Algorithmic Marketing with Data-Driven Simulations

By

Haifeng Zhang

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

August 11, 2017

Nashville, Tennessee

Approved:

Yevgeniy Vorobeychik, Ph.D.

Gautam Biswas, Ph.D.

Doug Fisher, Ph.D.

Bradley Malin, Ph.D.

William Rand, Ph.D.

To my great parents and loving sister.

To my darling, Ting, and my pearls, Emery and Ellene.

Thanks for always being there for me, and love me!

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my advisor Prof. Yevgeniy Vorobeychik for his continuous support of my Ph.D. study. Without his patience, inspiration, and guidance, I would not have accomplished so many in five years. I wish I could be mentored by him a little longer but it is time to begin my new adventure. I have been amazed by his novel way of formulating research questions, tackling technical challenges, and communicating scientific results. This influence will not simply disappear because of my leave but will perdure in the rest of my life.

I am thankful to the rest of my thesis committee: Prof. Gautam Biswas, Prof. Doug Fisher, and Prof. Bradley Malin and Prof. William Rand, for their helpful comments on the thesis proposal as well as the final dissertation. Special thanks to Prof. William Rand for traveling from NC State University for my defense and enlightening me the huge potential of agent-based models in fields like business and management.

I am grateful to Prof. Ariel Procaccia for his valuable advice while we collaborated on two past projects, and generous help in writing the papers, and for his strong support while I was hunting for an academic position.

I appreciate the time I took from Swetasudha Panda, Bo Li, and other lab mates. Thanks so much for their patience and giving the opportunities to challenge my expertise and to share valuable experience.

Finally, I would like to thank my parents and sister for their unconditional support and endless encouragement, my wife for taking care of me and two children willingly sacrificing her own future. Without their love, I would not have survived and got my Ph.D. in computer science.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 The Theory of Innovation Diffusion

One of the most fundamental questions in the field of marketing is to understand why and how products or services are adopted by consumers. At a higher level, any new product or service can be thought of as an *innovation*. Empirical studies of innovations, such as, new product, new technology, new practice and ideas had a long history even before 1962, when Everett M. Rogers published his book *Diffusion of Innovations*. Notably, the similarity of various kinds of innovations, for example, S-curve adoption, was not paid enough attention, until he conducted the meta-review of these case studies and invented a comprehensive *theory of innovation diffusion* [1]. He summarizes that there are five factors determining how fast an innovation is adopted, which are relative advantage, compatibility, complexity, trialability, and observability. His theory also emphasizes the impact from social peers on one's decision making with respect to network structure, social norms and leadership. One major theoretical contribution is the categorization of adopters, which posits that the population of adopters can be divided into five categories by *innovativeness* (measured by the time when the product was adopted): innovators, early adopters, early majority, late majority and laggards. Moreover, adopters in different categories can be differentiated by their social, economic, psychological characteristics. The theory has influenced marketing research that followed and been used in wide applications. In particular, the categorization of adopters motivates the use of *targeted marketing* (that takes consumer heterogeneity into account) and serves as the underlying framework for numerous studies that either conduct empirical analysis or construct computational models [2].

1.2 Modeling the Diffusion of Innovations

While qualitative insight into how an innovation is diffused within a social system is important, quantitative models that can characterize the diffusion process are extremely useful in marketing science [3, 4]. Although there are numerous and various types of mathematical models for product diffusion in literature, the *Bass model* is probably the most influential due to its wide application and impact on many other diffusion models [5, 6]. Basically, the model assumes that the *likelihood* for a consumer to adopt a new product given he/she has yet purchased is a *linear* function of the fraction of previous adopters. At any given time t, the temporal relationship can be described as follows

$$\frac{f(t)}{1-F(t)} = p + qF(t)$$

where $f(t)$ is the probabilty of adoption at time t, $F(t)$ is the accumulative probabilty as well as the fraction of adopters upto time t; p (*coefficient of innovation*) is often interpreted as external influence, e.g., advertising and other communications initiated by firms, and q (*coefficient of imitation*) refers to internal influence, e.g. interactions among adopters and potential adopters in the social systems [6]. Moreover, based on the assumption, the volume of sales as a function of time can be derived. Then, the time and volume of *peak* sale can be estimated, which are often interested by sales managers.

The Bass model is simple but remains a widely-used forecast tool in today's marketing industry. Particularly, once the model is calibrated by aggregate data of a sales time-series, it can replicate the common S-curved adoption trajectories for many consumer durables. However, it has a number of limitations. First, it assumes consumers are *homogeneous* with the same probability density functions and implicitly influence each other in a complete graph, which seems too ideal to explain any real diffusion process. Second, some important managerial variables, such as price and advertising cost, are in fact not explicitly modeled. Third, the predictive power of the model is often questioned due to the fact that to fit

the S-shaped adoption curve one needs data about several critical points that are often more interesting to predict. If an aggregate diffusion model, like the Bass model, is not ideal, then it begs the question: what kind of models are suitable for modeling innovation diffusion? Next, we introduce an innovative approach, namely, *agent-based modeling*.

1.3 Agent-based Modeling

Agent-based modeling (ABM) is a computational technique that simulates a system's behaviors by simulating the behaviors of individuals that compose the system. The technique is developed to study complex system properties emergent from interactions among agents [7, 8], but it has gained popularity in many scientific areas over the past decade [9, 10, 11]. As a bottom-up modeling method, modelers have to first specify rules that govern an agent's actions and decision making, then the system dynamics can be generated by simulating the interactions among agents. This distinguishes it from many other simulation techniques, for example, the Bass model, since heterogeneity, partial rationality and network-based interactions are generally easier to integrate into the system. Interestingly, traditional agent-based models often utilize *simple* rules [12, 13]. By controlling the complexity of the model, it is easier to interpret and understand, which seems to be an important design decision as the primary purpose of using agent-based models is to learn the causal relations between model's input (initial conditions, parameters) and output ("what-if" analysis). Traditionally simple agent-based models are used as learning tools to either test a modeler's hypotheses or uncover new insights about how an individual behaves and a system works.

Understanding how individual behaviors could induce the changes at the market level is, in fact, a preliminary step towards the design of effective marketing strategies. Marketing activities, such as, advertising and promotion, are common intervention policies that aim to increase either sales or brand awareness (outcomes at the system level) by influencing individual behaviors. To support the design of effective marketing strategies, the

agent-based models armed with simple rules seem very limited. On one hand, as the consumer decisions and their social interactions are complex, these simple models need to be extended so that they can capture more meaningful and important features. On the other hand, the ultimate goal, supporting decision making requires the agent-based model to be empirically grounded. There is growing body of work that calibrates agent-based models on real data and draws insight via what-if analysis [14]. Unfortunately, these models are often validated qualitatively by replicating stylish facts without evaluating the predictive power in a rigorous manner [2]. An agent-based model with lower prediction accuracy is a less valid representation of the modeled phenomenon and, therefore, less qualified to support any meaningful decisions by policy makers. Interestingly, few agent-based models have been developed explicitly for prediction and are validated rigorously. Surrounded by massive diffusion data, machine learning and data mining seemly provide us with efficient tools and algorithms to build high-predictive consumer behavioral models. If we could integrate these statistically-learned behavior models into multi-agent simulations, and validate them using empirical data, the qualification of agent-based models as decision support tools could be greatly enhanced.

1.4 Influence Maximization

The ultimate goal of any marketing activity is to promote the influence and adoption of products or services. A general marketing problem studied in the field of *information diffusion* is called *influence maximization* [15, 16], in which the "influence" can represent the aggregate adoption of ideas, beliefs or products. The influence is evaluated by two simple social contagion models: the Independent Cascading (IC) model [17] and the Linear Threshold (LT) model [18]. Both models are defined on a directed graph where activation is assumed to be monotonic: once a node is active (e.g., adopted, received information), it cannot become inactive.

Starting from a set of initial adopters, the diffusion process in the two models is as-

sumed to progress iteratively in a *synchronous* way along a discrete time-horizon, until no transmission is possible. IC is often considered *sender-centric*, while LT is *receiver-centric*. Particularly, in each iteration, a new active node in the IC model is given a single chance to activate its inactive neighbors independently with an exogenously specified probability (usually represented by the weight of the corresponding edge). In the LT model, in contrast, an inactive node will become active only if the sum of weights of its activated neighbors exceeds a predefined node-specific threshold, which is typically randomly assigned between 0 and 1 for each network node. Note that in both models a newly activated node becomes active immediately in the next iteration.

In either IC or LT model, *influence* is defined as the number of active nodes at the final stage, which depends on the initially seeded nodes S, denoted as $\sigma(S)$. The influence function $\sigma(S)$ has been proved to be *monotone* and *sub-modular*, i.e., diminishing returns to scale. The influence maximization problem in its simplest version asks: which are the best k nodes to seed so that to maximize $\sigma(S)$. It turns out that a straightforward greedy algorithm, i.e., iteratively add the a node with the largest marginal gain to the selected set, is guaranteed to obtain a solution that achieves an approximation, that is greater than $1-1/e$ (≈ 0.63) from the optimal. Notice that the algorithmic result relies on the assumption that either the IC or the LT model is used for influence function, which are unfortunately not guaranteed to be monotone and submodular [16]. In fact, the early stage of innovation diffusion is better characterized by a process showing increasing returns to scale. Consider the well-known logit function to define the adoption probability as a function of the number of previous adopters. When the probability is lower than 0.5 (evident for renewable technologies as seen in [19]), the likelihood of adoption grows faster as the number/fraction of adopters increases. This raises the question: how should we optimize marketing actions (*who* and *when* to target) in such a distinct setting, as opposed to the submodular influence maximization?

1.5 Submodular Optimization

Above we briefly mentioned that the influence function defined in either IC or LT model is submodular. Now we provide a formal definition of *submodularity*. Intuitively, a submodular function is a mathematical *set* function that exhibits the natural property of increasing returns scale. Let V be a collection of elements and $f : 2^V \rightarrow R_{\geq 0}$ a function over subsets of V, and assume that f is monotone increasing. For any set $X \subseteq V$, define $f(j|X) = f(X \cup \{j\}) - f(X)$, that is, the marginal improvement in f if element $j \in V$ is added to a set $X \subseteq V$. We say f is a *submodular* function if for all $S \subseteq T \subseteq V$, $f(j|S) \geq f(j|T)$. Submodular optimization is aimed to maximize the submodular function $f(X)$ given some constrains on X by selecting the optimal set X^*. The problem has received much attention due to its broad applications, such as viral marketing, information gathering, image segmentation, and document summarization [16, 20, 21].

Submodular optimization, particularly under cardinality or cost constraints, has received considerable attention in a variety of applications from sensor placement [22] to targeted marketing [16]. However, the constraints faced in many real domains are more complex. For instance, for the door-to-door marketing in a group of socially connected customers, a salesman would like to target a set of nodes subject to a budget constraint. To apply the *seeding* the salesman must pay the transportation cost and a node-based visit cost. In other words, the final cost with respect to a set of seeds can not be simply attributed to each node. If the cost function is additive, some cost discounted greedy algorithm with lower bound can find us a suboptimal solution [23, 24]. But, in the case of more general routing cost constraint, does there exist some approximation algorithm that guarantees a lower bound?

1.6 Multi-channel Marketing

Marketing operations are experiencing steady transformation from a single channel to multiple channels. The emergence of digital media, such as the World Wide Web, search engines, and online social networks, has opened up tremendous opportunities for today's marketers to look for prospects and engage existing customers. A mix of these innovative channels with traditional ones, such as TV, direct mailing, and door-to-door marketing, has been widely adopted by many companies to generate more sales, maintain stronger customer relationships, and achieve a higher customer retention rate [25]. Despite its benefits, this practice has also significantly increased operational complexity, making marketing one of the key managerial challenges [26, 27].

A critical decision faced by a multi-channel marketer is to determine the optimal budget allocation among the marketing channels [1]. To do so, the marketer needs a way to evaluate the effectiveness of alternative budget splits. This is often done by advanced simulation models, for example, some highly sophisticated agent-based models calibrated by and run with real data. The use of simulations, as compared to analytic objective functions (such as concave and continuous utility being maximized), introduces an important technical challenge: simulations are often slow, and computing parsimony is, therefore, crucial in query-based black-box optimization methods. A second technical challenge arises from the fact that the response function for each channel (e.g., the number of individuals who buy the product) commonly exhibits *budget complementarities*, requiring a non-trivial added expense on a channel to make a significant impact on the response function. In fact, if the effect of budget is given or is a continuous (concave) function of the budget, the problem can be solved with standard utility maximization techniques [29]. To the best of our knowledge, few existing budget optimization techniques take the simulation-related computational cost and budget complementarities into account.

[1] Lately, in marketing literature, extensive efforts on attribution modeling were made to quantify the importance of advertising impressions and channels [28]. Their results seem insightful but fail to provide a direct solution to the optimal allocation of a single budget over a mix of marketing channels.

1.7 Organization of the Thesis

The dissertation provides a novel computational method that leverages massive diffusion data and efficient algorithms to optimize marketing operations. The general framework is illustrated by Figure 1.1. The *algorithmic marketing* system is composed of two subsystems: *models* and *algorithms*. Models are mathematical systems that characterize either aggregate or individual adoption behaviors of innovations. Typical aggregate models include variations of the Bass model. Agent-based models are individual-based, but can also generate aggregate dynamics. Notably, models are *learned* from the diffusion data. By taking advantage of state-of-the-art machine learning techniques, the learning (training) process can be done quite efficiently when applying to large-scale data. Once a model is learned and validated (a step to ensure the model is representative and predictive using empirical data), it interacts with algorithms by evaluating effects of marketing actions. Algorithms utilize the computational model to solve for the optimal or near-optimal marketing plans, which can either assist decision making or be automatically executed. Since the nature of modeled phenomena (diffusion of innovations) could change over time, the computational models need to be updated accordingly, which is reflected by the cycle-wised system structure.

The thesis starts with a critical review of empirically-grounded agent-based models of innovation diffusion, followed by a proposal of an innovative data-driven agent-based modeling framework to attack the modeling challenge faced by marketing researchers and practitioners, and develops several efficient algorithms targeting realistic marketing problems and finally presents a novel optimization framework that addresses computing challenges arising from the interactions between simulation models and algorithms. We conclude that the ongoing efforts of algorithmic marketing with data-driven simulations will lead to the ultimate goal of *automated marketing*. Figure 1.1 illustrates how each chapter fits into our algorithmic marketing framework. A brief description of each chapter also follows.

In Chapter 2, we present a critical review of empirically grounded agent-based models

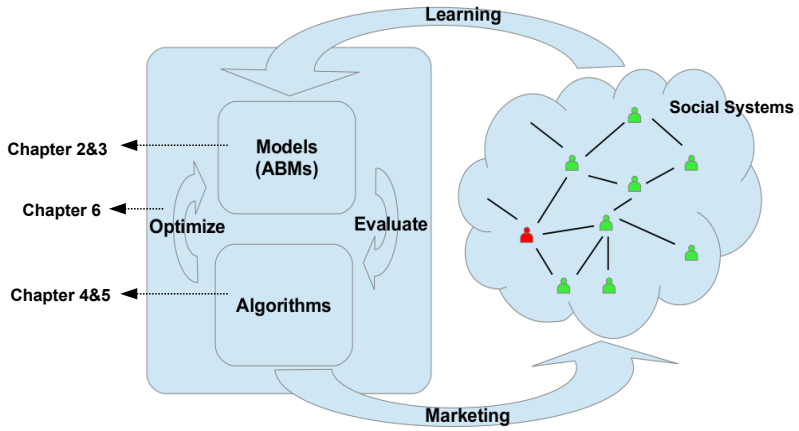


Figure 1.1: System Architecture of Algorithmic Marketing

of innovation diffusion, developing a categorization of this research based on types of agent models as well as applications. By connecting the modeling methodologies in the fields of information and innovation diffusion, we suggest that the maximum likelihood estimation framework widely used in the former is a promising paradigm for calibration of agent-based models for innovation diffusion. Although many advances have been made to standardize ABM methodology, we identify four major issues in model calibration and validation, and suggest potential solutions. This work is currently under journal review.

In Chapter 3, we present a novel data-driven agent-based modeling framework, in which individual behavior model is learned by machine learning techniques, deployed in multi-agent systems and validated using a holdout sequence of collective adoption decisions. We apply the framework to forecasting individual and aggregate residential rooftop solar adoption in San Diego county and demonstrate that the resulting agent-based model successfully forecasts solar adoption trends and provides a meaningful quantification of uncertainty about its predictions. Meanwhile, we construct a second agent-based model, with its parameters calibrated based on the mean square error of its fitted aggregate adop-

tion to the ground truth. Our result suggests that our data-driven agent-based approach based on maximum likelihood estimation substantially outperforms the calibrated agent-based model. Given the advantages over the state-of-the-art modeling methodology, we utilize our agent-based model to aid in the search for potentially better incentive structures aimed at spurring more solar adoption. Although the impact of solar subsidies is rather limited in our case, our study still reveals that a simple heuristic search algorithm can lead to more effective incentive plans than the current solar subsidies in San Diego County and a previously explored structure. Finally, we examine an exclusive class of policies that gives away free systems to low-income households, which are shown to be significantly more efficacious than any incentive-based policies we have analyzed to date. This work was published in the following journal:

- Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with application to rooftop solar adoption. *Autonomous Agents and Multi-Agent Systems*, 30(6):1023–1049, 2016

An earlier work appeared in the following conference:

- Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with application to rooftop solar adoption. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 513–521, 2015

In Chapter 4, we formulate a dynamic influence maximization problem under increasing returns to scale over a finite time horizon, in which the decision maker faces a budget constraint. We propose a simple algorithm in this model which chooses the best time period to use up the entire budget (called Best-Stage) and prove that this policy is optimal in a very general setting. We also propose a heuristic algorithm for this problem of which Best-Stage decision is a special case. Additionally, we experimentally verify that the proposed "best-time" algorithm remains quite effective, even as we relax the assumptions under which op-

timality can be proved. However, we find that when we add a "learning-by-doing" effect, in which the adoption costs decrease as a function of aggregate adoption, the "best-time" policy becomes suboptimal, and is significantly outperformed by our more general heuristic. This work was published in the following conference:

- Haifeng Zhang, Ariel D Procaccia, and Yevgeniy Vorobeychik. Dynamic influence maximization under increasing returns to scale. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 949–957, 2015

In Chapter 5, we investigate an important and very general class of problems of maximizing a submodular function subject to general cost constraints, especially focusing on costs coming from route planning. Canonical problems that motivate our framework include mobile robotic sensing and door-to-door marketing. We propose a generalized cost-benefit (GCB) greedy algorithm for a fundamental problem with only one actor, and use it to construct another algorithm based on sequential planning to solve a natural multi-actor extension. In both cases, we prove approximation guarantees under significantly weaker assumptions than those in prior literature. Experimental evaluation on realistic mobile sensing and door-to-door marketing problems, as well as using simulated networks, shows that our algorithm achieves significantly higher utility than state-of-the-art alternatives, and has either lower or competitive running time. This work has been submitted to a journal and preliminary results were published in the following conference:

- Haifeng Zhang and Yevgeniy Vorobeychik. Submodular optimization with routing constraints. In *AAAI Conference on Artificial Intelligence*, pages 819–825, 2016

In Chapter 6, we study the classical budget-constrained utility maximization problem but in a state-of-the-art multi-channel marketing context. Instead of assuming that the utility is a "nice" known analytic function, e.g., continuous and concave, we rely on a more reasonable assumption that the impact of allocating a fixed budget among alternatives (e.g., marketing channels) on outcomes is evaluated by sophisticated data-driven simula-

tions. While simulations enable high resolution evaluation of alternative budget allocation strategies, they significantly complicate the associated budget optimization problem. In particular, simulation runs are time consuming, significantly limiting the space of options that can be explored. An important second challenge is the common presence of budget complementarities, where non-negligible budget increments are required for an appreciable marginal impact from a channel. This introduces a combinatorial structure on the decision space. We propose to address these challenges by first converting the problem into a multi-choice knapsack optimization problem with unknown weights. We show that if the weights (corresponding to marginal impact thresholds for each channel) are well approximated, we can achieve a solution within a factor of 2 of the optimal, and that this bound is tight. We then develop several parsimonious query algorithms for achieving this approximation in an online fashion. Experimental evaluation demonstrates the effectiveness of our approach. This work will be published in the following conference:

- Haifeng Zhang, Yevgeniy Vorobeychik, and Ariel D Procaccia. Multi-channel marketing with budget complementarities. In *International Conference on Autonomous Agents and Multiagent Systems*, 2017, to appear

Chapter 2

Empirically Grounded Agent-Based Models of Innovation Diffusion: A Critical Review

Innovation diffusion has been studied extensively in a variety of disciplines, including sociology, economics, marketing, ecology, and computer science. Although aggregate models once dominated the diffusion literature, agent-based models are gaining popularity as it can easily incorporate heterogeneity and interactions among individuals. While most ABM work on innovation diffusion is theoretical, empirically grounded models are increasingly important, particularly in guiding policy decisions. The algorithmic marketing architecture proposed in Chapter 1 (Figure 1.1) is indeed empirically grounded. This chapter provides a critical review on calibration and validation of the empirically grounded agent-based models developed to represent the diffusion of a variety of innovations.

2.1 Introduction

2.1.1 Innovation Diffusion: Theoretical Foundations

Diffusion refers to the process by which an innovation is adopted over time by members of a social system [34, 35]. Commonly, an *innovation* refers to a new technology, but the conceptual notion can be applied far more broadly to consider the spread of ideas and practices. Rogers [34] laid down the theoretical foundations of innovation diffusion in his book, *Diffusion of Innovations*, in which he synthesizes studies in anthropology, sociology, and education, and proposes a generic theory to explain the diffusion of innovations among individuals and organizations. In addition to the five characteristics that determine the rate of adoption as mentioned in Chapter 1, he considers adoption decision of an innovation as a multi-stage process, involving five stages: *knowledge*, *persuasion*, *decision*, *implementation*, and *confirmation*. Most importantly, according to his theory, adopters can be classified into five categories: *innovators*, *early adopters*, *early majority*, *late majority*, and *laggards*.

In addition to these high-level considerations, much attention has been on the significance of social relationships and influence in innovation diffusion (in contrast with, or complementary to, economic considerations). Starting with early groundwork [36], there has now been extensive research on how social network structure, group norm, opinion leadership, weak ties, and critical mass impact the diffusion of innovations [37, 38].

2.1.2 Mathematical Models of Innovation Diffusion

Traditional mathematical models of innovation diffusion aim to model aggregate trends, rather than individual decisions. Numerous such models follow the framework of Bass model, which is one of the most influential models in marketing [39, 40]. The Bass model was originally designed for forecasting sales of new consumer durables. It can be calibrated with aggregate sales data, and Bass showed that it can qualitatively capture the S-shaped pattern of aggregate adoption over time [6].

The Bass model has a number of limitations. First, it does not capture individual interactions. Indeed, the model explicitly assumes a fully connected and homogeneous network. For innovation diffusion, this is an important drawback, as individual interdependence and communications are among the most significant aspects to understand innovation diffusion [35, 34]. The second criticism of the Bass model is that it does not include any decision variables that are of interest from a managerial perspective. In fact, Bass late extended his original model to incorporate marketing mix variables, price and advertising [41]. For an extensive review of research in this direction, we refer readers to [5, 42]. Nevertheless, these marketing mix variables are mostly designated to the entire market without a consideration of individual heterogeneity. Lastly, the predictability of the Bass model is often questioned. For example, [43] argue, that the model needs considerable data around the critical point at which diffusion accelerates to be effective, but once such data is available the value of the Bass model becomes limited.

2.1.3 Agent-Based Modeling for Innovation Diffusion

Agent-based modeling (ABM) has emerged as another natural approach to study innovation diffusion. Agent-based models are typically simulation models that capture dynamic interactions among a (usually large) collection of individuals. They were originally developed as a tool for complexity theory research [7, 8] and have gained popularity in many scientific areas over the past decade [9, 10, 11]. The ABM paradigm offers two advantages for the study of innovation diffusion: first, it facilitates the modeling of agent heterogeneity, and second, it enables fine-grained modeling of interactions mediated by social networks. Indeed, agent-based modeling has been applied in study of innovation diffusion to aid intuition, theoretical exploration, and to provide policy decision support [14].

Traditional agent-based models are largely conceptual [12, 13]. This use of ABMs as primarily conceptual tools is partly because they are commonly considered as ideal *learning tools* for scientists to understand a system under a variety of conditions by simulating the interactions among agents. As a consequence, the simplicity of agent rules is commonly a crucial consideration in the design of agent-based models. Such simplicity, however, has given rise to criticism of the ABM methodology as being "toy models" that do not reflect reality [11]. Moreover, an increasingly important criticism is that if ABMs are used in any policy decision support, the predictive validity of the model becomes paramount, and models that are primarily conceptual may be inadequate for such tasks.

It is this increasing use of agent-based modeling to obtain policy guidance that has motivated increasing use of empirically grounded agent-based models. Empirical agent-based models have recently experienced significant growth [14]. In these studies, empirical data are used to initialize simulation, parameterize agent-based models, and to evaluate model validity. The explosion of high-resolution data sets, coupled with advances in data analytics and machine learning have given rise to increased opportunities for empirically grounding agent-based models, and this trend is likely to continue. Our goal is to provide an overview of these empirically grounded agent-based models developed with the goal of

studying innovation diffusion. Through a careful examination of these studies, we also aim to identify potential methodological issues that arise, and suggest ways to address these.

2.1.4 Contributions

The diffusion of new products has been an important topic for decades [44, 5, 42, 43, 6]. The prevalence of the ABM approach can be glimpsed from a number of review papers from disciplines like sociology [45], ecology [46], and marketing [47, 4, 48]. For example, [47] describes potential uses of ABM in market research associated with innovations, exploring benefits and challenges of modeling complex dynamical systems in this fashion. [49] surveys agent-based models of innovation diffusion within a computational economics context. [6] reviews diffusion models in the context of a single market and cross-markets and brands. To the best of our knowledge, the closest work to ours is a review of agent-based simulations of innovation diffusion by [14], who survey both theoretical and empirical work. In comparison with these past reviews, we make the following novel contributions:

1. We provide a *systematic* review of the empirical agent-based models of innovation diffusion. This is in contrast to the narrative review of the applied work as provided in [14]. In particular, we offer a novel classification of agent adoption models as employed in the reviewed papers. By highlighting the adoption models and their parameterization methods, we aim to bridge methodological gaps among domains and applications. We identified the papers to include in a rigorous and systematic manner. In terms of scope, any work presenting an agent-based model using empirical data to simulate the diffusion of innovations was included. Our selection process combined results from multiple databases, including Google Scholar and ScienceDirect, with extensive search for relevant keywords, and back-tracking and forward-tracking reference lists, while carefully screening out non-candidates.

2. Our review is *comprehensive* and *updated*. The collection of reviewed papers spans a *superset* of the applications as covered in [14] and, indeed, a number of significant efforts have emerged after 2012. Notably, we also include a selection of papers from the literature on information diffusion, a fast-growing area. These models rely on principled machine learning techniques for model calibration based on empirical observations of diffusion traces. In addition, we exclude two (out of 15) papers from [14] which are not empirically grounded. In the end, we reviewed 43 papers, of which 30 (23 from years after 2011) were not included by [14].

3. We provide a *critical* review, assessing the strengths and weaknesses of the surveyed research. Almost all surveyed papers followed standard modeling steps and presented their results systematically. However, we conclude that the current literature commonly exhibits several major shortcomings in model calibration and validation[1]. Addressing these issues would significantly increase the credibility of agent-based models. We, therefore, devote a section to an overview of existing validation methods in the literature and an in-depth discussion of these issues and potential solutions.

2.2 Categorization of Empirically Grounded ABMs of Innovation Diffusion

We review the burst of recent developments of empirically grounded agent-based models, which are examined through two dimensions: models and applications. First, to facilitate methodological comparison, we group the papers into six categories which represent the specific approaches taken to model individual agent decision processes: *mathematical optimization based models*, *economic models*, *cognitive agent models*, *heuristic models*, *statistics-based models* and *social influence models*. Second, as we observe that modeling efforts span several domains, the next section offers an application-focused categorization.

The categorization in this section is aimed at qualitatively clustering the existing agent-based models with respect to their *modeling methods*, which can be further characterized

[1]The concepts of calibration and validation are explained in Section 2.5.1 below.

from several dimensions, such as behavioral assumption, data granularity, internal structure, calibration, and validation. The six categories we identified present a comprehensive picture and structured patterns of the different methods used to model individual agent decision processes seen in a variety of applications.

We review each paper in sequence and in some detail, providing sufficient depth in the review for a reader to understand the nature of each surveyed work. In particular, we focus on how data was used in the modeling process, and in particular, in initialization, calibration and validation steps. We attempt to draw connections among the papers using our categorization structure (i.e., by grouping them into the six categories based on the methodology used to model individual agent behavior). Table 2.1 shows how these survey articles are distributed across the categories and publication years[2]. Notice that this approach is different from the synthesis-based approach followed by other review papers, such as, [50], and [51], which generally draws conclusion for a collection of papers but does not provide sufficient detail to assess how data is used in these efforts.

Table 2.1: Distribution of surveyed papers over categories and years

Category by modeling methods	Distribution in year	Total Published
mathematical optimization based model	01,07(2),09,10,13	6
economic model	10, 11(2), 12, 13, 14(2), 15	8
cognitive agent model	02, 06, 09(2), 12, 13(2), 15(2), 16(2)	11
heuristic model	10, 11(2)	3
statistics-based model	07(2), 08, 09, 11, 12, 13, 14, 15, 16	10
social influence model	13(2), 14, 16, 17	5
Total		43

2.2.1 Mathematical Optimization (MO) Based Models

The MO-based models posit that agents (e.g., farmer households) are deliberate decision-makers who use sophisticated mathematical planning tools to assess the possible conse-

[2]For simplicity, we omit "20" and use the last two digits to denote a year. For example, "07(2)" stands for 2 publications in year 2007.

quence of actions. While agents may encounter uncertainty, incomplete information, and constraints, their final decisions to adopt innovations are determined by concrete optimization objectives. The use of complex mathematical programs is commonly justified by the fact that farmer agents often consider their farming decisions in terms of economic returns.

In a seminal paper, Berger [52] developed a spatial multi-agent mathematical programming (MP) model of diffusion of farming innovations in Chile. Production, consumption, investment, and marketing decisions of individual households are modeled using linear programming with the goal of maximizing expected family income subject to limited land and water assets. Moreover, in accordance with the literature on innovation diffusion, the model incorporates the effects of past experience, as well as observed experience by peers. This is done by imposing a precondition for the MP procedure that the net benefit is only calculated if peer adoption level reaches the predefined threshold. In addition to such contagion effects, agent interactions are also reflected by the feedback effects of land and water resources and return-flows of irrigation water, implemented by coupling the economic agent decision model with hydrological components. In simulation models, agents are cellular automata with each cell associated with biophysical and economic attributes, such as soil quality, water supply, land cover/land use, ownership, internal transport costs, and marginal productivity. *These agent properties are initialized using empirical data* derived from various data sources, including a survey that captures both agronomic and socio-economic features, and a spatial data set with information about land and water use. *Parameters were calibrated* in terms of closeness of simulation experiments and farm data at both macro and micro levels. *Validation* was then performed by regressing land use results based on the model on actual land use in the data. Although values of the slope of this regression are reported for both macro and micro levels, validation is incomplete. For instance, micro-validation is only conducted for the year when the simulation starts due to data availability. Finally, the fact that validation was not conducted on data independent from calibration is another important weakness. Later, Berger et al. [53] applied his MP-based agent-based

modeling approach to study the complexity of water usage in Chile. Unfortunately, that work still had the same issue on validation.

Schreinemachers et al. [54] adopted the MP-based approach to simulate soil fertility and poverty dynamics in Uganda, and analyze the impact on these of access to short-term credit and alternative technologies. At the heart of the model is a simulation of a farmer's decision process, crop yields, and soil fertility dynamics. The decision model is comprised of three parts: 1) a set of possible decisions related to agriculture, such as growing crops, raising livestock, and selling and purchasing agricultural products; 2) a utility function that determines how much the decisions contribute to the farmer's objectives; and 3) links among decision variables represented by a set of equations. Following Berger [52], a three-stage decision flow is defined that separates agent decisions into investment, production, and consumption. Moreover, the portion of the model capturing consumption includes econometrically-specified allocation of farm and non-farm income to saving, food, and other expenditures. Properties of the household agent, such as, quantity and quality of land, labor, livestock, permanent crops, and knowledge of innovation, are sampled from empirical distributions based on limited samples. Additional features include models of animal and tree growth, technology diffusion, demographics, and price changes. In technology diffusion, peer influence is captured in the same manner as Berger [52], but notably, each agent is assigned a threshold based on household survey data. The model was systematically validated in three steps: first, econometric models were validated for accuracy, then each component was validated independently, and finally the system as a whole. Similar to Berger [52], validation used the same data as calibration.

Schreinemachers et al. [55] studied diffusion of greenhouse agriculture, using bell pepper in a watershed in the northern uplands of Thailand as a case study. The work largely follows the MP-MAS (mathematical programing-based multi-agent systems) approach due to [52]. Notably, the author proposes calibrating the diffusion thresholds as described in [52] by using a binary adoption model (e.g., logistic regression), which is estimated from

farmer survey data. To obtain threshold values for individuals, the author first computes adoption probability for each agent based on a set of observable independent variables, and then ranks these, dividing them into the five categories of innovators due to Rogers [56]. Validation was carried by checking the value of R^2 associated with a regression of observed land use on its predicted value. The proposed validation method suffers from the same limitation as other related research in using the same data for calibration and validation.

Schreinemachers et al. [57] applied the MP-based approach to study the impact of several agricultural innovations on increasing profitability of litchi orchards in Northern Thailand. Unlike Schreinemachers et al. [55] that estimated a logistic regression model to assign agents to threshold groups, they assigned thresholds randomly due to the lack of relevant data. The model was validated using regression method as described in Schreinemachers et al. [55], and validation suggests that the model reasonably represents aggregate agent behavior, even while individual-level behavior is not well captured. As in prior work, calibration and validation used the same data.

Alexander et al. [58] developed an agent-based model of the UK perennial energy crop market to analyze spatial and temporal dynamics of energy crop adoption. The model includes the interaction of supply and demand between two agent groups: farmers and biomass power plant investors. The farmer agents have fixed spatial locations which determine the land quality and climate that in turn impact crop yields, and decide on the selection of crops via a two-stage approach similar to Berger [52], with peer influence again modeled through a threshold function. A farmer agent considers adoption only if the proportion of neighbors within a given radius with a positive adoption experience exceeds a threshold. When adoption is considered, a farm scale mathematical program is used to determine the optimal selection of crops that maximizes utility as described in Alexander et al. [59]. Calibration of the farm scale model is either informed by empirical data or in reference to previous studies. Validation involved checking model behaviors on simplified configurations, unit-testing of model components, and comparing simulation results against

empirical data. However, validation did not use independent data from calibration.

2.2.2 Economic Models

Unlike the MO based models in Section 2.2.1, the *economic* models use simpler rules with fewer constraints and decision variables. Particularly, agents commonly simply minimize cost, maximize profit, or, more generally, maximize personal utility.

2.2.2.1 Cost Minimization

Faber et al. [60] develop an agent-based simulation model for energy technologies, micro-CHP (combined heat and power) and incumbent condensing boilers, in competition for consumer demand. Consumer agents are classified by housing type, which is viewed as the most important factor in determining natural gas requirements for heating units. At each time step a consumer considers purchasing a new heating unit, and follows a three-step decision algorithm: 1) assess if a new unit is needed, 2) scan the market for "visible" heating units, where "technology awareness" is formulated as a function of the level of advertising, market share, and bandwagon effect, and 3) each consumer chooses the cheapest technology of those that are visible. The cost, which depends on the consumer's class, is comprised of purchase costs, subsidies, and use costs over the expected life of the technology. Some of the parameters are calibrated using empirical data, while others are set in an ad hoc fashion. Some validation was performed through the use of a sensitivity analysis of the variables such as market size, progress rate, and technology lifetime. However, no explicit model validation using empirical data was undertaken.

2.2.2.2 Profit Maximization

Sorda et al. [61] develop an agent-based simulation model to investigate electricity generation from combined heat and power (CHP) biogas plants in Germany. Instead of simulating farmer's individual decision whether to invest in a biogas plant, the model solves

a system-wide optimization problem from the perspective of a global planner. The model includes two types of agents: information agents, including federal government, bank, electric utility, and plant manufacturer, and agents making investment decisions, including the substrate supplier, district, decision-maker, and heat consumer. The core decision-making agent acts as a representative for investors in each community. The agent chooses to invest in a biogas facility whenever sufficient resources are available and the investment yields positive net present value. This work used multiple data sources to construct the simulation model. For example, plant operator guidelines and manufacturer specifications were used to obtain data about the characteristics of biogas plants. Although the model is thus informed by real data, it is not quantitatively validated.

2.2.2.3 Utility Maximization

Broekhuizen et al. [62] develop an agent-based model of movie goer behavior which incorporates social influence in movie selection decisions. Their study investigates two types of social influence: the influence of past behavior by others, and influence stemming from preferences of an individual's friends, such as group pressure to join others in seeing a movie. The main purpose of this work is to determine the degree to which different types of social influence impact inequality. In their model, an agent's decision-making is probabilistic and utility-driven. An agent first observes which movies are being shown in the marketplace with some probability. Next, with a specified probability, an agent is selected to consider seeing a movie. If selected, it goes to the movie that maximizes expected utility among all those it is aware of. Otherwise, it does not see any movie. *Utility* in this setting is a weighted sum of *individual* utility, which represents the alignment between individual's preferences and movie characteristics, and *social* utility which is a combination of the two types of social influence above. Some of the model parameters are either theoretically determined or empirically calibrated, while the variability of the rest is investigated by sensitivity analysis. Validation involved a cross-national survey, using

cross-cultural differences due to Hofstede's collectivism-individualism index to measure social influence. While the validation is based on an independent survey study, it is largely qualitative.

Günther et al. [63] introduce an agent-based simulation approach to support marketing activities. The proposed model was applied to the study of a new biomass-based fuel that would likely be introduced in Austria. Consumer agents are embedded in a social network, where nodes represent agents and edge weights determine the probability with which the connected agents communicate. The authors tested several network structures, including random (Erdos-Renyi) networks, small-world networks, and so-called "preferences-based" networks, where connections between agents are based on geographical and cognitive proximity as well as opinion leadership. Each agent is characterized by preferences, geographical position, tanking behavior, how informed they are about the product, and their level of social influence. Agents have preferences for several product attributes: price, quality, and expected environmental friendliness, which are initialized differently based on consumer type. Agents are geographically distributed in virtual space based on the spatial distribution of Austrian population, and their tanking behavior is a function of fuel tank capacity, travel behavior, and habits. Individual information level on the innovation at hand captures the knowledge about a product, which increases as a function of interpersonal communication and exposure to marketing activities. Influence level, on the other hand, represents an agent's expertise with the innovation and determines the amount of information received through communication. Upon interaction, an agent with lower information level learns from a more informed agent. Most importantly, the *utility* function for agent i at time t is given by $u_{i,t} = (1 - Price_t) \times w_{i,1} + Price_t \times w_{i,2} + ppq_{i,t} \times w_{i,3} + w_{i,4}$, where $0 \leq w_{i,k} \leq 1$ and $\sum_{k=1}^{4} w_{i,k} = 1$, and the first and second weights pertain to price, while the last two represent how strongly agents prefer quality and how willing they are to seek renewable energy sources for fuel, respectively. An agent is assumed to adopt if utility exceeds a specified individual threshold drawn for each agent from the uniform distribution. Moreover, the

perceived product quality, ppq_i is assumed to gradually converge the true product quality for adopters. The author briefly mentions these model parameters are set in reference to a prior case study. Apart from this, no detailed information is provided about how model parameters are actually calibrated in the setting. Moreover, the model was only validated qualitatively with subjective expert knowledge.

Holtz and Pahl-Wostl [64] develop a utility-based agent-based model to study how farmer characteristics affect land-use changes in a region of Spain. As relevant data are scarce, their model cannot be quantitatively calibrated and validated. Empirical data are used to initialize the model, deriving the initial crop distribution, and to assess the validity of the model qualitatively. In this model, an agent's utility is formulated as a Cobb-Douglas function by multiplying four influences: gross margin, risk, labor load, and regulatory constraints. Parameters associated with these influences differ with the types of farmers, for example, part-time, family, and business-oriented farmers would have distinct utility parameters. In the decision process, an agent chooses a land use pattern that maximizes its utility, where land use patterns involve a combination of crop and irrigation technology, constrained by policies. The diffusion of irrigation technology is simulated based on the concept that the more widely used a technology is, the more likely it is to be considered by individual farmers. Their experiments explore the importance of each influence variable in the utility function, as well as of farmer types, by qualitatively comparing the simulation results with empirical data.

Plötz et al. [65] propose a model for the diffusion of electric vehicles (EVs) to evaluate EV-related policies based driving data in Germany. The model determines the market shares of different technologies by simulating each driving profile as both EV and conventional vehicle, choosing the option which maximizes the driver's utility, and then extrapolating these agent-level choices to aggregate market shares. In modeling individual decisions, utility is defined as a function of *total cost of ownership* (TCO), choice of EV brands, and individual *willingness-to-pay-more* (WTPM). The authors combined survey re-

sults with driving profiles to derive four categories of agents (adopters), and assigned each driving profiles to one of these categories. Through simulating the *plug-in hybrid electric vehicle* (PHEV) share of the market as a function of annual average *vehicle kilometers traveled* (VKT) for medium-sized vehicles, the model was validated by comparing original group assignment with simulated outcomes and by examining simulated diesel market shares relative to actual values within different branches of industry. While validation is quantitative and rigorous, it does not use independent data. Moreover, the model does not capture social influence which is often a key aspect of innovation diffusion modeling.

McCoy and Lyons [66] develop an agent-based model of diffusion of electric vehicles among Irish households. Agents representing households are located at a regular lattice space. They are heterogeneous as suggested by their characteristics. Agents have two static attributes, *Income Utility* (IU) and *Environmental Utility* (EU), drawn independently from empirical distributions derived from a survey. In particular, IU is based on an agents social class, tenure type, and age, which are assumed to be highly correlated with income, whereas EU is based on the agent's past adoption of energy efficiency technologies and their attitude toward the environment. Each agent i has a unique threshold, θ_i, drawn from a distribution that is negatively correlated to IU, and adopts if $U_i(t) \geq \theta_i$ and $t \times crit \geq rand(0,1)$, where, *crit* is decimal value that is used to account for inertia that exists in early stage of technology adoption, while utility $U_i(t)$ is defined as $U_i(t) = \alpha_i IU_i + \beta_i EU_i + \gamma_i G_i(t) + \delta_i S(t)$, where, IU represents individual's preferences, G is social influence, and S is social norms, and $\alpha_i + \beta_i + \gamma_i + \delta_i = 1$. To allow these parameters to vary by agent, the authors specify four distinct consumer groups with different preferential weighting schemes. Although the agents in the simulation are initialized using empirical distributions, key parameters in the decision model are not derived empirically but are based on the authors' assumptions. Additionally, no rigorous validation is provided.

Palmer et al. [67] developed an agent-based model of diffusion of solar photovoltaic (PV) systems in the residential sector in Italy. The *utility* of agent j is defined as the

sum of four weighted partial utilities, i.e., $U(j) = w_{pp}(sm_j) \cdot u_{pp}(j) + w_{env}(sm_j) \cdot u_{env}(j) + w_{inc}(sm_j) \cdot u_{inc}(j) + w_{com}(sm_j) \cdot u_{com}(j)$, where $\sum_k w_k(sm_j) = 1$ for $k \in K : \{pp, env, inc, com\}$ and $w_k(sm_j), U(j) \in [0,1]$. From left to right the partial utilities are: (1) payback period of the investment, (2) environmental benefits, (3) household income, and (4) social influence. An agent chooses to invest in PV if its total utility exceeds an exogenously specified threshold. Thresholds above vary by agent's demographic and behavioral characteristics, sm_j. The four partial utilities are derived from empirical data. Specifically, the payback period is estimated based on investment costs, local irradiation levels, government subsidies, net earnings from generating electricity from the system vs. buying it from the grid, administrative fees, and maintenance costs. The environmental benefit is based on an estimate of reduced CO_2 emissions saved. Household income is estimated based on household demographics, such as age, the level of education, and household type. Finally, social influence is captured by the number of neighbors of a household within its social network who have previously adopted PV. The social network among agents is generated according to the small-world model [68], modified to account for socio-economic factors. The model parameters are calibrated by trying to match simulated adoption with the actual aggregate residential PV adoption in Italy over the 2006-2011 period. The model is then applied to study solar PV diffusion in Italy over the 2012-2026 period. However, no quantitative validation is offered.

2.2.3 Cognitive Agent Models

While both MO-based (Section 2.2.1) and economic (Section 2.2.2) models elaborate economic aspects of the decision process and integrate simple threshold effects, cognitive agent models aim to explicitly model how individuals affect one another in cognitive and psychological terms, such as opinion, attitude, subjective norm, and emotion. This category includes the *Relative Agreement Model*, the *Theory of Planned Behavior*, the *Theory of Emotional Coherence*, and the *Consumat Model*.

2.2.3.1 Relative Agreement Model

The *Relative Agreement Model* belongs to a class of opinion dynamics models [69] and addresses how opinion and uncertainty are affected by interpersonal interactions. Seminal work is due to Deffuant et al. [70], who investigate how the magnitude of thresholds, with respect to attitude difference, leads to group opinion convergence and extremeness. The relative agreement model is often known as "*Deffuant model*" in the literature.

Deffuant et al. [71] design an agent-based model to simulate organic farming conversion in France. To model impact of interactions on individual decision, they relied on the Deffuant model in which both opinion and uncertainty are continuous variables. In the diffusion model, farmer agent has an "interest" state with three possible values: *not-interested*, *uncertain*, and *interested*. The actual value is based on the agent's opinion (represented as a mean value and confidence) and economic consideration. The value of the interest state depends on the position of the global opinion segment compared to a threshold value. Agent changes opinion after discussing with peers using a variant of the *Relative Agreement* algorithm [72]. The farmers send messages containing their opinions and information, following a two-stage diffusion model of Valente [37], mediated by a network generated according to the Watts and Strogatz [68] model. These impact opinions of the recipients as a function of opinion similarity, as well as the confidence of the sender, with more confident opinions having greater influence. In addition, if the farmer agent is "interested" or "uncertain", he performs an evaluation of the economic criterion, and if he remains interested, he requests a visit from a technician. After this visit, the economic criterion is evaluated again under reduced uncertainty. Finally, the adoption decision is made when the farmer has been visited by a technician and remains "interested" for a given duration.

Many model parameters governing the decision and communication process are not informed by empirical data. The authors tested the sensitivity of the model by varying these variables, including the main parameters of the dynamics, the parameters of the initial opinion distribution average number of neighborhood and professional links, and variations of

the institutional scenario. Within this parametric space, they aimed to identify parameter zones that are compatible with empirical data. For each parameter configuration, the authors defined two error measures: the adoption error and the error of proximity of adopters to the initial organic farmers. A decision tree algorithm was then used to find the parameter zones where the simulated diffusion has an acceptable performance. While this sensitivity analysis step can be viewed as model calibration, it is distinct from classical calibration which aims at finding a single best parameter configuration. The model was not validated using independent data.

2.2.3.2 Theory of Planned Behavior

The *Theory of Planned Behavior (TPB)* postulates that an individual's *intention* about a behavior is an important predictor of whether they will engage in this behavior [73]. As a result, the theory identifies three attributes that jointly determine intention: *attitudes*, *subjective norms*, and *perceived behavioral control*. The relative contribution for each predictor is represented by a weight which is often derived empirically using regression analysis based on survey data.

