, & \text{if } z_j = 2 \\ 0, & \text{otherwise} \end{cases} \quad (6.2)$$

The objective function in Eq. (6.1) maximizes the sum of individual bidding probabilities for each worker's recommended tasks. The binary decision variable $x_{ij} \in \{0, 1\}$ is set to 1 if the task s_j is included in the list of worker w_i . Constraint (6.1a)

limits the length of each recommendation list to be less than K . In constraint (6.1b), we ensure that each task is recommended to at least $\psi = \left\lfloor \frac{|W|K}{|S|} \right\rfloor$ workers. Also, we ascertain that a minimum of $\frac{|V2G| \times K}{|S|}$ V2G tasks are also recommended to each workers in constraint (6.1c). Constraint (6.1d) requires the recommended tasks to consume no more than certain energy for each EV, ensuring that EV has sufficient energy after performing tasks to drive to charging location, if required. Finally, constraint (6.1e) ensures that only the tasks within λ distance from workers are recommended.

It is to be noted that the information on bidding probabilities is difficult to obtain *a priori* as it is specific for each worker and include elements of complex human psychology. Therefore, we assume that the preferences are initially unknown and are learned by observing the workers' behavior with respect to the assigned tasks. Recently, reinforcement learning mechanisms have been used extensively to learn the optimal policies in the run-time that gradually converge to take optimal actions based on feedback from the environment. In section 6.3, we present a *Combinatorial Multi-Armed Bandit (CMAB)*-based approach [45] that learns the preferences of workers over time while simultaneously recommending the optimal personalized list of tasks to them.

6.2.2 Winning Bid Selection and Final Payment Problem

After sending the personalized list of tasks to each worker, *e-Uber* collects the bids. Given the collected bids, e-Uber selects winning bids, i.e., the workers performing the tasks, by solving the Winning Bid Selection (*WiBS*) problem. This problem determines the best bids which minimize the total cost from perspective of task requesters. *WiBS* can then be formulated a constrained assignment problem as

follows:

$$\text{minimize} \quad \sum_{w_i \in \mathcal{W}} \sum_{s_j \in \mathcal{S}} b_{ij} q_{ij} \quad (6.3)$$

$$\text{s.t.} \quad \sum_{s_j \in \mathcal{S}} q_{ij} \leq 1, \quad \forall w_i \quad (6.3a)$$

$$\sum_{w_i \in \mathcal{W}} q_{ij} = 1, \quad \forall s_j, z_j < 2 \quad (6.3b)$$

$$\sum_{w_i \in \mathcal{W}} q_{ij} \leq 1, \quad \forall s_j, z_j = 2 \quad (6.3c)$$

$$\sum_{w_i \in \mathcal{W}} \sum_{s_j \in \mathcal{S}} g(z_j) l_{ij} q_{ij} \geq \mathcal{E}, \quad (6.3d)$$

$$q_{ij} \in \{0, 1\}, \quad \forall w_i, s_j \quad (6.3e)$$

The objective function in Eq. (6.3) minimizes the total cost of performing tasks from the collected bids. q_{ij} is the binary decision variable as defined in constraint (6.3e) that indicates whether a bid b_{ij} wins the auction and therefore the task s_j is assigned to worker w_i . Constraint (6.3a) ensures that a worker is assigned at most one task, while (6.3b) allows a ride-sharing and battery swapping tasks ($z_j < 2$) to be assigned to only one worker. Similarly, constraint (6.3c) ensures that a V2G task is assigned to at most one worker. Finally, constraint (6.3d), ensures that at least \mathcal{E} amount of energy will be supplied through V2G services. Note that the function $g(z_j) = 1$ if $z_j = 2$ (V2G task) and zero otherwise.

Following the selection of winning bids by solving the *WiBS* problem in Eq. (6.3), the final payment for each winning worker w_k assigned with task s_j is the second-to-the-selected bid received for that task. Since with the second price payment rule, the dominant strategy for all bidders is to bid truthful [151], it ensures rational workers will provide truthful bids.

Theorem 6. *WiBS problem defined in Eq. (6.3) is NP-hard.*

Proof. We provide a reduction from NP-Hard 0-1 min Knapsack (0-1 min-KP) problem [152]. In this problem, a set n items is provided, each item a_i has a value l_i and weight b_i . The goal is to select the subset of items that incurs minimum weight and has a value of at least \mathcal{E} .

Given a generic instance of min-KP, we construct an instance of our problem as follows. We only consider V2G tasks ($z_j = 2$). For each item a_i of min-KP we create a pair task-worker (s_{a_i}, w_{a_i}). We assume that worker w_{a_i} only submits one bid, and they bid for s_{a_i} for an amount b_i (the weight of a_i in min-KP). Additionally, the energy required by w_{a_i} to perform s_{a_i} is l_i (the value of a_i in min-KP). Finally, we set the energy requirement for V2G to \mathcal{E} .

Under these assumptions, the decision variable q_{ij} of our original problem can be reduced to q_i , since only one workers bid for one task and a task receives a bid only from one worker. Additionally, constraints (6.3a) and (6.3c) are trivially verified, since there is only task-worker pair, while constraint (6.3b) does not apply since we only have V2G tasks.

Solving our reduced problem instance finds the set task-worker pairs that minimize the sum of bids and meets the energy requirement \mathcal{E} . This corresponds (i.e., it can be translated in polynomial time) to the optimal solution of min-KP, i.e., the set of items with minimum weight that provide a value at least \mathcal{E} . As a result, our problem is at least as difficult as min-KP, and thus it is NP-Hard.

□

6.3 e-Uber Solution Approaches

6.3.1 CMAB-based Task Recommendation System

In order to solve the optimization problem in Eq. (6.1), it is necessary to have beforehand knowledge on the workers preferences. These are generally not known a priori in realistic settings. Therefore, it becomes necessary to learn these preferences during run-time, while simultaneously optimizing the task assignment. To this

purpose, we propose a reinforcement learning approach inspired by the Combinatorial Multi-Armed Bandit (CMAB) framework [45, 132].

Combinatorial Multi-Armed Bandit is a classic reinforcement learning problem that consists of setup where agents can choose a combination of different choices (i.e. certain decision-making *actions*) and observe a combination of linear *rewards* at each timestep. The long term objective for the problem is to find a strategy that maximizes such reward by selecting optimal actions. This strategy, better defined as *policy*, needs to be learned based on how the agents choose to interact with the system. The learning is carried out through *exploration vs. exploitation* trade-off. Since, at the beginning, the knowledge about how an agent chooses to engage with the system is not known, the system learns by allowing agent to choose from diverse options and therefore learning the user interaction accordingly, referred to as *exploration*. As the time passes, the system starts gathering information about agent’s behavior and therefore use that knowledge instead of sending out diverse range of choices, called *exploitation*. By balancing this exploration and exploitation mechanism over the course of time, the system eventually gathers sufficient information on agent’s behavior and learns optimal strategy for them. In our problem setting, the workers are the agents who needs to be sent out an optimal set of tasks so as to accumulate good quality bids from them. Specifically, the objective is to find the best possible task recommendations (actions) to be sent to each workers (agent) that will result in higher cumulative preferences for workers (reward).

Therefore in this section, based on this CMAB framework, we design an algorithm called *CMAB-based Algorithm for task Recommendation System (CARS)*. The pseudo code of *CARS* is shown in Alg. (7). *CARS* recommends the personalized tasks to each workers based on current estimation of worker preferences towards each task type. Note that the worker preference is defined as the bidding probability in section 4.2 that a worker will submit a bid for any task based on its type. The algorithm then

updates and learns these bidding probabilities based on the worker's engagement on the recommendation through bids. If the worker submits a bid, it is considered to be a preferred recommendation and opposite, if the worker chooses to ignore by not the submitting bid. Based on this information, the preference of workers towards each task type is updated.

Therefore, with \mathcal{F} as the overall solution space that consists of all feasible action matrices, the action matrix $\mathbf{A}(t) \in \mathcal{F}$ corresponds to the optimal set of recommendation lists for the timestep t . It consists of action values $x_{ij} \in \{0, 1\}$, which is same as the decision variable in POTR problem. Recall that it represents whether the task s_j is in personalized recommendation list of worker w_i for timestep t . Given this action matrix, the preference of worker w_i towards each task type z_j is modeled as a random variable $\bar{\alpha}_{iz_j}$ whose mean value is α_{iz_j} and is initially unknown. The current knowledge until timestep t for these random variables $\bar{\alpha}_{iz_j}$ is denoted by the estimated expected $\hat{\alpha}_{iz_j}$. The reward for the platform for selecting the action matrix $\mathbf{A}(t)$ at timestep t , is defined as the sum of the preferences to each workers:

$$\mathbf{R}_{\mathbf{A}(t)}(t) = \sum_{w_i, s_j} a_{ij}(t) \bar{\alpha}_{ij}(t) \quad (6.4)$$

Since the distribution of $\bar{\alpha}_{iz_j}$ is unknown, the goal of this CMAB-based approach is to learn the policy, that minimizes the overall *regret* up to time t . This regret is defined as the difference between expected reward with perfect knowledge of preferences and that obtained by the policy over time:

$$\mathcal{R}(t) = t\mathbf{R}_{\mathbf{A}(t)}^*(t) - \mathbb{E}\left[\sum_{t'=1}^t \mathbf{R}_{\mathbf{A}(t')}(t')\right], \quad (6.5)$$

where $\mathbf{R}_{\mathbf{A}(t)}^*(t)$ is the optimal reward obtained with perfect knowledge of the preference variables. Even though minimizing the regret is a difficult problem, *CARS* ensures that the regret is bounded, meaning the non-optimal actions will be picked

only a limited number of times and eventually the learned policy will converge towards optimal. We present a modified objective function from UCB1 algorithm to select the action matrix as follows.

$$\mathbf{A}(t) = \arg \max_{\mathbf{A} \in \mathcal{F}} \sum_{w_i \in \mathcal{W}} \sum_{s_j \in \mathcal{S}} a_{ij} \left(\hat{\alpha}_{iz_j} + \sqrt{\frac{(Q+1) \ln t}{m_{iz_j}}} \right) \quad (6.6)$$

where $Q = |\mathcal{W}| \times |z_j|$ is the total number of variables and m_{iz_j} is the number of observations so far for the variable $\bar{\alpha}_{iz_j}$.