Kaufmann et al. [74] build an agent-based simulation model on TPB to study the diffusion of organic farming practices in two New European Union Member States. Following the TPB methodology, each agent is characterized by three attributes: the attitude a_i, subjective norm s_i, and perceived behavioral control p_i, each ranging from -1 (extremely negative) to +1 (extremely positive). The intention I_i is defined as $I_i = w_i^a a_i + w_i^s s_i + w_i^p p_i$, where w_i^a, w_i^s, w_i^p are relative contribution toward intention. The weights for non-adopters and adopters are derived separately using linear regressions based on the survey data. If an agent's intention exceeds a threshold t it adopts, and does not adopt otherwise. The threshold is obtained from survey data as the average intention of non-adopters who have expressed a desire to adopt. In the simulation model, social influence is transmitted among network neighbors in each time step in a random order. Specifically, when one node speaks

to another, the receiver shifts its subjective norm closer to the sender's intention, following the relative agreement framework. Social networks are generated to reflect small-world properties [68] and a left-skewed degree distribution [75], with specifics determined by a set of parameters, which are set based on survey data (such as the average degree). While empirical data is thus used to calibrate parameters of the model, no quantitative validation was provided.

Schwarz and Ernst [76] propose an agent-based model of diffusion of water-saving innovations, and applied the model to a geographic area in Germany. Agents are households with certain lifestyles, represented by demographic and behavioral characteristics. They use two different decision rules to determine adoption: a cognitively demanding decision rule representing a *deliberate* decision and a simple decision *heuristic*. The particular decision rule to use is selected based on the agent's type and technology category. The deliberate decision-making algorithm is based on multi-attribute subjective utility maximization that integrates attitude, social norm, and perceived behavioral control. The heuristic decision rule makes decisions in greedy order of evaluation criteria based on innovation characteristics and social norms. Finally, if no clear decision can be made, agents imitate their peers, who are defined through a variation of a small-world network [68] which captures spatial proximity and lifestyle affinity in determining links among agents. The model was calibrated using data from a survey according to the framework of the Theory of Planned Behavior [73], with the importance of different decision factors derived by structural equation models or linear regressions for lifestyle groups. The model was validated using *independent* market research data at the household level. In addition, due to the lack of independent aggregated diffusion data, results of the empirical survey were used for validation.

Sopha et al. [77] present an agent-based model for simulating heating system adoption in Norway. Their model extends TPB to consider several contributing factors, such as household groups, intention, attitudes, perceived behavioral control, norms, and perceived

heating system attributes. Households are grouped using cluster analysis based on income level and basic values available in the survey data to approximate the influence of lifestyle on attitudes towards a technology. Attribute parameters are then estimated using regressions for each household cluster based on the household survey. Moreover, motived by the meta-theory of consumer behavior [78], the model assumes that a household agent randomly follows one of four decision strategies: repetition, deliberation, imitation, and social comparison, in accordance with empirical distribution based on survey data. Notably, this model is validated using *independent* data that is not used for calibration, examining how well simulation reproduces actual system behavior at both macro and micro level.

Rai and Robinson [79] develop an empirically grounded agent-based model of residential solar photovoltaic (PV) diffusion to study the design of PV rebate programs. The model is motived by TPB and assumes that two key elements determine adoption decision: attitude and (perceived) control. The authors calibrate population-wide agent attitudes using survey data and spatial regression. Following the opinion dynamics model in Deffuant et al. [72], at each time-step, agents' attitudes about the technology and their uncertainties are adjusted through interactions with their social network neighbors following the relative agreement protocol. Social influence is captured by households situated in small-world networks, with most connections governed by geographic and demographic proximity. In the "control" module, an agent i compares its perceived behavioral control pbc_i with the observed payback at the current time period PP_{it}. Then, if the agent exceeds its attitude threshold, it adopts when $PP_{it} < pbc_i$. pbc_i for each agent i, is calculated as a linear sum of financial resources, the amount of sunlight received, and the amount of roof that is shaded, while PP_{it} is calculated based on electricity expenses offset through the use of the solar system, the price of the system, utility rebates, federal investment tax credit, and annual system electricity generation. The six model parameters used to specify the social network, opinion convergence, the distribution of the behavioral control variable, and the global attitude threshold value, were calibrated by an iterative fitting procedure using historical adoption

data. The model was first validated in terms of predictive accuracy, comparing predicted adoption with empirical adoption level for the time period starting after the last date for the calibration dataset. Moreover, temporal, spatial, and demographic validation were conducted. However, validation was focused on aggregate (macro), rather than individual (micro) behavior.

Jensen et al. [80] develop an agent-based model to assess energy-efficiency impacts of an air-quality feedback device in a German city. A household agent makes two decisions: whether to adopt a feedback device and whether to practice a specific energy-saving behavior. The model involves simulating both the adoption of the feedback device and the heating behavior respectively. Two diffusion processes are connected based on the observation that the feedback device changes an agent's heating behavior, and eventually will form a habit. In the simulations, household agents are generated based on marketing data on lifestyle, and initial adopters of the heating behavior are selected based on a survey. The adoption of an energy-efficient heating behavior is triggered by external events, whose rate is estimated by historical data using Google search queries. Their survey reveals that both information and social influence drive behavior adoption. This insight is integrated into a decision-making model following the theory of planned behavior (TPB), in which information impacts the agent's attitude in each simulation step. On the other hand, the diffusion model of the feedback device is an adaptation of an earlier model also based on TPB. An adopter of the device is assumed to adopt the desired heating behavior with a fixed probability, which is informed by an empirical study. The space of model parameters is reduced by applying a strategy called "pattern-oriented modeling", which refines the model by matching simulation runs with multiple patterns observed from empirical data [81]. In their experiments, the authors calibrated several different models using empirical data and aimed to quantify the effect of feedback devices by comparing results generated by these models. However, no rigorous model validation is presented.

2.2.3.3 Theory of Emotional Coherence

When it comes to explaining and predicting human decisions in a social context, some computational psychology models also take emotional factors into account, which are often neglected by TPB-based models. Wolf et al. [82] propose an agent-based model of adoption of electric vehicles by consumers in Berlin, Germany, based on the *Theory of Emotional Coherence (TEC)*. The parameters of the model were derived based on empirical data from focus groups and a representative survey of Berlin's population. In particular, the focus group provided a detailed picture of people's needs and goals regarding transportation; the survey was designed to generate quantitative estimates of the beliefs and emotions people associate with specific means of transportation. The attributes of the agents include age, gender, income, education, residential location, lifestyle, and a so-called social radius, and are obtained based on the survey data. The social network structure is generated by similarities between these characteristics following the theory of *homophily* [83]; specifically, the likelihood of two individuals communicating with one other is a function of their similarity in terms of demographic factors. To validate the predictions made by the model, the authors regressed empirical data related to actual transportation-related decisions (e.g., weekly car usage) from the survey on the activation parameters resulting from simulations. However, validation did not use independent data.

2.2.3.4 Consumat Model

The Consumat Model is a social psychological framework, in which consumer agents switch among several cognitive strategies—commonly, *comparison*, *repetition*, *imitation*, and *deliberation*—as determined by need satisfaction and their degree of uncertainty [84]. Schwoon [85] uses an agent-based model (ABM) to simulate possible diffusion paths of fuel cell vehicles (FCVs), capturing complex dynamics among consumers, car producers, and filling station owners. In their model, the producers offer heterogeneous but similar cars, deciding in each period whether to change production to FCVs. Consumers have vary-

ing preferences for car attributes, refueling needs, and social influence factors. Although in a typical consumat approach [86], consumers follow one of four cognitive strategies on the basis of their level of need satisfaction and uncertainty, the author rules out repetition and imitation and argues that need satisfaction is rather low in their case. The consumer is assumed to maximize total expected utility, which is expressed as a function of car price, tax, the closeness between preferences and car characteristics, social need, as determined by the fraction of neighbors adopting each product type, and availability of hydrogen. In the model, individual preferences may evolve with time to be more congruent with the "average car", as determined by a weighted average of attributes of cars sold in the previous period, where weights correspond to market shares. The model is calibrated by trying to match main features of the German auto market. The network structure governing social influence is assumed to form a torus. The model does not attempt quantitative validation.

2.2.3.5 The LARA Model

LARA is the short for *Lightweight Architecture for boundedly Rational Agents*, a simplified cognitive agent architecture designed for large-scale policy simulations [87]. Comparing with existing complex psychological agent frameworks, LARA is more generalizable and easier to implement. We review two recent efforts motivated by the LARA architecture and grounded in empirical data.

Krebs et al. [88] develop an agent-based model to simulate individual's provision of neighborhood support in climate change adaptation. In their model, agents are assigned to lifestyle groups and initialized using spatial and societal data. Motivated by LARA, an agent makes decision in one of three modes: deliberation, habits, and exploration. In deliberation, an agent compares and ranks available options in terms of utility, which is the weighted sum of four goals: striving for effective neighborhood support, being egoistic, being altruistic, and achieving social conformity. The goal weights, which are different among lifestyle groups, are set based on expert ratings and the authors' prior work. A

probability choice model is used to choose the final option when multiple better options are available. An agent acts in deliberation mode if no experience is available (habitual behavior is not possible) and shifts to the exploratory mode with a predefined small probability. The network in which the agents are embedded is generated using lifestyle information. Simulation runs for an initial period from 2001 to 2010 provide plausible results on behavioral patterns in cases of weather changes. From 2011 to 2020, the authors examine the effects of two intervention strategies that mobilize individuals to provide neighborhood support. Some model parameters remain uncalibrated, and the entire model is not validated due to a lack of empirical data at the macro level.

Krebs and Ernst [89] develop an agent-based spatial simulation of adoption of green electricity in Germany. Each agent represents a household deciding to select between "green" and "gray" energy providers. Every agent is characterized by its geographical location and lifestyle group. Agents are initialized and parameterized by empirical data from surveys, psychological experiments, and other publicly available data. Following LARA, agents are assumed to make decisions either in a deliberative or habitual mode. Default agent behavior is habitual, and the agent transitions to a deliberative mode when triggered by internal and external events, such as a price change, personal communication, cognitive dissonance, need for cognition, and media events. An agent chooses an action that maximizes utility, which is a weighted sum of four goals: ecological orientation, economic orientation, social conformity, and reliability of provision. The goal weights depend on the lifestyle group and are derived from a survey and expert rating [90]. An artificial network that connects the agents is generated based on lifestyle and physical distance [90]. Once an agent decides to adopt green electricity, it chooses a service brand that is already known. The diffusion of the awareness of the brand is characterized by a simple word-of-mouth process. Validation focuses on two state variables of agent behavior: selected electricity provider and awareness of the brand, which involves comparing simulation results with historical data both temporally and spatially starting from aggregate to the individual level.

Unfortunately, validation was not conducted using independent data.

2.2.4 Heuristic Models

Heuristic adoption models are often used when modelers are not aware of any established theories for agent decision-making in the studied application. These models tend to give us an impression of being "ad-hoc", since they are not built on any grounded theories. More importantly, unlike the cognitive agent models such as the theory of planned behavior, there is no established or principled means to estimate model parameters. Therefore, model parameters are often selected in order to match simulated output against a realistic adoption level. Although heuristic-based model appears to be an inaccurate representation of agent decision-making, they are easy to implement and interpret.

Van Vliet et al. [91] make use of a *take-the-best* heuristic to model a fuel transportation system to investigate behavior of fuel producers and motorists in the context of diffusion of alternative fuels. In the model, producers' plant investment decision is determined by simple rules, and the same plant can produce multiple fuel types. Motorists are divided into several subgroups, each having distinct preferences. Each motorist is assumed to choose a single fuel type in a given year. Each fuel is assigned four attributes: driving cost, environment, performance, and reputation. Motorist preferences in the model are represented by two factors: 1) *priorities*, or the order of perceived importance of fuel attributes, and 2) *tolerance* levels, which determine how much worse a particular attribute of the corresponding fuel can be compared to the best available alternative to maintain this fuel type under consideration. The decision heuristic then successively removes the worst fuel one at a time in the order of attribute priorities. Due to the difficulty of obtaining actual preferences of motorists, the authors used the Dutch consumer value dispositions from another published model in literature as a proxy to parametrize the model. However, the model was not rigorously calibrated or validated using empirical data.

Zhao et al. [92] propose a two-level agent-based simulation modeling framework to analyze the effectiveness of policies such as subsidies and regulation in promoting solar photovoltaic (PV) adoption. The lower-level model calculates payback period based on PV system electricity generation and household consumption, subsidies, PV module price, and electricity price. The higher-level model determines adoption choices as determined by attributes which include payback period, household income, social influence, and advertising. A pivotal aspect of the model is the *desire* for the technology (PV), which is formulated as a linear function of these four factors, and an agent adopts if the desire exceeds a specified threshold. Survey results from a prior study were used to derive a distribution for each factor, as well as the membership function in a fuzzy set formulation. The agents in the model were initialized using demographic data, along with realistic population growth dynamics based on census data. Moreover, calibration of threshold value was conducted to match simulated annual rate of PV adoption with historical data. However, the model was not quantitatively validated using independent data.

A more complex TOPSIS (*Technique for Order Preference by Similarity to Ideal Solution*) model is a decision heuristic which selects an option from several alternatives that is the closest to the ideal option and the farthest from the worst possible option. Kim et al. [93] present agent-based automobile diffusion model using a TOPSIS approach to simulate market dynamics upon introduction of a new car in the market. The model integrates three determinants of purchasing behavior: (1) information offered by mass media, (2) relative importance of attributes to consumers, and (3) social influence. Individual agents rank products by considering multiple product attributes and choosing a product closest to an ideal. A survey was conducted to estimate consumers' weights on the car attributes and the impact of social influence. In the simulations, diffusion begins with innovators who try out new products before others; once they adopt, their social network neighbors become aware of these decisions, with some deciding to adopt, and so on. A *small-world* network structure was assumed for this virtual market, and choices of rewiring and connectivity were

determined by the model calibration step through comparing simulated results with historical monthly sales volumes of three car models. However, the model was not validated using independent data.

2.2.5 Statistics-Based Models

Statistics-based models rely on statistical methods to infer relative contribution of observable features towards one's decision whether to adopt. The estimated model is then integrated into an ABM. We review three subcategories of statistics-based methods for agent-based models of innovation diffusion: *conjoint analysis*, *discrete choice models*, and *machine learning*.

2.2.5.1 Conjoint Analysis

Conjoint analysis is a statistical technique used in market research to determine how much each attribute of a product contributes to consumer's overall preference. This contribution is called the *partworth* of the attribute. Combining with feature values of innovation obtained from the field study, one can construct a utility function accordingly.

Garcia et al. [94] utilize conjoint analysis to instantiate and calibrate an agent-based marketing model using a case study of diffusion of Stelvin wine bottle screw caps in New Zealand. With a particular emphasis on validation, the overall work follows Carley [95]'s four validation steps: *grounding*, *calibration*, *verification*, and *harmonizing* (the latter not performed, but listed as future work) to properly evaluate the model at both micro and macro levels. The model includes two agent types: wineries and consumers. In each period the wineries set the price, production level, and attributes of screw caps as a function of consumer demand. Consumers, in turn, make purchase decisions following their preferences. The model is calibrated using conjoint analysis, inferring *partworths* which determine consumer preferences in the model. Aggregate stylized facts were then replicated in the verification step. The work emphasizes the value of calibration, but pays less

attention to validation, which is merely performed at a face level rather than quantitatively.

Vag [96] presents a *dynamic* conjoint method that enables forecasts of future product preferences. The consumer behavior model considers many factors, including social influence, communication, and economic motivations. The author surveys behavior of individuals, such as their communication habits, and uses conjoint analysis to initialize preferences in the ABM. Notably, in this model agent priorities depend on one another, and the resulting social influence interactions may lead to large-scale aggregate shifts in individual priorities. To demonstrate the usability of their model, the study utilized empirical data on product preferences (in this case, preferences for mobile phones), consumer habits, and communication characteristics in a city in Hungary. Calibration of this model was only based on expert opinion and comparative analysis, rather than quantitative comparison with real data, and no quantitative validation was performed.

Zhang et al. [97] develop an agent-based model to study the diffusion of eco-innovations, which in their context are alternative fuel vehicles (AFVs). The model considers interdependence among the manufacturers, consumers, and governmental agencies in the automotive industry. The agents representing manufacturers choose engine type, fuel economy, vehicle type, and price, following a simulated annealing algorithm, to maximize profit in a competitive environment until a Nash equilibrium is reached [98]. The consumer agents choose which products to purchase. The partworth information in the utility function was derived by *choice-based conjoint* analysis using an empirical survey from Garcia et al. [94]. In particular, the probability of a consumer choosing a vehicle is formulated as a logit function of vehicle attributes, word-of-month, and domain-specific knowledge. The utility is modeled as a weighted sum of attributes, and parameters/partworth are estimated using hierarchical Bayes methods. The agent acting as "government" chooses policies aimed at influencing the behavior of both manufacturers and consumers. Model calibration involved conjoint analysis. However, the authors found that the ABM tended to overestimate the market shares of alternative fuel vehicles, which motivated them to adjust model pa-

rameters and to linearize the price parthworth in order to ensure that aggregate demand decreases with price. Like Garcia et al. [94], the authors follow the four steps of validation [95]. However, validation does not use data independent from calibration.

Lee et al. [99] introduce an agent-based model of energy consumption by individual homeowners to analyze energy policies in the U.K. The model utilizes historical survey data and choice-based conjoint analysis to estimate the weight of a hypothetical utility function, defined as the weighted sum of attributes. In the simulation, moving and boiler break-down events are assumed to trigger a decision by the household agent. In this case, a particular alternative is selected if its utility is higher than all other alternative as well as the status quo option. The model was populated with initial data based on a survey conducted in the U.K., and each agent was matched to a household type which can be further mapped to energy demand using energy consumption estimates. The authors then combined energy demand with fuel carbon intensity to determine annual household emissions. The model was calibrated by adjusting the weights in the decision model to match historic installation rates from 1996 to 2008 for loft insulation and cavity wall insulation. The model was not validated using independent data.

Stummer et al. [100] devise an agent-based model to study the diffusion of multiple products. Each product is characterized by a number of attributes determined by expert focus group discussion. True performance of each product attribute is unknown to consumers, and each agent, therefore, keeps track of the distribution of attribute values based on information previously received. This information is updated based on interactions with peers, advertising, or direct experience. Consumer agent behavior is governed by a set of parameters that capture heterogeneous preferences and mobility behavior. Agents have additive multi-attribute utilities, the weights of which were obtained from survey data using conjoint analysis. The authors adapt the preferential attachment algorithm introduced by [101] to generate networks in which the attachment probability depends on both node degree and geographic distance between nodes. Network parameters were determined by

taking into account additional information revealed in the consumer survey, such as the number of social contacts and communication frequency. An agent decides to purchase a product which maximizes utility. The model defines each advertising event to communicate a set of product attributes, which either increase product awareness or impact customer preferences. The model was validated extensively following [102], including conceptual validity, internal validity, micro-level external validity, macro-level external validity, and cross-model validity. The weakness of validation, however, is that it is only performed as an in-sample exercise without using independent data.

2.2.5.2 Discrete Choice Models

The *discrete choice* modeling framework, which originates in econometrics, is used to describe, explain, and predict agent choices between two or more discrete alternatives [103]. The approach has a wide range of applications, and we review several efforts targeted specifically at innovation diffusion.

Galán et al. [104] design an agent-based model to analyze water demand in a metropolitan area. This model is an integration of several sub-models, including models of urban dynamics, water consumption, and technological and opinion diffusion. The opinion diffusion model assumes that an agent's attitude towards the environment determines its water consumption, i.e., a non-environmentalist would use more water than an environmentalist. Accordingly, it is assumed that each agent can be in two states: *environmentalist* (E) or *non-environmentalist* (NE). The choice of a state depends on the agent's current state, the relative proportion of E and NE neighbors, and an exogenous term measuring the pressure towards E behavior. Transition probabilities between states E and NE are given in form of *logistic* functions. However, rather than using empirical data to estimate parameters of these functions, the authors parameterized the behavior diffusion model with reference to models in prior literature for other European cities. To determine adoption of water-saving technology, the opinion diffusion model is coupled with the technological diffusion

model, which is implemented by a simple agent-based adaptation of the Bass model following [105]. The model was validated qualitatively by domain experts, quantitatively calibrated based on the first quarter of 2006, and validated by comparing the model with actual adoption in the following two quarters. The authors demonstrate that simulation results successfully replicate the consequence of a water-saving campaign on domestic water consumption.

Dugundji and Gulyás [106] propose a computational model that combines econometric estimation with agent-based modeling to study the adoption of transportation options for households in a city in Netherlands. The presented discrete choice modeling framework aims to address interactions within different social and spatial network structures. Specifically, agent decision is captured using a *nested logit* model, which enables one to capture observed and unobserved behavior heterogeneity. Feedback effects among agents are introduced by adding a linear term (a so-called *field variable*) that captures proportions of an agent's neighbors making each decision to each agent's utility function. Because survey data on interactions between identifiable individuals was unavailable, this term only captured aggregate interactions among socioeconomic peers. The authors investigated simulated transition dynamics for the full model with two reference models: the first a nested logit model with a global field variable only and a fully connected network, and the second a multinomial logit model which is a special case to the full model. They found that simulated dynamics differ dramatically between the models. Given this lack of modeling robustness, no further validation was undertaken.

Tran [107] develops an agent-based model to investigate energy innovation diffusion. Agent behavior in this model is determined by the relative importance of technology attributes to the agents, and social influence. Social influence, in turn, takes two forms: indirect influence coming from the general population, and direct influence of social network neighbors. The author drew on ABM studies in the marketing literature, and formulated the adoption model as $Prob(t) = 1-(1-P_{ij})(1-Q_{ij})^{K_{ij}}$, where P_{ij} captures individual choice

using a discrete choice model of consumer decision-making, in which an agent's utility is defined as an inner product of coefficients and attributes. Coefficients are a random vector, with distribution different for different agents, capturing preference heterogeneity. Q_{ij} and K_{ij} is the indirect and direct network influence, respectively, captured as a function of the number of adopters at decision time. While the model was evaluated using simulation experiments, and the nature of the model makes it well suited for empirically grounded parameter calibration, it was not in actuality quantitatively calibrated or validated using empirical data.

2.2.5.3 Machine Learning Models

Machine learning (ML) is a sub-area of computer science that aims to develop algorithms that uncover relationships in data. Within a supervised learning paradigm which is of greatest relevance here, the goal is further to develop models that accurately predict the value of an outcome variable for unseen instances. To do so, a computer program is expected to recognize patterns from a large set of observations, referred to as a *training* process that is grounded in statistical principles and governed by intelligent algorithms, and make predictions on new, unseen, instances. This category of methods has recently drawn much attention in academia and industry due to tremendous advances in predictive efficacy on important problems, such as image processing and autonomous driving. Combining machine learning with agent-based modeling seems promising in the study of innovation diffusion since the two can complement each other. The former is specialized in building a high-fidelity predictive models, while the latter captures dynamics and complex interdependencies. Of particular relevance to combining ML and ABM is the application of machine learning to model and predict human behavior. Interestingly, relatively few attempts have been made to date to incorporate ML-based models of human behavior within ABM simulations.

Sun and Müller [108] develop an agent-based model that features Bayesian belief networks (BBNs) and opinion dynamics models (ODMs) to model land-use dynamics as they relate to payments for ecosystem services (PES). The decision model of each household is represented using a BBN, which were calibrated using survey data and based on discussions with relevant stakeholders, and incorporate factors such as income and land quality. Social interactions in decision-making are captured by ODM. The modeling framework was applied to evaluate China's Sloping Land Conversion Program (SLCP), considered among the largest PES programs. SLCP was designed to incentivize reforestation of land through monetary compensation. In their model, farmers make land-use decisions whether or not to participate in the SLCP program based on *internal* beliefs and *external* influences. External influences adjust internal beliefs cumulatively using a modified Deffuant model [72] within a community-based small-world social network. Initial model structures were obtained using a structural learning algorithm, with results augmented using qualitative expert knowledge, resulting in a pseudo tree-augmented naive Bayesian (TAN) network. The final BBN model was validated by using a sensitivity analysis, and measuring prediction accuracy and area under the curve (AUC) of the receiver operating characteristics (ROC) curve on a holdout test data set at both household and plot level. A crucial limitation of this work is that only the BBN model was carefully validated; the authors did not validate the full simulation model at either the micro or macro levels.

Zhang et al. [30] propose a data-driven agent-based modeling (DDABM) framework for modeling residential rooftop solar photovoltaic (PV) adoption in San Diego county. In this framework, the first step is to use machine learning to calibrate individual agent behavior based on data comprised of individual household characteristics and PV purchase decisions. These individual behavior models were validated using cross-validation methods to ensure predictive efficacy on data not used for model calibration, and were then used to construct an agent-based simulation with the learned model embedded in artificial agents. In order to ensure validation on independent data, the entire time series data of in-

dividual adoptions was initially split along a time dimension. Training and cross-validation for developing the individual-level models were performed only on the first (early) portion of the dataset, and the aggregate model was validated by comparing its performance with actual adoptions on the second, independent time series, into the future relative to the calibration data set. The authors thereby rigorously demonstrate that the resulting agent-based model is effective in forecasting solar adoption both at the micro and macro levels. To our best knowledge, this work proposed the first generic principled framework that combines ML and ABM in the study of innovation diffusion. Unlike most ABM studies we have reviewed, DDABM has the following features: 1) it does not make any assumptions on the structural features of social network, relying entirely on a data-driven process to integrate most predictive spatial and social influence features into the individual adoption model; 2) it does not rely on matching simulated dynamics with the empirical observations to calibrate the model, but instead parameterizes the model through a far more efficient statistical learning method at the level of individual agent behavior; and 3) validation is performed on *independent* data to evaluate the predictive effectiveness of the model. Moreover, validation is not only done at the macro-level by comparison with actual adoption traces, but also at the micro-level by means of the simulated *likelihood ratio* relative to a baseline model. To further justify the usefulness of ML-base approach, Zhang et al. [30] actually implement and compare their model with another agent-based model of rooftop solar adoption developed by [67], with parameters calibrated on the same dataset following the general aggregate-level calibration approach used by them. The result is very revealing, as it strongly suggests that aggregate-level calibration is prone to overfit the model to data, an issue largely avoided by calibrating individual agent behavior.

2.2.6 Social Influence Models

Our last methodological category covers several models looking specifically at social influence. These models are quite simple, abstract, but prevalent in the theoretical study of

innovation diffusion. Our purpose of discussing these is that there have been several recent efforts to calibrate these models using empirical data.

After analyzing an adoption dataset of Skype, Karsai et al. [109] develop an agent-based model to predict diffusion of new online technologies. Specifically, agents in their model are characterized by three states: *susceptible* (S), *adopter* (A), and *removed* (R). Susceptible refers to people who may adopt the product later. Adopter agents have already adopted. Finally, removed are those who will not consider adopting the product in the future again. The transition from S to A is regulated by *spontaneous adoption* and *peer-pressure*, from A to S by *temporary termination*, and from A to R by *permanent termination*, each of which is parametrized by a constant probability which is identical for all users. While some parameters, such as average degree and temporary termination probability, are estimated directly from observations, the remaining parameters are determined by simultaneously fitting the empirical rates using a bounded nonlinear least-squares method. The model is fit over a 5-year training period, and validation uses predictions over the last six months of the observation period. However, validation is somewhat informal, since the predictability of the model is evaluated on a part of the training data and there is no validation of micro-behavior. In a later work using the same Skype data, Karsai et al. [110] develop a threshold-driven social contagion model with only two states: susceptible and adopted. In addition, the model assumes that some fraction of nodes never adopt. The authors calibrated the value of this fraction by matching the size of the largest component of adopters given by the simulations with real data. In addition, the model assumes that susceptible nodes adopt with a constant probability, which is informed by empirical analysis. In their simulations, nodes have heterogeneous degrees and thresholds, which follow empirical distributions. However, validation was not performed using independent data.

Rand et al. [111] present two agent-based models of diffusion dynamics in online social networks. The first ABM is motivated by the Bass model, but time is discretized and each agent has two states: *unaware* and *aware*. At each time step, an unaware agent changes

state to aware as a function of two triggers: innovation arising from exogenous sources, such as advertising, and imitation, which comes from observing decisions by neighbors. The second model termed the *independent cascade* model, originating from Goldenberg et al. [17], has the similar structure to the agent-based Bass model, except that the imitation effect is formulated as a single probability with which each aware neighbor can independently change the state of an agent to *aware*. The author applied the two models in parallel to four diffusion data sets from Twitter, and calibrated parameters using actual aggregate adoption paths. Notably, validation is only performed at macro-level as an in-sample exercise, and shows that the two models behave similarly.

Using historical diffusion data of Facebook apps, Trusov et al. [112] introduce an approach that applies Bayesian inference to determine a mixture of multiple network structures. Notice that most ABMs we reviewed so far either assume a single underlying social network (with parameters determined in model calibration) or generate artificial networks based on empirical findings or social science theories. They first choose a collection of feasible networks that represent the unobserved consumer networks. Then, a simple SIR model (similar to the Bass ABM in [111]) is used to simulate the diffusion of products. The simulated time series are further transformed to multivariate stochastic functions, which provide priors to the Bayesian inference model to obtain the posterior weights on the set of feasible consumer networks. Like [111], the adoption model is calibrated from the aggregate output, rather than from observations of individual decisions.

Chica and Rand [113] propose an agent-based framework to build decision support system (DSS) for word-of-mouth programs. They developed a DSS to forecast the purchase of a freemium app and evaluate marketing policies, such as targeting and reward. The model captures seasonality of user activities by two probabilities for weekday and weekend respectively. The initial social network is generated by matching the degree distribution of the real network. Then, for each node, two weights are assigned to in- and out-edges, respectively, turning the network into a weighted graph that represents the heterogeneous

social influence among social neighbors. Specifically, two models are used to model the information diffusion. One is the Bass-ABM ([114]); the other is a contagion model (a threshold model but adding external influence). The parameters of the model were calibrated by a genetic algorithm [115], in which the fitness is defined based on the difference of simulated adoption from the historical adoption trajectory. Notably, the model was validated by a hold-out dataset, which is independent of the training data. For example, the entire 3 month period spanned by the data was divided into two: first 60 days for training, the last 30 days for validation.

The independent cascade model used by Rand et al. [111] and the threshold model used by Chica and Rand [113] are significant insofar as these connect to a substantial literature that has recently emerged within the Computer Science community on *information diffusion*, whereby information (broadly defined) spreads over a social network. We make this connection more precisely in Section 2.4 below.

2.3 Categorization of Innovation Diffusion Models by Application

Thus far, we followed a categorization of agent-based models of innovation diffusion focused on methods by which agent behavior is modeled. First, we observe that methods range from sophisticated mathematical optimization models (Section 2.2.1), to economic models (Section 2.2.2), to even simpler models based on heuristics for representing agent behavior (Section 2.2.4). While economic factors are dominant concerns in some applications, others emphasize the cognitive aspects of human decision-making (Section 2.2.3) and are frequently used to model influence over online social networks (Section 2.2.6). Second, we note that the method chosen to capture agent behavior also impacts the techniques used to calibrate model parameters from data. For example, cognitive models are often constructed based on detailed behavior data collected from field experiments and surveys, whereas models of agent behavior based on statistical principles rely on established statistical inference techniques for model calibration based on individual behavior data that

is either observational or experimental. Other modeling approaches within our six broad categories often do not use data to calibrate individual agent behavior, opting instead to tune model parameters in order to match aggregate adoption data.

We now offer an alternative perspective to examine the literature on empirical ABMs of innovation diffusion by considering applications—that is, what particular innovation is being modeled. A breakup of existing work using this dimension is given in Table 2.2. As shown in the first column, we group applications by broad categories: agricultural innovations and farming, sustainable energy and conservation technologies, consumer technologies and innovations, information technologies and social goods. Interestingly, the first two categories account for more than half of the publications in literature. This likely reflects the history of ABM as an interdisciplinary modeling framework for computational modeling of issues that are of great interest in social science. A closely related factor could be the relatively high availability of data in these applications generated by social scientists (e.g., through the use of surveys). Another interesting observation that arises is methodological convergence for given applications: relatively few applications have been modeled within different methodological frameworks as categorized above. Future research may explore the use of different methods for same application. Furthermore, comparison of different modeling methods is rare within a single work (except in [106, 30]), although such a methodological cross-validation is of importance as emphasized by some authors [95, 114].

Category	Application	Method	Citation
agricultural innovations and farming	agricultural innovations	mathematical programming	Berger [52], Schreinemachers et al. [54], Berger et al. [53], Schreinemachers et al. [55, 57], Alexander et al. [58]
		economic (utility)	Holtz and Pahl-Wostl [64]
	organic farming	cognitive model (Deffuant)	Deffuant et al. [71]
		cognitive model (TPB, Deffuant)	Kaufmann et al. [74]
	biogas plant	economic (profit)	Sorda et al. [61]
	payments for ecosystem services	machine learning	Sun and Müller [108]
sustainable energy and conservation technologies	water-saving technology	cognitive model (TPB)	Schwarz and Ernst [76]
		discrete choice model	Galán et al. [104]
	heating system	cognitive model (TPB)	Sopha et al. [77]
		conjoint analysis	Lee et al. [99]

		economic (cost)	Faber et al. [60]
	solar photovoltaic	heuristic	Zhao et al. [92]
		economic (utility)	Palmer et al. [67]
		cognitive model (TPB, Deffuant)	Rai and Robinson [79]
		machine learning	Zhang et al. [30]
	fuel cell vehicles	cognitive model (Consumat)	Schwoon [85]
	energy innovation	discrete choice model	Tran [107]
	electric vehicles	cognitive model (TEC)	Wolf et al. [82]
		economic (utility)	Plötz et al. [65]
		economic (utility)	McCoy and Lyons [66]
	alternative fuel vehicles	conjoint analysis	Zhang et al. [97]
	alternative fuels	heuristic	Van Vliet et al. [91]
		economic (utility)	Günther et al. [63]
		conjoint analysis	Stummer et al. [100]
	green electricity	cognitive model (LARA)	Krebs and Ernst [89], Ernst and Briegel [90]
	air-quality feedback device	cognitive model (TPB)	Jensen et al. [80]
consumer technologies and innovations	wine bottle closures	conjoint analysis	Garcia et al. [94]
	mobile phones	conjoint analysis	Vag [96]
	transportation mode	discrete choice model	Dugundji and Gulyás [106]
	new cars	Fuzzy TOPSIS (heuristic) Model	Kim et al. [93]
	movie	economic (utility)	Broekhuizen et al. [62]
information technologies	Skype	social contagion model	Karsai et al. [109, 110]
	Twitter	independent cascade model	Rand et al. [111]
	Facebook app	social contagion model	Trusov et al. [112]
	freemium app	social contagion model	Chica and Rand [113]
social goods	neighborhood support	cognitive model (LARA)	Krebs et al. [88]

Table 2.2: Categorization of surveyed work by Applications

2.4 Information Diffusion Models

Online social networks have emerged as an crucial medium of communication. It does not only allow users to produce, exchange, and consume information at an unprecedented scale and speed, but also speeds the diffusion of novel and diverse ideas [116, 117]. The emergence of online social networks and advances in data science and machine learning have nourished a new field: *information diffusion*. The fundamental problem in information diffusion is to model and predict how information is propagated through interpersonal connections over social networks using large-scale diffusion data. In fact, several authors have reviewed the topic of information diffusion over online social networks [118, 116, 119].

Our aim is not to provide a comprehensive review of this same topic. Instead, we are interested in building connections between the agent-based modeling approach to innovation diffusion, and the modeling methods in the field of information diffusion. Indeed, researchers in the ABM community have paid little attention to the existing methods for modeling information diffusion, and especially in the played by data science in this field, which has significant implications for ABM model calibration, as we discuss below.

2.4.1 Two Basic Models of Information Diffusion

Compared to agent adoption models in Section 2.2, the decision process in the information diffusion literature is typically very simple, following predominantly the social influence models. The two most common models in information diffusion are Independent Cascades (IC) [17] and Linear Threshold (LT) models [18]. These models are defined on a directed graph where activation is assumed to be monotonic: once a node is active (e.g., adopted, received information), it cannot become inactive. The diffusion process in both models starts with a few active nodes and progresses iteratively in a *discrete* and *synchronous* manner until no new nodes can be infected. Specifically, in each iteration, a new active node in the IC model is given a single chance to activate its inactive neighbors independently with an exogenously specified probability (usually represented by the weight of the corresponding edge). In the LT model, in contrast, an inactive node will become active only if the sum of weights of its activated neighbors exceeds a predefined node-specific threshold, which is typically randomly assigned between 0 and 1 for each network node. Note that in both models a newly activated node becomes active immediately in the next iteration. From an agent-based perspective, both IC and LT are generative models which define two diffusion mechanisms.

2.4.2 Learning Information Diffusion Models

Several efforts use empirical data to calibrate the parameters of the LT and IC models. Saito et al. [120] propose an asynchronous IC (AsIC) model, which not only captures temporal dynamics, but also node attributes. They show how the model parameters can be estimated from observed diffusion data using maximum likelihood estimation (MLE). The AsIC model closely follows the IC model, but additionally introduces a time delay before a newly activated node becomes active. The time delay is assumed to be exponentially distributed with a parameter that is defined as an exponential function of a feature vector (a composition of attributes associated with both nodes and edges). The transmission probability is then defined as a logit function of the feature vector. The data is given in the format of "diffusion traces", and each trace is a sequence of tuples which specify activation time for a subset of nodes. To learn the model using this data, the authors define the log-likelihood of the data given the model. The authors then demonstrate how to solve the resulting optimization problem using expectation-maximization (EM). While the proposed model is promising to be used for prediction, the learning method was only tested using synthetic data.

Guille and Hacid [121] show how to parameterize the AsIC model using machine learning methods based on Twitter data. In their model, the diffusion probability for information at any given time between two users is a function of attributes from three dimensions: *social*, *semantic*, and *time*, which group features with respect to social network, content and temporal property respectively. Four different classifiers were trained and compared in terms of cross-validation error: C4.5 decision tree, linear perceptron, multilayer perceptron, and Bayesian logistic regression. The last model mentioned above was finally used for prediction. Notably, time-delay parameter was determined separately in this work by comparing simulation results with actual diffusion dynamics, which is the same calibration method commonly used in ABM of innovation diffusion. Unlike [120], where all model parameters are inferred by MLE, here only a subset of model parameters are estimated

through established machine learning techniques, but the rest are calibrated by simulations. Their evaluation shows that the model accurately predicts diffusion dynamics, but fails to accurately predict the volume of tweets. In our ABM jargon, the model performs well at macro-level, but poorly at micro-level validation [95]. Another limitation of this work is that validation is only performed as an in-sample exercise, rather than using out-of-sample data.

Galuba et al. [122] propose two diffusion models with temporal features that are used to predict user re-tweeting behaviors on Twitter. Both models define the probability for a user to re-tweet a given URL to be a product of two terms: one is *time-independent*, the other is *time-dependent*. Both have the same time-dependent part which follows a log-normal distribution, but differ in the actual definitions of the time-independent part. In their first model termed At-Least-One (ALO), the time-independent component is defined as the likelihood of at least one of the causes: either one is affected by the agent it follows, or by the user tweets a URL spontaenously. The second, Linear Threshold (LT), model, posits that a user re-tweets a URL only if the cumulative influence from all the followees is greater than a threshold. The time-independent component in this model is given by a sigmoid function. In order to calibrate and validate the model, the data set was split along the time dimension into two parts. The model was calibrated by choosing parameters that optimize the estimated F-score using the gradient ascent method on the first (earlier) data set, and used to predict URL mentions in the second (later) data set. Their results show that the LT model achieves the highest F-score among all models and correctly predicts approximately half of URL mentions with lower than 15% false positives.

While all research reviewed so far assumes known network structure, a number of efforts deal with hidden network structures which must also be learned from data. The so-called *network inference problem* is to infer the underlying network given a complete activation sequence [116]. Gomez Rodriguez et al. [123] introduce a variant of the independent cascade model [124] adding time delay. Their problem is to find a directed graph with at

most k edges that maximizes the likelihood of a set of cascades for a given transmission probability and parameters of the incubation distribution, which is solved approximately using a greedy algorithm. Myers and Leskovec [125] propose a cascade model which is similar to Gomez Rodriguez et al. [123] but allows distinct transmission probabilities for different network edges. The goal is to infer the adjacency matrix (referring to the pairwise transmission probabilities) that maximizes the likelihood given a set of cascades, which is accomplished by solving a convex optimization problem derived from the problem formulation. Gomez Rodriguez et al. [126] develop a continuous-time diffusion model that unifies the two-step diffusion process involving both a transmission probability and time delay from Gomez Rodriguez et al. [123] and Myers and Leskovec [125]. The pivotal value is the *conditional* probability for a node i to be infected at time t_i given that a neighboring node j was infected at time t_j, which is formulated as a function of the time interval $(t_i - t_j)$ and parametrized by a pairwise transmission rate α_{ji}. Survival analysis [127] is used to derive the maximum likelihood function given a set of cascades, and they aim to find a configuration of all transmission rates that maximizes the likelihood. While most network inference algorithms assume static diffusion networks, Gomez Rodriguez et al. [128] address a network inference problem with a time-varying network. The resulting inference problem is solved using an online algorithm upon formulating the problem as a stochastic convex optimization.

2.4.3 Bridging Information Diffusion Models and Agent-Based Modeling of Innovation Diffusion

The methodological framework of the information diffusion inference problems discussed above is a natural fit for principled data-driven agent-based modeling. The information diffusion models characterized by transmission probabilities and time delay are essentially agent-based models. Given data of diffusion cascades, they can be constructed either using only the temporal event (adoption) sequence, or using more general node features,

social network, content, and any other explanatory or predictive factors. In fact, ABM researchers have started to apply similar statistical methods to develop empirical models (see Section 2.2.5). Notably, as shown by Zhang et al. [30], parametric probabilistic models of agent behavior can be estimated from observation data using maximum likelihood estimation methods. In addition, the approaches for network inference appear particularly promising in estimating not only behavior for a known, fixed social influence network, but for estimating the influence network itself, as well as the potentially heterogeneous influence characteristics.

A crucial challenge in translating techniques from information diffusion domains to innovation diffusion is that the latter only observes a single, partial adoption sequence, rather than a collection of complete adoption sequences over a specified time interval. As a consequence, the fully heterogeneous agent models cannot be inferred, although the likelihood maximization can still be effectively formulated by limiting the extent of agent heterogeneity (with the limit of homogeneous agents used by Zhang et al. [30]). In addition, the assumptions generally made in information diffusion models can also pose serious challenges to the transferability of the approach to agent-based modeling. Recall that information cascade models often assume that an adopter has a *single* chance to affect its inactive neighbors and a non-adopter is affected by its neighboring adopters *independently*. These assumptions simplify the construction of the likelihood function, but further justification is needed for them, especially when building empirical models that are expected to faithfully represent realistic social systems and diffusion processes. Note that rules that govern the interactions in agent-based models are quite flexible and can be very sophisticated, which is also one of the major advantages of agent-based computing over analytical models. Although one may be able to explicitly derive a parametric likelihood function given diffusion traces in more complex settings than existing information diffusion models do, this is sure to be technically challenging. Moreover, solving the resulting MLE can be computationally intractable. Therefore, to take advantage of MLE approach in information diffusion, ABM

researchers must make appropriate assumptions on agent interactions so that they can derive tractable likelihood functions without significantly weakening the model's explanatory and predictive power.

2.5 Discussion

2.5.1 Validation in Agent-Based Modeling

As agent-based modeling is increasingly called for in service of decision support and prediction, it is natural to expect them to be empirically grounded. An overarching consideration in empirically grounded agent-based modeling is how data can be used in order to develop reliable models, where reliability is commonly identified with their ability to accurately represent or predict the environment being modeled. This property of *reliability* is commonly confirmed through model *validation*. In social science, a number of authors have contributed to the topic of validation, from approaches for general computational models [95], to those focused on agent-based simulations [129, 130, 94, 131, 114], to specific types of agent-based models [132]. Outside of social science, validation of simulation systems has an even longer history of investigation [102, 133, 134, 135]. We now briefly review these approaches.

As previously mentioned, Carley [95] suggests four levels of validation: *grounding*, *calibration*, *verification*, and *harmonizing*. *Grounding* establishes reasonableness of a computational model, including face validity, parameter validity, and process validity; *calibration* establishes model's feasibility by tuning a model to fit empirical data; *verification* demonstrates how well a model's predictions match data; and *harmonization* examines the theoretical adequacy of a verified computational model.

More recently, drawing on formal model verification and validation techniques from industrial and system engineering for discrete-event system simulations, Xiang et al. [129] suggest the software implementation of agent-based model has to be verified with respect to

its conceptual model, and highlight a selection of validation techniques from Banks [133], such as face validation, internal validation, historical data validation, parameter variability, predictive validation, and Turing tests. Moreover, they suggest the use of other complementary techniques, such as model-to-model comparison [136] and statistical tests [134, 135].

For agent-based models in economics, Fagiolo et al. [130] proposed three different types of *calibration* methods: the *indirect calibration* approach, the *Werker-Brenner empirical calibration* approach, and the *history-friendly* approach. For example, Garcia et al. [94] adopt the last approach to an innovation diffusion study in New Zealand winery industry, using conjoint analysis to instantiate, calibrate, and verify the agent-based model qualitatively using stylized facts.

For agent-based models in marketing, Rand and Rust [114] suggest verification and validation as two key processes as guidelines for rigorous agent-based modeling. The use of term "verification" follows common understanding in system engineering [129]. In particular, the authors identify four steps for validation: micro-face validation, macro-face validation, empirical input validation, and empirical output validation using stylized facts, real-world data, and cross-validation. Note that the proposed validation steps echo the framework by Carley [95]: the first two steps correspond to grounding, the third to calibration, and the fourth roughly combines verification and harmonization. However, the cross-validation method mentioned in Rand and Rust [114] appears to suggest validation across models, whereas Carley [95] suggests validation across multiple data sets. The latter is consistent with the use of cross-validation in statistical inference and machine learning [137, 138].

Focusing specifically on empirically grounded ABMs, we suggest two pivotal steps in ensuring model reliability in a statistical sense: *calibration* and *validation*. By calibration, we mean the process of *quantitatively* fitting a set of model parameters to data, whereas validation means a quantitative assessment of the predictive efficacy of the model *using independent data*, that is, using data which was not utilized during the calibration

step. Moreover, insofar as a model of innovation diffusion is concerned with predicting future diffusion of an innovation, we propose to further split the dataset along a temporal dimension, so that earlier data is used exclusively for model calibration, while later data exclusively for validation. Starting with this methodological grounding, we now proceed to identify common issues that arise in prior research on empirically grounded agent-based models of innovation diffusion.

2.5.2 Issues in Model Calibration and Validation

Agent-based modeling research has often been criticized for lack of accepted methodological standard, hindering its acceptance in top journals by mainstream social scientists. One notable protocol due to Richiardi et al. [139] highlight four potential methodological pitfalls: *link with the literature*, *structure of the models*, *analysis*, and *replicability*.

A careful examination of the empirical ABM work on innovation diffusion through this protocol suggests that most of these issues have been addressed or significantly mitigated. For example, nearly all of the reviewed papers present theoretical background, related work, sufficient description of model structure, sensitivity analysis of parameter variability, a formal representation (e.g., UML[3], OOD[4]), and public access to source code. In spite of these improvements, however, there are residual concerns about systematic quantitative calibration and validation using empirical data.