At each timestep t , we solve the *POTR* problem with CMAB-based objective function in Eq. (6.6) instead of Eq. (6.1) and same constraints (6.1a)-(6.1f). By solving this modified problem, the sets of optimal actions (or recommendation lists) for each workers are selected based on current estimate of preferences until timestep $(t-1)$. For this purpose, we keep track of the $\hat{\alpha}_{iz_j}$, along with m_{iz_j} . These two information are then used to update the current estimation of the variable $\bar{\alpha}_{iz_j}$ at time t based on the worker's engagement with the recommendation i.e. whether the worker chooses to submit the bid or not. Needs to be noted that, if the worker chooses to submit the bid, they must complete the task if assigned.

$$\hat{\alpha}_{iz_j}(t) = \begin{cases} \frac{\hat{\alpha}_{iz_j}(t-1)m_{iz_j}(t-1) + \alpha_{iz_j}(t)}{m_{iz_j}(t-1) + 1} & \text{if } 0 < b_{ij} < \infty, \\ \hat{\alpha}_{iz_j}(t-1) & \text{otherwise.} \end{cases} \quad (6.7)$$

$$m_{iz_j}(t) = m_{iz_j}(t-1) + 1 \quad (6.8)$$

We present the *CARS* algorithm in Alg. 7. *CARS* begins by collecting information on workers and task in lines 1 – 2. It then sends out personalized recommendation to each worker by solving the optimization problem with Eq. (6.6) as objective function and constraints (6.1a)-(6.1f)(lines 3 – 4). Then, it collects the bids for recommended tasks from workers (line 5). Finally, the current knowledge on worker's bidding

probabilities are updated according to the Eqs. (6.7) and (6.8) based on how the workers respond to recommendations (lines 5 – 6). For the update process, the recommendations that receive a bid from workers are taken as positive reinforcement and the recommendations that do not receive bids as negative reinforcement. In the following, we prove that the Alg. 7 has a bounded regret and thus the algorithm eventually converges to optimal policy in finite time-steps.

Algorithm 7: CMAB-based Algorithm for task Recommendation System (CARS)

- 1 $\forall w_i \in \mathcal{W}_t$, collect the workers info $w_i = \langle c_i, e_i, r_i, r_i^{min} \rangle$;
 - 2 $\forall s_j \in \mathcal{S}_t$, collect the tasks $s_j = \langle z_j, c_j, d_j \rangle$;
 - /* Solve CMAB-based POTR problem */
 - 3 Select an action \mathbf{A} s.t. $\mathbf{A}(t) = \arg \max_{\mathbf{A} \in \mathcal{F}} \sum_{w_i} \sum_{s_j} a_{ij} \left(\hat{\alpha}_{iz_j} + \sqrt{\frac{(Q+1) \ln t}{m_{iz_j}}} \right)$;
 - 4 Send list of recommendations $\mathbf{A}(t)$ to the workers;
 - 5 Collect bids \mathcal{B}_t from workers based on $\mathbf{A}(t)$;
 - 6 Update $[\hat{\alpha}_{iz_j}]_{|\mathcal{W}| \times |z_j|}$ and $[m_{ij}]_{|\mathcal{W}| \times |z_j|}$ based on the collected bids using Eqs. (6.7) and (6.8);
-

Theorem 7. *CARS provides bounded regret given by:*

$$\mathcal{R}(t) \leq \left[\frac{4a_{max}^2 Q^3 (Q+1) \ln(t)}{(\Delta_{min})^2} + \frac{\pi^2}{3} Q^2 + Q \right] \Delta_{max}, \quad (6.9)$$

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 [132]. □

However, as shown in Theorem (6), finding the optimal solution for winner determination problem (*WiBS* problem Eqs. (6.3)-(6.3e)) is NP-Hard problem. Therefore, we devise a bipartite matching-based heuristic for winning bid determination with polynomial time complexity for worker-task assignment.

6.3.2 Winning Bid Selection using Weighted Bipartite Matching

The *WiBS* problem formulation in Eq. (6.3) is an extension of one-to-one weighted matching. However, this matching has to select minimum weighted edges for task allocation with energy budget constraints for V2G tasks. Therefore, we hereby develop a heuristic inspired by bipartite minimum weighted matching which can be solved in polynomial time using Karp’s algorithm [153]. To satisfy the energy budget constraint, we employ iterative matching that removes the highest weighted edges from the previous matching until the budget is met. Simply put, the algorithm runs the minimum weighted matching and if it does not satisfy the budget constraints, removes first z highest weighted edges connected to non-V2G tasks from the previous matching and then runs another round of matching until the feasible solution is found.

This *Bipartite Matching-based Winner selection (BMW)* algorithm is presented in Alg. 8. *BMW* takes set of available workers \mathcal{W} , tasks \mathcal{S} , and the set of bids \mathcal{B} as input and finds the winning bids with final pay P as the output. In line 1, the algorithm initializes the output graph Φ_{out} , a temporary graph Φ_{temp} for iterative matching purpose, and P . Then it creates a separate sets for V2G and non-V2G tasks as sets V and R in line 2 and collects the bids from all workers (line 3). With the information on bids, *BMW* generates a bipartite graph G between bipartite sets of workers \mathcal{W} and tasks \mathcal{S} , and adds edges between those nodes that have non-zero bids i.e. worker w_i with non-zero bid b_{ij} is connected with task s_j (lines 4 – 7). Now, it runs a bipartite matching iteratively with while loop in lines 8 – 15. Initially, both of the conditions for while loop are true and therefore the algorithm runs first round of Minimum Weighted Bipartite Matching on graph G (line 9). It then assigns the matched graph to the output graph Φ_{out} (line 10) and checks if the energy budget for V2G tasks is satisfied (line 11). If it is met in the first round, it breaks out of the while loop and determines final payment and task assignment. If it is not met, *BMW* removes the first z highest weighted edges in Φ_{out} from G that just meet the remaining

Algorithm 8: Bipartite Matching-based Winner selection (BMW)

Input : Sets of Workers (\mathcal{W}) and Spatial Tasks (\mathcal{S}), Bids (\mathcal{B})
Output: Winning bids with final pay (\mathbf{P})

```

/* Initialization */
1  $\Phi_{out} = \{\mathcal{W} \cup \mathcal{S}, E_{\Phi} = \emptyset\}; \Phi_{temp} = \emptyset; P = \emptyset;$ 
/* Generate bipartite graph  $G$  */
2  $\forall s_j \in \mathcal{S}$ , if  $g(z_j) = 1$  then  $V \leftarrow \{s_j\}$  else  $R \leftarrow \{s_j\};$ 
3  $\forall w_i \in \mathcal{W}$ , collect their respective bids  $\mathcal{B}_i$ ;
4 Build Bipartite Graph  $G = \{\mathcal{W} \cup \mathcal{S}, E_G = \emptyset\};$ 
5 for each  $w_i \in \mathcal{W}, s_j \in \mathcal{S}$  do
6   | if  $b_{ij} > 0$  then Add edge  $(w_i, s_j)$  to  $E_G$  with weight,  $b_{ij}$ ;
7 end
/* Run minimum weighted bipartite matching until termination */
8 while  $\sum_{(w_i, s_j) \in E_{out}} g(z_j)l_{ij} < \mathcal{E}$  or  $\Phi_{temp} \neq \Phi_{out}$  do
9   |  $E_{out} \leftarrow$  Perform Minimum Weighted Bipartite Matching on  $G$ ;
10  | Output graph  $\Phi_{out} = \{\mathcal{W} \cup \mathcal{S}, E_{out}\}$ , where  $E_{out} \subseteq E_G$ ;
    | /* Remove edges if V2G energy budget is not met, and run MWBM on reduced
    |     $G$  again */
11  | if  $\sum_{(w_i, s_j) \in E_{out}} g(z_j)l_{ij} < \mathcal{E}$  then
12  |   |  $Z \leftarrow$  Select the first  $z$  highest weight edges  $\in \Phi_{out}$  and  $R$  s.t.
    |   |  $\left( \sum_{(w_i, s_j) \in E_{out}} g(z_j)l_{ij} + \sum_{(w_i, s_j) \in Z} l_{ij} \right) \geq \mathcal{E};$ 
13  |   | if  $Z \neq \emptyset$  then Remove all edges  $\in Z$  from  $G$  and  $\Phi_{out}$  else  $\Phi_{temp} = \Phi_{out};$ 
14  | end
15 end
16  $\mathbf{q}^* = E_{out};$ 
/* Final Payment and Task Assignment */
17  $\forall w_k \in \mathcal{W}, P_k \leftarrow$  Second to the selected bid  $b_{kj}$ ;
18 Assign the tasks to winning workers along with final price  $\mathbf{P}$ ;

```

of energy budget not met (line 12 – 13). Then, since both of the conditions are still true, the algorithm runs another round of matching on reduced graph G . Eventually the final matching in output graph Φ_{out} is used as winning task assignments with final payment as per the bid (line 16 – 18).