We observe that different agent adoption models are calibrated differently. In the case of cognitive agent models (Section 2.2.3), such as the Theory of Planned Behavior and theory of emotional coherence, the individual model parameters are often estimated using survey data. Similarly, statistics-based models (Section 2.2.5) can be parametrized using either experimental or observational individual-level data. On the other hand, for conceptual

[3]The short for the Unified Modeling Language, developed by the Object Management Group: http://www.omg.org

[4]A standard to describe agent-based models originally proposed by Grimm et al. [140] for ecological modeling.

models, such as heuristic (Section 2.2.4) and economic models (Section 2.2.2), calibration is commonly done by iteratively adjusting parameters to match simulated diffusion trajectory to aggregate-level empirical data. Formally, we call the first kind of calibration "*micro-calibration*", as it uses individual data during calibration, whereas the second type "*macro-calibration*", as it uses aggregate-level data. Moreover, in many studies simulation parameters are determined using both micro- and macro-calibration. For example, since network structure is often not fully observed, and rules that govern agent interactions are assumed, parameters of these are commonly macro-calibrated. Our first concern is about macro-calibration.

Issue I: Potential pitfalls in macro-calibration. When a model has many parameters, over-fitting the model to data becomes a major concern [137, 138]. As Carley [95] suggests, "any model with sufficient parameters can always be adjusted so that some combination of parameters generates the observed data, therefore, large multi-parameter models often run the risk of having so many parameters that there is no guarantee that the model is doing anything more than curve fitting." Interestingly, the issue of over-fitting may even be a concern in macro-calibration when only a few parameters need to be calibrated. The reason is that agent-based models are highly non-linear, and even small changes in several parameters can give rise to substantially different model dynamics. This issue is further exacerbated by the fact that macro-calibration makes use of aggregate-level data, which is often insufficient in scale for reliable calibration of any but the simplest models, as many parameter variations can give rise to similar aggregate dynamics.

Addressing the issue requires greater care and rigor in applying macro-calibration. One possibility is that instead of choosing only a single parameter configuration, to select a parameter *zone* using a classifier such as decision trees [74] or other machine learning algorithms. Subsequently, the variability of parameters within this zone can be further investigated using sensitivity analysis. Another potential remedy is that instead of using only a single target statistic (e.g., average adoption rates) to use multiple indicators. A

relevant strategy to build agent-based models in the field of ecology is termed "pattern-oriented modeling", which utilizes multiple patterns at different scales and hierarchical levels observed from real systems to determine the model structure and parameters [81].

In addition, there are more advanced and robust techniques that can improve the rigor of macro-calibration. The modeling framework in [30] and statistical inference methods introduced in Section 2.4 propose methods which integrate micro and macro calibration into a single maximum likelihood estimation framework. Through well-established methods in machine learning, such as cross-validation, one can expect to parameterize a highly-predictive agent-based model and minimize the risk of over-fitting. Indeed, a fundamental feature of any approach should be to let validation ascertain the effectiveness of macro-calibration in generalizing beyond the calibration dataset. This brings us to the second common issue revealed by our review: lack of validation on independent data.

Issue II: Rigorous quantitative validation on independent data is uncommon. A common issue in the research we reviewed is that validation is often informal, incomplete, and even missing. The common reason for incomplete data-driven validation is that relevant data is simply unavailable. However, so long as data is available for calibrating the model, one can in principle use this data for both calibration and validation steps, for example, following cross-validation methods commonly utilized in machine learning. Several efforts seek to standardize the validation process for agent-based models, and computational models in general. However, few papers discussed explicitly follow any formalized validation approaches in this literature, although important exceptions exist [94, 97, 100].

Issue III: Few conduct validation at both micro-level and macro-level. There has been some debate about whether validation should be performed at both micro- and macro-level [95]. While arguments against the dual-verification often emphasize greater importance of model accuracy at the aggregate level, we argue that robust predictions at the aggregate level can only emerge when individual behavior is accurately modeled as well, particularly when policies that the ABM evaluates can be implemented as modifying indi-

vidual decisions.

Statistics-based models, such as machine learning, have well-established validation techniques which can be leveraged to validate individual-level models. One widely-used technique in machine learning and data mining is *cross-validation*. A common use of cross-validation is by partitioning the data into k parts, with training performed on $k-1$ of these and testing (evaluation) on the kth. The results are then averaged over k independent runs using each of the parts as test data. Observe that such a cross-validation approach can be used for models of individual behavior that are not themselves statistically-driven, such as models based on the theory of planned behaviors. Unfortunately, few of the surveyed papers, with the exception of statistics-based models, use cross-validation.

Issue IV: Few conduct validation of forecasting effectiveness on independent "future" data. One limitation of cross-validation techniques as traditionally used is that they provide an offline assessment of model effectiveness. To assess the predictive power of dynamical systems, the entire model has to be validated in terms of its ability to predict "future" data relative to what was used in calibration. We call this notion "*forward validation*". In particular, forward validation must assess simulated behaviors against empirical observations at both individual and aggregate levels *with an independent set of empirical data*. This can be attained, for example, by splitting a time-stamped data set so that calibration is performed on data prior to a split date, and forward validation is done on data after the split date [104, 30, 79, 113]. In this review, we do observe several approaches that are validated on independent data, but these either are not looking forward in time relative to the calibration data, or only focus on macro-level validation. Common argument for the use of in-sample data for forward validation is that new data is not available while the modeling task is undertaken. Notice, however, that any data set that spans a sufficiently long period of time can be split along the time dimension as above to effect rigorous forward validation.

2.5.3 Recommended Techniques for Model Calibration and Validation

We have identified several issues in calibration and validation which commonly arise in prior development of empirical agent-based models for innovation diffusion, and briefly discussed possible techniques that can help address these issues. We now summarize our recommendations:

Multi-Indicator Calibration. When macro-calibration is needed, the use of multiple indicators can help address *over-fitting*, whereby a model which appears to effectively *match* data in calibration performs poorly in prediction on unseen data. We suggest that such indicators are developed at different scale and hierarchical levels, so that models which cannot effectively generalize to unseen data can be efficiently eliminated.

Maximum Likelihood Estimation. When individual-level data are available, we recommend constructing probabilistic adoption models for agents, and estimating parameters of these models by maximizing a global likelihood function (see, for example, the modeling framework by Zhang et al. [30], and research discussed in Section 2.4.2). Doing so offers a principled means of calibrating agent behavior models from empirical data.

Cross Validation. This approach is widely used for model selection in the machine learning literature. Here, we recommend it for both micro-calibration and micro-validation of ABMs. Note that it does not only apply to statistics-based models, but can be used for any agent modeling paradigm where model parameters are calibrated using empirical data. The use of cross-validation in calibration can dramatically reduce the risk of over-fitting. Moreover, as it inherently uses independent data, such validation leads to more rigorous ABM methodology.

Forward Validation. This method involves splitting data into two consecutive time

periods. The modeler calibrates an agent-based model using data from the first period, and assesses the predictive efficacy of the model in the second period. More rigorously, validation of the model should be evaluated at both individual and aggregate levels.

2.6 Conclusions

We provided a systematic, comprehensive, and critical review of existing work on empirically grounded agent-based models for innovation diffusion. We offered a unique methodological survey of literature by categorizing agent adoption models along two dimensions: methodology and application. We identified six methodological categories: *mathematical optimization based models*, *economic models*, *cognitive agent models*, *heuristic models*, *statistics-based models* and *social influence models*. They differ not only in terms of assumptions and elaborations of human decision-making process, but also with respect to calibration and parameterization techniques. Our critical assessment of each work focused on using data for calibration and validation, and particularly performing validation with independent data. We briefly reviewed the most important work in the closely related literature on information diffusion, building connections between the innovation and information diffusion approaches. One particularly significant observation is that information diffusion methods rely heavily on machine learning and maximum likelihood estimation approaches, and the specific methodology used can be naturally ported to innovation diffusion ABMs. Drawing on prior work in validation of computational models, we discussed four main issues for existing empirically grounded ABM studies in innovation diffusion, and provided corresponding solutions.

On balance, recent developments of empirical approaches in agent-based modeling for innovation diffusion are encouraging. Although calibration and validation issues remain in many studies, a number of natural solutions from data analytics offer promising directions in this regard. The ultimate goal of empirically grounded ABMs is to provide decision

support for policy makers and stakeholders across a broad variety of innovations, helping improve targeted marketing strategies, and reduce costs of successful translation of high-impact innovative technologies to the marketplace.

Chapter 3

Data-Driven Agent-Based Modeling, with Application to Rooftop Solar Adoption

Our critical review in Chapter 2 summarized several categories of modeling methods in building empirically-grounded ABMs and suggested issues on model calibration and validation as well as their potential solutions. In dealing with large-scale diffusion data, the machine learning models seems quite efficient and more rigorous in terms of calibration and validation, which makes them ideal for building high-fidelity agent-based simulations to support decision making. This chapter takes a deep look into such a representative work that leverages machine-learning techniques to calibrate and validate agent-based models based on massive amount of individual adoption data and successfully applies it to forecast the adoption of renewable solar technology in San Diego county, US.

3.1 Introduction

The rooftop solar market in the US, and especially in California, has experienced explosive growth in last decade. At least in part this growth can be attributed to the government incentive programs which effectively reduce the system costs. One of the most aggressive incentive programs is the California Solar Initiative (CSI), a rooftop solar subsidy program initiated in 2007 with the goal of creating 1940 megawatts of solar capacity by 2016 [141]. The CSI program has been touted as a great success, and it certainly seems so: over 2000 megawatts have been installed to date. However, in a rigorous sense, success would have to be measured in comparison to some baseline; for example, in comparison to the same world, but without incentives. Of course, such an experiment is impossible in practice. However, in principle, insight can be drawn by sensitivity analysis based on hypothetical solar diffusion model. What is the most appropriate modeling methodology to build a highly robust solar diffusion model?

ABM has long been a common framework of choice for studying aggregate, or emergent, properties of complex systems as they arise from microbehaviors of a multitude of agents in social and economic contexts [142, 143, 114]. ABM appears well-suited to policy experimentation of just the kind needed for the rooftop solar market. Indeed, there have been several attempts to develop agent-based models of solar adoption trends [144, 67, 145]. Both traditional ABM, as well as the specific models developed for solar adoption, use data to calibrate aspects of the models (for example, features of the social network, such as density, are made to match real networks), but not the entire model. More importantly, validation is often qualitative, or, if quantitative, using the same data as used for calibration. The weakness of validation makes those models less eligible as a reliable policy experiment tool.

The emergence of "Big Data" offers new opportunities to develop agent-based models in a way that is entirely data-driven, both in terms of model calibration and validation. In the particular case of rooftop solar adoption, the CSI program, in addition to subsidies, also provides for a collection of a significant amount of data by the program administrators, such as Center for Sustainable Energy (CSE) in San Diego county, about specific (individual-level) characteristics of adopters. While by itself insufficient, we combine this data with property assessment characteristics for all San Diego county residents to yield a high-fidelity data set that we use to calibrate artificial agent models using machine learning techniques. However, the increasing availability of data from various sources in all levels, i.e., micro and macro levels, also poses significant computational challenge to any researcher who aims to study the phenomenon of solar diffusion. Machine learning and data mining provide us with efficient and scalable algorithms, well-principled techniques, such as cross validation, feature selection etc. A data-driven ABM is then constructed using exclusively such learned agent models, with no additional hand-tuned variables. Moreover, following standard practice in machine learning, we separate the calibration data from the data used for validation.

This chapter makes the following contributions:

1. a framework for data-driven agent-based modeling;

2. methods for learning individual-agent models of solar adoption, addressing challenges posed by the market structure and the nature of the data;

3. an adaptation of a recent agent-based model of rooftop solar adoption, used as a baseline, with an improved means for systematic calibration (systemitizing the approach proposed by Palmer et al. [67] (entirely new addition compared to our preliminary work [19]);

4. a data-driven agent-based model of solar adoption in (a portion of) San Diego county, with forecasting efficacy evaluated on data not used for model learning;

5. a comparison of the data-driven approach to the baseline adoption model (a new addition compared to our preliminary work [19]);

6. a quantitative evaluation of the California Solar Initiative subsidy program (including a significantly improved and extended approach to optimizing the solar discount policy relative to our preliminary work [19]), a broad class of incentive policies, and a broad class of solar system "seeding" policies.

3.2 Related Work

ABM methodology has a substantial, active, literature [142, 143, 114], ranging from methodological to applied. Somewhat simplistically, the approach is characterized by the development of models of agent behavior, which are integrated within a simulation environment. The common approach is to make use of relatively simple agent models (for example, based on qualitative knowledge of the domain, qualitative understanding of human behavior, etc.), so that complexity arises primarily from agent interactions among themselves and with the environment. For example, Thiele et al. [146] document that only 14% of articles

published in the Journal of Artificial Societies and Social Simulation include parameter fitting. Our key methodological contribution is a departure from developing simple agent models based on relevant *qualitative* insights to *learning* such models entirely on data. Due to its reliance on data about *individual agent behavior*, our approach is not universally applicable. However, we contend that such data is becoming increasingly prevalent, as individual behavior is now continuously captured in the plethora of virtual environments, as well as through the use of mobile devices. As such, we are not concerned about simplicity of agent models *per se*; rather, it is "bounded" by the amount of data available (the more data we have, the more complex models we can reliably calibrate on it).

Thiele et al. [146], as well as Dancik et al. [147] propose methods for calibrating model parameters to data. However, unlike our work, neither offers methodology for *validation*, and both operate at model-level, requiring either extremely costly simulations (making calibration of many parameters intractable), or, in the case of Dancik et al., a multi-variate Normal distribution as a proxy, losing any guarantees about the quality of the original model in the process. Our proposal of calibration at the *agent level*, in contrast, enables us to leverage state-of-the-art machine learning techniques, as well as obtain more reliable, and interpretable, models at the individual agent level. Recently, in field of ecology and sociology, there is rising interest to combine agent-based model with empirical methods [148]. Biophysical measurements, i.e., soil properties and socioeconomic surveys are used by Berger and Schreinemachers [149] to generate a landscape and agent populations which are consistent with empirical observation by Monte Carlo techniques. Notice that this is quite different application from ours, since we do not need to generate an agent population; rather we instantiate our multi-agent simulation with learned agents. Huigen et al. [150] visually calibrate a special agent-based model using ethnographic histories of farm households to understand linkage between land-use system dynamics and demographic dynamics. Happe et al. [151] instantiate an agent-based agricultural policy simulator with empirical data and investigate the impact of a regime switch in agricultural policy on struc-

tural change under various framework conditions. However, the model is not statistically validated. By populating ABM with a population of residential preferences drawn from the survey results, Brown and Robinson [152] explore the effects of heterogeneity in residential preferences on an agent-based model of urban sprawl, performing sensitivity analysis as a means of validation. In settings of public-goods games, Janssen and Ahn [153] compare the empirical performance of a variety of learning models with parameters estimated by maximum likelihood estimation and theories of social preferences. However, no systematic and rigorous validation is applied.

A number of ABM efforts are specifically targeted at the rooftop solar adoption domain [154, 144, 67, 145, 79, 155, 92]. Denholm et al. [144] and Boghesi et al. [154] follow a relatively traditional methodological approach (i.e., simple rule-based behavior model), and their focus is largely on financial considerations in rooftop solar adoption. Palmer et al. [67] and Zhao et al. [92], likewise use a traditional approach, but consider several potentially influential behavioral factors, such as social influence and household income. Palmer et al. calibrate their model using total adoption data in Italy (unlike our approach, they do not separate calibration from validation). Zhao et al. set model parameters based on a combination of census and survey data, but without performing higher-level model calibration with actual adoption trends. None of these past approaches makes use of machine learning to develop agent models (indeed, none except Palmer et al. calibrate the model using actual adoption data, and even they do not seem to do so in a systematic way, using instead "trial and error"). Much of this previous work on agent-based models of rooftop solar adoption attempts to use the models to investigate alternative policies. Unlike us, however, none (to our knowledge) consider the *dynamic* optimization problem faced by policy makers (i.e., how much of the budget to spend at each time period), nor compare alternative incentive schemes with "seeding" policies (i.e., giving systems away, subject to a budget constraint).

There have also been a number of models of innovation diffusion in general, as well as rooftop solar adoption in particular, that are not agent-based in nature, but instead as-

pire only to anticipate aggregate-level trends. Bass [39] introduce the classic "S-curve" quantitative model, building on the qualitative insights offered by Rogers [1] and others. In the context of rooftop solar, noteworthy efforts include Lobel and Perakis [156], Bollinger and Gillingham [157], and van Benthem et al. [158]. Lobel and Perakis calibrate a simple model of aggregate solar adoption in Germany on total adoption data; their model, like ours, includes both economics (based on the feed-in tariff as well as learning-by-doing effects on solar system costs) and peer effects. We therefore use their model, adapted to *individual* agent behavior, as our "baseline". Bollinger and Gillingham demonstrate causal influence of peer effects on adoption decisions, and van Benthem et al. focus on identifying and quantifying learning-by-doing effects.

Several related efforts are somewhat closer in spirit to our work. Kearns and Wortman [159] developed a theoretical model of learning from collective behavior, making the connection between learning individual agent models and models of aggregate behavior. However, this effort does not address the general problem of learning from a single observed sequence of collective behavior which is of key interest to us. Judd et al. [160] use machine learning to predict behavior of participants in social network coordination experiments, but are only able to match the behavior qualitatively. Duong et al. [161] propose history-dependent graphical multiagent models to compactly represent agent joint behavior based on empirical data from experimental cooperation games. However, scalability of this approach is quite limited. Another effort in a similar vein uses machine learning to calibrate walking models from real and synthetic data, which are then aggregated in an agent-based simulation [162]. Aside from the fundamental differences in application domains from our setting, Torrens et al. [162] largely eschew model validation, and do not consider the subsequent problem of policy evaluation and optimization, both among our key contributions. Most recently, Wunder et al. [163] fit a series of deterministic and stochastic models to data collected from on-line experimental public goods games. Like our approach, they make use of machine learning to learn agent behavior, and validate the model using out-of-sample

prediction. However, this work does not validate the model ability to forecast individual and aggregate-level behavior, since training and validation data sets are chosen randomly, rather than split across the time dimension (so that in many cases future behavior is used to learn and model is validated on "past" behavior). Moreover, the models are very simple and specific to the public goods game scenario, taking advantage of the tightly controlled source of data.

Finally, there has been substantial literature that considers the problem of marketing on social networks [16, 164]. Almost universally, however, the associated approaches rely on the structure of specific, very simple, influence models, without specific context or attempting to learn the individual behavior from data (indeed, we find that simple baseline models are not sufficiently reliable to be a basis for policy optimization in our setting). Moreover, most such approaches are static (do not consider the dynamic marketing problem, as we do), although an important exception is the work by Golovin and Krause [165], in which a simple greedy adaptive algorithm is proven to be competitive with the optimal sequential decision for a stochastic optimization problem that satisfies adaptive submodularity.

3.3 Data-Driven Agent-Based Modeling

The overwhelming majority of agent-based modeling efforts in general, as well as in the context of innovation/solar adoption modeling in particular, involve: a) *manual* development of an agent model, which is usually rule-based (follows simple behavior rules), b) ad hoc tuning of a large number of parameters, pertaining to both the agent behaviors, as well as the overall model (environment characteristics, agent interactions, etc), and c) validation usually takes the form of qualitative expert assessment, or is in terms of overall fit of aggregate behavior (e.g., total number of rooftop solar adoptions) to ground truth, *using the data on which the model was calibrated* [142, 143, 114, 154, 144, 67, 145, 92]. We break with this tradition, offering instead a framework for *data-driven agent-based modeling (DDABM)*, where agent models are learned from data about individual (typi-

cally, human) behavior, and the agent-based model is thereby fully data-driven, with *no additional parameters to govern its behavior*. We now present our general framework for *data-driven agent-based modeling (DDABM)*, which we subsequently apply to the problem of modeling residential rooftop solar diffusion in San Diego county, California. The key features of this framework are: a) explicit division of data into "calibration" and "validation" to ensure sound and reliable model validation and b) automated agent model training and cross-validation.

In this framework, we make three assumptions. The first is that time is *discrete*. While this assumption is not of fundamental importance, it will help in presenting the concepts, and is the assumption made in our application. The second assumption is that agents are *homogeneous*. This may seem a strong assumption, but in fact it is without loss of generality. To see this, suppose that $h(x)$ is our model of agent behavior, where x is *state*, or all information that conditions the agent's decision. Heterogeneity can be embedded in h by considering individual characteristics in state x, such as personality traits and socio-economic status, or, as in our application domain, housing characteristics. Indeed, arbitrary heterogeneity can be added by having a unique identifier for each agent be a part of state, so that the behavior of each agent is unique. Our third assumption is that each individual makes *independent* decisions at each time t, conditional on state x. Again, if x includes all features relevant to an agent's decision, this assumption is relatively innocuous.

Given these assumptions, DDABM proceeds as follows. We start with a data set of individual agent behavior over time, $D = \{(x_{it}, y_{it})\}_{i,t=0,\ldots,T}$, where i indexes agents, t time through some horizon T and y_{it} indicates agent i's decision, i.e., 1 for "adopted" and 0 for "did not adopt" at time t.

1. Split the data D into *calibration* D_c and *validation* D_v parts along the time dimension: $D_c = \{(x_{it}, y_{it})\}_{i,t \leq T_c}$ and $D_v = \{(x_{it}, y_{it})\}_{i,t > T_c}$ where T_c is a time threshold.

2. Learn a model of agent behavior h on D_c. Use cross-validation on D_c for model (e.g., feature) selection.

3. Instantiate agents in the ABM using h learned in step 2.

4. Initialize the ABM to state x_{jT_c} for all artificial agents j.

5. Validate the ABM by running it from x_{T_c} using D_v.

One may wonder how to choose the initial state x_{jT_c} for the artificial agents. This is direct if the artificial agents in the ABM correspond to actual agents in the data. For example, in rooftop solar adoption we know which agents have adopted solar at time T_c, and their actual housing characteristics, etc. Alternatively, one can run the ABM from the initial state, and start validation upon reaching time $T_c + 1$.

Armed with the underlying framework for DDABM, we now proceed to apply it in the context of spatial-temporal solar adoption dynamics in San Diego county.

3.4 DDABM for Solar Adoption

3.4.1 Data

In order to construct the DDABM for rooftop solar adoption, we made use of three data sets provided by the Center for Sustainable Energy: individual-level adoption characteristics of residential solar projects installed in San Diego county as a part of the California Solar Initiative (CSI), property assessment data for the entire San Diego county, and electricity utilization data for most of the San Diego county CSI participants spanning twelve months prior to solar system installation. Our CSI data, covering projects completed between May 2007 and April 2013 (about 6 years and 8,500 adopters), contains detailed information about the rooftop solar projects, including system size, reported cost, incentive (subsidy) amount, whether the system was purchased or leased, the date of incentive reservation, and the date of actual system installation, among others. The assessment data includes comprehensive housing characteristics of San Diego county residents (about 440,000 households), including square footage, acreage, number of bedrooms and bathrooms, and whether or

not the property has a pool. The CSI and assessment data were merged so that we could associate all property characteristics with adoption decisions.

3.4.2 Modeling Individual Agent Behavior

Our DDABM framework presupposes a discrete-time data set of individual adoption decisions. At face value, this is not what we have: rather, our data only appears to identify static characteristics of individuals, and their adoption timing. This is, of course, not the full story. Much previous literature on innovation diffusion in general [39, 166, 167, 1], and solar adoption in particular [157, 156, 168, 169], identifies two important factors that influence an individual's decision to adopt: economic benefits and peer effects. We quantify economic benefits using *net present value (NPV)*, or discounted net of benefits less costs of adoption: $NPV = \sum_t \delta^t (b_t - c_t)$, where b_t correspond to benefits (net savings) in month t, and c_t are costs incurred in month t; we used a $\delta = 0.95$ discount factor. Peer, or social, effects in adoption decisions arise from social influence, which can take many forms. Most pertinent in the solar market is *geographic* influence, or the number/density of adopters that are geographically close to an individual making a decision. Both economic benefits and peer effects are dynamic: the former changes as system costs change over time, while the latter changes directly in response to adoption decision by others. In addition, peer effects create interdependencies among agent decisions, so that aggregate adoption trends are not simply averages of individual decisions, but evolve through a highly non-linear process. Consequently, even if we succeed in learning individual agent models, this by no means guarantees success when they are jointly instantiated in simulation, especially in the context of a forecasting task. Next, we describe in detail how we quantify economic and peer effect variables in our model.

3.4.2.1 Quantifying Peer Effects

We start with the simpler issue of quantifying peer effects. The main challenge is that there are many ways to measure these: for example, total number of adopters in a zip code (a measure used previously [157]), fraction of adopters in the entire area of interest (used by [156]), which is San Diego county in our case, as well as the number/density of adopters within a given radius of the individual making a decision. Because we ultimately utilize feature selection methods such as regularization, our models consider a rather large collection of these features, including both the number and density of adoptions in San Diego county, the decision maker's zip code, as well as within a given radius of the decision maker for several radii. Because we are ultimately interested in policy evaluation, we need to make sure that policy-relevant features can be viewed as causal. While we can never fully guarantee this, our approach for computing peer effect variables follows the methodology of Bollinger and Gillingam [157], who tease out causality from the fact that there is significant spatial separation between the adoption decision, which is indicated by the incentive reservation action, and installation, which is used in measuring peer effects.

3.4.2.2 Quantifying Net Present Value

To compute NPV in our DDABM framework we need to know costs and benefits *that would have been perceived* by an individual i adopting a system at time t. Of course, our data does not actually offer such counterfactuals, but only provides information for adopters *at the time of adoption*. The structure of solar adoption markets introduces another complication: there are two principal means of adoption, buying and leasing. In the former, the customer pays the costs up-front (we ignore any financing issues), while in the latter, the household pays an up-front cost *and a monthly cost* to the installer. Moreover, CSI program incentives are only offered to system buyers, who, in the case of leased systems, are the installers. Consequently, incentives directly offset the cost to those buying the system outright, but at best do so indirectly for leased systems. In the case of leased systems, there

is also an additional data challenge: the system costs reported in the CSI data do not reflect actual leasing expenses, but the estimated market value, and are therefore largely useless for our purposes. Finally, both costs and benefits depend on the capacity (in watts) of the installed system, and this information is only available for individuals who have previously adopted.

Our first step is to estimate system capacity using property assessment features. We do so using step-wise linear regression [170], arriving at a relatively compact model, shown in Table 3.1. The adjusted R^2 of this model is about 0.27, which is acceptable for our purposes.

Table 3.1: Linear model of solar system capacity (size). All coefficients are significant at the $p = 0.05$ level.

Predictor	Estimate
(Intercept)	1.59
Owner Occupied (binary)	-0.25
Has a Pool (binary)	0.63
Livable Square Footage	7.58e-04
Acreage	1.32
Average Electricity Utilization in Zipcode	8.25e-04

Next, we use the system size variable to estimate system costs separately for the purchased and leased systems. For the purchased systems, the cost at the time of purchase is available and reasonably reliable in the CSI data, but only during the actual purchase time. However, costs of solar systems decrease significantly over time. A principal theory for this phenomenon is *learning-by-doing* [171, 172, 173, 156, 158], in which costs are a decreasing function of aggregate technology adoption (representing, essentially, economies of scale). In line with the learning-by-doing theory, we model the cost of a purchased system as a function of property assessment characteristics, predicted system size, and peer effect features, including total adoption in San Diego county. We considered a number of models for ownership cost and ultimately found that the linear model is most effective. In all cases, we used l_1 regularization for feature selection [137]. The resulting model is

shown in Table 3.2.

Table 3.2: Ownership cost linear model.

Predictor	Coefficient
(Intercept)	1.14e+04
Property Value	7.38e-04
Livable Square Footage	0.15
System Capacity	6.21e+03
Total Adoption in SD County	-1.06

In order to estimate total discounted lease costs, we extracted cost details from 227 lease contracts, and used this data to estimate the total discounted leasing costs $C^l = \sum_t \delta^t c_t$ through the duration of the lease contract in a manner similar to our estimation of ownership costs. One interesting finding in our estimation of lease costs is that they appear to be largely insensitive to the economic subsidies; more specifically, system capacity turned out to be the only feature with a non-zero coefficient (the coefficient value was 1658, with the intercept value of 10447). In particular, this implies that solar installers do not pass down their savings to customers of leased systems.

Having tackled estimation of costs, we now turn to the other side of NPV calculation: benefits. In the context of solar panel installation, economic benefits are monthly savings, which are the total electricity costs offset by solar system production. These depend on two factors: the size of the system, which we estimate as described above, and the electricity rate. The latter seems simple in principle, but the rate structure used by SDG&E (San Diego Gas and Electric company) makes this a challenge. The SDG&E rates have over the relevant time period a four-tier structure, with each tier depending on monthly electricity utilization relative to a baseline. Tiers 1 and 2 have similar low rates, while tiers 3 and 4 have significantly higher rates. Tier rates are marginal: for example, tier-3 rates are only paid for electricity use above the tier-3 threshold. The upshot is that we need to know electricity utilization of an individual in order to estimate marginal electricity costs offset by the installed solar system. For this purpose, we use the electricity utilization data for

the adopters. Here, we run into a technical problem: after running a regression model, we found that average predicted electricity rates for San Diego zip codes significantly exceed observed zip code averages—in other words, our data is biased. To reduce this bias, we modified the linear model as follows. Let (X, y) represent the feature matrix and corresponding vector of energy utilizations for a given month for adopters, and let $(\bar{X}, \bar{y})$ be the matrix of average feature values and average energy use for all San Diego county zip codes. A typical linear model chooses a weight vector w to minimize $(Xw-y)^T(Xw-y)$. In our model, we extend this to solve

$$\min_w \quad (Xw-y)^T(Xw-y) + \lambda(\bar{X}w-\bar{y})^T(\bar{X}w-\bar{y}),$$

which is equivalent to a linear regression with the augmented data set (Z, z), where

$$Z = \begin{pmatrix} X \\ \sqrt{\lambda}\bar{X} \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} y \\ \sqrt{\lambda}\bar{y} \end{pmatrix}.$$

When λ is small, our model is better able to capture fidelity of individual-level data, but exhibits greater bias. We used deviance ratio to choose a value of λ in the context of the overall individual-agent model.

Now that we can predict both system size and electricity utilization, we can correspondingly predict, for an arbitrary individual, their monthly savings from having installed rooftop solar. Along with the predicted costs, this gives us a complete evaluation of NPV for each potential adopter.

3.4.2.3 Learning the Individual-Agent Model

In putting everything together to learn an individual-agent model, we recognize that there is an important difference between the decision to buy and the decision to lease, as described above. In particular, we have to compute net present value differently in the

two models. Consequently, we actually learn two models: one to predict the decision to lease, and another for the decision to buy, each using its respective NPV feature, along with all of the other features, including peer effects and property assessment, which are shared between the models. For each decision model, we used l_1-regularized logistic regression. Taking x_l and x_o to be the feature vectors and $p_l(x_l)$ and $p_o(x_o)$ the corresponding logistic regression models of the lease and own decision respectively, we then compute the probability of adoption

$$p(x) = p_l(x_l) + p_o(x_o) - p_l(x_l)p_o(x_o),$$

where x includes the NPV values for lease and own decisions.

To train the two logistic regression models, we can construct the data set (x_{it}, y_{it}), where i correspond to the households in San Diego county and t to months, with x_{it} the feature vector of the relevant model and y_{it} the lease (own) decision, encoded as a 1 if the system is leased (owned) and 0 otherwise. To separate calibration and validation we used only the data through 04/2011 for calibration, and the rest (through 04/2013) for ABM validation below. The training set was comprised of nearly 7,000,000 data points, of which we randomly chose 30% for calibration (due to scalability issues of standard logistic regression implementation in R).[1] All model selection was performed using 10-fold cross-validation. Since leases only became available in 2008, we introduced a dummy variable that was 1 if the lease option was available at the time and 0 otherwise. We also introduced seasonal dummy variables (Winter, Spring, Summer) to account for seasonal variations in the adoption patterns. The final model for the propensity to purchase a solar system is shown in Table 3.3, and a model for leasing is shown in Table 3.4.

[1] In fact, we have sampled the process multiple times, and can confirm that there is little variance in the model or final results.

Table 3.3: Ownership Logistic Regression Model

Predictor	Coefficient
(Intercept)	-10.19
Owner Occupied (binary)	0.94
# Installations Within 2 Mile Radius	-3.05e-04
# Installations Within 1 Mile Radius	2.60e-03
# Installations Within $\frac{1}{4}$ Mile Radius	6.78e-03
Lease Option Available (binary)	0.69
Winter (binary)	-0.59
Spring (binary)	-0.19
Summer (binary)	-0.28
Installation Density in Zipcode	100.11
NPV (Purchase)	7.58e-06

Table 3.4: Lease Logistic Regression Model

Predictor	Coefficient
(Intercept)	-13.22
Owner Occupied (binary)	0.73
# Installations Within 2 Mile Radius	2.21e-03
# Installations Within $\frac{1}{4}$ Mile Radius	7.87e-03
Lease Option Available (binary)	1.65
Winter (binary)	-0.39
Spring (binary)	0.29
Summer (binary)	-0.20
Installation Density in Zipcode	85.69
NPV (Lease)	7.07e-06

3.4.3 Agent-Based Model

The models developed above were implemented in the Repast ABM simulation toolkit [174].

3.4.3.1 Agents

The primary agent type in the model represents residential households (implemented as a Java class in Repast [174]). In the ABM we do not make the distinction between leasing and buying solar systems, so that each agent acts according the the stochastic model $p(x_{it})$ derived as described in the previous section, where x_{it} is the system state relevant to agent

i's at time (iteration) t. In addition, in order to flexibly control the execution of simulation, we defined a special *updater* agent type which is responsible for updating state attributes of household agents x_{it} at each time step t.

3.4.3.2 Time Step

Time steps of the simulation correspond to months. The execution diagram for a single simulation run is presented in Figure 3.1. Initially, the simulation is populated by residential household agents that are characterized by GIS locations, home properties, and adoption states. At each tick of the simulation, updater agent first updates features x_{it} for all agents, such as purchase and lease costs, incentive (which may depend on time), NPVs, and peer effects, for all agents based on the state of world (e.g., the set of agents having adopted thus far in the simulation). Lease and ownership cost are computed using the lease and ownership cost models as described above, while the incentives may follow an arbitrary subsidy scheme, and in particular can mirror the CSI rate schedule. Next, each non-adopter household is asked to make a decision. When a household agent i is called upon to make the adoption decision at time t, this agent adopts with probability $p(x_{it})$ [2]. If an agent chooses to adopt, this agent switches from being a non-adopter to becoming an adopter in the simulation environment. Moreover, when we thereby create a new adopter, we also assign an installation period of the solar system. Specifically, just as in reality, adoption decision only involves the reservation of the incentive, while actual installation of the system takes place several months later. Since peer effect variables are only affected by completed installations, it is important to capture this lag time. We capture the delay between adoption and installation using a random variable distributed uniformly in the interval $[1,6]$, which is the typical lag time range in the training data. The simulation terminates in a user-specified number of steps.

[2] An agent decides to adopt solar panels if a system-generated random number is less than the adoption probability $p(x_{it})$.

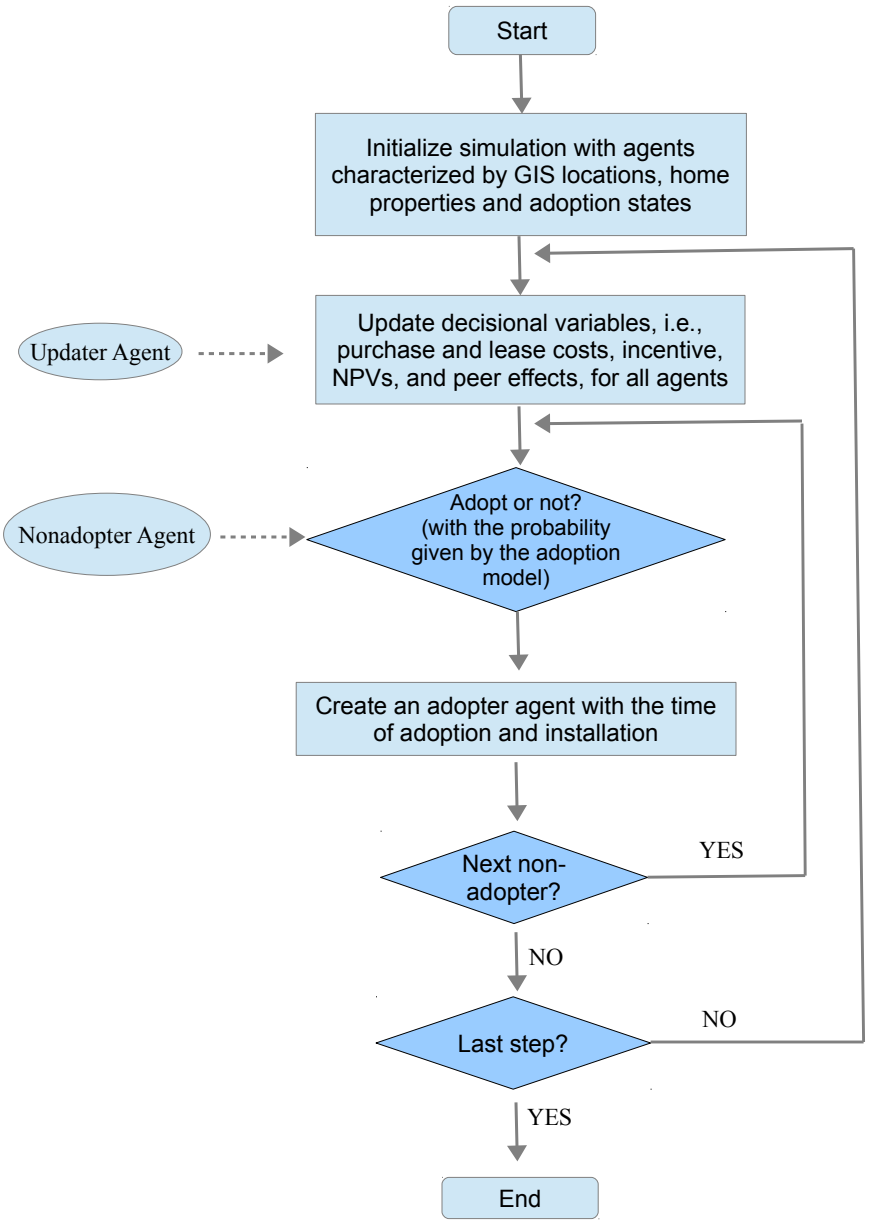


Figure 3.1: Execution diagram for a single simulation run.

3.4.3.3 Computing Peer Effect Variables

In order to compute geography-based peer effects, we need information about geographic location of the households. To this end we use a Repast GIS package. A naive way to compute peer effect variables would update these for each non-adopter agent in each iteration. However, this approach is very inefficient and scales poorly, as there are vastly more non-adopters than adopters in typical simulations. Therefore, we instead let adopter agents update peer effect variables for their neighbors at the time of system installation, dramatically reducing the corresponding overhead.

3.5 A State-of-the-Art Alternative Solar Adoption Model

Our model differs from most agent-based modeling approaches in the context of rooftop solar adoption on the following three principal dimensions: first, all features used for modeling agent behavior are carefully derived from available data, second, calibration is performed using the individual agent behavior, and third, the model is validated using data that is the "future" relative to the data used for model calibration.

In order to offer a principled baseline comparison of our model to "state-of-the-art", we implement a recent agent-based model that was also proposed in the context of rooftop solar adoption [67]. Our choice of the model was driven by the following considerations: a) the model was sufficiently well described for us to be able to independently replicate it, b) the model included an explicit section about parameter calibration, and c) it was possible for us to instantiate this baseline model, albeit somewhat imperfectly, using data available to us. Still, we faced several limitations, the most important of which being the difference between the targeted population (Palmer et al. model targeted Italy, whereas our model and data is for California) and available data (Palmer et al. utilized data not available to us, such as household income, as well as proprietary categorization of individuals into adoption classes).

In this section, we describe in detail our adaptation of the model by Palmer et al. [67], staying as close as possible to the original model. In addition, we describe a means of model calibration which was more systematic than the approach (trial-and-error) used by Palmer et al., but also uses as a calibration target aggregate adoption levels over time.

3.5.1 Consumer Utility Model

Strongly influenced by classical consumer theory, the agent in the Palmer et al. model makes adoption decision based on utility, i.e., to what extent the investment of solar would satisfy one's needs. The utility for an agent to install solar PV system i is defined as a weighted sum of four factors, or partial utilities:

$$U^i = w_{eco}u^i_{eco} + w_{env}u^i_{env} + w_{inc}u^i_{inc} + w_{com}u^i_{com} \tag{3.1}$$

where

$$\sum_f w_f = 1 \text{ for } f \in F : \{eco, env, inc, com\} \text{ and } w_f \in [0,1]$$

The four partial utilities are the economic benefit of the solar investment (u_{eco}), the environmental benefit of installing in a PV system (u_{env}), the utilities of household income (u_{inc}) and the influence of communication with other agents (u_{com}). Simply, agent decides to invest a PV system when one's utility surpasses a certain threshold. Notice also that the four weights in the model are identical for all agents, which along with the decision threshold are calibrated by matching the fitted aggregate adoption to the ground truth. [3]

[3] In the model developed by Palmer et al. [67], the weighs differ by agent's socio-economic group, derived using proprietary means. Since this categorization is not available in our case, and also to reduce the number of parameters necessary to calibrate (and, consequently, to reduce the amount of over-fitting), we use identical weights for all agents.

3.5.1.1 Economic Utility

Economic utility captures economic benefit/cost associated with solar installation. We use net present value of buying solar PV system to calculate the economic utility, which we normalize to have zero mean and unit variance:

$$u_{eco} = \frac{NPV^i_{buy} - \overline{NPV_{buy}}}{S(NPV_{buy})} \tag{3.2}$$

where $\overline{NPV_{buy}}$ and $S(NPV_{buy})$ are the sample mean and standard deviation of net present value of all potential adopters respectively.

3.5.1.2 Environmental Utility

The environmental utility ideally measures amount of CO_2 solar installation could save. Due to difficulty of obtaining this information, following Palmer et al.[67], we instead use expected solar electricity production to compute environmental utility.

$$u_{env} = \frac{E^i_{PV} - \overline{E_{PV}}}{S(E_{PV})} \tag{3.3}$$

where $E^i_{PV} = R^i_{CSI} * HR_{sun} * 30(days) * 12(months) * 20(years)$, or the total electricity production in 20 years. $\overline{E_{PV}}$ and $S(E_{PV})$ are sample mean and standard deviation of solar electricity generation for all potential adopters.

3.5.1.3 Income Utility

Income utility in agent model of Palmer et al. [67] is originally calculated by household income. Unfortunately, household income is not available in our current study, and we instead use home value that can be treated as a relatively reliable estimate of a household's income. Unfortunately, the home value in our original dataset are prices last time the home was sold, which can be significantly out of date. To compute home value more

accurately, we extract historical median home sale prices (merged both sold and list price in $dollar/ft^2$) of San Diego County from Zillow's on-line real estate database. Finally, the home value is recovered by multiplying the per-unit price with livable square feet. Similar to other utilities, the income utility of each agent is just the normalized home value, that is

$$u_{inc} = \frac{V^i_{home} - \overline{V_{home}}}{S(V_{home})} \tag{3.4}$$

where, $\overline{V_{home}}$ and $S(V_{home})$ denote sample mean and standard deviation of home value of all potential solar adopters.

3.5.1.4 Communication Utility

In Palmer et al. [67] work, the communication utility is calculated based on social economic status of each agent. Because the relevant information is unavailable, we turn to a simple variation, preserving the essence of their approach. Since, density of installation within 1-mile radius of a household is the most significant among all geology-based peer effect measures, we use it to derive the communication utility. In other sense, this is equivalent to assume that all agents within 1-mile radius of a household are in the same socio-economic group, which is a reasonable assumption since individuals with similar socio-economic status often live nearby. The communication utility is thus computed as follows.

$$u_{com} = \frac{F^i_{1-mile} - \overline{F_{1-mile}}}{S(F_{1-mile})} \tag{3.5}$$

where, $\overline{F_{1-mile}}$ and $S(F_{1-mile})$ denote sample mean and standard deviation of solar installation density within 1-mile radius for all potential adopters.

3.5.2 Calibration

Palmer et al. calibrated the parameters of their model using trial-and-error to explore the parameter space, and making use largely of a visual qualitative match between predicted

and observed adoption levels. We make use, instead, a more systematic calibration method, formulating as the problem of minimizing mean-squared error between predicted and actual adoption:

$$\theta^* = \arg\min_{\theta} \frac{1}{T} \sum_{t=1}^{T} (\hat{Y}^t - Y^t)^2 \tag{3.6}$$

where $\theta = (w_{eco}, w_{env}, w_{inc}, w_{com}, theshold)$, $\hat{Y}^t$ and Y^t are fitted and actual aggregate adoption at time t, which we take to be at monthly granularity.

To search for the optimal parameter, we implemented our adaptation of the Palmer et al. agent-based model in R. Specifically, at each tick, we compute utility of each agent and an agent will choose to install solar PV as long as its utility gets above the threshold. Because calibration of the entire dataset is computationally infeasible, we instead calibrate the model based on a random sample of 10% (about 44,000) of the households. Rather finding an ideal parameter by "trial and error", we here propose a more systematic way to search the parameter space. It is done through multiple iterations. In first iteration, it scans every possible parameters based on a relatively coarse discretization of parameter space and finds the optimal parameter with the minimum MSE. In the next iteration, it probes only a subspace of previous iteration around the best solution found so far, meanwhile, a more fine-grained discretization is applied. For example, Figure 3.2, one can see the most promising range of w_{env} is from 0 to 0.25, which is further examined in the next iteration. The process will terminate if no further improvement can be achieved by successive refinement. Notice, the approach involves checking a large number of candidate parameters. To tackle this, we run the calibration in parallel, each run instance examining a segment of entire search space. Table 3.5 shows parameter space, MSE, fitted percentage and number of parameters for each iteration. The final model (round 7) has the following parameters,

$$\theta^* = (w^*_{eco}, w^*_{env}, w^*_{inc}, w^*_{com}, theshold^*) = (0, 0.08, 0, 0.92, 0.9924)$$

achieving 82% of the observed aggregate adoption level. The model to some extent in-

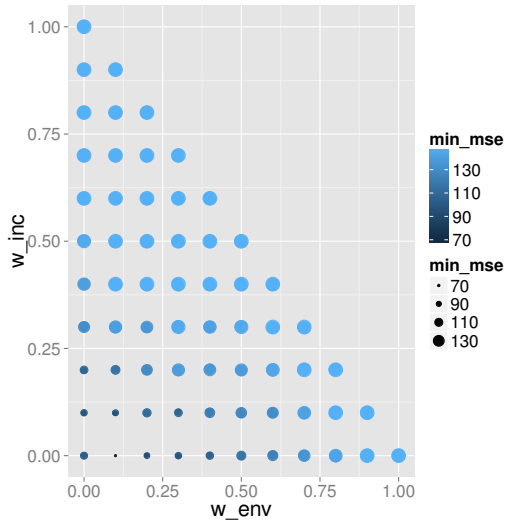


Figure 3.2: Utilities (MSE) of Parameters in 1st Iteration

Table 3.5: Iterative Localized Search

Round	w_{env}	Threshold	MSE	Fitted %	# of parameters
1	[0, 1]	[0.5, 1]	69.79	63	33000
2	[0, 0.25]	[0.98, 0.99]	82.64	78	6930
3	[0, 0.25]	[0.99, 1]	75.60	85	6930
4	[0.05, 0.11]	[0.991, 0.992]	67.21	88	7700
5	[0.05, 0.11]	[0.992, 0.993]	58.71	81	7700
6	[0.05, 0.11]	[0.9922, 0.9923]	51.96	84	7700
7	[0.05, 0.11]	[0.9923, 0.9924]	48.48	82	7700

dicates only environmental utility and communication utility are significant. Notably, the calibration process is extremely costly, i.e., each iteration takes about 6-7 hours with 70 processes running simultaneously. In contrast, the training procedure of our proposed DDABM only takes about 3 hours running on a sample of 30% entire data in a single process. For the calibrated model, the comparison between the fitted adoption and actual adoption are illustrated in Figure 3.3. The key takeaway is that the calibrated model achieves good performance with respect to the training (calibration) data. What remains to be seen is how it performs in the validation context, which is the subject of the next section.

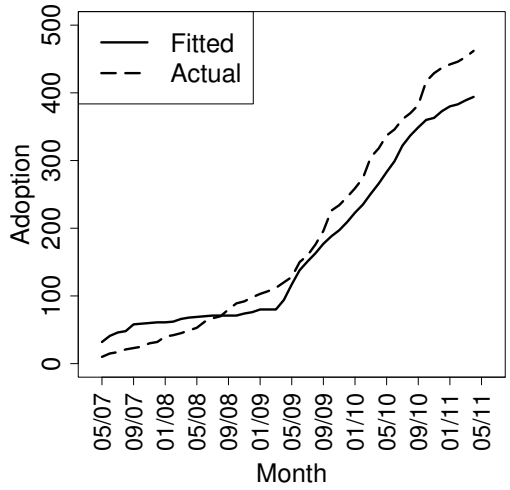


Figure 3.3: Cumulative adoption: Palmer et al. predicted vs. observed on calibration data.

3.6 ABM Validation

We have now reached Step 5 of the DDABM framework: validation. Our starting point is quantitative validation, *using data that is the "future" relative to the data used for model learning (calibration)*. Given that our agent model and, consequently, the ABM are stochastic, we validate the model by comparing its performance to a baseline in terms of *log-likelihood of observed adoption sequence* in validation data. Specifically, suppose that $D_v = \{(x_{it}, y_{it})\}$ is the sequence of adoption decisions by individuals in the validation data, where x_{it} evolves in part as a function of past adoption decisions, $\{y_{i,t-k}, \ldots, y_{i,t-1}\}$ (where k is the installation lag time). Letting all aspects relevant to the current decision be a part of the current state x_{it}, we can compute the likelihood of the adoption sequence given a model p as:

$$L(D_v; p) = \prod_{i,t \in D_v} p(x_{it})^{y_{it}} (1 - p(x_{it}))^{(1-y_{it})}.$$

Quality of a model p relative to a baseline b can then be measured using likelihood ratio, $R = \frac{L(D_v;p)}{L(D_v;b)}$. If $R > 1$, the model p outperforms the baseline. As this discussion implies, we need a baseline. We consider two baseline models: a NULL model, which estimates the

probability of adoption as the overall fraction of adopters, and a model using only the NPV and zip code adoption density features for the purchase and lease decisions (referred to as *baseline* below). The latter baseline is somewhat analogous to the model used by Lobel and Perakis [156], although it is adapted to our setting, with all its associated complications discussed above. As we found the NULL model to be substantially worse, we only present the comparison with the more sophisticated *baseline*.

To enable us to execute many runs within a reasonable time frame, we restricted the ABM to a representative zip code in San Diego county (approximately 13000 households). We initialized the simulation with the assessors features, GIS locations, and adoption states (that is, identifies of adopters) in this zip code. To account for stochasticity of our model, we executed 1000 sample runs for all models.

Figure 3.4 shows the likelihood ratio of our model (namely *lasso*) to the *baseline*. From this figure, it is clear that our model significantly outperforms the baseline in its ability to forecast rooftop solar adoption: the models are relatively similar in their quality for a number of months as the adoption trend is relatively predictable, but diverge significantly after 9/12, with our model ultimately outperforming the baseline by an order of magnitude.[4] In other words, both models predict near-future (from the model perspective) relatively well, but our model significantly outperforms the baseline in forecasting the more distance future.

Thus, quantitative validation already strongly suggests that the DDABM model we developed performs quite well in terms of forecasting the probability distribution of *individual decisions*.

In addition, we assess model performance in terms of aggregate behavior in more qualitative terms. Specifically we can consider Figure 3.5 , which shows *stochastic realizations* of our model (recall that agent behavior is stochastic), where heavier regions correspond to greater density, in comparison with the actual average adoption path. First, we can observe

[4] 9/12 is where the aggregate adoption becomes highly non-linear, so that the added value of the extra features used by our model sharply increases.

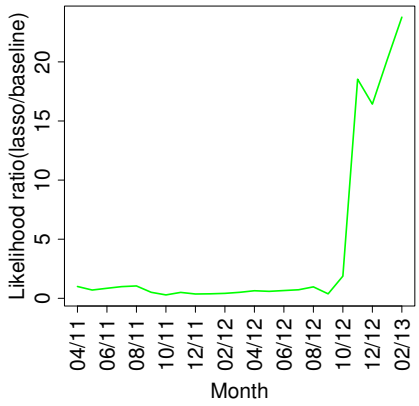


Figure 3.4: Likelihood ratio R of our model relative to the baseline.

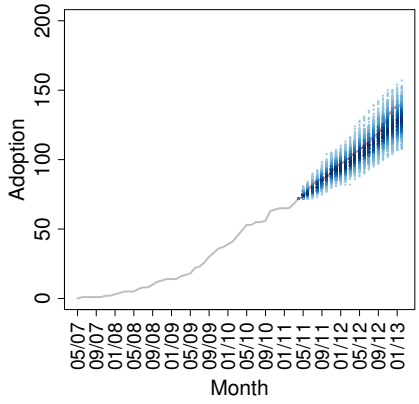


Figure 3.5: Spread of sample runs of our model, with heavier colored regions corresponding to higher density, and the observed average adoption trend.

that the actual adoption path is in the "high-likelihood" region of our model realizations. This is a crucial observation: when behavior is stochastic, it would be unreasonable to expect a prediction to be "spot-on": in fact, every particular realization of behavior path has a minuscule probability. Instead, model correctness is well assessed in terms of how likely observed adoption path is *according to the model*; we observe that our model is *very likely to produce an outcome similar to what was actually observed*. Second, our model offers a meaningful quantification of uncertainty, which is low shortly after the observed initial state, but fans out further into the future. Given that adoption is, for practical purposes, a stochastic process, it is extremely useful to be able to quantify uncertainty, and we therefore

view this as a significant feature of our model. Note also that we expect variation in the actual adoption path as well, so one would not therefore anticipate this to be identical to the model average path, just as individual sample paths typically deviate from the average.

Finally, we use the model developed in Section 3.5 to forecast adoption in the same zip code. Figure 3.6 compares the forecasting performance of the Palmer et al. model calibrated using aggregate-level adoption, and our DDABM model. While initially both models exhibit reasonable forecasting performance, after only a few months the quality diverges dramatically: the DDABM model is far more robust, maintaining a high-quality forecast at the aggregate level, whereas the baseline becomes unusable after only a few months. We propose that the primary reason for this divergence is over-fitting: when a model is calibrated to the aggregate adoption data, it is calibrated to a very "low-bandwidth" signal; in particular, there are many ways that individuals can behave that would give rise to the same *average* or *aggregate* behavior. Individual-level data, on the other hand, allows us to disentangle the microbehavior in much greater specificity and robustness, increasing the likelihood of meaningful behavior models that arise thereby, and reducing the chances of overfitting the parameters to a specific overall adoption trend.

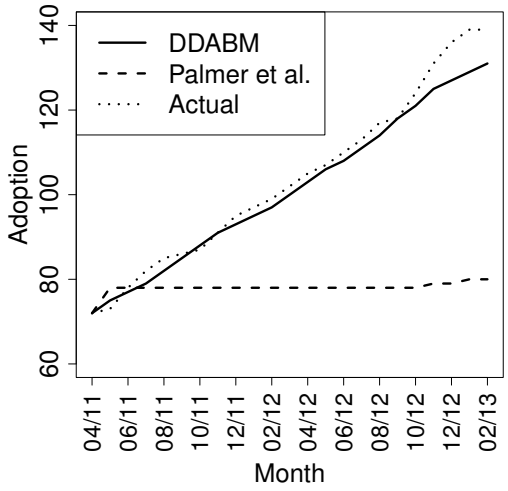


Figure 3.6: Expected Adoption: DDABM Model (mean squared error = 15.35) vs. Palmer et al. (mean squared error = 1045.30). Mean squared error measures forecasting error on evaluation data.

3.7 Policy Analysis

The model of residential rooftop solar we developed and validated can now be used both as a means to evaluate the effectiveness of a policy that had been used (in our case, California Solar Initiative solar subsidy program), and consider the effectiveness of alternative policies. Our evaluation here is restricted to a single representative zip code in San Diego county, as discussed above. We begin by considering the problem of designing the incentive (subsidy) program. Financial subsidies have been among the principal tools in solar policy aimed at promoting solar adoption. One important variable in this policy landscape is budget: in particular, how much budget should be allocated to the program to achieve a desired adoption target?

3.7.1 Sensitivity of Incentive Budget

Our first experiment compares the impact of incentive programs based on the California Solar Initiative, but with varying budget in multiples of the actual CSI program budget.[5] Specifically, we consider multiples of 0 (that is, no incentives), 1 (which corresponds to the CSI program budget), as well as $2, 4$, and 8, which amplify the original budget. To significantly speed up the evaluation (and reduce variance), rather than taking many sample adoption paths for each policy, we compare policies in terms of expected adoption path. This is done as follows: the simulation still generates 1000 sample "new" states, i.e., realizations of the probabilistic adoption decision, at each time step, but only uses the one with average number of adopters as initial state for the next time step.

Figure 3.7 shows the effectiveness of a CSI-based subsidy program on expected adoption trends over the full length of the program. As one would expect, increasing the budget uniformly shifts average adoption up. Remarkably, however, the shift is relatively limited,

[5] It is important to note that the CSI program has many facets, and promoting solar adoption directly is only one of its many goals. For example, much of the program is focused on improving marketplace conditions for solar installers. Our analysis is therefore limited by the closed world assumption of our simulation model, and focused on only a single aspect of the program.

even with 8x the original budget level. Even more surprisingly, the difference in adoption between no subsidies and incentives at the CSI program levels is quite small: only several more individuals adopt in this zip code, on average.

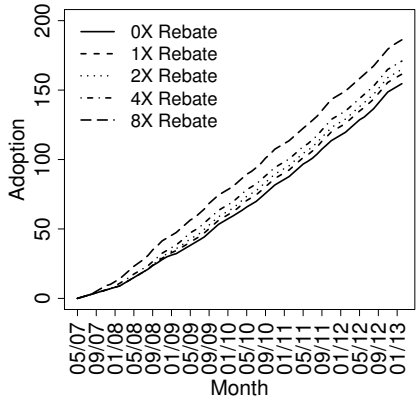


Figure 3.7: Adoption trends for the CSI-based subsidy structure.

3.7.2 Design of Incentive

Since we found that the CSI-like solar system subsidies have rather limited effect, a natural question is whether we can design a better subsidy scheme.

3.7.2.1 Problem Formulation

The incentive design problem can be formulated as follows. Assume we are given a fixed budget B, which supposed to subsidize solar adopters in T steps. The amount of incentive a household can get is simply multiplication of system capacity (kilowatt) and subsidy rate (dollar/watt). As a step-wise incentive structure, each step is associated with a fixed rate r_t and terminates as an accumulative target in megawatt m_t is achieved. Then, the subsidy program transits to a new step with a new rate and target. This is the exact structure of CSI program currently implemented in California shown in Figure 3.8.

Given this, the problem is to find an optimal incentive structure, $s^* = \{(r_t, m_t)\}_{0,\ldots,T}$,

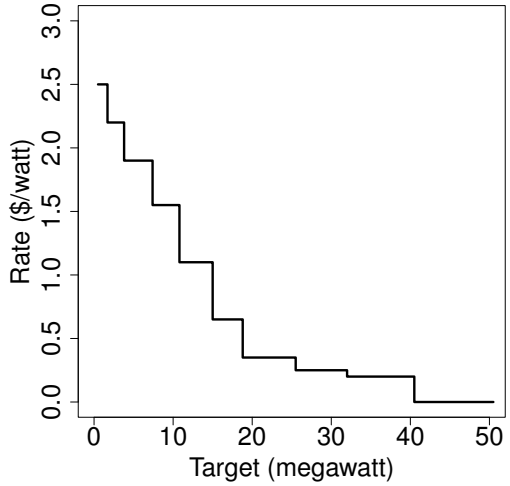


Figure 3.8: CSI Program Structure in California

which maximizes ultimate adoption simulated by ABM developed in Section 3.4,

$$s^* = \arg\max_s U_{abm}(s, B, T) \tag{3.7}$$

subject to two constraints: 1). budget constraint: $\sum_{i=0}^{T-1} r^i m^i \leq B$; and 2) non-increasing rates: $r^i \geq r^j, \forall i < j \in T$.

3.7.2.2 Parametric Optimization

We proceed by creating a parametric space of subsidy schemes that are similar in nature to the CSI incentive program. We restrict the design space by assuming that $r^{i+1} = \gamma r^i$ for all time steps i. In addition, we let each megawatt step m^i to be a multiple of the CSI program megawatt levels in the corresponding step, where the multiplicative factor corresponds to the budget multiple of the CSI program budget. This particular scheme gives rise to a set of incentive plans illustrated in Figure 3.9. With these restrictions, our only decision is about the choice of r^0, which then uniquely determines the value of γ based on the budget constraint. To choose the (approximately) optimal value of r^0, we simply considered a finely discretized space ranging from 1 to 8 \$/watt for 1x, 2x, and 4x

CSI budget. The results, in Figure 3.10 and 3.11 suggest that the impact of subsidies is quite limited even in this one-dimensional optimization context.

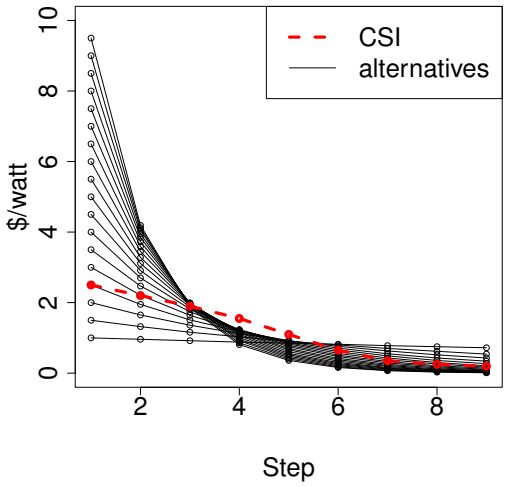


Figure 3.9: Parametric Incentive Plans

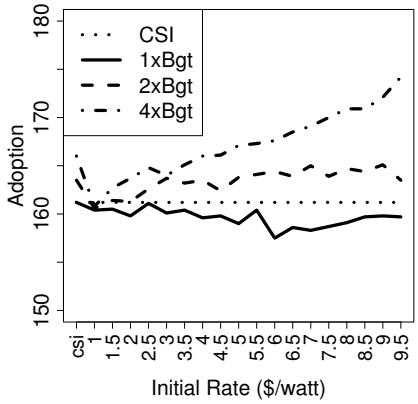


Figure 3.10: Expected Adoption over Different Initial Rates, where "kx Bgt" means k times as large as original CSI budget.

3.7.2.3 A Heuristic Search Algorithm

Given the challenge of finding effective incentive schemes, we now relax the restriction of the original CSI budget allocation pattern (see Figure 3.8), allowing now the proportion of the budget allocated each step to vary. To this end, we propose a simple heuristic

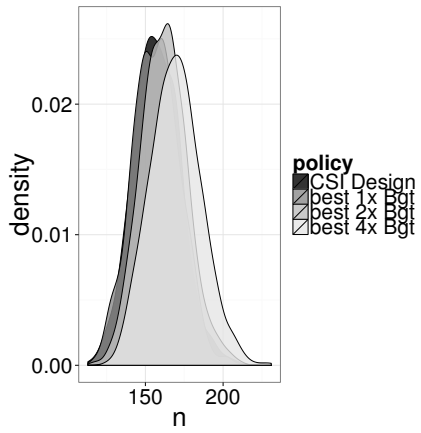


Figure 3.11: Comparison of distributions of the number of adopters (n) up to 4/13 for "optimal" incentive policies, where "kx Bgt" means k times as large as original CSI budget.

search algorithm. The algorithm is a step-wise greedy search method, with each step applying a heuristic which is learned from the previous step. The algorithm proceeds until no improvement can be achieved through the following series of steps:

1. Solve a basic one-stage incentive optimization problem, i.e., only one rate and one step, in other words, this is to uniformly spread the budget in one single term. As shown in Figure 3.12, for each r_1^i in the discretized space R_1 (i.e., equally divided 100 values in (0, 5]), we run our ABM to obtain utility $U(\{(r_1^i, m_1^i)\})$ for each policy correspondingly, s.t., $r_1^i m_1^i = B$. An optimal one-stage incentive optimization policy is defined as $s_1^* = \{(r_1^*, m_1^*)\}$, s.t., $U(\{(r_1^*, m_1^*)\}) \geq U(\{(r_1^i, m_1^i)\}), \forall \{(r_1^i, m_1^i)\} \neq \{(r_1^*, m_1^*)\}$

2. Solve a 2-stage incentive optimization problem. Rather than searching all possibilities in the discretized parameter space, the rate of the first stage for the 2-stage structure is fixed at r_1^*, as shown in Figure 3.13, by which we implicitly conjecture that r_1^* is superior to any other rates. For any possible proportion of B used in stage 1, say B_1^i, we can derive m_1^i accordingly from $r_1^* m_1^i = B_1^i$; then for each possible discretized rate r_2^i that is below r_1^*, we also determine m_2^i consequently by the budget constraint. Thus, for any arbitrary policy $s = \{(r_1^*, m_1^i), (r_2^i, m_2^i)\}$, we run ABM and

obtain its utility $U(s)$. The best policy should be

$$s^* = s(m_1^*, r_2^*) = \{(r_1^*, m_1^*), (r_2^*, m_2^*)\} = \arg\max_s U(s)$$

3. Solve a 3-stage incentive optimization problem. Similarly, as illustrated in Figure 3.14, the rate and megawatt target of the stage 1 are set to r_1^* and m_1^* respectively, and the rate of the 2nd stage is set to r_2^*. By the budget constraint, for any portion of budget B_2^i used in stage 2, one can derive m_2^i. Further, for any rate at stage 3, say r_3^i, which is below r_2^*, we can determine m_3^i similarly. Thus, for any 3-stage arbitrary policy $s = \{(r_1^*, m_1^*), (r_2^*, m_2^i), (r_3^i, m_3^i)\}$, or simply denote s as $s(m_2^i, r_3^i)$, we run ABM and obtain its utility $U(s)$. The best policy for the 3-stage problem is given by

$$s^* = s(m_2^*, r_3^*) = \{(r_1^*, m_1^*), (r_2^*, m_2^*), (r_3^*, m_3^*)\} = \arg\max_s U(s)$$

4. The algorithm will proceed unless no further utility improvement can be made in a step. The time complexity is $O(N_s N_b N_r)$, where N_s denotes number of steps in the worse case, N_b the number of discretized fractions of budget and N_r the number of discretized rates upper-bounded by the fixed rate in the preceding stage. Notice that there is also a constant factor involving running time of simulation for each parameter, but here we save it to highlight the main factors.

A comparison of expected adoption of different incentive structures is shown in Table 3.6, where "x-Budget" indicates the scale of budget relative to the original CSI subsidies, "OnePar" stands for incentive plans examined in Section 3.7.2.2 and "x-Rebate" refers to incentive structure discussed in 3.7.1. Our heuristic search method is able to find better alternative incentive plans for all budget levels. Moreover, the result suggests that an incentive plan with smaller number of steps, i.e., 1 to 3, may be better than spreading the whole budget in a large number of steps, say 10, which is currently deployed in California.

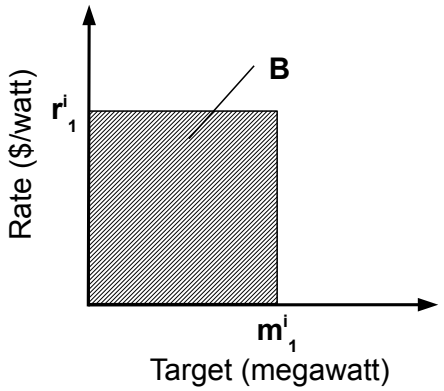


Figure 3.12: 1-stage Incentive Optimization

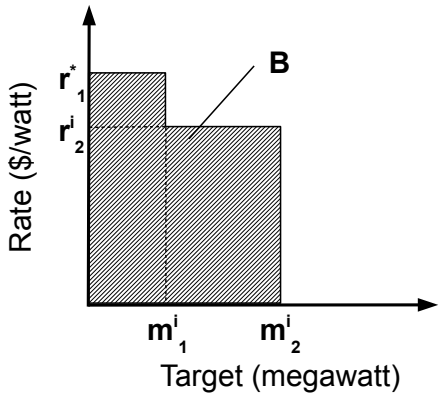


Figure 3.13: 2-stage Incentive Optimization

3.7.3 Seeding the Solar Market

Seeing a relatively limited impact of incentives, due to low sensitivity of our model to the economic variables, we also consider an alternative class of policy, called ”seeding”, which instead leverages the fact that peer effects have a positive and significantly stronger impact on adoption rates.

Suppose that we can give away free solar systems. Indeed, there are policies of this

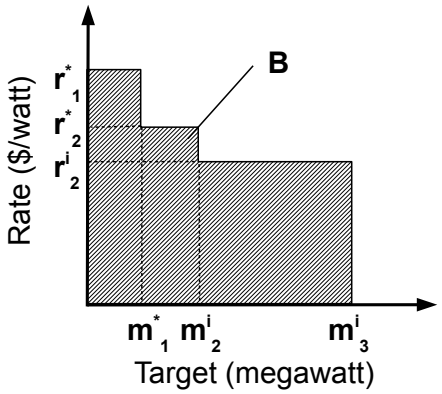


Figure 3.14: 3-stage Incentive Optimization

Table 3.6: A Comparison of Expected Adoption of Different Incentive Structures

x-Budget	OnePar	x-Rebate	1-Stage	2-Stage	3-Stage	4-Stage
1	159	161.5	163.2	163.9	-	-
2	163.8	165	166.7	-	-	-
4	167.1	170.9	171.9	172.2	172.3	-

kind already deployed, such as the SASH program in California [141], fully or partially subsidizing systems to low-income households. To mirror such programs, we consider a fixed budget B, a time horizon T, and consider seeding the market with a collection of initial systems in increasing order of cost in specific time periods (a reasonable proxy for low-income households). There is a twofold tension in such a policy: earlier seeding implies greater peer effect impact, as well as greater impact on costs through learning-by-doing. Later seeding, however, can have greater direct effect as prices come down (i.e., more systems can be seeded later with the same budget). We consider, therefore, a space of policies where a fraction of the budget α is used at time 0, and the rest at time $T-1$, and compute a near-optimal value of α using discretization.[6] Our findings, for different budget levels (as before, as multiples of the original CSI budget), are shown in Figure 3.15. We

[6]In fact, we optimize over discrete choices of alpha (at 0.1 intervals), and the optimal alpha varies with budget.

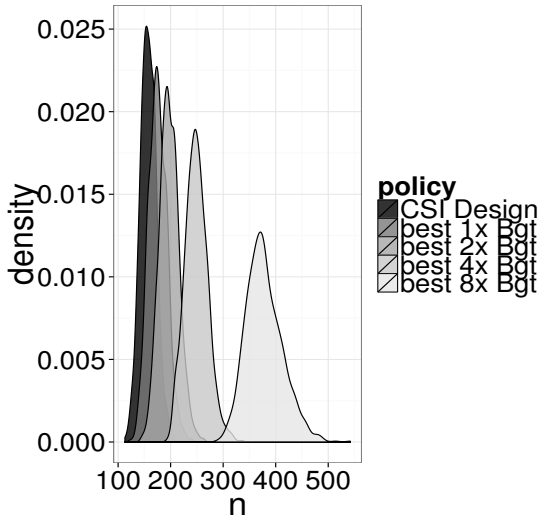


Figure 3.15: Distribution of final adoptions (n) for optimal split of the seeding budgets.

can make two key observations: first, we can achieve significantly greater adoption using a seeding policy as compared to the CSI program baseline, and second, this class of policies is far more responsive to budget increase than the incentive program.

3.8 Conclusion

We introduced a data-driven agent-based modeling framework, and used it to develop a model of residential rooftop solar adoption in San Diego county. Our model was validated quantitatively in comparison to a baseline, and qualitatively by considering its predictions and quantified uncertainty in comparison with the observed adoption trend *temporally beyond the data used to calibrate the model*. In the meantime, we developed a second agent-based model motived by state-of-the-art calibration methodology. It turned out this model severely underestimates solar adoption, poorly-performed compared to our developed agent-based model that is based on maximum likelihood estimation. We used our model to analyze the existing solar incentive program in California, as well as a class of alternative incentive programs, showing that subsidies appear to have little impact on adoption trends. Moreover, a simple heuristic search algorithm was deployed to identify

more effective incentive plans among all incentive structures we have explored. Finally, we considered another class of policies commonly known as "seeding", showing that adoption is far more sensitive to such policies than to subsidies.

Looking ahead, there are many ways to improve and extend our model. Better data, for example, electricity use data by non-adopters, would undoubtedly help. More sophisticated models of individual behavior are likely to help, though how much is unclear. Additionally, other sources of data can be included, for example, survey data about adoption characteristics, as well as results from behavior experiments in this or similar settings. The importance of promoting renewable energy, such as solar, is now widely recognized. Studies, such as ours, enable rigorous evaluation of a wide array of policies, improving the associated decision process and the increasing the chances of successful diffusion of sustainable technologies.

The DDABM framework is very promising in building high-precision ABMs from historical diffusion data, which was, unfortunately, impossible for traditional ABMs. Recall that our algorithmic marketing system not only consists of models but more importantly algorithms. Since we have already developed an effective computational approach to address the modeling challenges in marketing innovations, in the following chapters, our focus will shift to the design of efficient computer algorithms to solve for optimal or near-optimal marketing plans.

Chapter 4

Dynamic Influence Maximization Under Increasing Returns to Scale

While most diffusion models characterize the adoption of innovation as a process that shows decreasing returns to scale, i.e., the tendency for an individual to adopt is marginally decreasing as the number of adopters grows, the early stage of innovation is better represented by a process of increasing returns to scale. This is confirmed by the work presented in Chapter 3, in which the well-known logit function is used to model individual behaviors and the likelihoods are generally below 0.5 (within the convex section). Almost any marketing action comes with a cost, which is the also the case for seeding (offering free samples) marketing. Moreover, the cost often decreases over time due to technological updates and many other reasons. In these setting, this chapter studies an important marketing problem in which the marketer needs to decide number of seeds for each period in a discrete time horizon to maximize the number of adopters projected by simulation models.

4.1 Introduction

One of the important algorithmic questions in marketing on social networks is how one should leverage network effects in promoting products so as to maximize long-term product uptake. Indeed, a similar question arises in political science, if framed in terms of maximizing uptake of beliefs and attitudes, leading to particular voting behavior. Crucial to such problems is a model of social influence on networks, and a number of such models have been proposed [175, 18, 176, 16, 15, 164, 177, 178]. Some of the prominent models give rise to global influence functions (capturing the expected number of adopters as a function of a subset of initially targeted individuals) that are *submodular*, that is, possess a natural diminishing returns property: targeting a greater number of individuals yields a lower marginal increase in the outcome. Submodularity is a powerful property for algorithm de-

sign; in particular, a simple greedy heuristic has a provable approximation guarantee, and is typically very close to optimal in experiments [16, 164, 178, 179]. Submodularity as typically defined is only helpful in a static setting, that is, when we choose individuals to target in a single shot. But an extension, termed *adaptive submodularity*, was proposed to make use of greedy heuristics in dynamic environments where individuals can be targeted over time [165].

The diminishing returns feature naturally arises in many settings. However, early adoption trends can exhibit the opposite property of *increasing returns to scale*. For example, in the classic Bass model [39] the early portion of the famous "S-curve" is convex, and if one uses logistic regression to model individual adoption choice—a natural model that was used in recent work by the authors that learned individual rooftop solar adoption behavior [180]—the model is convex when probabilities are small. Arguably, early adoption settings are most significant for the development of effective product promotion strategies, since overall uptake is quite critical to the ultimate success of the product line. This is an especially acute concern in the "green energy" sector, where renewable energy technologies, such as rooftop solar, are only at a very early stage of overall adoption—indeed, adoption has been negligible except in a few states, such as California and Arizona.

We consider the problem of influence maximization with network effects by aggregating a social network into an "aggregate" adoption function which takes as input the number of adopters at a given time t and outputs the number of adopters at time $t+1$. Our main theoretical result is that when this function is convex and marketing budget can be reinvested at a fixed interest rate δ (equivalently, marketing or "seeding" costs decrease exponentially over time), influence over a finite time period can be maximized by using up the entire targeted marketing budget at a single time point (rather than splitting up the budget among multiple time periods); we refer to the resulting simple algorithm as the *Best-Stage* algorithm.

We study the degree to which the theoretical optimality of the best-time algorithm holds

in practice, using real data of rooftop solar adoption in San Diego county [180]. As a baseline, we develop a more general heuristic algorithm that splits the budget equally over a set of consecutive months, with the size of this set and the starting month allowed to vary. We find that investing the whole budget in a single month is indeed (almost) optimal under a variety of simplified seeding cost functions (exponential, polynomial, or even linear with time), despite a number of gaps between the theory and the experimental setup — suggesting that the theoretical model is somewhat robust.

In contrast, the best-time algorithm becomes suboptimal in an "ideal" model that was previously validated to confirm its predictive efficacy [180]. Through careful analysis, we find that this is the result of "learning-by-doing" effects, where marketing costs (for example, when actual products are given away for free) depend on the total product uptake in the marketplace.

4.2 Related Work

We build on the extensive literature of economics of diffusion with network effects [181, 182, 183, 184]. Largely, however, this literature is concerned with equilibria that arise, rather than algorithmic considerations. The latter are extensively studied in the literature on influence maximization on social networks. A number of models have been proposed to quantify influence diffusion on a network, perhaps the most prominent of which are linear threshold and independent cascade models [175, 18, 176, 16, 185, 186]. These and many related models give rise to a submodular "expected adoption" function. Many past approaches to "one-shot" influence maximization take advantage of this property in algorithmic development; in particular, a simple greedy algorithm is highly effective in practice, and offers a constant-factor approximation guarantee relative to a globally optimal solution to the associated combinatorial optimization problem [16, 178, 185].

While one-shot influence maximization on social networks has received much attention, significantly less work addresses the problem of dynamic influence maximization,

where individuals can be seeded in sequence. In an important effort in this vein, Golovin and Krause show that when dynamics (and uncertainty) satisfy the property of *adaptive submodularity*, a natural *dynamic* greedy heuristic is guaranteed to be only a constant factor worse than the optimal clairvoyant algorithm [165]. Our problem is distinct from this effort in several ways. The first, and key, distinction is that we are concerned with increasing returns to scale. The second is that we capture network effects simply in terms of total numbers of all past adoptions (thus, the social network is completely connected for our purposes). The third is that we introduce another key element of tension into the problem by supposing that there is a fixed total budget allocated for seeding, and either this budget (or any portion of it) can be set aside to collect interest, or the costs of seeding fall over time (commonly, costs of products are expected to fall over time as a result of learning-by-doing, or supply-side network effects where better processes and technology reduce production costs with increasing product uptake and experience in the marketplace). To our knowledge, we are the first to consider the algorithmic problem of dynamic influence maximization in such a context.

4.3 The Dynamic Influence Maximization Model

We consider a problem of adoption diffusion with network externalities. Our model of network externalities is simplified to consider only the aggregate adoption, which we denote by D. Specifically, we assume that the diffusion takes place in discrete time, and the aggregate adoption at time t, D_t, is a function only of adoption by the previous time step, D^{t-1}: $D_t = f(D^{t-1})$. In other words, adoption dynamics are deterministic and first-order Markovian. We make three assumptions on the nature of the diffusion function f:

1. f is strictly convex (i.e., increasing returns to scale),

2. f is strictly monotonically increasing, and

3. $\forall D > 0, f(D) > D$.

Assumptions (2) and (3) ensure that aggregate adoption increases over time (i.e., with every application of f); note that they are not redundant. Assumption (1) of increasing returns to scale captures a common model of early adoption dynamics (the convex portion of the "S-curve" [39], or the logistic regression model of adoption over a low-probability region discussed in our experiments below). We suppose that at time 0 (that is, initial decision time), there is some initial number of adopters in the population, $D^0 \geq 0$.

As stated, the problem poses no algorithmic tension: if one had a fixed budget for stimulating adoption, the entire budget should be used up immediately, as it would then take maximal advantage of network effects. We now introduce the principal source of such tension: an exponentially increasing budget. Specifically, suppose an agent is initially given a budget B^0 and any remaining budget will accrue by a factor of δ. For example, we can decide to invest residual budget at an interest rate δ. Alternatively, if we are giving away a product, its cost may decrease over time as technology matures, a process often referred to as "learning-by-doing" [171, 173, 158]. Such learning-by-doing effects are paramount when we consider technology evolution in its early stages, which is arguably the setting where we would be most interested in promoting the product by giving it away to a subset of the individuals in a population. Notice that either saving some of the budget at a fixed interest rate, or costs of seeding decreasing at a constant rate, both give rise to exponentially increasing purchasing power of the budget over time. This gives rise to a non-trivial tension between seeding early so as to maximally leverage network effects and seeding later so as to maximize the number of individuals seeded (as well as subsequent network effects). The algorithmic question we pose is: **how should we use a given budget B over a fixed time horizon T so as to maximize the total number of adopters at time T**? For simplicity, we assume unit cost for every seed; in other words, any budget B will create exactly B new adopters. This assumption will be relaxed in our experiments below.

Given the deterministic T-stage diffusion model and the budget accrual process described above, as well as an initial budget B^0 and aggregate adoption D^0, we can define a

"seeding" policy π as a sequence of fractions of the budget allocated in each stage, that is

$$\pi = (\alpha_0, \alpha_1, \ldots, \alpha_{T-1}), \quad \alpha_t \geq 0.$$

The dynamic influence maximization problem aspires to compute a policy π^* which leads to the maximum number of adopters based on the diffusion model f starting with initial adoption D^0, an initial budget B^0, and a budget growth rate δ. If we define the total number of adopters at time T as the *utility* of a policy, the problem can be written as

$$\pi^* = \arg\max_{\pi} U(f, \pi, D^0, B^0, \delta).$$

4.4 Algorithms for Dynamic Influence Maximization

To solve the dynamic maximization problem, it is convenient to view it as a T-stage decision process illustrated in Figure 4.1.

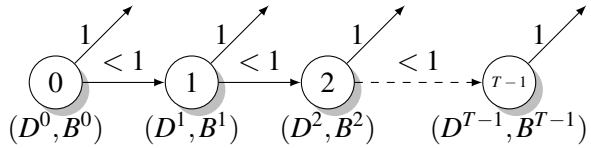


Figure 4.1: T Stage Decision Process

At any stage t, we consider two types of actions for an agent: spending all of the remaining budget ($\alpha = 1$) and fraction (or none) of budget ($0 \leq \alpha_t < 1$). Particularly, as long as one chooses $\alpha_t = 1$, the decision process is terminated and the utility can be obtained in terms of the number of final users. Otherwise, one should proceed until the budget is exhausted. Note that an agent will always spend whatever budget is left at $t = T - 1$ (since there is no utility from keeping a fraction of it intact thereafter).

In principle, one can solve the dynamic influence maximization problem using backward induction. For a T-stage decision problem, backward induction starts with considering optimal decisions at the last stage (i.e. $t = T - 1$) for all possible realizations of D

and B, then considers optimal choices at the second-to-last stage (i.e. $t = T - 2$) using the optimal decisions in the final stage, and so on. The process proceeds until the very first decision node (i.e. $t = 0$) is reached, and finally an optimal policy can be returned. However, such an approach quickly becomes intractable when the evaluation of f is very time consuming (for example, as in the instance described below, f corresponds to an agent-based simulation of individual adoption decisions). For example, if we suppose that each simulation takes only 1 second, our population involves 10,000 individuals, the budget can seed only 20 individuals, and our time horizon is 20 stages (e.g., 20 months), dynamic programming would take over 1000 hours, or about 1.5 months. Our primary goal is to develop algorithmic approaches which are orders of magnitude faster.

4.4.1 Optimal Algorithm

To start, consider Algorithm 1 (called *Best-Stage*), which simply finds a single best stage t at which to use up all of the budget accrued by this stage. Our main result in this section is that the *Best-Stage* algorithm is optimal.

Data: D^0, B^0
Result: t
$u^* = 0$;
$t^* = 0$;
for *t=0..T-1* **do**
    u = FindAdoptions($D^0, B^0\delta^t, t$);
    if $u > u^*$ **then**
        $u^* = u$;
        $t^* = t$;
    end
end

Algorithm 1: The *Best-Stage* algorithm. The function FindAdoptions(D, B, t) computes the total number of adoptions at time T when there are D adopters at time 0 and the entire budget B is used for seeding at time t.

Theorem 4.4.1. *Let t be the stage returned by the* Best-Stage *algorithm. Then the policy*

$\pi = (0,\ldots,0,B^0\delta^t,\ldots)$ *where the only non-zero entry is at time t, is optimal.*[1]

The proof of this result is rather involved, and we develop it in a series of steps.

Let us consider a general backward induction step from the descendant nodes at stage $t+1$ to their parent node at stage t. Suppose at stage t there are D^t users and B^t budget. Particularly, if the entire budget is used at t, an agent will have utility

$$U(\alpha_t = 1) = f^{(T-t)}(D^t + B^t) \tag{4.1}$$

where notation $f^{(m)}(\cdot)$ stands for applying f repeatedly for m stages. If some fraction $0 \leq \alpha_t < 1$ of the budget is used at stage t, there will be

$$D^{t+1} = f(D^t + \alpha_t B^t) \tag{4.2}$$

adopters and

$$B^{t+1} = (1+\delta)(1-\alpha_t)B^t \tag{4.3}$$

available budget in the next stage $t+1$. Let π^* be an optimal policy, and α_t^* the fraction of budget allocated to seeding under π^* in stage t. Our next series of results characterize the properties that π^* (and fractions α_t^*) must have, by virtue of its optimality.

Lemma 4.4.2. *Suppose that at some stage t the optimal policy has* $\alpha_t^* = 0$*. Further, suppose that there is some* $t^* > t$ *with* $\alpha_{t^*}^* = 1$ *and* $\alpha_{t'}^* = 0$ *for all* $t < t' < t^*$*. Then* $\alpha_{t-1}^* = 1$ *cannot be optimal.*

Proof. Let $\alpha_{t-1}^* = a$, and let D^t and B^t be the number of adopters and remaining budget at time t if the optimal policy π^* is followed. Notice that there are $T-t$ stages remaining

[1] Note that the nature of the policy after time t is irrelevant.

when we start at stage t. Thus, the utility in stage $t-1$ is given by

$$U(\alpha^*_{t-1} = a,\, \alpha^*_t = 0) = \\ f^{(T-t^*)}(f^{(t^*-t)}(D^t) + (1+\delta)^{t^*-t}B^t),$$

where $D^t = f(D^{t-1} + aB^{t-1})$, $B^t = (1+\delta)(1-a)B^{t-1}$ and $t^* > t$ is the stage of the optimal policy π^* after t at which the entire budget is used up. The lemma states that if $\alpha^*_t = 0$, then $\alpha^*_{t-1} = 1$ must not be. In formal notation, we can state this as $U(\alpha^*_{t-1} = a, \alpha^*_t = 0) > U(\alpha^*_{t-1} = 1)$ for $a < 1$. By monotonicity of f this is equivalent to

$$f^{(t^*-t)}(D^t) + (1+\delta)^{t^*-t}B^t > f^{(t^*-t+1)}(D^{t-1} + B^{t-1}). \tag{4.4}$$

By optimality of $\alpha^*_t = 0$, we further have

$$f^{(T-t^*)}(f^{(t^*-t)}(D^t) + (1+\delta)^{t^*-t}B^t) \geq f^{(T-t)}(D^t + B^t),$$

which by monotonicity of f is equivalent to

$$f^{(t^*-t)}(D^t) + (1+\delta)^{t^*-t}B^t \geq f^{(t^*-t)}(D^t + B^t). \tag{4.5}$$

If we could show that

$$f^{(t^*-t)}(D^t + B^t) > f^{(t^*-t+1)}(D^t + B^t)$$

or, equivalently,

$$D^t + B^t > f(D^{t-1} + B^{t-1}),$$

then combining this with with (4.5) will imply that (4.4) must hold. Let us rewrite (4.5) as follows

$$\frac{f^{(t^*-t)}(D^t + B^t) - f^{(t^*-t)}(D^t)}{B^t} \leq (1+\delta)^{t^*-t}, \tag{4.6}$$

or, equivalently, as

$$\frac{f^{(t^*-t)}(D^t+B^t)-f^{(t^*-t)}(D^t)}{f^{(t^*-t-1)}(D^t+B^t)-f^{(t^*-t-1)}(D^t)} \times \ldots \times \frac{f(D^t+B^t)-f(D^t)}{B^t} \leq (1+\delta)^{t^*-t}$$

Because $f^{(t^*-t)}(D^t) > \ldots > f(D^t)$ and by strict convexity, it follows that

$$\frac{f^{(t^*-t)}(D^t+B^t)-f^{(t^*-t)}(D^t)}{f^{(t^*-t-1)}(D^t+B^t)-f^{(t^*-t-1)}(D^t)} > \ldots > \frac{f(D^t+B^t)-f(D^t)}{B^t},$$

which in turn implies that

$$\frac{f(D^t+B^t)-f(D^t)}{B^t} < 1+\delta. \tag{4.7}$$

Additionally, observe that

$$D^t = f(D^t + aB^{t-1}) > D^t + aB^{t-1}$$

and

$$B^t = (1+\delta)(1-a)B^{t-1} > (1-a)B^{t-1}$$

By strict convexity, it then follows that

$$\frac{f(D^{t-1}+B^{t-1})-f(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}} < \frac{f(D^t+B^t)-f(D^t)}{B^t}$$

which together with (4.7) implies that

$$\frac{f(D^{t-1}+B^{t-1})-f(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}} < 1+\delta,$$

which is equivalent to

$$\begin{aligned} f(D^{t-1}+B^{t-1}) &< f(D^{t-1}+aB^{t-1})+(1+\delta)(1-a)B^{t-1} \\ &= D^t+B^t, \end{aligned}$$

completing the proof. □

The next lemma builds on Lemma 4.4.2 to significantly strengthen its result.

Lemma 4.4.3. *Suppose that at some stage t the optimal policy has $\alpha_t^* = 0$. Further, suppose that there is some $t^* > t$ with $\alpha_{t^*}^* = 1$ and $\alpha_{t'}^* = 0$ for all $t < t' < t^*$. Then $\alpha_{t-1}^* = 0$.*

Proof. In this lemma, we will show that it cannot be the case that $\alpha_{t-1}^* \in (0,1)$. Together with Lemma 4.4.2, it will imply the desired result.

We prove this lemma by contradiction. Suppose that the optimal choice is $\alpha_{t-1}^* = a$ with $0 < a < 1$. The optimal utility is then given by

$$U(\alpha_{t-1}^* = a, \alpha_t^* = 0) = f^{(T-t^*)}(f^{(t^*-t)}(D^t) + (1+\delta)^{t^*-t}B^t),$$

where t^* is the stage after t at which the entire budget is spent and

$$D^t = f(D^{t-1}+aB^{t-1}), B^t = (1+\delta)(1-a)B^{t-1}.$$

By optimality of π^*, this policy should be (weakly) better than a policy which spends nothing at stage $t-1$ and spends the entire budget in stage t^* (with the same t^* as above). The utility of such a policy is given by

$$U(\alpha_{t-1}^* = 0, \alpha_t^* = 0) = f^{(T-t^*)}(f^{(t^*-t)}(\bar{D}^t) + (1+\delta)^{t^*-t}\bar{B}^t),$$

where

$$\bar{D}^t = f(D^{t-1}), \bar{B}^t = (1+\delta)B^{t-1}.$$

Since we assume $U(\alpha^*_{t-1}=0,\alpha^*_t=0)\leq U(\alpha^*_{t-1}=a,\alpha^*_t=0)$,

$$f^{(t^*-t)}(f(D^{t-1}))+(1+\delta)^{t^*-t+1}B^{t-1}\leq$$
$$f^{(t^*-t)}(f(D^{t-1}+aB^{t-1}))+(1+\delta)^{t^*-t+1}(1-a)B^{t-1}$$

or

$$f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})-f^{(t^*-t+1)}(D^{t-1})$$
$$\geq(1+\delta)^{t^*-t+1}aB^{t-1},$$

which we can rewrite as

$$\frac{f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})-f^{(t^*-t+1)}(D^{t-1})}{aB^{t-1}}\geq(1+\delta)^{t^*-t+1}.$$

On the other hand, by strict convexity of f

$$\frac{f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})-f^{(t^*-t+1)}(D^{t-1})}{aB^{t-1}}$$
$$<\frac{f^{(t^*-t+1)}(D^{t-1}+B^{t-1})-f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}}.$$

Moreover, by the same argument as used in the proof of Lemma 4.4.2 to arrive at (4.6), the optimal choice a must satisfy

$$\frac{f^{(t^*-t+1)}(D^{t-1}+B^{t-1})-f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}}$$
$$\leq(1+\delta)^{t^*-t+1},$$

which implies that

$$\frac{f^{(t^*-t+1)}(D^{t-1}+aB^{t-1})-f^{(t^*-t+1)}(D^{t-1})}{aB^{t-1}}<(1+\delta)^{t^*-t+1},$$

a contradiction. □

Lemma 4.4.4. *Suppose that $\alpha_t^* = 1$. Then $\alpha_{t-1}^* \in (0,1)$ cannot be optimal.*

Proof. The utility of choosing $\alpha_{t-1}^* = a$, where $a \in [0,1)$ at stage $t-1$ is given by $U(\alpha_{t-1}^* = a, \alpha_t^* = 1) = f^{(T-t)}(D^t + B^t)$, where $D^t = f(D^{t-1} + aB^{t-1})$ and $B^t = (1+\delta)(1-a)B^{t-1}$. Suppose that $a \neq 1$. Then $U(a=1) \leq U(0 \leq a < 0, \alpha_t^* = 1)$, which means that

$$\begin{aligned} &f^{(T-t+1)}(D^{t-1} + B^{t-1}) \\ &\qquad \leq f^{(T-t)}(f(D^{t-1} + aB^{t-1}) + (1+\delta)(1-a)B^{t-1}), \end{aligned}$$

or, by monotonicity,

$$f(D^{t-1} + B^{t-1}) \leq f(D^{t-1} + aB^{t-1}) + (1+\delta)(1-a)B^{t-1}.$$

Rearranging, we obtain

$$\frac{f(D^{t-1} + B^{t-1}) - f(D^{t-1} + aB^{t-1})}{(1-a)B^{t-1}} \leq 1 + \delta. \tag{4.8}$$

On the other hand, the presumption of optimality of $\alpha_t^* = 1$ implies that $\alpha_t = 1$ is a (weakly) better choice than $\alpha_t = 0$. In other words, spending all budget at stage t is nevertheless better than saving it and spending at any other stages after t. This implies that for any $\tau > t$,

$$f^{(T-\tau)}(f^{(\tau-t)}(D^t) + (1+\delta)^{\tau-t}B^t) \leq f^{(T-t)}(D^t + B^t).$$

By monotonicity this is equivalent to

$$f^{(\tau-t)}(D^t) + (1+\delta)^{\tau-t}B^t \leq f^{(\tau-t)}(D^t + B^t),$$

which we can rewrite as

$$\frac{f^{(\tau-t)}(D^t+B^t)-f^{(\tau-t)}(D^t)}{B^t} \geq (1+\delta)^{\tau-t}.$$

In particular, if we let $\tau = t+1$, we have

$$\frac{f(D^t+B^t)-f(D^t)}{B^t} \geq 1+\delta.$$

By strict convexity and assumption (3) on f, it follows that

$$\frac{f(D^t+B^t)-f(D^t)}{B^t} > \frac{f(D^{t-1}+B^{t-1})-f(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}}.$$

We now show that if $a \neq 1$, it must be the case that $a = 0$. Notice that if $U(a=0,\alpha_t=1) \leq U(0<a<1,\alpha_t^*=1)$, it must be that

$$\begin{aligned} & f^{(T-t)}(f(D^{t-1})+(1+\delta)B^{t-1}) \\ \leq\ & f^{(T-t)}(f(D^{t-1}+aB^{t-1})+(1+\delta)(1-a)B^{t-1}), \end{aligned}$$

or, equivalently,

$$f(D^{t-1})+(1+\delta)B^{t-1} \leq f(D^{t-1}+aB^{t-1})+(1+\delta)(1-a)B^{t-1},$$

which can be written as

$$\frac{f(D^{t-1}+aB^{t-1})-f(D^{t-1})}{aB^{t-1}} \geq 1+\delta. \tag{4.9}$$

By strict convexity, we have

$$\frac{f(D^{t-1}+aB^{t-1})-f(D^{t-1})}{aB^{t-1}} < \frac{f(D^{t-1}+B^{t-1})-f(D^{t-1}+aB^{t-1})}{(1-a)B^{t-1}},$$

which, together with (4.8) and (4.9) implies that $1+\delta < 1+\delta$, a contradiction. Consequently, either $a = 0$ or $a = 1$. □

Armed with Lemmas 4.4.2-4.4.4, we are now ready to prove our main result.

of Theorem 4.4.1. We prove the theorem by induction.