Theorem 8. *The time complexity of the BMW algorithm is $O(|\mathcal{W}| \cdot |\mathcal{S}|^2 \cdot \log(|\mathcal{S}|))$.*

Proof. The complexity is dominated by the *while* loop (lines 10 – 17), executed at max $|\mathcal{S}|$ times. It involves running minimum weighted full matching as presented in [153], which has run time of $O(|\mathcal{W}| \cdot |\mathcal{S}| \cdot \log(|\mathcal{S}|))$. Therefore, the overall complexity of the BMW is $O(|\mathcal{W}| \cdot |\mathcal{S}|^2 \cdot \log(|\mathcal{S}|))$. \square

6.4 Experiment

In this section, we present the experimental details for the proposed system, comparison approaches and detailed study of performance of the algorithms.

6.4.1 Experimental Setup

Our experimental setup consists of modeling workers, tasks and the simulation platform. In case of workers, we gathered the publicly available data on 54 different EV models on battery size, range, charging power and charging speed, and formulated an individual profile for each EV in concern. Similarly for ride-sharing tasks, the high volume taxi trip data of New York City (NYC) from the year of 2013 [154] was used. The V2G tasks were generated from the 15 minutes energy consumption data from 25 NYC residences from PecanStreet [148]. In absence of real dataset on battery swapping tasks, half of the ride-sharing tasks were extracted as the battery swapping tasks, given their similar profile with batteries transported instead of passengers. These tasks are spatial, therefore, we collect the information on locations, distance, and time required to complete the tasks.

Furthermore, the simulation platform, e-Uber for crowdsourcing is developed using Python and Gurobi, NetworkX, and PyTorch libraries. We consider a reverse auction

period resolution of 15 minutes which corresponds to the standard set by grid for energy trading. This means that every 15 minutes the e-Uber algorithm will gather the tasks, push the personalized list of tasks to workers, collect the bids and assign the tasks to EV workers that minimizes the overall cost for the task requesters. We set the search radius for the tasks $\lambda = 10$ km and the maximum length of recommendation list $K = 5$. The energy budget for each 15 minutes time period was considered to be total of all 25 V2G tasks available. The user preferences were sampled uniformly from the set $\{0.1, 0.4, 0.5, 0.7, 0.9, 1.0\}$. The energy, time and location of the EVs are tracked and updated accordingly so as to simulate their real-world trip behavior. If the battery level of the cars fall below minimum level, they are considered for the charging for the next time-step.

For comparison approach, we use the task-centric winner selection algorithm as presented in [123] and refer it as *BG* for baseline greedy. This approach neither considers user-preference in the problem-setting nor it considers the personalized recommendation system. So for comparison purpose, we augment this method with perfect knowledge-based recommendation system that pushes K best tasks as recommendation to each workers. Then we implement the algorithm as presented in [123] that sorts the bids from lowest to highest for each tasks and assigns them one by one. Note that this approach may not guarantee a complete matching between workers and tasks as the tasks that are processed towards the end may not have any workers left to choose from because of limited number of bids and greedy selection approach. We use this *BG* as our baseline and compare the performance of our algorithms *CARS* and *BMW* along with their perfect knowledge variation *PK* which has the perfect knowledge on the worker preferences and thus do not involve learning, and *OPT* optimal solution to *WiBS* problem. The ride-share dataset in concern consists of actual ride-fare for specific car. However, we require bids from each vehicle for recommended tasks and a realistic model for bid generation is quite

difficult to obtain. Therefore we trained a Deep Neural Network with existing dataset for determining the ride fare of the given ride-sharing tasks, the details of which is presented in the following.

6.4.2 Results

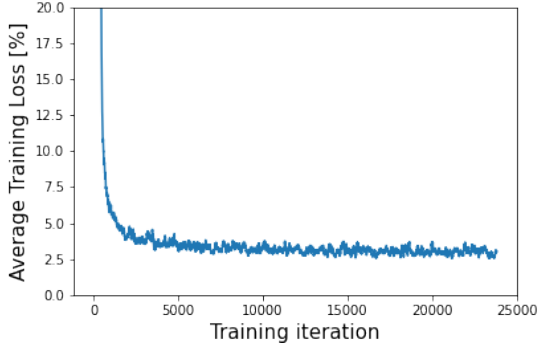


Figure 6.3: Training Loss %

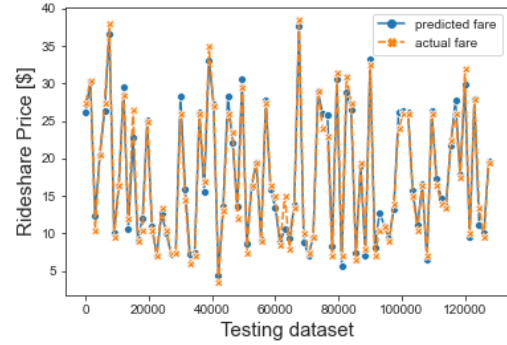


Figure 6.4: Bid Prediction test accuracy

Bid Generation DNN Model

We used 11 months of taxi data to train and test the DNN model with 80-20 train-test split. The DNN model consisted of 3 hidden layers of sizes (132, 132, 64). We employed ReLU activation function as well as one-hot encoding for the input features, and set the learning rate to 0.0001. The training was carried out for 3 epochs with 7974 training batches and batch size of 64. Consequently, the average training loss curve presented in Fig. 6.3, shows that the loss percentage reduces to $\sim 2.5\%$ after $\sim 12,000$ training iteration. On testing dataset, the bid generation DNN model, generated highly accurate fare prediction with 96.45% R^2 -score. This can also be observed in Fig. 6.4 which presents a plot of sample of prediction fares and actual fares to show testing accuracy.

This DNN model was then deployed in conjunction with the e-Uber to simulate the bidding action by each workers for each recommended tasks in the personalized list. In case of V2G tasks, the energy to be supplied by the EV was converted into its distance equivalent and fed into the DNN model along with other input features

to get the bids.

Experimental Observations

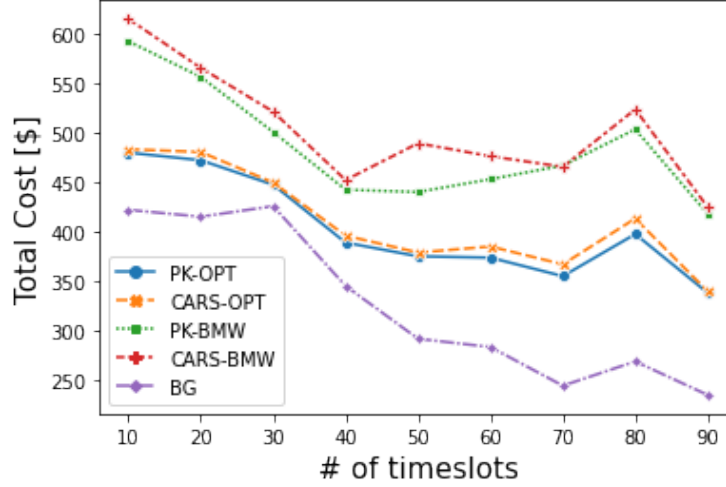


Figure 6.5: Snapshot of obj. values & matches vs. time

1. **Performance over time – Total Cost and # of Tasks:** In the first experimental scenario, we observe the performance of algorithms as a snapshot of objective values over 24 hours (i.e. $24 \times 4 = 96$ timeslots). We present the objective values from midnight to next midnight as a lineplot in Fig. 6.5 and cumulative bar plots of objective values (Fig. 6.6) and total tasks completed (Fig. 6.7) over a day. Although all the proposed approaches start from the same initial state (except for knowledge on preference), these algorithms may have different successive states since the solution is affected by the matching in previous timeslot, availability of specific workers for next round, and the distance travelled by these workers for previous assignment (or next assignment). Therefore, we employ cumulative objective values and cumulative tasks completed as the metric for a fair comparison of the approaches in Fig. 6.6. This cumulative objective value reflects the overall quality of task assignment made so far based on the total objective values to achieve the requirement while the cumulative tasks completed present the total number of matches made by the respective approach until the end of that timeslot.