Base Case: $T = 1$ For an 1-stage decision problem, it is clearly optimal to spend the entire budget budget, so the optimal policy is $\alpha_0^* = 1$.

Inductive Step: Suppose the argument holds for $T = K \geq 1$. That is for any K-stage decision problem, the optimal policy is using all budget at some stage t^*, s.t, $t^* \in 0, 1, \ldots, K-1$. Let us now consider a $T = K+1$-stage decision problem. Assume we are at the final backward induction step from nodes in stage $t = 1$ the root node (i.e. $t = 0$). Note that those different decision nodes at stage $t = 1$ are results from different values of α_0 picked at $t = 0$. Generally, for any node (D^1, B^1) in stage $t = 1$, by the inductive assumption, an optimal policy must spend the entire budget at some stage between stage 1 and $T-1$; in other words, starting in stage $t = 1$, *Best-Stage* is optimal. Such a policy can only be one of the following three types:

1. $\alpha_0^* = 1$: spend the entire budget at $t = 0$;
2. $\alpha_0^* = 0$: spend all budget at a single stage $t^* > 0$;
3. $0 < \alpha_0^* < 1$: spend only a fraction of budget at stage $t = 0$, and the rest at a single stage $t^* > 0$.

Clearly, we only need to rule out the third type.

Case 1: $\alpha_1^* = 0$ By Lemma 4.4.3, it must be the case that $\alpha_0^* = 0$.

Case 2: $\alpha_1^* = 1$ By Lemma 4.4.4, it must be the case that either $\alpha_0^* = 0$ or $\alpha_0^* = 1$. In either case, the optimal policy has the form of *Best-Stage*. □

The *Best-Stage* algorithm is clearly far faster than dynamic programming, with running time $O(T)$, compared to $O(DBT)$ for the former, where D is the size of the population, B

the maximum budget, and T the time horizon. In our example above, it will take only 20 seconds, as compared to 1.5 months!

4.4.2 A Heuristic Search Algorithm

The model we described above is clearly stylized. As with any stylized model, a natural question is how far its assumptions can be relaxed without losing the guarantees that come with them—in our case, a very simple and provably optimal algorithm for dynamic influence maximization. If we are to undertake such an analysis experimentally, it is simply not feasible to use dynamic programming as a baseline, for reasons made clear earlier. As an alternative, we propose a heuristic algorithm which generalizes the *Best-Stage* algorithm, but does not incur too prohibitive a cost in terms of running time. Rather than choosing a single best stage, this algorithm, which we term *Best-K-Stages* (see Algorithm 2) iterates over integers K between 1 and $T-1$, splitting the budget equally among K consecutive time steps. This algorithm only takes $O(T^2)$, which is still quite manageable (about 200 seconds in our example above).

$U \leftarrow -\infty$;
$\pi^* \leftarrow \emptyset$;
for *i =1,...,T-1* **do**
  for *j = 0,..., T-i-1* **do**
    $\pi \leftarrow (\alpha^j, \ldots, \alpha^{j+i-1})$;
    if $U(M, \pi, B^0, \delta) > U$ **then**
      $U \leftarrow max\{U, U(M, \pi, B^0, \delta)\}$;
      $\pi^* \leftarrow \pi$;
    end
  end
end
return π^*

Algorithm 2: The *Best-K-Stages* Algorithm.

4.5 Experiments

To verify the efficacy of our proposed algorithm, we utilize an agent-based simulation introduced by [180] which forecasts rooftop solar adoption in a representative zip code in San Diego county. This simulation has several features of relevance to us. First, it utilizes agent models which were learned from actual adoption, as well as auxiliary data, using a logistic regression model. Second, the model includes *network effects*: aggregate adoption is one of the variables that has a significant impact on individual adoption decisions. Third, the solar market is still very much in its developing stages, even in San Diego county, and consequently the relevant region of the logistic regression model is convex in the network effects variable. Fourth, while in the policy context of this model the budget remains fixed over the T-stage time horizon, costs decrease over time. In particular, in the previously validated version of the model [180], costs actually decrease as a function of "learning-by-doing", modeled as aggregate adoption in San Diego county (which in turn increases over time). Moreover, the costs decrease linearly, rather than exponentially. As a result, the cost in this simulation does not satisfy the assumptions that guarantee the optimality of the *Best-Stage* algorithm.

An additional deviation from the idealized model we consider above is that the system cost enters the consideration in two ways: first, decreasing cost translates into an effectively increasing budget, and second, cost enters the adoption model directly as a part of *economic* considerations in rooftop solar adoption. A final important imperfection is that the budget need not be perfectly divisible by cost. In this section, we systematically investigate the extent to which these deviations from the "ideal" modeled above impact the optimality of the *Best-Stage* algorithm.

Throughout, time is discretized in months, and we fix our time horizon at 24 months (that is, 24 stages). We use the *Best-K-Stages* as our benchmark, observing in each case below expected adoption at horizon T as a function of K. Thus, when we find $K = 1$ yields an optimal or a near-optimal solution, we conclude that the deviation from model assumptions

is not significant, while optimal $K > 1$ suggests the converse. We used the budget allocated by the California Solar Initiative (CSI) to the San Diego county incentive program as our baseline, and consider amplifications of this budget, denoted by "BX". For example, 50X budget means that 50 times the CSI budget was used in an experiment. Moreover, we explored the impact of the magnitude of the peer effect variable in the model by considering as a baseline the network effect coefficient produced by learning the adoption model from data, as well as its amplifications, denoted by "PX"; for example "2X" network effect means that we doubled the network effect coefficient.

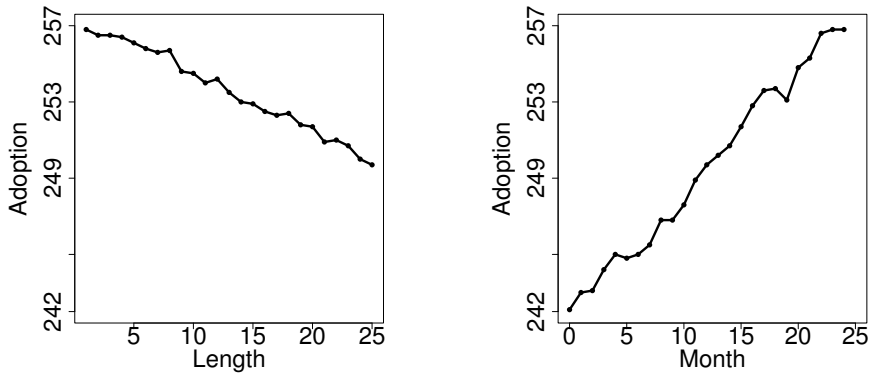
4.5.1 Exponential Cost

Our hypothesis is that the primary consideration is the nature of the cost model, with the remaining "imperfections" introduced by the complex considerations of the agent-based model in question being significantly less important. To investigate, we replace the cost model in the original model with a much simpler cost function which decreases exponentially with time t (equivalently, the budget is exponentially increasing over time): $C(t) = C_0 e^{\omega t}$, or, equivalently, $log(C) = log(C_0) + \omega t$. We used the rooftop solar cost data for San Diego county to estimate the parameters $log(C_0)$ and ω in this model (see Table 4.1). The new cost model was then "injected" into the otherwise unmodified agent-based model.

Table 4.1: Exponential Cost Model ($R^2 = 0.020$)

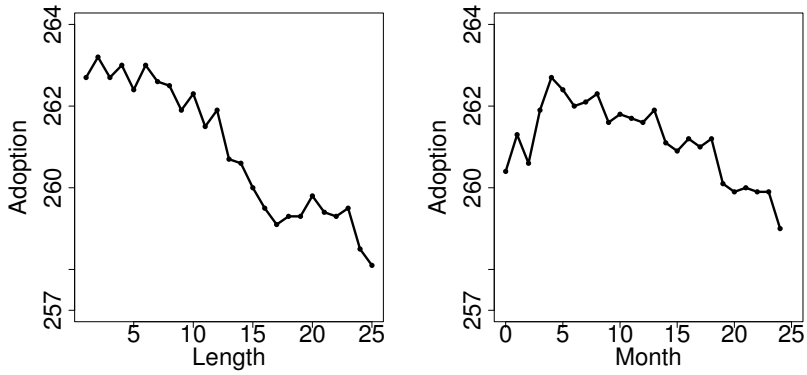
Parameter	Coefficient
$log(C_0)$	10.55
ω	-0.0059

The maximum expected adoption of seeding policies over different lengths (values of K in *Best-K-Stages*) is illustrated in Figure 4.2a. The results suggest that seeding in a single stage tends to be a better policy than splitting the budget over multiple stages. Moreover, for length-1 policies we find that in this case seeding at very end (i.e., in stage $T-1$) is



(a) Different Lengths (K values). (b) Different Seeding Month.

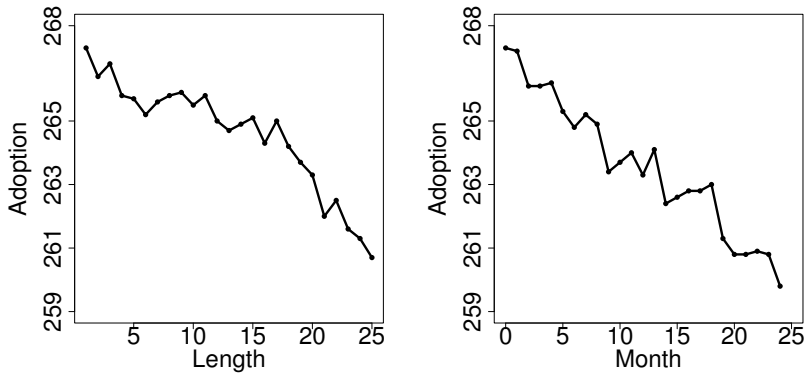
Figure 4.2: Exponential Cost: 50X Budget, 1X network effects.



(a) Different Lengths (K values) (b) Different Seeding Month

Figure 4.3: Exponential Cost: 50X Budget, 1.75X network effects.

optimal, as shown in Figure 4.2b, where the horizontal axis is the seeding month. This is likely because benefit of the exponential cost decay in this variation of the model exceeds the gains from seeding early due to network effects. As the importance of network effects increases, we expect that seeding earlier would become more beneficial. To investigate, we manually varied the coefficient of network effects in the adoption model, multiplying it by a factor of 1.75 and 2, and comparing the outcomes. Seeding in a single stage is still quite effective (Figure 4.3a and 4.4a; the jagged nature of the plots is likely due to the indivisibilities discussed above), but the optimal month to seed shifts earlier, as shown in Figure 4.3b and 4.4b.



(a) Different Lengths (K values). (b) Different Seeding Month

Figure 4.4: Exponential Cost: 50X Budget, 2X network effects.

4.5.2 Original Agent-Based Model

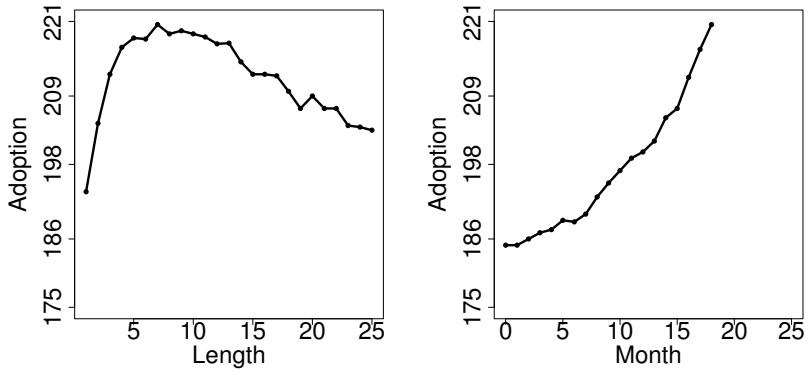
As shown in the previous section, the *Best-Stage* algorithm performs impressively despite a number of imperfections in the agent-based model of aggregate rooftop solar adoption. Next, we test its effectiveness in the context of an "ideal" model that was previously thoroughly validated in forecasting solar adoption. Significantly, the cost function used in this model includes a number of relevant parameters (such as system size), and in place of explicit dependence on time, it is a decreasing function of the overall uptake of solar systems in San Diego county (see Table 4.2). Finally, the cost function is modeled as linear in its parameters (with a fixed lower bound at zero).

Table 4.2: Cost function in the original agent-based model ($R^2 = 0.8399$).

Predictor	Coefficient
(Intercept)	11,400
Home Value	7.38e-04
Living Square Feet	0.15
System Capacity	6,210
San Diego County Adoption	-1.06

The results in Figure 4.5a are surprising: the *Best-Stage* algorithm performs *significantly* worse than policies that split budget over a relatively long contiguous series of

months ($K > 5$). Figure 4.5b further reveals that we optimally wish to push the initial month of this contiguous sequence rather late in the period; in other words, network effects are relatively weak. The key question is: what goes wrong? The observations in the previous section strongly suggest that it is the form of the cost function that is at the heart of the issue. We therefore proceed to carefully investigate what, precisely, about the nature of the cost function in this model is the cause of this qualitative change relative to our stylized model in Section 4.3. Specifically, we start with a simplified model of solar adoption that conforms to the assumptions of our main result (i.e., using only time as a variable), and incrementally relax it to bring it closer to the cost function actually used in the original simulation environment. In particular, we begin with a polynomial cost function, proceed to investigate a linear cost model (in time only), and finally consider a linear cost function that depends on aggregate product uptake (learning-by-doing) rather than time.



(a) Different Lengths (K values). (b) Different Start Month

Figure 4.5: Actual Cost Learning-by-doing: 10X Budget,1X network effects.

4.5.3 Polynomial Cost

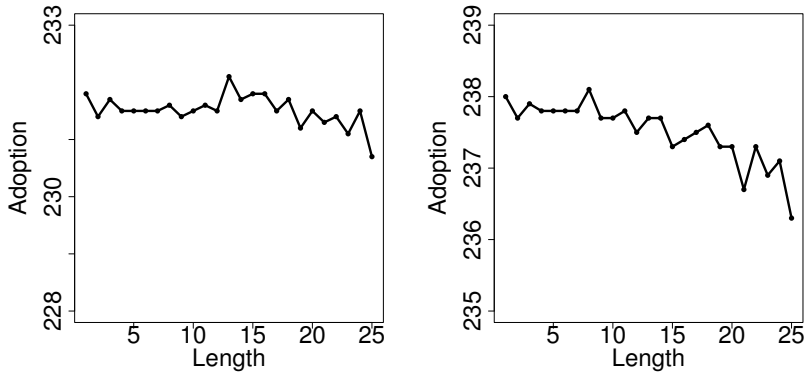
The exponential cost function, or exponentially increasing budget, seems quite dramatic. What if we slow this growth in budget buying power to be polynomial?

We formulate the following simple model of polynomial cost: $C(t) = C_0 t^{\theta}$, where t is time variable, and C_0 and θ are parameters. This function is equivalent to $\log(C(t)) =$

$\log(C_0) + \theta \log(t)$, which we can fit using linear regression. The resulting coefficients are given in Table 4.3.

Table 4.3: Polynomial Cost Model ($R^2 = 0.014$)

Parameter	Coefficient
$\log(C_0)$	10.70
$\log(t)$	-0.098



(a) 0.5X network effects.

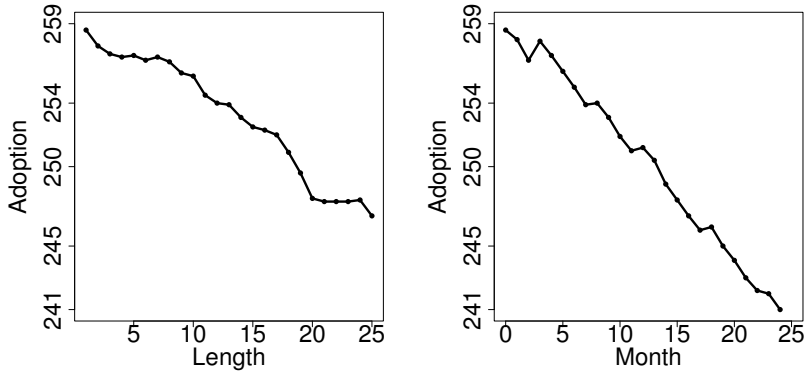
(b) 1X network effects.

Figure 4.6: Polynomial Cost: 50X Budget.

Figures 4.6a, 4.6b, and 4.7a suggest that *Best-Stage* is still a very good policy here, albeit the jagged nature of the plots (likely due to indivisibilities) makes this observation somewhat equivocal when network effects are very weak. However, as the magnitude of network effects increases, the advantage of *Best-Stage* over alternatives becomes more pronounced. On balance, it seems clear that the polynomial vs. exponential nature of the cost function does not give rise to a qualitative difference in the effectiveness of our underlying model.

4.5.4 Linear Cost

Given that the polynomial cost model did not appear to bring about a substantial difference, we proceed to relax to a linear cost model, inching even closer to the "ideal" model



(a) Different Lengths (K values). (b) Different Seeding Month.

Figure 4.7: Polynomial Cost: 50X Budget, 2X network effects.

used in the simulation. Our linear cost model has linear dependence on time, implying a slower decay rate than the polynomial cost function: $C(t) = a + bt$. As before, parameters a and b are estimated using solar system cost data for San Diego county, with the results given in Table 4.4. As before, we ran the experiments by "plugging in" this cost model into

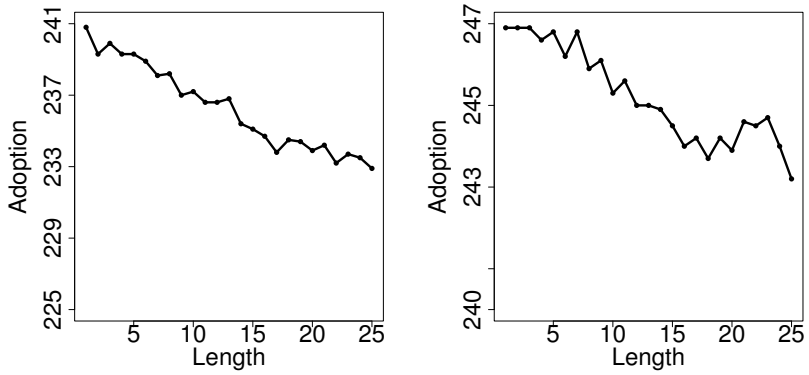
Table 4.4: Linear Cost Model ($R^2 = 0.012$)

Parameter	Coefficient
Intercept	42,053
t	-201

the simulation environment (retraining the individual adoption propensities). Our experiments show that seeding in a single month is, again, more effective than seeding in multiple consecutive stages (See Figure 4.8a and 4.8b), even as we vary the importance of network effects.

4.5.5 Linear Cost with Learning-by-Doing

Both "ideal" cost model and linear cost model are linear in their features. The key difference is that the ideal cost function depends on cumulative adoption, whereas the latter only depends on time. We now move yet closer to the ideal model, replacing the linear



(a) 1X network effects. (b) 2X network effects.

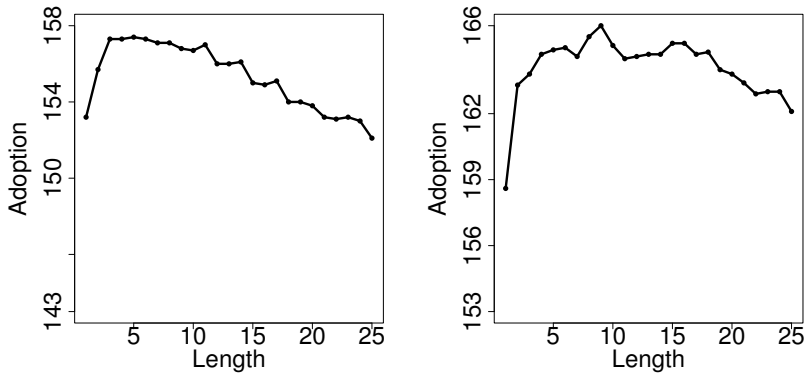
Figure 4.8: Linear Cost: 50X Budget.

dependence on time with linear dependence on aggregate solar system uptake in San Diego county: $C(t) = c + dy(t)$, where $y(t)$ is number of solar adoption up to time t. The parameters c and d are estimated via linear regression, and are given in Table 4.5.

Table 4.5: Linear Cost Model with Learning-by-Doing ($R^2 = 0.013$)

Parameter	Coefficient
(Intercept)	40,356
y	-1.74

Remarkably, the results now echo our observations in the "ideal" model (Figures 4.9a and 4.9b): *Best-Stage* is decidedly suboptimal, and policies that split the budget among $K = 5$ or more consecutive stages perform significantly better. Additionally, we can see that the "optimal" number of stages to seed (at least in our heuristic algorithm) increases as the magnitude of network effects increases (Figure 4.9a and 4.9b). The key distinction from the idealized model is that learning-by-doing makes the temporal benefit of waiting endogenous: now seeding earlier will directly reduce future seeding costs and, consequently, the effectiveness of residual budget. As a result, we observe what amounts to an "interior" optimum in budget allocation, with some of the budget used in seeding in order to make residual budget more valuable later.



(a) 1X network effects. (b) 2X network effects.

Figure 4.9: Linear Cost with Learning-by-Doing: 10X Budget

4.6 Conclusion

We formulate a novel dynamic influence maximization problem under increasing returns to scale and prove that the optimal policy must use up budget at a single stage, giving rise to a simple *Best-Stage* search algorithm. In addition, we propose a heuristic algorithm, *Best-K-Stages*, that includes *Best-Stage* as a special case. We experimentally verify that the proposed *Best-Stage* algorithm remains quite effective even as we relax the assumptions to different time-involved cost dynamics, i.e., polynomial and linear cost. On the other hand, we find that when we replace the time dependency of the cost function by cumulative adoption (learning-by-doing), *Best-K-Stages* significantly outperforms *Best-Stage*.

Looking forward, there are several possible directions we would like to pursue in the future. First, it is clear that there must exist even better policies remaining unexplored for the "ideal" cost model with learning-by-doing. Our heuristic search algorithm only covers a special subset of policies. Design of more efficient algorithms to find optimal solutions in such a realistic setting can be a meaningful extension to our current work. Second, the dynamic influence maximization problem proposed in this work assumes seeding does not discriminate among individuals. This is a very strong assumption, but enables us to make significant progress. Seeding in a social network with heterogeneous individuals has been

shown to be *NP*-hard even for "one-shot" decisions and a simple submodular diffusion model [16]. A relaxation to individual heterogeneity is sure to create further algorithmic challenges.

Chapter 5

Submodular Optimization with Generalized Cost Constraints

The diffusion model that exhibits increasing returns to scale as presented in Chapter 4 is critical for the design of effective marketing policies to promote innovations at their early stages. Note that most diffusion models still follow the classic property of decreasing return to scale, which mathematically has a submodular influence function. While the cost constraints studied in submodular influence maximization are often simple (cardinality constraint), in contrast, many real marketing applications need to consider more complex cost structures. For example, in door-to-door marketing, a salesman does not only incur a cost while visiting each customer but also expenses while traveling along the physical routes. To tackle the marketing challenge due to cost complexity, this chapter proposes an efficient heuristic algorithm and demonstrates its efficacy by extensive experiments.

5.1 Introduction

There has been much work on submoduar maximization with cardinality constraints [187] and additive/modular constraints [188, 189, 23, 24]. In many applications, however, cost constraints are significantly more complex. For example, in mobile robotic sensing domains, the robot must not only choose where to take measurements, but also plan a route among measurement locations, where cost constraints can reflect battery life. As another example, door-to-door marketing campaigns involve not only the decision about which households to target, but also the best route among them, and the constraint reflects the total time the entire effort takes (coming from work schedule constraints). Unlike the typical additive cost constraints, such route planning constraints are themselves *NP-Hard* to evaluate, necessitating approximation in practice.

We tackle the problem of maximizing a submodular function subject to a general cost

constraint, $c(X) \leq B$, where $c(X)$ is the optimal cost of covering a set X (for example, by a walk through a graph that passes all nodes in X). We propose a generalized cost-benefit greedy algorithm, which adds elements in order of marginal benefit per unit marginal cost. A key challenge is that computing (marginal) cost of adding an element (such as computing the increased cost of a walk when another node is added to a set) is often itself a hard problem. We therefore relax the algorithm to use a polynomial-time approximation algorithm for computing marginal cost. We then show that when the cost function is approximately submodular, we can achieve an approximation guarantee using this modified algorithm, which runs in polynomial time. While most work in the literature deals with only one actor, we show that the algorithm we developed for single-actor optimization can be extended to solve a multi-actor submodular maximization problem, also with provable approximation guarantees. To our knowledge, this offers the most generally applicable theoretically grounded approach in our domain known to date.

Our experiments consider two applications: mobile robotic sensing and door-to-door marketing.[1] In the former, we use sensor data on air quality in Beijing, China collected from 36 air quality monitoring stations, with a hypothetical tree-structured routing network among them. The objective in this case is to minimize conditional entropy of unobserved locations, given a Gaussian Process model of joint sensor measurements, subject to routing cost constraints (e.g., battery life or time). In the latter, we use rooftop solar adoptions from San Diego county as an example, considering geographic proximity as a social influence network and the actual road network as the routing network. The goal is to assign a subset of individuals on the social network to a single marketer or to a collection of marketers so as to maximize overall adoption given a fixed budget. In both these domains, for either the single or multi-actor settings, we show that the proposed algorithm significantly outperforms competition, both in terms of achieved utility, and, often, in terms of running time. Remarkably, this is true even in cases where the assumptions in our theoretical guarantees

[1] Due to the generality of submodular optimization, we evaluate our algorithms in multiple domains.

do not meaningfully hold.

In summary, this chapter makes the following contributions, building on work previously presented at AAAI 2016 [32]:

1. a formulation of submodular maximization under general cost constraints (routing constraints are of particular interest) including both single-actor and multi-actor scenarios (the latter is a natural extension of the fundamental problem studied in Zhang and Vorobeychik [32]);

2. a novel polynomial-time generalized cost-benefit algorithm with provable approximation guarantees for single-actor submodular maximization (specifically, we prove a non-bicriterion approximation guarantee, in contrast to the weaker bicriterion results in Zhang and Vorobeychik [32]);

3. a combination of a *sequential planning* algorithm and generalized cost-benefit algorithm to solve the multi-actor submodular maximization problem with provable approximation guarantees (this problem was not considered in Zhang and Vorobeychik [32]);

4. an application of our framework and algorithm to mobile robotic sensing, using real location and measurement data from air quality sensors in Beijing, China;

5. an application of our framework and algorithm to a *novel formalization* of a door-to-door marketing problem with social influence, using real road network and single-family home location data in San Diego county, CA from which we obtain a routing and social influence networks, respectively;

6. an evaluation of the performance of our algorithm using two approximate algorithms for computing a shortest walk: the nearest neighbor algorithm and the Christofides algorithm (only the nearest neighbor algorithm was used in Zhang and Vorobeychik [32]).

5.2 Related Work

Submodular optimization has received much attention due to its breadth of applicability, with applications including viral marketing, information gathering, image segmentation, and document summarization [20, 21]. A number of efforts consider submodular optimization under cardinality or additive cost constraints [187, 188, 189, 23, 24], demonstrating the theoretical and practical effectiveness of simple greedy and cost-benefit algorithms in this context. The problem of minimizing travel cost to cover a set of nodes on a graph, which gives rise to routing constraints that motivate our work, is a variant of the Traveling Salesman Problem (TSP), although in our variations we allow visiting the same nodes multiple times (this variation is sometimes called the Steiner TSP, or STPS) [190]. We adopt two well-known algorithms for approximating the shortest coverage route. The first is referred to as a nearest-neighbor heuristic [191], and is a $\log(n)$-approximation. The second is the well-known Christofides algorithm [192], which is slower, but yields a better 3/2 approximation factor. [2] Moreover, it is known that TSP has submodular walk length on tree-structured graphs [193], which motivates our relaxed submodularity assumption on the cost function due to Alkalay-Houlihan and Vetta [194].

A variation on the problem we study is the Orienteering Problem (OP), in which the goal is to maximize a total score collected from visiting vertices on a graph, subject to a travel time constraint [195, 196]. Chekuri and Pal [197] propose a quasi-polynomial time approximation algorithm that yields a logarithmic approximation guarantee for a more general *submodular* objective function. Singh et al. [198] show how this algorithm can be scaled up using *space decomposition* and *branch-and-bound* algorithms, and present results on planning informative paths for multiple robotic sensors. Note that nodes in our routing networks are not directly connected, and therefore traveling cost in each grid (if we decompose the space) cannot be ignored when the grid size is small. This violates

[2] We used the implementation of the Christofides algorithm available at https://github.com/faisal22/Christofides, in which matching is done greedily.

a key assumption of Singh et al. and poses a serious challenge as we attempt to apply their approach to our problem. Nevertheless, we utilize a variant of the recursive greedy algorithm with its running time boosted by a simple heuristic and use it as an alternative in our experiments. Notably, Singh et al. [198] prove that there exists an approximation guarantee when applying any approximation algorithm successively, known as *sequential-allocation*, to solve a multiple robots informative path planning problem. In this chapter, we use this result to derive the approximation guarantee for multi-actor optimization based on the single-actor algorithm we propose.

Perhaps the closest, and most practical, alternative to our algorithm is the framework proposed by Iyer and Bilmes [199]. Specifically, they consider submodular maximization under a submodular cost constraint, and propose several algorithms, including a greedy heuristic (GR) and iterative submodular knapsack (ISK) (their third proposed algorithm, involving ellipsoidal approximation of the submodular cost, scales poorly and we do not consider it). Our approach is a significant extension compared to Iyer and Bilmes [199] and Iyer [200]: we present a new generalized cost-benefit algorithm, and demonstrate approximation guarantees which relax the submodularity assumption on the cost function made by Iyer. The approximation guarantee is non-bicriterion in its form that complements a previous bi-criterion result due to Zhang and Vorobeychik [32]. This generalization is crucial, as routing costs are in general not submodular [193]. Moreover, we demonstrate that our algorithm outperforms that of Iyer and Bilmes [199] in experiments.

5.3 Problem Statement

Let V be a collection of elements and $f : 2^V \to R_{\geq 0}$ a function over subsets of V, and assume that f is monotone increasing. For any set $X \subseteq V$, define $f(j|X) = f(X \cup \{j\}) - f(X)$, that is, the marginal improvement in f if element $j \in V$ is added to a set $X \subseteq V$. Our discussion will concern *submodular functions* f, which we now define.

Definition 5.3.1. *A function* $f : 2^V \to R_{\geq 0}$ *is* submodular *if for all* $S \subseteq T \subseteq V$, $f(j|S) \geq f(j|T)$.

5.3.0.0.1 Single Actor Suppose there is only one actor, whose goal is to find a set $X^* \subseteq V$ which solves the following problem:

$$f(X^*) = \max\{f(X) \mid c(X) \leq B\}, \tag{5.1}$$

where $c : 2^V \to R_{\geq 0}$ is the cost function, which we assume is monotone increasing. As with f, we define $c(j|X) = c(X \cup j) - c(X)$, which denotes the marginal increase in cost when j is added to a set X.

An important motivating setting for this problem is when the cost function represents a least-cost route through a set of vertices X on a graph. Specifically, suppose that $G_R(V,E)$ is a graph in which V are nodes and E edges, and suppose that traversing an edge $e \in E$ incurs a cost c_e, whereas visiting a vertex $v \in V$ incurs a cost c_v. For a given set of nodes $X \subseteq V$, define a cost $c_R(X;s,t)$ as the shortest *walk* in G_R that passes through all nodes in X *at least* once, starting at node s and ending at t. The cost function for X then becomes

$$c(X;s,t) = c_R(X;s,t) + \sum_{x \in X} c_x,$$

that is, the total coverage cost (by a shortest walk through the graph), together with visit cost, for nodes in X, where start and end nodes are exogenously specified (for example, $s = t$ is the robot deployment location).[3]

5.3.0.0.2 Multiple Actor As a natural extension to the single-actor optimization problem, we suppose there are K actors, each of which covers a node set X_k, $k \in 1 \ldots K$, and is subject to its own budget constraint: $c(X_k) \leq B_k$. For simplicity, we hereafter assume all B_k are equal, that is, $\forall k, B_k = B$ for some B. The optimal set of nodes to be covered X^* is the

[3] A node in X may be visited more than once, but the visit cost is calculated only once.

union of coverage assignments X_k^* for each actor k, and solves the following maximization problem:

$$f(X^*) = \max\{f(X) \mid c(X_k) \leq B, \forall k \in 1,\ldots,K; X = X_1 \cup \ldots \cup X_K\}, \tag{5.2}$$

where $c : 2^V \to R_{\geq 0}$ is the monotonically increasing cost function.

5.4 Generalized Cost-Benefit Algorithm

Maximizing submodular functions in general is NP-hard [188]. Moreover, even computing the cost function $c(X)$ is NP-hard in many settings, such as when it involves computing a shortest walk through a subset of vertices on a graph (a variant of the traveling salesman problem). Our combination of two hard problems seems hopeless. We now present a general cost-benefit (GCB) algorithm (Algorithm 3) for computing approximate solutions to Problem (5.1). In the sections that follow we present theoretical guarantees for this algorithm under additional assumptions on the cost function, as well as empirical evidence for the effectiveness of GCB. At the core of the algorithm is the following simple heuristic: in each iteration i, add to a set G an element x_i such that

$$x_i = \arg\max_{x \in V \setminus G_{i-1}} \frac{f(x|G_{i-1})}{c(x|G_{i-1})}, \tag{5.3}$$

where $G_0 = \emptyset$ and $G_i = \{x_1,\ldots,x_i\}$. In Algorithm 3, f also involves an initially covered set C (which may be empty), and the cost function $c(\cdot)$ is parametrized with starting location s and end location t; we omit these complications for ease of exposition.

The greedy algorithm based on Equation (5.3) alone has an unbounded approximation ratio, which was shown for a modular cost function by Khuller et al. [188]. The key modification is to return the better of this solution and the solution produced by a greedy heuristic that ignores the cost altogether. Next, we observed that $c(\cdot)$ may not be computable in polynomial time. We therefore make use of an approximate cost function $\hat{c}(\cdot)$ in its place

which can be computed in polynomial time. The nature of our results will then depend on the quality of this approximation.

Data: s, t, B, C, V
Result: Selection $X \subseteq V$
$G \leftarrow \emptyset$;
$V^{'} \leftarrow V \backslash C$;
$\tilde{x} \leftarrow \arg\max\{f(x|C)|x \in V', \hat{c}_{s,t}(\{x\}|C) \leq B\}$;
while $V^{'} \neq \emptyset$ **do**
  foreach $x \in V'$ **do**
    $\Delta_f^x \leftarrow f(x|G \cup C)$;
    $\Delta_c^x \leftarrow \hat{c}_{s,t}(x|G \cup C)$;
  end
  $x^* \leftarrow \arg\max\{\Delta_f^x / \Delta_c^x | x \in V^{'}\}$;
  if $\hat{c}_{s,t}(G \cup \{x^*\}) \leq B$ **then**
    $G \leftarrow G \cup \{x^*\}$;
    $V^{'} \leftarrow V^{'} \backslash \{x^*\}$;
  else
    break;
end
return $\arg\max_{X \in \{\{\tilde{x}\}, G\}} f(X)$

Algorithm 3: Generalized Cost-benefit Algorithm: $GCB(s,t,B,C,V)$.

Observe that, the greedy solution in Algorithm 3 (i.e., solution ignoring the coverage cost), $\{\tilde{x}\}$ contains only a single element. A variant of this is to continue adding more elements in greedy order (again, ignoring costs) until we violate the budget constraint. Moreover, when the budget constraint in GCB is violated, alternatively, we can continue adding elements that are not *first-best*, until no more elements can be added. We show below that these variations yield the same approximation guarantees. Our implementations, therefore, use these enhancements.

5.5 Theoretical Analysis

A major limitation of prior work on cost-constrained submodular maximization is that the cost function itself was assumed to be submodular. In most practical problems where the cost function is generated by routing problems (such as coverage of nodes on a graph,

or TSP), the optimal cost is not submodular [193]. On the other hand, TSP has submodular special cases, such as when the graph is a tree [193]. Motivated by this observation, we make use of a natural relaxation of cost submodularity, α-submodularity.

Definition 5.5.1. *A cost function c over subsets of V is α-submodular if*

$$\alpha = \min_{x \in V \setminus T} \min_{S,T:S \subset T} \frac{c(x|S)}{c(x|T)}.$$

Clearly, $c(\cdot)$ is submodular *iff* $\alpha \geq 1$. In addition, we introduce a notion for a function c termed *curvature* which essentially measures deviations from linearity [201, 202].

Definition 5.5.2. *For a submodular function c over subsets of V,* curvature $k_c(X)$ *over a subset $X \subseteq V$ is defined as*

$$k_c(X) = 1 - \min_{j \in X} \frac{c(j|X \setminus j)}{c(j)},$$

where $c(j) = c(\{j\})$. We call $k_c \equiv k_c(V)$ the total curvature *of c.*

Below, a related notion of curvature, which we denote by $\hat{k}_c(X)$, will also be useful as a means to streamline the analysis:

$$\hat{k}_c(X) = 1 - \frac{\sum_{j \in V} c(j|X \setminus j)}{\sum_{j \in X} c(j)}.$$

As mentioned earlier, since optimal cost is often infeasible to compute, our GCB algorithm makes use of approximate cost function, $\hat{c}$. We now make this notion of approximation formal: we assume that $\hat{c}(X)$ is a $\psi(n)$-approximation of the optimal cost $c(X)$, where $n = |V|$. In other words, $c(X) \leq \hat{c}(X) \leq \psi(n)c(X), \forall X \subseteq V$. Below we use two algorithms to approximate coverage cost: *nearest neighbor*, which is fast and easy to implement, and has a $\log(n)$-approximation ratio, and the Christofides algorithm [192], which is rather complex but has a $3/2$-approximation guarantee. Finally, we introduce another

useful piece of notation, defining

$$K_c = \max\{|X| : X \subseteq V, c(X) \leq B\},$$

that is, K_c is the size of the largest set $X \subseteq V$ which is feasible for our problem.

Next we prove a novel non-bicriterion approximation guarantee for Algorithm 3, in contrast to our previous bi-criterion guarantee [32].

5.5.1 Building Blocks

Our first step is to connect $\hat{k}_c(X)$, $k_c(X)$, and k_c. The result is given in Lemma 5.5.1.

Lemma 5.5.1. *For any monotone α-submodular function and set $X \subseteq V$,*

$$\hat{k}_c(X) \leq k_c(X) \leq k_c$$

Proof.

$$\begin{aligned} 1 - k_c(X) &= \min_{j \in X} \frac{c(j|X \backslash j)}{c(j)} \\ &\leq \frac{c(j|X \backslash j)}{c(j)}, \forall j \in X \end{aligned} \tag{5.4}$$

Also we note that

$$\begin{aligned} 1 - \hat{k}_c(X) &= \frac{\sum_{j \in X} c(j|X \backslash j)}{\sum_{j \in X} c(j)} \\ &\geq \frac{\sum_{j \in X} (1 - k_c(X)) c(j)}{\sum_{j \in X} c(j)} \\ &= 1 - k_c(X) \end{aligned} \tag{5.5}$$

Thus, $\hat{k}_c(X) \leq k_c(X)$. By monotonicity, it holds that $k_c(X) \leq k_c$, since $X \subseteq V$. □

Next, we generalize the fundamental properties of submodular functions [187] to α-

submodular functions.

Lemma 5.5.2. *For any α-submodular function c, the following statements hold.*

(i). $c(j|S) \geq \alpha c(j|T), \forall S \subseteq T \subseteq V$ *and* $j \in V \backslash T$.

(ii). $c(T) \leq c(S) + \frac{1}{\alpha}\sum_{j\in T\backslash S} c(j|S) - \alpha \sum_{j\in S\backslash T} c(S\cup T\backslash j), \forall S,T \subseteq V$.

(iii). $c(T) \leq c(S) + \frac{1}{\alpha}\sum_{j\in T\backslash S} c(j|S), \forall S \subseteq T \subseteq V$.

(iv). $c(T) \leq c(S) - \alpha \sum_{j\in S\backslash T} c(j|S\backslash j), \forall T \subseteq S \subseteq V$.

Proof. (i) Since $\alpha = \min_{j\in V\backslash T} \min_{S,T:S\subseteq T} \frac{c(j|S)}{c(j|T)}$, we have $\frac{c(j|S)}{c(j|T)} \geq \alpha, \forall j \in V\backslash T$.

(i)$\Rightarrow$(ii). For arbitrary S and T with $T-S = \{j_1,\ldots,j_r\}$ and $S-T = \{k_1,\ldots,k_q\}$ we have

$$\begin{aligned} c(S\cup T) - c(S) &= \sum_{t=1}^{r} [c(S\cup\{j_1,\ldots,j_t\} - c(S\cup\{j_1,\ldots,j_{t-1}\}))] \\ &= \sum_{t=1}^{r} c(j_t|S\cup\{j_1,\ldots,j_{t-1}\}) \\ &\leq \frac{1}{\alpha}\sum_{t=1}^{r} c(j_t|S) \\ &= \frac{1}{\alpha}\sum_{j\in T\backslash S} c(j|S) \end{aligned} \tag{5.6}$$

where the inequality holds due to (i). Similarly, we know

$$\begin{aligned} c(S\cup T) - c(T) &= \sum_{t=1}^{q} [c(T\cup\{k_1,\ldots,k_t\} - c(T\cup\{k_1,\ldots,k_{t-1}\}))] \\ &= \sum_{t=1}^{q} c(k_t|T\cup\{k_1,\ldots,k_t\}\backslash k_t) \\ &\geq \alpha\sum_{t=1}^{q} c(k_t|T\cup S\backslash k_t) \\ &= \alpha\sum_{j\in S\backslash T} c(j|S\cup T\backslash j) \end{aligned} \tag{5.7}$$

By subtracting (5.6) and (5.7), we obtain (ii).

(ii)⇒ (iii). When $S \subseteq T, S \backslash T = \emptyset$, the last term of (ii) vanishes.

(ii)⇒ (iv). When $T \subseteq S, T \backslash S = \emptyset, S \cup T = S$, the second term of (ii) vanishes. □

Finally, based on Lemmas 5.5.1 and 5.5.2, the following holds.

Lemma 5.5.3. *Given a monotone α-submodular function c over subsets of X, it holds that*

$$\sum_{j \in X} c(j) \leq \frac{|X|}{1 + \alpha(|X| - 1)(1 - k_c(X))} c(X)$$

Moreover, it is also the case that,

$$\sum_{j \in X} c(j) \leq \frac{|X|}{1 + \alpha(|X| - 1)(1 - \hat{k}_c(X))} c(X)$$

Proof. It follows from Lemma 5.5.2 (iii) that

$$c(X) - c(x) \geq \alpha \sum_{j \in X \backslash x} c(j | X \backslash j), \forall x \in X$$

Summing over all instance of x, we get

$$
\begin{aligned}
|X|c(X) - \sum_{x\in X} c(x) &\geq \alpha \sum_{x\in X} \sum_{j\in X\setminus x} c(j|X\setminus j) \\
&= \alpha \sum_{x\in X} \left(\sum_{j\in X} c(j|X\setminus j) - c(x|X\setminus x) \right) \\
&= \alpha \left(\sum_{x\in X} \sum_{j\in X} c(j|X\setminus j) - \sum_{x\in X} c(x|X\setminus x) \right) \\
&= \alpha \left(|X| \sum_{j\in X} c(j|X\setminus j) - \sum_{j\in X} c(j|X\setminus j) \right) \\
&= \alpha(|X|-1) \sum_{j\in X} c(j|X\setminus j) \\
&= \alpha(|X|-1)(1-\hat{k}_c(X)) \sum_{j\in X} c(j) \\
&\geq \alpha(|X|-1)(1-k_c(X)) \sum_{j\in X} c(j)
\end{aligned}
$$

where the last equality holds due to definition of curvature $\hat{k}_c$ and last inequality follows from Lemma 5.5.1, as $1-\hat{k}_c(X) \geq 1-k_c(X) \geq 0$. The result follows after rearranging the terms. □

5.5.2 Proof of the Approximation Ratio

Suppose the GCB algorithm starts with an empty set $G_0 = \emptyset$, and keeps adding nodes the set by the greedy rule (Equation 5.3). It then generates a sequence of intermediate feasible sets, $G_1, \ldots, G_l$, until it violates the budget constraint in iteration $l+1$ with a set G_{l+1}.

Lemma 5.5.4. *For i=1, . . . , l+1, it holds that*

$$
f(G_i) - f(G_{i-1}) \geq \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB} (f(X^*) - f(G_{i-1}))
$$

where $\hat{c}$ is an α-submodular $\psi(n)$-approximation of the α-submodular function c, X^* is

the optimal solution of $\max\{f(X)|c(X) \leq B\}$ *and* $M = \frac{\psi(n)}{\alpha}\frac{K_c}{1+\alpha(K_c-1)(1-k_c)}$.