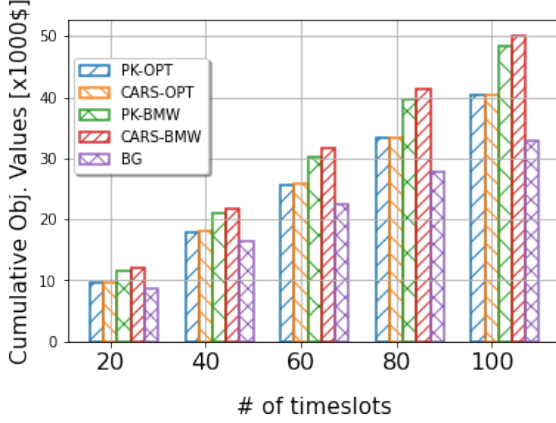


Figure 6.6: Cumulative obj. values

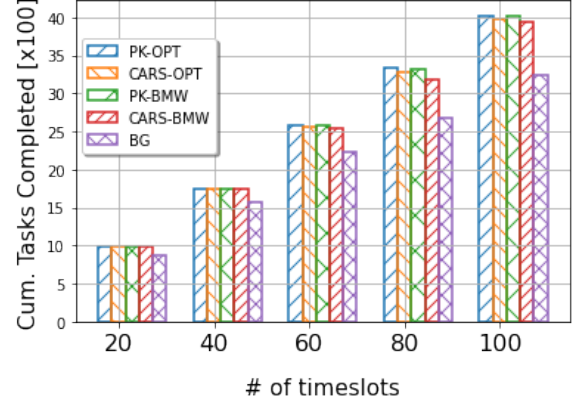


Figure 6.7: Cumulative tasks

As seen in the lineplot Fig. 6.5 and barplot Fig. 6.6, the solution generated by baseline greedy approach BG is the minimum one as it assigns task based on respective cheapest bid available but it doesn't meet the maximum number of matching possible unlike other approaches as shown in Fig.6.7. Therefore, BG mostly violates the V2G requirement, meaning it generates infeasible solutions and hence fails for this problem setting. The $PK - OPT$ produces the best result since it involves solving the $POTR$ and $WiBS$ problem optimally with perfect knowledge of the worker preferences. Following it, is the optimal solution OPT paired with our proposed learning framework for e-Uber, $CARS$, which performs close to optimal in terms of both objective values and number of tasks completed. Although this approach $CARS - OPT$ finds optimal solution, it does not have initial knowledge on preferences. Therefore, it generates sub-optimal recommendation list which then affects the solution to $WiBS$ problem and hence, the overall performance. However, even with online learning framework employed, it produces similar results to the $PK - OPT$. Also we observe similar pattern with $PK - BMW$ and $CARS - BMW$ since they both rely on bipartite matching-based approach to find feasible solution. Since $PK - BMW$ sends the optimal recommendation to workers for collecting bids, it therefore has higher overall performance compared to $CARS - BMW$ which learns

the preferences over time. The gaps between best performing $PK - OPT$ and worst performing $CARS - BMW$ however is less than \$150 which amounts to a price hike of $\sim \$3/\text{task}$ in the worst case with an average 50 tasks for a timeslot as in our case. We observe the cumulative objective values grow almost linearly for all approaches and as expected, the performance observed was better for $PK - OPT$ followed by $CARS - OPT$ and then $PK - BMW$ and finally $CARS - BMW$. However, the gap in cumulative objective value increased for the bipartite heuristic compared to optimal due to its sub-optimal performance. Note that the baseline BG generates less cumulative objective value but it fails to generate maximal matching as seen in Fig. 6.7. The number of tasks completed by the proposed approaches exceed 850 more than the BG in the span 24 hours.

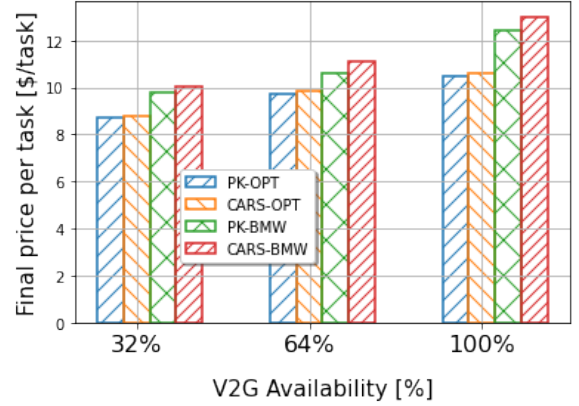
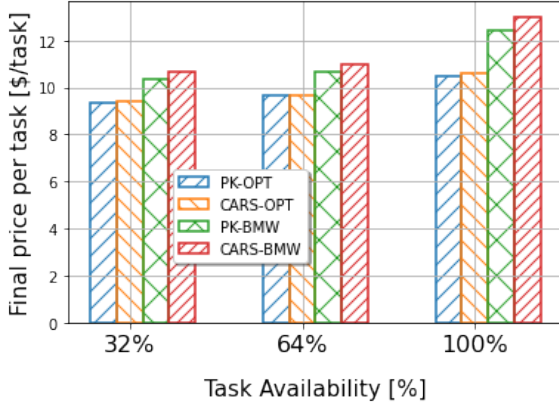


Figure 6.8: Avg. Price/task vs. Task(%) Figure 6.9: Avg. price/task vs. V2G (%)

2. **Average final price per task and scaling:** In this experiment, we track the average final price per task while scaling the available tasks from 32% to 64% and then at 100%. For scaling the tasks, we increase the number of each type of tasks proportionally. The result is plotted in Fig. 6.8. As the system scales, the average final price per task for all approaches rises since the overall cost for the system also increases with the tasks. However, it is also observed that $CARS - BMW$ and $BMW - PK$ suffer more as we scale the system. The margin between these and optimal approaches grows drastically up to $\sim \$2$. This can be attributed mainly to

the increased complexity of the problem as number of tasks is increased and hence the bipartite matching-based heuristic finds less efficient solution compared to optimal. The optimal solutions however have nominal increase in their average price per task ($\sim \$10$) even with scaling compared to rest.

We also study the effect of scaling V2G tasks to the average final price per task in Fig. 6.9. We observed similar trend to above but with noticeable gap between optimal and heuristic approaches when only 32% of V2G tasks are available. This results from the sub-optimal performance owing to less number of V2G tasks compared to rest and hence unequal rate of learning the preferences.

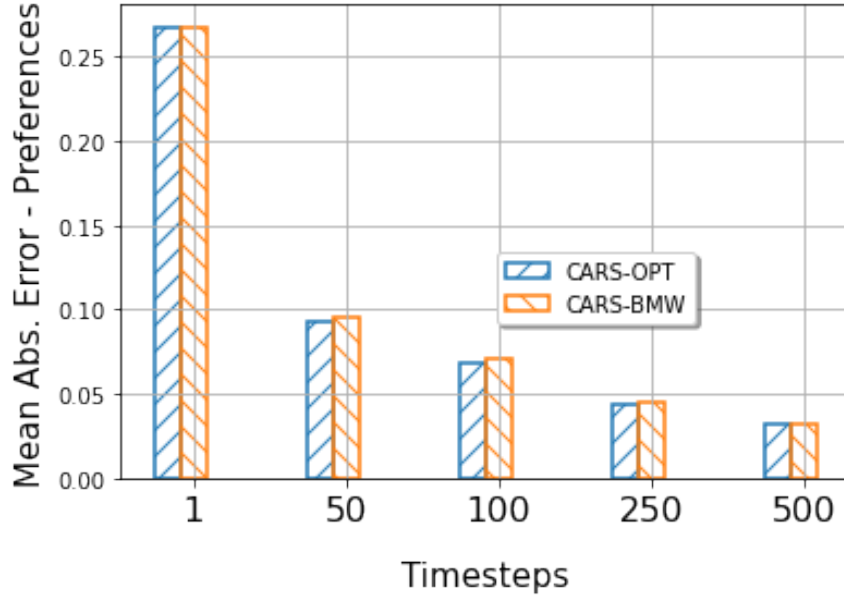


Figure 6.10: Mean Absolute Error vs. time

3. Learning accuracy for preferences – MAE: To study the quality of proposed CMAB-based learning algorithm *CARS* in conjunction with optimal and *BMW*, we use the Mean Absolute Error (MAE) of the learned preferences over time and present them in Fig. 6.10. Both approaches use same learning algorithm but the solution to *WiBS* problem differs and thus affects the learning performance. However, this difference is very negligible. Initially, the MAE is 0.28 and then rapidly decreases

to less than 0.05 for both approaches by 250 timesteps. The difference in learning efficacy between $CARS - OPT$ and $CARS - BMW$ reduces over the time and is almost same by 250 timesteps as seen in the graph. Since by 500 timesteps the system has garnered sufficient knowledge on workers preferences, MAE falls to 0.03 reflecting the efficacy of proposed CMAB-based preference learning. Furthermore, we present a cumulative reward plot in Fig. 6.11 that also shows the plots of both learning approaches converge after 200 timesteps.

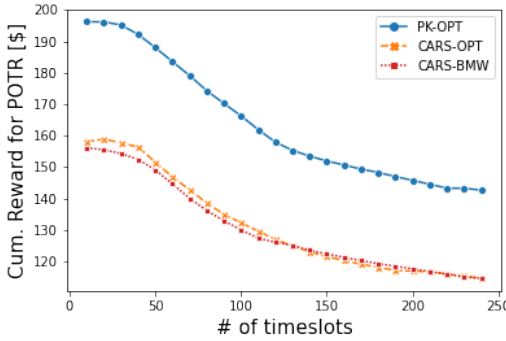


Figure 6.11: Cumulative reward plot

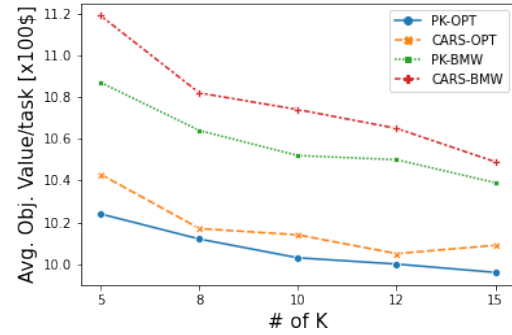


Figure 6.12: Obj. values/matching vs. K

4. **Dependency with K** : In this experiment, we discuss on the dependency of the performance of our proposed approach with recommendation length K , as presented in Fig. 6.12. Increasing the number of recommendation K means that the chance of receiving more bids with good quality from same number of workers at the same time increases. This in turn helps to find better solutions which reduce overall cost of the system. This is also verified from the observation in plot of Fig. 6.12. As we increase K , the objective values per task over a day's period reduces for all four approaches. Although the perfect and optimal optimal methods do not have significant difference in their performance with varied K , the effect is more pronounced in case of bipartite matching based $PK - BMW$ and $CARS - BMW$ where the learning of preferences is benefited by the increased number of bids to choose from with increasing K . However, it needs to be noted that pushing 10 recommends at each timestep can be very

intractable for workers and therefore, keeping the length of recommendation list as small as possible is desired.

6.5 Conclusion

e-Uber is a promising crowdsourcing platform for improving the efficiency and sustainability of ride-sharing and energy-sharing services through the use of EVs. It uses reverse auction mechanism to assign spatial tasks to EV drivers based on their preferences, battery level, and other realistic constraints like minimum energy requirement for grid and one-to-one assignment. To optimize the task recommendation process, the platform incorporates user behavioral models including worker preferences and bounded rationality. However, as these preferences are not known *a priori*, e-Uber uses reinforcement learning framework called combinatorial multi-armed bandit for learning the preferences at the runtime based on their feedback. We propose the *CARS* algorithm that finds optimal solution to both the *POTR* and *WiBS* problem. Since the *WiBS* problem is NP-hard, we propose another bipartite matching-based heuristic, called *BMW* that finds feasible solution to the winner selection while meeting the minimum V2G energy requirement. Experimental results and simulations demonstrate the effectiveness of e-Uber’s approaches, which outperform the baseline algorithm by serving more than 850 tasks within 24 hours of simulation. On top of that, the baseline often fails to find a feasible solution, rendering it inapplicable in this problem setting.

CHAPTER 7. CONCLUSION AND FURTHER RESEARCH

Given the increasing prosumer involvement in energy exchange process and limitations of existing energy market for utilizing excess energy at distribution side in recent years, this dissertation proposed an alternative modality of prosumer-centric P2P energy market. In this dissertation, a detailed discussion is presented on how such a localized energy exchange among prosumers provides a better way to address the problem at local level while also providing the prosumers with a platform to financially benefit from. Additionally, through the in-depth study of existing literature it was established that this market modality can prove beneficial to all the stakeholders in the energy market in long term. From the monetary incentives for local prosumers to minimizing the infrastructure cost and energy losses in transmission lines that is highly significant in existing power system architecture from the grid operators's perspective, the proposed P2P modality aims to benefit everyone involved in such energy market. Therefore, P2P energy trading market has a potential to transform the energy landscape towards a decentralized and open platform that allows energy exchanges among all the stakeholders with a profitable energy trading modality.

However, this modality also requires sustained participation from prosumers side, which might be overwhelming for prosumers owing to their limited time and cognitive capabilities. Through exhaustive literature review, we found that there is a lack of practical and realistic studies that could accommodate the behavior and perceptions of prosumers to accurately reflect their risk-sensitivity, loss-aversion character and preferences in a less demanding and flexible energy trading environment. Therefore, to address this problem, prosumer-centric framework was developed for the P2P trading system that is expected to be user convenient and less demanding of their active participation, through the incorporation of user behavioral modeling into the problem. For that reason, this work focused on developing a reliable and realistic

P2P energy trading mechanism that maximizes the local energy consumption among prosumers with due consideration to prosumers' preferences, energy trading behavior and perceived utility. With incorporation of user behavioral modeling and learning of this model through minimal active involvement from prosumers, the automated trading environment succeeded in reflecting the individual optimal trading behavior for all prosumers in the P2P energy trading setting.

7.1 Main Contribution

In chapter 4, a P2P energy exchange mechanism has been proposed that learns the user preference through daily interaction with prosumers. It also considers the bounded rationality in problem formulation that prevents the user from getting overwhelmed with the recommendations. Two algorithms for determining the energy allocation between prosumers is proposed that adopts reinforcement learning approach to simultaneously learn the user preference while finding the best match between sellers and buyers on the basis of their current estimate on preferences.

Similarly, moving a step forward, chapter 5 includes monetary incentives into the problem and presents fully automated mechanism to carry out energy trading between local prosumers in a P2P modality. The perceived utility of prosumers are captured through the use of novel-prize winning behavioral economics notion called prospect theory into the energy allocation framework between prosumers. A population-based metaheuristic algorithm is proposed to solve this non-linear energy allocation problem. A fully automated pricing mechanism for the P2P energy trading is proposed that dynamically updates the individual seller's selling price based on the feedback on the sold quantity of energy and profit made. Two novel reinforcement learning algorithms are proposed that take inspiration from standard Q-learning and Deep Q-Network approach by incorporating the risk-sensitivity and loss-aversion behavior to the RL framework.

In chapter 6, we propose a crowdsourcing mechanism that jointly enables ride- and

energy-sharing to provide a multifaceted solution to existing problems on efficiency and sustainability of transportation, energy management, and cost-effective demand response using EVs. *e-Uber* is the first work in that direction to the best of our knowledge. It works in three decision stages: calculate a personalized task recommendation for each EV worker, collect bids from workers, reverse auction-based winning bids selection for completing both ride-sharing and energy-sharing tasks. We propose a preference-aware optimal task recommendation system, and the reinforcement learning mechanism based on the work presented in chapter 4 to learn worker’s bidding probabilities. It solves the problem of task recommendation and updates the worker preferences based on their interaction with the recommendation. The reverse auction process is formalized for bidding and the winning bids are determined through an NP-hard optimization framework, and propose a bipartite matching-based heuristic for the problem.

To summarize, this dissertation has established that the P2P energy trading allows flexible and efficient way for individuals and small business with energy generation capabilities to sell excess energy directly to other consumers and generate revenue. In the context of P2P energy trading, as shown by the result and discussion in this dissertation, user behavioral modeling is significant to predict and influence how individuals will participate in the P2P market, and also to develop strategies for optimizing and sustaining their participation. It has the potential to increase the efficiency and sustainability of electricity markets while also providing newer opportunities and alternatives for retail and shared economy in changing energy landscape.

7.2 Further Research

In the following section, we shed light on some of the limitations of the works presented in this dissertation and future avenues for further research. It should be noted that the problem of devising a realistic, automated and decentralized P2P

energy trading market is quite challenging and this dissertation is just a nudge in that direction. Accommodating these limitations and adapting to the future need of changing energy market could be extensions or improvements to the framework presented in this dissertation or completely new approach to achieve localized energy exchange in a decentralized way.

7.2.1 Modeling complex dependencies of user behavioral parameters

The user behavior is subjective and qualitative trait that is guided by human psychology. The assumptions in quantifying these user behavioral parameters could have multilateral dependencies than what is apparent. For example, user preferences in P2P energy market setting could be shaped by sources of energy, trading price, proximity and many more. Similarly the degree of risk-sensitivity and loss-aversion considered in prospect theory model could differ by multitude of factors and scenarios including age, gender, way of living, and context. On top of that, user behavior is dynamic and capturing time-dependency into the model is necessary evil. Thus, in devising the P2P market models, these complex and inter-weaved dependencies of human behavior has to be modeled into the system in a way that does not cross the ethical boundaries. This will definitely improve the efficacy and efficiency of intelligent systems in P2P trading markets in general including energy market. Therefore, considering a more complex model for user behavior that reflects the irrational and variable nature of human decision in a holistic manner would be an interesting research direction for the future.

7.2.2 Fully decentralized energy allocation

It needs to be noted that although the works presented in this dissertation is "peer-to-peer" in terms of general mechanism and framework, the energy allocation solutions proposed is approached in centralized way. In order to render it fully decentralized, the work in this dissertation can be extended through a decentralized protocol to generate consensus in allocating energy between sellers and buyers in an iterative

way. The protocol should also consider the combinatorial nature of problem setting and should not demand active participation of prosumers for sustainability of energy market. Of course this brings a range of other economic, strategic and policy-making problems that could interestingly be researched further through the lenses of different stakeholders in the market.

7.2.3 Reward functions for RL frameworks

In reinforcement learning, the reward signal and function plays a pivotal role in overall learning process which acts as a quantitative feedback to the agent on the actions taken and therefore learn the best policy to adopt over long term. The reward functions considered in the RL frameworks in chapters 4 and 5 could be insufficient when implemented in real world and therefore room for improvement always exist in that regards and specially for deep reinforcement learning frameworks which can overshoot towards completely wrong trajectory. As an example a better reward function that also incorporates other features like excess energy generation and energy demand for next timestep, and battery energy storage level could guide the RL agent towards improved learning. Hence, determining the reward signals and functions that allow agents to converge faster towards the optimal policy is obviously an immediate next step to extend the works in this dissertation.