Proof. We have

$$\begin{aligned} f(X^*) - f(G_{i-1}) &\leq f(X^* \cup G_{i-1}) - f(G_{i-1}) \\ &= f(X^* \backslash G_{i-1} \cup G_{i-1}) - f(G_{i-1}) \end{aligned}$$

Assume that $X^* \backslash G_{i-1} = \{Y_1, \ldots, Y_m\}$ and let

$$Z_j = f(G_{i-1} \cup \{Y_1, \ldots, Y_j\} - f(G_{i-1} \cup \{Y_1, \ldots, Y_{j-1}\}, \forall j = 1, \ldots, m$$

then we have

$$\begin{aligned} \frac{Z_j}{\hat{c}(G_{i-1} \cup Y_j) - \hat{c}(G_{i-1})} &\leq \frac{f(G_{i-1} \cup Y_j) - f(G_{i-1})}{\hat{c}(G_{i-1} \cup Y_j) - \hat{c}(G_{i-1})} \\ &\leq \frac{f(G_i) - f(G_{i-1})}{\hat{c}(G_i) - \hat{c}(G_{i-1})} \end{aligned}$$

where first inequality holds due to submodularity and second inequality holds due to the greedy rule. Further, we know that

$$f(X^*) - f(G_{i-1}) \leq \sum_{j=1}^{m} Z_j \leq \sum_{j=1}^{m} [\hat{c}(G_{i-1} \cup Y_j) - \hat{c}(G_{i-1})] \frac{f(G_i) - f(G_{i-1})}{\hat{c}(G_i) - \hat{c}(G_{i-1})}$$

From α-submodularity of $\hat{c}$, by Lemma 5.5.2 (i), and the fact that $\hat{c}$ $\psi(n)$ approximates c, we have

$$\begin{aligned} \sum_{j=1}^{m} [\hat{c}(G_{i-1} \cup Y_j) - \hat{c}(G_{i-1})] &\leq \frac{1}{\alpha} \sum_{j=1}^{m} [\hat{c}(Y_j) - \hat{c}(\emptyset)] \\ &\leq \frac{1}{\alpha} \sum_{j \in X^*} \hat{c}(Y_j) \\ &\leq \frac{\psi(n)}{\alpha} \sum_{j \in X^*} c(Y_j) \end{aligned}$$

Since c is α-submodular, by Lemma 5.5.3, we know that $\sum_{j\in X^*} c(Y_j) \leq \frac{|X^*|}{1+\alpha(|X^*|-1)(1-k_c(X^*))} c(X^*)$. As $K_c \geq |X^*|$, $k_c \geq k_c(X^*)$ and $c(X^*) \leq B$, it follows that

$$\begin{aligned} f(X^*) - f(G_{i-1}) &\leq \frac{\psi(n)}{\alpha} \frac{|X^*|}{1+\alpha(|X^*|-1)(1-k_c(X^*))} c(X^*) \frac{f(G_i) - f(G_{i-1})}{\hat{c}(G_i) - \hat{c}(G_{i-1})} \\ &\leq \frac{\psi(n)}{\alpha} \frac{K_c}{1+\alpha(K_c-1)(1-k_c)} c(X^*) \frac{f(G_i) - f(G_{i-1})}{\hat{c}(G_i) - \hat{c}(G_{i-1})} \\ &\leq MB \frac{f(G_i) - f(G_{i-1})}{\hat{c}(G_i) - \hat{c}(G_{i-1})} \end{aligned}$$

where $M = \frac{\psi(n)}{\alpha} \frac{K_c}{1+\alpha(K_c-1)(1-k_c)} \geq 1$. □

Lemma 5.5.5. *For $i = 1, \ldots, l+1$ it holds that*

$$f(G_i) \geq \left[1 - \prod_{k=1}^{i} (1 - \frac{\hat{c}(G_k) - \hat{c}(G_{k-1})}{MB}) \right] f(X^*)$$

where $\hat{c}$ is an α-submodular $\psi(n)$-approximation of an α-submodular function c, X^* is the optimal solution of $\max\{f(X)|c(X) \leq B\}$ *and* $M = \frac{\psi(n)}{\alpha} \frac{K_c}{1+\alpha(K_c-1)(1-k_c)}$.

Proof. We prove this by induction. When $i = 1$, from Lemma 5.5.4, we know that

$$f(G_1) \geq \frac{\hat{c}(G_1) - \hat{c}(G_0)}{MB} f(X^*)$$

For $i > 1$, we have

$$\begin{aligned}
f(G_i) &= f(G_{i-1}) + [f(G_i) - f(G_{i-1})] \\
&\geq f(G_{i-1}) + \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB}(f(X^*) - f(G_{i-1})) \\
&= \left(1 - \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB}\right) f(G_{i-1}) + \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB} f(X^*) \\
&\geq \left(1 - \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB}\right) \left[\left(1 - \prod_{k=1}^{i-1}(1 - \frac{\hat{c}(G_k) - \hat{c}(G_{k-1})}{MB})\right) f(X^*)\right] \\
&+ \frac{\hat{c}(G_i) - \hat{c}(G_{i-1})}{MB} f(X^*) \\
&= \left(1 - \prod_{k=1}^{i}(1 - \frac{\hat{c}(G_k) - \hat{c}(G_{k-1})}{MB})\right) f(X^*)
\end{aligned}$$

□

Theorem 5.5.6. *The GCB algorithm obtains a set X such that*

$$f(X) \geq \frac{1}{2}(1 - e^{-\frac{1}{M}}) f(X^*),$$

where X^* is the optimal solution of $\max\{f(X) | c(X) \leq B\}$*, $\hat{c}$ is an α-submodular $\psi(n)$-approximation of an α-submodular function c, and* $M = \frac{\psi(n)}{\alpha} \frac{K_c}{1+\alpha(K_c-1)(1-k_c)}$.

Proof. It follows from Lemma 5.5.5 that

$$\begin{aligned}
f(G_{l+1}) &\geq \left[1 - \prod_{k=1}^{l+1}\left(1 - \frac{\hat{c}(G_k) - \hat{c}(G_{k-1})}{MB}\right)\right] f(X^*) \\
&\geq \left[1 - \prod_{k=1}^{l+1} e^{-\frac{\hat{c}(G_k) - \hat{c}(G_{k-1})}{MB}}\right] f(X^*) \\
&= \left[1 - e^{-\frac{1}{MB}\sum_{k=1}^{l+1} \hat{c}(G_k) - \hat{c}(G_{k-1})}\right] f(X^*) \\
&= \left[1 - e^{-\frac{\hat{c}(G_{l+1})}{MB}}\right] f(X^*) \\
&\geq \left[1 - e^{-\frac{1}{M}}\right] f(X^*)
\end{aligned}$$

where the second inequality is due to the fact that $1-x \leq e^{-x}$, and the last inequality holds as $\hat{c}(G_{l+1}) > B$ (notice that G_{l+1} violates budget constraint). Moreover, by submodularity, we note that $f(G_{l+1}) - f(G_l) \leq f(\{x_{l+1}\}) \leq f(\{\tilde{x}\})$, where $\tilde{x} = \arg\max_{x \in V} \{f(\{x\}) | \hat{c}(\{x\}) \leq B\}$. Therefore,

$$f(G_l) + f(\{\tilde{x}\}) \geq f(G_{l+1}) \geq (1 - e^{-\frac{1}{M}}) f(X^*)$$

and

$$\max\{f(G_l), f(\{\tilde{x}\})\} \geq \frac{1}{2}(1 - e^{-\frac{1}{M}}) f(X^*)$$

□

Having established a general approximation ratio for GCB, note that an exactly submodular cost function emerges as a special case with $\alpha = 1$, as does exact cost function when $\psi(n) = 1$, with the bound becoming tighter in both instances. Moreover, the bound becomes tighter as we approach modularity.

5.5.3 Multi-Actor Optimization

Multi-actor submodular maximization (Problem (5.2)) is a natural and important generalization of the single-actor optimization (Problem (5.1)). Previous work on orienteering and sensor placement shows that by sequentially applying an approximation algorithm for a single actor, for either modular/additive [203] or submodular [198] objective function, an approximation guarantee can be obtained. This provides us with a convenient way to construct an effective solution based on our GCB algorithm to tackle the multi-actor optimization problem.

5.5.3.1 Sequential Planning

The sequential planning algorithm successively applies a single actor submodular optimization algorithm multiple times, where each actor optimizes its actions given an initially-covered set resulting from all previous actors. Algorithm 4 demonstrates this using GCB

(Algorithm 3) for each single-actor subproblem, although any single-actor algorithm can be used.

Data: Number of actors K, pairs of starting and ending nodes $\{(s_i,t_i)\}_{i=1,\ldots,K}$, budget B, initial covered set C, set of all nodes V
Result: Selection $X \subseteq V$
$X_0 \leftarrow C$;
foreach $1 \leq i \leq k$ **do**
    $X_i \leftarrow GCB(s_i,t_i,B,X_{i-1},V)$;
    $X_i \leftarrow X_{i-1} \cup X_i$;
end
return X_i

Algorithm 4: Sequential planning algorithm using GCB as the single-actor optimization subroutine.

5.5.3.2 Approximation Guarantee

A general result that establishes the approximation guarantee of multi-actor optimization problem is due to Singh et al.:

Theorem 5.5.7. *[198] The sequential planning strategy will achieves an approximation guarantee of* $(1+\eta)$ *for the multi-actor submodular optimization problem, where* η *is the approximation ratio of the single-actor algorithm. Particularly, when all actor have same start nodes, i.e.,* $s_i = s_j$, $\forall i, j$ *and end nodes, i.e.,* $t_i = t_j$, $\forall i, j$, *the approximation guarantee improves to* $1/(1-\exp(-1/\eta)) \leq 1+\eta$.

In particular, when we sequentially apply GCB algorithm, the approximation ratio η becomes $\frac{2}{1-\exp(-1/M)}$. The next result then follows directly:

Corollary 5.5.8. *The sequential planning strategy with GCB algorithm will achieve an approximation guarantee of* $\frac{3-\exp(-1/M)}{1-\exp(-1/M)}$ *for the multi-actor submodular optimization problem. In the special case when all actors have the same start and end nodes, the approximation guarantee improves to* $1/(1-\exp(-\frac{1-\exp(-1/M)}{2}))$.

5.6 Applications

We apply the GCB algorithm to two important problems: mobile robotic sensing and door-to-door solar marketing, using both real and simulated networks. We show that the proposed GCB algorithm typically outperforms state-of-the-art alternatives, *particularly when routing problems do not yield a submodular optimal cost function*. Since both applications involve cost constraints arising from routing decisions, our GCB algorithm uses two polynomial-time algorithms for approximating the least-cost walk: the nearest neighbor heuristic [191] (GCB-nn) and the Christofides algorithm [192] (GCB-ct). We compare the two resulting GCB implementations to three state-of-the-art algorithms: the modified recursive greedy (RG) (see Algorithm 5 in Appendix 5.7), simple greedy (GR) [188], and iterative submodular knapsack (ISK) [199].[4]

All experiments were performed on an Ubuntu Linux 64-bit PC with 32 GB RAM and an 8-core Intel Xeon 2.1 GHz CPU. Each experiment used a single thread.

5.6.1 Case Study 1: Mobile Robotic Sensing

5.6.1.0.1 Single Robot We start with the following single-robot sensing problem. A mobile robot equipped with sensors and aims to optimally choose a subset of locations in 2-D space to make measurements so as to minimize the uncertainty about the unmeasured locations. Equivalently, the objective is to maximize entropy of selected locations [22]. We suppose that the robot faces two kinds of costs (e.g., reflecting battery life or time): costs of moving between a pair of neighboring locations, and costs of making measurements at a particular location, and must begin and end a walk at an exogenously specified location s. To reflect movement costs, we generated a hypothetical tree-structure routing network for the mobile robot (Figure 5.1), which corresponds to a minimum spanning tree based on the distance matrix of the original sensor locations.

[4] RG is not included in door-to-door marketing experiments because it does not scale to these problem instances.

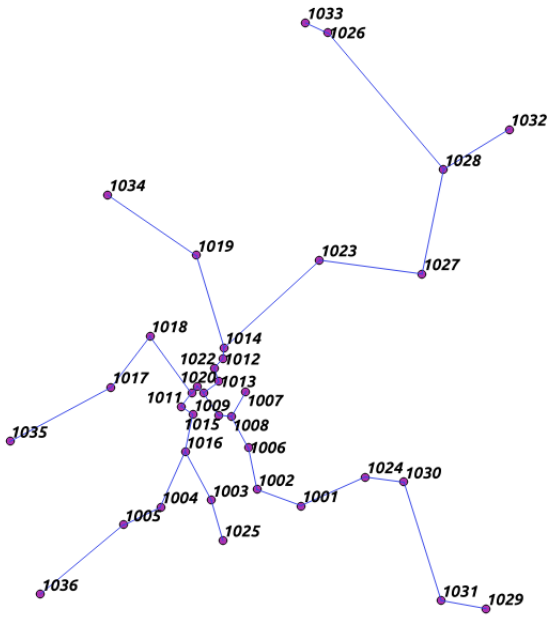


Figure 5.1: Mobile Robot Routing Network

Let A be the set of selected locations to be visited, and $H(A) = -\int p(\boldsymbol{x}_A)\log p(\boldsymbol{x}_A)d\boldsymbol{x}_A$ be the corresponding entropy, where $\boldsymbol{x}_A$ is the vector of random variables corresponding to A. H is known to be monotone increasing and submodular [22]. For any order of the elements in A, denote by A_i the set of locations that contains the first i elements, i.e., $A_i = \{x_1, \ldots, x_i\}$. Then, by the chain-rule of entropies, $H(A_i)$ can be computed by

$$H(A_i) = H(x_i|A_{i-1}) + \ldots + H(x_2|A_1) + H(x_1|A_0),$$

where the marginal benefit of adding x to A_i is $H(x|A_i) = \frac{1}{2}\log(2\pi e\sigma^2_{x|A_i})$, with $\sigma^2_{x|A_i} = \sigma^2_x - \Sigma_{xA_i}\Sigma^{-1}_{A_iA_i}\Sigma_{A_ix}$. In this expression, σ^2_x is the variance at location x, Σ_{xA_i} is a vector of covariances $Cov(x,u), \forall u \in A_i$, $\Sigma_{A_iA_i}$ the covariance submatrix corresponding to measurements A_i, and Σ_{A_ix} is the transpose of Σ_{xA_i}. The routing cost of the mobile robot for a set of locations A to be visited is $c(A) = c_R(A) + \sum_{i\in A} c_i$, where c_i is the cost of making a measurement at location i, and $c_R(A)$ is the cost of the shortest walk covering all locations in A. Formally, our single-robot mobile sensing problem solves for an optimal A^*, such that

$$H(A^*) = \max\{H(A) \mid c(A) \leq B\},$$

To evaluate performance of GCB in this application, we use sensor data representing air quality measurements for 36 air quality monitoring stations in Beijing, China [204], where we limit attention to temperature. We estimate the covariance matrix Σ used in computing entropy by fitting a multivariate Gaussian distribution to this data. As discussed above, we approximate the cost of the shortest walk $c_R(A)$ in two ways: using the nearest neighbor heuristic (GCB-nn) and using the Christofides algorithm (GCB-ct).

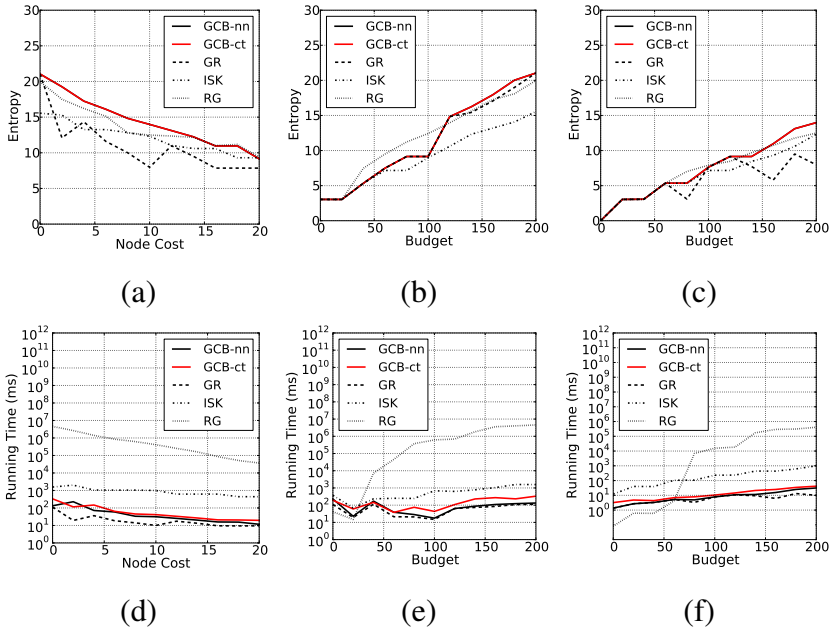


Figure 5.2: Entropy (a)-(c) & run time (d)-(f) comparison among algorithms for single robot mobile sensing scenario. (a), (d) As a function of visit (sensing) cost, fixing budget at 200. (b), (e) As a function of budget, fixing visit cost = 0. (c), (f) As a function of budget, fixing visit cost = 10.

Figure 5.2 shows the results of entropy (performance) and running time comparisons of GCB-nn and GCB-ct with three previous approaches (RG, GR and ISK). (In this figure, GCB-nn lines overlap with GCB-ct and are therefore not directly visible; the same issue obtains in Figure 5.3 below.) GCB-nn and GCB-ct achieve almost the same entropy, and while GCB-ct is slower than GCB-nn, it is considerably faster than other alternatives. GCB-nn and GCB-ct both nearly always outperform the other three in terms of entropy, often by

a large margin; GR is particularly weak in most comparisons. Moreover, the two GCB implementations have competitive running time with GR, and scale far better than ISK and RG. In some cases (see Figure 5.2(b)), RG yields higher entropy than GCB. However, its significantly higher running time (see Figure 5.2(e)), which is several orders of magnitude slower than all other algorithms, makes it impractical.

5.6.1.0.2 Multiple Robots As an extension of the single robot sensing problem, the multi-robot sensing problem solves $H(A^*) = \max\{H(A) \mid c(A_k) \leq B, \forall k \in 1,\ldots,K; A = A_1 \cup \ldots \cup A_K\}$, where each robot is subject to a budget B, and the goal is to assign a set of locations to each robot so that the entropy of the collective locations are maximized, subject to individual robot constraints on routing and measurement costs.

All algorithms in the single robot experiments are combined with the sequential planning strategy to solve multi-robot mobile sensor placement problem. Figure 5.3 shows the results of entropy and run time of multi-robot sensing optimization for three robots (i.e., $K = 3$). The GCB-based algorithms achieve the highest entropy in almost all cases. Both have nearly the same entropy, but GCB-nn is faster, and both scale better than ISK and RG. ISK seems the weakest one in terms of entropy, although it is several orders of magnitude faster than RG. RG can achieve comparable entropy to GCBs but is orders of magnitude slower.

5.6.2 Case Study 2: Door-to-door Marketing of Rooftop Solar Photovoltaic Systems

Our second application is to the door-to-door marketing problem, which we cast as social influence maximization [124] with routing constraints. We formalize this problem by considering two interdependent networks: the social influence network, which captures the influence adopters (of a product) have on the likelihood that others adopt, and the routing network, which represents routes taken by a marketer to visit households of choice. To our knowledge, ours is the first formal treatment of door-to-door marketing problem in a

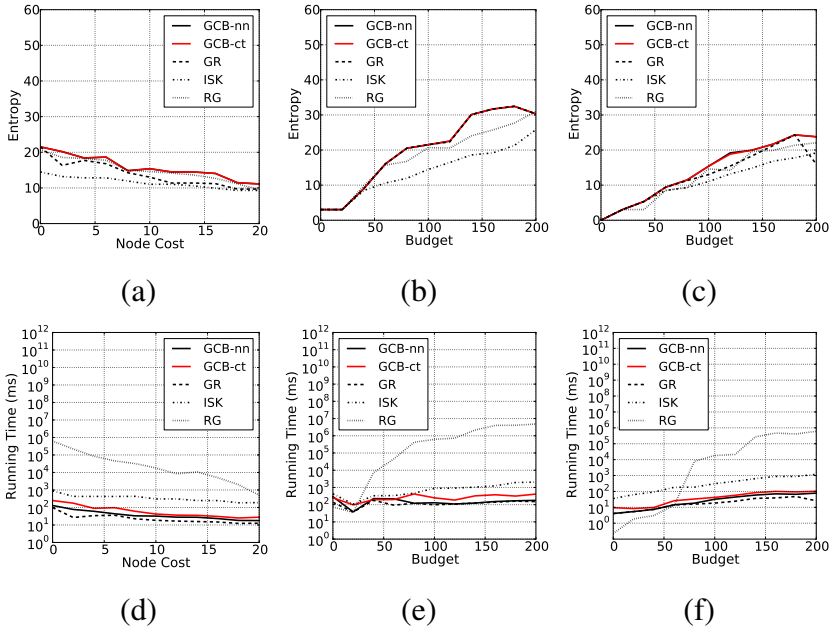


Figure 5.3: Entropy (a)-(c) & run time (d)-(f) comparison among algorithms for multiple robots sensing scenario. (a), (d) As a function of visit (sensing) cost, fixing budget at 100 for each robot. (b), (e) As a function of budget, fixing visit cost = 0. (c), (f) As a function of budget, fixing visit cost = 10.

constrained submodular optimization framework.

5.6.2.0.3 Single Marketer In the single-marketer problem, the goal is to choose a subset of individuals to maximize total adoption of the marketed product in a population as mediated by social influence. We model social influence as a diffusion process on a *social influence network* $G_S = (N, E)$, where N is the set of households (potential adopters) and E a directed graph with each edge $(i, j) \in E$ representing influence of node i on node j. We model diffusion of social influence using the well-known *independent cascade* (IC) model [124]. In the IC model, each neighbor j of an adopter i (corresponding to an edge $(i, j) \in E$) is independently influenced to adopt with probability p_{ij}. The diffusion process begins with a subset of nodes A initially visited by the marketing agent, and this adoption process iteratively repeats until no new nodes adopt. The expected final number of adopters after this process terminates is denoted by $I(A)$; we term this quantity *influence*

(of households A). Kempe et al. [124] showed that the function $I(A)$ is monotone increasing and submodular. In our experiments below, we used $p_{ij} = 0.1$ for all network edges (i, j). Since the IC model is stochastic in nature, we run it 1000 times and use a sample average to estimate $I(A)$.

Unique to a *door-to-door* marketing problem is the nature of costs which constrain which subsets of households A can be feasibly visited by the marketer. Specifically, we posit the existence of a routing network G_R of which household nodes N are a subset, and edges correspond to feasible routes, with costs c_{ij} of traversing an edge (i, j) corresponding, for example, to time it takes to use the associated route. We assume that the marketer faces a budget constraint B on the total routing and visit costs, stemming, for example, from constraints on normal working hours. Thus, the costs of a walk which visits all nodes in A is $c(A) = c_R(A) + \sum_{i \in A} c_i$, where c_i is the cost of visiting a household i, and $c_R(A)$ is the cost of the shortest walk covering all households in A. Formally, then, the marketer aims to choose a subset of households A^* solving

$$I(A^*) = \max\{I(A) \mid c(A) \leq B\}.$$

5.6.2.0.4 Multiple Marketers In the case of K marketers, we aim to maximize the total number of adopters induced by the union of households visited by all marketers. Formally, we solve

$$I(X^*) = \max\{I(A) \mid c(A_k) \leq B, \forall k \in 1 \ldots K; A = A_1 \cup \ldots \cup A_K\},$$

where each marketer k is subject to an identical budget constraint B.

5.6.2.1 Adoption of Visible Technology

We instantiate the door-to-door marketing problem in the context of marketing rooftop solar photovoltaics (PV). For solar PV, an important medium for social influence is its

visibility: specifically, Bollinger and Gillingham [157] showed that the number of solar systems installed in a neighborhood (more specifically, zip code) significantly impacts a household's likelihood to adopt it. Similarly, Zhang et al. [19] confirmed the significance of geographic proximity in determining the probability of solar PV adoption. Using these insights, we generate a social influence network based on a household location dataset for San Diego county, CA, inducing this network based on proximity as measured by a 165 foot radius, giving rise to the influence network shown in Figure 5.4 (top). Figure 5.4 (bottom)

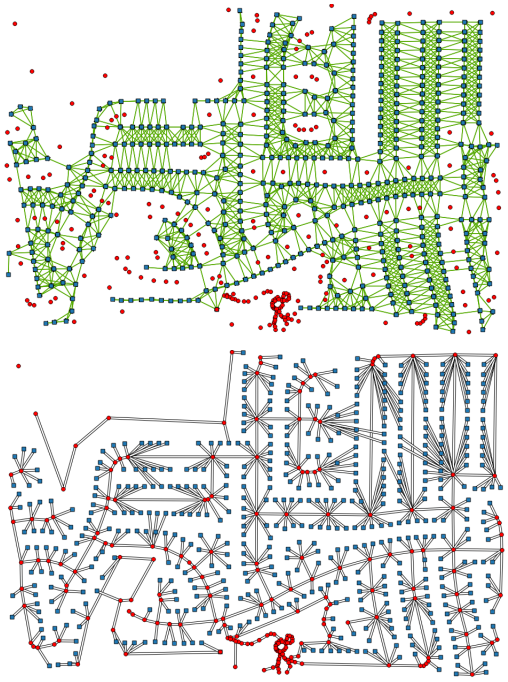


Figure 5.4: Top: social influence network arising from geographic proximity. Bottom: corresponding routing network.

shows the corresponding routing network obtained from OpenStreetMap website, where red dots are way points or intersections in the road networks. Each house is connected to its nearest way point, which finally gives rise to the routing network. The costs of edges in the routing network correspond to physical distance.

5.6.2.1.1 Single Marketer Figure 5.5 shows the results of comparing our GCB algorithms to GR and ISK, both in terms of achieved average influence and running time. In

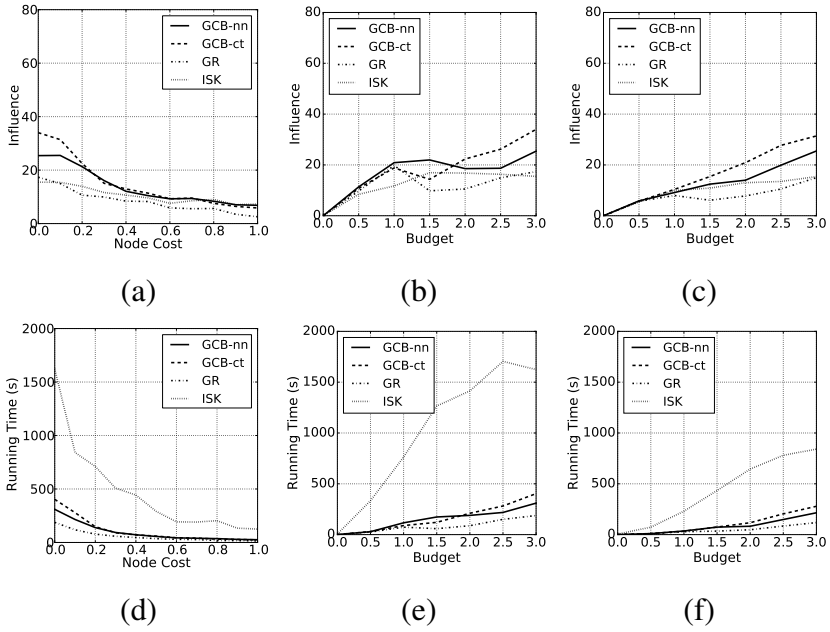


Figure 5.5: Influence σ (a)-(c) & run time (d)-(f) comparison among algorithms for single marketer door-to-door marketing scenario with visible technology. (a), (d) As a function of visit (sensing) cost, fixing budget at 3. (b), (e) As a function of budget, fixing visit cost = 0. (c), (f) As a function of budget, fixing visit cost = 0.1.

all cases, we can observe that GCBs outperforms the others on both measures, often by a substantial margin. Particularly striking is the running time comparison with ISK, where the difference can be several orders of magnitude. In addition, different from our finding in the mobile robotic sensing application, where both GCB algorithms achieve nearly the same objective value, here GCB-ct achieves significantly higher influence than GCB-nn, especially when the budget constraints allow for a coverage of a large set of nodes, for example, a small visit cost for a fixed budget in Figure 5.5 (a), or a large budget for a fixed visit cost, as in Figures 5.5 (b) and (c).

5.6.2.1.2 Multiple Marketers Figure 5.6 shows achieved average influence and run time comparison of GCB algorithms to GR and ISK solving multi-marketer influence maximization for three marketers ($K = 3$). GCB algorithms remain the most efficient among all candidates, achieving the highest influence and scaling better than the state-of-art ISK.

Similarly, we observe that GCB-ct outperforms GCB-nn in terms of average influence in cases of lower visit cost and larger budget, although it is slower than GCB-nn.

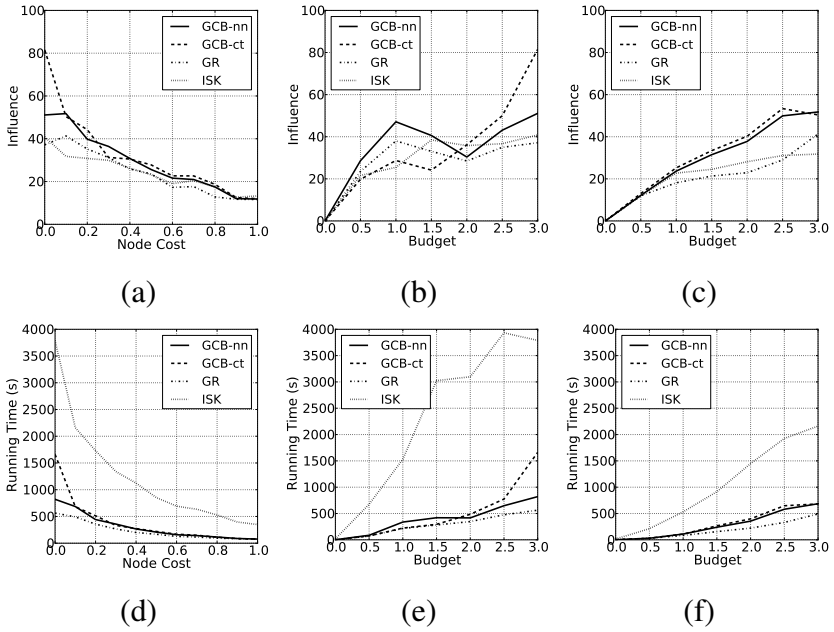


Figure 5.6: Influence σ (a)-(c) & run time (d)-(f) comparison among algorithms for multi-marketer door-to-door marketing scenario with visible technology. (a), (d) As a function of visit (sensing) cost, fixing budget at 3 for each marketer. (b), (e) As a function of budget, fixing visit cost = 0. (c), (f) As a function of budget, fixing visit cost = 0.1.

5.6.2.2 Experiments on Random Graph

Our second experimental investigation in the context of door-to-door marketing problems involves random graph models for both social influence propagation and routing. In particular, we use the well-known Barabasi-Albert (BA) model [205] to generate a random social network (a natural choice, since the BA model has been shown to exhibit a scale-free degree distribution, which is a commonly observed feature of real social networks), and the Erdos-Renyi (ER) model to generate the routing network [206]. The BA model is an iterative generative model which starts with a small seeded network (e.g., a triangle), and adds a node one at a time, connecting it to m vertices, with each edge using the new node as a source generated with probability proportional to the target node's degree. The ER model

is the simplest generative model of networks, where each edge is added to the network with a fixed probability p.

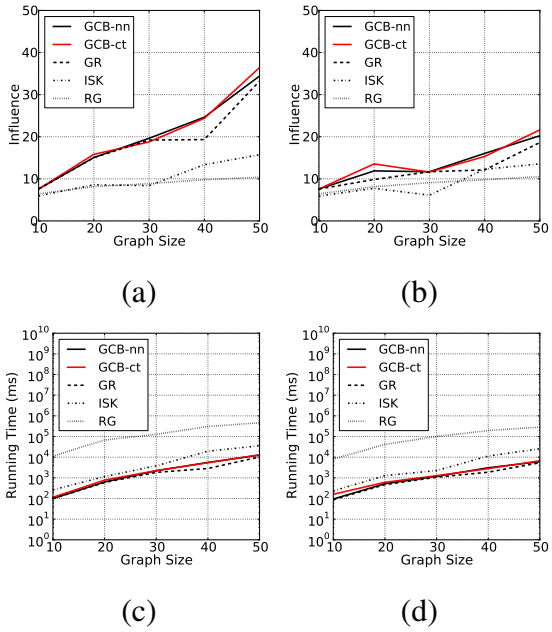


Figure 5.7: Entropy (a)-(b) & run time (c)-(d) comparison among algorithms for door-to-door marketing scenario over different sizes of random graph (p=0.17 in the ER model). (a), (c) As a function of graph size, fixing budget at 10 and visit cost at 0. (b), (d) As a function of budget, fixing budget at 10 and visit cost = 0.2.

Our first set of experiments investigates the scalability of all candidate algorithms. In the implementation, we generated BA models with size of 10, 20, 30 and 50, each adding $m = 2$ edges in every iteration. We only consider an ER model with $p = 0.17$.[5] Figure 5.7 shows the comparison of achieved influence and run time for the single actor influence maximization problem. The major takeaway is that RG is not a scalable solution as graph size increases, running roughly 100 times slower than the others.

Our second set of experiment compares two GCB algorithms with GR and ISK, excluding the non-scalable RG algorithm. In our implementation, we generated a BA social network of over 200 nodes, adding $m = 2$ edges in each iteration. Here, we considered ER models with $p \in \{0.01, 0.02, 0.03\}$. To generate routing costs, we randomly assigned

[5]Our experiments show that ER graph with this probability typically yields networks which are mostly connected, with only occasional isolated nodes.

coordinates for the 200 nodes in 2-D space, and use the Euclidean distance between nodes as a proximity measure.

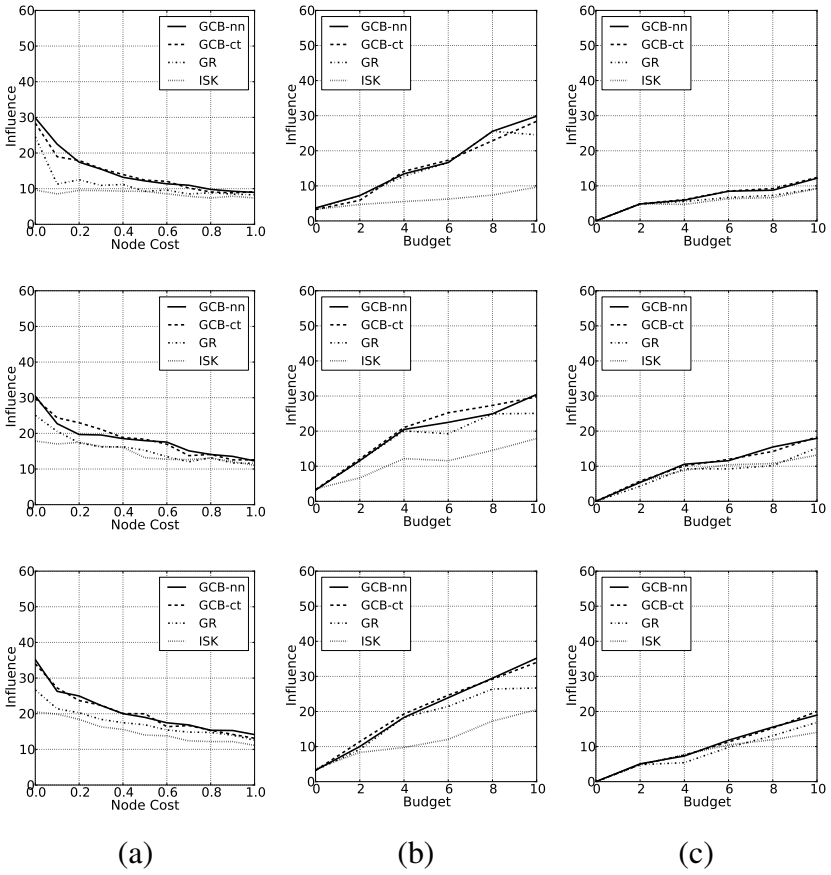


Figure 5.8: Influence σ comparison among algorithms for single marketer door-to-door marketing scenario on random graphs. Top row: $p = 0.01$ in the ER model. Middle row: $p = 0.02$. Bottom row: $p = 0.03$. (a) As a function of visit (sensing) cost, fixing budget at 10. (b) As a function of budget, fixing visit cost = 0. (c) As a function of budget, fixing visit cost = 0.5.

5.6.2.2.1 Single Marketer Figures 5.8 and 5.9 show the results for the random graph experiments of a single marketer, which are consistent with the observations so far: GCB-nn and GCB-ct tend to outperform alternative algorithmic approaches both in terms of objective value (influence, in this case), and in terms of running time (they are comparable to GR, and much faster than ISK). In all cases, varying either visit costs with a fixed budget, or varying the budget for a fixed visit cost, ISK turns out to be the worst in terms of achieved

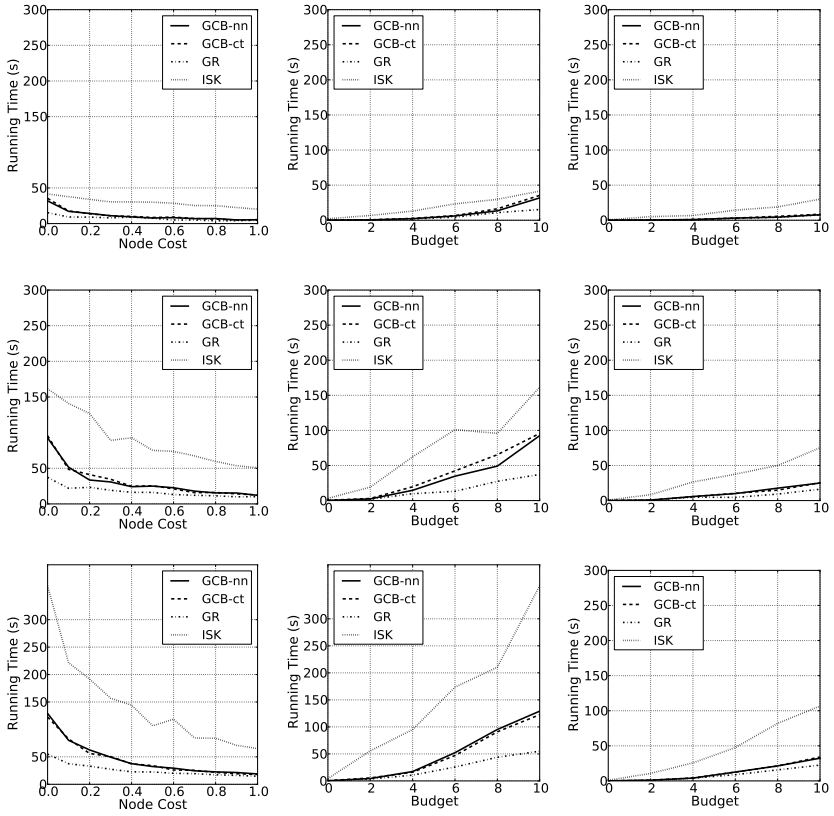
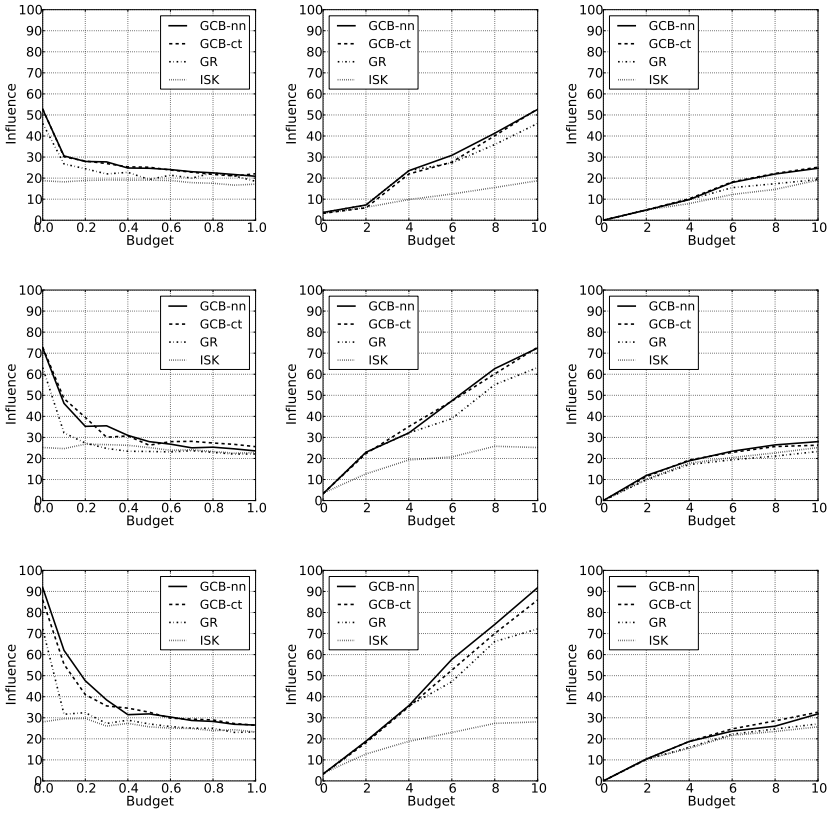


Figure 5.9: Run time comparison among algorithms for single marketer door-to-door marketing scenario on random graphs. Top row: $p = 0.01$ in the ER model. Middle row: $p = 0.02$. Bottom row: $p = 0.03$. (a) As a function of visit (sensing) cost, fixing budget at 10. (b) As a function of budget, fixing visit cost = 0. (c) As a function of budget, fixing visit cost = 0.5.

influence. Interestingly, we do not observe significant difference between GCB-nn and GCB-ct on both measures. We can also note that as p increases (and the routing network becomes denser), the running time of ISK increases rather dramatically, whereas both GCB and GR remain quite scalable.

5.6.2.2.2 Multiple Marketers Figures 5.10 and 5.11 show results on multi-marketer influence maximization on random graphs. Clearly, the two GCB algorithms both outperform all other alternatives on both achieved influence and run time, although GCB-nn and GCB-ct perform similarly. Again, ISK tends to be significantly outperformed on both



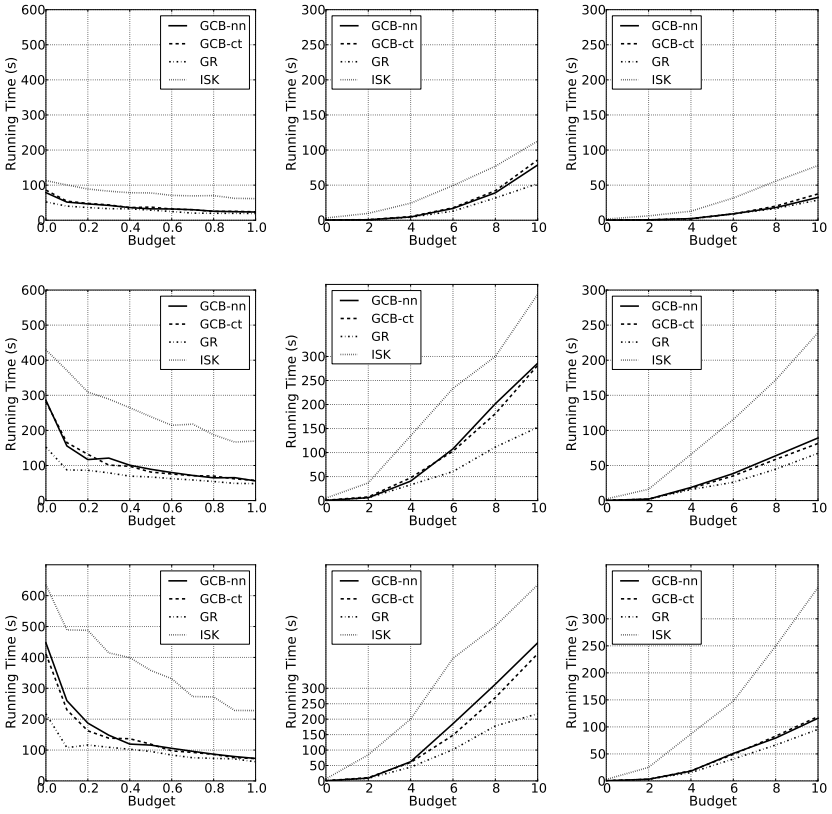
(a) (b) (c)

Figure 5.10: Influence σ comparison among algorithms for multi-marketer door-to-door marketing scenario on random graphs. Top row: $p = 0.01$ in the ER model. Middle row: $p = 0.02$. Bottom row: $p = 0.03$. (a) As a function of visit (sensing) cost, fixing budget at 10 for each agent. (b) As a function of budget, fixing visit cost = 0. (c) As a function of budget, fixing visit cost = 0.5.

measures.

5.7 Conclusion

We considered a very general class of problems in which a monotone increasing submodular function is maximized subject to a general cost constraint for both single-actor and multi-actor scenarios. This problem is motivated by two very different applications: one is mobile robotic sensing, in which a robot moves through an environment with the goal of making select sensor measurements in order to maximize the entropy of selected measure-



(a) (b) (c)

Figure 5.11: Run time comparison among algorithms for multi-marketer door-to-door marketing scenario on random graphs. Top row: $p = 0.01$ in the ER model. Middle row: $p = 0.02$. Bottom row: $p = 0.03$. (a) As a function of visit (sensing) cost, fixing budget at 10 for each agent. (b) As a function of budget, fixing visit cost = 0. (c) As a function of budget, fixing visit cost = 0.5.

ment sites, and another in door-to-door marketing. In both of these applications, the cost constraints arise from routing costs, as well as costs to visit nodes (e.g., to make sensor measurements or to make a marketing pitch). Our algorithmic contribution was a novel generalized cost-benefit algorithm, for which we showed approximation guarantees with a relaxed notion of cost submodularity as well as allowing optimal cost to be only approximately computed. Furthermore, this algorithm can be used to construct another efficient algorithm with provable approximation guarantees, making use of a sequential planning approach, to solve the multi-actor optimization problem. Through an extensive experimental evaluation on both real and synthetic graphs we showed that our approach, implemented

with two different cost approximation methods, significantly outperforms state-of-the-art alternatives in terms of objective value achieved, running time, or both.

APPENDIX: Recursive Greedy Algorithm

In the related literature investigating the so-called *Orienteering Problem*, a *recursive greedy* algorithm has been proposed for the single-actor submodular maximization problem (Problem in 5.1) [197]. Singh et al. [198] subsequently adapted this algorithm to solve the multiple-robot informative path planning problem. While somewhat older than some of the other state-of-the-art approaches, we make it one of the baselines for our experimental evaluation. However, the original algorithm does not account for visit cost, making it inapplicable to our target applications as is. Specifically, if we decompose the routing space like they do, nodes in some grids may not be locally connected, which violates their assumption that *the coverage cost within a grid can be ignored when the grid is small*. To address these challenges as well as provide a reasonable baseline for our GCB algorithm, we propose a *modified* recursive greedy algorithm (see Algorithm 5), which not only handles visit cost but also embodies a few novel ideas that allow us to improve its running time.

Like the original recursive greedy algorithm, Algorithm 5 recursively divides the path into two sub-paths, P_1 and P_2, and utilizes P_1 in a greedy fashion. It includes the following new features:

1. $c(s,t) = l(s,t) + c(t)$: the original recursive greedy algorithm considers only the length of the path $s-t$ (from s to t), $l(s,t)$, which we extend by including the visit cost of the end node $c(t)$.