7.2.4 Including energy storage and electric vehicles in P2P

P2P energy trading will be incomplete without including energy storage systems into concern. It could range from standard chemical battery packs to flywheel or compressed air energy storage to fuel cells like hydrogen. More flexible and mobile energy storage solution however can be powered through EVs that provide us with several novel possibilities to solve demand response and energy-sharing paradigms including Vehicle-2-Grid (V2G), Vehicle-2-House (V2H), and Vehicle-2-Vehicle (V2V) among others. In chapter 6, we discussed in detail about how the EVs can be realized in a P2P energy trading setting. Future research could focus on implementing and

evaluating e-Uber in real-world settings. This includes the assessment of the impact of different task recommendation and decision prediction algorithms, as well as the integration of new features such as real-time traffic and energy data and dynamic pricing. By exploring these areas, e-Uber has the potential to significantly improve the efficiency and sustainability of ride-sharing and energy-sharing services through the use of EVs. As a stand-out alternative medium of transportation in immediate future, EVs could be a pivotal enabler to bring the whole P2P energy market modality into accepted reality and therefore further research in this direction would not only be sensible but also appropriate and timely.

7.2.5 Blockchain technology for practical implementation

Implementation of such decentralized energy trading modalities in real world setting could lead to numerous privacy concerns and safety issues of the participants. Therefore, a proper way to address this privacy concern must accompany the practical implementation. This is where the blockchain technology could play a facilitating role to enable P2P energy trading in future. Privacy concerns can be easily addressed through the use of blockchain in addition to providing a secure trading platform for participating prosumers in P2P trading. Therefore, augmenting the decentralized P2P energy trading mechanism with blockchain technology would provide a secure and privacy-preserving platform to conduct financial transactions, which serves as the future research scope.

BIBLIOGRAPHY

- [1] Y. Parag and B. Sovacool, “Electricity market design for the prosumer era,” Nature Energy, vol. 1, p. 16032, March 2016.
- [2] “International energy agency.” <https://www.iea.org/reports/electricity-information-overview/electricity-production>.
- [3] “Energy information administration.” <https://www.eia.gov>.
- [4] “Renewable energy capacity highlights.” <https://www.irena.org/publications/2022/Apr/Renewable-Capacity-Statistics-2022>.
- [5] 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.
- [6] 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.
- [7] 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, p. 2106, 2017.
- [8] 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, pp. 231–254, July 2017.
- [9] W. Strielkowski, Social Impacts of Smart Grids: The Future of Smart Grids and Energy Market Design. Elsevier, 2019.
- [10] M. Nasimifar, V. Vahidinasab, and M. S. Ghazizadeh, “A peer-to-peer electricity marketplace for simultaneous congestion management and power loss reduction,” in 2019 Smart Grid Conference (SGC), pp. 1–6, IEEE, 2019.
- [11] “Freeing the grid: Best and worst practices in state netmetering policies and interconnection procedure,” 2009. <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, pp. 67–72, ACM, 2011.
- [13] 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.

- [14] 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, vol. 11, no. 4, pp. 3185–3200, 2020.
- [15] W. Tushar, T. K. Saha, C. Yuen, P. Liddell, R. Bean, and H. V. Poor, "Peer-to-peer energy trading with sustainable user participation: A game theoretic approach," IEEE Access, vol. 6, pp. 62932–62943, 2018.
- [16] 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.
- [17] G. El Rahi, W. Saad, A. Glass, N. B. Mandayam, and H. V. Poor, "Prospect theory for prosumer-centric energy trading in the smart grid," in 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5, IEEE, 2016.
- [18] U.S. Department of Energy, "The smart grid: An introduction," tech. rep., U.S. Department of Energy, November 2008.
- [19] 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.
- [20] 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.
- [21] "Brooklyn micogrid." <https://www.brooklyn.energy>.
- [22] "Vandebron." <https://vandebron.nl>.
- [23] "Piclo energy." <https://piclo.energy>.
- [24] 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, pp. 3569–3583, Sep. 2017.
- [25] S. Mullainathan and R. H. Thaler, "Behavioral economics," 2000.
- [26] H. A. Simon, Models of man; social and rational. Wiley, 1957.
- [27] D. Kahneman, "Maps of bounded rationality: Psychology for behavioral economics," American Econ. Rev., vol. 93, no. 5, pp. 1449–1475, 2003.
- [28] G. Gigerenzer and R. Selten, Bounded rationality: The adaptive toolbox. MIT press, 2002.

- [29] A. Szollosi and B. R. Newell, “People as intuitive scientists: Reconsidering statistical explanations of decision making,” Trends in Cognitive Sciences, 2020.
- [30] P. E. Earl, “Bounded rationality in the digital age,” in Minds, Models and Milieux, pp. 253–271, Springer, 2016.
- [31] D. E. Agosto, “Bounded rationality and satisficing in young people’s web-based decision making,” Journal of the American society for Information Science and Technology, vol. 53, no. 1, pp. 16–27, 2002.
- [32] A. Krajnović, D. Sikirić, and J. Bosna, “Digital marketing and behavioral economics,” CroDiM: International Journal of Marketing Science, vol. 1, no. 1, pp. 33–46, 2018.
- [33] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. Saha, H. V. Poor, and K. L. Wood, “Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches,” IEEE Signal Processing Magazine, vol. 35, no. 4, pp. 90–111, 2018.
- [34] M. I. Azim, S. Pourmousavi, W. Tushar, and T. K. Saha, “Feasibility study of financial p2p energy trading in a grid-tied power network,” in 2019 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5, IEEE, 2019.
- [35] 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), pp. 219–228, April 2013.
- [36] A. Paudel, L. Sampath, J. Yang, and H. B. Gooi, “Peer-to-peer energy trading in smart grid considering power losses and network fees,” IEEE Transactions on Smart Grid, vol. 11, no. 6, pp. 4727–4737, 2020.
- [37] A. Hariharasudan, I. Otolu, and Y. Bilan, “Reactive power optimization and price management in microgrid enabled with blockchain,” Energies, vol. 13, no. 23, p. 6179, 2020.
- [38] Y. Liu, C. Sun, A. Paudel, Y. Gao, Y. Li, H. B. Gooi, and J. Zhu, “Fully decentralized p2p energy trading in active distribution networks with voltage regulation,” IEEE Transactions on Smart Grid, 2022.
- [39] W. Tushar, T. K. Saha, C. Yuen, T. Morstyn, H. V. Poor, R. Bean, et al., “Grid influenced peer-to-peer energy trading,” IEEE Transactions on Smart Grid, vol. 11, no. 2, pp. 1407–1418, 2019.
- [40] 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), pp. 427–436, IEEE, 2020.

- [41] A. D. Rathnayaka, V. M. Potdar, O. Hussain, and T. Dillon, "Identifying prosumer's energy sharing behaviours for forming optimal prosumer-communities," in 2011 International Conference on Cloud and Service Computing, pp. 199–206, IEEE, 2011.
- [42] 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.
- [43] 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. 137063–137073, 2020.
- [44] Y. Yao, C. Gao, T. Chen, J. Yang, and S. Chen, "Distributed electric energy trading model and strategy analysis based on prospect theory," International Journal of Electrical Power & Energy Systems, vol. 131, p. 106865, 2021.
- [45] A. Timilsina, A. R. Khamesi, V. Agate, and S. Silvestri, "A reinforcement learning approach for user preference-aware energy sharing systems," IEEE Transactions on Green Communications and Networking, 2021.
- [46] A. Timilsina, "P2p energy trading in a smart residential environment with user behavioral modeling," in 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, IEEE, 2023.
- [47] A. Timilsina and S. Silvestri, "Prospect theory-inspired automated p2p energy trading with q-learning-based dynamic pricing," in GLOBECOM 2022-2022 IEEE Global Communications Conference, pp. 4836–4841, IEEE, 2022.
- [48] A. Timilsina and S. Silvestri, "P2p energy trading through prospect theory, differential evolution, and reinforcement learning," ACM Transactions on Evolutionary Learning and Optimization, 2023.
- [49] R. Alden, A. Timilsina, S. Silvestri, and D. M. Ionel, "V2g optimization for dispatchable residential load operation and minimal utility cost," in 2023 IEEE International Transportation Electrification Conference, IEEE, 2023.
- [50] A. Timilsina and S. Silvestri, "*e-Uber*: A crowdsourcing platform for electric vehicle-based ride-and energy-sharing," arXiv preprint arXiv:2304.04753, 2023.
- [51] 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.