2. $\tilde{V}$: the original recursive greedy algorithm examines all nodes in V on the graph, which is inefficient, since many middle nodes might not be beneficial (since they have lower marginal utilities). Our strategy (denoted by $topKnodes(V,s,t)$) is to choose a node from only the *top-k* most beneficial nodes determined by the GCB

Data: s, t, B, C, V, i
Result: Path P
if $c(s,t) > B$ **then return** infeasible;
$P \leftarrow s - t$;
if $i = 0$ **then return** P;
$m \leftarrow f(V(P)|C)$;
$\tilde{V} \leftarrow topKnodes(V,s,t)$;
foreach $v \in \tilde{V}$ **do**
  $\tilde{B} \leftarrow expSplit(B)$;
  foreach $B_1 \in \tilde{B}$ **do**
    $P_1 \leftarrow RG(s,v,B_1,C,i-1)$;
    $P_2 \leftarrow RG(v,t,B-B_1,C \cup V(P_1),i-1)$;
    if $f(V(P_1 \cdot P_2)|C) > m$ **then**
      $P \leftarrow P_1 \cdot P_2$;
      $m \leftarrow f(V(P)|C)$;
    end
  end
end
return P

Algorithm 5: Modified Recursive Greedy Algorithm: $RG(s,t,B,C,V,i)$

heuristic (in Section 5.4).

3. $\tilde{B}$: The original recursive greedy algorithm uses a linear scan of integers from 0 to B, which scales poorly for large budgets. Motivated by the method in [198], we utilize an exponential budget split, i.e., $0,1,2,4,\ldots,B$, which is denoted by $expSplit(\cdot)$.

In Algorithm 5, $V(P)$ refers to the set of nodes covered by a path P, $P_1 \cdot P_2$ is the concatenation of path P_1 and path P_2. As the budget is assumed to be an integer, like the original recursive greedy algorithm, the modified recursive greedy algorithm also requires rescaling the real-valued budget and costs. Moreover, the number of iterations i is often set to a value of $\lceil \log_2(B) \rceil$ [198]. It is not difficult to see that the running time of the algorithm is $O(|\tilde{V}||\tilde{B}|^{O(\log_2|\tilde{V}|)})$. In particular, when $\tilde{V} = V$ and $\tilde{B} = B$, i.e., without sub-approximation, the approximation guarantee of the modified algorithm remains the same as the original algorithm, as stated in the following corollary:

Corollary 5.7.1. *Let $P^* \in P(s,t,B)$ such that $P^* = (s = v_0, v_1, \ldots, v_k = t)$ be an optimal s-*

t path solution. Let P be the path returned by the modified recursive greedy algorithm $RG(s,t,B,C,V,i)$*, such that,* $\tilde{V} = V$ *and* $\tilde{B} = B$*. If* $i \geq \lceil 1 + \log k \rceil$*, then* $f(V(P)|C) \geq f(V(P^*)|C)/\lceil 1 + \log k \rceil$.

Proof. Proof. The result directly follows from the original proof where visit cost is zero due to Chekuri and Pal [197]. □

Chapter 6

Multi-Channel Marketing with Budget Complementarities

Having specialized algorithms that solve for optimal or near-optimal marketing strategies in specific settings like in Chapter 4 and Chapter 5 are not sufficient to build the algorithmic marketing system (see Figure 1.1). In fact, there remain other challenges that need to be addressed. First, even if we are able to optimize our actions in each marketing setting (or *channel*) for given budget, at a higher managerial level, it is not clear how to split budget among multiple marketing channels. Second, as the effect of each channel is generally evaluated by simulation models, any budget optimization algorithm that ignores the variability in running time of these simulations would fail to provide efficient marketing plans. Third, the effect of marketing actions does not follow the commonly-assumed smooth or continuous patterns instead highly discrete, which also challenges the existing convex optimization techniques. This chapter aims to address these challenges by introducing a powerful discrete budget optimization framework designed to interact with simulation models in an efficient manner providing satisfying marketing plans.

6.1 Introduction

The emergence of digital media, such as the world wide web, search engines, and online social networks, has opened up tremendous opportunities for today's marketers to look for prospects and engage existing customers. A mix of these innovative channels with traditional ones, such as TV, direct mailing, and door-to-door marketing, has been widely adopted by many companies to generate more sales, maintain stronger customer relationships, and achieve a higher customer retention rate [25]. Despite its benefits, this practice has also significantly increased operational complexity, making marketing one of the key managerial challenges [26, 27]. The demand for effective budget allocation solutions in

multi-channel marketing campaigns has in turn given rise to major software products aimed towards this goal, including those developed by SAS and IBM, among others.

In order to determine the optimal budget allocation among the marketing channels, the marketer needs a way to evaluate the effectiveness of alternative budget splits. Advanced simulation models, and abundant data that can be used to calibrate them, allow doing just that. The use of simulations, as compared to analytic objective functions (such as concave and continuous utility being maximized), introduces an important technical challenge: simulations are often slow, and parsimony is therefore crucial in query-based black-box optimization methods. A second technical challenge arises from the fact that the response function for each channel (such as the number of individuals who buy the product) commonly exhibits *budget complementarities*, requiring a non-trivial added expense on a channel to make a significant impact on the response function. For example, in door-to-door marketing, a budget increment needs to be sufficient to hire another salesman, or increase their working hours by a discrete amount. Similarly, in keyword auctions, moving up a slot requires a discrete added investment, the amount of which depends on specific pricing strategies and competition among bidders.

To address these challenges, we present a novel and powerful discrete budget optimization framework to generate near-optimal budgeting strategies when the budget allocation response is a *step function* represented by a simulator. We first show that the budget optimization problem can be readily cast into a multi-choice knapsack problem (MCKP), which admits effective state-of-the-art algorithms. Since the step-wise response function is represented using a simulator, the thresholds which identify the discrete jumps (serving as weights in the MCKP) are unknown, and a finite number of simulator queries can at best isolate these to small *intervals*. Consequently, the MCKP can at best be solved approximately. We show that under mild conditions, for sufficiently small bounds on weights, solving the MCKP with weight upper bounds yields a 2-approximation, and this bound is tight. Surprisingly, this bound holds even when the thresholds are not fully explored.

Next, we develop two efficient query algorithms that allow us to obtain tight intervals around MCKP weights (as well as associated response values). The first, Generalized Binary Query (GBQ) is a generalization of the classic binary search applied to the case of multiple thresholds, which we show to be more efficient than simple linear search. The second approach, namely, Heuristic Binary Query (HBQ), is designed to reduce the number of queries needed per iteration, with the help of the solution of an auxiliary optimization problem that corresponds to the best possible payoff in the next round.

Our framework is implemented in a simulated marketing environment that mimics a real-world multi-channel campaign in a targeted geographical area. We use this simulator to conduct extensive experiments to demonstrate the usability of the proposed framework and compare the performance of two query algorithms as well as a Simulated Annealing (SA) algorithm, which is a well-known stochastic local search method typically used for problems with highly non-linear objectives. Our results show that HBQ achieves payoffs only slightly lower than GBQ, but using significantly less time. Moreover, both HBQ and GBQ outperform SA in all experiments, reaching competitive payoff levels significantly faster.

In summary, we make the following contributions:

1. A novel discrete budget optimization problem with an application to multi-channel marketing, transformed into a multi-choice knapsack problem;

2. A theoretical analysis of the resulting problem in which weights (corresponding to steps in the response function) can only be bounded, showing that solving the approximate MCKP with upper bounds on weights yields a tight 2-approximation;

3. Two novel simulation query strategies for obtaining upper and lower bounds on MCKP weights: Generalized Binary Query and Heuristic Binary Query;

4. A simulation platform to evaluate the multi-channel marketing algorithms;

5. Extensive experiments across different marketing situations and a variety of budgets demonstrating the usability of the proposed framework and efficacy of proposed query methods, compared to a simulated annealing algorithm.

6.2 Related Work

Budget optimization is a classical problem in economics, operations research, and management science. The problem is traditionally tackled by maximizing specific objectives, e.g., sales, profit, or customer equity, based on a set of pre-specified constraints [26]. In practice, the real-world markets involve considerable complexity, for example, as a consequence of social interactions and influence [124]. Consequently, simulation-based optimization methods, such as system dynamic models, are often used to aid decision making [207]. Interestingly, most previous work assumes a continuous objective and solves the problem by smooth techniques which rely on computing a gradient or Hessian of the objective. In contrast, we present a novel combinatorial optimization framework in a setting where channel payoff exhibits strong budget complementarities which we model by considering a stepwise response function. Note that the non-smooth optimization techniques [208] such as the *sub-gradient* methods fail in our setting since the sub-gradient is not informative in step functions.

Recently, extensive work focused on budget optimization for a single marketing channel. For example, Yang et al. propose a hierarchical budget allocation framework for online advertising that links decisions at different decision levels, such as system, campaign, and keyword [209]. Using individual data, several authors model user responses to marketing actions as a Markov Chain and solve the budget optimization problem using a constrained Markov Decision Process (MDP). For instance, Abe et al. develop an MDP framework with reinforcement learning for direct mailing campaigns [210]. Under the assumption of positive carryover effects, Archak et al. propose an optimal greedy algorithm for online adverting in an MDP framework [211]. Boutilier and Lu address the allocation of a bud-

get among multiple MDPs representing different types of users or groups [212]. Zhang and Vorobeychik develop a route planner for door-to-door marketing based on submodular optimization [213]. These specialized budget optimizers based on empirical data are quite effective for targeted marketing, but are still specialized sub-problems of the overall problem of optimally allocating a budget among a collection of marketing channels, which we address.

The multi-choice knapsack problem (MCKP) is a variant of the simple knapsack problem in which a class can have multiple items but only one can be chosen, and has been extensively explored in the literature [214, 215, 216]. Our theoretical analysis of approximation bounds is related to *sensitivity analysis* in operations research [217], which examines the sensitivity of the optimal solution to changes in the coefficient matrix, cost, price, and budget. Hifi et al. provide sensitivity intervals for the 0-1 knapsack problem subject to changes of item weights [218]. In contrast, we address the question about the worst-case performance of the MCKP in which weights are tightly bounded, as a means to a broader end of multi-channel marketing budget allocation. Finally, our work is related to, but distinct from *robust optimization* [219, 220, 221, 222]. This line of work usually imposes a limit for the number of uncertain parameters (e.g., weights) to avoid overly conservative solutions, while we use the upper-bounds for all weights to secure a feasible solution. Goerigk et al. study query strategies for a robust knapsack problem, rather than a general MCKP as in our case [222]. They assume that a single query returns "true" weight; by contrast, we design a sequence of queries to efficiently approximate weight bounds, but cannot in general obtain true weights, as is commonly the case when marketing response is simulated.

6.3 Problem Statement

Suppose that a marketer is given a fixed *budget* B to advertise a new product over n marketing channels. Let $x = (x_1, \ldots, x_n)$ represent a budget split with x_i the amount of the

budget allocated to channel i. Let $r(x)$ be the net reward to the marketer (e.g., in terms of overall product uptake) given a budget split x. Our goal is to solve the following *multi-channel marketing optimization* (MCMO) problem:

$$\max_{x} r(x) \tag{6.1a}$$

$$\text{s.t.} : \sum_{i=1}^{n} x_i \leq B, \tag{6.1b}$$

that is, we aim to optimally split the budget B across the n channels to maximize the total net payoff. If $r(x)$ are concave and known and the budget divisible, as is commonly assumed in numerous related formulations, Problem 6.1 is straightforward to solve with the standard convex optimization methods (indeed, this is just the standard budget-constrained utility maximization problem in consumer theory [29]). What has not received much attention, and is of interest to us, is this problem in which (a) $r(x)$ exhibits strong, but imperfect, complementarities, and (b) $r(x)$ is not a priori known, but specified by a time consuming simulation model. For example, $r(x)$ may capture a complex social influence diffusion process which cannot be analytically characterized and is evaluated in simulations, as is the case for many important social influence models in the literature [124, 186]. Moreover, making a non-negligible impact on a given agent's decision (e.g., in seeding them by providing this agent a product at a low cost) incurs a non-zero cost which may be a complex function of contextual factors also embedded in a simulation, and therefore unknown a priori. As another example, online auction-based advertising channels (such as keyword auctions) require a sufficiently high investment to move into a higher priority slot, which makes a discontinuous impact on the expected number of clicks and, thus, conversions, and the precise amount of this investment is a complex function of bidding behavior by a collection of agents which can be captured in a simulation environment, but could be difficult to characterize in closed form.

In order to model such complementarities, we begin by assuming that $r(x) = \sum_i r_i(x_i)$,

with $r_i(x_i)$ increasing and $r_i(0) = 0$. For each channel i, we suppose that there is a collection of thresholds w_{ij} so that crossing a threshold results in a jump in $r_i(x_i)$. Formally, we assume that $r_i(x_i)$ is a step function of the following form:

$$r_i(x_i) = \begin{cases} 0, & x_i < w_{i1} \\ r_{i1}, & w_{i1} \leq x_i < w_{i2} \\ \ldots, & \ldots \\ r_{iJ_i}, & x_i \geq w_{iJ_i} \end{cases}$$

where there are $J_i(\geq 1)$ thresholds $\{w_{ij}\}_{j=1,\ldots,J_i}$ and non-zero payoff levels $\{r_{ij}\}_{j=1,\ldots,J_i}$.

As the first step, we transform Problem 6.1 with the structure just described into an equivalent multi-choice knapsack problem (MCKP). Since in our model any investment level not corresponding to a threshold is wasteful, the decision problem is to determine at which threshold level j we should allocate the budget for each channel i. We encode this decision as a binary variable y_{ij}, which is 1 whenever we allocate budget at threshold level j for channel i. The MCKP is then

$$\max_{y_{ij} \in \{0,1\}} \sum_{i=1}^{n} \sum_{j=1}^{J_i} r_{ij} y_{ij} \tag{6.2a}$$

$$\text{s.t.} : \sum_{i=1}^{n} \sum_{j=1}^{J_i} w_{ij} y_{ij} \leq B \tag{6.2b}$$

$$\forall i \in \{1, \ldots, n\}, \sum_{j=1}^{J_i} y_{ij} \leq 1, \tag{6.2c}$$

where the first inequality is the budget constraint, and the second implies that at most one budget level can be picked for any channel. Note that Problem 6.1 with known thresholds is harder than the MCKP, as a polynomial oracle for it can solve an arbitrary instance of MCKP. Throughout, it will be useful to denote the above MCKP as $MCKP(J_i, w_{ij})$, with a specified set of J_i weights for each i; the corresponding r_{ij} will be clear from context.

Armed with the MCKP formulation, we can now identify the key technical challenges: (1) w_{ij} can only be approximately determined from a finite number of queries, since these lie on a continuous interval, and (2) the problem parameters w_{ij} and r_{ij} must be obtained using time consuming simulations. We address these challenges below.

6.4 Approximate Multi-Choice Knapsack

We begin by addressing the first challenge above: the threshold values w_{ij} cannot be identified exactly. Surprisingly, despite considerable prior work on approximate and robust knapsack problems, this particular problem remains open, to the best of our knowledge. Our analysis of approximate MCKP may thus be of independent interest, but for us it is just an important piece of the puzzle. We subsequently take up the complementary piece: efficient query strategies for achieving good MCKP approximations.

Formally, suppose that w_{ij} are not known, but we have lower and upper bounds so that $w_{ij} \in [\underline{w}_{ij}, \overline{w}_{ij}]$, and let $\varepsilon := \max_{ij}\{\overline{w}_{ij} - \underline{w}_{ij}\} > 0$, which implies that $\overline{w}_{ij} - w_{ij} \leq \varepsilon$ for all i, j. Since w_{ij} are unknown, we propose to approximate the associated MCKP with $MCKP(J_i, \overline{w}_{ij})$. Next, we demonstrate that under a set of conditions which can be guaranteed with sufficiently many simulation queries, we can obtain a 2-approximation of $MCKP(J_i, w_{ij})$, and this approximation is tight unless the weights are known *exactly*.

First, notice that we have thus far implicitly assumed that *every interval contains exactly one threshold* w_{ij}. This is a significant challenge: even if we can guarantee that for a particular fixed ε all thresholds are bounded within intervals of length at most ε, we would still be unable to distinguish thresholds that all cluster within some such interval. Fortunately, even in such a case we do know that all such thresholds are in one of the intervals we have identified. This turns out to be sufficient to obtain the approximation guarantees.

Formally, suppose that there are $J'_i \leq J_i$ thresholds for channel i, and we solve $MCKP(J'_i, \overline{w}_{ij})$. For ease of exposition, let us denote the optimal value of Problem 6.2 as $OPT(J_i, w_{ij})$, while the optimal value of $MCKP(J'_i, \overline{w}_{ij})$ will be denoted by $OPT(J'_i, \overline{w}_{ij})$. Finally, let

$OPT(J_i', \underline{w}_{ij})$ be the optimal value of the problem $MCKP(J_i', \underline{w}_{ij})$, which uses lower-bound weights $\{\underline{w}_{ij}\}$ but upper-bound payoffs $\{r_i(\overline{w}_{ij})\}$ of J_i' intervals. The following is our key result:

Theorem 6.4.1. *Assume that* $\overline{w}_{ij} \leq B, \forall i \in 1, \ldots, n, \forall j \in 1, \ldots, J_i'$ *and denote* $\overline{w}_{min} = \min\{\overline{w}_{ij}\}_{i=1,\ldots,n; j=1,\ldots,}$ *If* $\varepsilon \leq \overline{w}_{min}/n$*, then, 1)* $OPT(J_i', \overline{w}_{ij}) \geq \frac{1}{2} OPT(J_i, w_{ij})$*; and 2) the bound is tight.*

We prove this theorem in a series of steps.

Lemma 6.4.2. $OPT(J_i, w_{ij}) \leq OPT(J_i', \underline{w}_{ij}), \forall J_i' \leq J_i$.

Proof. Since payoff increases with respect to weight, for channel i, any unexplored threshold h must be in one of the intervals we already discovered. Suppose it is in the interval $[\underline{w}_{ij}, \overline{w}_{ij}]$, where, $r(\underline{w}_{ij}) < r(\overline{w}_{ij})$. Clearly, option h $(w_{ih}, r(w_{ih}))$ (corresponding to threshold h) is dominated by option $(\underline{w}_{ij}, r(\overline{w}_{ij}))$, as the latter has lower cost but higher payoff. As to multiple channels, this suggests that $OPT(J_i, w_{ij})$ is always upper-bounded by $OPT(J_i', \underline{w}_{ij})$, although the number of intervals we identified is at most the number of thresholds: $J_i' \leq J_i$. □

Lemma 6.4.3. *Assume that* $\overline{w}_{ij} \leq B, \forall i \in 1, \ldots, n, \forall j \in 1, \ldots, J_i'$ *and denote* $\overline{w}_{min} = \min\{\overline{w}_{ij}\}_{i=1,\ldots,n; j=1,\ldots,J_i'}$*. If* $\varepsilon \leq \overline{w}_{min}/n$*, then,* $OPT(J_i', \underline{w}_{ij}) \leq 2OPT(J_i', \overline{w}_{ij}), \forall J_i' \leq J_i$.

Proof. Let $\dot{Y} = \{\dot{y}_{ij}\}$ and $\hat{Y} = \{\hat{y_{ij}}\}$ be the optimal solution that corresponds to $OPT(J_i', \underline{w}_{ij})$ and $OPT(J_i', \overline{w}_{ij})$ respectively. Note that $\varepsilon \geq \overline{w}_{ij} - \underline{w}_{ij}$, thus, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i'} \overline{w}_{ij} \hat{y}_{ij} \leq \sum_{i=1}^{n} \sum_{j=1}^{J_i'} \underline{w}_{ij} \dot{y}_{ij} + \varepsilon \sum_{i=1}^{n} \sum_{j=1}^{J_i'} \dot{y}_{ij} \leq B + n\varepsilon \tag{6.3}$$

where the last inequality holds due to the fact that $\dot{Y} = \{\dot{y}_{ij}\}$ is the optimal solution for $OPT(J_i', \underline{w}_{ij})$, which satisfies the budget constraint. By the definition of $\overline{w}_{min}$ and the as-

sumption

$$\varepsilon \leq \frac{\overline{w}_{min}}{n} \tag{6.4}$$

we have $n\varepsilon \leq \overline{w}_{min} \leq \overline{w}_{ij}$, $\forall i \in 1,\ldots,n$, $\forall j \in 1,\ldots,J'_i$.

Consider dropping any non-zero item s in $\dot{Y}$, the resulting solution must be feasible for the problem $MCKP(J'_i,\overline{w}_{ij})$ according to inequality (6.3), and bounded by $OPT(J'_i,\overline{w}_{ij})$. Thus, we have $\sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\dot{y}_{ij} - r_s \leq \sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}}$, and by rearranging we get that

$$\sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\dot{y}_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}} \leq r_s \tag{6.5}$$

We have assumed that $\overline{w}_i \leq B$, $\forall i \in 1,\ldots,n$, $\forall j \in 1,\ldots,J'_i$, so $\overline{w}_s \leq B$, which implies $r_s \leq \sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}}$. By inequality (6.5) we know that

$$\sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\dot{y}_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}} \leq \sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}}$$

and thus $\sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\dot{y}_{ij} \leq 2\sum_{i=1}^{n}\sum_{j=1}^{J'_i} r_{ij}\hat{y_{ij}}$. Hence, $OPT(J'_i,\overline{w}_{ij})$ is a 2-approximation of $OPT(J'_i,\underline{w}_{ij})$. □

of Theorem 6.4.1. Part 1 of the theorem follows directly from Lemmas 4.2 and 4.3. For Part 2, we show that the factor-2 bound is *tight*. Consider a simple example, in which we are given a budget of 1 to advertise in only two channels, such that, $w_1 = 1/2 - \delta$, $w_2 = 1/2 + \delta$, and, $r_1 = r_2 = 1$, and $0 < \delta < 1/2$. Recall that in our setting, both w_1 and w_2 are unknown, but instead we use their upper bounds $\overline{w}_1$ and $\overline{w}_2$. If the upper bounds are not identical to the actual weights, we can only choose one of the channels. Therefore, the approximate problem gives us at most a payoff of 1, whereas we can get a payoff of 2 when exact weights are known by choosing both channels. Moreover, the example also suggests that one will never be able to get a better than 2-approximation no matter how small ε is. □

Note that $MCKP(J_i', \overline{w}_{ij})$ is also an NP-hard problem. Suppose that we use a c-approximation algorithm to solve $MCKP(J_i', \overline{w}_{ij})$, then the statement below naturally follows from Theorem 6.4.1.

Corollary 6.4.4. *Assume that* $\overline{w}_{ij} \leq B, \forall i \in 1,\ldots,n,\ \forall j \in 1,\ldots,J_i'$ *and denote* $\overline{w}_{min} = \min\{\overline{w}_{ij}\}_{i=1,\ldots,n;j=1,\ldots,J_i'}$. *If* $\varepsilon \leq \overline{w}_{min}/n$, *a* c*-approximation algorithm for* $MCKP(J_i', \overline{w}_{ij})$ *achieves at least 1/2c of the optimal value of* $MCKP(J_i, w_{ij})$.

Proof. Let G be the near-optimal solution value given by the c-approximation algorithm for $MCKP(J_i', \overline{w}_{ij})$, then $G \geq \frac{1}{c} OPT(J_i', \overline{w}_{ij})$. From Theorem 6.4.1, we know that $OPT(J_i', \overline{w}_{ij}) \geq \frac{1}{2} OPT(J_i, w_{ij})$. Thus, it must be the case that $G \geq \frac{1}{2c} OPT(J_i, w_{ij})$. □

6.5 Query Strategies for Budget Allocation

Our analysis so far assumed that we have been given a set of intervals for MCKP weights w_{ij} (that is, thresholds at which the response function jumps in value) which are sufficiently tight, in the sense of Condition (6.4), to ensure a 2-approximation using just the interval upper bounds in an MCKP. The key next question, which we now address, is how to obtain such intervals efficiently using a sequence of simulation queries. First, observe that there is a straightforward query mechanism which can produce intervals of arbitrary width in linear time: finely discretize each channel in the interval $[0, B]$, and query each discrete value for each channel independently. However, this approach can be extremely wasteful: for example, one channel can yield a small response and require a minimal investment of B; in most cases, we can quickly discover this and ignore this channel altogether. We will propose a more intelligent query algorithm which *interleaves* MCKP computation with queries. This allows more efficient exploration of the allocation space, and early termination once a near-optimal allocation is found.

Data: maximum iteration K, total budget B, parameter θ
Result: budgeting plan $P_b = \{b_i\}_{i=1,\dots,n}$
$\mathscr{I}_i \leftarrow \emptyset, \forall i = 1,\dots,n$;
$b_i \leftarrow 0, \forall i = 1,\dots,n$;
$k \leftarrow 0$;
foreach *channel* $i \in 1,\dots,n$ **do**
    $v_0 \leftarrow (0,0)$;
    $v_B \leftarrow (B, r_i(B))$;
    $\mathscr{I}_i \leftarrow \mathscr{I}_i \cup (v_0, v_B)$;
end
while $k < K$ **do**
    $\{\overline{w}_{ij}\} \leftarrow U_{bs}(\mathscr{I}_i), \forall i \in 1,\dots,n$;
    $\{\underline{w}_{ij}\} \leftarrow L_{bs}(\mathscr{I}_i), \forall i \in 1,\dots,n$;
    $\underline{y} \leftarrow MCKP(J'_i, \overline{w}_{ij})$;
    $\bar{y} \leftarrow MCKP(J'_i, \underline{w}_{ij})$;
    $P_b \leftarrow \underline{y} \circ \bar{w}$;
    if $\bar{y} \cdot r - \underline{y} \cdot r \leq \theta$ **then**
        break;
    end
    foreach *channel* $i \in 1,\dots,n$ **do**
        $updateIntervals(\mathscr{I}_i)$;
    end
    $k \leftarrow k + 1$;
end
return P_b

Algorithm 6: Iterative Budgeting Algorithm.

6.5.1 Iterative Budgeting Algorithm

Now we present our primary algorithm, termed Iterative Budgeting (IB); it is given as Algorithm 6. In Algorithm 6, $\mathscr{I}_i$ stands for the set of intervals for channel i. Moreover, an interval is represented as a tuple (v_l, v_u), consisting of a lower bound $v_l = (w_l, r(w_l))$ and an upper bound $v_u = (w_u, r(w_u))$. $r_i(\cdot)$ is the step-wise payoff function for channel i evaluated by simulation. For any $\mathscr{I}_i$, $U_{bs}(\cdot)$ and $L_{bs}(\cdot)$ are simply functions that return a set of upper-bound and lower-bound weights respectively. $\underline{y}$ and $\bar{y}$ are the solutions for $MCKP(J_i', \overline{w}_{ij})$ and $MCKP(J_i', \underline{w}_{ij})$ respectively (e.g., using CPLEX). Each round, a new budget allocation plan $P_b = \underline{y} \circ \bar{w} = \{b_i = \sum_j \underline{y}_{ij} \bar{w}_{ij}\}$ is computed. The method $updateIntervals(\cdot)$ updates the set of intervals by sending more queries to the specified channel simulator, which will be described in more details in the next section.

The algorithm starts with one interval per channel and uses the upper-bound weights and payoffs to compute an initial solution. Next, in each iteration, it sends more queries to each channel and updates $\mathscr{I}_i$ with returned payoffs, and computes a new solution. As it runs more iterations, and the set of intervals is further refined, generally, the solved payoff $(\underline{y} \cdot r)$ will approach its upper-bound of the optimum $(\bar{y} \cdot r)$, where the notation $a \cdot b$ is the dot product of two vectors. θ is a parameter that controls the solution quality. Specifically, Lemma 6.4.2 shows that $OPT(J_i', \underline{w}_{ij})$ is an online upper bound for $OPT(J_i, w_{ij})$. This bound is very useful, since $OPT(J_i, w_{ij})$ is unknown; we can compute $OPT(J_i', \underline{w}_{ij})$ (suppose this is computationally feasible) and use

$$\Delta OPT = OPT(J_i', \underline{w}_{ij}) - OPT(J_i', \overline{w}_{ij})$$

to assess the quality of the approximate solution. In particular, when $\Delta OPT \leq OPT(J_i', \overline{w}_{ij})$, we are guaranteed to have a 2-approximation. This is because: $OPT(J_i', \underline{w}_{ij}) - OPT(J_i', \overline{w}_{ij}) \leq OPT(J_i', \overline{w}_{ij})$ and thus $OPT(J_i', \underline{w}_{ij}) \leq 2OPT(J_i', \overline{w}_{ij})$. As $OPT(J_i, w_{ij}) \leq OPT(J_i', \underline{w}_{ij})$, we have $OPT(J_i, w_{ij}) \leq 2OPT(J_i', \overline{w}_{ij})$. Consequently, we are guaranteed to get a 2-approximation

if we set $\theta = \bar{y} \cdot r = OPT(J_i', \overline{w}_{ij})$. In practice, we set a much smaller θ to obtain a better solution.

6.5.2 Query Strategies

Now we discuss how Algorithm 6 updates intervals by queries. We first notice that in the case of a single threshold per channel, binary search is more efficient than linear search. Given budget B, using binary search to obtain an interval width ε, one has to send at most $\lfloor \log(B/\varepsilon) \rfloor + 1$ queries (1 initial query for $r_i(B)$ and $r_i(0) = 0$). In particular, if $\exists i \in \mathbb{Z}^+$, such that $\varepsilon = B(\frac{1}{2})^i$, then we need exactly $i+1$ queries to guarantee ε, where $i = \log(B/\varepsilon)$. In contrast, using linear search, we need B/ε queries to achieve an interval width of ε, which is clearly less efficient than the binary search. The task is to extend binary search to handle the case that a channel could have a finite number of thresholds.

6.5.2.1 Generalized Binary Query Algorithm

We propose a Generalized Binary Query (GBQ) Algorithm (see Algorithm 7) to implement $updateIntervals(\cdot)$, which extends the binary search method to the case of multiple thresholds.

The algorithm scans each interval and creates a new query using a weight that is halfway between the lower and upper bound weights. In other words, it takes one binary search action within each interval available at an iteration. If the queried payoff equals to the lower (upper) bound payoff, then it updates the lower (upper) bound weight accordingly. Otherwise, a new interval is added to the current set of intervals $\mathscr{I}_i$. Therefore, the query action has two effects: 1) narrowing existing intervals, and 2) identifying new intervals (thresholds).

Note that, to obtain an interval width of ε, we could also send B/ε queries using a simple linear search. Suppose a channel has J thresholds, $\{w_i\}_{i=1,\cdots,J}$, such that, $w_1 < w_2 <, \cdots, < w_J$. Define $d_{min} = \min_j\{w_j - w_{j-1}\}$, which is the minimum distance between

Data: a set of intervals $\mathscr{I}_i$
Result: updated $\mathscr{I}_i$
$\mathscr{I}_i' \leftarrow \emptyset$
foreach *interval* (v_l, v_u) *in* $\mathscr{I}_i$ **do**
    $w' \leftarrow (w_l + w_u)/2$;
    $r' \leftarrow (w')$;
    if $r' = r_l$ **then**
        $v_l \leftarrow (w', r_l)$;
    else if $r' = r_u$ **then**
        $v_u \leftarrow (w', r_u)$;
    else
        $v' \leftarrow (w', r')$;
        $\mathscr{I}_i' \leftarrow \mathscr{I}_i' \cup (v_l, v')$;
        $\mathscr{I}_i' \leftarrow \mathscr{I}_i' \cup (v', v_u)$;
        $\mathscr{I}_i \leftarrow \mathscr{I}_i \backslash (v_l, v_u)$;
end
return $\mathscr{I}_i' \cup \mathscr{I}_i$
Algorithm 7: Generalized Binary Query (GBQ) Algorithm: updateIntervals($\mathscr{I}_i$).

two adjacent thresholds. Theorem 6.5.1 states that GBQ is more efficient than linear search.

Theorem 6.5.1. *Assume $\varepsilon = B(\frac{1}{2})^i$, where $i \in Z^+$. If $\varepsilon \geq d_{min}$, in the worst case, GBQ uses the same number of queries as linear search; however, if $\varepsilon < d_{min}$, GBQ always uses fewer queries.*

Proof. First, we notice that for a given ε that satisfies $\varepsilon = B(\frac{1}{2})^i$, the linear search with B/ε queries can be reimplemented as a naive binary search as follows: first, we query $r(B)$; next, $r(B/2)$; then, $r(B/4)$ and $r(3B/4)$, and so on. In total, we need $i+1$ iterations to achieve ε. In terms of queries, we need: $1+1+2+\ldots+2^{(i-1)} = 2^i$, which is exactly B/ε.

When $\varepsilon \geq d_{min}$, it is possible that each query will find a new interval, and thus 2^i queries may be required to obtain intervals of width ε in the worst case. However, when $\varepsilon < d_{min}$, we are guaranteed that GBQ has captured all thresholds at some iteration $\hat{i}+1$ before the iteration $i+1$. Otherwise, there must be two thresholds that are in the same interval, therefore, $d_{min} < \varepsilon$, which is a contradiction. First, we compare total number of queries used by GBQ and the naive binary search up to iteration $i+1$. Among these iterations, we are not guaranteed to capture all thresholds, so GBQ will need the same number of $(2^{\hat{i}})$

queries as the naive binary search in the worst case. Second, we compare the two algorithms for iterations after $\hat{i}+1$. Starting from iteration $(\hat{i}+2)$, GBQ only sends a fixed number of queries (equal to the actual number of thresholds, denoted by J), as the generalized binary search will skip intervals that do not have thresholds. The total number of queries used by GBQ after $(\hat{i}+1)$ is $J(i-\hat{i})$. By contrast, the naive binary search (equivalent to linear search as we have shown) has to explore all intervals available for each iteration. After iteration $\hat{i}+1$, it will send $(2^i-2^{\hat{i}})$ queries, which is *strictly* larger than $J(i-\hat{i})$. Notice that the number of queries used at iteration $\hat{i}+1$ by the naive binary search satisfies: $2^{\hat{i}-1} \geq J$, when all threshold are captured. In addition, the number of queries used by the naive binary search at iteration $(\hat{i}+2)$ is $2^{\hat{i}} (> 2^{\hat{i}-1})$, which implies that $2^{\hat{i}} > J$. Thus, we have $2^i - 2^{\hat{i}} = 2^{\hat{i}}(2^{i-\hat{i}}-1) \geq 2^{\hat{i}}(i-\hat{i}) > J(i-\hat{i})$. Therefore, if $\varepsilon < d_{min}$, GBQ always uses fewer queries. □

Notice that as the IB algorithm iterates using the GBQ strategy, in each iteration it maintains several intervals for each channel with uniform length (ε in Section 6.4). We would expect that in some iteration, it will satisfy Condition (6.4) and ensure the 2-approximation in Theorem 6.4.1. Formally, we show this to be the case in Corollary 6.5.2, where k is the number of iterations of IB.

Corollary 6.5.2. *When $k \geq \log \frac{Bn}{w_{min}}$, IB algorithm with GBQ strategy achieves a 2-approximation of MCKP(J_i, w_{ij}), where k is the number of iteration.*

Proof. Based on GBQ, we know that ε, the width of any interval satisfies: $\varepsilon = B/2^k$. As Condition 6.4 requires that: $\varepsilon \leq w_{min}/n$, we have $B/2^k \leq w_{min}/n$, which is $k \geq \log \frac{Bn}{w_{min}}$. □

6.5.2.2 Heuristic Binary Query Algorithm

Notice that using the query Algorithm 7, Algorithm 6 needs to send an increasing number of queries for each channel in each iteration. Although the number of queries required for an iteration will eventually be bounded by the actual number of thresholds, most queries

are wasteful, especially when the solution approaches its optimum. To improve efficiency of the query search we propose a Heuristic Binary Query Algorithm (HBQ), described in Algorithm 8. Notably, as distinct from GBQ, HBQ only sends *one* query per channel to the "most profitable" interval in each iteration, where this channel is determined by $\hat{y}$, computed as:

$$\hat{y} \leftarrow MCKP(J_i', (\underline{w}_{ij} + \overline{w}_{ij})/2).$$

That is, $\hat{y}$ is the solution of MCKP which uses the *average* of lower-bound and upper-bound weights, but the upper-bound payoffs of all intervals. Intuitively, $\hat{y}$ gives the highest payoff we would obtain in the next iteration if we query all current available intervals in a binary manner. Notably, if IB is to use HBQ, it is modified by using $\hat{y}$ in place of $\bar{y}$.

Data: a set of intervals $\mathscr{I}_i$, solution of best payoff in next iteration $\hat{y}$
Result: updated $\mathscr{I}_i$
$\mathscr{I}_i' \leftarrow \emptyset$
foreach *interval* (v_l, v_u) *in* $\mathscr{I}_i$ **do**
    if *item l is not selected in* $\hat{y}$ **then**
        continue;
    end
    $w' \leftarrow (w_l + w_u)/2$;
    $r' \leftarrow (w')$;
    if $r' = r_l$ **then**
        $v_l \leftarrow (w', r_l)$;
    else if $r' = r_u$ **then**
        $v_u \leftarrow (w', r_u)$;
    else
        $v' \leftarrow (w', r')$;
        $\mathscr{I}_i' \leftarrow \mathscr{I}_i' \cup (v_l, v')$;
        $\mathscr{I}_i' \leftarrow \mathscr{I}_i' \cup (v', v_u)$;
        $\mathscr{I}_i \leftarrow \mathscr{I}_i \backslash (v_l, v_u)$;
end
return $\mathscr{I}_i' \cup \mathscr{I}_i$

Algorithm 8: Heuristic Binary Query (HBQ) Algorithm: updateIntervals($\mathscr{I}_i$, $\hat{s}$).

6.6 Experiments

In order to evaluate our approach we developed a multi-channel marketing simulator. In this simulator, a fixed budget is allocated to advertise a technology over four channels: *door-to-door*, *keyword auction*, *direct mailing*, and *broadcast*, where the payoff for each channel is modeled by a step function of the allocated budget and corresponds to the total number of adopters of the technology. One of the sources of complexity is that response is a function not merely of those reached by marketing directly, but also by those indirectly affected through social influence. Below we describe the details of the simulator.

The target population was comprised of 536 households in San Diego county, CA which is shared among the marketing channels.[1] The full marketing campaign was restricted to 3 months. As our baseline we implemented the well-known simulated annealing (SA) algorithm which was tuned to our problem domain [223]. Our algorithm used CPLEX 12.6.1 through a Java API to solve MCKP, and experiments were run on an Ubuntu Linux 64-bit PC with 32 GB RAM and four 8-core Intel Xeon 2.1 GHz CPUs.[2]

6.6.1 Marketing Simulator

6.6.1.0.1 Door-to-Door Marketing In door-to-door marketing, a sales agent knocks on a customer's door and attempts to initiate a discussion that could eventually lead to a sale. To simulate relevant marketing decisions, we adopt the door-to-door marketing route planner developed by [32], which uses an independent cascades (IC) model of social influence [124] to model the spread of successful adoptions by individuals to their geographic neighbors. The key parameter in the IC model is the *transmission probability p* representing the likelihood of a newly adopted customer affecting its socially-connected neighbors. For a fixed budget allocated to this channel, we follow Zhang and Vorobeychik in approxi-

[1] Population was restricted primarily to restrict the time for each simulation run to enable a sufficient number of total runs for meaningful comparisons.

[2] Implementation of the marketing simulator and algorithms is available at: https://github.com/haffwin/mcmo.git.

mately computing the optimal routes for sales people to maximize influence using a greedy algorithm [213]. The step-function nature of the channel response arises from the fixed cost needed to add a sales person with sufficient allocated time to cover at least one household which has not already been covered by others.

6.6.1.0.2 Keyword Auction The keyword auction simulator mimics bidding decisions in a search engine keywords auction, analogous to Google AdWords. In the simulation, a marketer is given a fixed budget b to bid on k predefined keywords (relevant to the promoted product). A higher bid may help secure a higher position for one's advertised content in the search results page, and thus a higher *click-through* rate. However, the bid needs to be increased sufficiently to jump to a higher slot—hence the step-function nature of the response.

A keyword i is represented by a tuple (c_i, f_i), where c_i is its *cost-per-click* and f_i the fraction of users in the targeted region who click on the ads each day. c_i and f_i are determined as follows: $c_i = i + e_i^1$ and $f_i = \alpha(i + e_i^2)$ for all i, where, α is a pre-defined coefficient, and e_i^1 and e_i^2 are random variables drawn from the uniform distribution on $[0,1]$ i.i.d. for each i. In expectation, c_i is approximately proportional to f_i. Finally, let R_{oad} be the conversion rate, or fraction of clicks on an ad that result in a purchase.

We assume that for a given budget b, the marketer splits it equally over the 3-month period, and solves the following *fractional knapsack problem* every day to maximize the total number of clicks (since conversion rate is constant):

$$\max_x \sum_{i=1}^{k} x_i f_i \tag{6.6a}$$

$$M \sum_{i=1}^{k} x_i c_i f_i \leq b/d \quad \forall\, i, \tag{6.6b}$$

where x is the vector corresponding to the budget split, M is the total population size, and b/d the daily budget.

6.6.1.0.3 Direct Mailing A direct mailing marketer typically starts with a dataset of customer demographic information and purchase records, builds statistical models to predict *response rate*, and then ranks customers in descending order of response rate to send solicitations [224]. Our direct mailing simulator uses this strategy after randomly assigning a response rate to each customer from a uniform distribution on $[0, 2\bar{r}]$, where $\bar{r}$ is the pre-defined average response rate. In the simulation, an advertiser runs a direct mailing campaign weekly using an identical budget (i.e., the total budget is split equally among the 13 weeks comprising 3 months). In addition, we use a channel-specific conversion rate R_{dml} (probability of response to the ad).

6.6.1.0.4 Broadcast Marketing Simulator A marketer can also choose to advertise a product over a broadcast channel, such as TV or radio. In the simulation, we assume the marketer is facing J_b broadcast advertising options. Each option is represented by a tuple (c_j, r_j), with cost $c_j = e_j\bar{B}$ and response rate $r_j = \frac{1-e^{-\beta c_j}}{1+e^{-\beta c_j}}$ for each advertising option j, where, e_j is a random number drawn from a uniform distribution from 0 to 1, and $\bar{B}$ is the maximum budget allowed in the system. The distribution of r_j follows a *tanh* function with respect to c_j (essentially, a rescaled *logistic* function to ensure output is within [0,1] for positive values of cost c_j), with an exogenously specified coefficient β. The final parameter, conversion rate, is denoted by R_{brc}.

Similar to the other simulators, an advertiser divides the budget b equally into $w \leq 13$ weeks. Each week, with a budget b/w, the marketer chooses the best option (in terms of r_j). The expected response each week is the response rate multiplied by the market size. Finally, the marketer decides on an optimal duration of the marketing campaign in weeks, w^*.

6.6.2 Results

This section presents experimental results for 7 simulation configurations reflecting variability of relative channel efficacy for two channels, door-to-door marketing and keyword auctions. The parameter values are the transmission probability p in the former and conversion rate R_{oad} for the latter. The values of these for each configuration are given in Table 6.1.[3] Throughout, the conversion rates of direct mailing and broadcast marketing are 0.05 and 0.1 respectively.

configuration (ID)	door to door (p)	online ads (R_{oad})
0	0.1	0.05
1	0.08	0.05
2	0.12	0.05
3	0.12	0.03
4	0.12	0.07
5	0.12	0.01
6	0.14	0.05

Table 6.1: Parameter configurations.

Figure 6.1 shows the simulated payoffs as a function of budget for each channel for configuration 0, where thresholds correspond to the "steps", and each step-wise line is consists of the simulated payoffs for the corresponding budget. The step-wise output suggests the strong combinatorial nature of our optimized objective.

Figures 6.2 and 6.3 compare utility (overall rewards of the four channels) and running time of GBQ, HBQ and SA for different budgets, varying channel parameters of door-to-door marketing and keyword auction channels respectively. In all experiments, GBQ achieves the highest payoffs, but at a significant computational cost. Notably, HBQ is typically by far the most efficient in terms of computation time, and achieves utility that is

[3]Experimental results regarding other parameters are provided in Appendix.

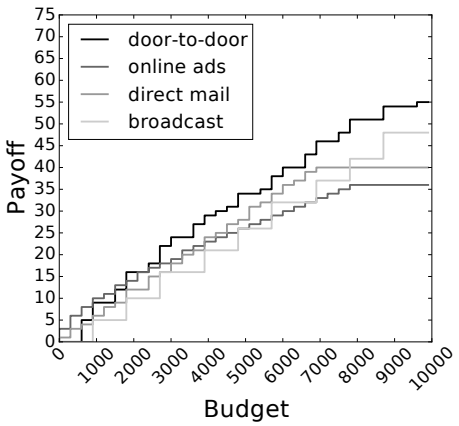


Figure 6.1: Simulated payoffs for Configuration 0 of four channels: door-to-door, online ads, direct mail and broadcast.

always nearly optimal. In addition, GBQ is typically significantly more time efficient than SA, which does eventually achieve a near-optimal utility as well, but after considerable computing time.

6.7 Conclusion

We presented a novel discrete optimization framework to address the budget allocation challenge faced in multi-channel marketing, where channel-wise payoffs exhibit strong budget complementarities. We showed that the budget optimization problem in our setting can be transformed into a well-known multi-choice knapsack problem, which can be solved effectively using state-of-the-art MILP solvers. We then introduced effective approximation and query schemes (GBQ and HBQ) when the response function of multi-channel marketing is represented using a simulation model, where weights of knapsack items can only be bounded through queries. We showed that the transformed (multi-choice) knapsack problem using the upper-bound knapsack weights is a tight 2-approximation to the optimum with exact thresholds, when the weights satisfy mild conditions. We implemented our framework in a simulated marketing platform motivated by real-world multi-channel marketing campaigns in a real geographical area. We conducted extensive experiments on

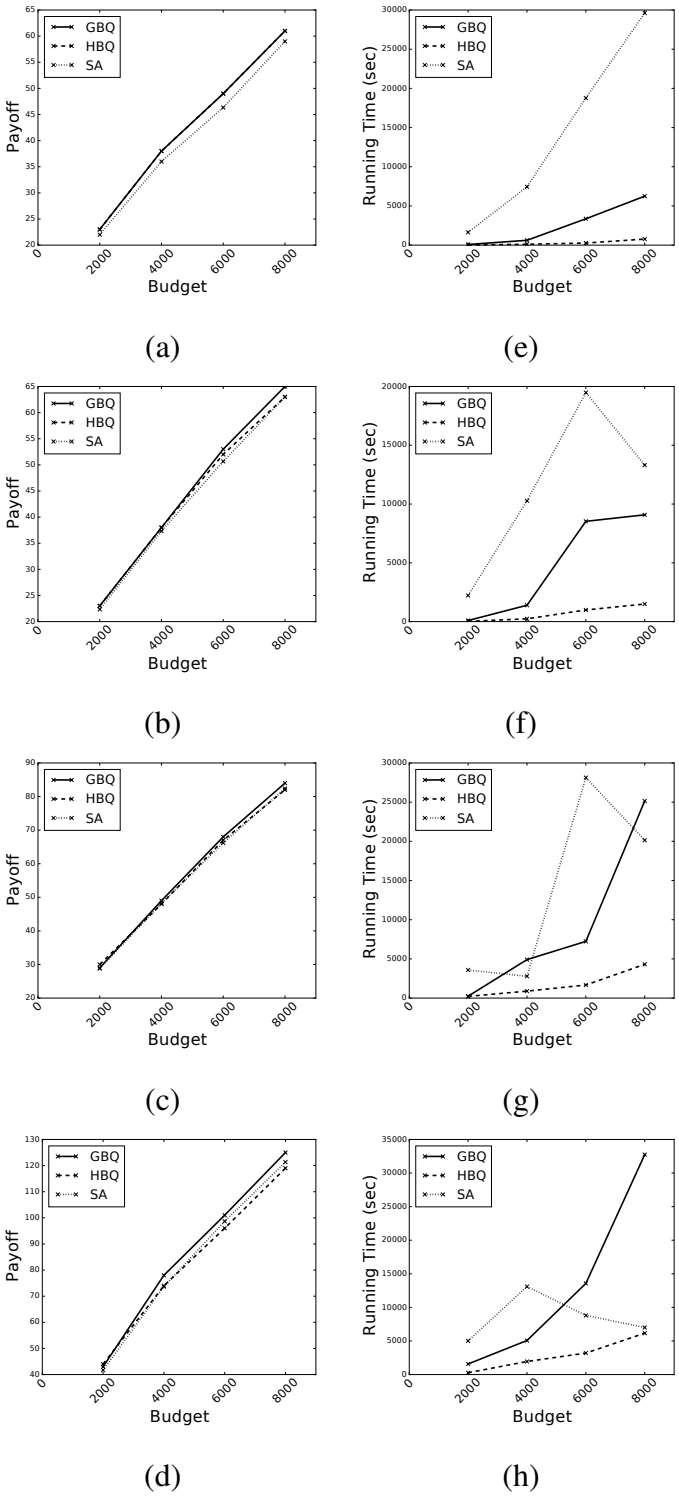


Figure 6.2: Payoff (a)-(d) & run time (e)-(h) comparison among algorithms over different budgets for different door-to-door marketing parameters. (a), (e) $p = 0.08$. (b), (f) $p = 0.1$. (c), (g) $p = 0.12$. (d), (h) $p = 0.14$.