- [52] M. Shurrab, S. Singh, H. Otrók, R. Mizouni, V. Khadkikar, and H. Zeineldin, "An efficient vehicle-to-vehicle (v2v) energy sharing framework," IEEE Internet of Things Journal, vol. 9, no. 7, pp. 5315–5328, 2021.
- [53] M. R. Sarker, H. Pandžić, and M. A. Ortega-Vazquez, "Optimal operation and services scheduling for an electric vehicle battery swapping station," IEEE transactions on power systems, vol. 30, no. 2, pp. 901–910, 2014.
- [54] T. Ackermann, G. Andersson, and L. Söder, "Distributed generation: a definition," Electric Power Systems Research, vol. 57, no. 3, 2001.
- [55] P. Asmus, "Microgrids, virtual power plants and our distributed energy future," The Electricity Journal, vol. 23, no. 10, pp. 72–82, 2010.
- [56] 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.
- [57] 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.
- [58] 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, pp. 1386–1397, July 2014.
- [59] 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, pp. 280–292, Jan 2019.
- [60] 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), pp. 1–7, IEEE, 2020.
- [61] E. A. Soto, L. B. Bosman, E. Wollega, and W. D. Leon-Salas, "Peer-to-peer energy trading: A review of the literature," Applied Energy, vol. 283, p. 116268, 2021.
- [62] E. Mengelkamp, J. Gärttner, K. Rock, S. Kessler, L. Orsini, and C. Weinhardt, "Designing microgrid energy markets: A case study: The brooklyn microgrid," Applied Energy, vol. 210, pp. 870–880, 2018.
- [63] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng, and Y. Zhang, "Consortium blockchain for secure energy trading in industrial internet of things," IEEE transactions on industrial informatics, vol. 14, no. 8, pp. 3690–3700, 2017.

- [64] J. Kang, R. Yu, X. Huang, S. Maharjan, Y. Zhang, and E. Hossain, “Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains,” IEEE Transactions on Industrial Informatics, vol. 13, no. 6, pp. 3154–3164, 2017.
- [65] T. Morstyn, A. Teytelboym, and M. D. McCulloch, “Bilateral contract networks for peer-to-peer energy trading,” IEEE Transactions on Smart Grid, vol. 10, no. 2, pp. 2026–2035, 2018.
- [66] L. Boratto and E. Vargiu, “Data-driven user behavioral modeling: from real-world behavior to knowledge, algorithms, and systems,” Journal of Intelligent Information Systems, vol. 54, no. 1, pp. 1–4, 2020.
- [67] S. Angeletou, M. Rowe, and H. Alani, “Modelling and analysis of user behaviour in online communities,” in International semantic web conference, pp. 35–50, Springer, 2011.
- [68] I. Zukerman and D. W. Albrecht, “Predictive statistical models for user modeling,” User Modeling and User-Adapted Interaction, vol. 11, no. 1, pp. 5–18, 2001.
- [69] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” in Handbook of the fundamentals of financial decision making: Part I, pp. 99–127, World Scientific, 2013.
- [70] N. Wilkinson and M. Klaes, An introduction to behavioral economics. Macmillan International Higher Education, 2017.
- [71] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
- [72] V. François-Lavet, P. Henderson, et al., “An introduction to deep reinforcement learning,” arXiv preprint arXiv:1811.12560, 2018.
- [73] F. Restuccia, P. Ferraro, T. S. Sanders, S. Silvestri, S. K. Das, and G. L. Re, “First: A framework for optimizing information quality in mobile crowdsensing systems,” ACM Transactions on Sensor Networks (TOSN), vol. 15, no. 1, pp. 1–35, 2018.
- [74] F. Restuccia, P. Ferraro, S. Silvestri, S. K. Das, and G. L. Re, “Incentme: Effective mechanism design to stimulate crowdsensing participants with uncertain mobility,” IEEE Transactions on Mobile Computing, vol. 18, no. 7, pp. 1571–1584, 2018.
- [75] W. Ai, T. Deng, and W. Qi, “Crowdsourcing electrified mobility for omni-sharing distributed energy resources,” in AI and Analytics for Smart Cities and Service Systems, pp. 365–382, Springer, 2021.

- [76] 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.
- [77] 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 Trans. on Sustainable Energy, 2020.
- [78] 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.
- [79] 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.
- [80] 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.
- [81] 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.
- [82] F. Plewnia, "The energy system and the sharing economy: Interfaces and overlaps and what to learn from them," Energies, vol. 12, no. 3, p. 339, 2019.
- [83] 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.
- [84] K. A. Melendez, V. Subramanian, T. K. Das, and C. Kwon, "Empowering end-use consumers of electricity to aggregate for demand-side participation," Applied Energy, vol. 248, pp. 372–382, 2019.
- [85] B. A. Bhatti and R. Broadwater, "Energy trading in the distribution system using a non-model based game theoretic approach," Applied Energy, vol. 253, p. 113532, 2019.
- [86] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized p2p energy trading under network constraints in a low-voltage network," IEEE Transactions on Smart Grid, vol. 10, no. 5, pp. 5163–5173, 2018.

- [87] K. Chen, J. Lin, and Y. Song, "Trading strategy optimization for a prosumer in continuous double auction-based peer-to-peer market: A prediction-integration model," Applied energy, vol. 242, pp. 1121–1133, 2019.
- [88] H. Liu, Y. Zhang, S. Zheng, and Y. Li, "Electric vehicle power trading mechanism based on blockchain and smart contract in v2g network," IEEE Access, vol. 7, pp. 160546–160558, 2019.
- [89] C. Long, J. Wu, Y. Zhou, and N. Jenkins, "Peer-to-peer energy sharing through a two-stage aggregated battery control in a community microgrid," Applied energy, vol. 226, pp. 261–276, 2018.
- [90] S. Nguyen, W. Peng, P. Sokolowski, D. Alahakoon, and X. Yu, "Optimizing rooftop photovoltaic distributed generation with battery storage for peer-to-peer energy trading," Applied Energy, vol. 228, pp. 2567–2580, 2018.
- [91] T. Baroche, P. Pinson, R. L. G. Latimier, and H. B. Ahmed, "Exogenous cost allocation in peer-to-peer electricity markets," IEEE Transactions on Power Systems, vol. 34, no. 4, pp. 2553–2564, 2019.
- [92] S. Wang, A. F. Taha, J. Wang, K. Kvaternik, and A. Hahn, "Energy crowdsourcing and peer-to-peer energy trading in blockchain-enabled smart grids," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 8, pp. 1612–1623, 2019.
- [93] C. Zhang, J. Wu, Y. Zhou, M. Cheng, and C. Long, "Peer-to-peer energy trading in a microgrid," Applied Energy, vol. 220, pp. 1–12, 2018.
- [94] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, p. 21260, 2008.
- [95] O. Utility, "A glimpse into the future of britain's energy economy," White Pap, pp. 1–25, 2016.
- [96] S. Silvestri, D. A. Baker, and V. Dolce, "Integration of social behavioral modeling for energy optimization in smart environments," in ACM Social Sense, pp. 97–97, 2017.
- [97] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in AEA Papers and Proceedings, vol. 108, pp. 22–27, 2018.
- [98] S. Barocas and A. D. Selbst, "Big data's disparate impact," Calif. L. Rev., vol. 104, p. 671, 2016.
- [99] S. Wendel, Designing for behavior change: Applying psychology and behavioral economics. O'Reilly Media, 2020.
- [100] 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.

- [101] 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), pp. 208–213, IEEE, 2020.
- [102] T. F. E. R. Commission, "Reports on demand response & advanced metering," tech. rep., The Federal Energy Regulatory Commission, December 2015.
- [103] V. Dolce, C. Jackson, S. Silvestri, D. Baker, and A. De Paola, "Social-behavioral aware optimization of energy consumption in smart homes," in 2018 14th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 163–172, IEEE, 2018.
- [104] W. Tushar, T. K. Saha, C. Yuen, T. Morstyn, M. D. McCulloch, H. V. Poor, and K. L. Wood, "A motivational game-theoretic approach for peer-to-peer energy trading in the smart grid," Applied energy, vol. 243, pp. 10–20, 2019.
- [105] W. Tushar, T. K. Saha, C. Yuen, M. I. Azim, T. Morstyn, H. V. Poor, D. Niyato, and R. Bean, "A coalition formation game framework for peer-to-peer energy trading," Applied Energy, vol. 261, p. 114436, 2020.
- [106] S. Dorahaki, M. Rashidinejad, S. F. F. Ardestani, A. Abdollahi, and M. R. Salehizadeh, "A peer-to-peer energy trading market model based on time-driven prospect theory in a smart and sustainable energy community," Sustainable Energy, Grids and Networks, vol. 28, p. 100542, 2021.
- [107] L. Xiao, N. B. Mandayam, and H. V. Poor, "Prospect theoretic analysis of energy exchange among microgrids," IEEE Transactions on Smart Grid, vol. 6, no. 1, pp. 63–72, 2014.
- [108] 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.
- [109] Y. Wang, W. Saad, N. B. Mandayam, and H. V. Poor, "Load shifting in the smart grid: To participate or not?," IEEE Transactions on Smart Grid, vol. 7, no. 6, pp. 2604–2614, 2015.
- [110] Y. Xia, Q. Xu, Y. Huang, Y. Liu, and F. Li, "Preserving privacy in nested peer-to-peer energy trading in networked microgrids considering incomplete rationality," IEEE Transactions on Smart Grid, vol. 14, no. 1, pp. 606–622, 2022.
- [111] 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.
- [112] 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.