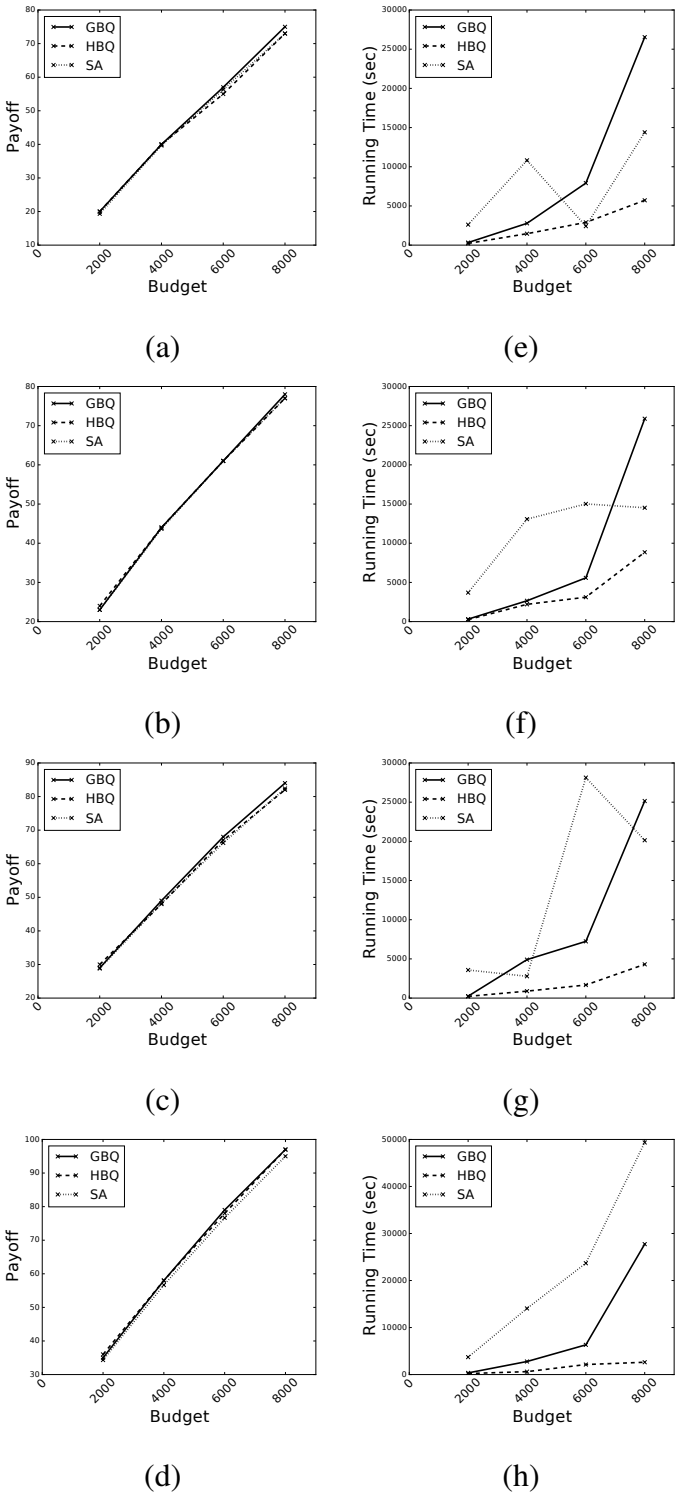


Figure 6.3: Payoff (a)-(d) & run time (e)-(h) comparison among algorithms over different budgets for different online ads marketing parameters. (a), (e) $R_{oad} = 0.01$. (b), (f) $R_{oad} = 0.03$. (c), (g) $R_{oad} = 0.05$. (d), (h) $R_{oad} = 0.07$.

different marketing configurations and showed that the proposed query algorithms significantly outperform a simulated annealing baseline.

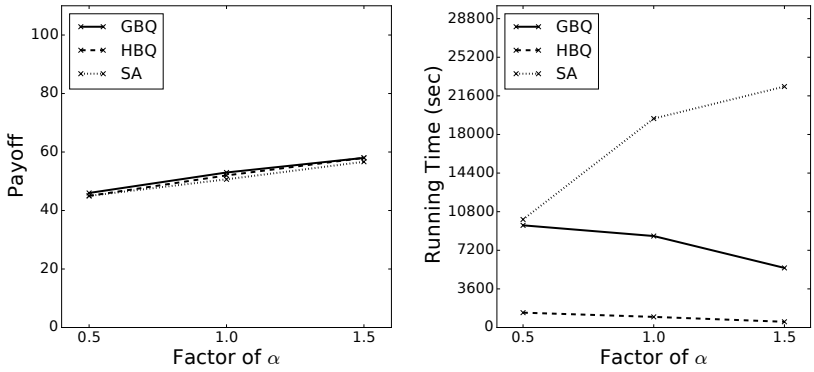
APPENDIX

Additional Experiments

Apart from the experimental results presented in Section 6.2, we also conducted a series of experiments regarding other parameters of the marketing simulator. Particularly, we set the default configuration to Configuration 0 (see Table 1) with a fixed budget 6000, vary one parameter a time and then compare the performance of GBQ, HBQ and SA in terms of utility and running time.

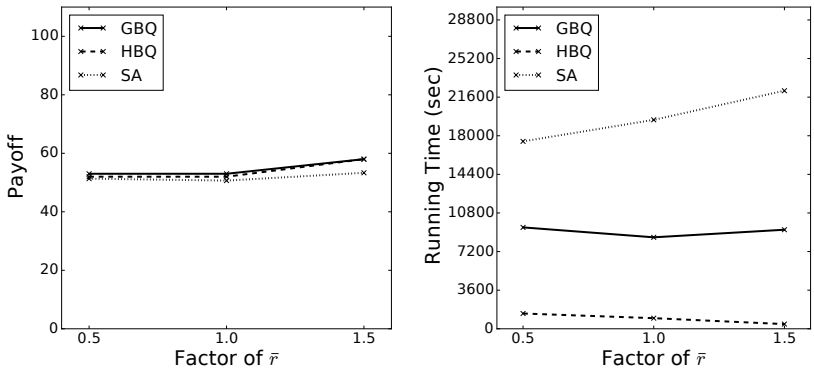
Figure 6.4 compares utility and running time of GBQ, HBQ and SA over different factors of α, which is a parameter defined in the keyword auction simulator. Note that when the factor of α equals to 1, it corresponds to default value of α in Configuration 0, whereas, factor of α equals to 0.5 and 1.5 are 0.5 times and 1.5 times of the default α respectively. Figure 6.5 compares utility and running time of all three algorithms over different factors of $\bar{r}$ for the direct mailing simulator. Similarly, factor of $\bar{r} = 1$ is the default setting. Figure 6.6 shows comparison among GBQ, HBQ and SA over different factors of β as defined in the broadcast marketing simulator, with factor of $\beta = 1$ the default value. Figure 6.7 displays comparison among GBQ, HBQ and SA over different options of R_{dml} for the direct mailing simulator. Figure 6.8 compares GBQ, HBQ and SA over different choices of R_{brc} for the broadcast marketing simulator.

Clearly, in all experimented cases, HBQ is the most efficient achieving competitive utilities significantly faster than GBQ, and both algorithms outperform the simulated annealing baseline.



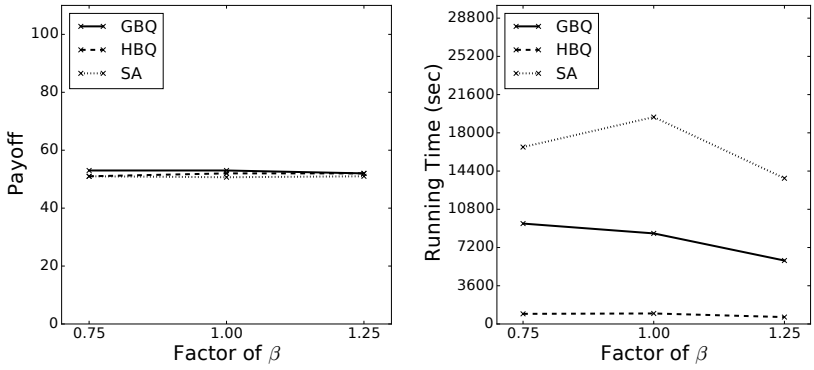
(a) Payoff (b) Run Time

Figure 6.4: Comparison among algorithms over different factors of α for online ads marketing.



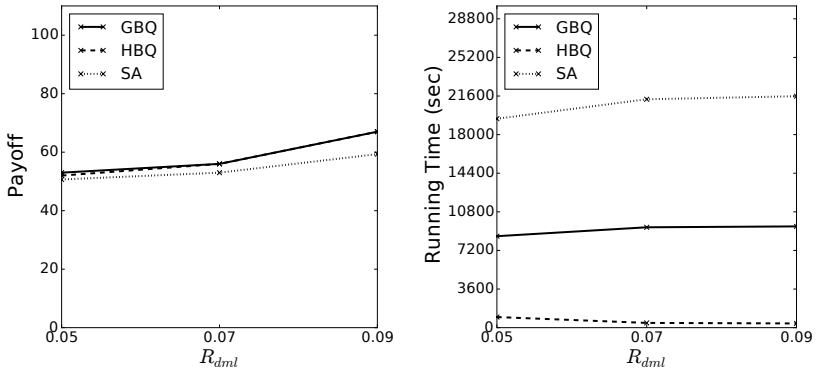
(a) Payoff (b) Run Time

Figure 6.5: Comparison among algorithms over different factors of $\bar{r}$ for direct mail marketing.



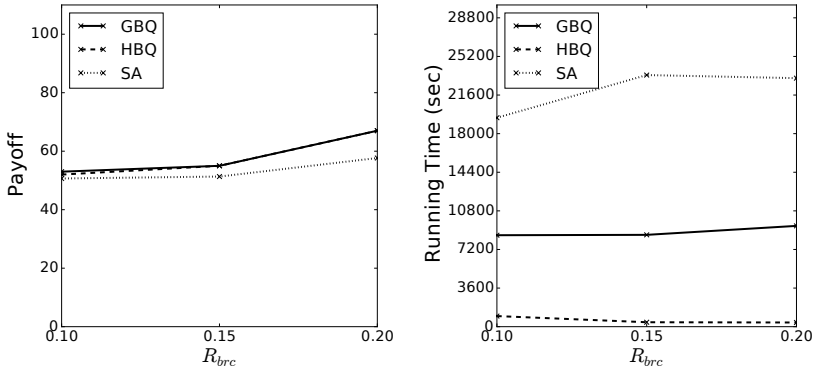
(a) Payoff (b) Run Time

Figure 6.6: Comparison among algorithms over different factors of β for broadcast marketing.



(a) Payoff (b) Run Time

Figure 6.7: Comparison among algorithms over different R_{dml} for direct mail marketing.



(a) Payoff (b) Run Time

Figure 6.8: Comparison among algorithms over different R_{brc} for broadcast marketing.

Chapter 7

Conclusions and Future Work

The general theory of innovation diffusion has been established for decades, however, modeling and simulating the diffusion process remains notoriously challenging. Lately, agent-based models (ABMs) have dominated traditional aggregate diffusion models, due to the remarkable advantage to capture individual heterogeneity and social and spatial interactions: the key to understanding complex systems. As for the study of innovation diffusion, empirical ABMs are particularly important to guide policy decisions. However, most work appears analytical and non-empirical. Our critical review of the empirically-grounded ABMs of innovation diffusion revealed that few such ABMs are calibrated properly, validated rigorously, and developed explicitly for prediction. This clearly limits their use in supporting decision-making in practice. The thesis contributed a rigorous data-driven agent-based modeling method to address the calibration and validation on massive, rich individual adoption data.

Generally, ABMs are used to answer "what-if" questions and draw insights on the efficacy of different policies, but, rarely provides *executable* and *quantitative* decisions. Mathematical optimization has been widely used to provide numerical solutions in many other domains. Interestingly, little work has coupled it with state-of-the-art ABMs. By solving marketing problems in several important settings, e.g., early adoption with increasing returns to scale, targeted marketing with routing constraints, multi-channel marketing with budget complementarities, the thesis demonstrated how efficient algorithms can help the design of intervention policies based on data-driven simulations, such as ABMs, providing optimal or near-optimal actionable plans for marketers.

Algorithmic marketing with data-driven simulations is a cutting-edge computational technique to model the common but highly complex diffusion process of innovations and

aid the design of effective marketing intervention policies. However, many challenges remain unresolved in order to build more credible and intelligent agent-based social simulations with application to innovation diffusion. In particular, *credible* data-driven ABMs require high accuracy, transparency, and efficiency; *intelligent* data-driven ABMs are expected to situate in the realistic marketing environment, actively learn from continuing new observations and adaptively update internal models. Thanks to the ever-expanding data availability and advances in data science, especially machine learning, now we are able to attack dynamics of innovation diffusion in a significantly more effective and novel way than our predecessors decades ago.

The optimization techniques combined with high-fidelity data-driven ABMs will also lead to the development of the *next-generation* marketing decision support systems. They are expected to be more usable by providing users with accurate predictions, meaningful insights, and actionable plans, and importantly more intelligent being able to automatically optimize marketing operations acting as agents (or delegates) for human marketers. Although our approach was originally proposed for marketing optimization, in the future, we plan to apply it to modeling and intervention design of innovations in other domains, such as medical and health-care, online communities.

BIBLIOGRAPHY

[1] Everett M. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, 2003.

[2] Haifeng Zhang and Yevgeniy Vorobeychik. Empirically grounded agent-based models of innovation diffusion: A critical review. 2016. URL http://arxiv.org/abs/1608.08517.

[3] Frank M Bass. Empirical generalizations and marketing science: A personal view. *Marketing Science*, 14(3_supplement):G6–G19, 1995.

[4] John Hauser, Gerard J Tellis, and Abbie Griffin. Research on innovation: A review and agenda for marketing science. *Marketing science*, 25(6):687–717, 2006.

[5] Vijay Mahajan, Eitan Muller, and Yoram Wind. *New-product diffusion models*, volume 11. Springer Science & Business Media, 2000.

[6] Renana Peres, Eitan Muller, and Vijay Mahajan. Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, 27(2):91–106, 2010.

[7] Roger Lewin. *Complexity: Life at the edge of chaos*. University of Chicago Press, 1999.

[8] John H Holland. *Hidden order: How adaptation builds complexity*. Basic Books, 1995.

[9] Nigel Gilbert and Klaus Troitzsch. *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[10] CM Macal and MJ North. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162, 2010.

[11] Rosanna Garcia and Wander Jager. From the special issue editors: Agent-based modeling of innovation diffusion*. *Journal of Product Innovation Management*, 28 (2):148–151, 2011.

[12] Robert M Axelrod. *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press, 1997.

[13] Joshua M Epstein. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60, 1999.

[14] Elmar Kiesling, Markus Günther, Christian Stummer, and Lea M Wakolbinger. Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research*, 20(2):183–230, 2012.

[15] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[16] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[17] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3): 211–223, 2001.

[18] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[19] Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with application to rooftop solar adoption. In *In-*

ternational Conference on Autonomous Agents and Multiagent Systems, pages 513–521, 2015.

[20] Satoru Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[21] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3:19, 2012.

[22] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

[23] Andreas Krause and Carlos Guestrin. A note on the budgeted maximization of submodular functions. *Technical Report, CMU-CALD-05-103*, 2005.

[24] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420–429. ACM, 2007.

[25] Arvind Rangaswamy and Gerrit H Van Bruggen. Opportunities and challenges in multichannel marketing: An introduction to the special issue. *Journal of Interactive Marketing*, 19(2):5–11, 2005.

[26] Scott A Neslin, Dhruv Grewal, Robert Leghorn, Venkatesh Shankar, Marije L Teerling, Jacquelyn S Thomas, and Peter C Verhoef. Challenges and opportunities in multichannel customer management. *Journal of Service Research*, 9(2):95–112, 2006.

[27] Scott A Neslin and Venkatesh Shankar. Key issues in multichannel customer man-

agement: current knowledge and future directions. *Journal of interactive marketing*, 23(1):70–81, 2009.

[28] P.K. Kannan, Werner Reinartz, and Peter C. Verhoef. The path to purchase and attribution modeling: Introduction to special section. *International Journal of Research in Marketing*, 33(3):449 – 456, 2016.

[29] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[30] Haifeng Zhang, Yevgeniy Vorobeychik, Joshua Letchford, and Kiran Lakkaraju. Data-driven agent-based modeling, with application to rooftop solar adoption. *Autonomous Agents and Multi-Agent Systems*, 30(6):1023–1049, 2016.

[31] Haifeng Zhang, Ariel D Procaccia, and Yevgeniy Vorobeychik. Dynamic influence maximization under increasing returns to scale. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 949–957, 2015.

[32] Haifeng Zhang and Yevgeniy Vorobeychik. Submodular optimization with routing constraints. In *AAAI Conference on Artificial Intelligence*, pages 819–825, 2016.

[33] Haifeng Zhang, Yevgeniy Vorobeychik, and Ariel D Procaccia. Multi-channel marketing with budget complementarities. In *International Conference on Autonomous Agents and Multiagent Systems*, 2017, to appear.

[34] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[35] Thomas W Valente. Network models and methods for studying the diffusion of innovations. *Models and methods in social network analysis*, 28:98, 2005.

[36] Bryce Ryan and Neal C Gross. The diffusion of hybrid seed corn in two iowa communities. *Rural sociology*, 8(1):15, 1943.

[37] Thomas W Thomas W Valente. *Network models of the diffusion of innovations*. Hampton Press, 1995.

[38] Thomas W Valente and Everett M Rogers. The origins and development of the diffusion of innovations paradigm as an example of scientific growth. *Science communication*, 16(3):242–273, 1995.

[39] Frank M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[40] Wallace J Hopp. Ten most influential papers of management science's first fifty years. *Management Science*, 50(12_supplement):1763–1763, 2004.

[41] Frank M Bass, Trichy V Krishnan, and Dipak C Jain. Why the bass model fits without decision variables. *Marketing science*, 13(3):203–223, 1994.

[42] Nigel Meade and Towhidul Islam. Modelling and forecasting the diffusion of innovation–a 25-year review. *International Journal of forecasting*, 22(3):519–545, 2006.

[43] Deepa Chandrasekaran and Gerard J Tellis. A critical review of marketing research on diffusion of new products. *Review of marketing research*, 3(1):39–80, 2007.

[44] Vijay Mahajan, Eitan Muller, and Frank M Bass. New product diffusion models in marketing: A review and directions for research. *The journal of marketing*, pages 1–26, 1990.

[45] Michael W Macy and Robert Willer. From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology*, pages 143–166, 2002.

[46] Robin B Matthews, Nigel G Gilbert, Alan Roach, J Gary Polhill, and Nick M Gotts. Agent-based land-use models: a review of applications. *Landscape Ecology*, 22(10): 1447–1459, 2007.

[47] Rosanna Garcia. Uses of agent-based modeling in innovation/new product development research*. *Journal of Product Innovation Management*, 22(5):380–398, 2005.

[48] Ashkan Negahban and Levent Yilmaz. Agent-based simulation applications in marketing research: an integrated review. *Journal of Simulation*, 8(2):129–142, 2014.

[49] Herbert Dawid. Agent-based models of innovation and technological change. *Handbook of computational economics*, 2:1235–1272, 2006.

[50] Paul Windrum, Giorgio Fagiolo, and Alessio Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007.

[51] Charles M Macal. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10(2):144–156, 2016.

[52] Thomas Berger. Agent-based spatial models applied to agriculture: a simulation tool for technology diffusion, resource use changes and policy analysis. *Agricultural economics*, 25(2-3):245–260, 2001.

[53] Thomas Berger, Regina Birner, Nancy Mccarthy, JosÉ DíAz, and Heidi Wittmer. Capturing the complexity of water uses and water users within a multi-agent framework. *Water Resources Management*, 21(1):129–148, 2007.

[54] Pepijn Schreinemachers, Thomas Berger, and Jens B Aune. Simulating soil fertility and poverty dynamics in uganda: A bio-economic multi-agent systems approach. *Ecological economics*, 64(2):387–401, 2007.

[55] Pepijn Schreinemachers, Thomas Berger, Aer Sirijinda, and Suwanna Praneetvatakul. The diffusion of greenhouse agriculture in northern thailand: Combining econometrics and agent-based modeling. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 57(4):513–536, 2009.

[56] Everett Rogers. Diffusion of innovations. *New York*, page 12, 1995.

[57] Pepijn Schreinemachers, Chakrit Potchanasin, Thomas Berger, and Sithidech Roygrong. Agent-based modeling for ex ante assessment of tree crop innovations: litchis in northern thailand. *Agricultural Economics*, 41(6):519–536, 2010.

[58] Peter Alexander, Dominic Moran, Mark DA Rounsevell, and Pete Smith. Modelling the perennial energy crop market: the role of spatial diffusion. *Journal of The Royal Society Interface*, 10(88):20130656, 2013.

[59] Peter Alexander, Dominic Moran, Pete Smith, Astley Hastings, Shifeng Wang, Gilla Sünnenberg, Andrew Lovett, Matthew J Tallis, Eric Casella, Gail Taylor, et al. Estimating uk perennial energy crop supply using farm-scale models with spatially disaggregated data. *GCB Bioenergy*, 6(2):142–155, 2014.

[60] Albert Faber, Marco Valente, and Peter Janssen. Exploring domestic micro-cogeneration in the netherlands: an agent-based demand model for technology diffusion. *Energy Policy*, 38(6):2763–2775, 2010.

[61] G Sorda, Y Sunak, and Reinhard Madlener. An agent-based spatial simulation to evaluate the promotion of electricity from agricultural biogas plants in germany. *Ecological Economics*, 89:43–60, 2013.

[62] Thijs LJ Broekhuizen, Sebastiano A Delre, and Anna Torres. Simulating the cinema market: How cross-cultural differences in social influence explain box office distributions. *Journal of Product Innovation Management*, 28(2):204–217, 2011.

[63] M Günther, C Stummer, LM Wakolbinger, and M Wildpaner. An agent-based simulation approach for the new product diffusion of a novel biomass fuel. *Journal of the Operational Research Society*, pages 12–20, 2011.

[64] Georg Holtz and Claudia Pahl-Wostl. An agent-based model of groundwater over-exploitation in the upper guadiana, spain. *Regional Environmental Change*, 12(1): 95–121, 2012.

[65] Patrick Plötz, Till Gnann, and Martin Wietschel. Modelling market diffusion of electric vehicles with real world driving datapart i: Model structure and validation. *Ecological Economics*, 107:411–421, 2014.

[66] Daire McCoy and Seán Lyons. Consumer preferences and the influence of networks in electric vehicle diffusion: An agent-based microsimulation in ireland. *Energy Research & Social Science*, 3:89–101, 2014.

[67] Johannes Palmer, Giovanni Sorda, and Reinhard Madlener. Modeling the diffusion of residential photovoltaic systems in italy: An agent-based simulation. *Technological Forecasting and Social Change*, 99:106–131, 2015.

[68] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[69] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[70] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

[71] Guillaume Deffuant, Sylvie Huet, JP Bousset, J Henriot, Georges Amon, Gérard Weisbuch, et al. Agent based simulation of organic farming conversion in allier département, 2002.

[72] Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation*, 5(4), 2002.

[73] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[74] Peter Kaufmann, Sigrid Stagl, and Daniel W Franks. Simulating the diffusion of organic farming practices in two new eu member states. *Ecological Economics*, 68 (10):2580–2593, 2009.

[75] Jason Noble, Simon Davy, and Daniel W Franks. Effects of the topology of social networks on information transmission. In *International Conference on Simulation of Adaptive Behavior*, pages 395–404, 2004.

[76] Nina Schwarz and Andreas Ernst. Agent-based modeling of the diffusion of environmental innovationsan empirical approach. *Technological forecasting and social change*, 76(4):497–511, 2009.

[77] Bertha Maya Sopha, Christian A Klöckner, and Edgar G Hertwich. Adoption and diffusion of heating systems in norway: coupling agent-based modeling with empirical research. *Environmental Innovation and Societal Transitions*, 8:42–61, 2013.

[78] Wander Jager. *Modelling consumer behaviour*. Rijksuniversiteit Groningen, 2000.

[79] Varun Rai and Scott A Robinson. Agent-based modeling of energy technology adoption: empirical integration of social, behavioral, economic, and environmental factors. *Environmental Modelling & Software*, 70:163–177, 2015.

[80] Thorben Jensen, Georg Holtz, Carolin Baedeker, and Émile JL Chappin. Energy-efficiency impacts of an air-quality feedback device in residential buildings: an agent-based modeling assessment. *Energy and Buildings*, 116:151–163, 2016.

[81] Volker Grimm, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310(5750):987–991, 2005.

[82] Ingo Wolf, Jochen Nuss, Tobias Schröder, and Gerhard de Haan. The adoption of electric vehicles: An empirical agent-based model of attitude formation and change. In *Conference of the European Association for Social Simulation*, pages 93–98, 2012.

[83] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[84] Wander Jager, MA Janssen, HJM De Vries, J De Greef, and CAJ Vlek. Behaviour in commons dilemmas: Homo economicus and homo psychologicus in an ecological-economic model. *Ecological economics*, 35(3):357–379, 2000.

[85] Malte Schwoon. Simulating the adoption of fuel cell vehicles. *Journal of Evolutionary Economics*, 16(4):435–472, 2006.

[86] Marco A Janssen and Wander Jager. Stimulating diffusion of green products. *Journal of Evolutionary Economics*, 12(3):283–306, 2002.

[87] Ramón Briegel, Andreas Ernst, Sascha Holzhauer, Daniel Klemm, Friedrich Krebs, and Aldo Martínez Piñánez. *Social-ecological modelling with LARA: A psychologically well-founded lightweight agent architecture*. PhD thesis, International Environmental Modelling and Software Society (iEMSs), 2012.

[88] Friedrich Krebs, Sascha Holzhauer, and Andreas Ernst. Modelling the role of neighbourhood support in regional climate change adaptation. *Applied Spatial Analysis and Policy*, 6(4):305–331, 2013.

[89] Friedrich Krebs and Andreas Ernst. A spatially explicit agent-based model of the diffusion of green electricity: Model setup and retrodictive validation. In *Social Simulation Conference*. European Social Simulation Association, 2015.

[90] Andreas Ernst and Ramón Briegel. A dynamic and spatially explicit psychological model of the diffusion of green electricity across germany. *Journal of Environmental Psychology*, 2016.

[91] Oscar Van Vliet, Bert De Vries, André Faaij, Wim Turkenburg, and Wander Jager. Multi-agent simulation of adoption of alternative fuels. *Transportation Research Part D: Transport and Environment*, 15(6):326–342, 2010.

[92] Jiayun Zhao, Esfandyr Mazhari, Nurcin Celik, and Young-Jun Son. Hybrid agent-based simulation for policy evaluation of solar power generation systems. *Simulation Modelling Practice and Theory*, 19:2189–2205, 2011.

[93] Shintae Kim, Keeheon Lee, Jang Kyun Cho, and Chang Ouk Kim. Agent-based diffusion model for an automobile market with fuzzy topsis-based product adoption process. *Expert Systems with Applications*, 38(6):7270–7276, 2011.

[94] Rosanna Garcia, Paul Rummel, and John Hauser. Validating agent-based marketing models through conjoint analysis. *Journal of Business Research*, 60(8):848–857, 2007.

[95] Kathleen M Carley. Validating computational models. *Paper available at http://www. casos. cs. cmu. edu/publications/papers. php*, 1996.

[96] Andras Vag. Simulating changing consumer preferences: a dynamic conjoint model. *Journal of Business Research*, 60(8):904–911, 2007.

[97] Ting Zhang, Sonja Gensler, and Rosanna Garcia. A study of the diffusion of al-

ternative fuel vehicles: An agent-based modeling approach. *Journal of Product Innovation Management*, 28(2):152–168, 2011.

[98] Jeremy J Michalek, Panos Y Papalambros, and Steven J Skerlos. A study of fuel efficiency and emission policy impact on optimal vehicle design decisions. *Journal of Mechanical Design*, 126(6):1062–1070, 2004.

[99] Timothy Lee, Runming Yao, and Phil Coker. An analysis of uk policies for domestic energy reduction using an agent based tool. *Energy Policy*, 66:267–279, 2014.

[100] Christian Stummer, Elmar Kiesling, Markus Günther, and Rudolf Vetschera. Innovation diffusion of repeat purchase products in a competitive market: an agent-based simulation approach. *European Journal of Operational Research*, 245(1):157–167, 2015.

[101] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173–187, 1999.

[102] Peter L Knepell and Deborah C Arangno. *Simulation validation: a confidence assessment methodology*, volume 15. John Wiley & Sons, 1993.

[103] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[104] José M Galán, Adolfo López-Paredes, and Ricardo Del Olmo. An agent-based model for domestic water management in valladolid metropolitan area. *Water resources research*, 45(5), 2009.

[105] Andrei Borshchev and Alexei Filippov. From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. In *International Conference of the System Dynamics Society*, volume 22. Citeseer, 2004.

[106] Elenna R Dugundji and László Gulyás. Sociodynamic discrete choice on networks in space: impacts of agent heterogeneity on emergent outcomes. *Environment and Planning B: Planning and Design*, 35(6):1028–1054, 2008.

[107] Martino Tran. Agent-behaviour and network influence on energy innovation diffusion. *Communications in Nonlinear Science and Numerical Simulation*, 17(9): 3682–3695, 2012.

[108] Zhanli Sun and Daniel Müller. A framework for modeling payments for ecosystem services with agent-based models, bayesian belief networks and opinion dynamics models. *Environmental modelling & software*, 45:15–28, 2013.

[109] Márton Karsai, Gerardo Iñiguez, Kimmo Kaski, and János Kertész. Complex contagion process in spreading of online innovation. *Journal of The Royal Society Interface*, 11(101):20140694, 2014.

[110] Márton Karsai, Gerardo Iñiguez, Riivo Kikas, Kimmo Kaski, and János Kertész. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific reports*, 6, 2016.

[111] William Rand, Jeffrey Herrmann, Brandon Schein, and Nea Vodopivec. An agent-based model of urgent diffusion in social media. *Journal of Artificial Societies and Social Simulation*, 18(2):1, 2015. ISSN 1460-7425.

[112] Michael Trusov, William Rand, and Yogesh V. Joshi. Improving prelaunch diffusion forecasts: Using synthetic networks as simulated priors. *Journal of Marketing Research*, 50(6):675–690, 2013.

[113] Manuel Chica and William Rand. Building agent-based decision support systems for word-of-mouth programs. a freemium application. *Journal of Marketing Research*, 2017.

[114] W. Rand and R.T. Rust. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193, 2011.

[115] Forrest Stonedahl and William Rand. *When Does Simulated Data Match Real Data? Comparing Model ¿ Calibration Functions using Genetic Algorithms*, pages 297–313. Springer Japan, Tokyo, 2014.

[116] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.

[117] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *International World Wide Web Conference*, pages 519–528. ACM, 2012.

[118] Francesco Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011.

[119] Paulo Shakarian, Abhinav Bhatnagar, Ashkan Aleali, R Guo, and E Shaabani. *Diffusion in Social Networks*. Springer (in press), 2015.

[120] Kazumi Saito, Kouzou Ohara, Yuki Yamagishi, Masahiro Kimura, and Hiroshi Motoda. Learning diffusion probability based on node attributes in social networks. In *Foundations of Intelligent Systems*, pages 153–162. Springer, 2011.

[121] Adrien Guille and Hakim Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *International World Wide Web Conference*, pages 1145–1152. ACM, 2012.

[122] Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. *WOSN*, 10:3–11, 2010.

[123] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028. ACM, 2010.

[124] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146. ACM, 2003.

[125] Seth Myers and Jure Leskovec. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems*, pages 1741–1749, 2010.

[126] M Gomez Rodriguez, D Balduzzi, B Schölkopf, Getoor T Scheffer, et al. Uncovering the temporal dynamics of diffusion networks. In *International Conference on Machine Learning*, pages 561–568. International Machine Learning Society, 2011.

[127] Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.

[128] Manuel Gomez Rodriguez, Jure Leskovec, and Bernhard Schölkopf. Structure and dynamics of information pathways in online media. In *ACM International Conference on Web Search and Data Mining*, pages 23–32. ACM, 2013.

[129] Xiaorong Xiang, Ryan Kennedy, Gregory Madey, and Steve Cabaniss. Verification and validation of agent-based scientific simulation models. In *Agent-Directed Simulation Conference*, pages 47–55, 2005.

[130] Giorgio Fagiolo, Paul Windrum, and Alessio Moneta. Empirical validation of agent-based models: A critical survey. Technical report, LEM Working Paper Series, 2006.

[131] Paul Ormerod and Bridget Rosewell. Validation and verification of agent-based models in the social sciences. *Epistemological Aspects of Computer Simulation in the Social Sciences*, pages 130–140, 2009.

[132] Daniel G Brown, Scott Page, Rick Riolo, Moira Zellner, and William Rand. Path dependence and the validation of agent-based spatial models of land use. *International Journal of Geographical Information Science*, 19(2):153–174, 2005.

[133] Jerry Banks. *Handbook of simulation: principles, methodology, advances, applications, and practice*. John Wiley & Sons, 1998.

[134] JPC Kleijen. Validation of models: statistical techniques and data availability. In *Winter Simulation Conference*, volume 1, pages 647–654. IEEE, 1999.

[135] SM Sanchez. Abc's of output analysis. In *Winter Simulation Conference*, volume 1, pages 30–38. IEEE, 2001.

[136] Robert Axtell, Robert Axelrod, Joshua M Epstein, and Michael D Cohen. Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory*, 1(2):123–141, 1996.

[137] J. Friedman, T. Hastie, , and R. Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

[138] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[139] Matteo G Richiardi, Roberto Leombruni, Nicole J Saam, and Michele Sonnessa. A common protocol for agent-based social simulation. *Journal of artificial societies and social simulation*, 9, 2006.

[140] Volker Grimm, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K Heinz, Geir Huse, et al. A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, 198(1):115–126, 2006.

[141] CPUC. California solar initiative program handbook, 2013.

[142] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(Supp 3):7280–7287, 2002.

[143] John H. Miller and Scott E. Page. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, 2007.

[144] Paul Denholm, Easan Drury, and Robert Margolis. The solar deployment system (SolarDS) model: Documentation and sample results. Technical report, National Renewable Energy Laboratory, 2009.

[145] Scott A Robinson, Matt Stringer, Varun Rai, and Abhishek Tondon. Gis-integrated agent-based model of residential solar pv diffusion. In *32nd USAEE/IAEE North American Conference*, pages 28–31, 2013.

[146] Jan C. Thiele, Winfried Kurth, and Volker Grimm. Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R. *Journal of Artificial Societies and Social Simulation*, 17(3), 2014.

[147] Garrett M. Dancik, Douglas E. Jones, and Karin S. Dorman. Parameter estimation and sensitivity analysis in an agent-based model of leishmania major infection. *Journal of Theoretical Biology*, 262(3):398–412, 2011.

[148] Marco A Janssen and Elinor Ostrom. Empirically based, agent-based models. *Ecology and Society*, 11(2):37, 2006.

[149] Thomas Berger and Pepijn Schreinemachers. Creating agents and landscapes for multiagent systems from random samples. *Ecology and Society*, 11(2):19, 2006.

[150] Marco GA Huigen, Koen P Overmars, and Wouter T de Groot. Multiactor modeling of settling decisions and behavior in the san mariano watershed, the philippines:

a first application with the mameluke framework. *Ecology and Society*, 11(2):33, 2006.

[151] Kathrin Happe, Konrad Kellermann, and Alfons Balmann. Agent-based analysis of agricultural policies: an illustration of the agricultural policy simulator agripolis, its adaptation and behavior. *Ecology and Society*, 11(1):49, 2006.

[152] Daniel G Brown and Derek T Robinson. Effects of heterogeneity in residential preferences on an agent-based model of urban sprawl. *Ecology and society*, 11(1): 46, 2006.

[153] Marco A Janssen and Toh-Kyeong Ahn. Learning, signaling, and social preferences in public-good games. *Ecology and society*, 11(2):21, 2006.

[154] Andrea Borghesi, Michela Milano, Marco Gavanelli, and Tony Woods. Simulation of incentive mechanisms for renewable energy policies. In *European Conference on Modeling and Simulation*, 2013.

[155] Scott A Robinson and Varun Rai. Determinants of spatio-temporal patterns of energy technology adoption: An agent-based modeling approach. *Applied Energy*, 151: 273–284, 2015.

[156] Ruben Lobel and Georgia Perakis. Consumer choice model for forecasting demand and designing incentives for solar technology. Working paper, 2011.

[157] Brian Bollinger and Kenneth Gillingham. Peer effects in the diffusion of solar photovoltaic panels. *Marketing Science*, 31(6):900–912, 2012.

[158] Arthur van Benthem, Kenneth Gillingham, and James Sweeney. Learning-by-doing and the optimal solar policy in california. *Energy Journal*, 29(3):131–151, 2008.

[159] Michael Kearns and Jennifer Wortman. Learning from collective behavior. In *Conference on Learning Theory*, 2008.

[160] Stephen Judd, Michael Kearns, and Yevgeniy Vorobeychik. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences*, 107(34):14978–14982, 2010.

[161] Quang Duong, Michael P Wellman, Satinder Singh, and Yevgeniy Vorobeychik. History-dependent graphical multiagent models. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 1215–1222. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[162] Paul Torrens, Xun Li, and William A. Griffin. Building agent-based walking models by machine-learning on diverse databases of space-time trajectory samples. *Transactions in GIS*, 15(s1):67–94, 2011.

[163] Michael Wunder, Siddharth Suri, and Duncan J Watts. Empirical agent based models of cooperation in public goods games. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 891–908. ACM, 2013.

[164] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 199–208, 2009.

[165] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

[166] P. A. Geroski. Models of technology diffusion. *Research Policy*, 29(4):603–625, 2000.

[167] K.U. Rao and V. Kishore. A review of technology diffusion models with special reference to renewable energy technologies. *Renewable and Sustainable Energy Reviews*, 14(3):1070–1078, 2010.

[168] Varun Rai and Ben Sigrin. Diffusion of environmentally-friendly energy technologies: buy versus lease differences in residential pv markets. *Environmental Research Letters*, 8(1):014022, 2013.

[169] P. Zhai and E.D. Williams. Analyzing consumer acceptance of photovoltaics (pv) using fuzzy logic model. *Renewable Energy*, 41:350–357, 2012.

[170] N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, 2nd edition, 1981.

[171] Kenneth J. Arrow. The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173, 1962.

[172] C. Harmon. Experience curves of photovoltaic technology. Technical report, International Institute for Applied Systems Analysis, 2000.

[173] A. McDonald and L. Schrattenholzer. Learning rates for energy technologies. *Energy Policy*, 29(4):255–261, 2001.

[174] M.J. North, N.T. Collier, J. Ozik, E. Tatara, M. Altaweel, C.M. Macal, M. Bragen, and P. Sydelko. Complex adaptive systems modeling with repast simphony. In *Complex Adaptive Systems Modeling*. Springer, 2013.

[175] S. Morris. Contagion. *Review of Economic Studies*, 67, 2000.

[176] Duncan Watts. A simple model of global cascades in random networks. *Proceedings of the National Academy of Sciences*, 99:5766–5771, 2002.

[177] Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In Noam Nissan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 24, pages 613–632. Cambridge University Press, 2007.

[178] Bo Liu, Gao Cong, Dong Xu, and Yifeng Zhang. Time constrained influence maximization in social networks. In *International Conference on Data Mining*, pages 439–448, 2012.

[179] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[180] H. Zhang, Y. Vorobeychik, J. Letchford, and K. Lakkaraju. Predicting rooftop solar adoption using agent-based modeling. In *AAAI Fall Symposium on Energy Market Prediction*, 2014.

[181] Joseph Farrell and Garth Saloner. Standardization, compatibility, and innovation. *Rand Journal of Economics*, 16(1):70–83, 1985.

[182] L.M.B. Cabral. On the adoption of innovation with network externalities. *Mathematical Social Sciences*, 19:299–308, 1990.

[183] Emmanuelle Auriol and Michel Benaim. Standardization in decentralized economies. *American Economic Review*, 90(3):550–570, 2000.

[184] Jacomo Corbo and Yevgeniy Vorobeychik. The effects of quality and price on adoption dynamics of competing technologies. In *International Conference on Information Systems*, 2009.

[185] Dong Li, Zhi-Ming Xu, Nilanjan Chakraborty, Anika Gupta, Katia Sycara, and Sheng Li. Polarity related influence maximization in signed social networks. *PLoS ONE*, 9(7):e102199, 2014.

[186] Huiyuan Zhang, Thang N. Dinh, and My T. Thai. Maximizing the spread of positive influence in online social networks. In *International Conference on Distributed Computing Systems*, pages 317–326, 2013.

[187] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

[188] Samir Khuller, Anna Moss, and Joseph Seffi Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[189] Maxim Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004.

[190] Fumei Lam and Alantha Newman. Traveling salesman path problems. *Mathematical Programming*, 113(1):39–59, 2008.

[191] Daniel J Rosenkrantz, Richard E Stearns, and Philip M Lewis, II. An analysis of several heuristics for the traveling salesman problem. *SIAM journal on computing*, 6(3):563–581, 1977.

[192] Nicos Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, DTIC Document, 1976.

[193] Yale T Herer. Submodularity and the traveling salesman problem. *European journal of operational research*, 114(3):489–508, 1999.

[194] Colleen Alkalay-Houlihan and Adrian Vetta. False-name bidding and economic efficiency in combinatorial auctions. In *AAAI Conference on Artificial Intelligence*, pages 538–544, 2014.

[195] Bruce L Golden, Larry Levy, and Rakesh Vohra. The orienteering problem. *Naval Research Logistics (NRL)*, 34(3):307–318, 1987.

[196] Pieter Vansteenwegen, Wouter Souffriau, and Dirk Van Oudheusden. The orienteering problem: A survey. *European Journal of Operational Research*, 209(1):1–10, 2011.

[197] Chandra Chekuri and Martin Pal. A recursive greedy algorithm for walks in directed graphs. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 245–253. IEEE, 2005.

[198] Amarjeet Singh, Andreas Krause, Carlos Guestrin, William J Kaiser, and Maxim A Batalin. Efficient planning of informative paths for multiple robots. In *IJCAI*, volume 7, pages 2204–2211, 2007.

[199] Rishabh K Iyer and Jeff A Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. In *Advances in Neural Information Processing Systems*, pages 2436–2444, 2013.

[200] Rishabh Krishnan Iyer. *Submodular Optimization and Machine Learning: Theoretical Results, Unifying and Scalable Algorithms, and Applications*. PhD thesis, University of Washington, 2015.

[201] Michele Conforti and Gérard Cornuéjols. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete applied mathematics*, 7(3):251–274, 1984.

[202] Rishabh K Iyer, Stefanie Jegelka, and Jeff A Bilmes. Curvature and optimal algorithms for learning and minimizing submodular functions. In *Advances in Neural Information Processing Systems*, pages 2742–2750, 2013.

[203] Avrim Blum, Shuchi Chawla, David R Karger, Terran Lane, Adam Meyerson, and Maria Minkoff. Approximation algorithms for orienteering and discounted-reward tsp. In *Annual IEEE Symposium on Foundations of Computer Science*, pages 46–55. IEEE, 2003.

[204] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444. ACM, 2013.

[205] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[206] P ERDdS and A R&WI. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[207] Alan K Graham and Carlos A Ariza. Dynamic, hard and strategic questions: using optimization to answer a marketing resource allocation question. *System Dynamics Review*, 19(1):27–46, 2003.

[208] Adil Bagirov, Napsu Karmitsa, and Marko M Mäkelä. *Introduction to Nonsmooth Optimization: theory, practice and software*. Springer, 2014.

[209] Yanwu Yang, Jie Zhang, Rui Qin, Juanjuan Li, Fei-Yue Wang, and Wei Qi. A budget optimization framework for search advertisements across markets. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(5):1141–1151, 2012.

[210] Naoki Abe, Naval Verma, Chid Apte, and Robert Schroko. Cross channel optimized marketing by reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–772. ACM, 2004.

[211] Nikolay Archak, Vahab Mirrokni, and S Muthukrishnan. Budget optimization for online advertising campaigns with carryover effects. In *Ad Auctions Workshop*, 2010.

[212] Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

[213] Haifeng Zhang and Yevgeniy Vorobeychik. Submodular optimization with routing constraints. In *AAAI Conference on Artificial Intelligence*, pages 819–825, 2016.

[214] Prabhakant Sinha and Andris A Zoltners. The multiple-choice knapsack problem. *Operations Research*, 27(3):503–515, 1979.

[215] David Pisinger. A minimal algorithm for the multiple-choice knapsack problem. *European Journal of Operational Research*, 83(2):394–410, 1995.

[216] M.E. Dyer, W.O. Riha, and J. Walker. A hybrid dynamic programming/branch-and-bound algorithm for the multiple-choice knapsack problem. *Journal of Computational and Applied Mathematics*, 58(1):43–54, 1995.

[217] Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[218] Mhand Hifi, Hedi Mhalla, and Slim Sadfi. Sensitivity of the optimum to perturbations of the profit or weight of an item in the binary knapsack problem. *Journal of Combinatorial Optimization*, 10(3):239–260, 2005.

[219] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

[220] Mustafa Ç Pínar. A note on robust 0-1 optimization with uncertain cost coefficients. *4OR*, 2(4):309–316, 2004.

[221] Michele Monaci and Ulrich Pferschy. On the robust knapsack problem. *SIAM Journal on Optimization*, 23(4):1956–1982, 2013.

[222] Marc Goerigk, Manoj Gupta, Jonas Ide, Anita Schöbel, and Sandeep Sen. The robust knapsack problem with queries. *Computers & Operations Research*, 55:12–22, 2015.

[223] Dimitris Bertsimas, John Tsitsiklis, et al. Simulated annealing. *Statistical science*, 8(1):10–15, 1993.

[224] Indranil Bose and Xi Chen. Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, 195(1):1–16, 2009.