- [113] C. R. Fox and R. A. Poldrack, "Prospect theory and the brain," in Neuroeconomics, pp. 145–173, London, UK: Elsevier, 2009.
- [114] A. Ghasemi, A. Shojaeighadikolaei, K. Jones, M. Hashemi, A. G. Bardas, and R. Ahmadi, "A multi-agent deep reinforcement learning approach for a distributed energy marketplace in smart grids," in 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6, IEEE, 2020.
- [115] R. Leo, R. Milton, and A. Kaviya, "Multi agent reinforcement learning based distributed optimization of solar microgrid," in 2014 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–7, IEEE, 2014.
- [116] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," Machine learning, vol. 49, no. 2, pp. 267–290, 2002.
- [117] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," Neural computation, vol. 26, no. 7, pp. 1298–1328, 2014.
- [118] Y. Zhao, K. Zheng, H. Yin, G. Liu, J. Fang, and X. Zhou, "Preference-aware task assignment in spatial crowdsourcing: from individuals to groups," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 7, pp. 3461–3477, 2020.
- [119] S. Sodagari, "Trends for mobile iot crowdsourcing privacy and security in the big data era," IEEE Transactions on Technology and Society, vol. 3, no. 3, pp. 199–225, 2022.
- [120] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), pp. 344–353, IEEE, 2016.
- [121] Y. Xing, L. Wang, Z. Li, and Y. Zhan, "Multi-attribute crowdsourcing task assignment with stability and satisfactory," IEEE Access, vol. 7, pp. 133351–133361, 2019.
- [122] M. Xiao, K. Ma, A. Liu, H. Zhao, Z. Li, K. Zheng, and X. Zhou, "Sra: Secure reverse auction for task assignment in spatial crowdsourcing," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 4, pp. 782–796, 2019.
- [123] Y. Liu, X. Xu, J. Pan, J. Zhang, and G. Zhao, "A truthful auction mechanism for mobile crowd sensing with budget constraint," IEEE Access, vol. 7, pp. 43933–43947, 2019.

- [124] H. Hong, X. Li, D. He, Y. Zhang, and M. Wang, "Crowdsourcing incentives for multi-hop urban parcel delivery network," IEEE Access, vol. 7, pp. 26268–26277, 2019.
- [125] Y. Wang, J. Jiang, and T. Mu, "Context-aware and energy-driven route optimization for fully electric vehicles via crowdsourcing," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, pp. 1331–1345, 2013.
- [126] Y. He and C. Csiszár, "Model for crowdsourced parcel delivery embedded into mobility as a service based on autonomous electric vehicles," Energies, vol. 14, no. 11, p. 3042, 2021.
- [127] A. Yassine, M. S. Hossain, G. Muhammad, and M. Guizani, "Cloudlet-based intelligent auctioning agents for truthful autonomous electric vehicles energy crowdsourcing," IEEE Transactions on Vehicular Technology, vol. 69, no. 5, pp. 5457–5466, 2020.
- [128] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," IEEE Transactions on Smart Grid, vol. 10, no. 1, pp. 841–851, 2017.
- [129] P. E. Greenwood and M. S. Nikulin, A guide to chi-squared testing, vol. 280. John Wiley & Sons, 1996.
- [130] 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, pp. 329–338, 1993.
- [131] 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, pp. 611–620, Society for Industrial and Applied Mathematics, 2006.
- [132] 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, pp. 1466–1478, Oct 2012.
- [133] 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.
- [134] 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.
- [135] 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.

- [136] Z. Galil, “Efficient algorithms for finding maximum matching in graphs,” ACM Computing Surveys (CSUR), vol. 18, no. 1, pp. 23–38, 1986.
- [137] “Pecan street inc..” <https://www.pecanstreet.org>.
- [138] “Solar Resource Data.” <https://pvwatts.nrel.gov/pvwatts.php>.
- [139] 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.
- [140] L. Gurobi Optimization, “Gurobi optimizer reference manual,” 2020. <https://www.gurobi.com>.
- [141] M. O. Rieger, M. Wang, and T. Hens, “Estimating cumulative prospect theory parameters from an international survey,” Theory and Decision, vol. 82, no. 4, pp. 567–596, 2017.
- [142] V. Baláž, V. Bačová, E. Drobná, K. Dudeková, and K. Adamík, “Testing prospect theory parameters,” Ekonomicky časopis, vol. 61, pp. 655–671, 2013.
- [143] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, Nonlinear programming: theory and algorithms. John Wiley & Sons, 2013.
- [144] R. Storn and K. Price, “Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces,” Journal of global optimization, vol. 11, no. 4, pp. 341–359, 1997.
- [145] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., “Human-level control through deep reinforcement learning,” nature, vol. 518, no. 7540, pp. 529–533, 2015.
- [146] E. Casella, E. Sudduth, and S. Silvestri, “Dissecting the problem of individual home power consumption prediction using machine learning,” in 2022 IEEE International Conference on Smart Computing (SMARTCOMP), (Finland), pp. 156–158, IEEE, 2022.
- [147] T. H. Cormen, Introduction to algorithms. MIT press, 2009.
- [148] P. S. Inc., 2019.
- [149] NREL, “Solar resource data,” 2019.
- [150] E. Casella, A. R. Khamesi, S. Silvestri, D. A. Baker, and S. K. Das, “Hvac power conservation through reverse auctions and machine learning,” in 2022 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 89–100, IEEE, 2022.

- [151] A. A. Lazar and N. Semret, “The progressive second price auction mechanism for network resource sharing,” in 8th International Symposium on Dynamic Games, Maastricht, The Netherlands, 1998.
- [152] J. Csirik, “Heuristics for the 0-1 min-knapsack problem,” Acta Cybernetica, vol. 10, no. 1-2, pp. 15–20, 1991.
- [153] R. M. Karp, “An algorithm to solve the $m \times n$ assignment problem in expected time $O(mn \log n)$,” Networks, vol. 10, no. 2, pp. 143–152, 1980.
- [154] C. of New York Taxi and L. Commission, “New york city taxi and limousine commission (tlc) trip record data of the year 2013,” 2019.

VITA

Ashutosh Timilsina

Education

- Ph.D. in Computer Science from the University of Kentucky. Expected Graduation May, 2023.
- B.E. in Electrical Engineering from the Tribhuvan University, Institute of Engineering, Pulchowk Campus. December, 2016.

Professional Experience

- August 2019 to May 2023: Graduate Research Assistant at Cyber Physical Systems Lab (under Dr. Simone Silvestri), University of Kentucky, Lexington, USA.
- April 2019 to August 2019: Electrical Engineer at the Nilgiri Khola Hydropower Company Ltd.
- May 2017 to May 2019: Electrical Engineer at the Mandu Hydropower Ltd.
- Dec. 2016 to May 2018: Technical Officer at the H.I.F. Renewable Energy Ltd.

Research Interests

User Behavioral Modeling, eCommerce, Mathematical Optimization, Artificial Intelligence, Machine Learning, Reinforcement Learning, Blockchain

Technical Skills

- Languages: Python, C/C++, MATLAB, SQL
- Libraries: Gurobi, NetworkX, PyTorch, Keras, TensorFlow, MPI, OpenMP
- Software: LaTeX, AutoCAD, KiCAD, SolidWorks, PVSyst

Scholastic and Professional Honors

- Outstanding Student Paper Award – Winner, Department of Computer Science (2022)
- Member of the Year Award – Winner, GSACS, University of Kentucky (2022)
- Recipient of four Student Travel Grants for attending ACM NanoCom, IEEE GLOBECOM, IEEE PerCom, IEEE TPEC
- Top 5 poster award at 6th Annual Commonwealth Computational Summit, University of Kentucky 2022
- Recipient of UKY GSC Student Conference Award (2022)
- LOCUS 2015: Electrical Project Competition - Winner (2015)
- LOCUS 2014: Electrical Project Competition – Appreciation (2014)
- Outstanding High School Student (2012)

Publications

- “e-Uber: A Crowdsourcing Platform for Electric Vehicle-based Ride- and Energy-sharing" (Under Review), A. Timilsina and S. Silvestri, 2023.
- “V2G Optimization for Dispatchable Residential Load Operation and Minimal Utility Cost", R. Alden, A. Timilsina, S.Silvestri, D.Ionel, 2023 IEEE ITEC, Michigan, 2023.
- “P2P Energy Trading in a Smart Residential Environment with User Behavioral Modeling", A. Timilsina, 2023 PerCom PhD Forum, Georgia, 2023.
- “P2P Energy Trading through Prospect Theory, Differential Evolution, and Reinforcement Learning", A. Timilsina and S. Silvestri, 2023
- “Prospect Theory-inspired Automated P2P Energy Trading with Q-learning-based Dynamic Pricing", A. Timilsina and S. Silvestri, IEEE Globecom, Rio de Janeiro, Brazil, 2022
- “A Reinforcement Learning Approach for User Preference-aware Energy Sharing Systems", A. Timilsina et al, IEEE Transactions on Green Communication & Networks, 2021
- “Comparative Analysis of Cell Balancing Topologies in Battery Management Systems", A. Timilsina et al, IoE Graduate Conference Summer, Kathmandu, Nepal, 2019
- “Technical Design of a Grid-Connected Photovoltaic System and Its Challenges in Nepalese Power Scenario", A. Timilsina and B. Paudyal, IEEE ICPS, Pune, India, 2017
- “A Novel Approach for Wireless Power Transfer using Magnetic Resonant Method", A. Timilsina et al, OKRP Conference, Kathmandu, Nepal, 2016

Leadership & Volunteering Experiences

- Peer Reviewer for over 18 articles and journal papers
- October 2022: International Conference on Network Protocols (ICNP’22) - Volunteer
- May 2022 – May 2023: University of Kentucky Graduate Student Congress – Representative